2022

Tackling Big Data Variety using Metadata

Vranopoulos, Georgios

http://hdl.handle.net/10026.1/20099

http://dx.doi.org/10.24382/474

University of Plymouth

COPYRIGHT STATEMENT

UNIVERSITY OF PLYMOUTH

Tackling Big Data Variety using Metadata

By

Georgios E. Vranopoulos

A thesis submitted to the University of Plymouth in partial

fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Engineering, Computing & Mathematics

December 2022

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Dr. Shirley Atkinson and Prof. Nathan Clarke, for their guidance and support throughout this research. Many thanks to my wife and daughter, Kyriaki and Natalia, for their support and patience while pursuing this initiative. Last but not least I am grateful to my friends, with special thanks to Dr. Athanasios Triantafyllidis, for their moral support during challenging days.

AUTHOR'S DECLARATION

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Doctoral College Quality Sub-Committee.

Work submitted for this research degree at the University of Plymouth has not formed part of any other degree either at the University of Plymouth or at another establishment.

Publications (or public presentation of creative research outputs):

Georgios Vranopoulos, Nathan Clarke, Shirley Atkinson (2022). Addressing Big Data Variety Using an Automated Approach for Data Characterization. *Springer Journal of Big Data*, vol. 9, no. 8, doi:10.1186/s40537-021-00554-3. https://doi.org/10.1186/s40537-021-00554-3.

Georgios Vranopoulos, Nathan Clarke, Shirley Atkinson (2022). Big Data Confidentiality: An Approach toward Corporate Compliance using a Rule-Based System. Under Review.

Word Count of main body of thesis: 35,069

Signature :

Date : 21-Dec-2022

ABSTRACT

Georgios E. Vranopoulos

Tackling Big Data Variety using Metadata

The creation of new knowledge from manipulating and analysing existing knowledge is one of the primary objectives of any cognitive system. Out of the Big Data governing Vs, namely Volume, Velocity, Variety, Veracity, Validity, Volatility and Value, the first three are considered the primary ones. Most of the effort on Big Data research has been focussed upon Volume and Velocity, while Variety, "the ugly duckling" of Big Data, is often neglected and difficult to solve. A principal challenge with Variety is being able to understand and comprehend the data in gaining insight. Organisations have been investing in analytics relying on internal and external data to gain a competitive advantage. However, the legal and regulatory acts imposed nationally and internationally have become a challenge.

The approach focuses on the use of self-learning systems that will enable automatic compliance of data against regulatory requirements along with the capability of generating valuable and readily usable metadata towards data classification. While for data confidentiality, a framework that utilises algorithmic classification and workflow capabilities is proposed. Such a rule-based system, implementing the corporate data classification policy, will minimise the risk of exposure by facilitating users to identify the approved guidelines and enforce them quickly.

Two experiments towards confidential data identification and data characterisation were conducted in evaluating the feasibility of the approach. The focus of the experiments was to confirm that repetitive manual tasks can be automated, thus reducing the focus of a

Data Scientist on data identification and thereby providing more focus towards the extraction and analysis of the data itself. In addition to that, a survey with subject matter experts, a diverse audience of academics and senior business executives in the fields of security and data management, was conducted featuring and evaluating a working prototype. The proof-of-concept showcased the model's capabilities and provided a hands-on experience for expert to better understand the proposal.

The experimental work confirmed that: a) the use of algorithmic techniques attributed to the substantial decrease in false positives regarding the identification of confidential information; b) evidence that the use of a fraction of a data set, along with statistical analysis and supervised learning is sufficient in identifying the structure of information within it; c) the model for corporate confidentiality is viable and the proposed features of the system are of value.

With this proposal, the issues of understanding the nature of data can be mitigated, enabling a greater focus on meaningful interpretation of the heterogeneous data, while at the same time the organisations can secure their data and confirm data confidentiality and compliance.

TABLE OF CONTENTS

Copyright Statement i
Acknowledgements i
Author’s Declaration i
Abstract i
Table of Contents iii
Table of Figures v
Table of Tables vii
Table of Abbreviations / Acronyms ix
Chapter 1. Introduction 1
1.1. Evidence to the Problem 1
1.2. Big Data in Brief 4
1.3. The Variety Barrier 8
1.4. Research Aims and Objectives 19
1.5. Thesis Overview 21
Chapter 2. Big Data 24
2.1. Historical Review 24
2.2. Defining Big Data 26
2.3. The V’s Theory 31
2.4. Tools & Techniques 35
2.5. Business and Societal Perspective 43
2.6. Sector Analysis 55
2.7. Conclusion 61
Chapter 3. Dealing with Big Data Vs 63
3.1. Introduction 63
3.2. The Review Approach 65
3.3. Current State 72
3.4. Conclusion 91
Chapter 4. A Novel Approach to Tackling Variety 92
4.1. Introduction 92
4.2. Problem Definition 92
4.3. The Novel Proposal 94
4.4. Summary 110
Chapter 5. Optimising Confidential Data Identification - The “Booster Metrics” ... 112
5.1. Introduction 112
5.2. Experimental Methodology 113
5.3. The IS Overview 116

5.4. Configuration 117
5.5. Process Data Set / File(s) 123
5.6. Results 129
5.7. Conclusion 131
Chapter 6. Automating The Dataset Characterisation 133
6.1. Introduction 133
6.2. Experimental Methodology 134
6.3. IS Overview 140
6.4. Configuration 141
6.5. Process Data Set / File(s) 143
6.6. Results 148
6.7. Conclusion 161
Chapter 7. The “Confidentiality Preservation System (BD-CPS)” 163
7.1. Introduction 163
7.2. The Big Data – Confidentiality Preservation System (BD-CPS) 164
7.3. The Prototype 174
7.4. Evaluation 192
7.5. Conclusion 201
Chapter 8. Conclusion 202
8.1. Research Achievements 202
8.2. Research Limitations 204
8.3. Future Directions 205
References i
Appendix 1
I. Publications 1
II. IT Milestones over the Ages 2
III. Big Data Programming Environments 5
IV. DFD, ERD, ETL and NF 7
V. “Booster Metrics” Configuration sample XML 18
VI. Dataset Characterisation sample JSON configuration 26
VII. Information Form 27
VIII. Consent Form 29
IX. BD-CPS PoC Entities Data Samples 30
X. BD-CPS PoC Additional Approval Journeys 39
XI. BD-CPS PoC Design UI/UX Elements 40

TABLE OF FIGURES

Figure 1. DB Challenges presented....6
Figure 2. Google Trends....15
Figure 3. The initial 3 v's....31
Figure 4. All 7 V's....33
Figure 5. MapReduce Framework....38
Figure 6. MapReduce Execution Overview....39
Figure 7. Hadoop Solutions Q1 '14....42
Figure 8. Organisational Culture and the 3v's....46
Figure 9. Companies with at least 100 terabytes of stored data....55
Figure 10. Types of Data Stored....56
Figure 11. An Integrated view of Big Data....76
Figure 12. Proposition Steps/Stages....97
Figure 13. Confidentiality Process Flow....102
Figure 14. Columnar log file sample (3 columns)....102
Figure 15. Delimiter Determination Process Flow....104
Figure 16. Data Confidentiality Stage....112
Figure 17. "Booster Metrix" high-level process....116
Figure 18. Data Set Element Processing....124
Figure 19. Element Validations/Checks....125
Figure 20. Windowed/Block Processing....126
Figure 21. "Card" & "List" Processing....127
Figure 22. Absolute XML Processing....128
Figure 23. Relative XML Processing....128
Figure 24. Data format & Break Down Identification Stages....133
Figure 25. Experimental data composition concerning Origin, Volume and Size....135
Figure 26. The experimental configuration of the ANN....138
Figure 27. Training performance graphs....139
Figure 28. Neural Network Result Adjustment....139
Figure 29. Dataset Characterisation high-level process....140
Figure 30. File processing....144
Figure 31. File content processing....145
Figure 32. Statistics Calculations....146
Figure 33. Pre-Processor Functions....147
Figure 34. Neural Network Results in Cross-Tabulation Heat Map sample (1 out of 30)....154
Figure 35. Comparative analysis of 1%-10% of error for 1%-3% of Lines Read....159
Figure 36. Aggregated Set Results Heat Map....160
Figure 37. ANN functions Differentiation phasing out plot....161
Figure 38. BD-CPS overview....166
Figure 39. Action Strength Calculation....172
Figure 40. Web & Mobile Application....177
Figure 41. System Parameterisation....180
Figure 42. Entities Edit & Approval process....180
Figure 43. Data Set Elements Actions Selection UI....182
Figure 44. BD-CPS ERD....187
Figure 45. BD-CPS Web Interface Screens Flow....189
Figure 46. BD-CPS Mobile Interface Screens Flow....190
Figure 47. Business Entity Approval screens flow....191
Figure 48. Data set approval screens flow....192

Figure 49. Rating heat map 197
Figure 50. IDC's Digital Universe Study 3
Figure 51. ETL Processing Framework 16

TABLE OF TABLES

Table 1. Concept Matrix 67
Table 2. Aggregated Concept Matrix 67
Table 3. Big Data References Classification –Aggregated Concept Matrix 68
Table 4. Credit Sample RegEx 99
Table 5. Personally Identifiable Information Sample RegEx 100
Table 6. Classification to "booster metrics" association 101
Table 7. Data Set Characteristics 113
Table 8. Sample Metrics Definition 114
Table 9. General Configuration Parameters 118
Table 10. “Card” Classifier Configuration Parameters 119
Table 11. “List” Classifier Configuration Parameters 120
Table 12. “XML tag” Classifier Configuration Parameters 121
Table 13. Block/Windowed Reader Configuration Parameters 122
Table 14. Command-line parameters 122
Table 15. Results file structure 123
Table 16. False Positive percentages 130
Table 17. Additional Hits using XML tags 130
Table 18. File Lines % 135
Table 19. Dataset Lines No Classification 136
Table 20. Delimiters Confidence Level Weights 136
Table 21. Metrics Formulas 137
Table 22. Block/Windowed Reader Configuration Parameters 142
Table 23. Command-line parameters 142
Table 24. FileInfoOutputFileName structure 143
Table 25. StatsOutputFileName structure 143
Table 26. Data Sets File Content Classification 149
Table 27. File Encoding Classification 149
Table 28. Datasets Execution Timing 150
Table 29. Datasets Execution trend lines 151
Table 30. Meta-Heat Map for 10% Mismatch of the Neural Network Results 157
Table 31. DB-CPS Functional Requirements/Characteristics & Business Aims 165
Table 32. Terms Dictionary 167
Table 33. Combined Entity Minimum Levels Algorithms 170
Table 34. Combined Entity Algorithm example 170
Table 35. Interviewees Mix 193
Table 36. Rating & Open-Ended Questions 195
Table 37. DFD Notations 8
Table 38. Features in Data Models 11
Table 39. ERD Notations 12
Table 40. Most Common ERD Cardinality Links Representations 13

TABLE OF ABBREVIATIONS / ACRONYMS

Acronym	**Description / Definition**
AaaS	Application as a Service
ACID	Atomicity, Consistency Isolation and Durability
AD	Anno Domini, which is Latin for "year of our Lord"
AI	Artificial Intelligence
AICPA	American Institute of Certified Public Accountants
AML	Anti-Money Laundering
ANN	Artificial Neural Network
API	Application Programming Interface
ASG	SQL Access Group
ATM	Automatic Teller Machine
BASEL	The banking supervision Accords (recommendations on banking regulations)—Basel I, Basel II and Basel III—issued by the Basel Committee on Banking Supervision (BCBS). They are called the Basel Accords as the BCBS maintains its secretariat at the Bank for International Settlements in Basel, Switzerland and the committee normally meets there. The Basel Accords is a set of recommendations for regulations in the banking industry.
BC	Before Christ
BCNF	Boyce-Codd Normal form
BDaaS	Big Data as a Service
BI	Business Intelligence
BIN	Bank Identification Number
CAD	Computer-Aided Design
CAM	Computer-Aided Manufacturing
CAP	Theorem states that with consistency, availability and partition only two can be optimised at any time
CASE	Computer-Aided Software Engineering
CICA	Canadian Institute of Charted Accountants
CIO	Chief Information Officer
COBIT®	COBIT® framework was developed by ISACA® and the latest iteration was released in 2012
CODASYL	Conference on Data Systems Language
COO	Chief Operating Officer
CORBA	Common Object Request Broker Architecture
DaaS	Data as a Service
DBA	DataBase Administrator
DBMS	Database Management System
DCOM	Distributed Component Object Model
DFD	Data Flow Diagram
DFS	Distributed File Systems
DLP	Data Loss Prevention
DNA	Deoxyribonucleic acid
DW	Data Warehouse
EDD	Enterprise Data Dictionary
EDI	Electronic Data Interchange
EME	Enterprise Metadata Environment

Acronym	Description / Definition
ER	Entity Relationship
ERD	Entity Relationship Diagram
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load process
FRBCS	Fuzzy Rule Based Classification Systems
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
GIS	Geographic Information Systems
GPS	Global Positioning System
GUI’	Graphical User Interface
IaaS	Infrastructure as a Service
IBAN	International Bank Account Number
IMS	Information Management Systems
IoT	Internet of Things
IPR	Intellectual Property Rights
ISACA®	Information Systems Audit and Control Association, an independent, non-profit, global association, which engages in the development, adoption and use of globally accepted, industry-leading knowledge and practices for information system
ISAM	Indexed Sequential Access Method
JSON	JavaScript Object Notation
KM	Knowledge Management
MDM	Master Data Management
MiFID	Markets in Financial Instruments Directive
MIME	Multipurpose Internet Mail Extensions
ML	Machine Learning
MME	Managed Metadata Environments
MVCC	Multi-Version Concurrency Control
NF	Normal Form
NLP	Natural Language Processing
NoSQL	Not Only SQL
NSA	United States National Security Agency
ODBC	Open DataBase Connectivity
OLAP	OnLine Analytical Processing
OODBMS	Object Oriented Database Management System
PaaS	Platform as a Service
PCI	Payment Card Industry
PDA	Personal Digital Assistant
PII	Personally Identifiable Information
POC	Proof Of Concept
POS	Point Of Sale
PSI	Public-Sector Information
RAD	Rapid Application Development
RDBMS	Relational Database Management System
RegEx	Regular Expression
ROI	Return On Investment
SaaS	Software as a Service
SEQUEL	Structured English QUEry Language

Acronym	Description / Definition
SGML	Standardised Generalised Markup Language
SN	Social Networks
SNS	Social Network Site
SSADM	Structured Systems Analysis and Design Method
SWOT	Strengths, Weaknesses, Opportunities and Threats Analysis
TCO	Total Cost of Ownership
UI	User Interface
VSAM	Virtual Storage Access Method
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

1.1. Evidence to the Problem

Since 2010 data creation and subsequently storage and processing have been growing with exponential rates (Bedi et al., 2014; Kalambe et al., 2015; Kaur Sandhu. Amanpreet, 2022; Krawczyk et al., 2015; L. Zhang, 2014). In fact, in 2010, the world produced more than 1ZB of data (Villars et al., 2011). EMC, with research and analysis by IDC, claim in their 2014 report that the digital universe, which much like the physical universe is vast and in which bits are as many as the stars, is growing with a yearly rate of 40%, expecting to reach 44 zettabytes (44 trillion gigabytes) by 2020, while it is expected to double in size from 2022 to 2026 (Gantz & Reinsel, 2012, 2013; Sivarajah et al., 2017). The datasphere is expected to reach 175 zettabytes by 2025 (Reinsel et al., 2018; Rydning, 2022). This explosion of data led to the development of new technologies and concepts, summarised under the term “Big Data,” that came from large enterprises like Yahoo, Google, Facebook, LinkedIn etc., in their attempt to analyse large amounts of data (Devakunchari, 2014).

The first definition of the term, although at the time named “3D Data Management,” was provided by Doug Laney. Laney identified three integral parts in his definition, namely Volume, Velocity, and *Variety* (Laney, 2001). Technology advancements have mainly addressed volume and velocity, the first two attributes/capabilities (Kumar, 2013; Young et al., 2021). Distributed computing and cost reductions in storage aid companies in coping with large data sets, but the diverse types of data is the biggest challenge for data scientists (McElhenny, 2014). *Variety* is the different manifestation by which data are added to the digital universe. Indicative examples would include devices (IoT) that share the same information in different formats (JSON, XML, fixed-length etc.), comprehensive data sets that will include data sources extending to unstructured or semi-structured data like video, email, pictures etc.

Mastering data *Variety* rather than just coping with volume will be the main focus in attaining valuable insight through Big Data initiatives (Baker, 2015). Data analysts spend most of their time preparing data instead of mining them for business incentives (Fayyad & Uturusamy, 2002; Warden, 2011). Volume and analytics are addressed in several research and implementation projects; nonetheless, little effort has been attributed to making the data available for analysis (Mao et al., 2015). The task remains time-consuming and labour-intensive (Assunção et al., 2014; Lassoued et al., 2021). Data analysts are assisted in mining data with resources concerning distributed storage and parallel processing. However, they are left "on their own" when populating the data repository and their automation expectation is far greater than the current availability, thus making data *Variety* increasingly daunting and costly to wrestle (Baker, 2015; D. Wang et al., 2021).

In a Paradigm4 survey entitled "Leaving Data on the Table: New Survey Shows Variety, Not Volume, is the Bigger Challenge of Analysing Big Data," 91% of the respondents are either managing or plan to kick off a significant data initiative, and 71% highlighted *Variety* being more of a problem than Volume (E. Brown, 2014). It is not simply an academic research area but an actual drawback in business acceptance. Big Data cannot attribute to benefit realisation unless *Variety* is tackled, and this is identified not only by scholars but also by business IT experts (E. Brown, 2014; Jensen et al., 2021; Kumar, 2014; Shacklett, 2014).

Limited, if any, technological tools are available in addressing the issue of *Variety*, the proliferation of data from many sources, internal and external, public and private and in numerous formats (Kimura, 2014). Research is limited and actual implementations scarce in respect to software solutions with automated procedures for analysis and documentation that will minimise the effect of *Variety*. Unless people's skills, particularly analytical and synthesis data scientists skills, are detached from efforts in normalising data, business acceptance will remain low, and the full potential of Big Data will not be realised (Kumar, 2013). Further evidence to the lack of the appropriate technology for coping with *Variety* is Gartner's survey, "Big Data Drives Rapid Challenges in Infrastructure and $232 Billion in IT Spending Through 2016," stating that spending on services outweighs spending on software by a ratio of nine to one. Worldwide IT spending is projected to be $4.7 trillion in 2023 out of which $1.3 trillion (8.5%) will be allocated for IT Services (Rimol, 2022). Specifically for Big Data services, it is estimated that by 2026 the market will grow with a 32.3% increase rate to reach the size of $4.2billion (*Big Data Consulting Market: Market Size, Industry Outlook, Market Forecast, Demand Analysis ,Market Share, Market Report 2021-2026*, 2021; Kumar, 2013). The costs in services denote that tools are not available and that specialists are also scarce and thus well paid. Due to high costs, it is difficult for individual companies to have such experts in-house. Consultants will also impose high costs on any Big Data initiative. It seems that unless there are specific innovations in the "battle with Variety," companies are expected to invest heavily, internally by employing data analysts or externally by hiring consultants. These costs can diminish the expected net gain from any Big Data initiative, thus making *Variety* a vital cost factor for as long as it remains "tied" with people's skills. Once the tacit knowledge of consultants is recorded and depicted in a methodology alongside software implementation, in the form of a tool, it could lower costs, thus making Big Data more attractive for implementation.

1.2. Big Data in Brief

In understanding Big Data, basic concepts and definitions will be outlined. An in-depth analysis will be provided in the second chapter. At the same time, in this section, key definitions will be addressed in order to assist in understanding the complexity and turmoil posed by *Variety*.

1.2.1. The V's Definition

Based on Laney's first definition, it has become a trend to define Big Data in terms of V's. The initial approaches identified 3 V's by which characterised the term:

- Volume: Refers to the amount of data created and stored in the digital universe (Ali-Ud-Din Khan et al., 2014).
- Velocity: In Big Data environments, the speed of data change is quite high.
- **Variety**: This characteristic has to do with the data itself and the manifestations it can pertain to. Sensors, IoT, database records, video and audio have different formats and standards in addition to diverse communication protocols used to propagate the data.

Early approaches mainly were technical/technology oriented, but amendments were made to the V's theory as the concept gained business acceptance. V's with business essence were identified as follows:

- Veracity: identifying the lack of the required governance and homogeneity.
- Validity: validity is concerned with correctness and accuracy (Ali-Ud-Din Khan et al., 2014).

- Volatility: refers to how long are the data valid for and for how long they should be stored since it is a fact that storage resources, even with the use of clouds, are finite (Ali-Ud-Din Khan et al., 2014; L. Zhang, 2014; Zheng, 2006).
- Value: Business is concerned with income and realisation of competitive advantages; thus, bringing business value to the organisation poses a significant challenge.

These business concept extensions were incorporated in the V's theory to extend the technocratic viewport into a more business-related one. It is easier for corporate managers to understand terms like value and governance rather than velocity. In this way, business acceptance could be attained since the sheer technical innovation would not appeal to business implementation.

"Big Data" is defined in terms of V's. Technological advancements have enabled the use of low-cost commodity servers in utilising parallel computing, thus providing a cost-efficient horizontal expansion of processing power and storage.

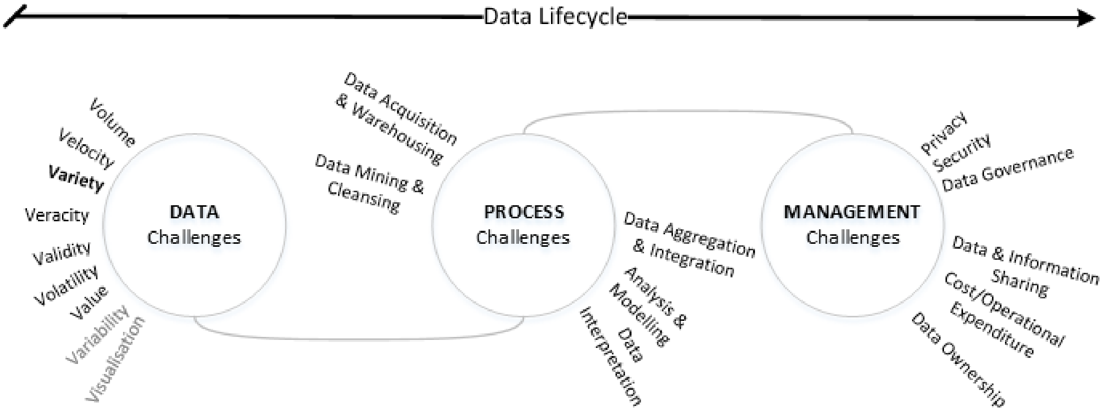
The Big Data ecosystem is identified, and several projects have been introduced in an attempt to bring the value of analytics one step further from the analysis of historical data, data warehouses (H. Jain & Gosain, 2012), in real-time predictive analytics towards "future prediction" (Kakish & Kraft, 2012; C.-N. Wang et al., 2022). The definition of Big Data and the technologies related to them are progressing and evolving; considerable research is done in the field; nonetheless, most of the effort is spent on Volume and Velocity and little work is done regarding *Variety* (Rui et al., 2015).

1.2.2. Big Data Challenges

In getting an initial understanding of the landscape of challenges that have emerged alongside the opportunities of Big Data, the classification by Sivarajah based on Akerkar and Zicari is presented (Corporation MarkLogic, 2012; Sivarajah et al., 2017).

As presented in Figure 1, three categories of challenges can be identified in the data lifecycle:

- Data challenges, are related to the basic characteristics of Big Data that have been presented and analysed in detail in 1.2.1
- Process challenges, address difficulties of handling data due to the complex nature posed by the Big Data environment.
- Management challenges, have to do with data privacy, governance, ownership, stewardship etc., covering the rules in handling the data. Another important aspect is the skillset and knowledge required to manipulate, analyse, and understand the respective data sets.



-

Source: Sivarajah et al., 2017

Figure 1. DB Challenges presented

Potential threats in fully utilising the wealth of Big Data could also include heterogeneity, scale, timelines, and personal privacy (Oguntimilehin & Ademola, 2014).

1.2.3. A Glimpse into the Technology Infrastructure

Having "defined" the term Big Data, the respective technologies utilised will be described, whilst a detailed presentation can be found in Chapter 2. The technology employed is primarily targeted in "cheaper," "bigger," and "faster," leaving a technological void in respect to *Variety*. The following paragraphs will present a non-exhaustive enumeration of the existing technologies targeting the "main 3 V's." Through this short presentation, it can be observed that *Variety* is the last and probably least addressed V, although addressed to some extent.

Towards Volume and Velocity, the processing power of the underlying infrastructure had to be fortified. To this end, and in an attempt to minimise investment, the need to harness the power of parallel computing by utilising commodity servers became a necessity. In attaining this goal and successfully distributing data and processing, several Distributed File Systems (DFS) were implemented (see 2.4). In this initiative by large organisations like Google, Amazon, etc., *Variety* is "absent." What is of importance is to attain power in terms of processing power and disk storage space at a minimal cost.

Once the infrastructure to utilise parallel processing became available, data stores that harvested this new technology emerged. The NoSQL data stores (see 2.4), apart from utilising the infrastructure in respect to Volume and Velocity, can also be considered a first attempt to manage, through distributed computing and highly optimised append operations, *Variety* (Berg et al., 2007). These stores have loosely defined fields versus tables and fields of an RDBMS to accommodate the diverse data types. To that extent, they can cater to semi-structured or unstructured data storage. *Variety*, nonetheless, is not simply a storage problem; it has to do with quality, proliferation, outliers, context, coherence, interactivity and much more not addressed by the Data Stores infrastructure.

Techniques like Sharding and Map Reduce (see 2.4) are utilised in Big Data environments to utilise parallel processing for speed and distribution of data across nodes. In addition to that, the most popular implementations of Big Data frameworks (see 2.4) and programming environments (see Appendix III) provide no out-of-the-box tools for *Variety*.

1.3. The Variety Barrier

As previously mentioned, *Variety* posed one of the most important factors of Big Data and was identified from the infancy of the Big Data era. It was one of the V's in the first definition of the environment, and over the years, it has sustained its strength since it is identified in almost all subsequent definitions of the term over the years.

1.3.1. Variety Real Life Examples

Data representation, data formats, non-aligned data structures, inconsistent data semantics and many other terms are employed in defining *Variety* (Kaisler et al., 2013), but what is it all about in plain words? Some examples from real environments are outlined to understand *Variety*'s effect in day-to-day analytics.

In the wholesale industry, a wholesaler tries to analyse customers' orders and cross-reference them with the records obtained by the customers' sales systems. *Variety* can take the following formats:

- The part numbers used by the wholesaler and retailer are different, making it almost impossible to uniquely identify the products.
- Sales and orders are in different metrics; while sales are recorded in tens of thousands, orders are aggregated in hundreds of thousands.

- Sales are in the clients’ currency, supposing a web-store, whilst orders are in local currency.
- Different decimal points and date formats are utilised based on the clients' locality.
- Every retailer has agreed to disclose their data, but a common file structure is not identified; thus, several files have different formats, i.e. some are fixed length, others are delimited (of course with different delimiters), and some are self-documented - that is tagged files like XML or JSON.

A security department has diverse data needs; in this case, the *Variety* problem manifests through the different data formats. The data store should be able to accommodate for:

- Text data, for instance, records from the access control system outlining which card was used in opening which door.
- Video data, from surveillance, closed circuits, or public traffic circuits.
- Voice data from the telephony subsystem, i.e. calls for bomb threats etc.
- Relational Data, from the company’s employee registry to identify access cardholders.

In the health sector, many records are produced automatically from the respective microbiological equipment and scanners (magnetic, tomographic, C-scans, etc.). Here *Variety* can be identified in the following occurrences.

- Different types of data, some data are images (i.e. scans) whilst others are numeric (i.e. blood test results).
- Not all equipment uses the same scale to produce results; thus, the same metric, such as Na-Sodium or Triglycerides, can be reported in percentage (32%) or decimal values (0.32).

- Metadata concerning the patient may or may not be available depending on the machinery. In some cases, the patient Identification Document (ID) can be identified whilst other systems provide the reference number of the test.

The banking sector, in identifying the withdrawals made on an ATM, *Variety* is hidden in the following formats (Vranopoulos et al., 2016):

- When the ATM is offline, due to network connectivity, no record is recorded; thus, misleading records show that there were no withdrawals at that day or time interval.
- When the ATM is out of money, the withdrawal attempts are again not recorded; thus, the reading is misleading.
- Disputes by customers are recorded in other systems, thus making it almost impossible to adjust the withdrawals levels.
- Apart from the systems logs and the transaction logs, video feeds must also be cross-referenced in cases of disputes; thus, video and withdrawal records must be stored and timestamp referenced.

Semantic problems intensify *Variety*. Many companies try to get information from social media in an attempt to identify public opinion and increase retention rates.

- Not all social media disclose all information; thus, “the whole picture” is not present in some cases.
- It is usual practice for people to express their dislike more often than their satisfaction, in which case sentiment seems to be constantly “negative.” Lack of data can express “half-truth.”
- Every provider supplies its data stream in a different format, thus requiring different “agents” in acquiring the data.

- Natural language algorithms in identifying sentiments and sentiment analysis can be misleading and inaccurate, especially with not widely used languages.
- Outliers might indicate an error or a reality that must be addressed—identifying what, can cause “headache” to analysts.
- Cross-referencing public data, social media, internal records, and customer databases can be pretty tricky since it is difficult to uniquely identify customers based on their names.

The manufacturing industry, a company with several factories, tries to consolidate the requirements for raw materials in an attempt to gain from economies of scale through bulk purchases. *Variety* here is associated with:

- Not all factories use the same metrics in recording the consumption of materials.
- Depending on the factory's location, different legal requirements pose for different safety levels, which are depicted in the reorder levels but are not documented.
- Seasonality effects are not documented; thus, cross-referenced orders depict high differences and rigorous fluctuations.
- Different machinery requires minor adjustments (i.e. texture, length, etc.) in the raw materials that are not documented since local suppliers are aware of the irregularities.
- Not all factories have complete records; thus, many materials show lower consumption rates from reality.
- In most cases, different part numbers are used based on the local suppliers' coding systems; thus, there are several part numbers for the same material.
- In some cases, the superintendent is responsible for manually entering the consumed materials, making the process vulnerable to human error.

Universal abstraction of a wide range of data types, even with the uses of multidimensional types, is a troublesome and effort-intensive task (Rui et al., 2015). Abstraction is a process of generalisation, and as such, it is efficient when aggregating information within a predefined and well-structured environment. When it comes to Big Data, the environment becomes larger; thus, generalisation needs to consider a multitude of factors that cannot be easily handled in many cases.

The sheer complexity of data sources, accurate and inaccurate data mixed together, and multiple formats and units of measurement, pose another risk in handling data within a Big Data implementation (Kumar, 2013). With the evolution of the internet and the widespread use of mobile devices, the magnitude of data has risen. Alongside the Volume and Velocity of data generated and need to be stored and analysed, the confidence of data has also declined. Since the environment is loosely coordinated, in contrast to a well-structured business environment, each application or device and eventually any data generator can provide data in forms convenient to the producer (programmer, manufacturer etc.) instead of the data consumer (data analyst). Thus, data can be erroneous and come in different formats and units of measurement. Erroneous, inaccurate, out-of-date or incomplete data may significantly affect any corporate database, including a Big Data lake (Mohamed & Noordin, 2011). Furthermore, inaccurate, out-of-date or incomplete data can significantly impact corporate status due to economic and social impacts that can affect brand loyalty (R. Y. Wang et al., 1995).

The proper business context is essential to ask the correct analytical questions to yield business value (Shacklett, 2014). Due to “siloed” systems, most data series utilised from the web or even within the organisation lack such context. Several systems can handle the same kind of information, e.g. there might be several purchasing systems to fit the exact need of different departments within an organisation. Putting data into context and generating information considering each department's “special needs” is of utmost importance. Machine learning algorithms are employed in “curating” the data (Shacklett, 2014). By “going out of the box,” that is, manipulating data series from systems outside the company, this effect is magnified, making indexing and unification of data a challenge on its own, and it is exhibited that diverse data types are the biggest challenge for data scientist causing them to “leave data on the table” (McElhenny, 2014).

Unstructured data can be troublesome since there is little knowledge in handling them (E. Brown, 2014). Most organisations, meaning systems and people, are accustomed to dealing with well-defined and well-structured data; when unstructured data “comes into play,” there is little corporate knowledge in handling them. Knowledge curves can be steep, and alongside resistance to change, there might be avoidance or misuse of such data, resulting in erroneous analysis.

Big Data repositories by design do not have fixed schemas to cater to *Variety*; nonetheless, this poses a challenge when keeping in sync the data schema and the respective application code (Cerqueus et al., 2015). Much of the “logic” concerning the schema is transferred from the database layer to the application layer. When developing on these principles and constant application upgrades occur due to environmental changes, there is a high probability of errors in case precise versioning is not available. It is not only the data but the way they are handled that can deliver erroneous results, especially when environment condition/context change.

Well established techniques for analysing data can fail or produce results of diminished confidence when the data are not validated. Noise and vagueness being inherent in available data can increase the uncertainty of extracted information, undermining the effectiveness of Fuzzy Rule-Based Classification Systems (FRBCSs), which are popular tools for pattern recognition and classification (del Río et al., 2015). It might be the case that this factor can be catered for by analysing a more extensive data set; nonetheless, it is of utmost importance to consider the assumptions of the analysis and document it appropriately.

From the above information, it can be inferred that too many things can "go wrong" when data is collected, aggregated and investigated upon. That is why *Variety* is a factor that cannot be easily addressed and solved programmatically (Kumar, 2013).

1.3.2. Limited Research

It is identified that most research is primarily done in the areas of distribution, either of storage or processing power, in an attempt to lower costs by distributing computing across inexpensive, redundant components (Kumar, 2013). Little work has been done in the proliferation of data (Rui et al., 2015).

A quick look in Google trends, which is a public facility by Google in identifying the aggregate occurrence of a search-term in relation to the total search-volume across the world, can confirm that *Variety* is less addressed in comparison to Volume and Velocity, nonetheless it seems to gain momentum (Lennard, 2014). The graph is primarily presented in visually depicting the substantial distance/gap between Volume and Variety.

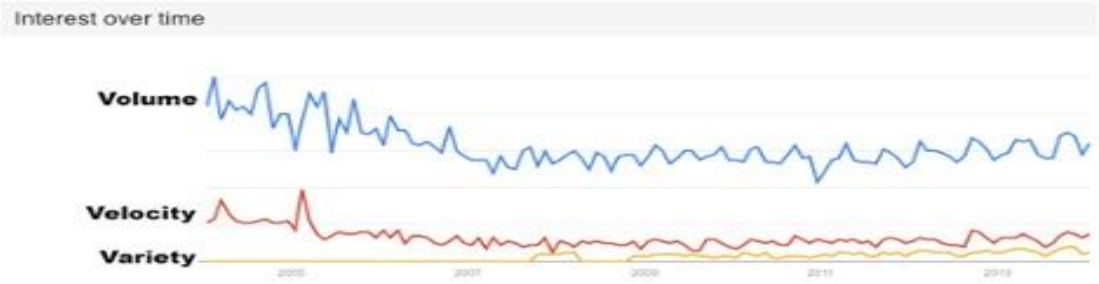


Figure 2. Google Trends

Another indicator is that many start-ups engage in the field of Big Data, but only a few are addressing *Variety*. Out of "The 10 coolest Big Data Startups of 2015" (Whiting, 2015), only one is engaged in "battling the evils of 'schema proliferation'." In its "Top Ten Most Funded Big Data Startups for 2015," Forbes features one company that specialises in *Variety* (GilPress, 2015). Although the company provides a product addressing *Variety,* on a closer look at the definition, "our engineers integrate and map all of the relevant source data" (Palantir, 2015), it is easily identified that it is more of a service rather than a tool. The same applies for companies featured in the top 13 and top 100 Big Data companies where, although they focus on data ingestion and the *Variety* challenges, they are offering services instead of a product as per their declarations; "IT staff augmentation" and "a skilled data management team" (*Top 100 Big Data Companies of 2022*, 2022; *Top 13 Best Big Data Companies of 2022*, 2022).

Research in academia and business can be identified as being limited; nonetheless, it is promising since there appears to be a trend towards addressing the problem. It can be argued that *Variety* gaining momentum is a consequence of the Big Data evolution. Technology infrastructures are prerequisites in order for *Variety* to become a concern. First, the data should reside somewhere e.g. distributed file systems. Then data must be accessible – parallel computing and newly evolved programming languages. Finally, data are analysed, and erroneous outcomes are identified, *welcome to the world of Variety…*

1.3.3. Other V’s Intensifiers

It is imperative to emphasise that the *Variety* challenge is intensified by Volume and Velocity when addressed in Big Data environments. Volume effect can be easily understood with the use of a simple example. It is convenient for an analyst to identify “outliers” in the distribution of the withdrawals of one ATM for one month, posed by hardware problems. Suppose the analysis is extended to all ATMs of a bank, ranging from a couple of hundred to several thousand ATMs depending on the bank's size, for a couple of years. In that case, it becomes an impossible task for the analyst. Time is also of the essence; Velocity can make the difference since data is produced in high frequencies and quickly becomes outdated. In a traditional environment utilising a relational database, it might take several weeks for administrators to alter/adjust an existing data warehouse schema to accommodate changes in data structures. A MapReduce solution might reduce this time to hours, thus permitting actual processing before data become outdated (Rehman et al., 2020; Villars et al., 2011).

It is of utmost importance to keep in mind the intensifying effect of Velocity and Volume when trying to understand and handle *Variety*. It might be the case that *Variety* is very low or even non-existent in a prototype with a data subset. However, when transferred to production, several problems might manifest due to the increase in Volume and Velocity.

1.3.4. The Indicators

An indicator of the complexity is that companies today spend 70%-80% of their time modelling and preparing data rather than interacting with the data in generating business insight (Baker, 2015). In “predicting,” the future companies rely on past data accumulated in discrete systems (silos). Then all information is transferred into the data warehouse, where with the employment of Extract, Transform, Load process (ETL)[1], data become available for further analysis. From that point onwards, context and data integration are usually done by data scientists specialised in each sector by employing their extensive, in-depth knowledge in the specific industry (Kumar, 2013).

As long as context and the definition of meaningful relations are tasks assigned primarily to humans, and *Variety* remains resilient to software solutions (Trader, 2014), scalability and universality will pose a challenge for Big Data. Since only by utilising the diverse types of data will businesses unlock and harness the enormous potential of analytics (McElhenny, 2014), it is crucial to enhance data scientists' capabilities with automated processes or a methodology in facilitating their initial task of data integration, enabling them to focus in the actual data analysis.

Machine learning and advanced algorithms that "curate" data across multiple sources into a single unified view can be utilised in an attempt to minimise human interaction (Shacklett, 2014). Classifying Neural Networks and other techniques can be employed to address the issue; nonetheless, the change in the mind-set of the problem will make the difference. As Big Data environments evolved in accommodating the shortcomings of RDBMSs, it is only logical that if the same approach is employed in utilising this new technology, the end result will be a hybrid in which technology "can provide" whilst processes fall back. The first step is already taken, from ETL used in the RDBMS environments to ELT utilised in Big Data infrastructures (Berisha et al., 2022; Michel et

[1] Extract, Transform, Load process. A variation suggested for Big Data implementation is known as ELT, where the Loading phase precedes the Transformation in order to harness the infrastructure capabilities in the later phase.

al., 2014). The following steps should lead to a new methodology that will harness the power of technology and address *Variety* and Velocity problems.

1.3.5. Where do we Stand?

“Big Data is here to stay” (Bughin et al., 2021; Conrad & Vault, 2014; Gregory, 2013; Miettinen & Tergujeff, 2021; Newell & Marabelli, 2015). Academia and business are embarking on a voyage in breaking the frontiers of data processing. The path is paved with failure and lost opportunities since more than 30% will never attain a competitive advantage (Gregory, 2013). In 2022 only a 20% of analytic insight is expected to produce any real business value (Reggio & Astesiano, 2020). Simply investing does not guarantee success (Gupta, 2014). However, as shown by Gartner and MIT studies, a substantial number of companies have started or are planning to start Big Data initiative (Gregory, 2013).

The main focus of evolvement is placed upon technological advancements to address the basic needs of storage and computing power. This trend is closely related to Total Cost Of Ownership (TCO)[2] per terabyte (TB), ranging from $20,000 to $100,000 in RDBMS. Big Data infrastructures like Hadoop range from $333 to $1,000 (Hogan & Jovanovic, 2015). Technology is becoming, by the day, more stable and trustworthy. Business implementations evolve from proofs-of-concept to actual production tools. The infrastructure is becoming available, and the tools hiding the underlying complexity are available for none-IT experts to handle such frameworks. Everything falls into place to identify and “battle” with a “forgotten” enemy, *Variety*. Standardisation and rigid

[2] TCO spans the entire product’s lifecycle and in many cases can be more accurate than ROI in determining the value of an investment (Bigelow, 2021).

structure of RDBMS environments have disguised *Variety*, but the whole new aspect of flexibility and openness of the Big Data ecosystem has revived it.

There is a need to address the essence of *Variety* and identify techniques and processes to offload the burden posed on the data scientist. Only then will they actually be able to gain insight on data and produce the long-sought competitive advantage that the companies can capitalize upon. A change in mind-set has been identified by adapting existing RDBMS technics into Big Data, i.e. from ETL to ELT. A trend is developing in more academic and business researchers tackling problems posed by *Variety*. Since RDMBS have a "solid ground" and have proven their worth for all kinds of business, it is only fair to review, transform and adapt as many proven techniques and process into Big Data.

1.4. Research Aims and Objectives

In the previous paragraphs, it has been identified that *Variety* poses a threat to any Big Data implementation. It was one of the initial V's that, in essence, define Big Data. There seems to be limited research compared to other V's. Primarily *Variety* cannot be resolved through a technological evolution like Volume, where technology advancements and new hardware could provide a solution. *Variety* is all about controversy and diversity, which is a complex enigma to solve. Just like in nature, almost all data elements are unique pertaining to a specific sector, company, department or even individual employee.

The aim of this research is to provide with robust novel framework that will cater to and facilitate the following:

- Efficient and effective confidential data identification and recording system using metadata in describing feeds and data lakes.
- An auto-classification mechanism that will:
    - Identify/classify data formats.
    - Identify the nature of data attributes by comparing them with existing data feeds.
    - Correlate new data series with existing data in the lake.
    - Perform primary (mean, median, max, min, standard deviation, etc.) and more complex (extrapolation, projection, etc.) statistical analysis to identify the metadata that can provide with useful insight in further processing.
- Easy web-enabled system to govern data in conformance to Data Loss Prevention (DLP):
    - Classify the information about business metrics.
    - Govern the process with automated checks and controls.
    - Record any assumptions made during data clearance.
    - Record the outcomes of the respective analysis.

The aims of this thesis will be addressed throughout the upcoming chapters. Initially the framework is presented and explained. Subsequently, the confidentiality and auto-classification mechanism are presented along with their respective quantitative experiments. Finally, the DLP governance system is presented along with a qualitative expert's evaluation.

The objective is to provide data scientists with “the framework,” being the novel contribution, and a tool that will formalise how analysis is performed and how data are disseminated to external or internal entities.

In this way, it is expected that the lead time required for a Data Scientist to perform an analysis will be drastically reduced since:

- Data Identification time will be reduced by having the engine identify basic features like format, structure, data types, naming conventions, etc.
- Data Profiling time will be reduced since initial statistical analysis and correlation to existing data will be readily available.
- Data Reusability will be enhanced since all information concerning prior usage and analysis of the data, along with assumptions and results, will be available and interrelated in a common interphase.
- Data Governance will be enhanced since there will be an impartial way of disseminating and proliferating data.

1.5. Thesis Overview

Evidence of the problem related to Big Data and *Variety* have been exhibited along with a brief review of the Big Data concepts and basic definitions. The challenges posed by *Variety* were enumerated and showcased. The objectives of the research towards data Identification, profiling, reusability, and governance have been illustrated and the review methodology explained.

The Big Data concepts (see Chapter 2) will be presented in detail focussing not only on technology and definition but also on the socioeconomic and business aspects. A historical review will provide context while the definition of Big Data and the V’s theory will provide with ecosystem understanding. The available tools and techniques related to Big Data implementations will be presented in describing the associated technology. The corporate, social, legal and ethical impact and interrelation with Big Data is showcased along with a sectoral analysis in understanding the magnitude of the extended ecosystem.

The prior art review is presented (Chapter 3) in understanding what has been done in the past and identify any gap that would lead to a new approach. The literature review epistemology and methodology for the research are presented and a time series analysis in respect to *Variety* is provided so that the current state is identified, and the challenges highlighted.

A novel approach (Chapter 4) in tackling *Variety* is presented which will be the bases of the research and the work done towards the previously identified challenges. The approach is focussing on the data ingestion process and a proposal that can be adopted in minimising the *Variety* effects.

Laboratory experiments related to confidential data identification (Chapter 5) and dataset characterisation (Chapter 6) will be presented along with their outcome and results. The first experiment is focusing on confidential data identification by introducing an optimization methodology and proving its effectiveness and efficiency. The second experiment is targeting the data characterisation and with the use of an algorithmic methodology is proving that dataset can be processed and adequately characterised. The efficiency and applicability of the methodology is proven based on the implemented approach.

The design and implementation of a prototype information system for confidentiality preservation (Chapter 7), will be showcased along with the results of a survey of experts in evaluating it. The system is designed to minimise the data disclosure risk an organisation might face by providing a standard methodology and hands-on approach. Data confidentiality is preserved using a structured way throughout the organisation and in a consistent and self-explanatory style.

In all work the practicality and feasibility of the approach will be confirmed from the experimental quantitative and survey qualitative analysis. By providing with a functional and viable paradigm the cumulative knowledge of the field will be advanced. In addition to the expansion of body of knowledge, the work is aiming to provide an implementable approach from which data scientists and data driven organisations can benefit from its simplicity and robustness in achieving better, quicker and secure data insights.

Before dwelling on the “Variety Problem,” the Big Data ecosystem will be addressed to understand the concepts, technology and terminology. In attaining a 360^{o} view, business, ethics and sociological effects and challenges will be addressed. The historical review and evolution of IT in respect to data management is showcased (Section 2.1) to understand the progression to the Big Data era. The Big Data basic concepts and definitions are showcased (Sections 2.2 and 2.3) along with the tools and techniques used (Section 2.4). Following the definitions, Big Data impact and interdependence is presented along two axis, a) the aspects of the organisation and corporate culture, the socioeconomic surrounding and legal framework are discussed (Section 2.5) and b) a corporate sectoral analysis where the size along with impact of Big Data per sector is identified (Section 2.6). Finally focus points in relation to the *Variety* challenge, in respect to the current Big Data business exposure, are outlined.

2.1. Historical Review

Historical events pose a fair insight whenever an attempt is made to understand the present or the future. Although other socioeconomic and business evolution factors might intensify or weaken the effect, historical milestones depict and can accurately outline a trend. In understanding why IT has embarked on the voyage leading to Big Data, some of the major milestones in the history of computing with primary emphasis on databases (*A Timeline of Database History*, 2015; *Database Systems: A Brief Timeline*, 2000; Boyd & Crawford, 2012) will be outlined.

Databases are the main focus since there is an analogy to Big Data. Databases evolved in order to process more data as opposed to simple files. The business shifted from single files into databases into storing transactional and historical data. In the beginning, the database adoption rate was limited, but with the development of tools and the increase in the stability of the infrastructure today, it has become a standard. Like databases, the Big Data era came about since data processing, and information requests are never enough compared to limited resources available (Komar et al., 2022; Trifu & Ivan, 2014).

In antiquity, people started to store information once writing was invented, more than 6000 years ago (Smitha, 1998). The actual revolution in data accumulation happened in the 1950s with the introduction of the commercial computer. In the decades to come, the database concepts were identified and evolved. Client development tools and productivity tools came into existence whilst another significant milestone is the invention of the internet and subsequently mobile technologies. *Appendix II. IT Milestones over the Ages* outlines the highlights of technological advancements which can be classified in eras, with Mainframe being the first and Mobile Devices and Web 2.0 being the last - Web 3.0 is "still happening."

What can be deducted from all previous eras is that, even in the case of leveraging Big Data, both IT investment and managerial innovation are required. The era of "Big Data" is here… (Boyd & Crawford, 2012). It can be seen that there is momentum in implementing Big Data environments, although Big Data technologies are still in their infancy (O'Driscoll et al., 2013). A Gartner study showed that out of 720 companies 460 have already invested in Big Data. In addition to that an MIT study on the "Fortune 500 companies" showed that 85% have launched, or have planned for, a Big Data initiative (Gregory, 2013; Mikalef et al., 2020). This trend has provided the required funding in IT investments. However, since the success rate is limited, only 67% are gaining a competitive edge from the use of Big Data analytics (Gregory, 2013), the managerial innovation is quite limited. In order for a Big Data project to prosper, organisations should acquire new technologies and invest in a new set of human capabilities - the “data scientist” (Davenport, 2012). Of course, human capabilities can be limiting to the framework's evolution if utilised in “massaging” data rather than generating business-critical insight.

2.2. Defining Big Data

In the business world and the IT ecosystem, there are several definitions for Big Data. The very first was by Laney in 2001, saying:

> *“3D Data Management: Controlling Data Volume, Velocity, and Variety. Current business conditions and mediums are pushing traditional data management principles to their limits, giving rise to novel, more formalized approaches.”* (Laney, 2001)

This definition was primarily interested in the IT infrastructure's limitations back in 2001. Storage was expensive, and parallel processing could be achieved with proprietary hardware at relatively high costs. However, technology advances and solutions for cheap storage and parallel computing on commodity hardware becomes a realisation. In adapting to this new evolvement, we have a new definition for Big Data addressing these capabilities.

> *"Big Data should be defined at any point in time as "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time." [...] Nowadays, it may mean data that is too large to be placed in a relational database and analyzed with the help of a desktop statistics/visualisation package—data, perhaps, whose analysis requires massively parallel software running on tens, hundreds, or even thousands of servers."* (Jacobs, 2009)

Once the web came into everyone's life, the term gained another definition incorporating the power of social networks, geotagging and IoTs as formulated below.

> *"Big Data - extremely large sets of data related to consumer behavior, social network posts, geotagging, sensor outputs, and more."* (Johnson, 2012)

Nonetheless, once business value realisation becomes evident, another definition comes into existence incorporating the business flavour.

> *"The concept of "Big Data" burst onto the technology and business scene in 2010, and since then has excited many executives with its potential to transform businesses and organisations. The concept refers to data that is either too voluminous, too unstructured, or from too many diverse sources to be managed and analyzed through traditional means."* (Davenport, 2012)

The first "clouds" in Big Data become apparent. It is more complex than initially thought to implement such environments, and thus a new definition gains acceptance. It implies both the power and the challenge of Big Data initiatives.

> *"Big Data means data that cannot be handled and processed in a straightforward manner."* (Fisher et al., 2012)

Implementations have realised little success; thus, yet another definition comes to identify this trend. Big Data is no longer simply a technological effect; it is "touching ground" with society – real people's lives.

> *"We define Big Data as a cultural, technological, and scholarly phenomenon that rests on the interplay of:*
>
> *(1) Technology: maximising computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.*
>
> *(2) Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.*
>
> *(3) Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy."* (Boyd & Crawford, 2012)

Cell phones and mobile technologies have become apparent, and these applications tend to create vast amounts of data. To that end, the definition of Big Data is further refined to incorporate the newly emerging technology.

> *“Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is Big Data.”* (IBM, 2015)

In all the referenced definitions found in the literature, it is easy to spot keywords that underline the essence and “texture” of Big Data. In attaining a high-level understanding of the V’s theory, prominent words shall be used in identifying trends.

The first group of indicators refers to Volume, 2.5 quintillion bytes of data, enormous sets of data, too voluminous, datasets whose size is beyond the ability of typical database software, speaking about petabytes and exabytes of data, cannot load into a computer’s working memory, data that is too large, analyse significant amounts of unstructured data, very large volumes, large data sets.

In this set of indicators, wording easily points us to the direction of volume, the enormous data size that must/should be handled—considering projections about data growth, it is understood that there is no mystery why they are referred to as Big Data.

The second set of indicators refers to Velocity, data comes from everywhere, too many diverse sources, beyond the ability of typical database software tools to capture, data they are generating every second, enabling high-velocity capture, discovery, and/or analysis.

The time dimension can be identified in this set as frequency or velocity, commonly known in the Big Data world. This factor has a dual perspective, the timeframe in which data must be acquired and, more importantly, the lifetime of the data. With the creation of multitudes of data, the lifetime of data sets tends to be smaller, thus requiring immediate processing.

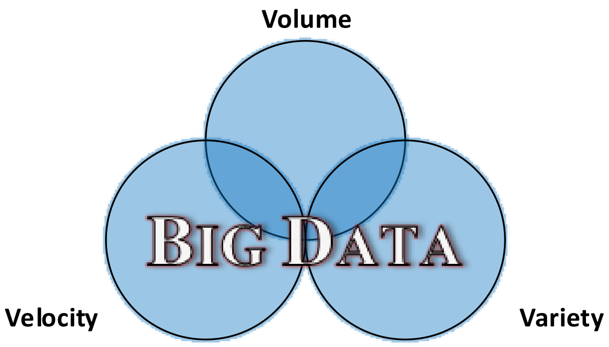
The third group of indicators refers to *Variety*, data comes from everywhere, social network posts – geotagging - sensor outputs – and more, from too many diverse sources, structured, semi-structured and unstructured data, ample amounts of unstructured data and a wide variety of data.

This group of indicators obviously refers to *Variety*; the data sets are too complex in relation to their content. The data can be structured (e.g. transactional data), semi-structured (e.g. XML files, logs), or unstructured (e.g. images, video) and their sources vary in nature. Sources can be sensors, IoTs, web, social media or public databases, each having its own structure and, in many cases, utilising different communication protocols.

Although Boyd and Crawford propose a more holistic approach that entails technology, analysis and methodology (Fosso Wamba et al., 2015), it is more common to identify and interpret Big Data in terms of V's, based on the first article on the subject by Laney (Laney, 2001).

2.3. The V's Theory

When Laney wrote the "3D Data Management," the term Big Data did not exist, but he was the first to identify the first set of V's, shown in Figure 3, that characterise the term.



Source: Rodríguez, 2013

Figure 3. The initial 3 v's

Volume: Refers to the amount of data created and stored in the digital universe (Ali-Ud-Din Khan et al., 2014). These data may be created from all kinds of sources like text, audio, video, database records, social media/networks, medical data, DNA sequencing, geolocation, music etc. In order to sustain, store, analyse and ultimately use this "ocean of data" (Hussein, 2020; Trifu & Ivan, 2014), organisations utilise petabytes of local storage in addition to cloud storage (Elragal, 2014). Nowadays, the yardstick in measuring corporate data shifted from gigabytes to petabytes (Kataria & Mittal, 2014; Munawar et al., 2020).

Velocity: Big Data is characterised as being high in velocity (Elragal, 2014), which means that the speed at which data change is quite high. In addition to change, velocity also incorporates the essence of real-time (or near real-time) processing of data streams that takes data from input through to the decision (Ali-Ud-Din Khan et al., 2014; Jabbar et al., 2020). The essential issue, posing a severe bottleneck, is that analytical resources, human and infrastructure, are limited whilst information requests are constant and varying, thus unlimited (Trifu & Ivan, 2014). To this end, almost all Big Data infrastructures utilise distributed processing and multi-user accessibility. Velocity can be subdivided into two types: between-subjects velocity and within-subjects velocity. The first refers to large-scale data collection from mass efforts, while the second is concerned with rapid data collection over time for an individual (McAbee et al., 2017).

Variety: This characteristic has to do with the data itself and the flavours it can pertain to. Sensors, IoT, database records, video and audio have different formats and standards, let alone the fact that alternative communication protocols must be used to disseminate the data streams in many cases. It is very complicated and “resource hungry” to analyse text, records, audio, video in the same context and acquire reliable results (Trifu & Ivan, 2014). Data variety directly affects data integrity, and thus it is a great challenge to deliver systems that can integrate such a diverse data mix (Ali-Ud-Din Khan et al., 2014). It is evident that data are heterogeneous (Kataria & Mittal, 2014), and apart from the specific formats (Anusha et al., 2021) used to encode the data (e.g. mp3, mp4, avi, text delimited, etc.), they can be distinguished into three categories (Elragal, 2014):

- Structured Data: mainly data coming from the corporate databases, data warehouses or data marts. This type of data is highly structured since the RDBMS systems impose a very strict schema, coherent and reliable.
- Semi-Structured Data: this category identifies data that have some kind of structure that nonetheless is not fixed and can vary in respect to size (length), multiplicity and content. Generally, this type of data tends to be self-defined to its structure. Possible occurrences of such data are emails, tweets, logs, XML,JSON files etc.
- Unstructured Data: All data that cannot be structurally classified are included here. That is data with primarily binary content, which is almost impossible to identify at “first glance.” Primarily in this category, images, audio and video are classified.

Since 2001 when these 3 V’s were identified, Big Data has come a long way, and in the process, new V’s were added, as shown in Figure 4. In the early years of Big Data, analysts and scholars were primarily focused on interpreting the term based on IT and the technical implications posed. As the technical challenges were tackled upon, business became the primary focus, thus enriching the V’s by incorporating business essence.

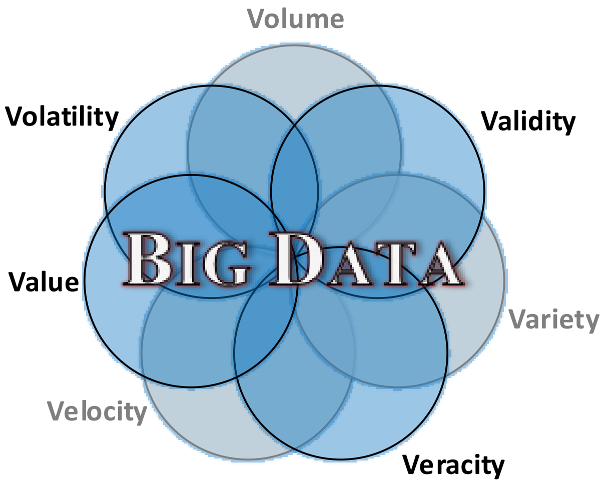


Figure 4. All 7 V's

Veracity: since many data come from sources outside the organisation, they tend to lack the required governance and homogeneity. Veracity is all about credibility, truthfulness and suitability (Elragal, 2014; Seliya et al., 2021). In other words, are the data what they claim to be? How certain can we be about this data? It is of great importance to employ data cleansing, utilising sophisticated tools and algorithms to ensure data coherence and erroneous data exclusion (Ali-Ud-Din Khan et al., 2014).

Validity: going “hand-to-hand” with veracity, validity is concerned with correctness and accuracy (Ali-Ud-Din Khan et al., 2014; Saeed & Husamaldin, 2021). The main difference with veracity is that data are evaluated in a context for their correctness. For example, veracity is concerned if the data stream claiming to provide customer geolocation has GPS coordinates and that these coordinates are existing locations. Validity would superimpose the intended use, for instance, location-based marketing. In this case, data would be screened in relation to existing customers and the intended geo-boundaries of the campaign.

Volatility: Is a by-product of the primary 3 V's, Volume – Velocity – *Variety*, since it is defined as the retention rate of data (Ali-Ud-Din Khan et al., 2014; Kamble et al., 2021). It is a fact that storage resources, even with the use of clouds, are finite. Given that and having taken into consideration the 3V's, it is evident that it makes a significant difference if there is a need to recall information for one month, one year or ten years. Of course, these are business decisions that can doom a Big Data project into failure. Imagine investing in infrastructure, software, people and business cultural changes only to find out that the available "window of data" is too small to have valid results. There are specific regulations related to data retention policies in specific industries being regulated (e.g., medicine, banking). For instance, in the Greek Banking sector, the retention rate for customer transactions is imposed by the Bank of Greece. It is set to five years after the termination of the customer's cooperation with the institute (ΠΔ/ΤΕ 2577 - Annex 8, 2006).

Value: Business is concerned with income and the realisation of competitive advantages to accumulate wealth. Bringing business value to the organisation poses a significant challenge and the same time, opportunity (Elragal, 2014). For any project to be successful, its Return On Investment (ROI)[3] must be positive. The same applies in Big

[3] Stands for Return On Investment. The simple formula in order to calculate this metric is $ROI = \frac{(Gain\ from\ Investment - Costs\ of\ Investment)}{Costs\ of\ Investment} \times 100$, whilst in most cases NPV (Net Present Value) is used

Data; data value and the respective information produced must exceed its cost, ownership or management (Oesterreich et al., 2022). The actual value lies in the eyes of the internal or external customer of business data (Ali-Ud-Din Khan et al., 2014). It would be in vain to create a sophisticated environment for Big Data that its users underutilise both in terms of data analysis as well as in terms of business initiatives based on the analytics produced. Value can be realised only if both IT and Business utilise effectively and efficiently the respective resources. Another factor that should be addressed to minimise possible loss of value is risk; thus, a governance mechanism can significantly enhance the final outcome (Ali-Ud-Din Khan et al., 2014).

2.4. Tools & Techniques

In this section, some of the most well-known concepts, tools, products and open source projects are referenced to get a glimpse of the Big Data ecosystem and jargon.

The first concept to be discussed is the Distributed File System (DFS). In order to harness the power of parallel computing by utilising commodity servers into a powerful Big Data infrastructure, one of the basic requirements is the ability to distribute data across these nodes. To this end, several Distributed File Systems (DFS) were implemented.

- Google File System: A proprietary system developed by Google (Manyika et al., 2011).
- S3: A proprietary system developed by Amazon (Warden, 2011).
- HDFS: An open-source implementation hosted by the Apache Software Foundation. Files are divided into blocks of 64MB by default, and stored on different nodes (Bedi et al., 2014). Data replication is utilised to compensate for hardware failures and

since investments span in more than one years with the following formula $ROI = \frac{(NPV\ Adjusted\ Total\ Inflows - NPV\ Adjusted\ Total\ Outflows)}{NPV\ Adjusted\ Total\ Outflows} \times 100$ (Audit IT, 2011; M. Schmidt, 2015).

performance bottlenecks (Warden, 2011). The infrastructure utilises two different types of nodes, the Namenode and the Datanode. As their names suggest, Datanodes store the actual data blocks while Namenodes store the metadata about the block's location within the Datanodes (Bedi et al., 2014).

The second concept is related to the data stores and the use of NoSQL which relates to how data is stored. Unlike RDMBS, these data stores utilise techniques to compensate for structure flexibility, large data sets querying performance and horizontal stability (Cattell, 2010). Many of these new systems are known as NoSQL databases, which provide high scalability and flexibility by exploiting parallel computing (Speegle & Baker, 2014). In contrast to RDBMS, ACID[4] has to be sacrificed, and the CAP[5] theory has to be utilised where at least one dimension, namely Consistency, Availability or Partition Tolerance, is underutilised (Padhy et al., 2011; Tannahill & Jamshidi, 2014). Although NoSQL databases are not destined to replace RDBMS, they are a better option for certain types of projects (Leavitt, 2010). The most popular types are the following:

- Document-oriented Databases: These database stores organise and store data as a collection of documents instead of structured tables (Leavitt, 2010). They are primarily concerned with high query performance rather than consistency and read-write performance (Han et al., 2011). Document stores do not require tables' structure to be defined beforehand since they store the information as a series of names and associated values. Theoretically, each and every record in such a structure could have a different structure (Warden, 2011).

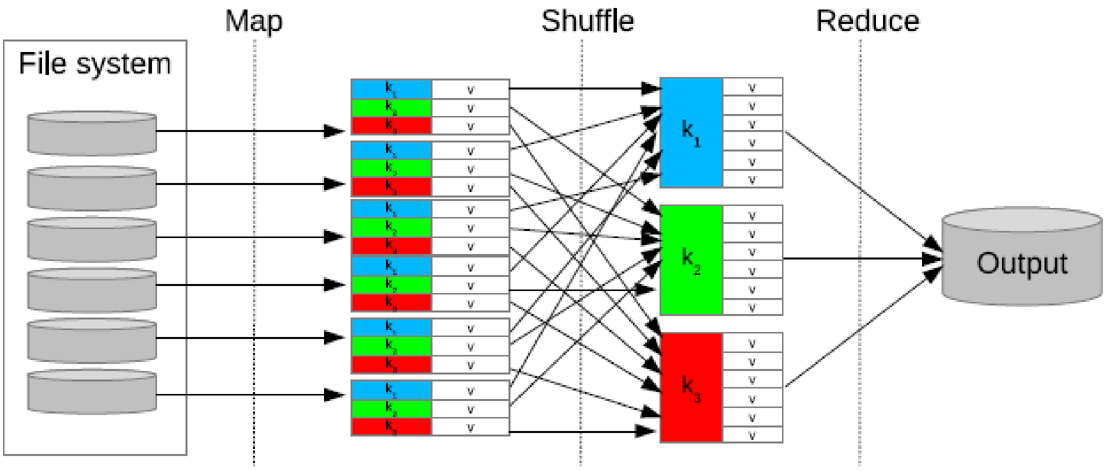
[4] ADIC stands for Atomicity, Consistency Isolation and Durability (Cattell, 2010).
[5] CAP Theorem states that with consistency, availability and partition only two can be optimised at any time (Bakshi, 2012).

- Key/Value Stores: They are similar to document-oriented databases in that they also store information in a key-value format. Their main difference is that there is no secondary index to the values in facilitating application retrieval (Han et al., 2011). The ancestor of the store is the Memcached system introduced to improve the performance of the website LiveJournal (Jose et al., 2011) and was mainly utilised by web developers (Warden, 2011). Unlike Memcached, which is an in-memory cache distributed store (Jose et al., 2011), these stores provide for persistence and replication, versioning, locking and other RDBMS like features (Cattell, 2010).
- Column-oriented Databases: These systems are similar to the RDBMS in that they tend to have a defined, although very flexible, structure (Han et al., 2011). The columns groups must be pre-defined whilst the actual columns/attributes can be easily redefined at any time (Cattell, 2010). The fundamental difference with RDBMS is that tables cannot be interrelated and that data, as the name implies, is not stored in records (rows) but in columns (Han et al., 2011). When the store is queried, only the requested column(s), instead of all columns from a table as in RDBMS, are used to provide the results. The most significant advantage is memory utilisation and speed in querying large data sets since, in most implementations, data is directly accessed as memory arrays without any deserialization (Floratou et al., 2011). Nonetheless, in such stores, write operations are generally considered problematic (Abadi et al., 2009).

From the respective implementation it is understood that the systems tend to move towards an unstructured storage representation. This is denoting that *Variety* has been acknowledged and is taken into consideration when storing the information, but none of the respective systems present the means of countering the effects of *Variety* before storing or processing the information.

The Big Data ecosystem is more than storing capabilities and options. Thus, it is essential to refer to basic techniques and methodologies in data placement and retrieval in a distributed environment.

- Sharding: Any database that employs multiple nodes (servers) to store data requires a way of identifying which node should be used to store and eventually retrieve any specific piece of data (Warden, 2011). It facilitates scalability by distributing data across nodes, whilst data replication ensures availability and automatic failover (MongoDB, 2013). Sharding is nothing more than an algorithm utilising a value (key) to distribute the data across the infrastructure. It can be user or application defined. Like RDBMS indexes and partitioning, the biggest problem is distribution and cardinality to evenly split the data across repositories (Warden, 2011).

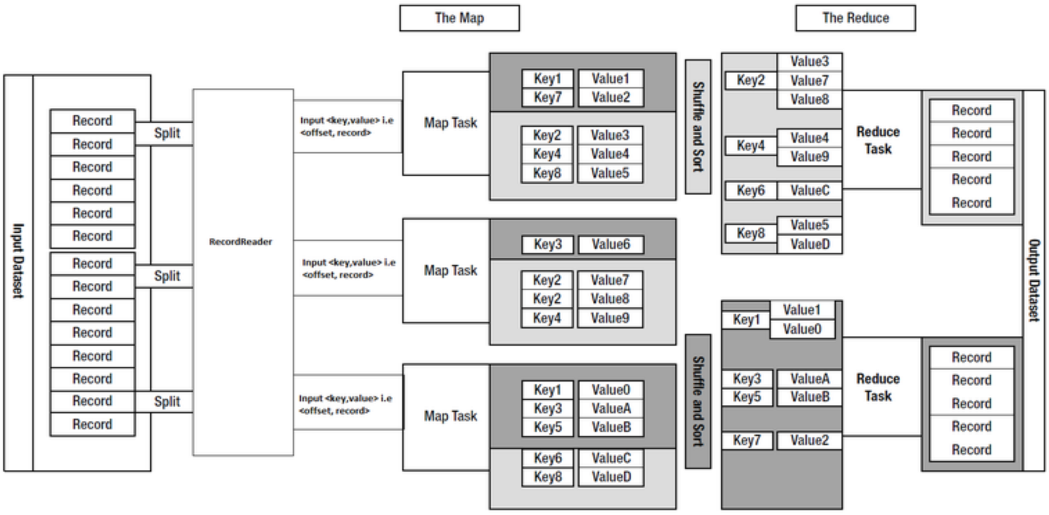


Source: Triguero et al., 2014

Figure 5. MapReduce Framework

- Map Reduce: This is a widely used framework for distributed computing (Bakshi, 2012). This model consists of two functions written by the user and three phases (Schneider, 2012; Warden, 2011).

- Map: in this phase, information of interest is identified by applying the map() function. The map() function takes a pair as input and produces intermediate key/value pairs (Dean & Ghemawat, 2004).
- Sort/Shuffle: in this phase, data prepared by the mapper tasks are moved to the nodes where the reducer task is to be executed (DeRoos, 2014). All intermediate values for a given key are combined into a list.
- Reduce: This phase applies the reduce() function to combine the list's intermediated values into one or more final values for the same key. The reduce function takes an intermediate key to the respective set of values and tries to merge these values into a smaller set (Dean & Ghemawat, 2004).



Source: Deepmohan, 2011

Figure 6. MapReduce Execution Overview

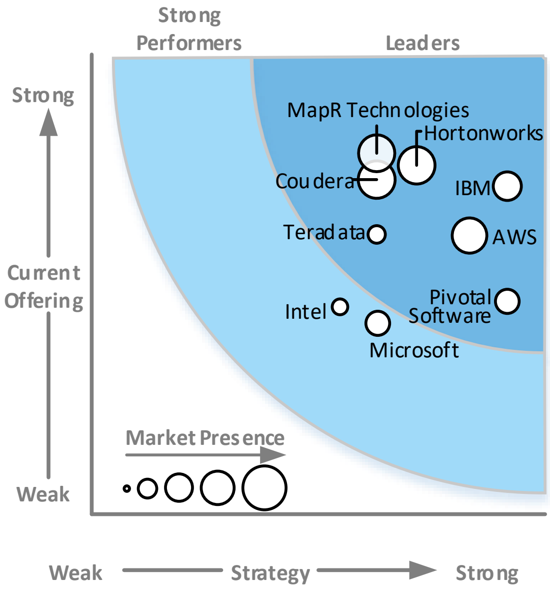
Finally, having gained familiarity with the basic concepts of Big Data, specific implementations and products available will be outlined. A growing number of technologies and implementations are used to manage, manipulate, and analyse Big Data. However, this cannot be a complete list since more technologies continue to develop (Manyika et al., 2011).

- MongoDB: This is a document-oriented database with records simulating the JSON structure that provides high performance, availability and scalability (Warden, 2011). It is written in C++, and the JSON like storage implementation is called BSON (Cattell, 2010). This implementation is very close to the relational model since it encloses a powerful query language, high speed to mass data[6] and complex data types (Han et al., 2011).
- CouchDB: An Apache project since 2008 which is written in Erlang. Queries are performed through the so-called "views" written in javascript (Cattell, 2010). The major limitation of the implementation is its interphase which is limited to HTTP REST API (Han et al., 2011).
- SimpleDB: Has been part of Amazon's proprietary cloud package since 2007 and, as its name suggests, is simpler than most document-oriented stores since it does not allow for nested documents (Cattell, 2010). SimpleDB structures data in subdivisions referred to as "domains" which in turn store a set of records (Padhy et al., 2011). There are important limitations as far as scaling is concerned; read can be horizontally scaled by reading replicas whilst write cannot scale. Built-in constraints like 10GB per domain, 100 active domains, 5 seconds limit on queries are extremely limiting (Cattell, 2010).
- Cassandra: This is a database system initially developed by Facebook, now managed by Apache open source, designed to handle a massive amount of data on a distributed system (Manyika et al., 2011). Its main characteristics include a very flexible schema, range querying and high scalability (Han et al., 2011). Cassandra is written in Java, and although it supports partitioning, replication, automatic failure detection and recovery, its main weakness is consistency (Cattell, 2010).

[6] MongoDB databases that exceed 50GB is 10 times faster than MySQL.

- Redis: An implementation initially introduced in 2009 utilising an in-memory key-value structure written in C (Cattell, 2010). Since it is an in-memory implementation, it is evident that it is limited by the amount of physical memory available on the hardware. Consequently, although it can handle more than 100,000 read/write operations per second, it is not destined for Big Data due to poor scalability (Han et al., 2011).
- BigTable: One of the pioneering alternative databases was designed by Google to handle substantial data loads utilising the underlying Google File System (Warden, 2011). It remained a proprietary infrastructure and was the inspiration for HBase (Manyika et al., 2011). It can readily scale into thousands of nodes and petabytes of data, performing millions of read/write transactions per second (Padhy et al., 2011).
- HBase: Is an open-source project handled by Apache Software Foundation as part of the Hadoop project (Manyika et al., 2011). It is written in Java utilising the HDFS and was first introduced in 2007. Partitioning and distribution are transparent, while B-trees provide high-speed range querying (Cattell, 2010).
- Hypertable: It is very similar to HBase and BigTable. It is written in C++ utilising an underlying distributed file system. Although it is open-sourced by Zevents, its popularity is relatively low (Cattell, 2010). Still in its infancy, it was designed for the proportion of 1,000 nodes, and the performance is not bad at all. The write rate per node can reach 7MB/sec, and the read speed can top at 1M calls/sec (Han et al., 2011).
- Dynamo: proprietary distributed data storage developed by Amazon (Manyika et al., 2011).
- Voldemort: The open-source clone of Amazons' Dynamo created by LinkedIn implementing a classic key/value architecture (Warden, 2011). Written in Java, it encompasses an automatic sharding mechanism and multi-version concurrency

control (MVCC), but since replicas are updated asynchronously, it does not guarantee consistency (Cattell, 2010).



Source: Experfy Editor, 2014b

Figure 7. Hadoop Solutions Q1 '14

- Hadoop: Initially developed by Yahoo!, inspired by Google's MapReduce and Google Distributed File Systems, it is now an open-source project managed by the Apache Software Foundation (Manyika et al., 2011). Hadoop's widespread implementation and recognition have made it a synonym for Big Data implementations. It is pretty easy to install and "play around" since it requires limited resources and can simulate pseudo-distributed mode on a single node (Bedi et al., 2014). Several "sibling" projects exist in the Apache Software Foundation that leverage its power and ease of use. MapR, Cloudera and Hortonworks, which were recently bought out, are the commercial edition distributions of Hadoop (Experfy Editor, 2014).
- Azure DocumentDB: a JSON document database utilising JavaScript. It is provided as-a-Service through the Azure portal. The infrastructure provides for seamless scale

and offers four distinct levels of consistency (Strong, Bounded Staleness, Session and Eventual) (CrowCour, 2014).

- Zookeeper: Reading and understanding the configuration information of a single machine is a moderately simple task; nonetheless, when faced with a distributed environment, this task can become quite difficult, if not impossible. In order to facilitate such environments, this framework was initially built by Yahoo! In essence, it is a very specialised key/value data store (Warden, 2011).
- Tokyo Cabinet: was a sourcefourge.net project now licensed and maintained by FAL Labs (Cattell, 2010) initially developed for Japans' largest Social Network Site (SNS) mixi.jp (Han et al., 2011). Tokyo Cabinet and Tokyo Tyrant are written in C, and there are six different variations of the backend server (Cattell, 2010). It is a very powerful storage engine since it can handle 4-5M read/write operations/sec and at the same time safeguard consistency, but if the data grows into billions, concurrent write performance will decline significantly (Han et al., 2011).

The aforementioned implementations are mainly targeting Volume and Velocity in the same way that presented NoSQL systems are not explicitly facilitating for *Variety*. There are many programming interfaces and tools developed primarily as supplements that make the developers' "lives easier" and can minimise the knowledge/learning curve by providing programmers with environments similar to the ones they are used to working with (Manyika et al., 2011; Warden, 2011). Detailed information on such environments can be found in *Appendix III. Big Data Programming Environments*.

2.5. Business and Societal Perspective

In the business environment, corporations engage in certain activities to gain a competitive advantage in their struggle to be profitable (Gregory, 2013). Big Data and analytics can provide such an advantage; nonetheless, certain aspects must be considered, namely Corporate/Organisational culture, Social and Legal. Identifying the organisation itself, the surrounding cultural and social environment and finally, the legal framework in which the organisation operates.

2.5.1. Corporate Aspects

Organisations that utilise Big Data attempt to attain value by adopting a flexible and multidisciplinary approach (Padhy et al., 2011). To this end, corporations should re-evaluate their structures in preparing to adopt the new management styles and embrace significant changes in the organisational culture required by Big Data (Dutta & Bose, 2014).

It is essential that in any case of investment in Big Data, there is a purpose, a definitive task in which the project has to deliver a solution (Gregory, 2013). If the corporation embarks on a journey to implement the technology and at the same time identify possible areas of interest, the project is likely to fail. This failure can be attributed to diminished top management sponsorship or apophenia practice[7] (Boyd & Crawford, 2012). In order to gain long-term success, apart from top management commitment and sponsorship, solid and continuous cooperation should exist between business and IT (Turner et al., 2013).

An interesting approach concerning internally versus externally focused corporations with different management styles towards stability and flexibility, in conjunction with the 3 V's theory, is formulated by Gupta. It attempts to identify the most prominent

[7] The occasion where, due to enormous quantities of data, patterns are identified where none exist.

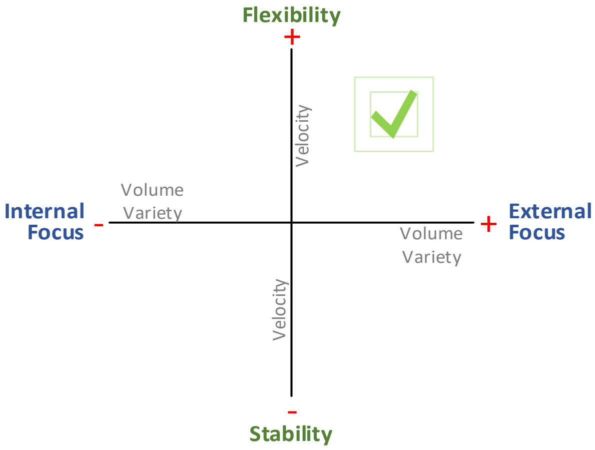
organisational cultures for implementing Big Data projects. The approach differentiates company cultures into (Gupta, 2014):

- Internally-focused organisations that invest in internal competencies in their pursuit of value. They establish family-like workplace environments and focus on harmony amongst organisational members.
- Externally-focused organisations are primarily concerned with the survival and growth of the organisation itself and closely monitor competition.
- Stable organisational structure implies an authoritative leadership in which employees are expected to adhere to decisions made by their managers.
- Flexible organisational structure suggests organic structures in which members, irrespective of their rank, are motivated to share their views and opinions freely.

Since externally-focused organisations monitor their business environment closely, it is relatively safe to argue that they are data prone and try to gather any piece of information in their ecosystem, thus positively influencing Volume and *Variety*. On the same principle, internally-focused organisations are primarily interested in their own logistics and operations. It is highly possible that these organisations pay little attention to external data and thus negatively influence Volume and *Variety*.

On the other hand, as far as organisational structure is concerned, stable, structured organisations tend to have rigid and tall schemas; thus, decision-making is time-consuming. In flexible organisational structures where everyone is considered equal, spontaneous decisions because of insight are more probable. Having defined Velocity, not only in relation to change but also in respect to the time needed to attain knowledge and utilise it, it could be argued that flexible-structured organisations would positively influence Velocity, while stable structured would have a negative one.

The diagram of Figure 8 summarises Guptas' approach by visualising it, using the Cartesian coordinates. It is evident that quadrant II assembles positively (+) in all of the 3V's, thus suggesting that flexible and externally focused companies have a competitive, or cultural, advantage in successfully completing Big Data initiatives.



Source: Gupta, 2014

Figure 8. Organisational Culture and the 3v's

Concerning *Variety*, the main focus of concern; externally focused companies seem to understand the effect since, on a smaller scale, they have been struggling with diverse data sets from clients/customers and vendors for some time now. In that manner, it is much easier for them to quantify the effort in contrast to internally focused companies, which have much more control over the processed datasets. Furthermore, although organisational structure mainly affects Velocity with respect to propensity and reaction to change, it can also be a valuable asset in the challenges of *Variety* since multi-talented pools can be assembled and tasked to address the proliferation of data. In the case of organisations with limited expenditure, such as non-profit organisations or small to medium sized corporations, Big Data instead of being an opportunity will become a handicap since they do not possess the tooling nor the intellectual capacity to manipulate Big Data towards a productive and developmental process (Gordo, 2017).

2.5.2. Social Aspects

Apart from the actual corporations interested in a profiting by employing Big Data technologies, society is also greatly affected. Very little is understood about the ethical implications that underpin the Big Data phenomenon (Boyd & Crawford, 2012). Discussing privacy and personal impact is a new ethical and sociological challenge posed by Big Data (Krawczyk et al., 2015). A statement by the Commissioner of the Federal Trade Commission highlights this aspect by stating:

> *"The potential benefits of Big Data are many, consumer understanding is lacking, and the potential risks are considerable"* (Fordham University School of Law, 2012)

Critics of Big Data worry that abuse of inferential data could seriously undermine personal privacy, civil liberties and consumer freedom (Bollier, 2010). They are concerned that Big Data is taking the principle of "knowing your customer" a step too far, thus feel unsettled about organisations knowing "that much" about them and their habits (Gregory, 2013). On the other hand, organisations seem to ignore consumers' desires for privacy (Kshetri, 2014).

In the attempt to "manipulate" *Variety*, automated techniques should be employed in minimising human intervention and thus allow the data scientists to visualise and identify insights. In this way, the process will not solely depend on humans and their respective skills which will in turn enable better business penetration and adoption (Assunção et al., 2014; Kumar, 2013). In such an attempt, Machine Learning (ML), Neural Networks

(NN), Artificial Intelligence (AI) and advanced algorithms can be utilised in classification, peak elimination etc. All the techniques mentioned above are based on statistics and probabilities, which can be acceptable in generalised terms. However, when it goes from a percentage to human lives, it becomes a reality. Anonymous and generalised statistics can be used and referred to without jeopardising any personal or social damage. However, when it becomes customer-centric or people-oriented, the damage affects the individual. The following examples, from the early days of data (predictive) analytics, have become "urban myths" amongst the data scientists and data analysis practitioners. a) The case of "My TiVO Thinks I'm Gay" is a manifestation of predictive analytics that "labelled" a TiVO customer as gay and, once he/she tried to alter his viewing preferences, "suggested" he/she is a Third Reich follower (Bollier, 2010; Zaslow, 2002). b) The case of TARGET, a consumer behaviour analysis/correlation system, which identified a teen as pregnant and triggered a baby care promotional letter which was intercepted by her father, who was not informed of the pregnancy yet. (Duhigg, 2012; Hill, 2012). It is vital to bear in mind, especially when utilising the results, that the outcomes are suggestions and not "definitive truths" and should be treated as such.

The "war" against *Variety* is primarily a way of thinking, an approach. In order to facilitate the process of giving data context and unification, multiple "battles" have to be fought using elaborate algorithms in the "battle fields" of statistical analysis, artificial intelligence, cognitive processing, and natural language processing, along with security and data confidentiality. In such a process the data scientist should always verify coherences and indicate possible risk factors. In presenting the final insight, the scientist ought to identify every assumption in all stages of data processing, including the ones employed in limiting the *Variety* effect.

It is argued by researchers that:

- With enough data, even depersonalised data can be processed to infer and identify actual individuals, referred to as the "de-" process (Narayanan & Shmatikov, 2009; Zimmer, 2008).
- Likes can be used to accurately predict highly personal attributes like sexual orientation, ethnicity, religion, intelligence, welfare etc. (Kosinski et al., 2013)
- The paradox of "My TiVo Thinks I'm Gay" can lead to erroneous human labelling and, in some cases, can harm the reputation and be the source of physiological turmoil (Zaslow, 2002).

Concerning Big Data, privacy/confidentiality is one thing that seems to be a source of turmoil in society, and discrimination is the second. "Digital Shadow" is a recent term introduced in identifying information that, although deemed to be public, in some cases, individuals would instead prefer that they remain private (Gantz & Reinsel, 2011). Organisations traditionally discriminate to maximise market penetration and market share, with so-called "targeting" - "personalized marketing" and so on. Based on analytics of Big Data, companies can observe trends and so discriminate between groups of individuals, which can lead to customers' unfair treatment. This aspect is exemplified by the EU case, in which insurance companies are banned from using statistical evidence concerning gender to differentiate premiums (Newell & Marabelli, 2015). In order to ensure "sensible" and "safe" usage of data by organisations without the possibility of potential harm to society, it is widely argued that there should be governmental intervention.

Big Data is characterised by volume, and it is suggested that more data will ensure better decision quality in the long run (Sterner & Franz, 2017). Nonetheless, from a societal point of view, it is argued that although the volume is substantial, the composition available does not adequately represent all parts of society across the world. For that matter, many social and ethnic populations have limited access to the internet (Gordo, 2017). In order to have “socio-profiled” and “neutral” data based on equal access to online services, internet penetration must cover all social groups and geolocations.

2.5.3. Legal Aspect

There are no rights in data itself, but extensive rights and obligations arise about data, mainly by intended use (Kemp, 2014). In several heavily regulated sectors like Banking, Insurance or Health, mechanisms have been implemented, politely stated for imposed, to safeguard both the customer and the institutions. Nevertheless, the Big Data initiative affects all sectors (Manyika et al., 2011). There might be sector-specific cases where legal aspects are not fully covered by the general EU or local legislation; it is argued that collecting, processing, and consuming such vast amounts of personal data are certainly uncompliant with the existing EU data protection laws (Helbing et al., 2017). Most organisations do not employ systemic and systematic mechanisms and procedures to ensure controlled data access (Kshetri, 2014).

Big Data has changed the way data is used to acquire knowledge. With the utilisation of inference, there may be legal implications in the results' use, visualisation, or communication. Rights and duties that arise about data are both valuable and potentially onerous and, as a matter of law, are developing rapidly (Kemp, 2014)

“Data law” is emerging as a new entity concerned with Intellectual Property Rights (IPR). Contract and regulation are primarily concerned with the six-level data stack's three middle layers (Kemp, 2014). In attempting to minimise the *Variety* effect, governance and data stewardship will have to be addressed; thus, the six levels mentioned will be referenced and explained in identifying basic legal understanding.

- Level 1 - Platform Infrastructure: Interested in the physical infrastructure and the respective software utilised. It is primarily concerned with software copyrights and rightful ownership.
- Level 2 – Information Architecture: The interest resides with infrastructure documentation and is easily overlooked. For instance, the database schema is protected in the EU under Chapter II, Article 3 of the Database Directive.
- Level 3 – Intellectual Property Rights: IPR concerning data are copyright, database rights and confidentiality. Copyright protects the form of expression of information, not the underlying information itself. Database rights arise when the maker has made a substantial investment in obtaining, verifying or presenting the information, and confidentiality is about trust and non disclosure.
- Level 4 – Contracting for Data: Contract law has strong and enforceable obligations since liability is strict. It has merit on its own since a judge in UK High Court in 2006 said that:

"I agree with BHB that it is entitled, in principle, to impose a charge for the use of its prerace data by, and for the benefit of, overseas bookmakers, whether or not BHB has IP rights in respect of the data, and, in particular, database rights under the Databases Directive and the Databases Regulations or copyright, and irrespective of the extent of any such rights." (The High Court of Justice - Chancery Division, 2007)

- Level 5 – Data Regulation: Data governance, privacy, and protection are essential integrals in the policy utilised by corporations. Client confidentiality rules have been the cornerstone of the legal profession for generations; nonetheless, digitisation and sector-specific mandates have altered the landscape. Examples of private information delivered to external committees are data required by the financial sector institutions to comply with the Markets in Financial Instruments Directive (MiFID) and Anti-money Laundering (AML).
- Level 6 – Information Management and Security: Several standards have evolved to protect data. PCI is the standard that financial institutions should adhere to in relation to Debit and Credit Cards. For example, one of the primary directives is that the credit card number should be masked or hashed in all the institutions' systems, thus avoiding employee fraud.

These levels provide with an understanding of the law behind information stewardship and governance thus presenting the regulatory framework under which data confidentiality is becoming very important.

2.5.4. Ethical Aspect

Technology itself is ethics-agnostic (Chessell, 2014); thus, data and information can be argued to be ethics-agnostic, being source and product of an agnostic process. On the other hand, gathering, usage and intent have ethics written all over them. Big Data are revolutionary, but it is imperative to ensure that this is a revolution that people agree upon and that long-cherished values like privacy, identity and individual power are preserved (Richards & King, 2013).

Big Data sceptics argue that gathering data without people's knowledge or consent cannot meet the ethical obligation of treating people with "justice, beneficence and respect" (Crawford et al., 2014). Three paradoxes have been identified (Richards & King, 2013):

- The Transparency Paradox: Big Data analytics proclaim to use data for transparency. Nonetheless, its collection is, at large, done invisibly without people knowing it. Furthermore, action plans are based on the analysis results; discussions are made about people, and these people have the right to know the basis on which these decisions are made.
- The Identity paradox: Big Data seeks to identify, but also threatens Identity. "I am" and "I like" are under the risk of being transformed into "You are" and "You will like," respectively. Even further, without proper identity protection, there is a high probability of mutation into "You cannot" and "You will not," respectively, giving birth to a world like the one presented in the film Gattaca[8].
- The Power Paradox: Big Data will create winners and losers, and it is likely to favour organisations exploiting the tools over individuals being mined, analysed and sorted. It is very similar to the famous phrase by Winston S. Churchill, "History is written by the victors."

[8] Columbia Pictures 1997.

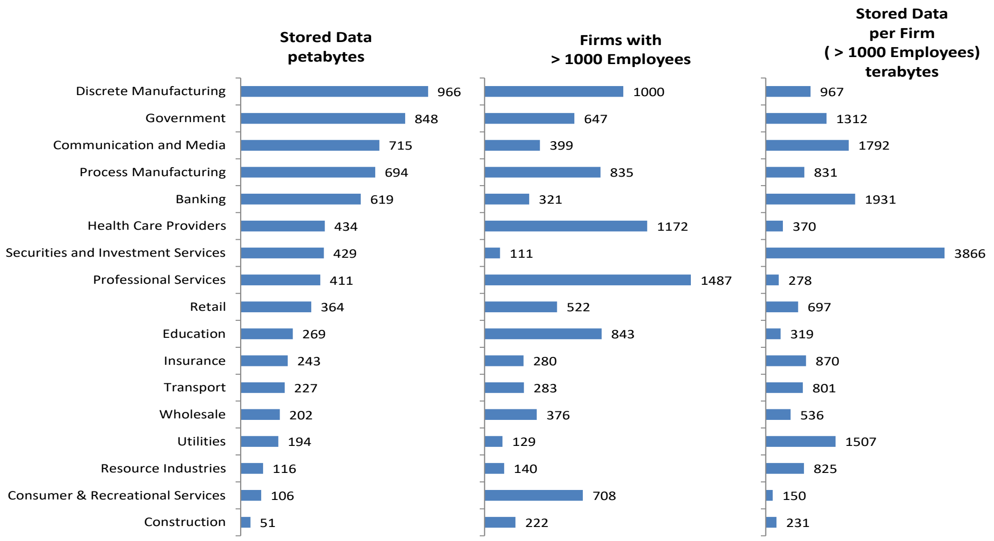
The UK and Ireland Technical Consultancy Group developed the ethical awareness framework to help the industry follow “a righteous path” by developing ethical policies for Big Data and analytics. The following facets comprise the framework (Chessell, 2014):

- Context: What was the original purpose for collecting the data? What is the current usage of the data? How far apart are the two? Is this appropriate?
- Consent and Choice: What choices does the affected party have? Is it to their knowledge that they are making a choice? Do they really understand to what they are agreeing? Is there the possibility of disagreement or decline? Are there alternatives offered?
- Reasonable: Is the depth and breadth of data employed along with the derived results reasonable for its intended use?
- Substantiated: Are the data sources used in a timely, complete and appropriate manner?
- Owned: Who owns the results? What is their responsibility to privacy protection?
- Fair: Are all parties equally compensated? How equitable are the results to all parties?
- Considered: Can the consequences of the analysis and results be identified and quantified?
- Access: Who can access data? Can subjects access data completely and transparently?
- Accountable: Can the affected parties validate the correctness of the results? Were any errors or unintended mistakes identified, and how were they resolved?

The ethical questions posed are augmenting the legal requirements in presenting with soft-law principles. In many cases, even on a corporate level, confidentiality will require one to understand and incorporate such principles, possibly within a rule-based identification system.

2.6. Sector Analysis

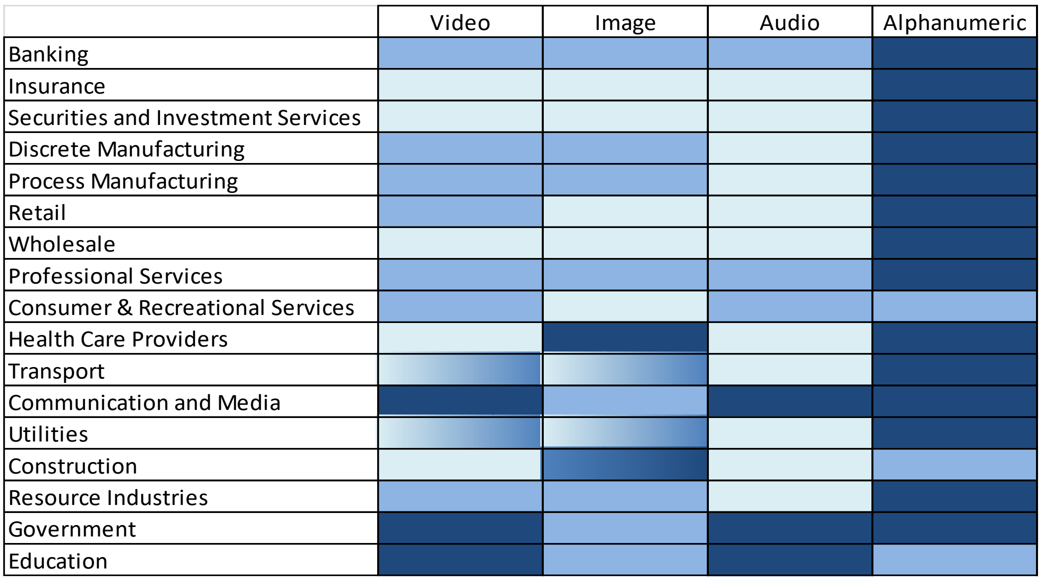
McKinsey Global Institute states that data growth is phenomenal in almost all industry sectors and further provides a sectoral analysis of the global economy.



Source: Manyika et al., 2011

Figure 9. Companies with at least 100 terabytes of stored data

In Figure 9, a graphical representation, in the form of a bar chart, of the companies with at least 100 terabytes of storage in the US identifies sectors with a higher potential to capture value from Big Data. The types of data utilised by each sector is shown in Figure 10, in the form of a heat-map where the intensity of the colour in the heat-map is denoting the volume of data under the respective category. The darker the shade, the more data under that category the respective sector has.



Source: Manyika et al., 2011

Figure 10. Types of Data Stored

Based on the information from the two figures it can be identified that a) financial services, b) government, c) retail, d) health care and e) manufacturing exhibit high volumes and differentiation in types of data and thus seem to be the most prominent sectors for Big Data implementation, requiring further investigation (Manyika et al., 2011):

a) Financial Services

The banking sector is one of the major players in IT implementations globally. Indicatively, the trading platforms, which account for a small part of the entire sector, are approximately a $25 billion market (Kemp, 2014). Paul Scholten, Chief Operating Officer (COO) of ABN AMRO’s retail and private banking, states one of the most common practices in the banking and financial industry; financial data analysis (Briody, 2011). Traditionally, financial institutions collect transactional data and derive several metrics in relation to the customer's organisation (e.g. liquidity report, debt exposure, etc.) (e.g. customer profitability, credit control, etc.). Structured data is second nature to the sector, but when it comes to unstructured data, there seems to be "sufficient confusion" (Briody, 2011).

Financial and pharmaceutical are supposed to be the most regulated industries. Since the financial crises in 2008, several rulebooks, like BASEL and MiFID, emerged and were imposed on the industry (Kemp, 2014). Also, regulatory bodies have expanded from national to international and global boundaries to govern and possibly prevent another crisis in the future.

The main focal points in implementing analytics are driving revenue, controlling costs, and mitigating risks (McKinsey & Company, 2013). Controls, procedures, and know-how are already present in the sector. Financial institutions are excellent candidates for Big Data projects, along with extensive data sets with a high degree of conformity, coherence, and correctness. For instance, in the banking industry, extensive IT resources are utilised in attending to their information needs, with in-house technical and analytical resources having accumulated extensive knowledge over the years. This corporate knowledge will enable a smooth transition in employing Big Data initiatives.

Nonetheless, it is of utmost importance to extend Big Data analytics by reaching out to data outside their "comfort zone," namely social and web-based data (Manyika et al., 2011). Simply replacing existing infrastructures with more efficient Big Data will realise benefits, but there is more to it. Only by integrating internal transactional data with external social datasets can there be significant value gained (Tankard, 2012). Mining and analytics to understand customer needs, predict their wants and demands and optimize the use of resources (Assunção et al., 2014).

The social, ethical, and legal implications are resolved due to heavy regulation but only cover traditional financial data. If institutions are to utilise and integrate external sources, matters come to a new level of complexity that must be seriously addressed. For instance, what if a bank had access to health records and utilised it for credit scoring?

b) Government (EU)

Governmental and Public Sector Administration worldwide are under enormous pressure towards efficiency, effectiveness, and general productivity. Due to the resection trends experienced worldwide, significant budgetary cuts are imposed to reduce the national debt. It is estimated that the European public sector can realise even more than €150-€300 billion of value by employing Big Data and increasing transparency (Manyika et al., 2011).

Possible areas of immediate value can be identified in operational efficiency, reduced costs attributed to errors, and a substantial increase in tax receipts. By utilising Big Data analytics, other opportunities can be accrued in value realisation. Accurate projections of workforce dynamics, management of public resources like transport and fixed infrastructures and last but not least, crises management preparedness can be realised with the use of extended what-if analysis and simulation (Kambatla et al., 2014).

c) Retail (US)

Retailers traditionally mined datasets concerning customers, attaining valuable insight that assisted in managing supply chain, merchandising and pricing. The sector has utilised this kind of practice since the introduction of Point-of-Sale (POS) devices in the 1970s

(Manyika et al., 2011). Since then, several means have been utilised to identify personalized buying trends and preferences; loyalty cards have been the most common implementation. Cross-selling and market segmentation are also of great interest to the retail industry, where web-based stores are utilised extensively to maximise targeting and efficiency.

Apart from direct sales, though, retailers can utilise analytics in order to fine-tune inventory and logistics, placement practices and in-store behavioural analysis (Manyika et al., 2011) and cut down on financial costs by imposing “pressure” on their suppliers with respect to payment terms and actual prices.

Even though companies “are in it for the profit,” Big Data can substantially benefit customers by providing better-suited products for their needs (Manyika et al., 2011). Of course, as in all analytics prospects, the legal and ethical implications should be addressed, and companies must respect their customers’ rights to privacy and disclosure (Hoy, 2014).

It is challenging to quantify possible value gain from the initiative of implementing Big Data in this sector because the realisation of success is closely correlated with customer behaviour. Each organisation can utilise the technology in diverse perspectives whilst customer response is unpredictable in most cases. To put it simply, a firm might do all the right things for the right reasons but attain a minimal number of new customers, because of external factors, such as acceptance due to talent, cultural and technological obstacles (Manyika et al., 2011). Nonetheless, an estimate of 15% to 20% regarding ROI is argued to be attainable (Fosso Wamba et al., 2015).

d) Health Care (US & UK)

Health care is one of the largest sectors of the US economy, accounting for approximately 17% of Gross Domestic Product (GDP)[9] in 2015 and raising to 19.7% by 2020 while expected to reach one fifth (1/5) of the GDP by 2028 (The World Bank, 2015; J. Yang, 2022), estimated[10] at 4.7 trillion US dollars and employing 11% of the country's workforce. As far as the UK economy is concerned, the sector accounted for 8% of the GDP in 2014, estimated at £120 billion and is expected to increase to 12.8% by 2020 (Kemp, 2014; J. Yang, 2021).

Thus far, health care has lagged behind in adopting technology-enabled processes to improve operations. Nonetheless, it is estimated that if Big Data opportunity is realised, more than 300 billion US dollars can be captured in new value with two-thirds (2/3) accounting for reductions in expenditure (Manyika et al., 2011). McKinsey & Company has estimated that the usage of Big Data could lead to a cost reduction of approximately $450 billion in the sector (Wu et al., 2017).

Possible areas of application can be identified in public health, disease research, drug research, genetic engineering etc. The inefficient paper-based record keeping can be digitized in the assembly of data polls for insight evaluation (Bollier, 2010); however, the appropriate legislative initiatives must be taken to ensure privacy, objectivity and accuracy (Hong, 2014). Wearables that constantly collect medical information can also

[9] Gross Domestic Products, “an aggregate measure of production equal to the sum of the gross values added of all resident, institutional units engaged in production (plus any taxes, and minus any subsidies, on products not included in the value of their outputs).” (Dawson et al., 2010).

[10] US Bureau of Economic Analysis estimates US GDP for 2022 at 24.380.000 millions (US Bureau of Economic Analysis, 2022).

provide valuable personalisation in the healthcare industry and increase the risk of privacy information leakage (Wu et al., 2017).

e) Manufacturing (Global)

Manufacturing was one of the pioneer sectors in implementing IT from the very beginning. The returns realised in past decades concerning production growth and automation were attributed to operational improvements leading to efficiency in the manufacturing process and improvements in product quality (Baily et al., 2013). The sector is identified as the backbone of many developed economies and an important driver for GDP (Manyika et al., 2011).

In order to achieve value improvements from Big Data initiatives, the fragmentation of the organisations must be addressed by shifting mindsets and behaviours toward abolishing silos. Areas of application include Research and Development (R&D), after-sales support, factory layout and, product life cycle (Manyika et al., 2011; Tankard, 2012).

Concerning value realised, it is argued that the application of Big Data analytics can reduce R&D costs by 20% to 50%. At the same time, supply chain optimization could yield a 2 to 3 percentage point improvement in profit margin (Baily et al., 2013).

2.7. Conclusion

The Big Data era came about, and in doing so, it brought new techniques and tools into the IT landscape. Distributed data management and distributed computing are utilised in augmenting computer power. Volume and Velocity are being researched, and technical innovation is leading to new implementations, whilst *Variety* is seldom addressed. In addressing *Variety*, it is identified that there has to be a change in mentality and not only a technological advancement.

When the business world started investing in Big Data, challenges manifested in an attempt to gain competitive advancement by utilising insight from the information availability. Traditional organisational structures were proven inefficient in catering for the skill augmentation and elasticity required for ad-hock multi-talented scram oriented teams needed to take a Big Data initiative to successful completion. Social and Ethical dilemmas surfaced and identified the need to preserve human rights. Understanding further the effect of Big Data on society and human relations will eliminate paradoxes like Identity or Power, which can erode the social structures as we know them. As mentioned above, the dilemmas and challenges gave birth to the “Data law” since there must be a legally documented approach to handle disputes on both a personal and corporate level.

Variety manifests itself in all business, cultural, social, ethical and technological conceptualisations of Big Data. From the definition of a data set to the definition of human behaviour, we can detect anomalies mainly attributed to *Variety*. Nonetheless, it is evident that several business sectors are investing in Big Data. Implementations are primarily aimed towards revenue increase, cost cuts or risk mitigation. The amounts to be realised by investing in Big Data is estimated in the hundreds of billions, thus making Big Data initiatives essential for all organisations. This need for Big Data skills fuels a lucrative industry to be exploited by services, consulting and IT companies in addressing and implementing these initiatives.

A hybrid approach will present the related work based on the research performed. It will have a content subdivision with respect to V's, and within it, the time dimension will be utilised in putting it into perspective.

The references in this chapter will be limited to the implementations and solutions identified in the respective areas (V's) rather than the theoretical and documentary approach which was used in the previous chapters. The approach used in performing the research in regard to epistemology and methodology are outlined (Section 3.2), followed by the prior art being detailed (Section 3.3).

3.1. Introduction

Doug Laney, with his analysis, "3D Data management," in 2001, is considered the "Father" of Big Data. He was the first to identify and describe the basic V's that would become the cornerstone of the Big Data ecosystem. In his document, he identifies the need to harness information from several sources across different applications, structures, and formats. In addition to that, the timeline element is introduced by identifying current and historical data needs. In his analysis, at that time, e-commerce was the source of an explosion of data concerning three dimensions, Volume, Velocity and *Variety*. Since then, these 3V's have become the defining dimensions of Big Data.

The IT and business landscape has evolved since 2001. Concepts like social networks, personalised marketing, Omni channels, and IoT have come into existence and are further attributed to the exponential data growth in all three dimensions (Gopal et al., 2022; Krawczyk et al., 2015; L. Zhang, 2014). In addition to that, Big Data theory has also evolved, and further V's were identified. By 2014 V's had increased to seven, expanding to Veracity, Validity, Volatility and Value. In Khan’s conference paper “Seven V’s of Big Data,” a comprehensive analysis and discussion on the 7 V’s is conducted. Referring to the www, smartphones, corporate relational databases and, social media is used to identify, correlate, and understand the V’s. Of importance is that in his paper, he makes reference to the most common RDMBS commercial systems (Microsoft SQL Server and Oracle) and identifies that these systems are incapable of catering to the Big Data ecosystem. What seems to be missing are ethical, social, and legal dimensions.

In 2012 Boyd & Crawford, in their publication “Critical Questions for Big Data,” identified Big Data as a cultural phenomenon addressing the socio-ethical aspect of the ecosystem. Of great importance in their work is the emphasis given on the correlation between Big Data and the traditional way research is performed. As they denote, “Big Data stakes out new terrains of objects, methods of knowing, and definitions of social life” since they argue that the way data is analysed in essence shapes the result, just like accounting is not only recording financial measures but also shaping them.

Bollier, in his work “The promise and peril of Big Data,” published in 2010, also begins his analysis based on the scientific substance of Big Data by elaborating on Chris Anderson's statement “the data deluge makes the scientific method obsolete.” The criticism identified and discussed mainly concerns using data through a correlation methodology in corroborating evidence towards a self-fulfilling prophecy. The importance of this publication is that it is concerned with personal freedom and civil liberties that can be breached with the use of Big Data. Boiller is taking one step further than the Big Data human labelling paradox discussed in Zaslow’s article, “If TiVo Thinks You Are Gay, Here's How to Set It Straight” in The Wall Street Journal in 2002. The latest addition in 2017, by Gordo, is the inequality of data contribution based on the ethnic aspect. It is argued that there is injustice and prejudice in Big Data since many countries, especially the so-called "developing," have limited internet access. Legal aspects came into the picture a little after since there has to be adoption before there is any legal dispute.

Although scholars have been working in socio-ethical aspects of Big Data since 2002, we seem to identify major legal concerns after 2014. In Richard Kemp's article, "Legal aspects of managing Big Data," published in 2014, a 6 level data stack is introduced to "regularise" intellectual property rights, contracts and regulations in the Big Data context. His work is comprehensive toward identifying a legal framework and can be adopted by most business entities. The governance pillar is identified and elaborated on to provide a step-wise approach towards implementation. What is not identified or measured is the cost of compliance, whether monetary or time, associated with governance. These costs could render the aforementioned implementation unsustainable for many organisations, especially those with smaller sizes.

One of the latest developments in the legal area, from April 2018, is forming the Social Media Working Group by the European Union to investigate any private data harvesting from social media (Fioretti, 2018). This act follows the Cambridge Analytica scandal in which tens of millions of Facebook users' private data were found in possession of the political consultancy agency without users' consent (Kleinman, 2018).

3.2. The Review Approach

Before describing the methodology employed in researching and understanding prior knowledge, the theory of thinking will be presented.

3.2.1. Epistemology

Although in almost all text books there is a chapter on literature review, the focus would be more on the topic's importance rather than providing actionable guidelines (Boote & Beile, 2005). In achieving a thorough review, it is important to first define the literature review in terms of its properties (Schryen et al., 2015):

- Synthesis and Interpretation. The synthesis, which will provide a detailed summary of the existing material available, is an integral part of the review. In addition to that many scholars argue that interpretation in identifying less researched areas and future research is also very important.
- Focus on Domain Knowledge. The review should be focussed on the knowledge relevant to the particular field of study.
- Comprehensiveness. The review should be comprehensive, spanning across different sets of journals and different geographic regions (Webster & Watson, 2002).

Based on these properties, the definition of the scope of a literature review can be defined as follows: "*A literature review provides both a comprehensive synthesis and an interpretation of the body of knowledge of a specific domain*" (Schryen et al., 2015).

In order to advance collective understanding any researcher or scholar should understand prior work by reviewing the literature of what was done before in the relevant field(s), commonly referred to as "synthesis" (Boote & Beile, 2005). An equally important contributor is the adoption of a new perspective where the literature is reviewed under a new prism, identifying views that were previously not exploited which could lead to the identification of research gaps (Schryen et al., 2015).

The review is concept centric (Webster & Watson, 2002). In presenting the information tabulations with respect to identified concepts are useful, but since the articles may very well be quite many, it can be further aggregated in to a concept – count matrix.

Table 1. Concept Matrix

Article No	**Concepts**				
	A	B	C	D	…
1		✓	✓		✓
2	✓	✓			
3					
…			✓	✓	

Source: Webster & Watson, 2002

Table 2. Aggregated Concept Matrix

Concept	**Number of Articles**
A	1
B	2
C	3
D	1
…	

3.2.2. Methodology

A thorough analysis and understanding of the Big Data environment is required in order to be able to identify the current state of the art. The methodology used was an iterative task in which the following three steps were applied:

- Planning
- Execution
- Documentation

In planning the review, certain areas of Big Data were selected to have a holistic understanding of the ecosystem. Apart from the technological aspects, the following areas were also researched:

- Corporate and Business Sector Analysis in relation to Big Data adoption and prospect.
- Social paradigm and dilemmas introduced by Big Data.
- Legal challenges in utilising Big Data.
- Ethical concepts that might slow down the adoption process.
- RDBMS and related tools were also investigated to identify the adoption process and make parallelism.

- Metadata standards and technologies were investigated to understand and identify possible usage in the Big Data ecosystem.

To identify the current state of the art for *Variety* (concerning Big Data), a broad review of the prior art was undertaken using Elsevier's ScienceDirect, Springer Link, IEEE Xplore, and Google Scholar. The methodology used included multiple rounds of research, evaluation and classification of Big Data reference material. The documents were initially evaluated based on the search engine results. The second and third screening extended to the abstract, introduction and conclusions along with a complete document review, respectively. Based on these iterations, a set of four hundred and fourteen (414) articles were identified and categorised, as shown in Table 3.

Table 3. Big Data References Classification –Aggregated Concept Matrix

Subject	Number of Resources
Big Data in General	152
Business / Corporate	37
RDBMS	40
Legal, Social, Ethical	23
Metadata	56
Productivity Tools	42
Volume	23
Velocity	14
Variety	13
Heterogeneous Data (in Big Data)	14

Notable within this analysis, whilst a good number of research studies and publications have focused on Big Data, a relatively small number have specifically focused upon *Variety* – even though the issues concerning *Variety* have been well established. Unfortunately, an analysis of the thirteen (13) papers specifically on *Variety* revealed that all were focused on confirming the challenge rather than proposing or validating solutions. To ensure the review captured all relevant research, it was decided to undertake further searches with a broadened set of keywords – including related terms like

“Heterogeneous Data.” One hundred and seventy (170) studies were initially identified as relevant and further reviewed from the new result set. This resulted in twelve (12) papers being identified as relevant, focusing on challenges originating from automatic schema management, federated data sets, missing and erroneous data and the interdependence of heterogeneity and volume. Additional searches utilising “heterogeneity” were executed. However, because most of the searched and reviewed documents were primarily related to specific case studies on datasets pertaining to Biology (proteins), Genetics (Gene sequences), Medical (Cancer & C-Scan Recognition), and Disaster (floods, droughts, earthquakes), it was decided to further limit the search by applying filtering terms including “RDBMS,” “Big Data” and “Database.” From the new pool of references, one hundred (100) were reviewed. However, only two (2) met the criteria since most of the references focused on storage implications, NoSQL Databases, filtering and contrasting rather than the *Variety* challenges like integration and unification.

Timeline analysis of the references denotes the researchers have mainly identified the challenges but have offered limited solutions, if any. Kumar, in his study, identified that the cost to be paid on services in addressing *Variety* in any Big Data project would be the “key cost element” and suggested that internal resources should be used in lowering the costs (Kumar, 2013). On the other hand, Lennard, Shacklett and Brown all agree, in independent publications, that specialised personnel have to be utilised and that the respective specialisation is not present in many organisations (E. Brown, 2014; Lennard, 2014; Shacklett, 2014). These studies help to substantiate that human capital is a critical and scarce resource. Being complex and costly is not the only challenge. It is identified that errors and controversial results will be the outcome of *Variety* (Ali-Ud-Din Khan et al., 2014), and organisations have understood the value of minimising the effect. As Kimura denotes, 74% of businesses would like to harmonise their data (Kimura, 2014). However, *Variety* will be the biggest challenge, according to Sharmila Mulligan, CEO and founder of ClearStory Data. They, in turn, identified that *Variety* is the biggest challenge in providing valuable business insight (Baker, 2015). Mao is very specific in his work "*However, most of the effort was spent on volume and velocity, but not as much on variety*" in naming the research inadequacy (Rui et al., 2015).

Given the lack of relevant literature within Big Data *Variety*, it was deemed prudent to broaden the problem and search criteria. In investigating the issue of data heterogeneity, a wider body of literature was identified. The topic dates back to the middle 1970s (Luo et al., 2008). Initially, methodologies like Common Object Request Broker Architecture (CORBA), Distributed Component Object Model (DCOM) and Electronic Data Interchange (EDI) were devised and employed. However, they all focused on defining a clear and rigid communication framework that would prevent *Variety* by abolishing any diversification (Luo et al., 2008). This was arguably sufficient for the early years. However, when the World Wide Web came into existence and was eventually extensively adopted, such methodologies had to be abandoned since it was impossible to enforce such rigid communication protocols. With the exponential increase of the generated data, it became apparent that standardisation of the data was required after it was generated and not beforehand, as in prior methodologies. Towards this post-publication data integration approach, several frameworks, like ARTEMIS-MOMS, Cupid, "similarity-flooding," and iMAP, were suggested to have reconciliation at the target schemas and have federated queries executed across all of them (Banek et al., 2007). These methodologies/techniques utilise linguistic matching, translation into graphs and semantic matching; however, as

Banek highlights, all of these techniques will produce candidates that will have to be confirmed or “taught” using examples (e.g. Neural Network training). Full automation is almost impossible since human intervention is required. The same applies to matching diverse information within the data lake.

Instead of federated queries, the concept of an aggregated model/meta-model can be used (Banek et al., 2007). This model will contain information on linking the metadata of the distinct models into an aggregate model, which will be used in the reconciliation or definition of the heterogeneous schemas/information storage. The Heterogeneous Data Quality Methodology (HDQM) presents an approach where the focal element is the Conceptual Entity, an abstraction of any single phenomenon of the real-life instantiation of an entity of interest (Batini et al., 2011). If the concept is adapted to Big Data, it could address some of the *Variety* challenges, and an essential factor would be how to automate such an identification process.

Wang explicitly refers to Big Data Heterogeneous Data, but the research is mainly confined to identifying/raising the challenges instead of proposing alternatives for overcoming them. Of interest is the new term *Data Swamp,* which is used to identify ungoverned Data Lakes, where data is dumped without any metadata which will lead to confusion and limited usage since the semantics and structure of the data will be unknown (L. Wang, 2017). The data swamp can have devastating implications for an organisation due to the regulatory implications of data confidentiality. In addition, fuzzy data sets are of little use to data scientists since they will have to first format and break down the data before using them.

3.3. Current State

The following sections will present the advancements and current state in academia concerning Big Data V's subject areas. It is essential to cross-reference with actual production implementations in the market and business world. In this way, not only the theory but also the implementation and value of the solutions can be identified.

In order to understand the *Variety* challenge and verify the possible contribution to science, the basic V's will be presented based on their "representation" in literature. Volume, Velocity and *Variety* are consistently identified in documents covering Big Data. Thus, it is no surprise that the terms can be identified, through a search mechanism, in all four hundred and seventy-five (475) documents present in the project library. The context and extent to which they are investigated, though, vary, and for that, it is essential to actually go into the documents themselves.

3.3.1. Volume

Volume has been a field of research for IT with the primary manifestation the Teradata established in 1979 and Netezza founded in 1999, which IBM acquired in 2010. These companies were responsible for providing the first proprietary distributed storage and computing power systems. Their architecture mainly consisted of coupling disks with CPUs to manage segmenting and processing high volumes of data with better performance than conventional RDBMSs (e.g. Oracle, SQL Server etc.). Nevertheless, the problem was not fully addressed since data kept growing and the proprietary technology was expensive. Alternatives were sought and identified. It is not a coincidence that pioneers in this implementation were "web-tycoons" like Google, Amazon and Apache. The sheer Volume existing in the World Wide Web enforces these companies to invest and identify solutions for scalability. Once the respective companies identified the solution, it was a matter of time before they would market their new product. Apart from solving their own Volume problem, they would offer this capability to the world.

The technique used to address Volume challenges was known based on the early work of proprietary systems, but the concept is very old. The ancient Greek politics & military technique formulated by Philip of Macedonia, father of Alexander the Great, also utilised by Cesar and Napoleon, of "Διαίρει και Βασίλευε" / "Divide and Rule/Conquer" was employed and became the core concept in taming Volume. As Jayakumar, Patil, Singh and Joshi describe in their publication "Evaluation Parameters of Infrastructure Resources Required for Integrating Parallel Computing Algorithm and Distributed File System" in 2015, parallel computing along with Distributed File Systems were researched and evolved into viable production solutions.

Hadoop, which is often utilised as a synonym for Big Data, is one of the new implementations introducing a distributed file system and parallel processing capabilities. Sivaraman and Manickachezian, in their 2014 conference paper "High Performance and Fault-Tolerant Distributed File System for Big Data Storage and Processing using Hadoop," identify MapReduce and HDFS implementation in the Hadoop ecosystem as a fast-growing ecosystem gaining wide acceptance. Nevertheless, as indicated by Gartner, Big Data innovation was not assimilated by the business world. Gartner Hype Cycles for Emerging Technologies for 2012 and 2013 identified that the prediction of Big Data adoption had changed from 2-5 years into 5-10 years, indicating that there will be a delay since the respective "new invention or innovation" is not mature enough for business adoption.

Although Sivaraman’s and Manickachezian’s paper is very detailed, providing technical and technological details along with real-life business implementation, it does not contain any financial data. As a result, it is challenging to evaluate the solution proposed in the paper in terms of value brought into the business by implementing such systems.

In trying to address the cost of processing power and storage, several solutions were identified, and the first was to move away from proprietary hardware to “vanilla machines.” In this way, the cost of the technology stack in respect to hardware was substantially lowered by utilising PCs/servers that could be bought off the shelf at low consumer prices. The next step identified and implemented was to utilise the “Cloud.” In this case, the TCO was even lower since the resources were utilised as needed without permanent investments. The cost-benefit analysis in most cases proves that hosting an environment in the cloud would minimise CAPEX whilst OPEX would be similar to on-premises implementations (Fujitsu, 2014). Researchers like Zhang and Kaur have worked on the efficiency and benchmarking of Big Data cloud solutions (Kaur & Sood, 2017; X. Zhang et al., 2016).

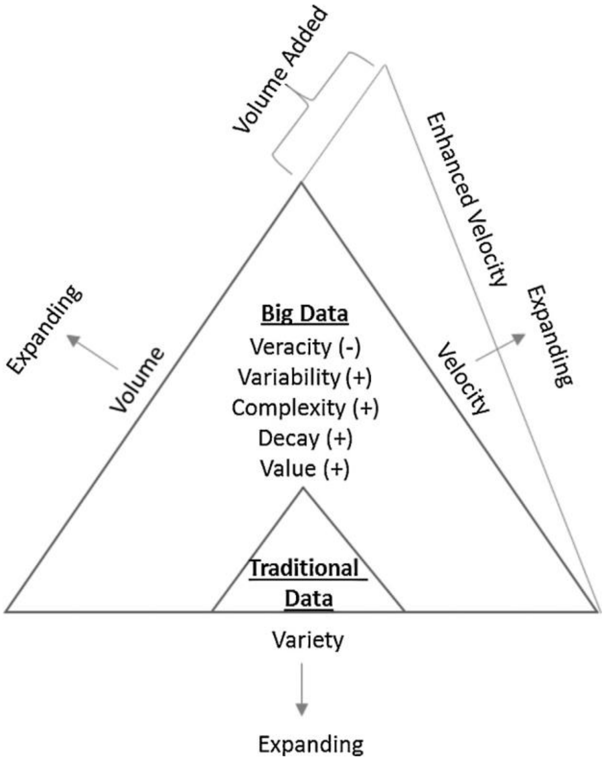
With the aforementioned innovative implementations in respect to commodity servers and the cloud, business adoption started. The cloud proposition can be efficient in terms of performance and cost thus a viable option in implementing a Big Data solution. However, many businesses, especially the highly regulated ones, e.g. banking, finance, medical sectors, had to satisfy governance and security. Sakharkar (Sakharkar et al., 2017) proposes the combination of private and public clouds to address security but most notably, the use of cloud and SaaS by medium or small enterprises to harness the power of Big Data with affordable investment. In the paper, a value proposition is available in affordably minimising security risk, but actual CAPEX, OPEX or ROI are not identified and discussed. Further research reveals that security is gaining the attention of researchers since it is a business challenge that has to be addressed. Based on this trend Elsevier, in 2017, has attributed a special issue of Digital Communications and Networks journal in realising and exploiting “Big Data security and privacy” (Yu et al., 2017) while Hadeer, Hagazy and Khafagy in their paper (Mahmoud et al., 2018) try to address the security of Hadoop implementations by incorporating encryption techniques in the MapReduce framework. In line with this approach for governance and enhanced security, “Corporate Flavoured” Hadoop implementations have been marketed. Such implementations that have gained acceptance are Cloudera (Gartner, 2018), Hortonworks, Confluent and Waterline (Brust, 2018). It is also important to mention here that large IT technology providers like Microsoft, Oracle, and SAP have also embraced Hadoop technologies and have their products integrated with them. Traditional pioneers in the appliance data space like Teradata and Netezza have also adopted Hadoop technologies, namely Hortonworks and Cloudera.

Researches have further investigated Volume by elaborating on the scale (Cheriere & Antoniu, 2017), flexibility (Segura et al., 2015), in-memory optimization (Zhaowei et al., 2017), a large number of small files (T. Wang et al., 2015), high performance (B. Schmidt & Hildebrandt, 2017) and other technical subjects pertaining to performance, reliability and cost. Based on its technical and well-manifested nature, Volume is constantly being researched in identifying new techniques and mechanisms of further enhancement. Siddiqa introduces a paper that stands out from such technological content research; Karim and Gani based on their technological survey on storage technologies, identified an analogy to RDBMS and presented a SWOT analysis. The paper focuses on Big Data taxonomy by identifying all available technologies and the consistency (Brewer’s CAP) theorem.

Based on the respective papers and previous chapters' references to technological innovation, investment and implementations, it is evident that not only in academia but also in the IT companies, Volume is researched and many implementations have “seen the light” in addressing the challenges.

3.3.2. Velocity

It is important to understand that all three V’s are interrelated and affect each other. Based on the “Integrated view of Big Data” provided by Lee (Lee, 2017), each V will, in essence, intensify the other V’s as shown in Figure 11. In this way, it is understandable that the Volume added will result in an expansion of Velocity.



Source: Lee, 2017

Figure 11. An Integrated view of Big Data

Velocity is the speed of change and represents the requirement of real-time processing. Speed, real-time and infrastructure bottlenecks are some of the terms used to describe the challenges posed by Velocity. In essence, what Velocity is all about is performance. The Volume challenge and its proposed implementations are not efficient enough to cater to the data processing before they become outdated.

To better understand the challenge, the geofencing paradigm utilised by financial institutions in their Big Data implementations will be described. As a credit card transaction is happening; it is captured by the switching system of the organisation and forwarded to the real-time decision engine to provide the customer with a personalized marketing message, offering another purchase/offer at a close-by shop. The decision has to be made in sub-seconds for the customer to be presented with the message. For a decision to be made, the entire transactional behaviour and customer position with the institution must be analysed to identify trends and offer the most intriguing sale proposition. This information would be irrelevant once the customer has left the location, and the marketing lead will be rendered useless. Similarly, in the case of web searches or online purchases, the lead will be short-lived before the customer closes the browser or makes another purchase, once again triggering the process. Suppose the concept is augmented with the number of users. In that case, it is understandable that Volume (e.g. the number of transactions X number of customers X historical consumer behaviour) will attribute to Velocity, and any latency in the systems will result in outdated processing, rendering the investment void and zeroing the corporate profit.

From the aforementioned case examples, it is understandable that Velocity and Volume go hand in hand, and in essence, most research for Volume is related to Velocity. Literature also indicates the same, Rustem Dautov is identifying, in IEEE 2017 publication “Quantifying volume, velocity, and variety to support (Big) data-intensive application development," that velocity is directly related to Volume since it refers to the rate at which data volumes are generated (Dautov & Distefano, 2017). Both Volume and Velocity are technology-oriented (L. Zhou et al., 2017), and as such, technological advancements can contribute to resolving the challenges. Performance tuning of DFS, parallel processing, in-memory processing (Kataria & Mittal, 2014), incremental clustering techniques (Srivastava & Dong, 2013) are techniques and technological implementations in addressing both.

On the other hand, technology means cost since there has to be an investment in IT. To reduce TCO, just like in the case of Volume, several researchers argue that the cloud could be the solution for Big Data processing. Chaowei Yang expresses an interesting view in "Big Data and cloud computing: innovation opportunities and challenges," stating that Application as a Service (AaaS) and Big Data as a Service (BDaaS) can support further infrastructure and business modelling. What is proposed is that optimised algorithms can be utilised and disbursed from such platforms to provide "ready-made" models for Big Data Analytics (C. Yang et al., 2017). The viewpoint expressed is a very interesting extension to infrastructure augmentation and minimisation of cost of ownership but what is not identified in this research is the value - especially when it is identified that Value has become one of the essential V's.

The usage of cloud based infrastructure and services is identified in providing viable alternatives to on-premises investment. The market for Big Data as a service is expected to reach $61.42 billion by 2026 and would include the layers of Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Data as a Service (DaaS) and Big Data Business Functions as a Service (BDaaS) / Data Analytics as a Service (DAaaS) (Kataria & Mittal, 2014; Rake & Kumar, 2020). Apart from the cost of ownership, in this review, it is identified that 90% of data is noise – pertaining to *Variety* – and that compliance and confidentiality can pose a challenge for many institutes since data on the cloud are "outside" the company. A solution is presented using hybrid clouds, consisting of a public and private cloud blend. This approach is utilised widely in highly regulated industries like banking, medicine, etc., where customer information is privileged. In addition to that based on the General Data Protection Regulation (GDPR) (EU) 2016/679 (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Da*, 2016), which prohibits sharing of customer information without customer consent private clouds to hold such information are constantly gaining acceptance.

Corporations are profit-oriented; thus, while they are interested in reducing TCO, they are primarily interested in revenue. Diversification and competitive advantage are tools in augmenting revenue, and to utilise Big Data, corporates must act on the produced insight. Manjul Gupta, in his 2014 paper "Organisational Culture and the Three V's of Big Data," presents a different perception of Velocity in relation to business, the essence of "reaction." He is identifying that Velocity is not simply the speed at which information is generated but also the speed at which the organisations can identify and act upon the intelligence by processing the data (Gupta, 2014). This link is crucial since it is the first time that Big Data is analysed under the prism of corporate structures and is also highlighted in Power's and McAbee's work (McAbee et al., 2017; Power, 2014). The research correlates the organisational culture with the V's to understand "success factors" in a Big Data initiative. What is missing from such research is the consideration of factors like compliance and international operations that can determine to a great extent the corporate culture. According to Munesh Kataria in 2014 article “Big Data: A review,” the answer to prompt business reaction is to utilise in-memory computing, thus promptly processing Big Data in realising “real-time” benefits.

Minimising TCO and “reacting” upon Big Data insight is important, but the time element should also be considered. Acting promptly is significant in acquiring and retaining the competitive advantage. In 2017 Brock’s article entitled “Are enterprises ready for Big Data analytics? A survey-based approach” is concerned with Velocity. After identifying the issues with networks and computing power, an important analogy is made. For analytics to be of business usage, it is identified that they have to be time-relevant. This resembles the 1980’s Japanese break-through of Just-In-Time (JIT) management style (Brock & Khan, 2017), where goods were presented when the need arose to minimise costs. In this case, data will be utilised in real-time to present insight through analytics as the challenge itself manifests in an attempt to minimise any consequences. Towards this real-time processing, Ayse Selcuk proposes near-real-time processing using “lazy updates” of data in the data sets, which is essentially an incremental approach (Selcuk et al., 2015). This incremental suggestion coincides with Storey’s analysis, where it is suggested that filtering and conceptual modelling can be used to extract and process important information (Storey & Song, 2017). It is interesting to observe that Bertil Schmidt and Andreas Hildebrandt, in their 2017 review, have not even mentioned variety as a “motivator” for the use of clouds and Big Data technologies. They only consider Volume and Velocity as the main drivers and refer to the existing implementations, like Map-Reduce, which are already available.

Once again, just like Volume, we can see that there are several implementations and alternatives to overcoming the challenges presented by Velocity. Technologies would present the main alternative, where software and hardware augmentation can be easily implemented at affordable rates. Academic and field research is performed to further enhance performance and benchmark existing implementations along with innovative alternative solutions. Also, it is imperative to mention that compliance is factored into the presented solutions, making it easier for business adoption.

3.3.3. Variety

Although there is adequate literature and research on Volume and Velocity, the review on *Variety* revealed a different trend in the documents. In this case, it is identified that researchers are not investigating but rather highlighting the fact that *Variety* is not researched enough whilst at the same time highlighting the challenge posed. In identifying the “Current State” for *Variety*, “Heterogeneous Data” was used to uncover papers and relevant research. The initial results were massive, which was an encouraging finding after the limited information gathered for variety. After an in-depth analysis with documents overview and full document evaluation, it was identified that most of the research was specific to instantiations of heterogeneity.

In many cases, specific datasets were combined. The field case studies were reflected in the research. On many other occasions, the focus is on storage and structure availability to store diverse information. In essence, although they are entitled to or have utilised heterogeneity in the keywords, the content was related to Volume and NoSQL solutions already presented. Of course, some publications provided insight and will be analysed in the following paragraphs.

We shall split the cited work accordingly to utilise the initial categorisation of the topic and time dimension into a) Variety and b) Heterogeneity. *Variety* related services in any Big Data project will be the “key cost element” of the TCO (Kumar, 2013). Services will be part of the initial investment (CapEx) and will have no tangible result but are a prerequisite to incorporating the raw data into the Data Lake and eventually utilising them in analysis, thus any cost approval will not be readily justifiable, especially if it is a substantial percentage of the overall budget. In certain corporate cultures, even though it might not be the case, it is considered that internally allocated resources will have lower costs. In the case of *Variety*, Lennard, Shacklett and Brown all agree that effort has to be put in by specialised personnel, and the respective specialisation is not present in many organisations, thus resulting in hiring, which will augment the costs as identified in the following analysis of their work.

Lennard has named *Variety* “the ugly duckling” of Big Data since it is the most neglected and difficult to solve (Lennard, 2014). Nonetheless, it is stated that organisations should not be discouraged and try to leverage their data power by implementing a data landscaping exercise in which the used data formats are identified. A limitation to this approach is that such an exercise can be challenging, requiring specialised skills and be time-consuming, especially when we superimpose the other V’s effects. Shacklett is verifying in the respective paper that it could be challenging to implement for any organisation if they have to evaluate and classify all existing corporate data. It could be impossible to classify all business-related external data (Shacklett, 2014).

Up until this point in time, the researchers have considered *Variety* a “difficulty” in merging and classifying data. However, researchers diversify in arguing that the impact is far more significant in the following years. It is identified that errors and controversial results will be the outcome of *Variety* (Ali-Ud-Din Khan et al., 2014). The arrival to erroneous conclusions based on data *Variety* was also identified in a case study on the development of an ATM Cash Replenishment Artificial Neural Network (Vranopoulos et al., 2016). It was proven that the quality of results dramatically improved after data normalisation. Organisations have understood the value of minimising the effect; as Kimura denotes, 74% of businesses would like to harmonise their data (Kimura, 2014).

Nevertheless, Kimura is also introducing another perspective to *Variety*; it is suggested in his work that Value can be harnessed from *Variety's* identification. Instead of focusing on analytics in analysing the data, it is suggested that the differences and diversity can be analysed to better understand the business environment or the natural/biophysical environment. This approach can be utilised, but it is not a viable solution to our understanding. Although it can provide insight, it is at the "end of the road." No business will undertake the "Big Data Voyage" to solely or primarily realise benefits from better understanding its environment diversification.

In the subsequent years, the trend of identifying the issue continued with Baker signifying that *Variety* would be the biggest challenge by quoting Sharmila Mulligan, CEO and founder of ClearStory Data, who in turn identified *Variety* as the biggest challenge in providing valuable business insight (Baker, 2015). In his 2015 publication in IEEE, Mao is particular "*However, most of the effort was spent on volume and velocity, but not as much on variety*" (Rui et al., 2015). Universalisation, abstraction and standardisation are suggestions towards minimising the effect, but no business cases or adoption are identified in the research. Moreover, even in most recent publications, it is identified that the solution is not evident. The challenge has to be addressed, like in a IEEE publication in 2017 by Rustem Dautov, who identifies that *Variety* is “more difficult” to measure and quantify (Dautov & Distefano, 2017) even though the research sheds some light on quantifying the *Variety* effect on a scale of 0 to 1, whilst previous work and research is mainly limited to identifying the challenge.

Data Heterogeneity is not something new, and the study of data integration started in the middle 1970s (Luo et al., 2008). Methodologies like CORBA, DCOM, EDI were devised and employed, but they all focused on preventing *Variety*. This was sufficient for the early years, but such methodologies were abandoned when the world wide web came into existence and was extensively adopted. With the exponential increase of the generated data, it is crucial to standardise data after it is generated and shared rather than beforehand, thereby standardising the communication layer with protocols and conversions. An approach suggested by Banek is to reconcile the target schemas of different databases and have federated queries executed across all of them. Sub-queries will be constructed and executed based on the initial user query for each schema based on previous mappings. Several approaches have been developed to match schemas like ARTEMIS-MOMS, Cupid, “similarity-flooding,” and iMAP (Banek et al., 2007). The aforementioned methodologies/techniques utilise linguistic matching, translation into graphs and semantic matching. However, most importantly, as Banek highlights, all these techniques will produce candidates that will have to be confirmed or will have to be “taught” using examples (e.g. Neural Network training). Full automation is almost impossible since human intervention is required. The same applies to matching diverse information within the Date Lake. An automation process can be devised in helping data scientists. Also, it is imperative to minimise the repetition of this task by standardising and disseminating this information to other data scientists, thus minimising re-work.

In sharing and reusing domain knowledge, ontologies can be used. In their work, Luo – Dang and Mao define ontology as the canonical and precise description of one domain that explicitly denotes the class and attributes of various features, conditions and properties related to conception. This paper is important since it links ontology to object-oriented models and highlights XML as the standard that can make the combination of heterogeneous data easy (Luo et al., 2008).

In essence, Banek, Luo, Dang and Mao suggest the use of meta-models which will be used in the reconciliation or definition of the heterogeneous schemas/information storage. This concept is further analysed by Batinmi, Barone, Cabitza and Grega in their work "A Data Methodology for Heterogeneous Data," where they present a Heterogeneous Data Quality Methodology (HDQM). In their work, they rely on meta-model definition and once again identify the importance of XML as the representation means. After investigating methodologies like TDQM, TIQM, AIMQ, CIHI, ISTAT, COLDQ, DaQuinCIS and CDQ, they have concluded that limitations to the number of attributes or type of data or datatypes exist and propose HDQM as a superset in addressing these limitations. The fundamental concept of the HDQM meta-model is the "conceptual entity," an abstraction of any single phenomenon of the real-life instantiation of an entity of interest (Batini et al., 2011). This concept is fundamental since it can be applied in the Big Data ecosystem and used to describe the data conceptually, regardless of their nature, in an attempt to contain the effect of *Variety* in respect to defining and understanding the data. Also, the work identifies that data quality can best be described in two dimensions:

accuracy and currency. Based on prior references throughout this document, we can easily relate these two dimensions to the underlying challenges posed by *Variety* and Velocity.

Further research in meta-models has been identified, but a paper that stands out for its relevance to Big Data Volume is “Beyond 100 Million Entities: Large scale Blocking-based Resolution for Heterogeneous Data.” Although Papadakis and his team labelled it as “Heterogeneous Data,” their work refers to Heterogeneous Schemas by identifying a methodology to match a large number of entities. Their proposition is based on the sorted neighbourhood approach and suggests that instead of working on the entire collection, it is sufficient to handle data inside each block (Papadakis et al., 2012). This approach is quite similar to Map Reduce and can thus be adapted to the Big Data ecosystem.

Mylka is introducing the concept of a knowledge base. The respective paper represents the X2R methodology which is used in defining an ontology knowledge base to describe the conceptual entities of an RDBMS, XML and LDAP data source (Mylka et al., 2012). Although it cannot be directly related to Big Data since all data sources have a high degree of structure, the research exhibits the sue of high volumes and the aging of source data, thus suggesting the usage of the meta-model.

Up to this point in time, most research highlights the merge or reconciliation/matching of schemas in addressing data heterogeneity. Abdullin and Nasraoui, in their work, extend on Papadakis’s work in clustering and instead of performing it on schemas, try to access the data itself. Semi-supervised learning of neural networks implementations is used in clustering labelled and unlabelled data (Abdullin & Nasraoui, 2012). Olaru, in the PhD thesis is utilising existing entity representation methodologies like Entity Relationship Diagram (ERD), Data Flow Diagram (DFD) and Data Normal Forms -3rd Normal Form, see details in Appendix IV - DFD, ERD, ETL and NF, in representing conceptual schemas (Olaru, 2014), which we believe can be also utilised in the Big Data ecosystem to present the aggregated structures/no-structures into a knowledge base.

More relevant to the Big Data structures is the work of Karpathiotakis and his team, who suggest that although Map Reduce and NoSQL systems can attribute substantial benefits to the data-driven analysis, it would be more efficient to have a federated query system that will benefit from the capabilities of each subsystem rather than trying to bring everything into one repository. In this way, their suggested system, Prometheus, utilises query algebra in translating the respective initial query to the downstream specific system enquiry, thus harnessing the native power of each infrastructure (Karpathiotakis et al., 2016). RDBMS, NoSQL, Map Reduce and other technologies provide the baseline to be accessed via a federated query mechanism. This approach is exciting since, in essence, it limits the exposure to *Variety* by utilising natively any structured resource (e.g. RDBMS) rather than having the data moved into any Big Data implementation infrastructure. Oracle and Microsoft implementations are following the exact opposite route by importing the data into a Data lake and providing “relational query” interfaces to it. So there seems to be a “deviation” between the academia and market approach; thus, no conclusive decision can be presented in the approach that must be followed.

Wang explicitly refers to Big Data and Heterogeneous Data, but apart from identifying/raising the challenges, there are limited alternatives proposed to overcome them. Of interest is the new term “Data Swamp,” which is used to identify ungoverned

Data Lakes where data is dumped without any metadata that will lead to confusion and limited usage since the semantics and structure of the data will be unknown (L. Wang, 2017). Also, there is an attempt to explain *Variety* in terms of heterogeneity types, where the following are identified:

- Syntactic, refers to language differentiation.
- Conceptual/semantic/logical, referring to differences in the modelling of the domain under investigation.
- Terminological, where the same entities are referred to under different naming conventions.
- Semiotic/pragmatic, which has to do with the differences in the human interpretation of the entities.

Another pillar of *Variety* that refers to erroneous and incomplete data, sometimes referred to as "noise-data," is investigated by Nazabal. In the respective work, it is identified that there is no clear description in the literature in incorporating missing data (Nazabal et al., 2020). The paper suggests the usage of Variational AutoEncoders (VAEs) in completing and extrapolating missing information with the use of ANN. The initial results show that the methodology can be utilised in future business-related implementations.

Variety refers to the multiplicity of the sources and formats, but it also refers to the diverse context and usage of the data element. The plethora of different formats, standards and notations have increased the risk of identifying personal information, thus augmenting the risk of disclosure (Cuquet et al., 2017). In minimising such risks, the governance of data custodians' activities through legislative frameworks has increased, and information is increasingly being regulated in multiple sectors (OKeefe, 2017). These acts include, the Personally Identifiable Information (PII), Public-Sector Information (PSI) Directive, General Data Protection Regulation (GDPR) law, and Payment Card Industry (PCI) security standards along with anonymization standards like European Medicines Agency Policy 0070 or the Health Insurance Portability and Accountability Act along with "soft law," as classified by the Organisation for Economic, Cooperation and Development (OECD), like the International privacy framework standard IS/EEC 29100 or the Generally Accepted Privacy Principles (GAPP) developed by the American Institute of Certified Public Accountants (AICPA) and the Canadian Institute of Charted Accountants (CICA) (Brandon & De Souza, 2017). These regulatory acts can have an immediate financial impact in the form of fines thus constituting the Data Loss Prevention (DLP) risk, making Data Confidentiality a candidate for automated identification.

To date, scholars have suggested and implemented several techniques in ensuring Big Data privacy throughout the Big Data life cycle, which is composed of different stages, i.e. generation, collection, storage, processing, analytics, utilisation and destruction (P. Jain et al., 2016; Koo et al., 2020). In the initial stages of generation and collection, most scholars would limit the risk exposure by restricting access or restricting information (OKeefe, 2017). The restriction of information, which is essentially nothing more than abolishing or falsifying confidential and personal information, is usually suggested by many authors to be done with anonymization of data (P. Jain et al., 2016; Rai & Sharma, 2020). In achieving anonymization or depersonalization of the data, many techniques are available. From deleting or hashing the data to the more sophisticated techniques of micro-aggregation, e.g. l-diversity, t-closeness, matrix anonymization, k-Anonymity etc. (Domingo-Ferrer & Rebollo-Monedero, 2009; Ninghui et al., 2007; Rumbold & Pierscionek, 2018; Sei et al., 2019; B. Zhou & Pei, 2010). The use of such quantitative techniques will bring us to the Statistical Disclosure Control, which seeks to balance between data disclosure and data loss (Domingo-Ferrer & Mateo-Sanz, 2002; Gouweleeuw et al., 1998; Oganian & Domingo-Ferrer, 2001; Rumbold & Pierscionek, 2018). This will lead us to human intervention in deciding what and how the data will be anonymised or depersonalised.

In implementing any anonymization or depersonalization approach the possible data elements of interest must be identified. The most popular methodologies for data matching is to use either regular expressions or proprietary parsers (Dalvi et al., 2021). Although in academia regular expressions are criticised as hand-optimised methods, they serve as a fast and simple implementations with minimal infrastructure requirement as opposed to Machine Learning (ML) models (Markov et al., 2021). But there could be a serious limitation from the high number of false positives RegEx might match against invalid instantiations (Saha et al., 2020). In particular when working with confidential information and data leak detection a common problem is the quantum of false positives generated (Lounici et al., 2021). This high number, if combined with the Volume of Big Data will result in making the identification highly unsuccessful if not impossible. In addition to that the different origins and formats that constitute *Variety* tend to make the RegEx identification patterns complex and ineffective due to programming and syntactic errors (Spishak et al., 2012).

For any organisation to effectively and efficiently manage the process of data sharing, whether internally, e.g. test environments, or externally, to support vendors, based on the aforementioned challenges about Data Confidentiality, there has to be a framework or a formulated process. This research seeks to provide the guidelines for having an enterprise-wide precise, and safeguarded operational procedure. As mentioned above, people have conflicting interests, so the ratio between rendering the data unusable and data disclosure will have to be formulated in a systemically facilitated and implemented strategy. In this way, multiple iterations on the nature and usage of the data can be avoided.

3.4. Conclusion

An in-depth analysis and presentation of certain aspects of Big Data were performed. The basic V's, Volume, Velocity and *Variety* were presented as identified throughout the literature. Identifying and correlating the strengths and shortcomings of publications was performed in order to identify the V's research depths.

Although there are many business and technical references in relation to Volume and Velocity, it was identified through this research that there is a limited investigation in trying to address *Variety*. Whilst there are references where the authors identify the magnitude of the challenge, minimal solutions are presented. Abstraction, generalisation, identification, classification are several techniques mentioned, but no business case is made, and more importantly, there is no framework in bringing them together and automating the task. Even investigating the long-standing issue of heterogeneous data did not help much apart from giving hints towards metadata utilisation and use of taxonomies and ontologies.

Since the web was the first source of "ambiguity" in the structured world of IT, techniques that were used in classifying the www could be utilised in addressing data heterogeneity. Building on such techniques and methodologies as ontology, it might be possible to understand and address some of the *Variety* challenges.

Chapter 4. A Novel Approach to Tackling Variety

4.1. Introduction

In the previous chapters, the Big Data ecosystems were presented along with the prior art in understanding the technologies and concepts relevant to *Variety*. It was established that *Variety* is a source of challenges, and alternatives to address the respective issues were sought. The academia and respective resources were also presented in identifying the current state.

An approach towards identifying a methodology along with an experimental approach will be presented to facilitate the countering of the identified *Variety* challenges. The data ingestion journey is enhanced with techniques that will aid the data scientist to perform their work faster, by automating parts of it, and safer by identifying and managing confidential data. Section 4.2 sets the stage by identifying the challenges that will be addressed in Section 4.3 with a resolution approach.

4.2. Problem Definition

In several places, it was identified that *Variety* is a challenge, and it was also established that limited solutions had been identified. Due to the limited literature identified under *Variety*, data heterogeneity has also been researched, and similar behaviour existed. In most cases, the problem identification was limited to its actual instantiation, e.g. correlating specific data sets, rather than providing a methodology.

Before suggesting any course of action, it is important to recap the challenges identified. The following list cannot be considered exhaustive but will provide an excellent basis for understanding.

- Storing semi-structured and unstructured data was one of the initial *Variety* manifestations. Although new types of databases like NoSQL have emerged in addressing storage, other dimensions like quality, proliferation, outliers, context, and coherence could not be resolved by simply enhancing the storage process.
- Implementation of tools is relatively limited, if any, even though there are implementations of algorithms like Sharding and Map Reduce that utilise parallel processing to tackle Velocity.
- Abstraction is a suggested approach in tackling *Variety*, but it is a human dependent and effort-intensive task. Thus, making this approach costly since specialists must be employed to perform such tasks.
- The multiplicity of formats and units of measurement is also a known risk that stems from *Variety*, leading to erroneous conclusions.
- Outdated data related to the Velocity can also pose a threat leading to incorrect decisions since the ecosystem has changed and the decision itself has become outdated.
- Context and coherence within an environment are substantial issues since it is known that "siloed" data will lead to a partial understanding of the factors. Also, series with no context will transform the Data Lake into a "Data Swamp."
- Unification of data and the diversity of datatypes is considered one of the "greatest" challenges that will cause data scientists "not to touch the data."
- Unstructured data is relatively new in the data space; thus, knowledge of handling them is limited. The knowledge curve for practitioners is quite steep, and this has a financial and timeline impact on Big Data initiatives.

All above mentioned will contribute to making *Variety* a challenge, but if further analysis is performed, it can be identified that there are other challenges, none-IT related. Corporate culture can affect Big Data initiatives and will influence all V's, including Variety, making it even harder to overcome. Social unjustness and personal bias on behalf of data scientists can pose a social issue related to Variety. In their attempt to normalize the data, subjective decisions are made, which in many cases are not even recorded for future reference. Acquiring data without consent can be both a legal and ethical issue, but tempering with it in trying to minimise Variety is more treacherous; thus, there has to be proper governance and control when "tempering" with the data.

Last but not least, on top of these general concerns, there are sector related issues and regulations that will further intensify the effect of Variety. Any proposed solution should also consider the aggregated business world and cater to specific industry issues, especially in the case of highly regulated environments like banking or health.

4.3. The Novel Proposal

In the research performed to date, it was identified that not many scholars have been working on resolving the issue. Instead, most are trying to formulate and describe the phenomenon. Even when data heterogeneous concepts were referred to, which have been known and tackled since the '70s, limited aggregation and standardisation were identified. References to metadata, context, aggregation, conceptualisation, taxonomies, and query synthesis are identified as possible solutions, but there seems to be no methodology available. Most research is performed using a siloed approach to solve the problem at hand. For example, merging two database/dataset schemas, identifying the correlation of two specific datasets, identifying risk of non-compliance, increased TCO due to knowledge, and other instantiations.

The literature available was mainly evaluated under the prism of business acceptance and possible implementation. The experience of data manipulation and usage in a highly regulated environment was utilised in qualifying the solution viability. Also, based on this experience, governance, security, risk and audit angles were evaluated and highlighted. Another parameter was the actual cost. In recent years, we have faced monetary crises in several countries, e.g. Cyprus, Greece, Spain, and Italy; cost and TCO are used in evaluating any proposition or suggestion. Further to these, going into the details and cross-referencing possibilities and concepts were proven helpful in identifying a possible course of action.

The work contained in this thesis suggests and further analyses a framework that, based on contextual data, metadata, and existing techniques, will try to standardise and automate the process. In this way, methodologies like DFD, ERD, and others will be used to depict and define the datasets and their properties, see details in Appendix IV - DFD, ERD, ETL and NF, whilst technologies like Neural Networks and basic scripting will be used to automate some tasks.

The standardisation will provide a more governed and robust environment where there will be traceability and reusability. The automation will have a dual effect; a) it will provide data scientists with a “jump-start” where they will have the capability to get some pre-populated information instead of starting from scratch, which will give them more “quality time” interpreting rather than identifying the data, and b) it will reduce the amount of prejudice and personal influence that can “pollute” the data.

When it comes to Data Confidentiality, it is suggested that a framework for cross-corporate standardisation will be proposed. The rule-based system will be capable of providing an impartial approach using workflows and algorithmic enforcement of corporate-wide policies.

Big Data is the trend in data science, and *Variety* has come into the picture. The dimensions and challenges are many and diverse, but it is believed that a methodology can help. It is not suggested that this approach and methodology will fully solve the issue but it will help identify and understand the issue better. Also, it will provide a framework in which business environments and the user will have some reference and thus serve as implementation guidelines.

The business world is risk-averse, and *Variety* is increasing, by nature, the possibility of risk manifestation, may it be by introducing false positives, data source structural uncertainty or data loss. It is imperative, if not to solve, to at least identify and quantify the possible risks of a Big Data initiative for decision-makers in the company to absorb or not. If a structured approach and a methodology are identified, this process of risk identification will be supported, thus making such initiatives more tangible.

Data Origination, Data Format & Breakdown and Data confidentiality are depicted as part of the Big Data ingestion process in Figure 12. The proposal utilised existing elements from the data management space and builds upon them in counteracting the *Variety* effects.

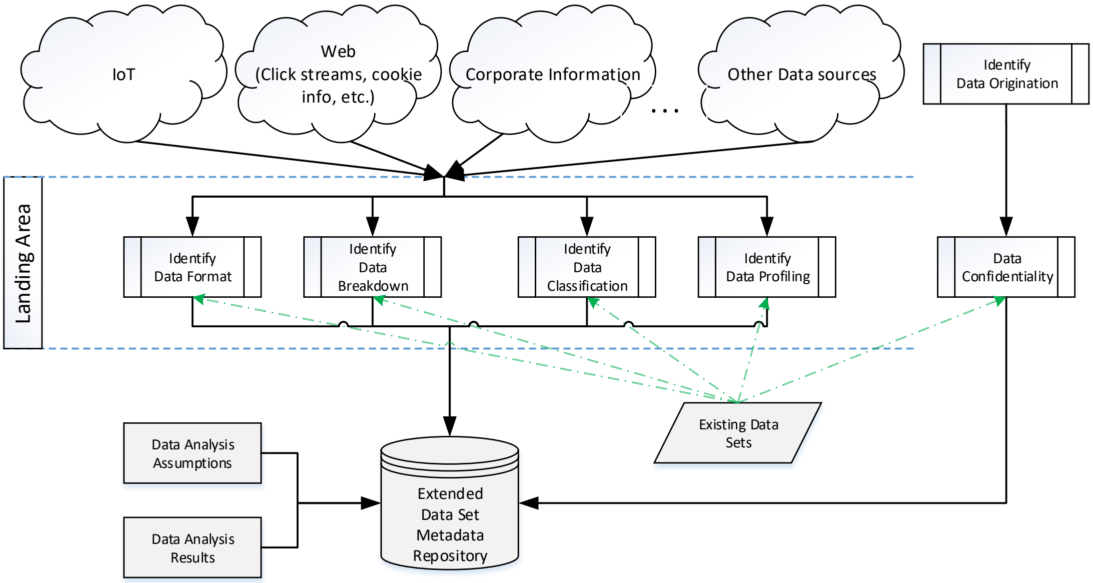


Figure 12. Proposition Steps/Stages

As mentioned earlier the Web, IoT, internally created corporate information and other data sources compose the surrounding environment and is the stimuli for data ingestion. In managing the information flow, the concept of a landing area is utilised so that information is properly processed before being incorporated into the Big Data lake. By doing so, the data integrity, confidentiality, coherence and validity is safeguarded in making sure that the lake is not gradually transformed into a swamp. Building upon these concepts the following sub-processes of ingestion are highlighted:

a) Identify Data Origination, which, although unmanageable will provide contextual information for the processes to follow in regard to the sanitisation of the data and possibly the structure/format.

b) Identify Data Format is related to automatically identifying the data format, whether it is binary information like pictures video or sound, textual information like data files, emails, reports and documents or event-based information like click streams, logs and of course a number of other formats and flavours which constitute part of the *Variety* challenge.

c) Identify Data Breakdown is used in disassembling the information into logical chunks that can be autonomously used in data mining.
d) Identify Data Classification is used in to further break down, analyse, and understand the nature and correlation of the data.
e) Identify Data Profiling is used in validating the work done in prior stages, where the employment of data visualisation, can confirm the usability of the data before being released for actual insight processing.
f) Data Confidentiality is an integral process where the incoming information has to be identified and the metadata information is recorded in a consistent manner.

Existing information from prior analysis, being data sets or metadata about the sets, is retrofitted into the system and can provide the basis for automated AI solutions. Additionally the repository holds the analyst’s prior work, cataloguing assumptions and prior decisions about the data, such as the confidentiality classification or the data set statistical information and rulesets used.

Throughout the data ingestion journey outlined, this work is focused on confidential data identification optimization, data characterisation and data confidentiality by tackling the respective challenges.

4.3.1. The False Positives Challenge

Empirical evidence suggests that most of the existing systems available in the market mainly rely on RegEx and lexical lists, which will produce a high number of false positives (Koenig, 2019; Solbers, 2012). Looking to improve the hit-rate and expand on the art in regular expressions, this research sought to introduce means of refining the outputs and reduce the misclassifications and possible errors.

An example of low hit-rate, would be the pattern of a credit cards. In that case, any number of 15 or 16 digits will be identified as a possible candidate; samples are available in Table 4 (Goyvaerts, 2021; *Regex - Algorithms for Detecting Credit Card Number Reducing False Positives/Negatives - Stack Overflow*, 2013). By reporting any random 15 or 16 digit number as a possible credit card, many false positives will arise since not all such numbers are credit cards. Any customer number or account number will be reported as confidential information event if it is a simple sequence number of minimal significance with no governance requirements. Such "noise" can increase the risk of missing actual occurrences of confidential data, which could lead to a compliance incident.

Table 4. Credit Sample RegEx

Confidential Data	**Regular Expression**
mastercard	(?:5[1-5][0-9]{2}|222[1-9]|22[3-9][0-9]|2[3-6][0-9]{2}|27[01][0-9]|2720)[0-9]{12}
VISA	4[0-9]{12}(?:[0-9]{3})?
AMERICAN EXPRESS	3[47][0-9]{13}
Diners Club	3(?:0[0-5]|[68][0-9])[0-9]{11}
DISCOVER FINANCIAL SERVICES	6(?:011|5[0-9]{2})[0-9]{12}
JCB	(?:2131|1800|35\d{3})\d{11}

The same would apply for PII governed elements such as identification card (ID) numbers, passport numbers or social security numbers, e-mail, physical address, internet protocol (IP) address, media access control (MAC) address. For more generic data elements like account numbers, International Bank Account Number (IBAN) and customer numbers which tend to be sequential numbers, this will produce even more false positives since any number would qualify; samples are available in Table 5 (Farenda, 2017; *IBAN Regex Design - Stack Overflow*, 2017; *RegExp Library Formats*, 2019; Santiago, 2005).

Table 5. Personally Identifiable Information Sample RegEx

<table>
<tr><th>Confidential Data</th><th>Regular Expression</th></tr>
<tr><td>Persian Gulf Countries Civil ID</td><td>\d{1} (?!00)\d{2} (?!00)\d{2} (?!00)\d{2} (?!0000)\d{4}</td></tr>
<tr><td>Danish Civil ID</td><td>[0-3][0-9][0-1]\d{3}-\d{4}</td></tr>
<tr><td>Greek Civil ID</td><td>[A-Ω]{1,2}[0-9]{6}</td></tr>
<tr><td>Social Security Number</td><td>(\d{3}-\d{2}-\d{4})|(\d{3}\d{2}\d{4})$</td></tr>
<tr><td>United Kingdom Passport</td><td>[0-9]{10}GBR[0-9]{7}[U,M,F]{1}[0-9]{9}</td></tr>
<tr><td>International Passport</td><td>[A-Z0-9&lt;]{9}[0-9]{1}[A-Z]{3}[0-9]{7}[A-Z]{1}[0-9]{7}[A-Z0-9&lt;]{14}[0-9]{2}</td></tr>
<tr><td>Indian Passport</td><td>[A-Z]{1}-[0-9]{7}</td></tr>
<tr><td>IBAN</td><td>[a-zA-Z]{2}[0-9]{2}[a-zA-Z0-9]{4}[0-9]{7}([a-zA-Z0-9]?){0,16}</td></tr>
<tr><td>eMail</td><td>(?:[a-z0-9!#$%&amp;'*+/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&amp;'*+/=?^_`{|}~-]+)*|\"(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])*\")@(?:(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\[(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.){3}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?|[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])</td></tr>
<tr><td>IP Address (v4)</td><td>([01]?[0-9]{1,2}|2[0-4][0-9]|25[0-5])\.([01]?[0-9]{1,2}|2[0-4][0-9]|25[0-5])\.([01]?[0-9]{1,2}|2[0-4][0-9]|25[0-5])\.([01]?[0-9]{1,2}|2[0-4][0-9]|25[0-5])</td></tr>
<tr><td>MAC Address</td><td>([0-9A-Fa-f]{2}[:-]){5}([0-9A-Fa-f]{2})</td></tr>
</table>

A confidence level is proposed to reduce the number of false positives and provide an accurate metric on the content of confidential data. In calculating the level, the existing RegEx methodology will be extended with multiple other metrics, which will serve as “booster metrics” in increasing the confidence of the match. Metrics like Soundex, proximity and structural confirmation/check digit calculations will be used. In further extending the system, a self-learning module for auto-identification of new RegEx patterns based on reoccurring probable Soundex matches can be added.

In addressing the false positives impediment in a structured manner, the system utilises four main classifications to accommodate the different types of confidential information to be identified, as depicted in Table 4 and Table 5. The “booster metrics” devised to implement the extended confidence level and their associated classification are presented in Table 6.

Table 6. Classification to "booster metrics" association

Classification	Search Condition	Booster Metrics
Card	list of RegEx expressions to get the initial data match	linguistic boundary characters, e.g. space, comma, quotation Luhn algorithm, for check digit verification institutional bank identification numbers (BINs)
Lists	list of RegEx expressions to get the initial data match	monitor terms proximity. The distance of the occurrence with words like password, account, card, credit, id etc. is calculated
Absolute XML	-	list of specific xml tags e.g.<CivilId>ID123456</CivilId>
Relative XML	-	list of XML tags containing terms<*Passport*> was * indicates any number of any character

Extending on the previously mentioned example of the credit card’s low hit-rate, and utilising the card classification from Table 6, not all 15 or 16 digit numbers will be reported as hits. With the use of “booster metrics” like the proximity of the words “Credit Card” or “CC” and the check digit calculation, the system will automatically filter out any low confidence hits. In this way the number of false positives will be reduced in an attempt to remove “the noise” from the data results.

The system was extended with a pre-processor and a co-processor to ensure the textual nature of the data fed into the system and compensatory controls for possible feed structural differences. In Figure 13, the pre-processor and a co-processor are depicted as part of the entire system process.

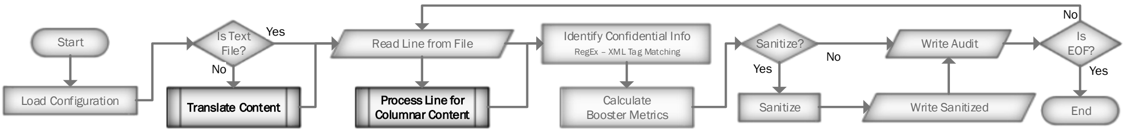
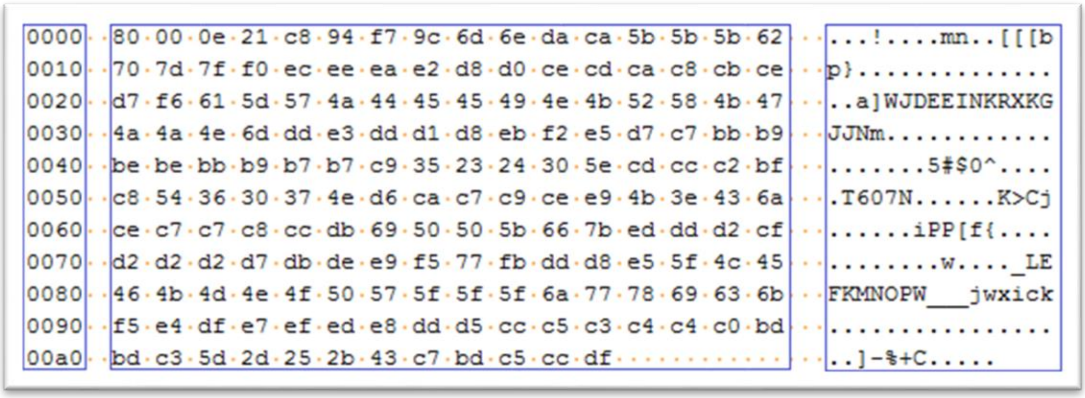


Figure 13. Confidentiality Process Flow

For the pre-processor, the Apache Tika project was used to ensure that any non-text feed or file, e.g. pdf, MsWord, MsExcel etc., were converted into a text format before being sent into the matching subsystem. For the co-processor, alternative readers were introduced on top of the traditional line reader, where block reading was enabled. In this way, for example, characters from line 1 from column 5 to 30 were merged with characters from line 2 from column 5 to 30 to make a logical sequence. This kind of formatting, shown in Figure 14, is most common in trace files where there are two columns, one with hex and one with text and in documents or pixel-perfect reports that utilise columnar writing styles.



```
0000  80 00 0e 21 c8 94 f7 9c 6d 6e da ca 5b 5b 5b 62   ...!....mn..[[[b
0010  70 7d 7f f0 ec ee ea e2 d8 d0 ce cd ca c8 cb ce   p}..............
0020  d7 f6 61 5d 57 4a 44 45 45 49 4e 4b 52 58 4b 47   ..a]WJDEEINKRXKG
0030  4a 4a 4e 6d dd e3 dd d1 d8 eb f2 e5 d7 c7 bb b9   JJNm............
0040  be be bb b9 b7 b7 c9 35 23 24 30 5e cd cc c2 bf   .......5#$0^....
0050  c8 54 36 30 37 4e d6 ca c7 c9 ce e9 4b 3e 43 6a   .T607N......K>Cj
0060  ce c7 c7 c8 cc db 69 50 50 5b 66 7b ed dd d2 cf   ......iPP[f{....
0070  d2 d2 d2 d7 db de e9 f5 77 fb dd d8 e5 5f 4c 45   ........w...._LE
0080  46 4b 4d 4e 4f 50 57 5f 5f 5f 6a 77 78 69 63 6b   FKMNOPW___jwxick
0090  f5 e4 df e7 ef ed e8 dd d5 cc c5 c3 c4 c4 c0 bd   ................
00a0  bd c3 5d 2d 25 2b 43 c7 bd c5 cc df               ..]-%+C.....
```

Figure 14. Columnar log file sample (3 columns)

In this way, by means of having better confidence levels through the "booster metrics" identifying more data with additional XML elements identifiers and columnar reading, the system presented a viable proposition for confidentiality enhancement automation.

4.3.2. The Data Identification and Breakdown Challenge

When a new data set is introduced, *Variety* will be in play, and the data scientist will consume time identifying the nature of the data and its structure. An automated system is introduced in standardising the process and identifying the incoming data. A file/set can be delimited with any character or sequence of characters. Common delimiters across the industry are comma, semicolon (comma-delimited files) or tab (tab-delimited files). This poses a challenge since these characters are also used as common characters in many digital forms, including documents, textual data, etc. In an attempt to incorporate this manifestation of *Variety* where the same character can have different contextual importance, the possible file delimiters were given different weights given their probable occurrence as punctuation marks in ordinary text. For instance, a comma would be more likely to be a punctuation mark than a tilde. As a result, tilde would have a higher weight attributed to it. The rationale behind this is to increase the significance of a reoccurring rare character combination as a separator against common character separators. Detailed information on the delimiters and associated weights attributed are provided in Section 6.2.

The system used multiple independent components, as seen in Figure 15, feeding information to each other. Pre-processors would be used in harmonising the data and performing the initial set of classifications:

a) Identify the nature of the file (Text Vs Binary) using MIME.

b) Identify the files encoding in order for the system to be able to identify the content language and accordingly configure the subsequent input readers (e.g. an ANSI reader will not be able to load data for UTF encoded files correctly) for subsequent parsing

c) Removal of special characters where quotations, JSON and XML notation characters were nullified since they tend to "break" the parsers,

d) Adjust for specific set characteristics, like long paragraphs embedded in the files along with lists and other structures.

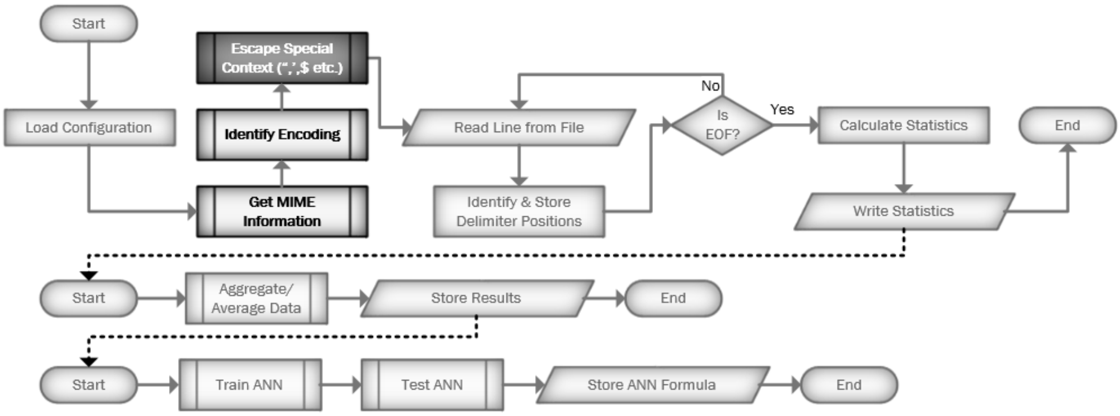


Figure 15. Delimiter Determination Process Flow

Based on the preliminary results of the experiment, some inconsistency was observed, and after investigation, a further component was introduced into the framework. It became apparent that there was a difference in one of the experimental datasets that exhibited low accuracy in predicting the correct file delimiter. *Variety* was in play; something was different with this set that was not considered as an initial variable in the definition of the experiment. The pre-processor labelled "*Escape Special Context*" was incorporated as a compensating control for the following characteristics that attributed to the differentiation:

- The set utilised multiple delimiters for segmentation. Although the file, for instance, was delimited with a comma, one of the fields had in it multiple values delimited with semicolons. As a result, both delimiters exhibited high degrees of conformity.

- The records were extended into multiple lines while being enclosed in double quotes.
- Many of the fields had long, document–sized text, that had a very high degree of variance in length stretching from a couple of lines to hundreds of lines.

The following set of primary metrics were used and calculated to map the characteristics of the delimiters' properties identified in each file, or each line read:

- The number of lines read from the file.
- The number of lines having the same number of columns.
- Min, Max and Mean of Standard Deviation for the position of the delimiter across all lines read per position.
- Min, Max and Mean of Coefficient of Variation for the delimiter position across all lines read per position.
- Min, Max and Mean of Standard Deviation for the relative position (distance) of the delimiter from the previous delimiter across all lines read per position.
- Min, Max and Mean of Coefficient of Variation for the relative position (distance) of the delimiter from the previous delimiter across all lines read per position.

The primary set of metrics were averaged out in aggregating the data from the line level to the file level, considering each delimiter found. The derived set of metrics include:

- The number of identified delimiters in the file.
- Whether the number of columns is consistent across all lines read (Boolean metric).
- The average Standard Deviation for the absolute position of the delimiter.
- The average Coefficient of Variation for the absolute position of the delimiter.
- The average Standard Deviation for the relative position of the delimiter.
- The average Coefficient of Variation for the relative position of the delimiter.

These metrics are subsequently used as the input parameters into a supervised feed-forward multi-layer perception (MLP) neural network to classify the file format without any human input (K. A. Brown et al., 2020). The neural network will be used to verify, based on the delimiter characteristics set by the calculated metrics, if the file was delimited with the respective delimiter. In this way, the identification process will be automated and shall require no human intervention or supervision.

Through automating the task, the required scope of data scientists services will be limited to the most important task of interpreting the data instead of identifying it. This synergy will result in multiple perspective benefits, including a) adaptability and responsiveness to changing business environments since efforts required in incorporating new data will be less (Greenbaum, 2008), b) profitability, by allocating internal or external resources towards data mining activities, instead of data identification, which will identify a competitive advantage (Gregory, 2013), c) a lower TCO, and in turn achieving higher business adoption rates (Kumar, 2013).

4.3.3. The Data Confidentiality Challenge

As it is imperative to govern data custodians' activities through legislative frameworks, information is increasingly being regulated in multiple sectors (OKeefe, 2017). Examples include the Personally Identifiable Information (PII) privacy act 5 u.s.c. 552a 2020 edition, Public-Sector Information (PSI) Directive, General Data Protection Regulation (GDPR) law and Payment Card Industry (PCI) Security Standards along with anonymization standards like European Medicines Agency Policy 0070 or the Health Insurance Portability and Accountability Act (European Commission, 2022; European Medicines Agency, 2017; *General Data Protection Regulation (GDPR) Compliance*

Guidelines, 2020; *Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC*, 1996; *Official PCI Security Standards Council Site - Verify PCI Compliance, Download Data Security and Credit Card Security Standards*, 2006; U.S. Department of Justice - Office of Privacy and Civil Liberties, 2020). These regulatory acts can have an immediate financial impact in the form of fines that can reach 4% of the annual global turnover. A recent example is Facebook, with a confirmed $5 billion fine and another €56 million potential fines depending on the outcome of 11 ongoing GDPR investigations (Cristina Abellan Matamoros, 2019; Lovejoy, 2019). Due to the Data Loss Prevention (DLP) risk, data confidentiality is a candidate for further automation, especially when all these rules and regulations impose a highly complicated legal and compliance framework to which organisations must comply. At the same time, reliance on data-driven analysis and visualisation is increasing, as confirmed by scholars and the business community, leading to the addition of Value as one of the essential V's of Big Data (Davenport, 2012; C. Yang et al., 2017). In an attempt to increase value through Big Data, utilisation of data will intensify the use of data, which will increase the risk of confidential information being disseminated without proper controls. Once this risk is combined with the *Variety* effect of Big Data, it is evident why data confidentiality is considered the most crucial aspect of Big Data protection (Rawat et al., 2019).

Scholars have suggested and implemented several techniques to ensure Big Data privacy throughout the Big Data life cycle (P. Jain et al., 2016; Koo et al., 2020). In achieving anonymization or depersonalisation of the data, many techniques are available including simple processes like deleting or hashing the data or more sophisticated techniques of micro-aggregation (e.g. l-diversity, t-closeness, matrix anonymization or k-Anonymity) (Domingo-Ferrer & Rebollo-Monedero, 2009; Ninghui et al., 2007; Rumbold &

Pierscionek, 2018; Sei et al., 2019; B. Zhou & Pei, 2010). Business experts, along with scholars, have highlighted that the use of such quantitative techniques will point to the known controversy of the Statistical Disclosure Control, where scientists propose approaches for reaching the balance between data disclosure and data loss (Domingo-Ferrer & Mateo-Sanz, 2002; Gouweleeuw et al., 1998; Oganian & Domingo-Ferrer, 2001; Rumbold & Pierscionek, 2018). This delicate balance between the usability of the data and the preservation of the compliance frame is achieved with human intervention in deciding what and how the data will be anonymised or depersonalised, which is resource-intensive and prone to human errors and omissions.

The complexity of compliance requirements is critical, whilst at the same time, the need to derive value through the use of data for analytics and visualisation is also critical. In archiving both, an organisation will have to have in place a framework to protect against data loss and at the same time preserve the usefulness of data. To do so, any movement of data for subsequent analysis a) within the organisation (e.g. from production to test/development environments) or b) externally to it (e.g. sharing information with vendors or competitors) will have to be closely monitored and managed to ensure Data Loss Prevention (DLP). Data confidentiality in archiving DLP will have to be governed and implemented throughout the organisation with a mechanism that will ensure coherence and ease of use. Most of the research identified towards corporate guidelines for confidentiality, is focussed on "the how.” Algorithm implementation, rationalisation, and optimisation are being researched, focusing on minimising the data loss, but "the what," which should be safeguarded, has not been referenced in the research. Research and most academic work focus on achieving anonymization with sophisticated techniques rather than identifying the element to be anonymised. Theoretical and practical implementations regarding data confidentiality and their enforcement are unsatisfactory (Sabelfeld & Myers, 2003). Label systems mainly focus on enforcement of security and data access rather than identification and dissemination (Zheng, 2006). Classification levels associated with labelling are also available in academia. However, they focus mainly on four principles: labelling, binding, change management, and processing, which focus on access prevention rather than secure data proliferation (Blažič & Šaljić, 2010). The identification of data elements and the decision on what should be done with/on them is imperative for any organisation since it is the primary measure in countering the risk of noncompliance.

The work in the later part of this thesis proposes a corporate approach towards DLP through a systemic extension to the currently employed corporate policies for information classification. The prerequisites for data automation, confidentiality and stewardship have been met by discussing in previous subsections a) how to overcome the false positives challenge and identify confidential information, with higher accuracy, in addition to b) an automated system to auto-identify and auto-characterise the data sets' structures. At this point the thesis will focus on extending the novel framework proposed with an information system to facilitate and govern business processes. There has to be impartiality, accountability, standardisation, and corporate cultural awareness in doing so. The "Big Data - Confidentiality Prevention System" (BD-CPS) is proposed to achieve this. The system is equipped with workflows, algorithmic safeguards and real-time rule-based enforcement engines.

BD-CPS is a data driven system which uses structural information metadata to define the information's confidentiality level. The levels are addressing three major viewpoints: a) the owner’s perception of the information, b) the regulators viewpoint of the information and c) the risk of identifying information after concealing confidential information. The system will facilitate the rating by providing a corporate wide ontology and approach that will educate and envision users. Algorithms will be used in correlating and aggregating the confidentiality metadata while workflows with predefined authority levels will facilitate the approval process. Real-time representation of the configured confidentiality rules guides users in achieving minimisation of errors and accidental disclosure. The system also provides an extensive auditing and reference framework that enables users to identify any decision made through the process of confidentiality characterisation along with the reasoning behind it.

4.4. Summary

Variety poses a challenge in identifying the information characteristics. These characteristics can be related to a) the actual instantiation of the data, like the format or the structure, or b) the “weight” the data carried regarding disclosure. Understanding the data and using them safely and securely is essential for all organisations. Most importantly, for heavily regulated sectors where governance and compliance can result in heavy financial penalties or even revocation of the licence to operate.

Several enhancements to the process will provide a cohesive and standardised way of performing te tasks to better identify and understand the data during the ingestion process. Automation, accountability and standardisation shall be the drivers for the organisation to decrease the TCO and increase ROI by better utilising the data scientists. Thus allocating time to the mining of the data for valuable insight instead of having to identify the structure and be concerned with hidden confidential information within the sets or putting the organisation at data loss risk by disseminating privileged information.

The approach will try to facilitate standardisation and thus reuse of work along with all required controls and procedures for safeguarding the organisation. An additional benefit is the corporate user awareness that is built through the proposed framework. Through automation, workflows, and rules, the suggested systems will be available to familiarise and educate an extensive number of participants and cultivate a data and confidentiality knowledge culture within the organisation.

A novel framework is suggested in trying to address the aforementioned challenges. In all cases a systemic approach, suggesting an IS that will automate the process, is proposed. The framework consisted of several components to be analysed, experimented upon and evaluated for their effectiveness and efficiency. The components of the framework, showcased in Figure 12, can be outlined as follows: a) a system that will minimise false positives by targeting the early stages of data identification and confidentiality; b) the automatic characterisation system that comes into play to facilitate the data format identification and data break down; and c) finally, a corporate system to enhance awareness and DLP is proposed and evaluated.

5.1. Introduction

Following the ingestion journey of data, see Figure 12, the first step would be to understand where the data is coming from. This step, named "*Data Origination*," has little room for automation since it is outside the realm of the actual ingestion but is vital since it will provide contextual information. Contextual information would include whether the data set is public or private, identifying the second milestone in ingestion: *Data Confidentiality*. In this stage, highlighted in Figure 16, the data is inspected for any private or confidential information and flagged accordingly for subsequent actions. The experiment presented in this chapter aims to enhance the process of confidential information identification by addressing the false positives challenge. The confidentiality rules will apply for both the sender and receiver of the information, in the first case for action and in the second more in terms of awareness.

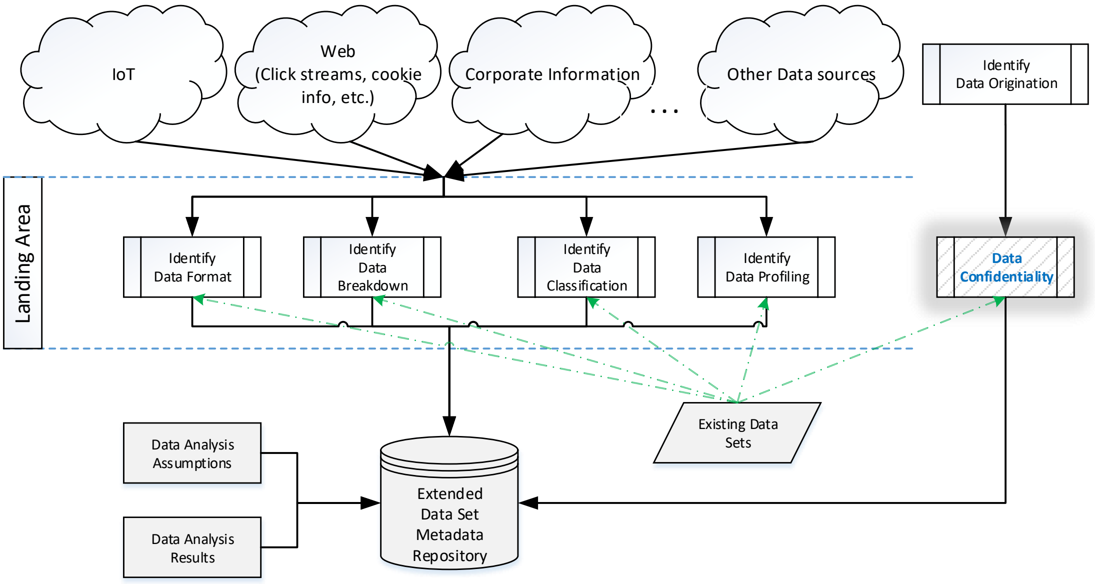


Figure 16. Data Confidentiality Stage

The experiment details are presented in the sections that follow. The methodology used is initially presented (Section 5.2) followed by the actual implementation and configuration details (Sections 5.3 and 5.4 respectively). Section 5.5 details the processing and the experiments. Subsequently the results are presented (Section 5.6).

5.2. Experimental Methodology

In the Data Origination stage, we are interested in identifying the source and content of the data so that the accidental use of private, confidential, and regulated data is minimised if not abolished. With standard user-generated files like word processor, spreadsheet, emails, etc., logs were utilised to investigate and validate the proposed approach. Logs were included since they are widely used in the industry for security analysis, click-stream insights, and other analytics insights (Fan et al., 2015; Mcdaniel et al., 2013). This research has utilised logs from applications and captures from test environments with dummy/seeded data, so there is no compliance concern. The composition and sizes of the files used are presented in Table 7.

Table 7. Data Set Characteristics

File Type	# Files	Size
Network Capture	41	2Gb
Line of Business Application log	3	1Gb
Click Stream Application log	4	10Mb

The four major (4) classifications, along with lists and parameters (RegEx, bin numbers, associated "booster" add-on percentages, sanitisation options etc.), were stored in an XML configuration file, making the system highly flexible. In this manner, the system could quickly be adapted to any required changes or extensions regarding the required configurations. Part of the system initialisation is to set the "booster metrics" configuration, along with the location to be searched and the flag for whether the data should be sanitised. Sanitisation depends on the classification configuration where the technique to be used is defined as mask, hash, encrypt, or replace/truncate the identified information. For each classification, the percentage contribution of each booster metric is configured, and the confidence level is mentioned in screening the results as shown in the examples in Table 8 , for respective RegEx, refer to Table 4.

Table 8. Sample Metrics Definition

Occurrence	Classification	Metrics Definition	Value / Add-on Contribution to Confidence Level
Credit Card	Card	RegEx Identified	40%
		linguistic boundary	20%
		no linguistic boundary	10%
		Luhn algorithm	40%
		exists in institutional BINs	5%
		Sanitisation method	Masking (first 6 and last 3 chars)
		Confidence Level	60%
PII	List	RegEx Identified	40%
		linguistic boundary	20%
		no linguistic boundary	10%
		Proximity	10%
		Sanitisation method	Hash
		Confidence Level	50%
PII	Absolute XML	RegEx Identified e.g. (<CIVIL_ID>)	100%
		Sanitisation method	Truncate
		Confidence Level	50%
PII	Relative XML	RegEx Identified e.g. (<*ID*>)	50%
		Sanitisation method	Truncate
		Confidence Level	50%

An example will be used to understand how the percentages are working together to form an algorithmic sequence. Taking into consideration the first case exhibited, that of a credit card the systems will implement the following processes:

- The moment the system identifies a RegEx match it will increment the confidence level to 40%.
- If the specific match is surrounded by spaces, making it an exact linguistic match, the confidence level will be incremented by 20%, thus becoming 60%.
- If the specific match is compliant with the Luhn algorithm for the check digit, the confidence level will be further incremented by 40%, reaching 100%

- Since the cut-off percent is set to 60% and the calculated confidence level has reached 100%, exceeding the cut-off, the system will consider this a positive hit.
- Likewise, in case none of the buster metric applied, the calculated confidence level would remain at 40%, which is below the cut-off and thus the system would not consider this a match.

For ease of use, the system will recursively search any path provided for all available files in any folder depth, thus allowing for one or multiple sets to be processed in a single execution. By adding the contributions of each metric, the system will arrive at a total confidence percentage per occurrence. If the respective percentage of the instantiation exceeds the defined high watermark for confidence level, as defined previously in 4.3.1, the entry will be considered confidential. The confidence levels are parameterised in the system and should be higher than the initial contributor, in our case RegEx, and lower than the sum of contributors. In filtering out false positives, the level can be increased depending on how many contributors the analyst would like to consider. Multiple post-processors are used depending on the configuration to: a) sanitise the data by applying the requested algorithm; b) remove the data for following classification occurrence so that multiple recordings of the identical instantiations are not recorded, e.g. credit card is not also identified as a debit card c) record all identified values from similar occurrences of a tag value in the Relative XML class to be forwarded to an external AI system for identifying new RegEx patterns.

All the system results are recorded in a log file so that the lineage and detailed results of identified and qualified entries are available for review. If the parameterisation of the system indicated the requirement to sanitise the file, a new file would be created so that the original file is preserved whilst the new one can be released for future usage. In this way, the data scientists can tune the system by altering the contribution percentage and fine-tune the high-water marks per dataset and containing entities.

5.3. The IS Overview

The system is developed with the use of Java so that it can be easily ported to any Big Data platform. The process, shown in Figure 17, will load the configurations from a file so that the data scientists can customise the behaviour without making code changes. Then it will start processing the data set. If the given data set is a path, the system will iterate along the entire tree folder structure to process all the contained information.

The information for each dataset partition/file will be processed, and there will be an explicit memory garbage collection so that the memory and all relevant structures are cleared before processing the next partition. This is done since memory can be a challenge for Big Data environments with limited in-memory processing.

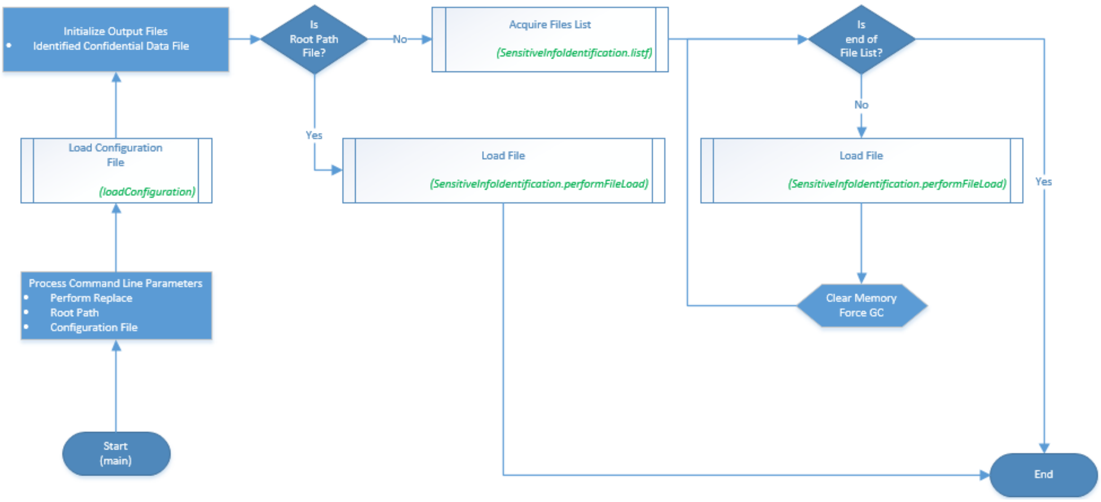


Figure 17. "Booster Metrix" high-level process

The system is implementing inheritance and object-oriented principles by minimising and reusing the structures defined for pattern matching. the first step would be to identify the data set content and, if required, reformat it into a text format so that the RegEx matching can be applied. The second module is the actual identification module where the regular expressions are utilised in pointing out elements of interest that will be further analysed by the additional metrics. The module will take input information for the linear as well as the block processor, and by doing so, the confidence levels are calculated. Last but not least is the results module where all the information about the processing, the results, and the replacements is recorded. In this way the system, depending on the parameterisation/configuration, will provide a report on the findings and the anonymized outcome, while at the same time, if required, present low level information for debugging and optimization purposes.

5.4. Configuration

The configuration is stored in an XML file and will contain information for the parameters that the user can easily change. It was decided to use XML instead of JSON in order for the files' structure to be self-explanatory and reduce the complexity since it has multiple entries. Several distinct sets of parameters are used in the system, out of which one is related to the general configurations and the rest are related to the "booster metrics" classification, as outlined in Table 8. The system can be configured with as many instantiations of any classifier as required and with multiple scrolling windows for vertical reading in terms of text or hexadecimal "windows." A complete configuration file sample is available in Appendix III. Big Data Programming Environments.

a) General Parameters

The system uses this set of parameters in configuring its basic behaviour and logging parameterisation, as shown in Table 9.

Table 9. General Configuration Parameters

Parameter Name	Sample	Description
DebugMode	True/False	This variable is used in producing extensive log and audit information in order to troubleshoot or get a detailed understanding of the functions executed and their outcomes.
DefaultEncoding	ISO-8859-7	If a file's encoding cannot be determined, the system will use this value in trying to interpret the content.
OutputFileName	Results.csv	The filename where the system will store all the matches.
DefaultMaskingCharacter	*	The character to be used in order to mask information.
EncryptionKey	1234-xyz	The key to be used for symmetric encryption on the identified values.
DisplayPlainInfo	True/False	It will determine if the result file will record/show the actual text identified. It is essential to set it to false in many cases so that the privileged information is not revealed to the reviewer.
DisplayMaskedInfo	True/False	It will determine if the result file will record/show the masked text of the information identified.
DisplayHashedInfo	True/False	It will determine if the result file will record/show the hashed text of the information identified.
DisplayReplacedInfo	True/False	It will determine if the result file will record/show the replaced text of the information identified.
DisplayEncryptedInfo	True/False	It will determine if the result file will record/show the encrypted text of the information identified.
ReplaceFileExtension	.rpl	The file that will be prepared by the system while applying any privacy policies.
TextProximity	25	The number of characters to be searched for related information.
BlockMaxLines	10	The number of maximum carriage returns that can be considered a single unit of the data text window element.

b) "Card" Classifier Parameters

The card parameters can have multiple instantiations and are distinguished in the configuration by the classification attribute of the XML element being set to "CARD." Each tag will additionally have the description attribute for reference and readability purposes. A sample tag for the definition of debit card would be *"<Config description = "Debit Cards" classification = "Card">*" where the description and classification are easily identified. The required parameters under a "Card" classifier are described in Table 10.

Table 10. "Card" Classifier Configuration Parameters

Parameter Name	Sample	Description
RegExs	<Dscr> Debit Card </Dscr> <regEx> [4]{1}[0-9]{15} </regEx>	This element will have n number of child nodes which will in turn have two tags. The description which is used in the results file to identify the rule under which the specific entry matched and the RegEx expression to be used in identifying the entry.
MyBins	123152	It is used to parameterise the institutional BINs so that if there is a match, the booster metric is triggered.
Masking	<ShowFromStart> 6 </ShowFromStart> <ShowFromEnd> 3 </ShowFromEnd	The masking that will have to be applied in case the replacement option is "Mask".
ReplacementOption	Truncate (Default) Mask Hash Encrypt	Is the algorithm to be used in anonymizing the results in both the results identification file and the new anonymized/depersonalised file. If the entry is missing, the system assumes the truncate option, which will not display any data.
ReplacementStrengthBoarder	50	Is the Confidence level which will be used in reporting an occurrence as a positive match.
AlgorithmicStrength	0-100	Holds the parameterisation of the percentages used for each booster metric. The list would include, as described in Table 8: BoundTextPercent, linguistic boundary UnoundTextPercent, no linguistic boundary LuhnAddOn, Luhn algorithm inMyBinsAddon, exists in institutional BINs.

c) "List" Classifier Parameters

The list parameters can have multiple instantiations and are distinguished in the configuration by the classification attribute of the XML element being set to "LIST." Similarly to the card classifier, the list tag will have a description attribute. A sample tag for the definition of IDs could include a passport number, a civil ID, etc. An identification number would be *"<Config description = "IDs" classification = "List">"* where the description and classification are easily identified. The required parameters under a "List" classifier are described in Table 11.

Table 11. "List" Classifier Configuration Parameters

Parameter Name	Sample	Description
RegExs	<Dscr> ID </Dscr> <regEx> [4]{1}[0-9]{15} </regEx>	This element will have n number of child nodes which will in turn have two tags. The description which is used in the results file to identify the rule under which the specific entry matched and the RegEx expression to be used in identifying the entry.
TextProximityEnhancers	Passport Identity	It is a list of literal that will be searched in the surrounding text; see general parameter "TextProximity" of the RegEx match.
Masking	<ShowFromStart> 6 </ShowFromStart> <ShowFromEnd> 3 </ShowFromEnd	The masking that will have to be applied in case the replacement option is "Mask".
ReplacementOption	Truncate (Default) Mask Hash Encrypt	Is the algorithm to be used in anonymizing the results in both the results identification file and the new anonymized/depersonalised file. If the entry is missing, the system assumes the truncate option, which will not display any data.
ReplacementStrengthBoarder	50	Is the confidence level which will be used in reporting an occurrence as a positive match.
AlgorithmicStrength	0-100	Holds the parameterisation of the percentages used for each booster metric. The list would include, as described in Table 8: BoundTextPercent, linguistic boundary UnoundTextPercent, no linguistic boundary ProximityAddOn, Proximity text exists.

d) XML Tags Classifier Parameters

There are two instantiations for XML tags identification—absolute and relative. Like the previous definitions, the element will have a classification and a description attribute. Although there is no difference in the definition of the parameters, the logic applied for both is different since the "tag" parameter is used in the first case of "absolute" as an exact textual match, whilst in the case of "relative," the system will try to identify if the "tag" is part of any literal, even as a partial match. The parameters of the XML classifiers are depicted in Table 12.

Table 12. “XML tag” Classifier Configuration Parameters

Parameter Name	Sample	Description
Dscr	Customer FullName	The description used in the results file to identify the rule under which the specific entry matched.
Tag	FullName	The text to be searched as an absolute or relative tag within the source XML.
RelOpt	Truncate (Default) Mask Hash Encrypt Replace	Is the algorithm to be used in anonymizing the results in both the results identification file and the new anonymized/depersonalised file. If the entry is missing, the system assumes the truncate option, which will not display any data.
ShowFromStart	0-n	The masking, start and end that will have to be applied in case the parameter RelOpt is “Mask”.
ShowFromEnd	0-n	
Replacement	xyz	Static value to be used in replacing the content of the matched entry if parameter RelOpt is Replace.
RegExCandidate	True/False	If set to true, the system will record all occurrences of the respective entry so that a secondary process can try to identify a possibility of a new RegEx expression.

e) Block/Windowed Reader Parameters

The network traces, as described earlier, can contain two vertical blocks. For the system to understand the boundaries and decode the recorded information, several parameters must be in place. In this case, the configuration uses two classifiers and a description to identify the content. Thus, if a text block is identified, the classifier will be “TXT,” whilst if it is a hexadecimal block, the classifier would be “HEX.” The description can be the same since they are both parts of the same aggregation. If, for example, a WireShark network trace file was depicted, the following two entries would be defined a) “*<BlockConfig description = "WireShark Data Block" classification = "TXT">*” and b) “*<BlockConfig description = "WireShark Data Block Hex" classification = "HEX">*.” The parameters for each block can be seen in Table 13.

Table 13. Block/Windowed Reader Configuration Parameters

Parameter Name	Sample	Description
BlockFromColumn	56	The column/character from the start of the line where the block starts.
BlockWidth	16	The number of columns/characters from the start of the block, see parameter BlockFromColumn, for which this block extends.
BlockTerminator	|	A character that will identify the end of a block. By default, an empty line will be considered as a block terminator.
BlockWindoeLines	10	If a block terminator is not encountered, the system will consider a "hard stop" the respective n number of lines read.

Block/windowed reading functionality should be used in specific configurations and not be part of a generic template since they will initiate multiple scrolling windows to read the content of the data set, that will entail a substantial increase in memory and CPU consumption.

f) Command Line Parameters

The system for every execution requires some parameters to be provided, which can be identified in Table 14.

Table 14. Command-line parameters

Parameter Name	Sample	Description
replaceData	True/False	The parameter denotes if the system should produce a new file with all the confidential information de-personalized / anonymized.
dataFileName	C:\xxx\yyy\	The location of the data set or individual element to be processed.
configFileName	C:\xxx\yyy\abc.xml	The configuration file.

g) Output

The system provides an output file that denotes the identified confidential information. The file structure is parameterised using the configuration file parameters, see Table 9. The complete set of columns is exhibited in Table 15.

Table 15. Results file structure

Column Name	Sample	Description
File	C:\xxx\yyy\abc.txt	the full path/location of the file as it is read by the system.
Encoding	UTF-8	The file encoding identified.
Line	1-n	The line where the confidential information was identified.
MatchDscr	Social Security Number	The description of the matching criteria.
MatchText	1-n	The confidential text identified.
MatchTextMasked	xyz	The confidential text as produced after applying the mask rule.
MatchTextEncypted	xyz	The confidential text as produced after applying the encryption rule.
MatchTextHashed	xyz	The confidential text as produced after applying the hash rule.
MatchTextReplaced	xyz	The confidential text as produced after applying the replace rule.
MatchTextStartPosition	1-n	The start location of the identified text.
MatchTextEndPosition	1-n	The end location of the identified text.
MatchTextWidth	1-n	The width of the identified text.
MatchStrength	1-100	The strength of the identified test based on the “Booster Metrics.”

Depending on the user request, the system can also generate a file similar to the source file but with all identified confidential information de-personalized/anonymized.

5.5. Process Data Set / File(s)

Once the configuration is read and the system has identified the list of files/data set elements that will have to be processed, it will process each data set element sequentially. In Figure 18, there is a flowchart representation of the flow. The system will initially try to identify the Multipurpose Internet Mail Extensions (MIME) in order to identify if a conversion is required with the use of the TIKA libraries. Once the conversion is finished, the system will inspect the output file to confirm that it can be further processed. If the user has requested an anonymized file, the system will initialise the output and results files. In addition to that, based on the configuration, multiple readers will be initialised for each block/widowed configuration set. As long as there is content in the pipeline, the system (shown in Figure 18) will process each element from the linear and block readers, store the results, replace the information as part of the anonymization and write the output data.

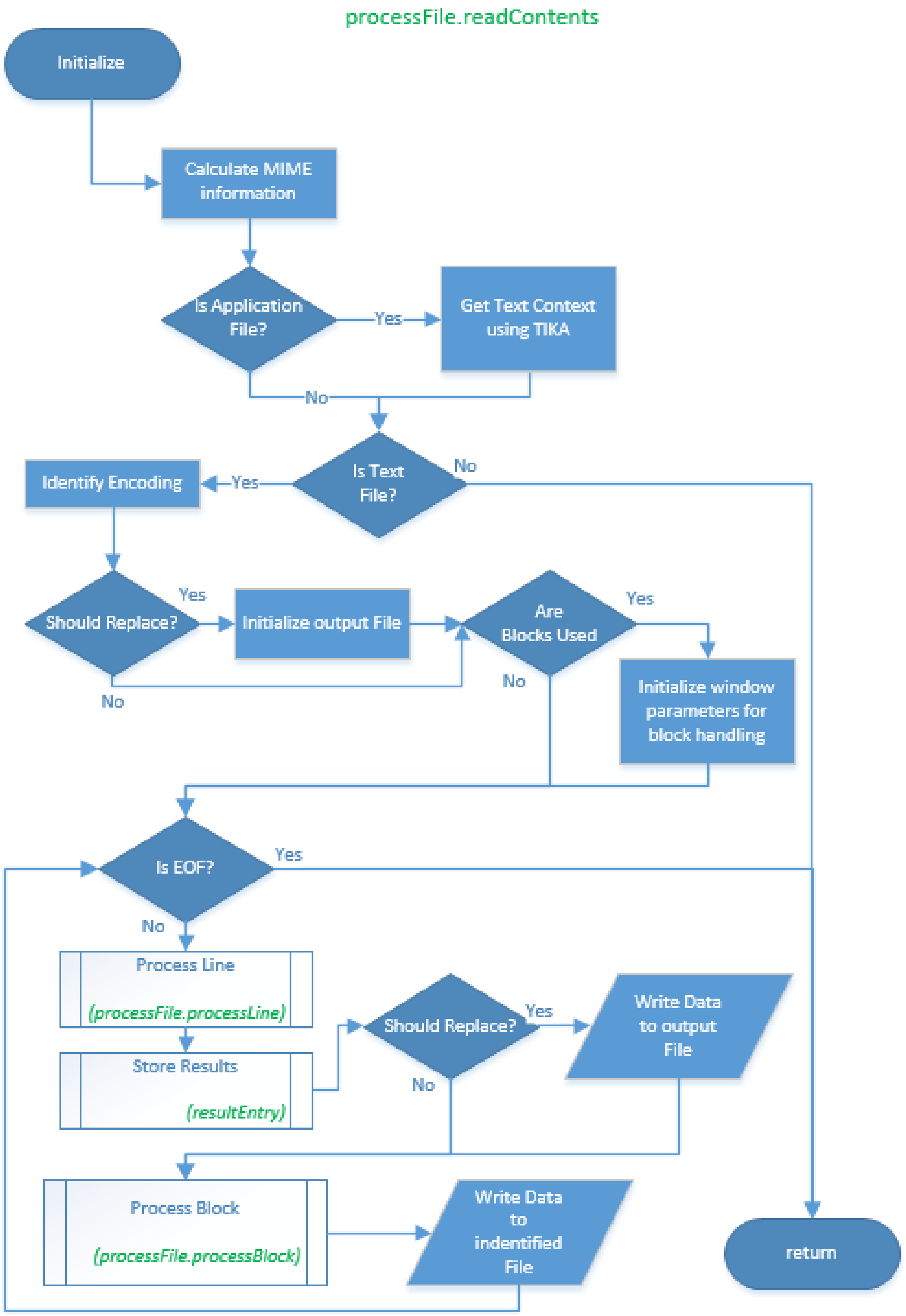


Figure 18. Data Set Element Processing

As stated, there is a linear and a block process for each element. As shown inFigure 19, the linear will iterate all types of classifier instantiations and identify if there is a match.

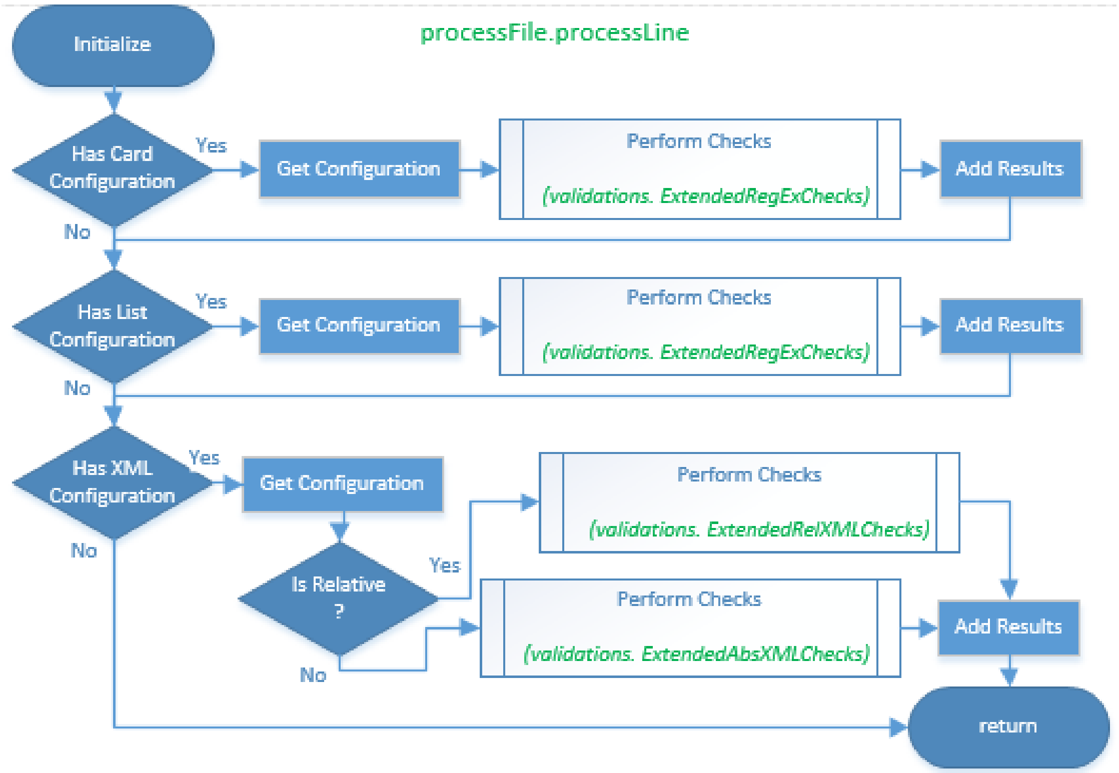


Figure 19. Element Validations/Checks

The block process, on the other hand, as shown in Figure 20, will utilise the prior buffered information in concatenating into a new string and apply the linear process to the new aggregated information.

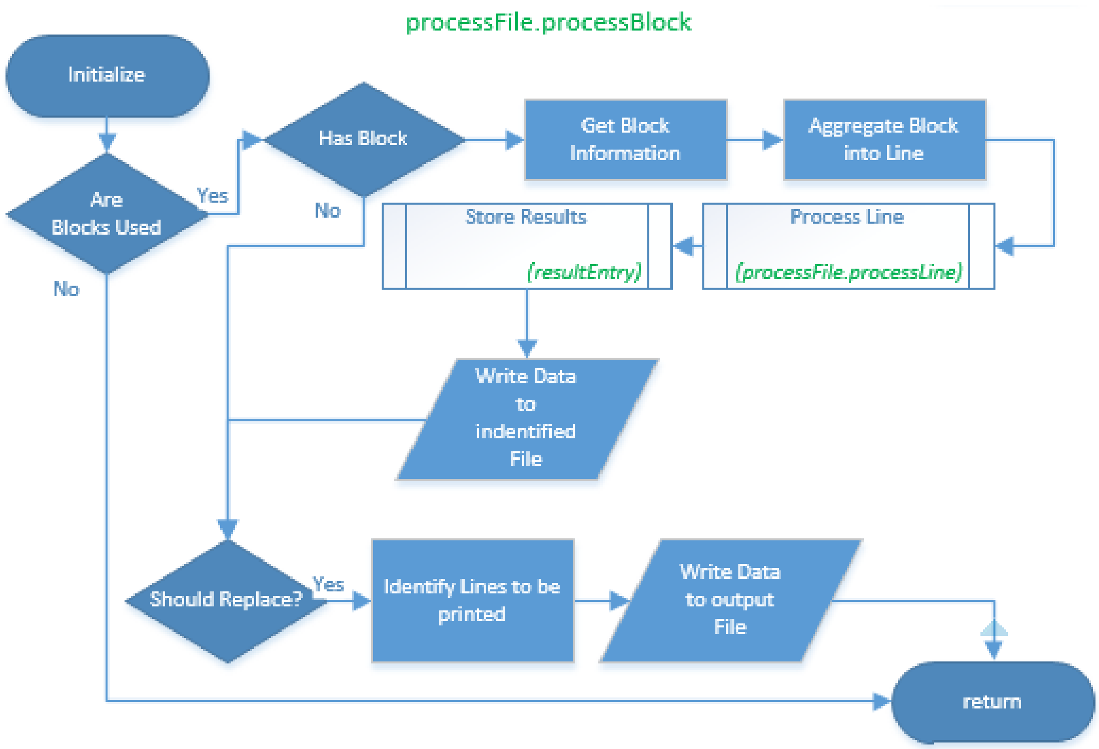


Figure 20. Windowed/Block Processing

Depending on the type of classifier instantiation under execution, there are three distinct checks and attempts to confirm a “hit” on the data based on the input parameters. The first is the case of “List” and “Card” classifiers depicted in Figure 21, and the other two are the “Absolute” and “Relative XML classifiers presented in Figure 22 and Figure 23, respectively. In all processes, the data after any match are masked so that subsequent executions of other classifiers or block readers do not generate duplicate findings.

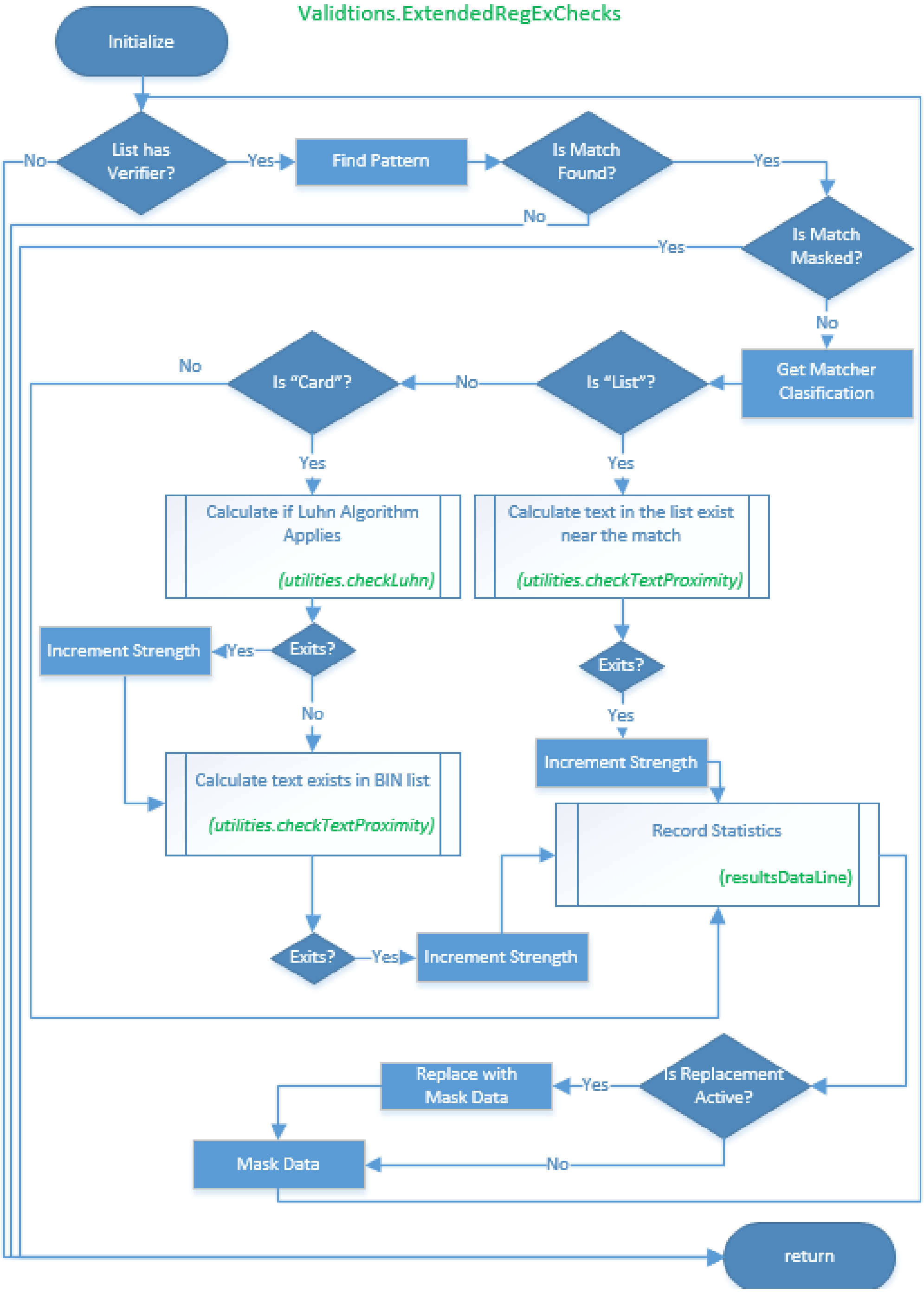


Figure 21. "Card" & "List" Processing

Different calculations are implemented when the classifier is “Card” or “List” in relation to proximity, Luhn and BINs. Any match from the RegEx will be complemented by the respective executions in increasing the confidence/strength percentage.

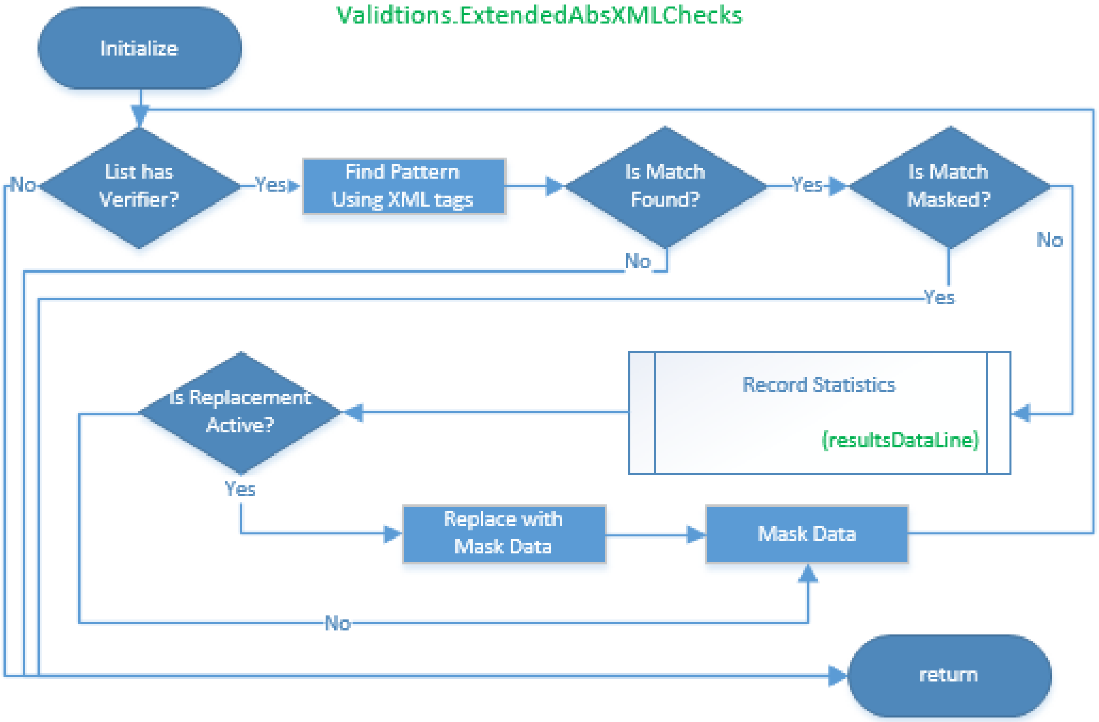


Figure 22. Absolute XML Processing

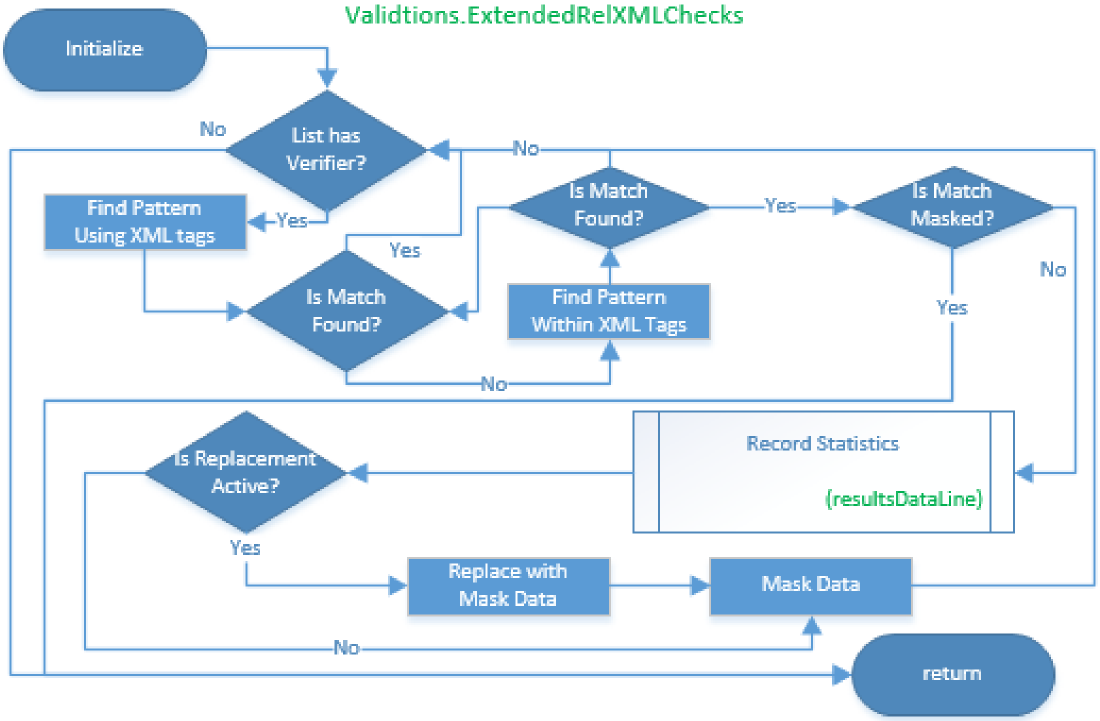


Figure 23. Relative XML Processing

The Relative XML processing is similar to the Absolute, with the difference that in relative, the sought literal is being identified either as an exact match or as any part of the XML tag element name.

All gathered information in case of a "hit" is output in a result file. Depending on the parameterisation of the system (see Table 15), this result file will have the following elements:

- The name of the data set element or file that the match was identified.
- The character encoding of the data set element or file.
- The line in which the match was identified.
- The description of the classifier that produced the "hit."
- (Optional) The actual text that produced the match.
- (Optional) The masked text that produced the match.
- (Optional) The encrypted text that produced the match.
- (Optional) The hashed text that produced the match.
- (Optional) The replaced / static text that produced the match applicable only for XML.
- The position where the matching occurred. Start, width and end.
- The matching strength/confidence percentage.

5.6. Results

The purpose of the experiment was to confirm the capability of minimising the number of false positives and identifying more occurrences of confidential data using extended pattern matching. In verifying the method used, the results were compared to a traditional simple RegEx match. Files with a combination of actual positives and false positives were utilised in identifying the value added by the proposal. In order to calculate the performance against the simple RegEx, the 50% confidence level was considered a benchmark. Based on the parameters, standard RegEx would always yield values lower than 50%, whilst the "boosters" would elevate the respective to higher percentages. Any value below 50% should be considered a “false positive,” since no “booster” was utilised to verify its validity. Utilising the “un-boosted” RegEx matching methodology, 5.4 million occurrences were identified. Based on a sample verification, 3.8million occurrences were confirmed hits, which would coincide with the >50% “boosted” result set. On average, it can be identified that there was an improvement of ≈35% in filtering out false positives, as shown in Table 16.

Table 16. False Positive percentages

Classification	Standard RegEx	“Boosted” RegEx confidence level ≤ 50%	“Boosted” RegEx confidence level >50%
Cards	18,422	7,337 (39.83%)	11,085 (60.17%)
Lists	5,394,547	1,565,160 (29.01%)	3,829,387 (70.99%)
Total	5,412,969	1,572,1497	3,840,472

In calculating the system's performance for attaining a better match by introducing Absolute and Relative XML, all 450K additional matches are considered; see details in Table 17. This would constitute an increase in the hit ratio of ≈12%. Out of these, by utilising the aforementioned 50% confidence rule, the immediate contribution would be ≈3% without false positives. The remaining 9%, which in essence is the 325K pertaining to Relative XML without a “booster,” can be further mined for identifying new RegEx expression or new “booster metrics.”

Table 17. Additional Hits using XML tags

Classification	“Boosted” Absolute and Relative XML confidence level ≤ 50%	“Boosted” Absolute and Relative XML confidence level >50%
Absolute XML	N/A	22,741
Relative XML	325,530	105,758

Having understood the benefit of having a higher confidence level, it would be beneficial to address the challenge from a legal point of view. The legal requirement will always be to have 100% identification and removal of confidential information in order to minimise the risk of any data exposure. The 50% confidence level and the 100% identification are complementary in nature, and in essence, the confidence level is aiming towards facilitating the 100% identification. The 50%, 60%, or any other high watermark the data scientist will set, will only filter out confirmed cases of false positives, it will not filter out in cases where there is probability of 50%. For example, if a 15 digit number is not calculated to a valid check digit, it is 100% sure it is not a credit card and by attaining only a 40% confidence level (see Table 8), it should be excluded. Similarly it is "too much of a coincidence" to have the word(s) "credit card" beside a 15 digit number with a valid bin attaining a confidence level of 65% (see Table 8) and it cannot be excluded, although the check digit might not match. The confidence level percentage will only "exclude" confirmed hits of noncompliance with the implemented metrics. In this way, the real cases of confidential data will be uncovered within the "sea of information" and dealt with, thus driving towards the 100% identification legal requirement.

5.7. Conclusion

The results have confirmed that with the use of "booster metrics" the accuracy of identifying confidential data will be enhanced. The use of automated processes can identify confidential information, reduce regulatory risk of noncompliance, and minimise possible data loss. By minimising the false-positives the data scientists and data governance teams can focus on the actual data that require anonymization and direct their efforts towards a smaller and more focused data set.

By employing such a solution, the organisation will be able to increase the safety level of identifying confidential information while at the same time lowering the effort and involvement of the analysts. Awareness will also be increased since more accurate results will caution all involved parties to seriously take into consideration any findings, thus minimising the risk of accidental use or distribution of confidential information.

6.1. Introduction

Following the ingestion journey of data, see Figure 12, and having cleared the confidentiality barrier to ensure compliance (see Chapter 5), the data should be broken down and imported into a landing area. This step is the *Data Format Identification* followed by the *Data Breakdown* phase, as highlighted in Figure 24. The proposed approach focuses on viability and accuracy in addressing the heterogeneity challenge. As Volume can be an impediment when addressing *Variety* - as analysing complete data sets can often be impossible/infeasible, the proposed approach seeks to achieve format identification through an analysis of the delimiters presence on fragments of the original data set.

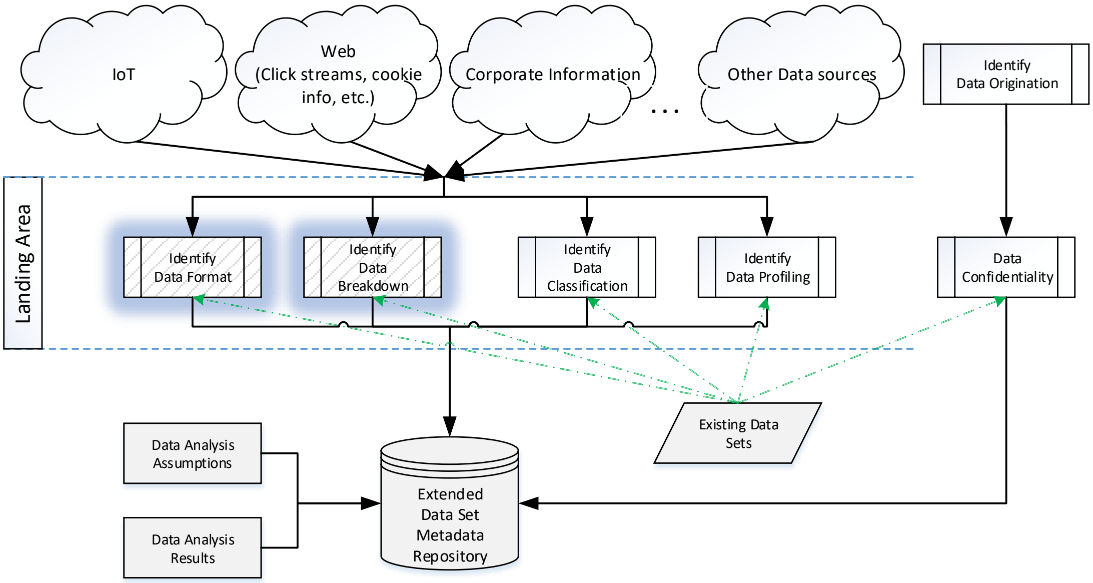


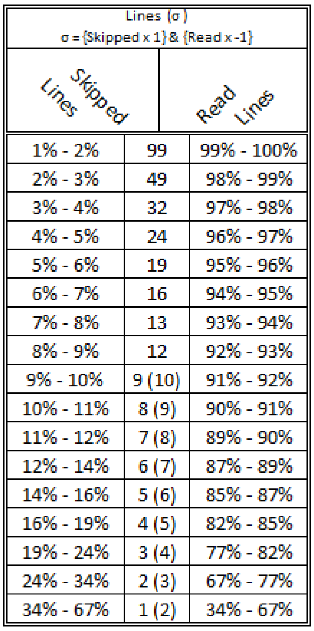
Figure 24. Data format & Break Down Identification Stages

The Information System (IS) presented automates the ingestion process, and the experiment conducted will confirm its accuracy and viability in the Big Data ecosystem. The proposal will assist data scientists in ingesting information in the landing area faster and more accurately, thus enabling them to focus on the actual data mining. Towards that end, Section 6.2 is identifies the means by presenting the methodology followed by the design, implementation and configuration of the system (Sections 6.3 and 6.4). The data set flowing through the system and step-wise processing are exhibited (Section 6.5) with the results being highlighted at the end (Section 6.6).

6.2. Experimental Methodology

This Proof of Concept (PoC) was used to confirm that an automated solution for the data format is feasible. The experiment would have to identify the quantity of data that should be processed to attain reasonable confidence in the data set. In attaining a uniform read pattern across the data set, the system will have to process all parts of the file instead of the common practice of processing only the beginning of the file. Microsoft tools like MS Access, MS Excel and SSIS read a chunk, depending on the tool, that may vary from 100 to 100000 lines, in presenting the user with a sample for import processing, which will arguably not suffice in Big Data sets. A failsafe was implemented within the configuration; apart from the file percentage to be read, the data scientist can enforce a certain number of lines to be read from the beginning of the file. For the purpose of the experiment, the initial set of lines to be mandatorily read was set to 100 lines. Since we are dealing with Big Data, it is imperative that a fraction of the data set is read. In identifying whether a line read should be considered in the analysis as input data, a skip line algorithm was devised. Table 18 illustrates the algorithm results for a sample of 100 lines of input.

Table 18. File Lines %



Lines (σ) σ = {Skipped x 1} & {Read x -1}		
Skipped Lines		Read Lines
1% - 2%	99	99% - 100%
2% - 3%	49	98% - 99%
3% - 4%	32	97% - 98%
4% - 5%	24	96% - 97%
5% - 6%	19	95% - 96%
6% - 7%	16	94% - 95%
7% - 8%	13	93% - 94%
8% - 9%	12	92% - 93%
9% - 10%	9 (10)	91% - 92%
10% - 11%	8 (9)	90% - 91%
11% - 12%	7 (8)	89% - 90%
12% - 14%	6 (7)	87% - 89%
14% - 16%	5 (6)	85% - 87%
16% - 19%	4 (5)	82% - 85%
19% - 24%	3 (4)	77% - 82%
24% - 34%	2 (3)	67% - 77%
34% - 67%	1 (2)	34% - 67%

$$x = \left(\begin{array}{ccc} & & l \ < \ \mu \\ Or & \left(\begin{array}{cc} if & (\sigma \ > 0) \\ then & \left((l \ \ modulo\ \sigma) = 0 \right) \\ else & \left((l \ \ modulo\ \sigma) \neq 0 \right) \end{array} \right) & \end{array} \right)$$

x: Should Process Line
l : Current Read Line Number/Counter
μ: Mandatory Lines to Read
σ: Skip Lines Number

The system should accurately identify data formats irrespective of volume or individual element size of data sets. To that end, the approach investigated data sets with different characteristics in terms of origin, file size and the number of files in each set, as shown in Figure 25.

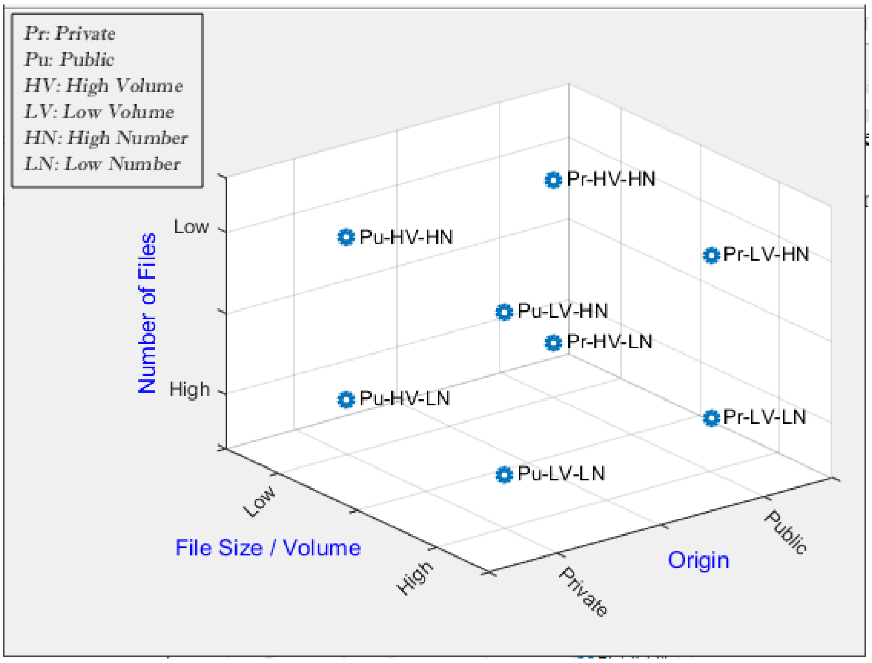


Figure 25. Experimental data composition concerning Origin, Volume and Size

The selected datasets constitute a representative sample where all three combinations of the two-volume measures are identified in addition to the origin factor: a) Public, High number of files and Low volume (NCDC) b) Private, Moderate number of files and Moderate volume (ODS) c) Public, Low number of files and High volume (CDC). Although the ratio of file numbers to their sizes indicates the number of lines per file, it is crucial to better understand since processing many small files compared to a limited number of files with a high number of line presents different challenges (e.g. memory constraints). The classification of lines number bands and respective counts for each dataset are available in Table 19. The presented count and percentage are based on the text files identified in each set and exclude the binary files count, which eventually are excluded from the experiment.

Table 19. Dataset Lines No Classification

Number of Lines	**NCDC**		**CDC**		**ODS**	
	# of Files	**% of Data set**	**# of Files**	**% of Data set**	**# of Files**	**% of Data set**
0-100	7.225	52%	84	15	3.737	63
101-500	2.119	15%	102	18	330	
501-10,000	3,482	25%	246	44	1.592	27
10,001-100,000	903	6%	68	12	185	3
100,001-10,000,000	290	2%	59	11	74	1

In an attempt to auto-detect delimiters, the system was configured with different confidence levels for possible delimiters, as shown in Table 20. Weights were granted based upon the usage of the respective character(s) combination. Thus a comma “,” which is extensively used, has the lowest weight compared to a “~||~” set, which is rarely used in everyday communications.

Table 20. Delimiters Confidence Level Weights

Delimiter	**Weight**	**Delimiter**	**Weight**
Comma	1	Tilde	3
Semicolon	2	Tilde Pipe Tilde	4
Tab	2	Tilde Pipe Pipe Tilde	5

Once the unit testing of the system was concluded and several pre-processors were developed, multiple executions were conducted to incorporate different sampling sizes, as identified in Table 18 for each set. The formulas used to calculate the metrics are available in Table 21.

Table 21. Metrics Formulas

Metric	Formula
Mean (μ)	$\mu = \frac{\sum x_i}{n}$
Standard Deviation (σ)	$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$
Coefficient of Variation (C_v)	$C_v = \frac{\sigma}{\mu}$

The primary derived metrics and features like data set name, and delimiter name were input to the next component of the systems, a multi-layer perceptron (MLP) neural network, to approximate the relation between the delimiter characteristics and the delimiter (Pijanowski et al., 2014). This component was implemented in MatLab with a graphical illustration of the network shown in Figure 26. In reducing model overfitting, the train-validate-test technique is used, and the input data set is divided into three parts accordingly, with a ratio of 70%, 15% and 15% (NcCaffrey, 2015). Although there is no clear ratio rule, most researchers tend to use this mix (Draelos Rachel, 2019; *Train Test Validation Split Python*, 2020). The Bayesian Regularization Back Propagation (trainbr) training function was used since it presents several advantages, and due to that, it is extensively used (Livingstone, 2008; Yue et al., 2011). The network utilised the Mean Squared Normalized Error (mse) performance as is typical (Christiansen et al., 2014). In respect to the executions, the limiting number of epochs was set to 10,000, and the hidden neurons were set to 6, using the “rule of thumb,” $\frac{2\times(\#\,Input+\#\,Output)}{3}$ (Y’barbo, 2012).

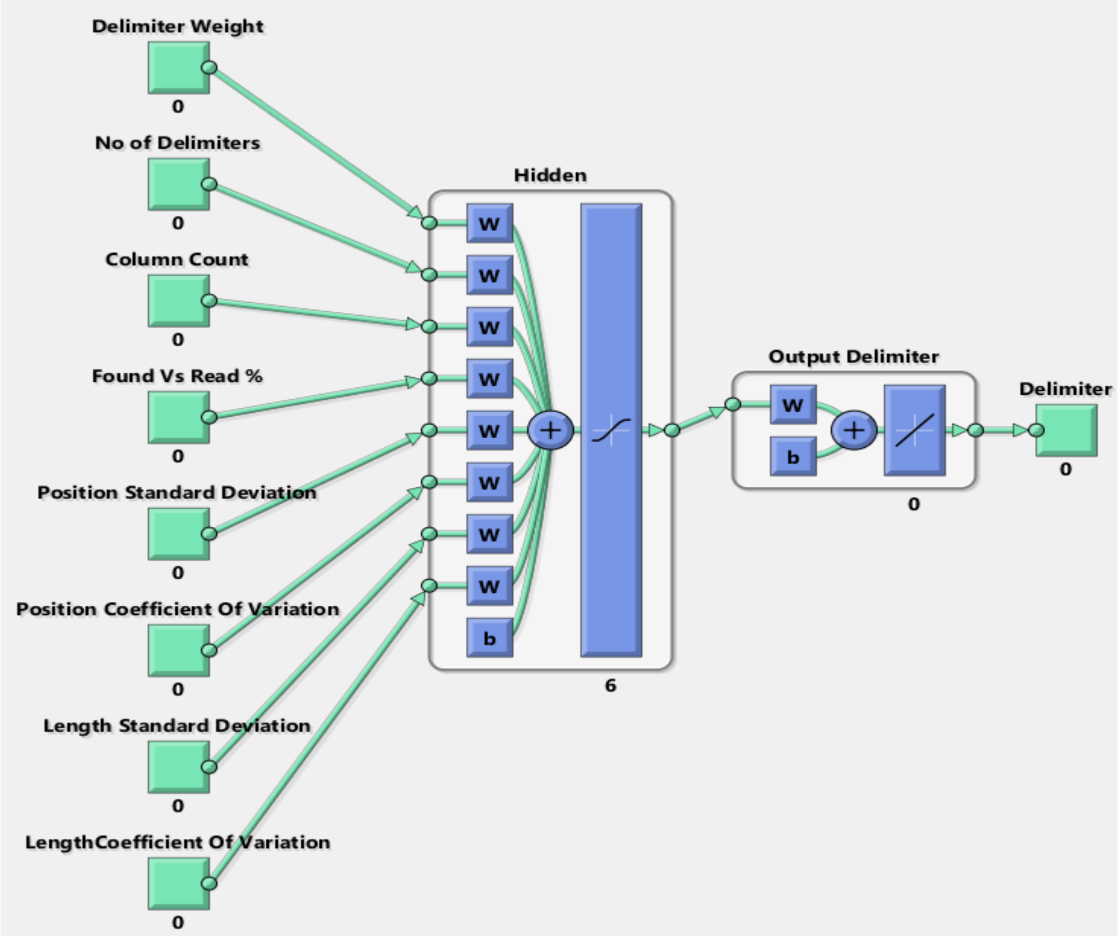

Delimiter Weight
0
No of Delimiters
0
Column Count
0
Found Vs Read %
0
Position Standard Deviation
0
Position Coefficient Of Variation
0
Length Standard Deviation
0
LengthCoefficient Of Variation
0
Hidden
W
W
W
W
W
W
W
W
b
6
Output Delimiter
W
b
0
Delimiter
0

Figure 26. The experimental configuration of the ANN

To train the neural network, the training data was augmented with target outputs using the values zero (0) and one (1). One (1) was used to represent valid combinations of file and delimiter, whilst all other combinations, representing false positives, were assigned zero (0). The performance graphs of a sample training from one of the input datasets are presented in Figure 27.

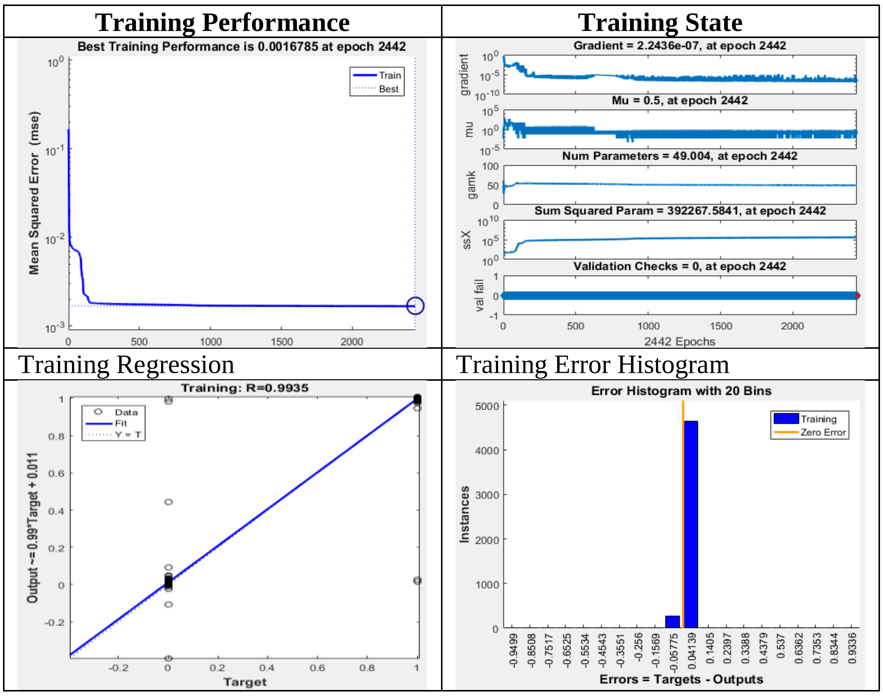


Figure 27. Training performance graphs

Post-training and the final calculation of the neural network confidence level confirm whether the file was delimited with the specific delimiter, and then the neural network result/output is adjusted to the range of -1 and 2 as shown in Figure 28.

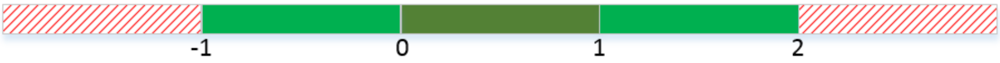


Figure 28. Neural Network Result Adjustment

Multiple runs were executed with different sampling sizes. A neural network was created, trained, and used to predict the delimiter's correctness based on each set of calculated metrics. In this manner, the automation of the process would be achieved whilst its viability would be guaranteed by lowering the effect of the Big Data *Volume* dimension with the use of undersized samples in correctly discovering the structure of a delimited file.

6.3. IS Overview

The system is developed with the use of Java so that it can be easily ported to any Big Data platform. The process, shown in Figure 29, will load the configurations from a file so that the data scientists can customise the behaviour without having to make code changes. Then it will start processing the data set. If the given data set is a path, the system will iterate along the complete tree to process all the contained information.

The information for each dataset partition/file will be processed, and there will be an explicit memory garbage collection so that the memory and all relevant structures are cleared before processing the next partition. This is done since memory can be a challenge for Big Data environments with limited in-memory processing.

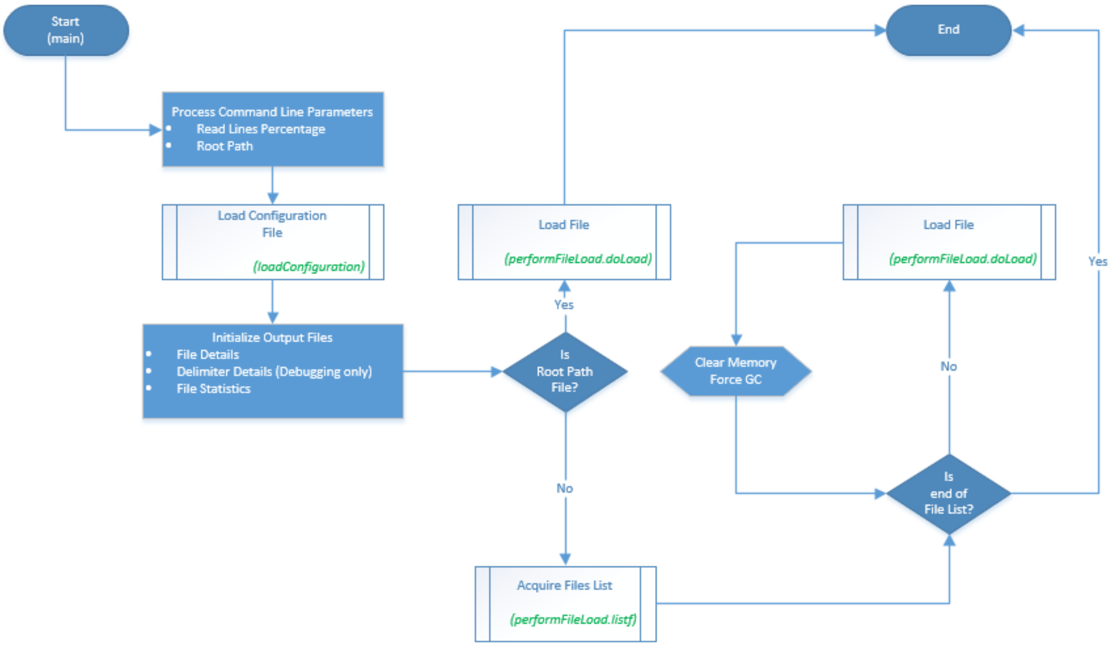


Figure 29. Dataset Characterisation high-level process

The system will process each set in order to calculate statistical information based on the positions of possible delimiters. Based on these metadata the system will perform an aggregation and present a set of dimensions per set processed. The experiment identified variations thus it was deemed necessary to subject the datasets to a pre-processor that would sanitise the input sets. Having consistent information across different types of sets led to the next step of the distributed system where a neural network would be trained and tested to identify the set's structure.

6.4. Configuration

The configuration is stored in a JSON file and contains information for the parameters that the user can easily change. The system can be configured manually with multiple delimiters or with the use of another automated system that will deduce new possible delimiter character sets. A complete configuration file sample is available in the Appendix VI. *Dataset Characterisation sample JSON configuration*.

a) General Parameters

The set of parameters used for the experiment is outlined in Table 22.

Table 22. Block/Windowed Reader Configuration Parameters

Parameter Name	Sample	Description
DebugMode	True/False	This variable is used in producing extensive log and audit information in order to troubleshoot or get a detailed understanding of the functions executed and their outcomes.
DefaultEncoding	ISO-8859-7	If a file’s encoding cannot be determined, the system will use this value in trying to interpret the content.
Delimiters	{ "regEx" : ";", "Dscr" : "Semicolon" }	It is an array of a set of 2 parameters that will serve as configurations in identifying the possible structure of the data set.
Delimiters - regEx	"\~\\|\\|\~"	Is the actual regular expressions delimiter text used to identify the data set structure.
Delimiters - Dscr	Comma	Is the description/identifier to be used against which the calculated metrics will be stored.
LinesToReadFromStartOfFile	100	Is the minimum number of lines that the system will read regardless of the data read percentage set via the command line.
StatsOutputFileName	DelimiterStatResults	Is the filename to be used in order to record statistical details calculated by the system, the structure is available in Table 25.
FileInfoOutputFileName	DelimiterFileInfo	Is the filename to be used in order to record information about the files that the system has processed, the structure is available in Table 24.
OutputFileExtension	.txt	The output files extension.
FileSizeLimit	104,857,600	The maximum data set chunk size that the system will process.

b) Command Line Parameters

The system for every execution requires some parameters to be provided, which can be identified in Table 23.

Table 23. Command-line parameters

Parameter Name	Sample	Description
fileReadLinesPrc	1-100	The percentage of the file to be read and considered for processing.
dataFileName	C:\xxx\yyy\	The location of the data set or individual element to be processed.

c) Output

There are multiple output files configured in the system, with diverse structures representing different types of collected information and metrics. In Table 24 and Table 25, the respective file structures are explained.

Table 24. FileInfoOutputFileName structure

Column Name	Sample	Description
File	C:\xxx\yyy\abc.txt	The full path/location of the file as the system reads it.
MIME Type	Binary / Text	The MIME type of the file as identified by the system. The type denotes if the file is in text format or binary format.
MIME Subtype	Rdf, xml, txt etc.	The MIME subtype of the file as identified by the system. The subtype denotes what kind of a file is identified, meaning if it is a report file, an XML file, an word processing file, a worksheet etc.
Encoding	UTF-8	Is the file encoding identified.
Lines no	1-n	Number of lines of the file as read by the system.

Table 25. StatsOutputFileName structure

Column Name	Sample	Description
Delimiter Info	Comma, Tab	The description of the delimiter for which the metrics are calculated.
Lines Read	1-n	Number of lines processed by the system based on the read percentage set.
Length	1-n	The total number of lines read from the file.
Min	1-n	Are the statistics, as the column name denotes, that are calculated for each set per delimiter for the delimiter position.
Max	1-n	
Mean	0-n	
Sum	0-n	
Standard Deviation	0-n	
Coefficient Of Variation	0-n	
Min	1-n	Are the statistics, as the column name denotes, that are calculated for each set per delimiter relative position (distance) of the delimiter from the previous delimiter.
Max	1-n	
Mean	0-n	
Sum	0-n	
Standard Deviation	0-n	
Coefficient Of Variation	0-n	

6.5. Process Data Set / File(s)

The PoC developed comprises two autonomous systems that complement each other in reaching into an automated process of identifying the file structures. The first component will process the data set and calculate a set of statistical metrics. Complimentary, the second will utilise these metrics in training a neural network, which will eventually identify the delimiter.

6.5.1. Data set statistical analysis

Having read the configuration file and identified the command line parameters, the system will identify the files required to be processed and start processing them one by one. As shown in Figure 30, the system will read the file content and calculate the respective statistics for each file. An essential part of this process is nullifying variables and clearing memory since the systems exhausted all the available memory in several runs with large files.

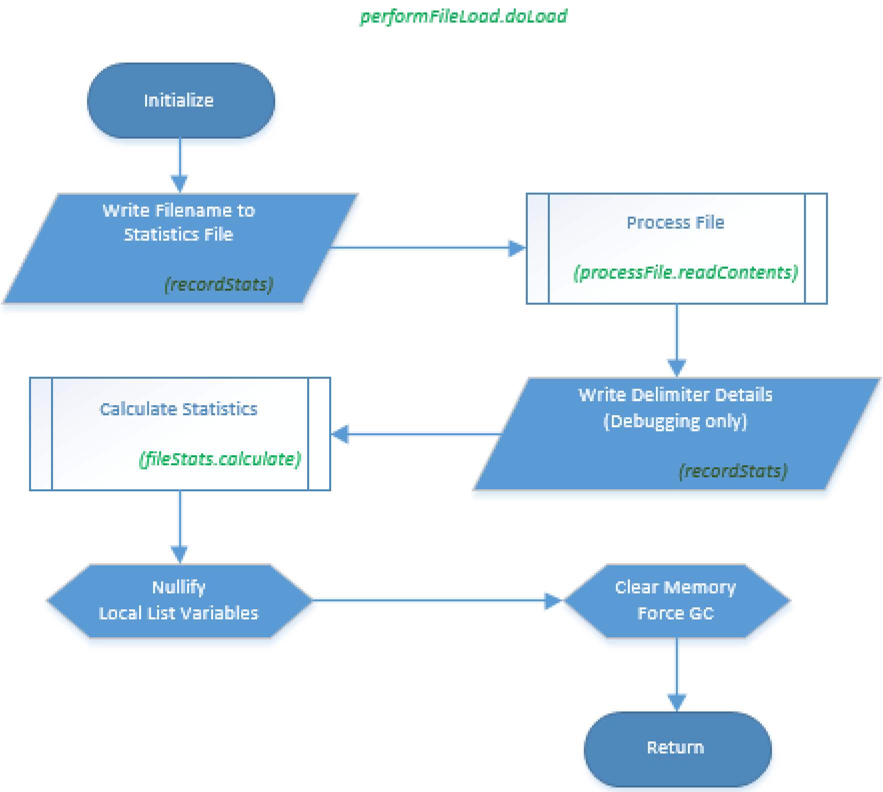


Figure 30. File processing

The file contents are read only if the file is a text file. As depicted in Figure 31, based on the percentage of the file to be read, the system will skip or continue processing each line. Once the line is in memory, the system will identify the positions of each delimiter within the text and store them for further processing in calculating the statistics. For each occurrence of the delimiter, two values will be set: a) the actual position in the string and b) the distance position, which in essence is the current position minus the prior position.

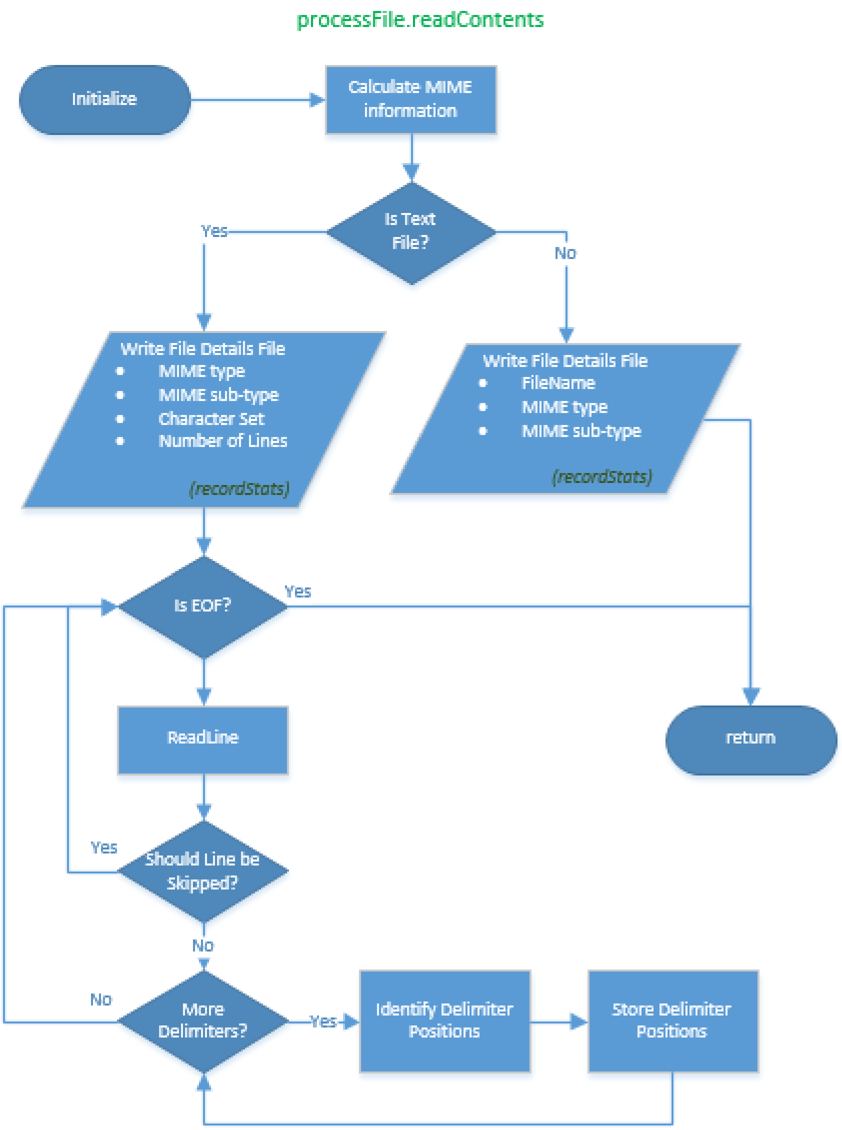


Figure 31. File content processing

Once the entire file has been processed, depending on the percentage of the file that will have to be read, the system will iterate all the delimiters and for each will calculate the statistics, as shown in Figure 32. The corresponding array populated for the delimiter will be identified, and two sets of calculations will be executed in identifying the metrics for the actual positions and the distance positions.

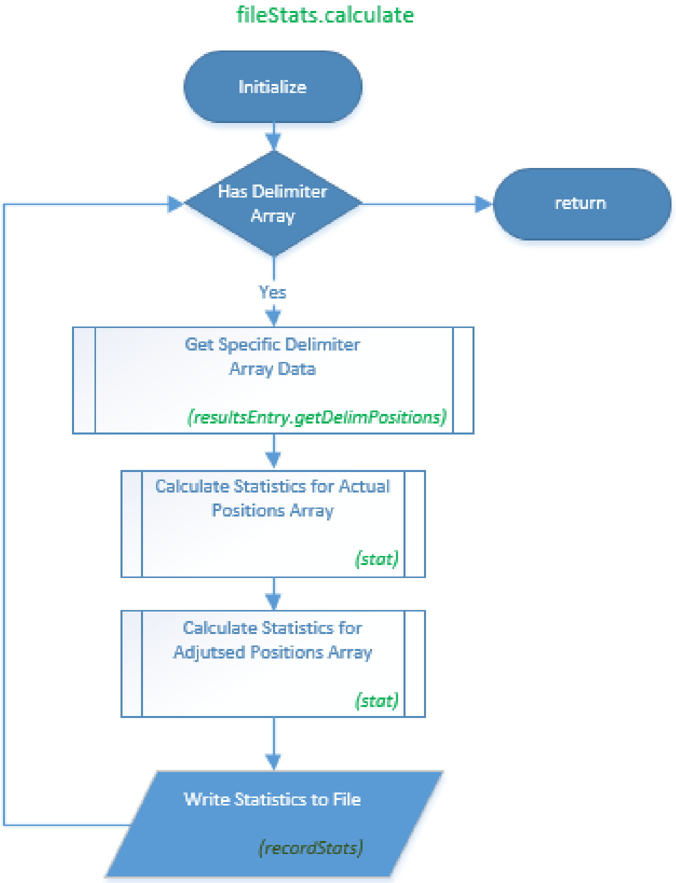

fileStats.calculate
Initialize
Has Delimiter Array
return
Yes
Get Specific Delimiter Array Data
(resultsEntry.getDelimPositions)
Calculate Statistics for Actual Positions Array
(stat)
Calculate Statistics for Adjutsed Positions Array
(stat)
Write Statistics to File
(recordStats)

Figure 32. Statistics Calculations

For each file and delimiter, the following set of metrics (formulas are available in Table 21) are calculated on the corresponding array:

- Length/Size.
- Sum, Minimum and Maximum.
- Mean, Standard Deviation and Coefficient of variance.

6.5.2. Data set Pre-Processor

The primary function of the pre-processor is to equalise the data within the data set and cater for anomalies that could distort the statistical results and invariably affect the calculation of the Neural Network function. The pre-processor, as shown in Figure 33, will identify and eliminate:

- multiple occurrences of quotation.

- escape special characters, e.g. dollar sign.
- escape delimiters within literals.

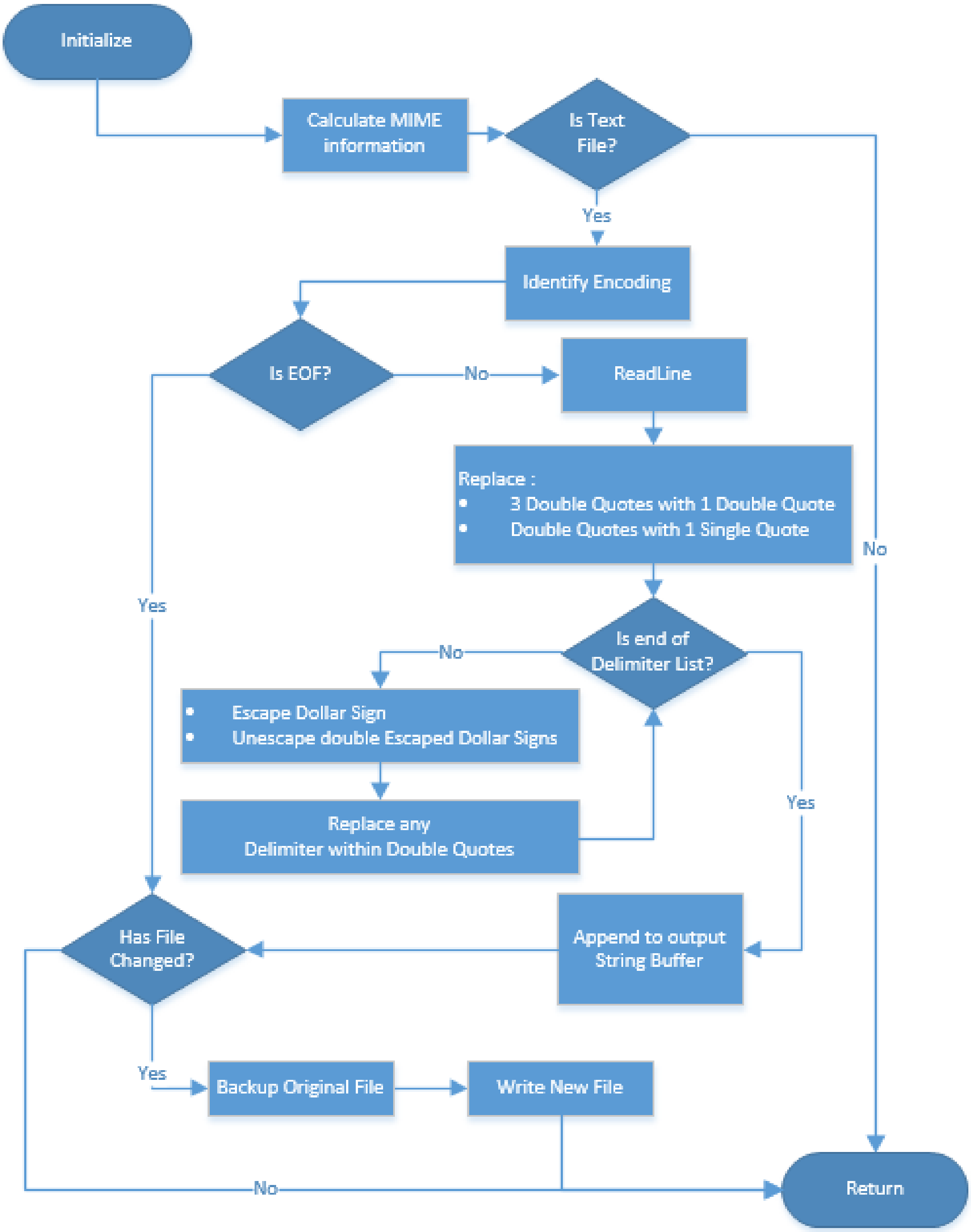

Initialize
Calculate MIME information
Is Text File?
Yes
No
Identify Encoding
Is EOF?
No
ReadLine
Replace :
3 Double Quotes with 1 Double Quote
Double Quotes with 1 Single Quote
Yes
Is end of Delimiter List?
No
Escape Dollar Sign
Unescape double Escaped Dollar Signs
Replace any Delimiter within Double Quotes
Yes
Has File Changed?
Append to output String Buffer
Yes
Backup Original File
Write New File
No
Return

Figure 33. Pre-Processor Functions

6.5.3. Neural Network processing

The neural network will read the data set statistics for each delimiter and the indicator if this is the delimiter used in the file. The training phase, utilising the configurations mentioned in Section 6.2, would calculate a formula that would most closely depict the outcome, 0 or 1, depending on whether this is the correct delimiter, having as input the set of metrics calculated from the statistical analysis of the data set. Once training is finished, the formula is ready to be used in "predicting" the correct delimiter based on the statistical values inputted.

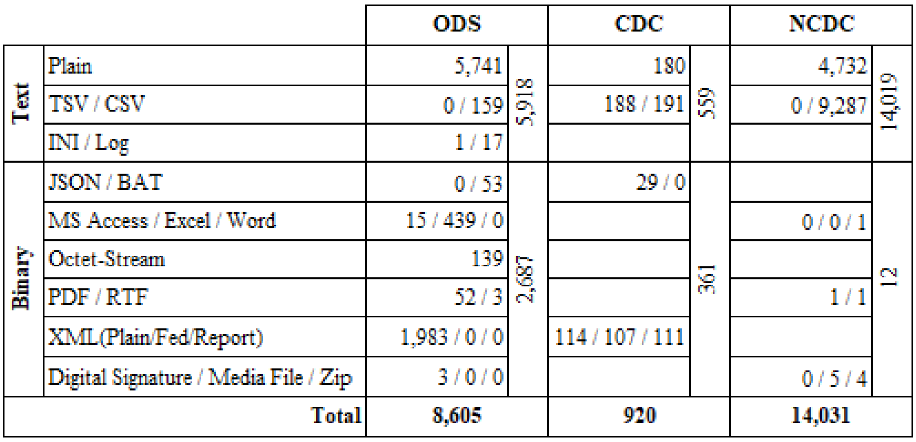
As part of the PoC, the process of creating a NN, training it and testing it was automated. The reason behind creating a script was to accommodate for the large number of iterations needed (per data set, per read percentage) with the use of nested loops. Since the number of executions was high, the process could not be human-driven to run day and night without intervention. Due to the extended number of iterations, the execution time was prolonged. As a result, the process was prone to different kinds of failures, e.g. automated widows patching restarts, power failures, bugs, and others. In making sure the process could be restarted in a consistent manner, an auditing mechanism was built into the scripts, which enabled the process owner to identify the failure and resume from the appropriate stage in case of failure.

6.6. Results

This experiment sought to explore the viability of identifying data using machine learning and to establish how much data needs to be processed to achieve this. The results of this experiment are presented in a structured, sequential manner since each result set contributes as an input into the next phase of the analysis.

The first result set consists of information about the identification of the files. For a source to be eligible for further processing, it should be identified as text. In Table 26, the classification text vs binary and the respective file MIME are presented. The files identified were: 559 CDC, 5,918 ODS and 14,019 NCDC; a total of 20,496 will participate in the subsequent analysis step. In Table 27, the encoding distribution that was identified for all the text files is presented.

Table 26. Data Sets File Content Classification



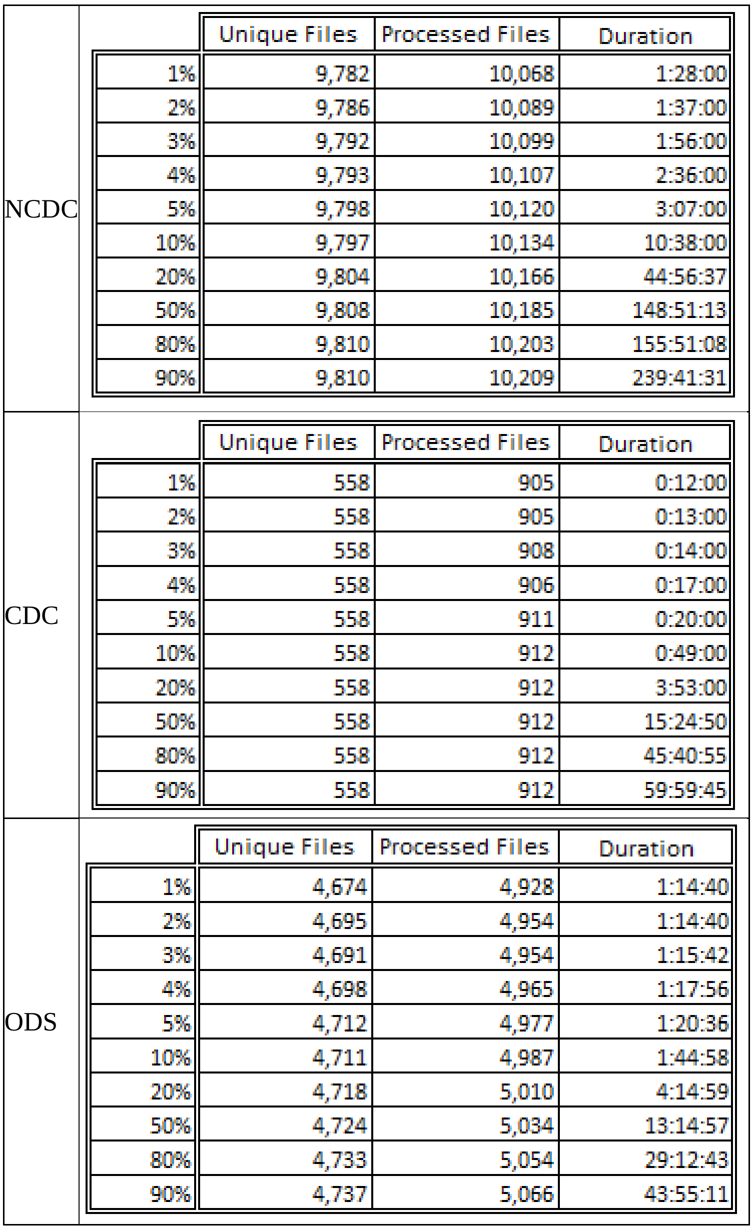
		ODS		CDC		NCDC	
Text	Plain	5,741	5,918	180	559	4,732	14,019
	TSV / CSV	0 / 159		188 / 191		0 / 9,287	
	INI / Log	1 / 17					
Binary	JSON / BAT	0 / 53	2,687	29 / 0	361		12
	MS Access / Excel / Word	15 / 439 / 0				0 / 0 / 1	
	Octet-Stream	139					
	PDF / RTF	52 / 3				1 / 1	
	XML(Plain/Fed/Report)	1,983 / 0 / 0		114 / 107 / 111			
	Digital Signature / Media File / Zip	3 / 0 / 0				0 / 5 / 4	
	Total	**8,605**		**920**		**14,031**	

Table 27. File Encoding Classification

File Encoding	ODS	CDC	NCDC
ISO-8859-1	5,625	126	13,664
ISO-8859-8	1		
Windows-1252	35		253
KOI8-R	1		
MACCYRILLIC	7		
UTF-16LE	205		
UTF-32LE	1		
UTF-8	43	443	102
Total	**5,918**	**559**	**14,019**

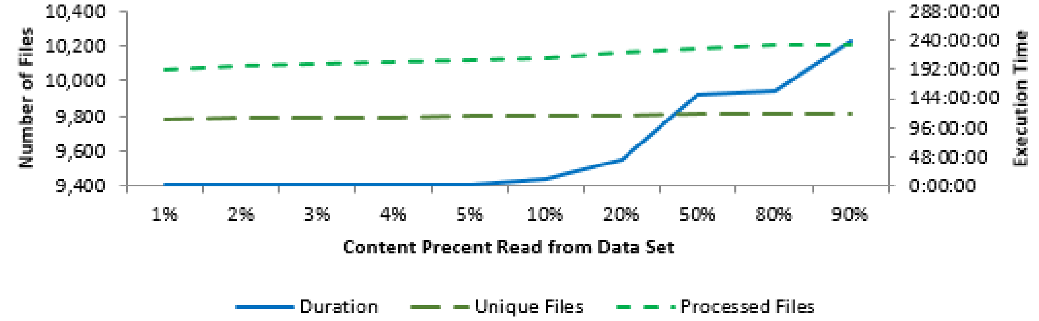
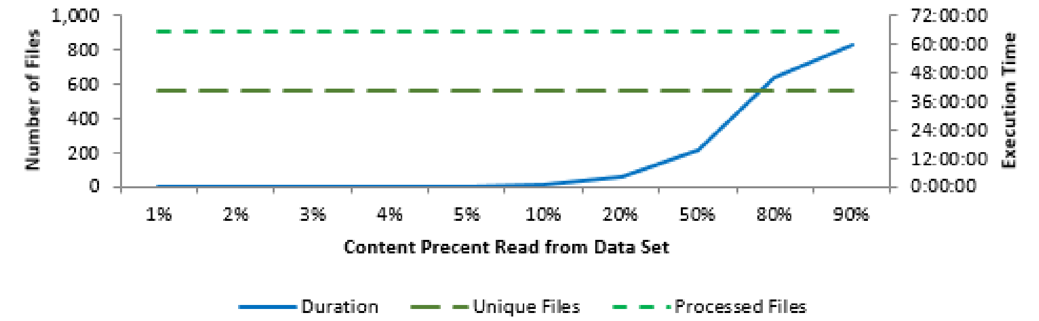
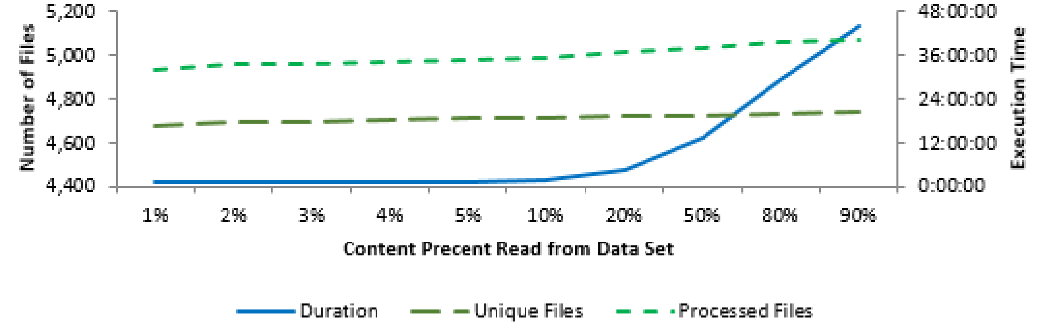
The execution time regarding the results is the next metric that was evaluated. The execution time for each percentage of line reads was measured and compared with the number of delimiters found in the files. “Unique Files” stands for the actual number of files in the dataset, whilst “Processed Files” stands for the actual number of files processed. The difference is that files identified to have multiple delimiters will have to be processed by the system as many times as the identified delimiters in order to produce the required metrics and statistics for each delimiter. As expected, the results showed a gradual increase in execution time whilst the read percentage increased. The exciting outcome was that the variation in the delimiter identification was minimal throughout all read percentages. The detailed information per dataset can be seen in a tabular format in Table 28 and using a graphical representation in Table 29.

Table 28. Datasets Execution Timing



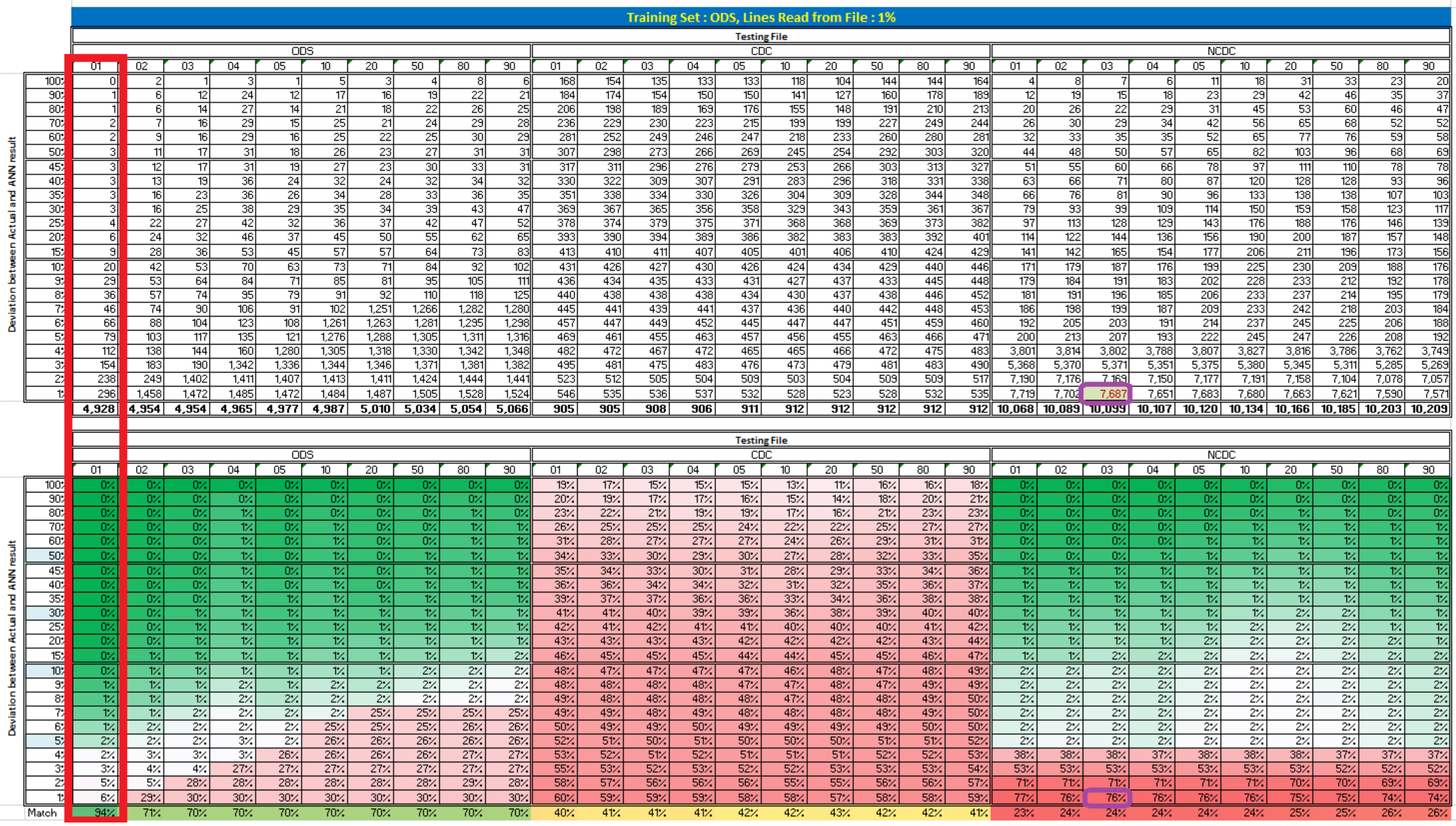
		Unique Files	Processed Files	Duration
NCDC	1%	9,782	10,068	1:28:00
	2%	9,786	10,089	1:37:00
	3%	9,792	10,099	1:56:00
	4%	9,793	10,107	2:36:00
	5%	9,798	10,120	3:07:00
	10%	9,797	10,134	10:38:00
	20%	9,804	10,166	44:56:37
	50%	9,808	10,185	148:51:13
	80%	9,810	10,203	155:51:08
	90%	9,810	10,209	239:41:31
CDC	1%	558	905	0:12:00
	2%	558	905	0:13:00
	3%	558	908	0:14:00
	4%	558	906	0:17:00
	5%	558	911	0:20:00
	10%	558	912	0:49:00
	20%	558	912	3:53:00
	50%	558	912	15:24:50
	80%	558	912	45:40:55
	90%	558	912	59:59:45
ODS	1%	4,674	4,928	1:14:40
	2%	4,695	4,954	1:14:40
	3%	4,691	4,954	1:15:42
	4%	4,698	4,965	1:17:56
	5%	4,712	4,977	1:20:36
	10%	4,711	4,987	1:44:58
	20%	4,718	5,010	4:14:59
	50%	4,724	5,034	13:14:57
	80%	4,733	5,054	29:12:43
	90%	4,737	5,066	43:55:11

Table 29. Datasets Execution trend lines

NCDC	
CDC	
ODS	

Based on the results, it can be concluded that the most interesting percentages are between 1 and 3, since they have low execution timings and relatively high accuracy. Nevertheless, it is imperative to validate other metrics, like the efficiency of the location of delimiters and the file structure's actual breakdown, before concluding. For that reason, further statistical analyses were performed and were imported into a neural network as input parameters. The network was utilised in accurately identifying the delimiter of each file. In matching the results of the ANN with the actual data, a manual and semi-manual process was employed to verify the files delimiters and cross-reference them with the ANN output. The initial series of tests, unit testing the implementation, with one dataset whilst reading 1% of the file's content, was quite revealing and unexpected. The accuracy was so high that in certain cases where the ANN result was in conflict with the manually identified delimiter, it was confirmed that there was a human error in the manual classification, and the ANN identified the correct delimiter. At that time, having confidence in the results from the concluded unit testing, the experiment was expanded to cover all content reading percentages (1%, 2%, 3%, 4%, %5, 10%, 20%, 50%, 80%, 90%) and all the datasets (ODS, CDC, NCDC) which resulted in a set of 30 files to be referred to as “% read” files. The training and final calculation of the ANN function for each set, resulting in 30 neural networks, were stored in independent files. Each derived ANN was then applied into the respective 30 files from the “% read” files,” producing a set of 900 result files, one file per neural network per “% read” file. A cross-tabulation and the related heat map were created for each of the 30 training networks and their associated 30 files. Figure 34 illustrates the ODS-based network's output for 1% lines read against all other sets and read percentages. The first cross-tabulation shows the number of unmatched files, and the second cross-tabulation heat-map shows the match percentage. We have the input file tested per data set and per percentage of lines read on the horizontal axis. On the vertical axis, we have an inverted percentage representing the percentile difference per file percentage that was exhibited when comparing the ANN result with the actual delimiter value of the file. This percentile labelled “Deviation between Actual and ANN results” indicates how much difference is exhibited between the actual value (0 or 1) and the network calculated value (between -1 and 2 as shown in Figure 28), which in essence is the neural network performance. The ANN formula is defined by training the ANN using the data collected from processing the ODS data set and reading 1% of each file’s content. The formula identified is then used to predict the confidence level of the delimiter using the CDC and NCDC data sets and for all content read percentages (1%, 2%, 3%, 4%, 5%, 10%, 20%, 50%, 80%, 90%). The data elements represented are counts and their respective percentile values.

- In the upper part of the figure cross-tabulation, there is a list of the number of files that fall into that category. The highlighted data element (in purple) represents that 7,687 files exhibit a difference less than or equal to 1% when comparing the ANN result with the actual delimiter. The tested ANN formula is calculated based on the ODS dataset having read 1% of the file content and tested against the NCDC data set while reading 3% of the file content.
- In the heat-map, the percentage of files that fall into that category is shown. The counts from the upper cross-tabulation are depicted as percentages (count/file number). Taking the same example, 76% (in purple) of the files exhibit a difference less than or equal to 1% when comparing the ANN result with the actual delimiter.



Training Set : ODS, Lines Read from File : 1%

Deviation between Actual and ANN result

	Testing File																													
	ODS										CDC										NCDC									
	01	02	03	04	05	10	20	50	80	90	01	02	03	04	05	10	20	50	80	90	01	02	03	04	05	10	20	50	80	90
100%	0	2	1	3	1	5	3	4	8	6	168	154	135	133	133	118	104	144	144	164	4	8	7	6	11	18	31	33	23	20
90%	1	6	12	24	12	17	16	19	22	21	184	174	154	150	150	141	127	160	178	189	12	19	15	18	23	29	42	46	35	37
80%	1	6	14	27	14	21	18	22	26	25	206	198	189	169	176	155	148	191	210	213	20	26	22	29	31	45	53	60	46	47
70%	2	7	16	29	15	25	21	24	29	28	236	229	230	223	215	199	199	227	249	244	26	30	29	34	42	56	65	68	52	52
60%	2	9	16	29	16	25	22	25	30	29	281	252	249	246	247	218	233	260	280	281	32	33	35	35	52	65	77	76	59	58
50%	3	11	17	31	18	26	23	27	31	31	307	298	273	266	269	245	254	292	303	320	44	48	50	57	65	82	103	96	68	69
45%	3	12	17	31	19	27	23	30	33	31	317	311	296	276	279	253	266	303	313	327	51	55	60	66	78	97	111	110	78	78
40%	3	13	19	36	24	32	24	32	34	32	330	322	309	307	291	283	296	318	331	338	63	66	71	80	87	120	128	128	93	96
35%	3	16	23	36	26	34	28	33	36	35	351	338	334	330	326	304	309	328	344	348	66	76	81	90	96	133	138	138	107	103
30%	3	16	25	38	29	35	34	39	43	47	369	367	365	356	358	329	343	359	361	367	79	93	99	109	114	150	159	158	123	117
25%	4	22	27	42	32	36	37	42	47	52	378	374	379	375	371	368	368	369	373	382	97	113	128	129	143	176	188	176	146	139
20%	6	24	32	46	37	45	50	55	62	65	393	390	394	389	386	382	383	383	392	401	114	122	144	136	156	190	200	187	157	148
15%	9	28	36	53	45	57	57	64	73	83	413	410	411	407	405	401	406	410	424	429	141	142	165	154	177	206	211	196	173	156
10%	20	42	53	70	63	73	71	84	92	102	431	426	427	430	426	424	434	429	440	446	171	179	187	176	199	225	230	209	188	176
9%	29	53	64	84	71	85	81	95	105	111	436	434	435	433	431	427	437	433	445	448	179	184	191	183	202	228	233	212	192	178
8%	36	57	74	95	79	91	92	110	118	125	440	438	438	438	434	430	437	438	446	452	181	191	196	185	206	233	237	214	195	179
7%	46	74	90	106	91	102	1,251	1,266	1,282	1,280	445	441	439	441	437	436	440	442	448	453	186	198	199	187	209	233	242	218	203	184
6%	66	88	104	123	108	1,261	1,263	1,281	1,295	1,298	457	447	449	452	445	447	447	451	459	460	192	205	203	191	214	237	245	225	206	188
5%	79	103	117	135	121	1,276	1,288	1,305	1,311	1,316	469	461	455	463	457	456	455	463	466	471	200	213	207	193	222	245	247	226	208	192
4%	112	138	144	160	1,280	1,305	1,318	1,330	1,342	1,348	482	472	467	472	465	465	466	472	475	483	3,801	3,814	3,802	3,788	3,807	3,827	3,816	3,786	3,762	3,749
3%	154	183	190	1,342	1,336	1,344	1,346	1,371	1,381	1,382	495	481	475	483	476	473	479	481	483	490	5,368	5,370	5,371	5,351	5,375	5,380	5,345	5,311	5,285	5,269
2%	238	249	1,402	1,411	1,407	1,413	1,411	1,424	1,444	1,441	523	512	505	504	509	503	504	509	509	517	7,190	7,176	7,169	7,150	7,177	7,191	7,158	7,104	7,078	7,057
1%	296	1,458	1,472	1,485	1,472	1,484	1,487	1,505	1,528	1,524	546	535	536	537	532	528	523	528	532	535	7,719	7,702	7,687	7,651	7,683	7,680	7,663	7,621	7,590	7,571
	4,928	**4,954**	**4,954**	**4,965**	**4,977**	**4,987**	**5,010**	**5,034**	**5,054**	**5,066**	**905**	**905**	**908**	**906**	**911**	**912**	**912**	**912**	**912**	**912**	**10,068**	**10,089**	**10,099**	**10,107**	**10,120**	**10,134**	**10,166**	**10,185**	**10,203**	**10,209**

Deviation between Actual and ANN result

	Testing File																													
	ODS										CDC										NCDC									
	01	02	03	04	05	10	20	50	80	90	01	02	03	04	05	10	20	50	80	90	01	02	03	04	05	10	20	50	80	90
100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	19%	17%	15%	15%	15%	13%	11%	16%	16%	18%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
90%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	20%	19%	17%	17%	16%	15%	14%	18%	20%	21%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
80%	0%	0%	0%	1%	0%	0%	0%	0%	1%	0%	23%	22%	21%	19%	19%	17%	16%	21%	23%	23%	0%	0%	0%	0%	0%	0%	1%	1%	0%	0%
70%	0%	0%	0%	1%	0%	1%	0%	0%	1%	1%	26%	25%	25%	25%	24%	22%	22%	25%	27%	27%	0%	0%	0%	0%	0%	1%	1%	1%	1%	1%
60%	0%	0%	0%	1%	0%	1%	0%	0%	1%	1%	31%	28%	27%	27%	27%	24%	26%	29%	31%	31%	0%	0%	0%	0%	1%	1%	1%	1%	1%	1%
50%	0%	0%	0%	1%	0%	1%	0%	1%	1%	1%	34%	33%	30%	29%	30%	27%	28%	32%	33%	35%	0%	0%	0%	1%	1%	1%	1%	1%	1%	1%
45%	0%	0%	0%	1%	0%	1%	0%	1%	1%	1%	35%	34%	33%	30%	31%	28%	29%	33%	34%	36%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
40%	0%	0%	0%	1%	0%	1%	0%	1%	1%	1%	36%	36%	34%	34%	32%	31%	32%	35%	36%	37%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
35%	0%	0%	0%	1%	1%	1%	1%	1%	1%	1%	39%	37%	37%	36%	36%	33%	34%	36%	38%	38%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
30%	0%	0%	1%	1%	1%	1%	1%	1%	1%	1%	41%	41%	40%	39%	39%	36%	38%	39%	40%	40%	1%	1%	1%	1%	1%	1%	2%	2%	1%	1%
25%	0%	0%	1%	1%	1%	1%	1%	1%	1%	1%	42%	41%	42%	41%	41%	40%	40%	40%	41%	42%	1%	1%	1%	1%	1%	2%	2%	2%	1%	1%
20%	0%	0%	1%	1%	1%	1%	1%	1%	1%	1%	43%	43%	43%	43%	42%	42%	42%	42%	43%	44%	1%	1%	1%	1%	2%	2%	2%	2%	2%	1%
15%	0%	1%	1%	1%	1%	1%	1%	1%	1%	2%	46%	45%	45%	45%	44%	44%	45%	45%	46%	47%	1%	1%	2%	2%	2%	2%	2%	2%	2%	2%
10%	0%	1%	1%	1%	1%	1%	1%	2%	2%	2%	48%	47%	47%	47%	47%	46%	48%	47%	48%	49%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
9%	1%	1%	1%	2%	1%	2%	2%	2%	2%	2%	48%	48%	48%	48%	47%	47%	48%	47%	49%	49%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
8%	1%	1%	1%	2%	2%	2%	2%	2%	2%	2%	49%	48%	48%	48%	48%	47%	48%	48%	49%	50%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
7%	1%	1%	2%	2%	2%	2%	25%	25%	25%	25%	49%	49%	48%	49%	48%	48%	48%	48%	49%	50%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
6%	1%	2%	2%	2%	2%	25%	25%	25%	26%	26%	50%	49%	49%	50%	49%	49%	49%	49%	50%	50%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
5%	2%	2%	2%	3%	2%	26%	26%	26%	26%	26%	52%	51%	50%	51%	50%	50%	50%	51%	51%	52%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
4%	2%	3%	3%	3%	26%	26%	26%	26%	27%	27%	53%	52%	51%	52%	51%	51%	51%	52%	52%	53%	38%	38%	38%	37%	38%	38%	38%	37%	37%	37%
3%	3%	4%	4%	27%	27%	27%	27%	27%	27%	27%	55%	53%	52%	53%	52%	52%	53%	53%	53%	54%	53%	53%	53%	53%	53%	53%	53%	52%	52%	52%
2%	5%	5%	28%	28%	28%	28%	28%	28%	29%	28%	58%	57%	56%	56%	56%	55%	55%	56%	56%	57%	71%	71%	71%	71%	71%	71%	70%	70%	69%	69%
1%	6%	29%	30%	30%	30%	30%	30%	30%	30%	30%	60%	59%	59%	59%	58%	58%	57%	58%	58%	59%	77%	76%	76%	76%	76%	76%	75%	75%	74%	74%
Match	94%	71%	70%	70%	70%	70%	70%	70%	70%	70%	40%	41%	41%	41%	42%	42%	43%	42%	42%	41%	23%	24%	24%	24%	24%	24%	25%	25%	26%	26%

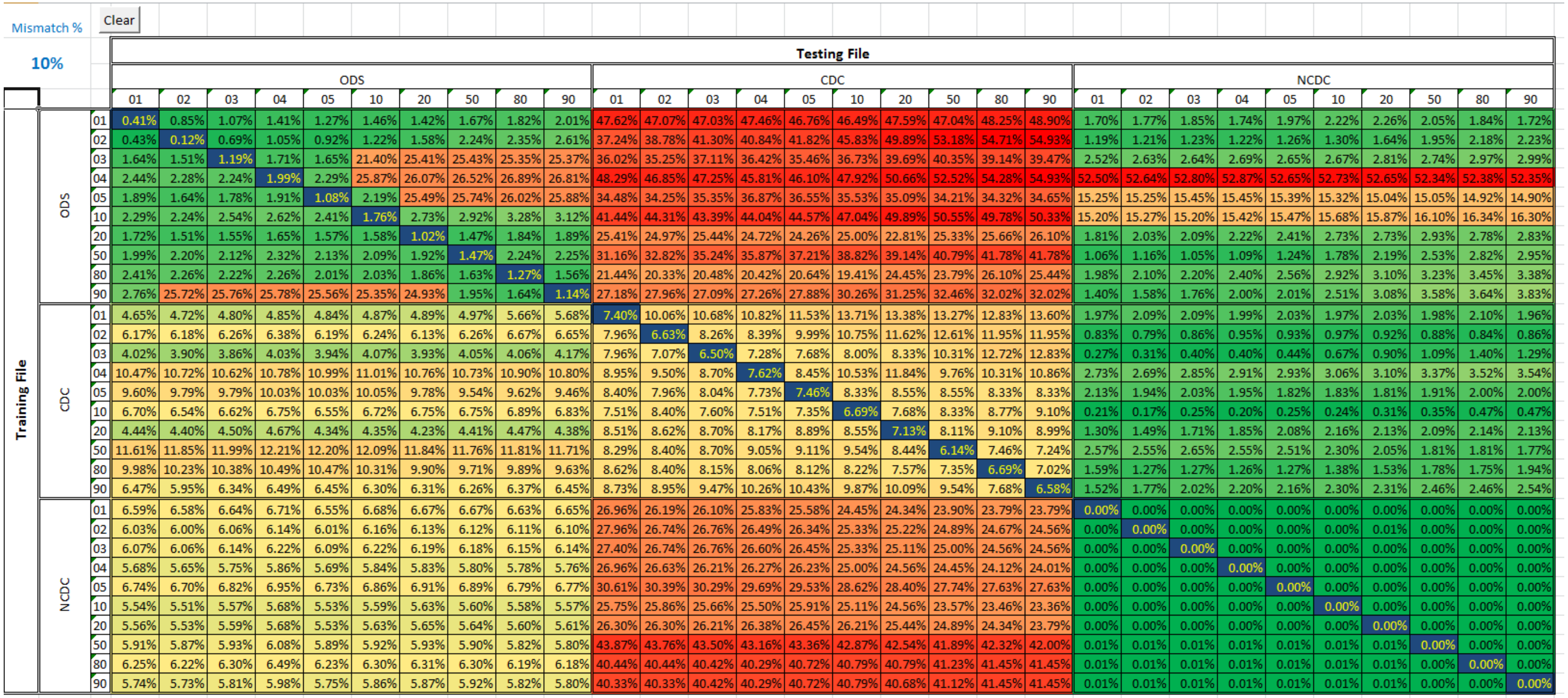
Figure 34. Neural Network Results in Cross-Tabulation Heat Map sample (1 out of 30)

The heat map is used to visualise the progression of differentiation, exhibiting the margin of error. Based on the colour coding, the map indicates that although there is high differentiation when testing with NCDC (only 23%-26% matches), the bulk number of erroneously classified files is between 1% and 4% since, after that, the number of files is dramatically reduced in contrast to CDC where although 40% is matched the deviation levels remain high throughout.

One column would be invalid for each of the aforementioned cross-tabulations since the training and testing files would be the same. In the sample depicted in Figure 34, the column ODS / 01 with 94% match, highlighted in the red parallelogram, will be invalid since the training and testing data for the ANN is an ODS dataset file which has read 1% of the file content. It quickly became apparent that the difference mentioned above in CDC variation was exhibited in all 30 cross-tabulations. The exhibited difference ranged from 49% and reached a rate of 98%, depending on the data set and read percentage used to train the ANN, which indicated an error or a peculiarity with the specific dataset. Having confirmed the calculations had no error, it became apparent that *Variety* was in play, and the “Escape Special Context” pre-processor as described in Section 6.5.2 was implemented in adjusting the model to cater to the CDC set's unique characteristics. As a result of the pre-processor, the effect was minimised, having values from 10% to 60%, but it was still apparent that the specific set was acting differently. This preliminary differentiation is vital since it is an indication that the model can auto-adjust itself and perform with high accuracy even though an outlier dataset was introduced into the ecosystem.

A meta-heat map was created to understand the results further by, aggregating the previous results from the independent cross-tabulations. Given an acceptable percentage for the probability of error, the heat map aggregated the data from the 30 individual maps by displaying the respective slice of each table. In Table 30, we can see the heat map for a mismatch of 10%. Just like in the original cross-tabulation, the cases where training and testing were performed on the same dataset should not be considered (the values in blue).

Table 30. Meta-Heat Map for 10% Mismatch of the Neural Network Results



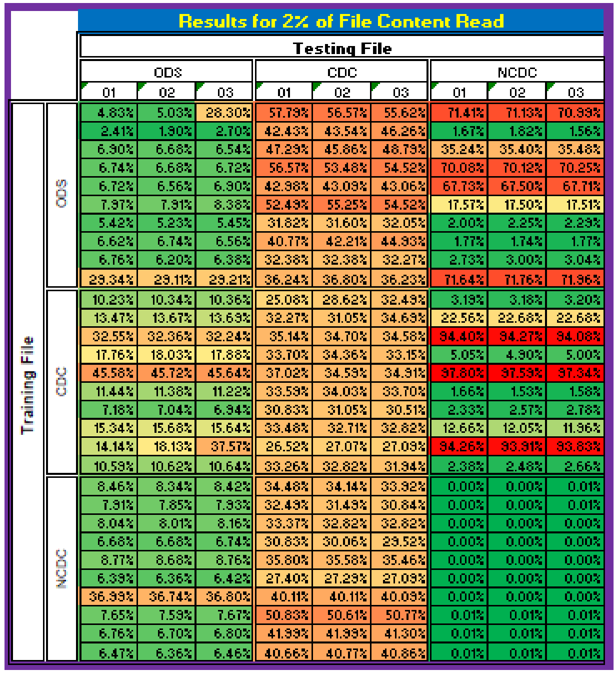
Mismatch %

10%

Clear

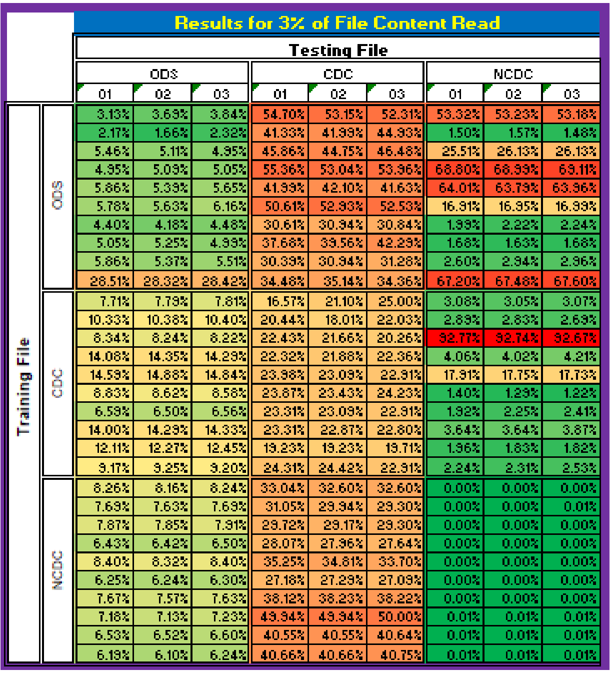
		Testing File																													
		ODS										CDC										NCDC									
		01	02	03	04	05	10	20	50	80	90	01	02	03	04	05	10	20	50	80	90	01	02	03	04	05	10	20	50	80	90
Training File	ODS	01	0.41%	0.85%	1.07%	1.41%	1.27%	1.46%	1.42%	1.67%	1.82%	2.01%	47.62%	47.07%	47.03%	47.46%	46.76%	46.49%	47.59%	47.04%	48.25%	48.90%	1.70%	1.77%	1.85%	1.74%	1.97%	2.22%	2.26%	2.05%	1.84%	1.72%
		02	0.43%	0.12%	0.69%	1.05%	0.92%	1.22%	1.58%	2.24%	2.35%	2.61%	37.24%	38.78%	41.30%	40.84%	41.82%	45.83%	49.89%	53.18%	54.71%	54.93%	1.19%	1.21%	1.23%	1.22%	1.26%	1.30%	1.64%	1.95%	2.18%	2.23%
		03	1.64%	1.51%	1.19%	1.71%	1.65%	21.40%	25.41%	25.43%	25.35%	25.37%	36.02%	35.25%	37.11%	36.42%	35.46%	36.73%	39.69%	40.35%	39.14%	39.47%	2.52%	2.63%	2.64%	2.69%	2.65%	2.67%	2.81%	2.74%	2.97%	2.99%
		04	2.44%	2.28%	2.24%	1.99%	2.29%	25.87%	26.07%	26.52%	26.89%	26.81%	48.29%	46.85%	47.25%	45.81%	46.10%	47.92%	50.66%	52.52%	54.28%	54.93%	52.50%	52.64%	52.80%	52.87%	52.65%	52.73%	52.65%	52.34%	52.38%	52.35%
		05	1.89%	1.64%	1.78%	1.91%	1.08%	2.19%	25.49%	25.74%	26.02%	25.88%	34.48%	34.25%	35.35%	36.87%	36.55%	35.53%	35.09%	34.21%	34.32%	34.65%	15.25%	15.25%	15.45%	15.45%	15.39%	15.32%	15.04%	15.05%	14.92%	14.90%
		10	2.29%	2.24%	2.54%	2.62%	2.41%	1.76%	2.73%	2.92%	3.28%	3.12%	41.44%	44.31%	43.39%	44.04%	44.57%	47.04%	49.89%	50.55%	49.78%	50.33%	15.20%	15.27%	15.20%	15.42%	15.47%	15.68%	15.87%	16.10%	16.34%	16.30%
		20	1.72%	1.51%	1.55%	1.65%	1.57%	1.58%	1.02%	1.47%	1.84%	1.89%	25.41%	24.97%	25.44%	24.72%	24.26%	25.00%	22.81%	25.33%	25.66%	26.10%	1.81%	2.03%	2.09%	2.22%	2.41%	2.73%	2.73%	2.93%	2.78%	2.83%
		50	1.99%	2.20%	2.12%	2.32%	2.13%	2.09%	1.92%	1.47%	2.24%	2.25%	31.16%	32.82%	35.24%	35.87%	37.21%	38.82%	39.14%	40.79%	41.78%	41.78%	1.06%	1.16%	1.05%	1.09%	1.24%	1.78%	2.19%	2.53%	2.82%	2.95%
		80	2.41%	2.26%	2.22%	2.26%	2.01%	2.03%	1.86%	1.63%	1.27%	1.56%	21.44%	20.33%	20.48%	20.42%	20.64%	19.41%	24.45%	23.79%	26.10%	25.44%	1.98%	2.10%	2.20%	2.40%	2.56%	2.92%	3.10%	3.23%	3.45%	3.38%
		90	2.76%	25.72%	25.76%	25.78%	25.56%	25.35%	24.93%	1.95%	1.64%	1.14%	27.18%	27.96%	27.09%	27.26%	27.88%	30.26%	31.25%	32.46%	32.02%	32.02%	1.40%	1.58%	1.76%	2.00%	2.01%	2.51%	3.08%	3.58%	3.64%	3.83%
	CDC	01	4.65%	4.72%	4.80%	4.85%	4.84%	4.87%	4.89%	4.97%	5.66%	5.68%	7.40%	10.06%	10.68%	10.82%	11.53%	13.71%	13.38%	13.27%	12.83%	13.60%	1.97%	2.09%	2.09%	1.99%	2.03%	1.97%	2.03%	1.98%	2.10%	1.96%
		02	6.17%	6.18%	6.26%	6.38%	6.19%	6.24%	6.13%	6.26%	6.67%	6.65%	7.96%	6.63%	8.26%	8.39%	9.99%	10.75%	11.62%	12.61%	11.95%	11.95%	0.83%	0.79%	0.86%	0.95%	0.93%	0.97%	0.92%	0.88%	0.84%	0.86%
		03	4.02%	3.90%	3.86%	4.03%	3.94%	4.07%	3.93%	4.05%	4.06%	4.17%	7.96%	7.07%	6.50%	7.28%	7.68%	8.00%	8.33%	10.31%	12.72%	12.83%	0.27%	0.31%	0.40%	0.40%	0.44%	0.67%	0.90%	1.09%	1.40%	1.29%
		04	10.47%	10.72%	10.62%	10.78%	10.99%	11.01%	10.76%	10.73%	10.90%	10.80%	8.95%	9.50%	8.70%	7.62%	8.45%	10.53%	11.84%	9.76%	10.31%	10.86%	2.73%	2.69%	2.85%	2.91%	2.93%	3.06%	3.10%	3.37%	3.52%	3.54%
		05	9.60%	9.79%	9.79%	10.03%	10.03%	10.05%	9.78%	9.54%	9.62%	9.46%	8.40%	7.96%	8.04%	7.73%	7.46%	8.33%	8.55%	8.55%	8.33%	8.33%	2.13%	1.94%	2.03%	1.95%	1.82%	1.83%	1.81%	1.91%	2.00%	2.00%
		10	6.70%	6.54%	6.62%	6.75%	6.55%	6.72%	6.75%	6.75%	6.89%	6.83%	7.51%	8.40%	7.60%	7.51%	7.35%	6.69%	7.68%	8.33%	8.77%	9.10%	0.21%	0.17%	0.25%	0.20%	0.25%	0.24%	0.31%	0.35%	0.47%	0.47%
		20	4.44%	4.40%	4.50%	4.67%	4.34%	4.35%	4.23%	4.41%	4.47%	4.38%	8.51%	8.62%	8.70%	8.17%	8.89%	8.55%	7.13%	8.11%	9.10%	8.99%	1.30%	1.49%	1.71%	1.85%	2.08%	2.16%	2.13%	2.09%	2.14%	2.13%
		50	11.61%	11.85%	11.99%	12.21%	12.20%	12.09%	11.84%	11.76%	11.81%	11.71%	8.29%	8.40%	8.70%	9.05%	9.11%	9.54%	8.44%	6.14%	7.46%	7.24%	2.57%	2.55%	2.65%	2.55%	2.51%	2.30%	2.05%	1.81%	1.81%	1.77%
		80	9.98%	10.23%	10.38%	10.49%	10.47%	10.31%	9.90%	9.71%	9.89%	9.63%	8.62%	8.40%	8.15%	8.06%	8.12%	8.22%	7.57%	7.35%	6.69%	7.02%	1.59%	1.27%	1.27%	1.26%	1.27%	1.38%	1.53%	1.78%	1.75%	1.94%
		90	6.47%	5.95%	6.34%	6.49%	6.45%	6.30%	6.31%	6.26%	6.37%	6.45%	8.73%	8.95%	9.47%	10.26%	10.43%	9.87%	10.09%	9.54%	7.68%	6.58%	1.52%	1.77%	2.02%	2.20%	2.16%	2.30%	2.31%	2.46%	2.46%	2.54%
	NCDC	01	6.59%	6.58%	6.64%	6.71%	6.55%	6.68%	6.67%	6.67%	6.63%	6.65%	26.96%	26.19%	26.10%	25.83%	25.58%	24.45%	24.34%	23.90%	23.79%	23.79%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		02	6.03%	6.00%	6.06%	6.14%	6.01%	6.16%	6.13%	6.12%	6.11%	6.10%	27.96%	26.74%	26.76%	26.49%	26.34%	25.33%	25.22%	24.89%	24.67%	24.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%
		03	6.07%	6.06%	6.14%	6.22%	6.09%	6.22%	6.19%	6.18%	6.15%	6.14%	27.40%	26.74%	26.76%	26.60%	26.45%	25.33%	25.11%	25.00%	24.56%	24.56%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		04	5.68%	5.65%	5.75%	5.86%	5.69%	5.84%	5.83%	5.80%	5.78%	5.76%	26.96%	26.63%	26.21%	26.27%	26.23%	25.00%	24.56%	24.45%	24.12%	24.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		05	6.74%	6.70%	6.82%	6.95%	6.73%	6.86%	6.91%	6.89%	6.79%	6.77%	30.61%	30.39%	30.29%	29.69%	29.53%	28.62%	28.40%	27.74%	27.63%	27.63%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		10	5.54%	5.51%	5.57%	5.68%	5.53%	5.59%	5.63%	5.60%	5.58%	5.57%	25.75%	25.86%	25.66%	25.50%	25.91%	25.11%	24.56%	23.57%	23.46%	23.36%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		20	5.56%	5.53%	5.59%	5.68%	5.53%	5.63%	5.65%	5.64%	5.60%	5.61%	26.30%	26.30%	26.21%	26.38%	26.45%	26.21%	25.44%	24.89%	24.34%	23.79%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
		50	5.91%	5.87%	5.93%	6.08%	5.89%	5.92%	5.93%	5.90%	5.82%	5.80%	43.87%	43.76%	43.50%	43.16%	43.36%	42.87%	42.54%	41.89%	42.32%	42.00%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%
		80	6.25%	6.22%	6.30%	6.49%	6.23%	6.30%	6.31%	6.30%	6.19%	6.18%	40.44%	40.44%	40.42%	40.29%	40.72%	40.79%	40.79%	41.23%	41.45%	41.45%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%
		90	5.74%	5.73%	5.81%	5.98%	5.75%	5.86%	5.87%	5.92%	5.82%	5.80%	40.33%	40.33%	40.42%	40.29%	40.72%	40.79%	40.68%	41.12%	41.45%	41.45%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%	0.00%	0.00%	0.00%

Depending on the risk “appetite,” the data analyst/scientist can identify the percentages that each ANN function would yield against each set and retrieve the respective heat map. Taking into consideration that a) an error level of 1% to 10% would be a viable risk for analysis and b) the timings of reading percentages for 1% to 3% were the fastest; a subset of the heat map in Table 30 is created for each percentage, e.g. the third heat map of Figure 35. The comparative analysis confirmed that the best results yielded from 2% to 3%; see details in the first and second heat maps of Figure 35. Since 2% has a lower execution time than 3% for all dataset processing, see Table 28, it is concluded that the optimal percentage would be 2%.



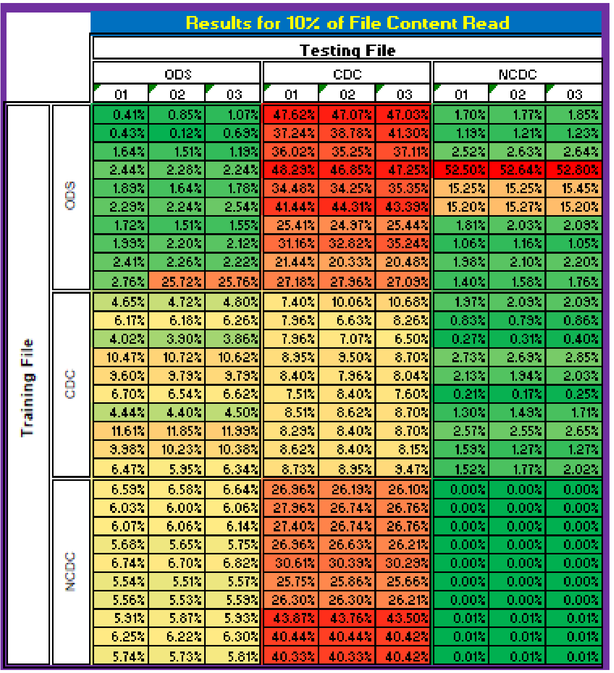
Results for 2% of File Content Read

		Testing File								
		ODS			CDC			NCDC		
		01	02	03	01	02	03	01	02	03
Training File	ODS	4.83%	5.03%	28.30%	57.79%	56.57%	55.62%	71.41%	71.13%	70.99%
		2.41%	1.90%	2.70%	42.43%	43.54%	46.26%	1.67%	1.82%	1.56%
		6.90%	6.68%	6.54%	47.29%	45.86%	48.79%	35.24%	35.40%	35.48%
		6.74%	6.68%	6.72%	56.57%	53.48%	54.52%	70.08%	70.12%	70.25%
		6.72%	6.56%	6.90%	42.98%	43.09%	43.06%	67.73%	67.50%	67.71%
		7.97%	7.91%	8.38%	52.49%	55.25%	54.52%	17.57%	17.50%	17.51%
		5.42%	5.23%	5.45%	31.82%	31.60%	32.05%	2.00%	2.25%	2.29%
		6.62%	6.74%	6.56%	40.77%	42.21%	44.93%	1.77%	1.74%	1.77%
		6.76%	6.20%	6.38%	32.38%	32.38%	32.27%	2.73%	3.00%	3.04%
		29.34%	29.11%	29.21%	36.24%	36.80%	36.23%	71.64%	71.76%	71.96%
	CDC	10.23%	10.34%	10.36%	25.08%	28.62%	32.49%	3.19%	3.18%	3.20%
		13.47%	13.67%	13.69%	32.27%	31.05%	34.69%	22.56%	22.68%	22.68%
		32.55%	32.36%	32.24%	35.14%	34.70%	34.58%	94.40%	94.27%	94.08%
		17.76%	18.03%	17.88%	33.70%	34.36%	33.15%	5.05%	4.90%	5.00%
		45.58%	45.72%	45.64%	37.02%	34.59%	34.91%	97.80%	97.59%	97.34%
		11.44%	11.38%	11.22%	33.59%	34.03%	33.70%	1.66%	1.53%	1.58%
		7.18%	7.04%	6.94%	30.83%	31.05%	30.51%	2.33%	2.57%	2.78%
		15.34%	15.68%	15.64%	33.48%	32.71%	32.82%	12.66%	12.05%	11.96%
		14.14%	18.13%	37.57%	26.52%	27.07%	27.09%	94.26%	93.91%	93.83%
		10.59%	10.62%	10.64%	33.26%	32.82%	31.94%	2.38%	2.48%	2.66%
	NCDC	8.46%	8.34%	8.42%	34.48%	34.14%	33.92%	0.00%	0.00%	0.01%
		7.91%	7.85%	7.93%	32.49%	31.49%	30.84%	0.00%	0.00%	0.01%
		8.04%	8.01%	8.16%	33.37%	32.82%	32.82%	0.00%	0.00%	0.00%
		6.68%	6.68%	6.74%	30.83%	30.06%	29.52%	0.00%	0.00%	0.00%
		8.77%	8.68%	8.76%	35.80%	35.58%	35.46%	0.00%	0.00%	0.00%
		6.39%	6.36%	6.42%	27.40%	27.29%	27.09%	0.00%	0.00%	0.00%
		36.99%	36.74%	36.80%	40.11%	40.11%	40.09%	0.00%	0.00%	0.00%
		7.65%	7.59%	7.67%	50.83%	50.61%	50.77%	0.01%	0.01%	0.01%
		6.76%	6.70%	6.80%	41.99%	41.99%	41.30%	0.01%	0.01%	0.01%
		6.47%	6.36%	6.46%	40.66%	40.77%	40.86%	0.01%	0.01%	0.01%



Results for 3% of File Content Read

		Testing File								
		ODS			CDC			NCDC		
		01	02	03	01	02	03	01	02	03
Training File	ODS	3.13%	3.69%	3.84%	54.70%	53.15%	52.31%	53.32%	53.23%	53.18%
		2.17%	1.66%	2.32%	41.33%	41.99%	44.93%	1.50%	1.57%	1.48%
		5.46%	5.11%	4.95%	45.86%	44.75%	46.48%	25.51%	26.13%	26.13%
		4.95%	5.09%	5.05%	55.36%	53.04%	53.96%	68.80%	68.99%	69.11%
		5.86%	5.39%	5.65%	41.99%	42.10%	41.63%	64.01%	63.79%	63.96%
		5.78%	5.63%	6.16%	50.61%	52.93%	52.53%	16.91%	16.95%	16.99%
		4.40%	4.18%	4.48%	30.61%	30.94%	30.84%	1.99%	2.22%	2.24%
		5.05%	5.25%	4.99%	37.68%	39.56%	42.29%	1.68%	1.63%	1.68%
		5.86%	5.37%	5.51%	30.39%	30.94%	31.28%	2.60%	2.94%	2.96%
		28.51%	28.32%	28.42%	34.48%	35.14%	34.36%	67.20%	67.48%	67.60%
	CDC	7.71%	7.79%	7.81%	16.57%	21.10%	25.00%	3.08%	3.05%	3.07%
		10.33%	10.38%	10.40%	20.44%	18.01%	22.03%	2.89%	2.83%	2.69%
		8.34%	8.24%	8.22%	22.43%	21.66%	20.26%	92.77%	92.74%	92.67%
		14.08%	14.35%	14.29%	22.32%	21.88%	22.36%	4.06%	4.02%	4.21%
		14.59%	14.88%	14.84%	23.98%	23.09%	22.91%	17.91%	17.75%	17.73%
		8.83%	8.62%	8.58%	23.87%	23.43%	24.23%	1.40%	1.29%	1.22%
		6.59%	6.50%	6.56%	23.31%	23.09%	22.91%	1.92%	2.25%	2.41%
		14.00%	14.29%	14.33%	23.31%	22.87%	22.80%	3.64%	3.64%	3.87%
		12.11%	12.27%	12.45%	19.23%	19.23%	19.71%	1.96%	1.83%	1.82%
		9.17%	9.25%	9.20%	24.31%	24.42%	22.91%	2.24%	2.31%	2.53%
	NCDC	8.26%	8.16%	8.24%	33.04%	32.60%	32.60%	0.00%	0.00%	0.00%
		7.69%	7.63%	7.69%	31.05%	29.94%	29.30%	0.00%	0.00%	0.01%
		7.87%	7.85%	7.91%	29.72%	29.17%	29.30%	0.00%	0.00%	0.00%
		6.43%	6.42%	6.50%	28.07%	27.96%	27.64%	0.00%	0.00%	0.00%
		8.40%	8.32%	8.40%	35.25%	34.81%	33.70%	0.00%	0.00%	0.00%
		6.25%	6.24%	6.30%	27.18%	27.29%	27.09%	0.00%	0.00%	0.00%
		7.67%	7.57%	7.63%	38.12%	38.23%	38.22%	0.00%	0.00%	0.00%
		7.18%	7.13%	7.23%	49.94%	49.94%	50.00%	0.01%	0.01%	0.01%
		6.53%	6.52%	6.60%	40.55%	40.55%	40.64%	0.01%	0.01%	0.01%
		6.19%	6.10%	6.24%	40.66%	40.66%	40.75%	0.01%	0.01%	0.01%

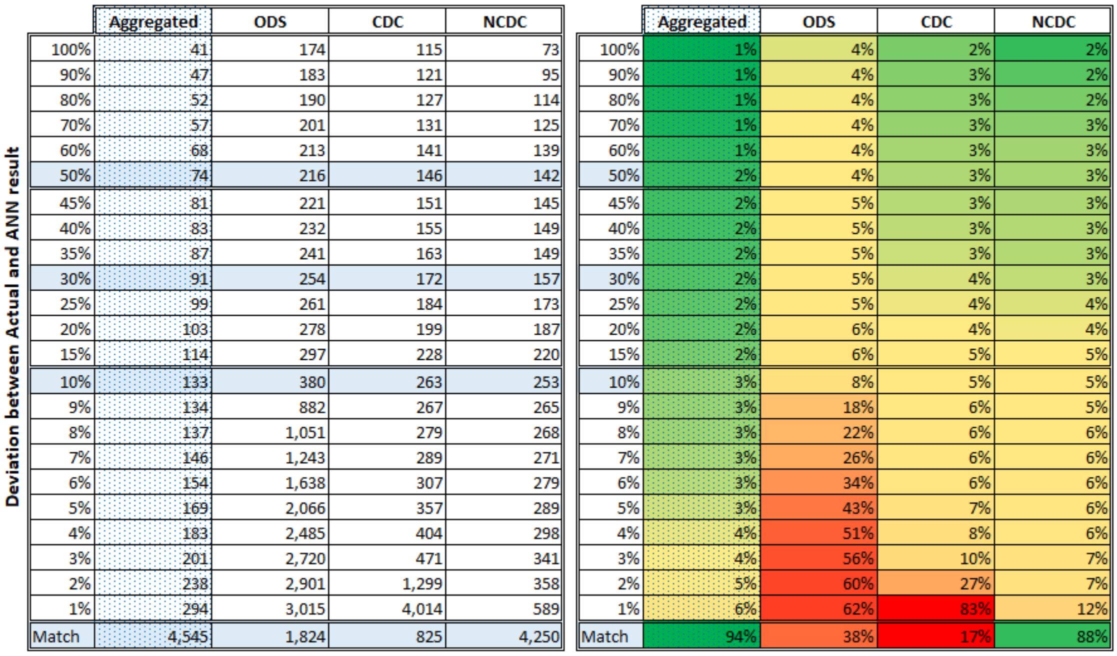


Results for 10% of File Content Read

		Testing File								
		ODS			CDC			NCDC		
		01	02	03	01	02	03	01	02	03
Training File	ODS	0.41%	0.85%	1.07%	47.62%	47.07%	47.03%	1.70%	1.77%	1.85%
		0.43%	0.12%	0.69%	37.24%	38.78%	41.30%	1.19%	1.21%	1.23%
		1.64%	1.51%	1.19%	36.02%	35.25%	37.11%	2.52%	2.63%	2.64%
		2.44%	2.28%	2.24%	48.29%	46.85%	47.25%	52.50%	52.64%	52.80%
		1.89%	1.64%	1.78%	34.48%	34.25%	35.35%	15.25%	15.25%	15.45%
		2.29%	2.24%	2.54%	41.44%	44.31%	43.39%	15.20%	15.27%	15.20%
		1.72%	1.51%	1.55%	25.41%	24.97%	25.44%	1.81%	2.03%	2.09%
		1.99%	2.20%	2.12%	31.16%	32.82%	35.24%	1.06%	1.16%	1.05%
		2.41%	2.26%	2.22%	21.44%	20.33%	20.48%	1.98%	2.10%	2.20%
		2.76%	25.72%	25.76%	27.18%	27.96%	27.09%	1.40%	1.58%	1.76%
	CDC	4.65%	4.72%	4.80%	7.40%	10.06%	10.68%	1.97%	2.09%	2.09%
		6.17%	6.18%	6.26%	7.96%	6.63%	8.26%	0.83%	0.79%	0.86%
		4.02%	3.90%	3.86%	7.96%	7.07%	6.50%	0.27%	0.31%	0.40%
		10.47%	10.72%	10.62%	8.95%	9.50%	8.70%	2.73%	2.69%	2.85%
		9.60%	9.79%	9.79%	8.40%	7.96%	8.04%	2.13%	1.94%	2.03%
		6.70%	6.54%	6.62%	7.51%	8.40%	7.60%	0.21%	0.17%	0.25%
		4.44%	4.40%	4.50%	8.51%	8.62%	8.70%	1.30%	1.49%	1.71%
		11.61%	11.85%	11.99%	8.29%	8.40%	8.70%	2.57%	2.55%	2.65%
		9.98%	10.23%	10.38%	8.62%	8.40%	8.15%	1.59%	1.27%	1.27%
		6.47%	5.95%	6.34%	8.73%	8.95%	9.47%	1.52%	1.77%	2.02%
	NCDC	6.59%	6.58%	6.64%	26.96%	26.19%	26.10%	0.00%	0.00%	0.00%
		6.03%	6.00%	6.06%	27.96%	26.74%	26.76%	0.00%	0.00%	0.00%
		6.07%	6.06%	6.14%	27.40%	26.74%	26.76%	0.00%	0.00%	0.00%
		5.68%	5.65%	5.75%	26.96%	26.63%	26.21%	0.00%	0.00%	0.00%
		6.74%	6.70%	6.82%	30.61%	30.39%	30.29%	0.00%	0.00%	0.00%
		5.54%	5.51%	5.57%	25.75%	25.86%	25.66%	0.00%	0.00%	0.00%
		5.56%	5.53%	5.59%	26.30%	26.30%	26.21%	0.00%	0.00%	0.00%
		5.91%	5.87%	5.93%	43.87%	43.76%	43.50%	0.01%	0.01%	0.01%
		6.25%	6.22%	6.30%	40.44%	40.44%	40.42%	0.01%	0.01%	0.01%
		5.74%	5.73%	5.81%	40.33%	40.33%	40.42%	0.01%	0.01%	0.01%

Figure 35. Comparative analysis of 1%-10% of error for 1%-3% of Lines Read

Having identified the optimal read percentage to be 2%, the experimental work resumed to further investigate how the impediments presented by the CDC dataset could affect the framework's accuracy and adaptability. In introducing *Variety* to the training of the ANN, two new sets were devised that consisted of files from all three datasets, where each set had an equal contribution of 33%. The first, which was used for training, was composed of 4,801 files, whilst the second, used for testing, had 4,839 files. Precautions were exercised so that there were no common files between the two sets. A new neural network was implemented using a 2% reading of lines. Similarly to prior tabulations and heat maps, Figure 36 presents the results from this investigation.



Deviation between Actual and ANN result

	Aggregated	ODS	CDC	NCDC
100%	41	174	115	73
90%	47	183	121	95
80%	52	190	127	114
70%	57	201	131	125
60%	68	213	141	139
50%	74	216	146	142
45%	81	221	151	145
40%	83	232	155	149
35%	87	241	163	149
30%	91	254	172	157
25%	99	261	184	173
20%	103	278	199	187
15%	114	297	228	220
10%	133	380	263	253
9%	134	882	267	265
8%	137	1,051	279	268
7%	146	1,243	289	271
6%	154	1,638	307	279
5%	169	2,066	357	289
4%	183	2,485	404	298
3%	201	2,720	471	341
2%	238	2,901	1,299	358
1%	294	3,015	4,014	589
Match	4,545	1,824	825	4,250

	Aggregated	ODS	CDC	NCDC
100%	1%	4%	2%	2%
90%	1%	4%	3%	2%
80%	1%	4%	3%	2%
70%	1%	4%	3%	3%
60%	1%	4%	3%	3%
50%	2%	4%	3%	3%
45%	2%	5%	3%	3%
40%	2%	5%	3%	3%
35%	2%	5%	3%	3%
30%	2%	5%	4%	3%
25%	2%	5%	4%	4%
20%	2%	6%	4%	4%
15%	2%	6%	5%	5%
10%	3%	8%	5%	5%
9%	3%	18%	6%	5%
8%	3%	22%	6%	6%
7%	3%	26%	6%	6%
6%	3%	34%	6%	6%
5%	3%	43%	7%	6%
4%	4%	51%	8%	6%
3%	4%	56%	10%	7%
2%	5%	60%	27%	7%
1%	6%	62%	83%	12%
Match	94%	38%	17%	88%

Figure 36. Aggregated Set Results Heat Map

It can be observed, in contrast to CDC, that there is a high match when it comes to the NCDC set, 4,250 out of 4,839, and a relative high match when it comes to ODS, 1,824 out of 4,839, where the result calculated from the ANN exactly matches the actual file delimiter. Confirming the prior experiment results, the network exhibits a high differentiation with the CDC dataset that is steeply phased out compared to the ODS-based network with lower differentiation but is phased out relatively slower. In contrast, the network based on NCDC and the Aggregated set exhibits high accuracy and constant smooth phasing (see Figure 37).

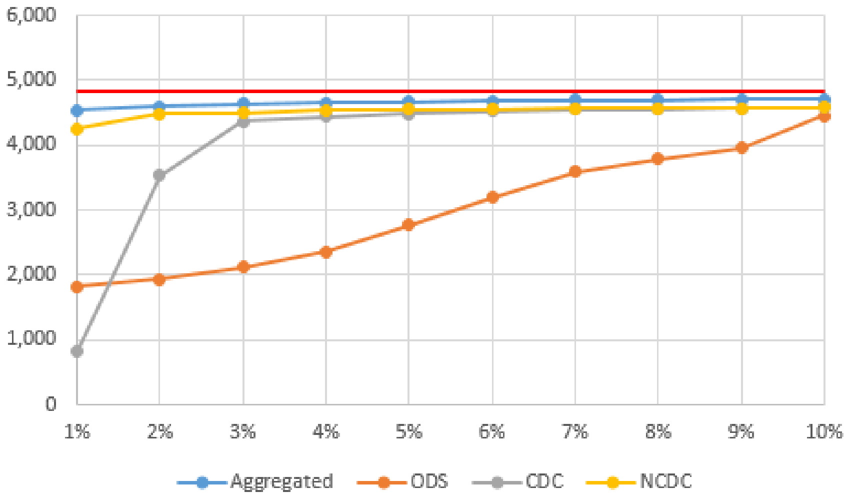


Figure 37. ANN functions Differentiation phasing out plot

Based on these new results from the aggregated sets, it was proven that the Automated Approach for Data Characterisation framework could accurately classify structures even when the data sets exhibit unique trends and characteristics. The results have shown that: a) by incorporating multiple sets, the framework is capable of absorbing *Variety* and transforming it into an asset by auto-calibrating the neural network that caters for the new multi-sourced set (e.g. 94% accuracy on the aggregated set) b) the neural network can efficiently and effectively classify other independent sets (e.g. classify NCDC based on ODS) or multiple sets (e.g. classify all three sets – aggregated set based on NCDC).

6.7. Conclusion

The results have shown that data scientists will be assisted in their data classification and identification in an automated manner using the proposed approach. Even though the experiment showed that human interaction is recommended as analysts will be required to understand and interpret the results, the work itself can be automated. The approach would remove much of the labour-intensive part of the data origination and data format tasks, leaving only critical decision-making to the data scientists. In this way, the initial implications of *Variety* in staging and ingesting the data can be minimised utilising an algorithmic approach. The approach also confirmed its viability in a Big Data environment by utilising 2% of the dataset data whilst producing high-quality results that can address the challenge at hand.

Automation and minimal human interaction will enable the capability of adopting Big Data and enhance the success ratio which is currently hindered by the *Varity* challenges. The organisation will be able to allocate the highly skilled resource involved into more productive tasks, thus lowering the TCO.

Chapter 7. THE "CONFIDENTIALITY PRESERVATION SYSTEM (BD-CPS)"

7.1. Introduction

Organisations need to ensure that personal information is not shared, but protecting everything in Big Data ecosystems can be challenging since it is almost impossible (Rawat et al., 2019). In an approach to adequately protect data, it is imperative to know the data characteristics and understand the aspects/dimensions of Big Data (Cuquet et al., 2017).

Having identified confidential information more accurately with the use of "Booster Metrics" (Chapter 5) and having automation of the ingestion of data with the use of dataset characterisation (Chapter 6), it is important to be able to manage the data confidentially throughout the data journey. Data are disseminated and used across the organisation, and in many cases, it is published to the public domain or to entities external to the organisation,. The following chapter proposes a framework for maximising Big Data's utility by mitigating the risk presented by preserving data confidentiality at a corporate level. The approach focuses upon the requirements of a rapid turnaround of processing requests for data dissemination, information coverage, automation, standardisation, flexibility, accountability, and traceability.

The system governing principles, algorithms, and workflows are showcased and explained (Section 7.2). Following, in Section 7.3, a PoC is presented showcasing a real-life implementation of the system, which is adds to the understanding and sets the stage for a hands-on presentation of the design principles and concepts. Providing the means to convey the system through a tangible prototype, the presentation methodology and evaluation by experts are showcased (Section 7.4).

7.2. The Big Data – Confidentiality Preservation System (BD-CPS)

This section, towards the novel contribution, proposes a framework that will enable the organisation to implement a comprehensive, standardised and usable compliance approach toward data confidentiality and data loss prevention. The primary objectives are:

- to suggest automation of the process with a software-driven, algorithmic rule-based system.
- to suggest alternatives for minimising the Data Loss risk for an organisation through standardisation
- to investigate the feasibility of transforming the repetitive classification work before distributing any data internally or externally to the organisation

The intention is to transform the process from a human labour-intensive, non-standardised, and error-prone effort into a corporate-wide, standardised, and automated process. Towards these objectives, the Big Data – Confidentiality Preservation System (BD-CPS) system was developed to provide a consistent and robust corporate data confidentiality rule-based framework. Based on the business analysis and corporate best practices regarding resistance to change, management, and auditability, the framework considered the functional requirements/characteristics, and business aims presented in Table 31. The system will store the definition of all the data elements of the corporate data dictionary in order to formulate a data confidentiality corporate taxonomy/ontology, utilising an automated ingestion and identification process.

Table 31. DB-CPS Functional Requirements/Characteristics & Business Aims

Characteristic	Objective
Time-to-Market	Fast turnaround, in a matter of hours, towards quick data dissemination for business use.
Accessibility	Enhanced accessibility with the use of technology (e.g. mobile implementation).
Information Completeness	Sufficient information completeness that will allow for one-stop decision making.
Automation	Use of algorithmic automation in attaining minimal user judgment in interpreting corporate policies.
Training	Provide an informational framework that will guide and educate users.
Standardisation	Standardisation through a corporate-wide repository that will reflect all acceptable policies along with any approved deviations.
Flexibility	Flexibility in adapting to the constantly changing corporate ecosystem and accommodating exceptions through a parameterised environment.
Accountability	Accountability and segregation of duty are required for compliance management systems where different roles and duties will be available.
Traceability	Traceability will be available since all actions via the information system will be audited and can always be back-traced.

The corporate data elements will be abstracted through the use of entities in order to be able to implement the corporate strategy on a less granular level. For example, if we consider the mobile number for abstraction, we can refer to the superset as "party mobile.” In this way, the mobile telephone number element is generically represented in multiple data sets (e.g. customer mobile, vendor mobile, mobile used to login to an application, One-Time-Password (OTP) delivery number). In some instances, the presence of multiple entities as a group might necessitate a different approach to risk; thus, the concept of a combined entity to encapsulate multiplicity is also introduced. A set of dimensions/classification attributes is required for each entity to define the entity quantitatively. This is required since entities must participate in numerical calculations when implementing algorithmic preventive controls and safeguards. Appropriate attributes were sought and identified in adequately describing an entity in the corporate environment concerning confidentiality.

Having identified the entities and the classification attributes, it is imperative for transparency to have multiple business bodies/departments review and confirm the ratings for each attribute of every entity. To that end, the system (illustrated in Figure 38) will provide a workflow mechanism referred to as the *Entities Classification Workflow* so that the rating of the attributes is inputted and approved. Similar mechanisms for approval will be employed in releasing a data set in the "*Data Release Workflow,*" where the system will algorithmically leverage the defined entity attributes. The calculated indexes will be utilised in visualising and enforcing the organisational DLP strategy in real-time.

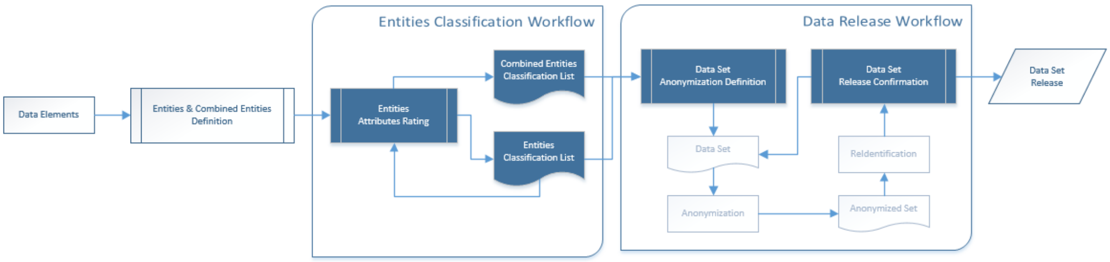

Data Elements
Entities & Combined Entities Definition
Entities Classification Workflow
Entities Attributes Rating
Combined Entities Classification List
Entities Classification List
Data Release Workflow
Data Set Anonymization Definition
Data Set Release Confirmation
Data Set
Anonymization
ReIdentification
Anonymized Set
Data Set Release

Figure 38. BD-CPS overview

The proposed system provides a configurable environment where the different subject matter experts from business, risk, security and other control disciplines, for example internal audit, will be able to depict and review the corporate confidentiality rules. The data elements under management for the organisation are mapped to entities which in turn are classified based on the following three classification attributes:

- Business Classification, which the data owner provides, indicates the operational use of the data.
- Regulatory Classification is the classification derived from industry, country or international regulatory requirements, acts, and laws.
- Anonymization Risk will provide classification relevant to anonymization and re-identification risks.

Once the corporate classification policy is enriched with the classification and depicted, the system will employ algorithmic rules to identify if the proposed anonymization or depersonalisation action taken against each element or a combination of elements is adequate. There will be little ambiguity in this way, and the process will be secure and robust.

The BD-CPS spans multiple business areas, introducing complexity in understanding the terms used. To make it easy to follow and understand the terms used and their context in the subsequent paragraphs, a data dictionary and some examples are provided in Table 32.

Table 32. Terms Dictionary

Term	Description
Data Set	It is a set of information (Fields / Business Entities) that will be considered one unit when handling the information, e.g. for distribution. Examples would include Credit Card Transactions, EOD Account Balances, Customer Master, Mobile Clickstream Etc.
Business Entity	Any information available in the organisation's Data Domain is reflected in business terms. Examples would include Party Name, Party Gender, Party Date of Birth, Party Nationality, Party ID Number, IBAN Account Number, Credit Card Number, Application Details, Balance Etc.
Data Set Fields	The actual instantiation of Business Entity as part of a specific Data Set. Examples would include Customer Name (Party Name), Merchant Name (Party Name), Account Balance (Balance), Form Name (Application Details) Etc.
Combined Business Entity	Is multiple Business Entities within a Data Set that will change the Classification Attributes once associated. For example, "Home Address" will not identify the individual since, e.g. in any address, there are multiple tenants, but once the "Date of Birth" is added, the re-identification probability is redefined since the chance of a person living in an area having the exact date of birth is relatively lower, thus the probability of identifying the person higher.
Anonymization Depersonalisation Actions	Is the action/algorithm to be applied on any piece of information in anonymising or depersonalise it as part of a Data Set to be distributed. Examples would include Pain Text, Mask, Encrypt, Hash, Micro-Aggregate, Delete / Static Value, Delete / Random Data.

The BD-CPS system is equipped with multiple features: a) algorithms for decision making; b) workflows and automated orchestrations for collaboration and structured processing; c) user guiding intricate UI to educate and facilitate the users; d) parameterisation and configuration screens and features to provide flexibility to the users; e) user role and access management. The system is adhering to the latest IT application development standards incorporating an n-tier architecture, which can be on-premises or cloud-enabled, along with the use of mobile technologies and distributed network access. An information loaded user interface is available so that the user is presented with an adequate amount of information for decision-making without requiring him/her to use other systems, which could be overwhelming and lead to confusion and information overload. BD-CPS can be delivered using mobile technologies. The system features a detailed auditing system to ensure the recording of invaluable information regarding forensics, traceability, and accountability.

The *Entities Classification Workflow*, where multiple levels of approvals can be accommodated, is based on the concept of segregation of duty, commonly referred to as "duality" (Manning, 2020). In implementing segregation of duty and preserving the integrity and accuracy of the data in the system, independent parties will have to review and inspect using the service world standard of the n-eyes principle (Lamberti, 2013). In this system, the 6-eyes principle instead of the 4-eyes principle is proposed due to its importance and criticality. Additionally, the workflow will enable the engagement of all related parties in a structured dialogue through the approval process. The roles suggested for duty segregation are: a) initiator/inputter; b) 1st level approver; and c) 2nd level approver. In this way, three independent bodies can be mapped to corporate control functions like Compliance, Risk, Security or Audit. Any role can be assigned to any user, and in this way, the system can be parameterised in any manner the organisation deems appropriate.

In each approval level of the workflow, there is a capability to further restrict the initial classifications to avoid multiple iterations. In addition, when it comes to assigning an element's anonymization or depersonalisation action, the user can request an exception to the predefined calculated rule, which will again have to be ratified using the approval process.

A systemic algorithmic implementation is suggested to minimise the corporate policies' user perception and possible bias. The system will automate the process and provide the user with the required guidelines for better understanding and engagement along with corporate-wide alignment, and standardisation BD-CPS is designed for a) aggregating underlined entity ratings for combined entities and b) auto-calculating the baseline for element anonymization action as a corporate standard and as a run-time feature for each dataset evaluation.

The first suggested algorithmic implementation is the calculation of the minimum levels for combined entities, which is proposed to facilitate the user when multiple interrelated attributes come into play. The system is automatically trying to provide suggestions concerning the possible classification level. Additionally, the system will compel the user to select a "safe" level by aggregating the undelaying attribute classifications. In this way, BD-CPS facilitates the organisation in ensuring that the users have a clear metric to follow and minimises the risk of misclassification. The combined entity's classification algorithms are presented in Table 33. They are calculated based on the maximum of all underlying entities for business and regulatory classification and with the advancement of one level for anonymization risk. For anonymization risk, the advancement of one level was implemented since the combination of multiple underlying elements will increase the exposure risk (Armando et al., 2015).

Table 33. Combined Entity Minimum Levels Algorithms

Classification Attribute	Algorithm	$\boldsymbol{e}$ → is the individual entities of the combined entity $\boldsymbol{l}$ → is the level of the metric for e
Business	$\max_{e_1 \dots e_n}(\{l_1, \ \dots, \ l_n\})$	
Regulatory	$\max_{e_1 \dots e_n}(\{l_1, \ \dots, \ l_n\})$	
Anonymization Risk	$\max_{e_1 \dots e_n}(\{l_1, \ \dots, \ l_n\}) + 1\ level$	

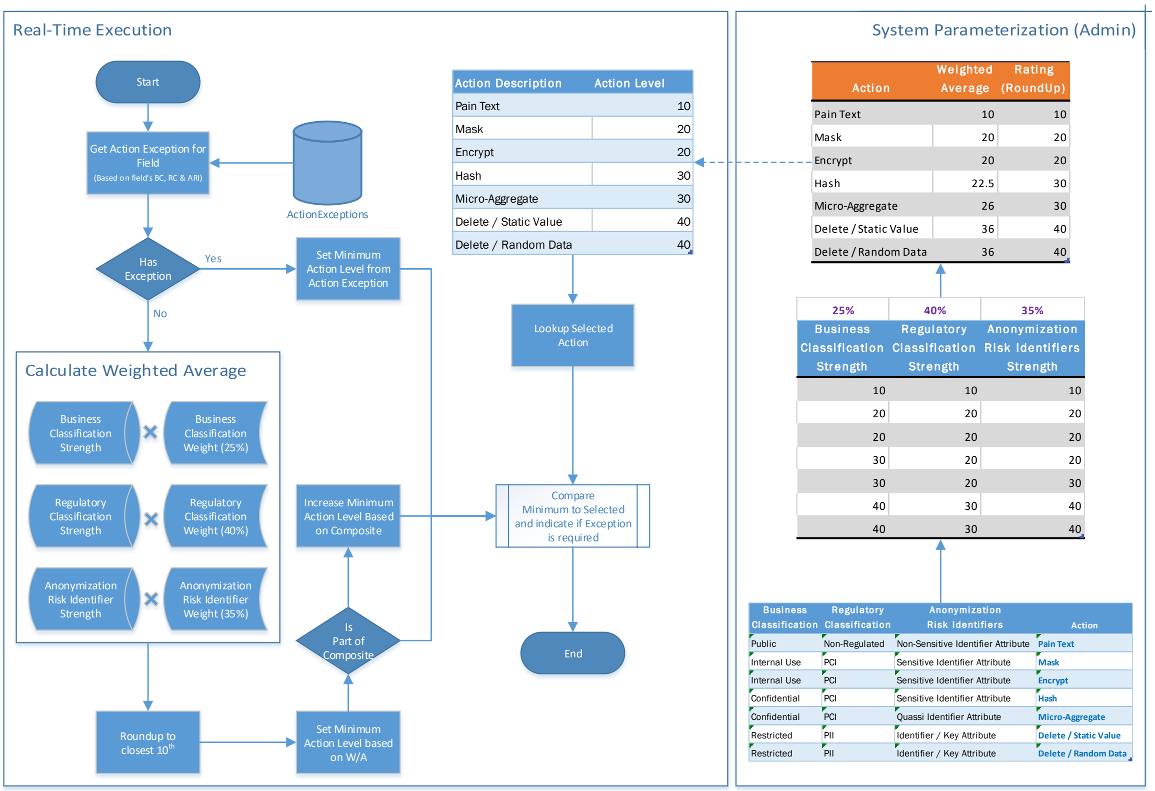
A working hypothesis example of the calculation is presented in Table 34. In the example, a scale of 10 to 40 in increments of ten is employed, and the combined entity is assumed to have three referenced business entities. The calculated levels are proposed to the user and are limiting in their nature regarding minimal compliance. However, the user can always propose a more restrictive profile if deemed fit.

Table 34. Combined Entity Algorithm example

	Business	Regulatory	Anonymization Risk
Referenced Entity 1	20	30	10
Referenced Entity 2	30	10	30
Referenced Entity 3	20	20	20
Combined Entity	30	30	40 (30+1 Level)

The second proposed algorithmic implementation is related to the automatic suggestion of the minimum required level for anonymization. Similar to combined entities, when the anonymization actions have to be selected, the system will suggest a minimal level restriction based on the configuration (see Figure 41).

These functionalities and automation are essential for the BD-CPS since they will a) enhance knowledge and awareness, b) increase productivity, and c) protect the organisation. It is argued that the BD-CPS will cultivate a cultural change towards understanding and embracing the corporate confidentiality policies through exposing users to the corporate policies on anonymization via the automated approach. The users will be able to immediately get feedback on organisationally approved practices and guidelines for the selected data element anonymization actions. This way, they will be empowered to make decisions without lengthy reviews. By having an independent non-human operated algorithmic rule enforcement engine, as exhibited in Figure 39, the organisation has simply to define the rules, and enforcement will be non-bias, thus mitigating the risk of data loss. The parameterisation as shown in the “System Parameterisation” figure section and enforcement as shown in the “Real-Time Execution” section are achieved in a two-step process using two independent calculations, which both utilise the concept of an "action strength.” Each action is associated and classified with respect to the three classification attributes (showcased bottom-up in the right parallelogram of Figure 39). Based on the underlying numeric equivalent (see Appendix IX. BD-CPS PoC Entities Data SamplesIX), a weighted average is calculated. The weights (shown in purple in Figure 39) for the weighted average calculation can be parameterised in the system (see Figure 41). The calculated values, after being rounded up so that we always minimise the risk of disclosure by applying a more restrictive action, will be used as a benchmark against the data scientist's proposed action. In real-time (showcased top-to-bottom in the left parallelogram of Figure 39), the system will calculate each entity's "action strength" by performing a similar weighted average calculation based on the entity's classification attributes. The calculated action strength will be compared to the strength assigned to the selected action. An indicator provides the user with feedback in real-time on compliance. In case of non-compliance, the list of compliant actions will be provided to the user to facilitate the process. Should the user decide to retain a non-compliant action, the exception process for the specific instantiation will have to be initiated, followed by all the required approvals based on the workflow.


Real-Time Execution
Start
Get Action Exception for Field
(Based on field's BC, RC & ARI)
ActionExceptions
Has Exception
Yes
No
Set Minimum Action Level from Action Exception
Calculate Weighted Average
Business Classification Strength
×
Business Classification Weight (25%)
Regulatory Classification Strength
×
Regulatory Classification Weight (40%)
Anonymization Risk Identifier Strength
×
Anonymization Risk Identifier Weight (35%)
Increase Minimum Action Level Based on Composite
Is Part of Composite
Roundup to closest 10th
Set Minimum Action Level based on W/A
Lookup Selected Action
Compare Minimum to Selected and indicate if Exception is required
End
System Parameterization (Admin)

Action Description	Action Level
Pain Text	10
Mask	20
Encrypt	20
Hash	30
Micro-Aggregate	30
Delete / Static Value	40
Delete / Random Data	40

Action	Weighted Average	Rating (RoundUp)
Pain Text	10	10
Mask	20	20
Encrypt	20	20
Hash	22.5	30
Micro-Aggregate	26	30
Delete / Static Value	36	40
Delete / Random Data	36	40

25%	40%	35%
Business Classification Strength	Regulatory Classification Strength	Anonymization Risk Identifiers Strength
10	10	10
20	20	20
20	20	20
30	20	20
30	20	30
40	30	40
40	30	40

Business Classification	Regulatory Classification	Anonymization Risk Identifiers	Action
Public	Non-Regulated	Non-Sensitive Identifier Attribute	Pain Text
Internal Use	PCI	Sensitive Identifier Attribute	Mask
Internal Use	PCI	Sensitive Identifier Attribute	Encrypt
Confidential	PCI	Sensitive Identifier Attribute	Hash
Confidential	PCI	Quassi Identifier Attribute	Micro-Aggregate
Restricted	PII	Identifier / Key Attribute	Delete / Static Value
Restricted	PII	Identifier / Key Attribute	Delete / Random Data

Figure 39. Action Strength Calculation

Rule enforcement is achieved in real-time alongside immediate visualisation of any user interaction. These are critical for the success of the BD-CPS since it is of utmost importance for the user to have relevant and in-time feedback and guidance. The interface should at least include the following set of functionalities: a) immediate interactive calculations, without going to the backend server for recalculation; b) capability to request or revoke exceptions; c) capability to view the approval level an exception is pending; d) enquiring for prior rejections of exceptions for any element; and e) easy and graphics-oriented journeys possibly using gamification.

The *Data Set Release Workflow* is composed of three stages. These stages are aligned to the segregation of duties principle, and the stages data undergoes before it can be released to the public. In order for the set to progress to the next stage, the aforementioned 6-eyes principle approval workflow is imposed.

- Release Definition is the stage the data analysts / scientist define the anonymization or depersonalisation action to be assigned against each field of the data set.
- Release Anonymization is the stage where the anonymization or depersonalisation actions are implemented. The implementation can be done with corporate proprietary systems or by subcontracting to an external, trusted entity. Since the data set is not yet anonymised or depersonalised, strict rules and contractual agreements should apply in the case of external entities' involvement. In this stage, the re-identification or de-anonymization strength is also calculated using specialised services. The respective percentage is inputted as a metric to proceed to the next level.
- Public Release Confirmation is when the approvals are to be given, post anonymization and/or depersonalisation so that the data set is distributed for its intended use.

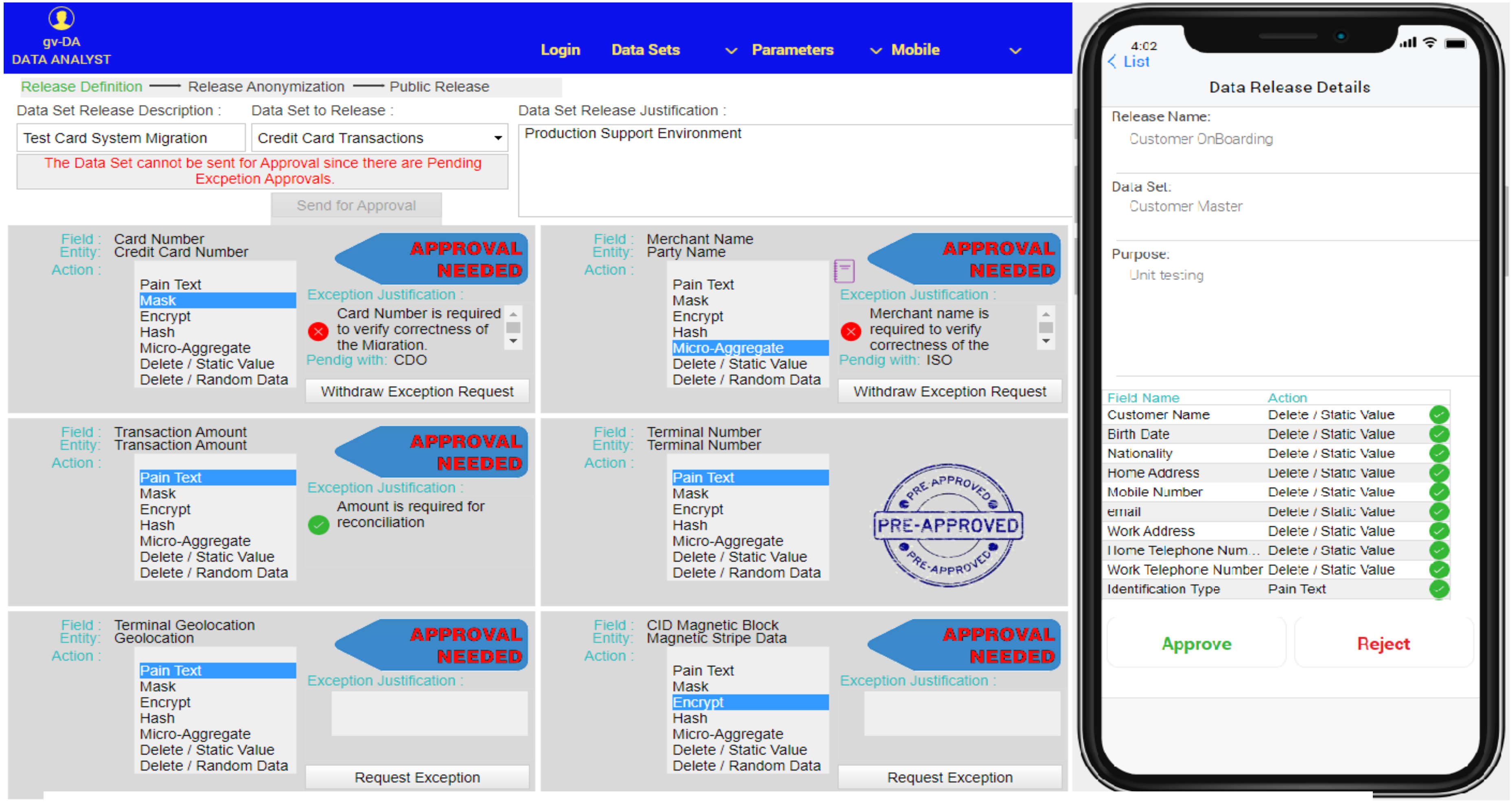
An exception to this 3-stage process is the case where an already approved data set is utilised. In this case, the existing release definition is used, but the subsequent two stages remain in place. The reasons for not bypassing all the stages and directly distributing the data set are primarily two a) to avoid the distribution of data sets for different usage due to negligence or unlawful intent, and b) to re-anonymise and re-depersonalise with the current and possibly more advanced techniques and algorithms.

The administrative and configuration user interfaces/forms should be accessed using privileged accounts that will only have the capability of parametrising the system. There will be no conflict of interest since the administrative users will have no access to the actual approval process or any application data-related information. These functionalities would cover the parameterisation of the weights between the classification attributes and the association of the anonymization/depersonalisation actions to the classification attributes. The user management should utilise a different role and a new set of administrative user accounts to further segregate the roles and responsibilities. If required by the organisation, 4-eyes principle can be implemented for all administrative and parameterisation functionalities by segregating the inputter and authoriser functions. In this way, the two-step process will further minimise the risk of erroneous or unlawful alterations regarding access and global BD-CPS parameterisations.

7.3. The Prototype

The purpose of the experiment and PoC using a working prototype is to evaluate the value a Data Confidentiality framework would bring to an organisation and its users. The organisation utilising the framework should be able to minimise the Data Loss risk and thus mitigate any regulatory fines and brand-related losses. In addition to that, the framework will provide better awareness and understanding by serving as a hands-on training instrument.

The fact that it is a PoC suggests that not all the functionality has been implemented, e.g. the user management module, mobile push notifications, or elaborate audit, but that fundamental elements of the proposed information system have been implemented. Web applications facilitating n-tier architectures have become the standard for application development (Hieatt & Mee, 2002; Shan & Hua, 2006), but the adoption of mobile technology is also very high, and COVID-19 has intensified the adoption (Shaw et al., 2022; Taylor et al., 2011). Based on the current trends and the increasing mobile adoption, the system comprises two interrelated applications, using different technologies for enhanced responsiveness and mobility (as illustrated in Figure 40). The web application targets users who use their PCs, and its User Interface (UI) is elaborate. The application provides capabilities for managing the configurations, data entry for classification, data request, and release process. The mobile application is developed primarily for mobility, so that approvals will not require a PC but rather a mobile device. In this way, the approvers can easily and quickly process any required requests.



Data Set Release Description : Test Card System Migration

Data Set to Release : Credit Card Transactions

Data Set Release Justification : Production Support Environment

The Data Set cannot be sent for Approval since there are Pending Excpetion Approvals.

Field : Card Number
Entity: Credit Card Number
Action : Pain Text / **Mask** / Encrypt / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data
Exception Justification : Card Number is required to verify correctness of the Migration.
Pendig with: CDO

Field : Merchant Name
Entity: Party Name
Action : Pain Text / Mask / Encrypt / Hash / **Micro-Aggregate** / Delete / Static Value / Delete / Random Data
Exception Justification : Merchant name is required to verify correctness of the
Pendig with: ISO

Field : Transaction Amount
Entity: Transaction Amount
Action : **Pain Text** / Mask / Encrypt / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data
Exception Justification : Amount is required for reconciliation

Field : Terminal Number
Entity: Terminal Number
Action : **Pain Text** / Mask / Encrypt / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data

Field : Terminal Geolocation
Entity: Geolocation
Action : **Pain Text** / Mask / Encrypt / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data
Exception Justification :

Field : CID Magnetic Block
Entity: Magnetic Stripe Data
Action : Pain Text / Mask / **Encrypt** / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data
Exception Justification :

Data Release Details

Release Name: Customer OnBoarding

Data Set: Customer Master

Purpose: Unit testing

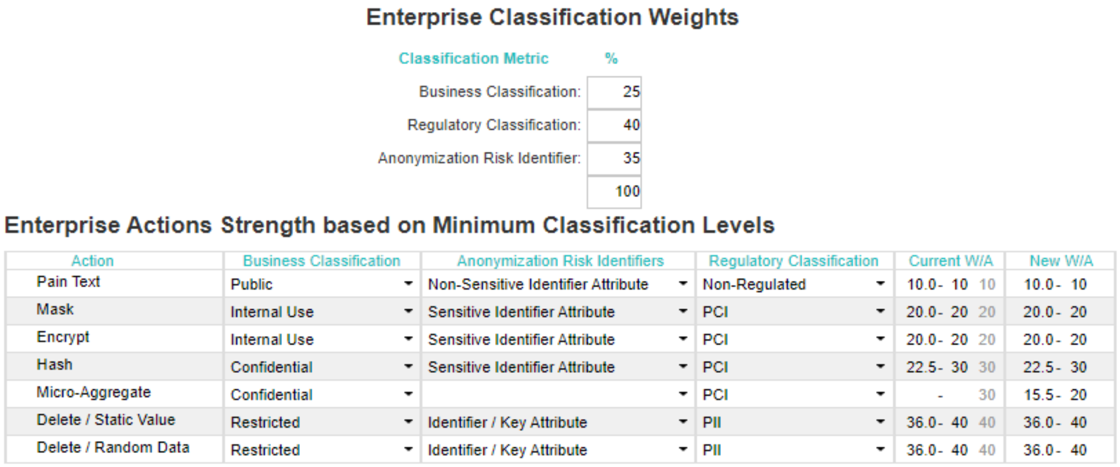
Field Name	Action	
Customer Name	Delete / Static Value	✓
Birth Date	Delete / Static Value	✓
Nationality	Delete / Static Value	✓
Home Address	Delete / Static Value	✓
Mobile Number	Delete / Static Value	✓
email	Delete / Static Value	✓
Work Address	Delete / Static Value	✓
Home Telephone Num…	Delete / Static Value	✓
Work Telephone Number	Delete / Static Value	✓
Identification Type	Pain Text	✓

Figure 40. Web & Mobile Application

A use case has been utilised to better understand and provide a walkthrough of the proposed implementation. The selected use case is conducted in a highly regulated environment (e.g. banking sector) to highlight the risk reduction.

Before users can utilise the system, certain parameters and configurations will have to be implemented. The parameterisation will be performed with the use of an administrative account, which will be in the custody of a control agency like the Information Security Office:

a) Post initial configuration of the system roles, which will be Data Analyst/Scientist (DA) - inputter, Chief Data Office (CDO) -first approval level and Information Security Office (ISO) – second approval level. The roles will be assigned to users operating the system with the associated privileges.
b) Parameterisation of the actions weighted average percentages, considering a highly regulated industry like banking or health, for the PoC and its evaluation, it was suggested to use a mix of 25% for business, 40% for regulatory and 35% for anonymization risk. The mix is suggested since the impact of any regulatory or anonymization risk in that order can have profound legal/compliance implications and penalties.
c) Define the minimum levels of association between the actions and the classification attributes. A visualisation from the PoC is provided in Figure 41.



Enterprise Classification Weights

Classification Metric	%
Business Classification:	25
Regulatory Classification:	40
Anonymization Risk Identifier:	35
	100

Enterprise Actions Strength based on Minimum Classification Levels

Action	Business Classification	Anonymization Risk Identifiers	Regulatory Classification	Current W/A	New W/A
Pain Text	Public	Non-Sensitive Identifier Attribute	Non-Regulated	10.0- 10 10	10.0- 10
Mask	Internal Use	Sensitive Identifier Attribute	PCI	20.0- 20 20	20.0- 20
Encrypt	Internal Use	Sensitive Identifier Attribute	PCI	20.0- 20 20	20.0- 20
Hash	Confidential	Sensitive Identifier Attribute	PCI	22.5- 30 30	22.5- 30
Micro-Aggregate	Confidential		PCI	- 30	15.5- 20
Delete / Static Value	Restricted	Identifier / Key Attribute	PII	36.0- 40 40	36.0- 40
Delete / Random Data	Restricted	Identifier / Key Attribute	PII	36.0- 40 40	36.0- 40

Figure 41. System Parameterisation

Based on *Data Origination* of the data ingestion journey (see Figure 12), the fields and entities of a dataset can be automatically identified. The respective identified information can become an automated feed to the proposed system. In addition, information available through the corporate Information Classification Policy (ICP) can be imported into the system. The next configuration level will have to be performed with the basic information available, configurations and fields/entities. If available, the classifications can be imported from the institute's already-defined Data Classification Policy (DCP); otherwise, the evaluation and classification of the entities concerning the aforementioned *Classification Attributes* will commence manually in the system.

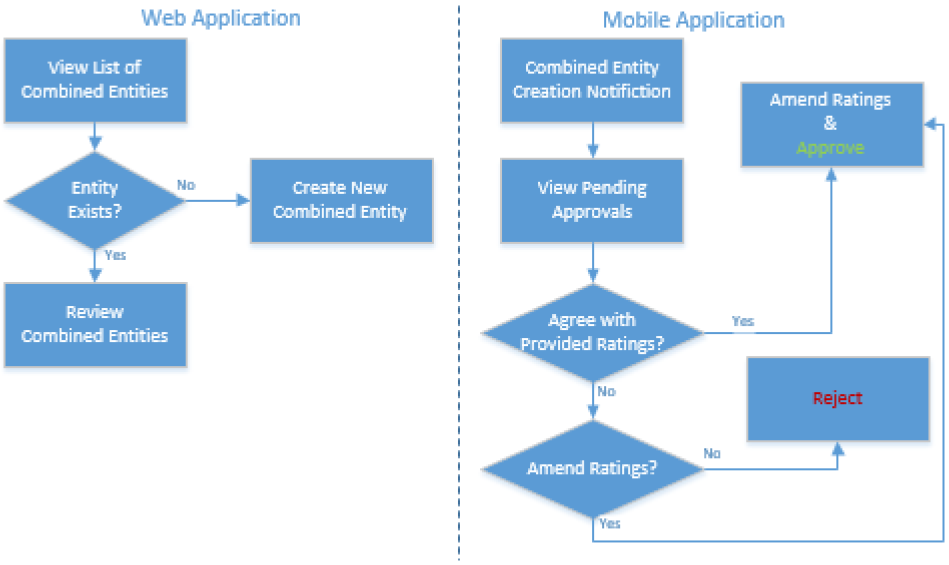
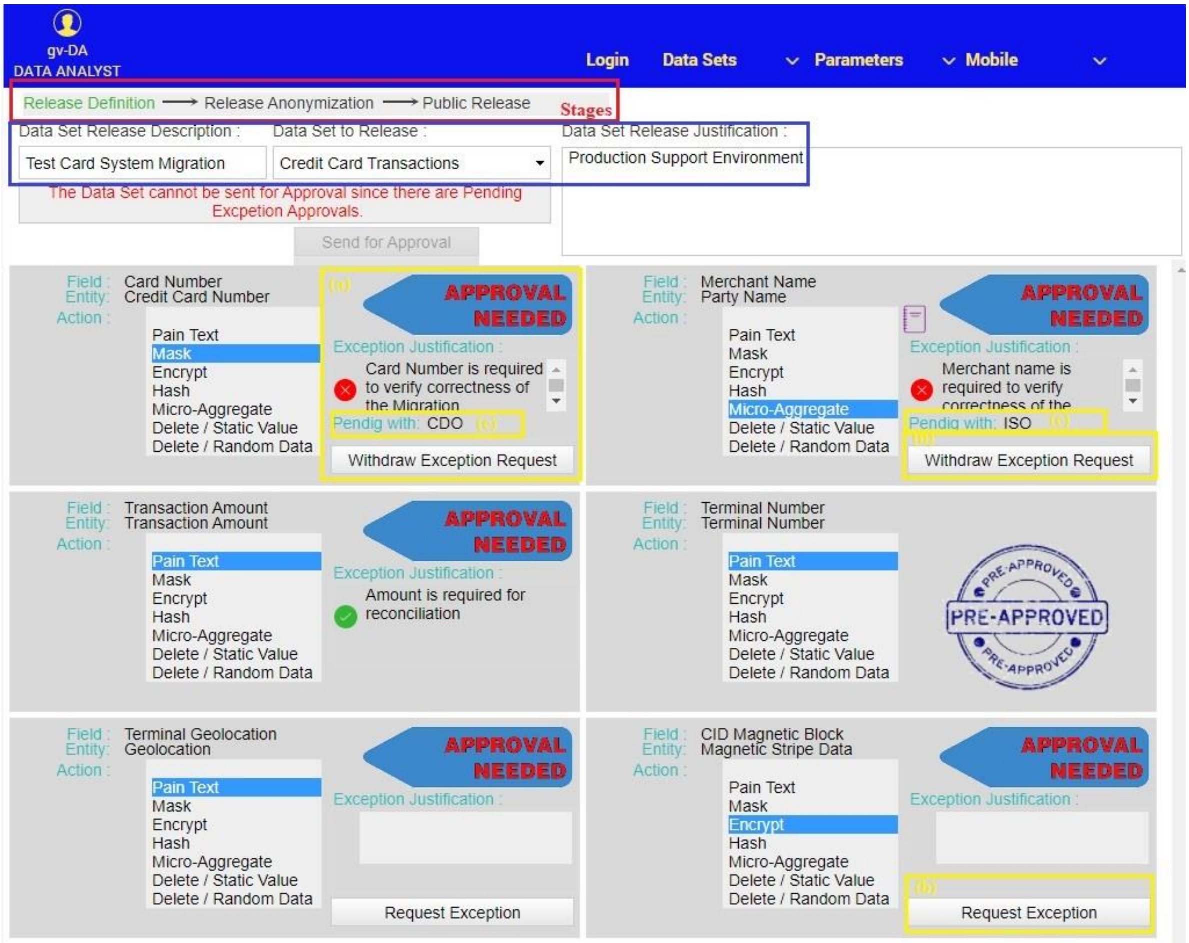


Figure 42. Entities Edit & Approval process

Using the web interface, see flow in Figure 42, the DA will search for any entity and define the Business, Regulatory and Anonymization Risk. Once the levels have been identified, an approval request will be sent to the CDO-1st level approval- on the mobile application using push notifications. The CDO will be using the available deep-link to go directly to the required entity and view the request. If it is deemed necessary, the approver can edit the classification, only increasing the level and thus making the data policy more restrictive. Once reviewed and confirmed, the approver (CDO) will record the approval. If the evaluator is not in agreement, there is a rejection option where the evaluator can also record the comments and the rejection justification. The same approval process will be automatically triggered for the next level of approval, ISO-2nd level approval. The system is fully parameterised in implementing the rule-based corporate-wide evaluation of datasets for internal or external dispatching.



Data Set Release Description : Test Card System Migration

Data Set to Release : Credit Card Transactions

Data Set Release Justification : Production Support Environment

The Data Set cannot be sent for Approval since there are Pending Excpetion Approvals.

Field : Card Number
Entity: Credit Card Number
Action : Pain Text / **Mask** / Encrypt / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data
Exception Justification : ✗ Card Number is required to verify correctness of the Migration
Pendig with: CDO

Field : Merchant Name
Entity: Party Name
Action : Pain Text / Mask / Encrypt / Hash / **Micro-Aggregate** / Delete / Static Value / Delete / Random Data
Exception Justification : ✗ Merchant name is required to verify correctness of the
Pendig with: ISO

Field : Transaction Amount
Entity: Transaction Amount
Action : **Pain Text** / Mask / Encrypt / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data
Exception Justification : ✓ Amount is required for reconciliation

Field : Terminal Number
Entity: Terminal Number
Action : **Pain Text** / Mask / Encrypt / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data

Field : Terminal Geolocation
Entity: Geolocation
Action : **Pain Text** / Mask / Encrypt / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data
Exception Justification :

Field : CID Magnetic Block
Entity: Magnetic Stripe Data
Action : Pain Text / Mask / **Encrypt** / Hash / Micro-Aggregate / Delete / Static Value / Delete / Random Data
Exception Justification :

Figure 43. Data Set Elements Actions Selection UI

A credit card migration project is utilised to showcase the data dissemination process and prototyped UI, exhibited in Figure 43. In this use case (highlighted in the blue parallelogram), the credit card transaction will have to be moved from the production environment to a lower environment (development or test environment) to facilitate and verify the migration process's correctness. In authorising the data sharing, the approval level is implemented in three stages and visualised on the form in the red parallelogram. For all the stages of approvals, the mobile application is used, whilst for editing the dataset confidentiality attributes, the web application is used. The web application was selected since many calculations are to be implemented, and a substantial amount of information is to be exhibited. A mobile application interface would be complex and fragmented, leading to user confusion, omissions, or errors. The form, presented in Figure 43, showcases (highlighted in yellow parallelograms with labelling) the following capabilities a) immediate interactive calculations, without going to the backend server for recalculation b) capability to request or withdraw an exception c) capability to view at which approval stage the exception is pending, and d) enquiring for prior exception approvals/rejections for the element.

The stage is available on the top of the screen, red parallelogram in Figure 43 so that the DA can move to the next level after implementing each level's requirements. In the 1st stage, the DA will have to define all the necessary actions to be taken per field. The system on the fly will inform the user of a) the required adjustment, b) the required exception, c) recommendations or d) confirmation to proceed. When all fields are associated with a pre-approved action or the required approvals for an exception are in place, the DA will request the initiation of the approval to move to the next stage. The approval process is similar to the approval process mentioned above for the (Combined) Business Entities. However, the approvers cannot edit the proposed action levels in this case. The reason is that the approvers have already accepted any deviations from the corporate policy by approving the individual exceptions and now will have to approve the entire set. Suppose the DA is trying to distribute an already approved release of a dataset for the same or different purposes. In that case, this step is obsolete, and the process will automatically move to the next level.

The 2nd stage is related to the implementation of the suggested action. The set will be anonymised/depersonalised based on the actions set. This process is external to the system and can be undertaken using proprietary corporate mechanisms or using a contractually bound partner since the dataset contains all the privileged information at this stage. Folowing the anonymization/depersonalisation process; the set will have to be submitted to an external engine, which can be one of the available public service providers, to evaluate the anonymised/depersonalised data and calculate the re-identification factor/percentage of the set. The DA will have to input the respective factor as evidence and initiate the approval process so that the control function can validate and confirm that the required level of anonymization/depersonalisation is achieved. The set has reached its final (3rd) stage, where it is ready to be distributed.

What has been exhibited so far, is that the system will provide the configuration reusability so that human intervention is limited. The entities will be parameterised once and reviewed at regular intervals but will not be required to be evaluated for each data set release. The users will get accustomed to the security concepts with the structured interactions, exceptions, and approvals, as well as the control functions, thus increasing their data loss prevention awareness and education. In addition to that, the corporation can record all interactions and have full accountability and responsibility for any data release, along with the assurance that the system will preserve the minimal levels already defined in the system.

7.3.1. Data Design Elements

Regarding the implemented PoC, several data events have been defined to represent the information required to be managed by the system. An Entity Relationship Diagram (ERD) is presented in Figure 44, whilst the entity descriptions are as follows. Sample data for all entities are available in Appendix IX. BD-CPS PoC Entities Data Samples:

- Data Sets: The description of the data sets that are managed by the system.
- Data Set Fields: The fields/data elements contained in each set.

- Data Set Releases: The instantiation of a release request is managed through the system's approval process.
- Data Sets Release Fields: These are the fields of the data set to be released. It will contain the fields associated with the data set, as per the data set fields entity, with additional information regarding the specific distribution, e.g. the actions taken for each field in the specific release.
- Classification Actions: These are the possible actions an analyst can take to preserve the data confidentiality of a data element/field.
- Actions: These are the actual actions that an analyst took to preserve the data confidentiality of a data element/field.
- Exception: refers to the exception implemented in a specific element of a distribution set.
- Business Entities: The actual instantiation of a Business Entity as part of a specific data set. Examples would include Customer Name (Party Name), Merchant Name (Party Name), Account Balance (Balance), Form Name (Application Details) etc.
- Combined Business Entities: Are multiple Business Entities within a data set that will change the classification attributes once associated. For example, "Home Address" will not identify the individual since, in any address, there are multiple tenants, but once the "Date of Birth" is added, the re-identification probability is redefined since the chance of a person living in an area having the exact date of birth is relatively lower, thus the probability of identifying the person higher.
- Combined Business Entities – Entities: This is an intermediate entity to facilitate the many-to-many relationship between Combined Business Entities and Business Entities.

- Business Classifications: Contain the definition of the “business classification,” as shown in Section 7.2, along with their strength.
- Regulatory Classifications: Contain the definition of the “Regulatory Classification,” as shown in Section 7.2, along with their strength
- Anonymization Risk Identifiers: Contain the definition of the “Anonymization Risk Identifier,” as shown in Section 7.2, along with their strength.
- Users: The information related to the system’s users.
- User Roles: The possible roles that a user can have.

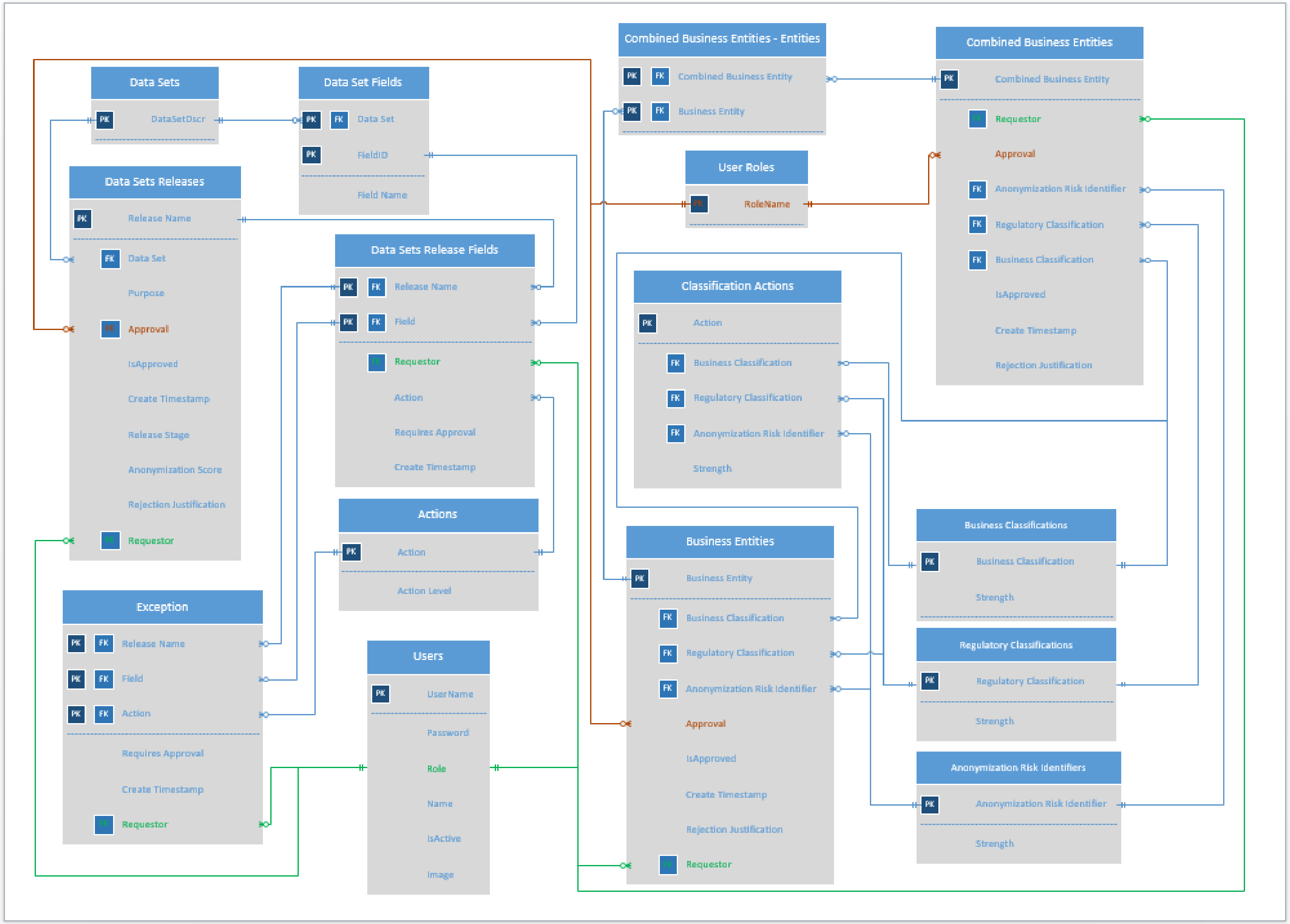


Figure 44. BD-CPS ERD

7.3.2. System Screen Flows

The system contains several flows regarding its capabilities, and the same is depicted in the interface and approval diagrams.

The web interface, showcased in a diagrammatic form in Figure 45, will require the user to login. The interface focuses on data entry and the system's data manipulation using complex IU/UX. After login, the user can perform operations related to the parameterisation of the system, including creating new, viewing or editing a) Business Entities, b) Combined Entities, c) Data Sets, or d) Data Set Requests. Additionally, the user can view the user information from any screen.

The mobile interface, showcased in a diagrammatic form in Figure 46, primarily focuses on the approval process. The user will receive a push notification, and by interacting with it, will be deep-linked into the required approval process. If the user is not authenticated, the system will prompt with a login screen. The user using the mobile interface can perform limited alterations to the suggested ratings of any entity and can approve or rejectany request. The approval workflows are presented in detail.

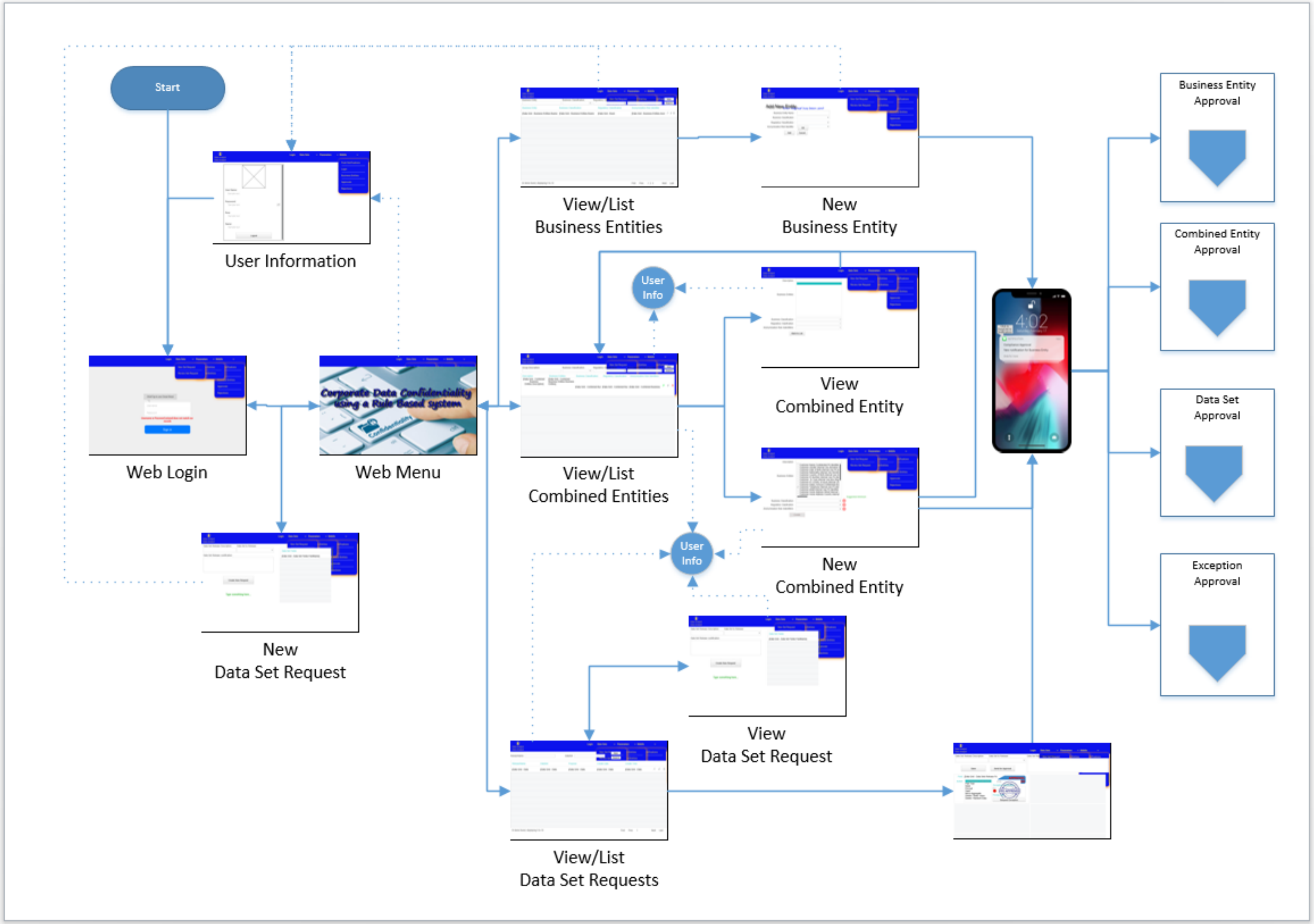


Figure 45. BD-CPS Web Interface Screens Flow

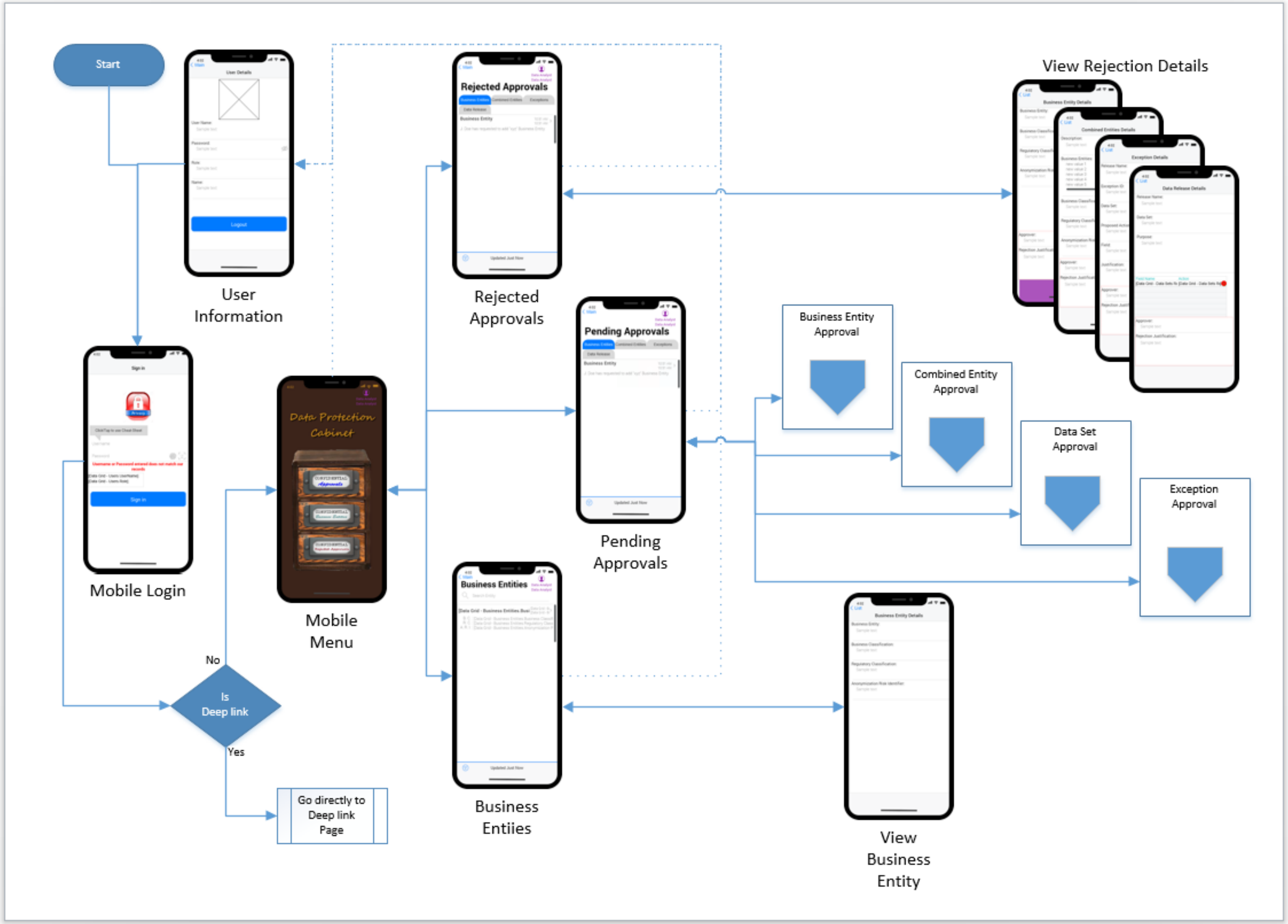


Figure 46. BD-CPS Mobile Interface Screens Flow

The approval workflow is similar to all processes except the business entity approval. The evaluator, or approver, can directly approve on the approval list by swiping left or by proceeding to the detailed approval screen for any request, as shown in Appendix X. BD-CPS PoC Additional Approval Journeys. The business entity approval journey, shown in Figure 47, provides the additional capability of editing the entity information before the approval. This functionality was introduced to reduce the approval cycle timelines by allowing the approver to make the necessary alterations and approvals instead of sending it back to the requestor for editing and resubmitting.

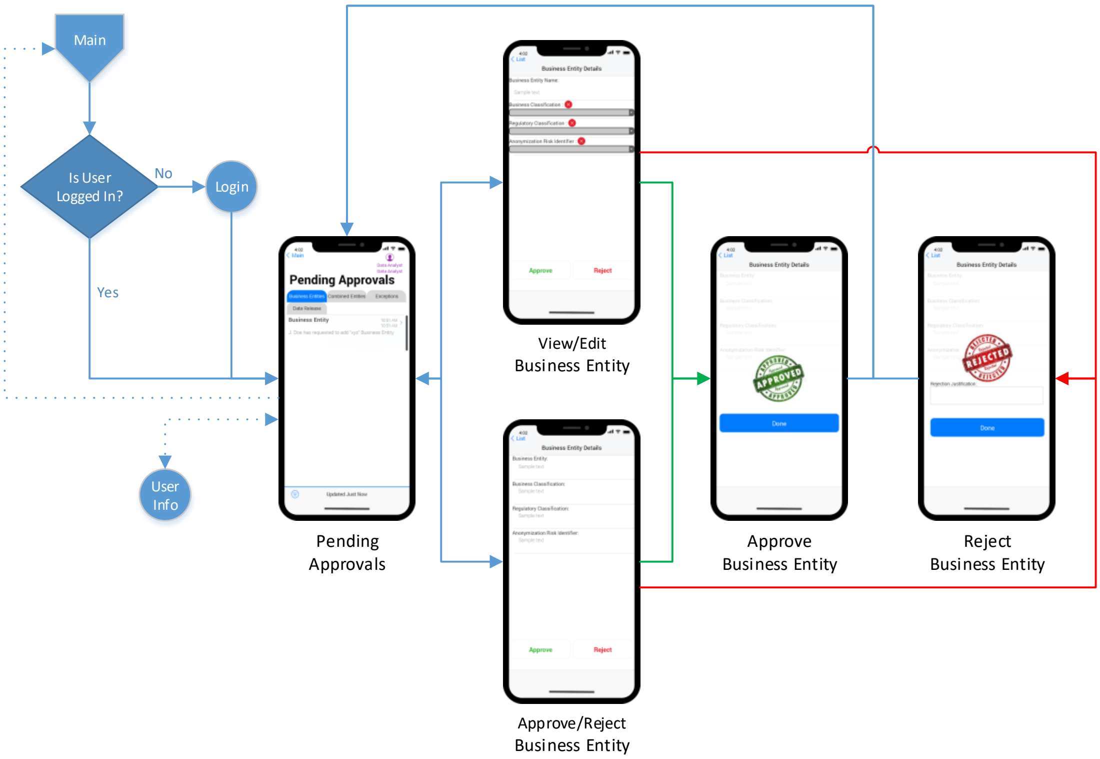


Figure 47. Business Entity Approval screens flow

An additional automation is also provided in the data set approval. In this flow, the system will provide an indicator and prevent any approval if there are pending approvals for exceptions related to the specific data set distribution.

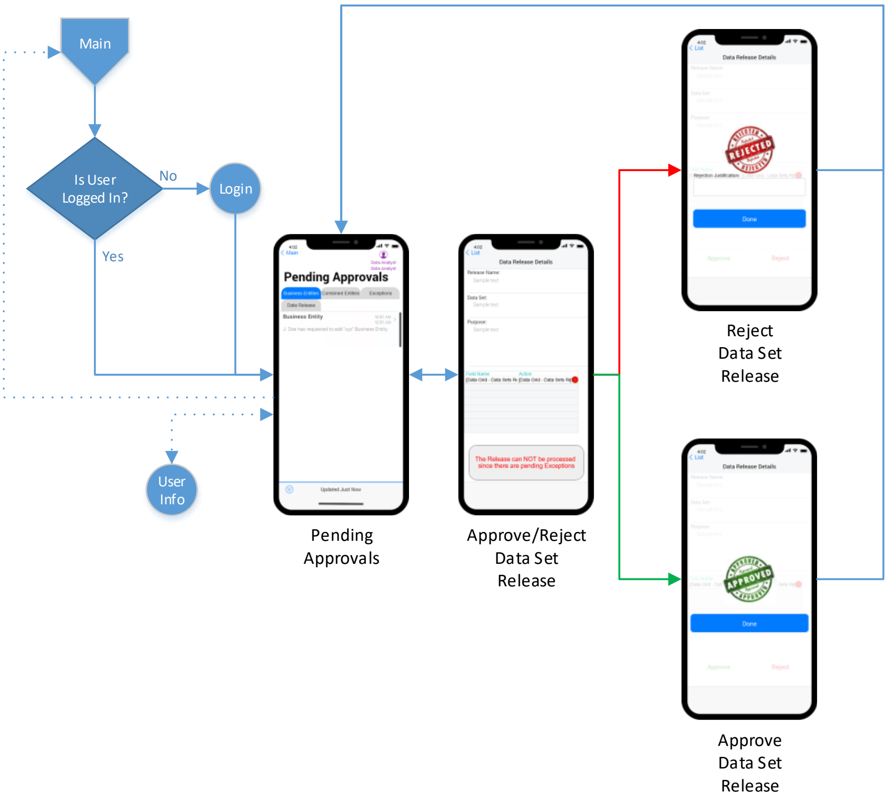


Figure 48. Data set approval screens flow

7.4. Evaluation

7.4.1. Methodology

In validating any proposal or suggested approach, there must be an evaluation process to measure the effect and possible impact. The proposal has to be gauged regarding innovation and value along with suggestions towards advancements and realignment. There can be a quantitative or qualitative measurement approach towards any such engagement. The differences between the approaches are many. An important one is the participants' number where the qualitative will study fewer people in more depth. For this reason, the qualitative approach was selected in order to be able to get a rich response from experts with in-depth retrospect and outlook on the proposed concepts designed (Hyett et al., 2014; Majid et al., 2017; Swanborn, 2010). In ensuring a holistic review, experts from both academia and industry were invited. The diverse origin and different accumulated experiences of both academics and senior executives were important in understanding the proposal's value regarding theoretical validity and practical applicability. In addition to the different knowledge backgrounds and experiences, the geolocation diversity and cultural background were considered by involving experts working in multiple countries, namely the UK, India, Kuwait, and Greece.

In order to have a constructive discussion, a PoC utilising a working prototype was implemented, available on https://rb.gy/ekm5j1. Using the prototype, several videos were created in showcasing the concepts. A structured approach in delivering the concept was designed using a presentation. The delivered presentation exhibited an overview of the information presented in Sections 7.2 and 7.3, along with video clips showcasing the process of utilising the publicly available functional prototype (https://rb.gy/ekm5j1). Finally, the attendees were involved in a discussion, driven by a predefined set of questions available in Table 36. The audience mix, see Table 35, was decidedly diverse in order to cover both academic and business perspectives. The expert commentators' number was initially driven by the 10±2 rule and finalised based on the exhibited saturation of the responses during the progress of the interviews (Creswell, 2013; Guest et al., 2006; Hwang & Salvendy, 2010).

Table 35. Interviewees Mix

Expert Commentator Role	**Number of Interviewees**	**Average Years of Experience**	**Total Years of Experience**
Academic (Data Related)	3	18	55
Academic (Security Related)	2	24	47
Chief Data Office / Data Protection Office	3	29	87
Information Security Office	2	24	47
Information Technology (Data Related)	2	29	58
Totals	12		294

Ethical approval was sought and granted from the Faculty of Science & Engineering Research Ethics and Integrity Committee of the University of Plymouth (Ref: 2862). The invites were delivered through emails along with the information form (available in Appendix VII. Information Form) outlining the initiative's details and the consent form (available in Appendix VIII. Consent Form). The interviews were performed using teleconferencing facilities due to different geolocations and Covid-19. In evaluating the framework and system at the end of the 45 minute presentation, a set of questions were discussed.

In acquiring insight on the occurrence and mitigation of particular challenges pertaining to Big Data and Data Confidentiality in particular, participants were asked a set of rating questions in addition to a series of open-ended questions, see Table 36. The questionnaire was formulated to facilitate the discussion and acquire insight from the experts. During the discussion, the focus of the insight was to explain the concepts, acquire feedback on the subject from the experts, and drive experts to provide further suggestions for enhancements. A 5-point scale similar to psychometric Likert and Five-Star quality rating grade were selected since it is easy and many people have prior experience with it. The five level gauge instead of three was used, although it has minimal statistical impact or value since we are interested in individual behaviour (Dawes, 2008; Friedman & Amoo, 1999; Malhotra & Peterson, 2006).

Table 36. Rating & Open-Ended Questions

Question	Type
Data under management are too many (Volume)	Rating
Data under management are changing very quickly (Velocity)	Rating
Too many "flavours" of Data under management, input files, output files, reports, emails, DBs, Legacy [M/F] etc. (Variety)	Rating
There is a classification of Data under management in terms of Business & Regulatory context	Rating
A robust anonymization and depersonalisation strategy is defined and implemented	Rating
There is sufficient understanding of which Data is used for each business function under which circumstances	Rating
The toolset in place for Data Confidentiality (identify, approve etc.) has limited capabilities or insufficient automation	Rating
Do you believe that in the era of Big Data there are challenges in managing Data Confidentiality? If so, have you faced such challenges?	Open-Ended
Do you believe that the introduction of a Mobile App for approvals will enhance responsiveness? Will it be sufficient/practical to use the Mobile App or the approvers will need to access the Web App in getting more details and context before approving?	Open-Ended
Would you alter the weights allocated for the calculation of the minimum actions? If so why?	Open-Ended
Do you think that "Business Classification", "Regulatory Classification" and "Anonymization Risk Identifiers" can be used to adequately characterise a data element in terms of confidentiality? Would you suggest and additional Classification?	Open-Ended
Are the anonymization/depersonalisation "Actions" identified sufficient? Would you suggest any other "Action" to be added?	Open-Ended
Would you suggest any additional Automation(s)? In which area (e.g. User Interface, Calculations, Approval Workflows)? What would that Automation(s) be?	Open-Ended
Would you suggest such a methodology in addressing Data Confidentiality issues? If so is there an immediate benefit you can think of?	Open-Ended

The results are presented in narratives as per the interviewees' responses. Direct quotations will be used to present the subject matter experts' views to comprehend the response content and the tone and emphatic nature of the responses. The ranking provided by the commentators will be presented in heat maps in order to visually highlight the concentration of the responses. The actual colour and intensity are used to reveal the progression/intensity. The selected visualisation uses three colours: green denotes high concentration, blue is the middle ground with moderate concentration, and ivory is the unused ratings.

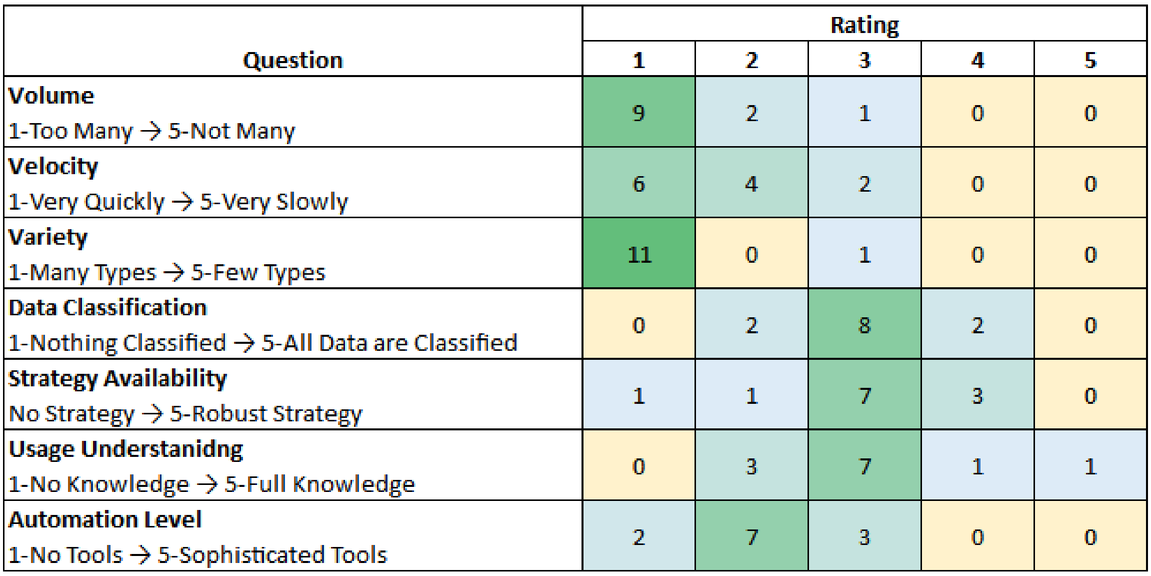
In implementing and evaluating the PoC, a desktop PC with Inter® Core™ i7-6700 CPU @ 3.40 GHz and memory of 16GB was utilised. The system and application software used would include Windows 10 (64-bit), Ms Office (2013, 2016), JustInMind prototyping platform.

7.4.2. Results

In putting the theory to the test, a total of twelve experts from several fields across academia and business were invited. By being senior executives and academics, the experts exhibit on average 25 years of experience in their fields while bringing a total experience of approximately three centuries to the case study. They were presented with the framework and asked to comment on a series of rating and open-ended questions as described in the Methodology section (Section 7.4.1).

All commentators confirmed that managing data confidentiality is an existing day-to-day challenge. The intensity and awareness were exhibited with the use of strong words like "*obviously,*" "*definitely,*" "*indeed,*" and "*of course*" when the commentators were describing the data confidentiality challenge as the "*most important aspect of information management*" which is "*pretty much impossible to guarantee*.” Everyone had faced the issue, and different perspectives were given based on the commentator's role and experience. The academics were more on the receiving end, where the data shared was inconsistent or depleted due to the anonymization, thus often reducing their value. On the other hand, business originating commentators were on the sharing side where the confirmed concerns stemmed from the regulatory and security perspectives.

In acquiring further context, the commentators were asked pointed questions to which they had to provide a rating, see Table 36. The responses are tabulated in the heat map presented in Figure 49. We can see the topic of the related question and the number of participants that gave the respective rating.



Question	Rating				
	1	2	3	4	5
Volume 1-Too Many → 5-Not Many	9	2	1	0	0
Velocity 1-Very Quickly → 5-Very Slowly	6	4	2	0	0
Variety 1-Many Types → 5-Few Types	11	0	1	0	0
Data Classification 1-Nothing Classified → 5-All Data are Classified	0	2	8	2	0
Strategy Availability No Strategy → 5-Robust Strategy	1	1	7	3	0
Usage Understanidng 1-No Knowledge → 5-Full Knowledge	0	3	7	1	1
Automation Level 1-No Tools → 5-Sophisticated Tools	2	7	3	0	0

Figure 49. Rating heat map

From the colour coding and the rating distribution, it is evident that all commentators acknowledge the challenges posed by the Big Data basic 3V dimensions. *Variety* is confirmed to pose the highest challenge by exhibiting the highest ranking amongst the three, with the highest concentration on the slightest grade. The following three rankings are about data classification, strategy availability, and usage and seem to concentrate toward the middle of the scale. This is of vital importance to the data confidentiality framework. In essence, these three dimensions are the prerequisites in identifying, defining and classifying the data elements along with the proposed classification attributes, namely business, regulatory, and anonymization risk. Before going into the details of each metric, it is also important to mention that since the ratings are in the middle, there is obvious room for improvement. That is the reason why commentators, while discussing these points, confirmed that the framework would also serve as a training and awareness tool. Having a good understanding of the business and regulatory classification will be the basis for the framework where the entities/data elements will be easily and quickly classified. Having such information readily available and documented will ensure consensus among all parties, and the process will be smooth. In addition to that, if there is a high-level strategy for anonymization and depersonalisation, it would mean that the involved parties have prior experience and understanding and are seasoned enough to take the next step in automating the process. It was also confirmed that the data usage is well known, enhancing the classification by attributing to the anonymization risk. When the usage is known, it will be easier for each party to associate the risks and identify the required policy to mitigate them. Last but not least - if not most important - is the existence of tools and automation. Most of the participants have confirmed that existing tools are in their early stages and lack sophistication and automation. This fact is crucial in confirming the novelty of the proposed framework, suggesting a rule-based automated and algorithmically driven system.

The system, and to a certain extent the framework, in order to be adopted, will have to be easy to use and provide value to the users and the organisation. In facilitating the user experience and the web application, a mobile application was introduced. The experts welcomed the introduction of the mobile application and confirmed that it would enhance responsiveness. Statements like "*will definitely help,*" "*it will enhance the responsiveness 100%,*" and "*I would give priority to the mobile app*" indicated the enthusiasm and confidence of the experts towards the use of the proposed mobile application. It was also pointed out that the value of the mobile application will increase throughout the time when the data elements will be stable, and the distribution of the sets will mainly focus on approvals rather than the definition of the anonymization or depersonalisation actions.

The three classification attributes of the data confidentiality framework were confirmed to be sufficient and well equipped to provide a holistic understanding and classification. The commentators stating, "*I really like these suggestions because they are clear*" or "*100% sufficient,*" confirmed that using these attributes would effectively and efficiently characterise the data elements in terms of confidentiality. Regarding the percentages allocated for each attribute towards the weighted average, the experts affirmed the research suggestion, in which *Regulatory* is the highest, followed by *Risk Anonymization* and *Business*. The wording indicates the consent: "*I would stick to the ones you have put together*." Nevertheless, all the commentators pointed out that the respective percentages are organisation and sector-specific and applauded, "*as long as it is an option I decide*" the availability of a system capability to parameterise them through the administrative screens. In addition to that, they affirmed the concept of an upward (ceiling function) roundup in increasing controls and reducing the risk of the assigned Action Strengths. For the anonymization or depersonalisation actions, comments like "*I think is a good set,*" "*I have nothing to add,*" or "*I do not think we need to add anything more*" were indicative of the experts' acknowledgement and confirmed as being a representative set which would cover most, if not all, the Data Confidentiality requirements.

Towards the end of the discussion, the experts were asked their opinion on the presented automation, calculations, workflows, and the framework in general. Commentators identified the suggested framework as a viable and complete proposition while at the same time confirming that all BD-CPS functional characteristics, as exhibited in Table 31, were showcased and would have a positive impact on corporate DLP challenges. "*For sure, the work adds significant value to the business sector*" and will prove to be helpful to the users in their day-to-day operations and preserve the organisation's interests towards the threat of Data Loss.

The framework proposed was accredited by all expert commentators, and it was confirmed that it could be a valuable addition to any organisation's arsenal toward Data Loss Prevention. Throughout the process, it was exciting to observe a diverse set of experts converging in identifying similar challenges and confirming the framework's suitability for a diverse set of organisational applications towards data confidentiality.

It was suggested that the system could become a Software as a Service (SaaS), a proposal where the organisations engaged can, should they choose to, share information amongst themselves. In this way, based on a sectoral classification, the system can provide templates and proposed values, percentages, classifications, actions or levels to the participants by aggregating existing similar prior input from other participants.

Another future evolvement could include an automatic calculation where the system, considering the classification attributes, the anonymization factor, and the intended use of the data, will provide a risk factor. The factor will then be used to differentiate the approval workflows and define different roles and approval levels. The risk factor can be further augmented using the history of the approved and shared datasets, where the concentration risk can be identified. In this calculation, the system will further aggregate the data that a destination already has and warn on possible exposures from the combination of seemingly unrelated distributions. For a tool to gain acceptance in the corporate world, it has to integrate and interact with existing office productivity tools

(Collins, 2007); towards that end, the system will have to provide hooks or addins for the most commonly used business applications.

7.5. Conclusion

With the use of Business Classification, Regulatory Classification and Anonymization Risk attributes, the DB-CPS seeks to classify all elements in the corporate data domain and automate the process of data safe distribution. The approach was proven effective and efficient by a group of expert commentators based on their evaluation of a working prototype that showcased the framework.

The adoption of such a system in the corporate environment will enhance user awareness with the means of hands-on interaction. In addition to that, it will provide a clear guideline across the organisation and facilitate a consistent approach in managing data confidentiality. Last but not least, DB-CPS, will serve as a reference point for the governance and control units of the organisation in respect to approvals and data dissemination. By providing a centralised system for policy making, approvals and all the forensic data related to data set confidentiality management, DB-CPS will become the heart of the corporate data policy.

Chapter 8. CONCLUSION

This research is focused on identifying means of minimising the effects of Big Data *Variety*. The areas of confidential data identification, dataset categorisation and data loss prevention have been investigated. In this chapter the research achievements are presented (Section 8.1) followed by a discussion of the limitations and risks (in Section 8.2). The thesis finishes in Section 8.3 with an outline of the future directions and enhancements of the research.

8.1. Research Achievements

This research is focused on feasibility, applicability and attainability in three areas.

The main objective in regards to Optimising Confidential Data Identification, was to increase the recognition accuracy in such a manner that the process can be used in a Big Data environment. False positives will hinder the results and hide important characteristics and information of the data set, thus elevating the accuracy is a show stopper if not dealt with. The proposed extension of the classic regular expressions with the use of “Booster Metrics” has proven to increase a) efficiency by more than 30% due to accurately identifying confidential information and thus removing false positives, and b) effectiveness by more than 10% by widening the horizon and identifing more confidential information. Combining the two will provide a superior approach in identifying privileged and regulated information that can enable detection with minimal effort and limited human resources.

For Automating Data Characterisation, the viability of countering *Variety* with such an approach is the focal point. Using a fragment of the data in conjunction with highly accurate prediction has proven to be feasible. The use of statistical analysis along with machine learning can result in an approach that will yield accuracy levels near to 90% while the data consumed will remain as low as 2% of the overall data set. An important factor to also take into consideration is that the designed and implemented approach can be utilised in diversified and heterogeneous data sets. This cross-heterogeneity attribute of the proposal is an indicator that *Variety* effects have been countered to an extent. Reflecting upon the costs related to implementation, since the system is auto-calibrated with self-learning capabilities the human capital and invested man-hours are less thus lowering the TCO and making the Big Data initiatives more attractive.

The last central point of this research was Data Confidentiality Preservation, where an approach is exhibited and a system is showcased and tested. The capability of organisations to counter the regulatory requirements triggered by legal or soft-law requirements is the main driver. To that end, the system proposal should be consistent across the organisation but flexible enough to cater for any adjustments required for particular confidentiality requirements. The BD-CPS, with the use of a PoC, has demonstrated its capabilities in being a system that will help business adoption towards DLP. The suggested implementation and concepts can adequately classify confidential information along three pivotal axes, namely Business, Regulatory and Risk Anonymization. It can increase awareness and become a tool for hands-on training throughout the organisation. Its algorithmic and workflow capabilities along with the real-time classification representation of the rule-based confidentiality policies, appear to be a valuable proposition for any organisation. The extensive auditing capabilities facilitate accountability and traceability, whilst the metadata gathered on all approvals processed creates a valuable knowledge pool. BD-CPS can help organisations of all sizes to minimise the risk of data exposure as confirmed by a diverse set of experts from academia and business.

8.2. Research Limitations

Although no direct limitation was caused by the COVID-19 pandemic on the research, indirectly it was affected due to the socioeconomic ripples of the pandemic. The world research initiatives in information technology were not mainly focused on Big Data but rather trying to address sustainability issues like remote working. Similarly investments slowed down since almost all business areas suffered from the prolonged economic recession imposed by the pandemic.

In all experiments the dimensions of Volume and Velocity pose a limitation since all systems will require a set of high-end processing power which could lead to the risk of failures. With the adoption of cloud computing and cloud infrastructures these can be countered but this will also pose a new risk of data disclosure and data leakage. For example certain countries will not allow the use of clouds residing outside their borders, thus substantially limiting the implementation capabilities and options.

Coming to the research topics at hand, each area has identified certain limitations and risks. When referring to ”Optimising Confidential Data Identification,” the initial set of matching rules and “Buster Metrics” algorithms will have to be identified. Although this is mainly a one-time effort, there is a risk of not being able to employ knowledgeable resources to undertake such specialised task. The limitations might be related to cost or availability, since the respective human capital should be knowledgeable in IT and possess in-depth business functions knowledge. For Automating Data Characterisation a substantial number of executions on a large sample of diverse datasets have to happen in order to be able to retrofit the data onto the system and further validate and enhance it. The risk of not having access to such an extensive repository, could have an impact on fine-tuning the system. Additionally, the risk of the neural network becoming outdated is apparent, if it is not retrained during the adoption of new sets. The BD-CPS requires extended timelines to be implemented across any organisation; thus, it can be side tracked in case of revenue earning projects, which will take precedence. Any corrective action and possible realignment of the system might have to be done later in the day since evaluation can happen only once BD-CPS is rolled out to a substantial part of the organisation. The rework might be relatively high and to that end it is proposed to start with an extensive PoC across the organisation that will quickly provide adequate feedback.

All three proposals in countering *Variety*, although they exploit many automated techniques, lead to a supervised process since the data acientists interactions are imperative in fine-tuning the models, monitoring the systems and understanding the exceptions. The presented risks and limitations can slowdown or even stop adoption, thus should be taken into consideration towards successful implementations.

8.3. Future Directions

Information technology is a fast growing research field thus predicting future trends and possible implementations can be difficult. Based on the research performed and current trends, possible future advancements and extensions to the work at hand will be proposed. The Optimising Confidential Data Identification can be extended by employing AI and utilising advanced algorithms for pattern matching. For Automating Data Characterisation use of advanced techniques for data clustering and identification can be beneficial. Employment of Natural Language Processing and linguistics can also extend and enhance the accuracy of categorisation. The adoption of unsupervised machine learning algorithms, in which models are not supervised with the use of training data sets but instead dwell on the data for hidden patterns, associations, and insights, similar to human brain learning, could also be investigated while the AI and ML technics are advancing. The BD-CPS can be extended by implementing templates for sectoral ratings in order to provide a jump start for any organisation embarking on the implementation. Metadata can be utilised as input into visual data analytics in understanding trends and could also be combined across the community in identifying ratings for collective intelligence. The implementation of a BD-CPS plugin to be integrated with existing business toolsets and office automation can be an additional track to pursue in increasing adoption. Also, additional matrices and indicators can be developed to augment the system's capabilities e.g. creation of conformance rates for a set or aggregated conformance for the organisation.

REFERENCES

A Timeline of Database History. (2015). Intuit QuickBase. http://quickbase.intuit.com/articles/timeline-of-database-history

Abadi, D. J., Boncz, P. a., & Harizopoulos, S. (2009). Column-oriented database systems. *Proceedings of the VLDB*, 1664–1665. https://doi.org/10.14778/1687553.1687625

Abdullin, A., & Nasraoui, O. (2012). *Clustering Heterogeneous Data Sets*. https://doi.org/10.1109/LA-WEB.2012.27

Agrawal, D., Bernstein, P., & Bertino, E. (2011). Challenges and Opportunities with Big Data. *Proceedings of the VLDB Endowment*, 1–16. http://dl.acm.org/citation.cfm?id=2367572%5Cnhttp://docs.lib.purdue.edu/cctech/1/

Ali-Ud-Din Khan, M., Uddin, M. F., & Gupta, N. (2014). Seven V's of Big Data understanding Big Data to extract value. *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education - "Engineering Education: Industry Involvement and Interdisciplinary Trends", ASEE Zone 1 2014*. https://doi.org/10.1109/ASEEZone1.2014.6820689

Anusha, Y., Visalakshi, R., & Srinivas, K. (2021). *Big Data Quality-A Survey paper to attain Data quality*. https://doi.org/10.52458/978-81-95502-00-4-71

Armando, A., Bezzi, M., Metoui, N., & Sabetta, A. (2015). Risk-Based Privacy-Aware Information Disclosure. *International Journal of Secure Software Engineering*.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. a. S., & Buyya, R. (2014). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*. https://doi.org/10.1016/j.jpdc.2014.08.003

Audit IT. (2011). *Return on Investment (ROI)*. Ready Ratios. http://www.readyratios.com/reference/profitability/return_on_investment_roi.html

Baily, M. N., Manyika, J., & Gupta, S. (2013). U. S. Productivity Growth: An Optimistic Perspective. *International Productivity Monitor*, *23*, 3–12. http://www.csls.ca/ipm/25/ipm-25-baily-manyika-gupta.pdf

Baker, P. (2015, January 15). *Variety, not volume, biggest big data challenge in 2015*. FireceBigData. http://www.fiercebigdata.com/story/variety-not-volume-biggest-big-data-challenge-2015/2015-01-14

Bakshi, K. (2012). Considerations for big data: Architecture and approach. *IEEE Aerospace Conference Proceedings*, 1–7. https://doi.org/10.1109/AERO.2012.6187357

Banek, M., Vrdoljak, B., Tjoa, A. M., & Skočir, Z. (2007). Automating the Schema Matching Process for Heterogeneous Data Warehouses. *Springer*, *4654*, 45–54.

ΠΔ/ΤΕ 2577 - Annex 8, Bank Of Greece - ΠΔ/ΤΕ (2006).

Batini, C., Barone, D., Cabitza, F., & Grega, S. (2011). A Data Quality Methodology for Heterogeneous Data. *International Journal of Database Management Systems*, *3*. https://doi.org/10.5121/ijdms.2011.3105

Bedi, P., Jindal, V., & Gautam, A. (2014). Beginning with Big Data Simplified. *IEEE*.

Berisha, B., Mëziu, E., & Shabani, I. (2022). Big data analytics in Cloud computing: an overview. *Journal of Cloud Computing, 11,* 24. https://doi.org/10.1186/s13677-022-00301-w

Big Data Consulting Market: Market size, Industry outlook, Market forecast, Demand Analysis ,Market Share, Market Report 2021-2026. (2021). Furion Analytics Research & Consulting LLP. https://www.industryarc.com/Report/17928/big-data-consulting-market.html?utm_source=FreePR&utm_medium=Sharath&utm_campaign=Sharat

h

Bigelow, S. (2021). *TCO (total cost of ownership)*. TechTarget SearchDataCenter. http://searchdatacenter.techtarget.com/definition/TCO

Biscobing, J. (2019). *Entity Relationship Diagram (ERD)*. TechTarget SearchDataCenter. http://searchcrm.techtarget.com/definition/entity-relationship-diagram

Blažič, A. J., & Šaljić, S. (2010). Confidentiality Labeling Using Structured Data Types. *IEEE*, 182–187. https://doi.org/10.1109/ICDS.2010.70

Bollier, D. (2010). The Promise and Peril of Big Data. In *Program*. http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/InfoTech09.pdf

Boote, D. N., & Beile, P. (2005). Scholars Before Researchers: On the Centrality of the Dissertation Literature Review in Research Preparation. *Educational Researcher*, *34*(6), 3–15.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society - Routledge Tailor & Francis*, *15*(June 2012), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brandon, H., & De Souza, L. (2017). *Big Data Analytics and privacy & data protection*.

Briody, D. (2011). Big data - Harnessing a game-changing asset. *The Economist*. http://www.sas.com/resources/asset/SAS_BigData_final.pdf

Brock, V. F., & Khan, H. U. (2017). Are enterprises ready for big data analytics? A survey-based approach. *International Journal of Business Information Systems*, *25*(2), 256–277. https://doi.org/10.1504/IJBIS.2017.10004408

Brown, E. (2014). *Big Data Problems - Variety not Volume*. Big-Data Forum. http://www.big-dataforum.com/605/big-data-problems-variety-not-volume

Brown, K. A., Brittman, S., Maccaferri, N., Jariwala, D., & Celano, U. (2020). Machine Learning in Nanoscience: Big Data at Small Scales. In *Nano Letters* (Vol. 20, Issue 1, pp. 2–10). American Chemical Society. https://doi.org/10.1021/acs.nanolett.9b04090

Brust, A. (2018). *Hortonworks, Confluent and Waterline attempt to make Big Data easier*. ZDNet. https://www.zdnet.com/article/hortonworks-confluent-and-waterline-make-big-data-easier/

Bughin, J. R. J., Cincera, M., Reykowska, D., & Ohme, R. (2021). Big data is decision science: The case of COVID-19 vaccination. In *Handbook of Research on Applied Data Science and Artificial Intelligence in Business and Industry* (pp. 126–150). IGI Global.

Cattell, R. (2010). Scalable SQL and NoSQL Data Stores. *SIGMOD Record, 39*(4), 12. https://doi.org/10.1145/1978915.1978919

Cerqueus, T., Almeida, E. C. De, & Scherzinger, S. (2015). Safely Managing Data Variety in Big Data Software Development. *IEEE*, 7. https://doi.org/10.1109/BIGDSE.2015.9

Chapple, M. (2022, February 22). *The Basics of Database Normalization*. Databases.about.Com. http://databases.about.com/od/specificproducts/a/normalization.htm

Chen, P. P.-S. (1976). The Entity-Relationship Model - Toward a Unified View of Data. *ACM Transactions on Database Systems*, *1*(1), 9–36. https://doi.org/10.1145/320434.320440

Cheriere, N., & Antoniu, G. (2017). How Fast Can One Scale Down a Distributed File System? *IEEE International Conference on Big Data*.

Chessell, M. (2014). Ethics for Big Data and Analytics. *IBM*. http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG

Study Report - Ethics for BD&A.pdf

Christiansen, N. H., Erlend, P., Voie, T., Winther, O., & Høgsberg, J. (2014). Comparison of Neural Network Error Measures for Simulation of Slender Marine Structures. *Journal of Applied Mathematics*. https://doi.org/10.1155/2014/759834

Codd, E. F. (1970). A relational model of data for large shared data banks. *Commun. ACM*, *13*(6), 377–387. https://doi.org/10.1145/357980.358007

Codd, E. F. (1974). Information Retrieval - A Relational Model of Data for Large Shared Data Banks. In *Communications of the AMC* (Vol. 13, pp. 377–387). Proc. 1974 Congress (Stockholm, Sweden, 1974).

Collins, C. (2007). *History of ODBC*. WordPress. https://ccollins.wordpress.com/2007/06/03/history-of-odbc/

Conrad, C., & Vault, A. (2014). Big data: an information security context. *Network Security*, *2014*(1), 18–19. https://doi.org/10.1016/S1353-4858(14)70010-8Feature

Corporation MarkLogic. (2012). *Top 3 Ways Big Data Impacts Financial Services* (Issue MarkLogic).

Crawford, K., Miltner, K., & Gray, M. L. (2014). Critiquing Big Data : Politics , Ethics , Epistemology. *International Journal of Communication*, *8*, 1663–1672. ijoc.org/index.php/ijoc/article/download/2167/1164

Creswell, J. (2013). *Qualitative Inquiry & Research Design*.

Cristina Abellan Matamoros. (2019, July 24). *Facebook to pay record $5 billion fine over privacy violations, but are they getting off lightly?* Euronews - REUTERS. https://www.euronews.com/2019/07/24/facebook-to-pay-record-5-billion-fine-over-privacy-violations-but-are-they-getting-off-lig

CrowCour, R. (2014, August 21). *Introducing Azure DocumentDB – Microsoft's fully managed NoSQL document database service*. Microsoft. http://blogs.msdn.com/b/documentdb/archive/2014/08/22/introducing-azure-documentdb-microsoft-s-fully-managed-nosql-document-database-service.aspx

CS Odessa Corp. (1993). *Structured Systems Analysis and Design Method (SSADM)*. CS Odessa Corp. http://www.conceptdraw.com/How-To-Guide/ssadm

Cuquet, M., Vega-Gorgojo, G., Lammerant, H., Finn, R., & Hassan, U. (2017). Societal impacts of big data: challenges and opportunities in Europe. *ArXiv - Cornell University*.

Dalvi, A., Siddavatam, I., Thakkar, V., Jain, A., Kazi, F., & Bhirud, S. (2021). Link Harvesting on the Dark Web. *2021 IEEE Bombay Section Signature Conference, IBSSC 2021*, 7–11. https://doi.org/10.1109/IBSSC53889.2021.9673428

Database Systems: A Brief Timeline. (2000).

Dautov, R., & Distefano, S. (2017). Quantifying Volume, Velocity, and Variety to Support (Big) Data-Intensive Application Development. *2017 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/BigData.2017.8258252

Davenport, T. H. (2012). The Human Side of Big Data and High-Performance Analytics. *International Institute of Analytics*.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, *50*(1).

Dawson, G., Mackintosh, M., & Paul, A. (2010). *Economics and Economic Change*. Academic Internet Publishers.

Dean, J., & Ghemawat, S. (2004). MapReduce: Simplied Data Processing on Large Clusters. *6th Symposium on Operating Systems Design and Implementation*, 137–149. https://doi.org/10.1145/1327452.1327492

del Río, S., López, V., Benítez, J. M., & Herrera, F. (2015). A MapReduce Approach to

Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules. *International Journal of Computational Intelligence Systems*, *8*(3), 422–437. https://doi.org/10.1080/18756891.2015.1017377

DeRoos, D. (2014). *The Shuffle Phase of Hadoop's MapReduce Application Flow*. Wiley - For Dummies. http://www.dummies.com/how-to/content/the-shuffle-phase-of-hadoops-mapreduce-application.html

Devakunchari, R. (2014). Analysis on big data over the years. *International Journal of Scientific and Research Publications, 4*(1), 1–7.

Domingo-Ferrer, J., & Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, *14*(1), 189–201. https://doi.org/10.1109/69.979982

Domingo-Ferrer, J., & Rebollo-Monedero, D. (2009). Measuring Risk and Utility of Anonymized Data Using Information Theory. *2009 EDBT/ICDT Workshops*.

Draelos Rachel. (2019). *Best Use of Train/Val/Test Splits, with Tips for Medical Data – Glass Box*. Glassbox Medicine. https://glassboxmedicine.com/2019/09/15/best-use-of-train-val-test-splits-with-tips-for-medical-data/

Duhigg, C. (2012, February 16). *How Companies Learn Your Secrets*. The New Yourk Times Magazine. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Dutta, D., & Bose, I. (2014). Managing a Big Data project: The case of Ramco Cements Limited. *International Journal of Production Economics*. https://doi.org/10.1016/j.ijpe.2014.12.032

Elragal, A. (2014). ERP and Big Data: The Inept Couple. *Elsevier - Procedia Technology, 16*, 242–249. https://doi.org/10.1016/j.protcy.2014.10.089

European Commission. (2022, July). *Implementation of the Public Sector Information Directive*. https://digital-strategy.ec.europa.eu/en/policies/public-sector-information-directive

European Medicines Agency. (2017). *External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use*. www.ema.europa.eu/contact

Experfy Editor. (2014). *Cloudera vs Hortonworks vs MapR: Comparing Hadoop Distributions*. http://www.experfy.com/blog/cloudera-vs-hortonworks-comparing-hadoop-distributions/

Fan, S., Lau, R. Y. K., & Zhao, J. L. (2015). Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix. In *Big Data Research*. https://doi.org/10.1016/j.bdr.2015.02.006

Farenda. (2017). *Java regex matching IP Address - Yet another programming solutions log*. Yet Another Programming Solutions Log. https://farenda.com/java/java-regex-matching-ip-address/

Fayyad, U., & Uturusamy, R. (2002). Evolving Data Mining into Solutions for Insights. *Communications of the ACM*, *45*(8).

Fioretti, J. (2018). EU privacy watchdogs to look into harvesting of data from social media. *Reuters*. https://www.reuters.com/article/us-facebook-privacy-eu/eu-privacy-watchdogs-to-look-into-harvesting-of-data-from-social-media-idUSKBN1HJ15O

Fisher, D., DeLine, R., Czerwinski, M., & Drucker, S. (2012). Interactions with big data analytics. *Microsoft - Interactions*, *19*, 50. https://doi.org/10.1145/2168931.2168943

Floratou, A., Patel, J., Shekita, E., & Tata, S. (2011). Column-Oriented Storage

Techniques for MapReduce. *VLDB Endowment*, 419–429. http://arxiv.org/abs/1105.4252

Fordham University School of Law. (2012). *Big Data, Big Issues*. 1–5.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*. https://doi.org/10.1016/j.ijpe.2014.12.031

Friedman, H., & Amoo, T. (1999). Rating The Rating Scales. *The Journal of Marketing Management, Winter,* 114–123.

Fujitsu. (2014). *Solution Approaches for Big Data*.

Gantz, J., & Reinsel, D. (2011). Extracting Value from Chaos. *IDC*. http://www.emc.com/digital_universe.

Gantz, J., & Reinsel, D. (2012). THE DIGITAL UNIVERSE IN 2020: Big Data , Bigger Digital Shadow s , and Biggest Grow the in the Far East. *IDC*, 1–16.

Gantz, J., & Reinsel, D. (2013). THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East - Uni ted States. *IDC*. http://www.emc.com/leadership/digital-universe/iview/index.htm.

Gartner. (2018). *Gartner Peer Insights “Voice of the Customer”: Data Management Solutions for Analytics Market*. Gartner Peer Insights. https://www.gartner.com/doc/reprints?id=1-5AP8NEV&ct=180804&st=sb

General Data Protection Regulation (GDPR) Compliance Guidelines. (2020). GDPR.EU. https://gdpr.eu/

GilPress. (2015). *Top Ten Most Funded Big Data Startups January 2015 - Forbes*. Forbs. http://www.forbes.com/sites/gilpress/2015/01/31/top-ten-most-funded-big-data-startups-january-2015/

Gopal, P. R. C., Nripendra, ·, Rana, P., Thota, ·, Krishna, V., Ramkumar, · M, Nripendra, B., Krishna, T. V., & Ramkumar, M. (2022). Impact of big data analytics on supply chain performance: an analysis of influencing factors. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04749-6

Gordo, B. (2017). “Big Data” in the Information Age. *City & Community*. https://doi.org/10.1111/cico.12219

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., & De Wolf, P.-P. (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, *14*(4), 463–478.

Goyvaerts, J. (2021). *Finding or Verifying Credit Card Numbers*. Regular-Expresions.Info. https://www.regular-expressions.info/creditcard.html

Greenbaum, J. (2008). Adding Business Value to Database Consolidation. *Enterprise Applications Consulting*.

Gregory, M. (2013). *Strategies for Implementing Big Data Analytics*. Richard Linowes, Kogod School of Business.

Guest, G., Bunce, A., & Johnson, L. (2006). How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods*, *18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Gupta, M. (2014). Organizational Culture and the Three V’s of Big Data. *MWAIS 2014 Proceedings*, Paper 10.

Han, J., Haihong, E., Le, G., & Du, J. (2011). Survey on NoSQL database. *2011 6th International Conference on Pervasive Computing and Applications, ICPCA 2011*, 363–366. https://doi.org/10.1109/ICPCA.2011.6106531

Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC. (1996). https://www.cdc.gov/phlp/publications/topic/hipaa.html

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., Van Den Hoven, J., Zicari, R. V, & Zwitter, A. (2017). Will Democracy Survive Big Data and Artificial Intelligence? *Scientific American's*. https://www.scientificamerican.com/article/will-democracy-survi...

Hieatt, E., & Mee, R. (2002). Going faster: Testing the Web application. *IEEE Software*, *19*(2), 60–65. https://doi.org/10.1109/52.991333

Hill, K. (2012, February 16). *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did*. Forbes. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=bbff29f66686

Hogan, M. T., & Jovanovic, V. (2015). ETL Workflow Generation for Offloading Dormant Data from the Data Warehouse to HADOOP. *Issues in Information Systems*, *16*(I), 91–101.

Hong, L. (2014). Philosophical Reflections on Data. *Elsevier - Procedia Computer Science*, *30*, 60–65. https://doi.org/10.1016/j.procs.2014.05.381

Hoy, M. B. (2014). Big Data: An Introduction for Librarians. *Taylor & Francis - Medical Reference Services Quarterly*, *33*, 320–326. https://doi.org/10.1080/02763869.2014.925709

Hussein, A. A. (2020). Fifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). *Journal of Information Security*, *11*, 304–328. https://doi.org/10.4236/jis.2020.114019

Hwang, W., & Salvendy, G. (2010). Number of People Required for Usability Evaluation: The 10±2 Rule. *Communications of the ACM*, *53*(5), 130–133. https://doi.org/10.1145/1735223.1735255

Hyett, N., Kenny, A., & Dickson-Swift, V. (2014). Methodology or method ? A critical review of qualitative case study reports. *International Journal of Qualitative Studies on Health and Well-Being*, *1*, 1–12. https://doi.org/10.3402/qhw.v9.23606

IBAN Regex design - Stack Overflow. (2017). StackOverflow. https://stackoverflow.com/questions/44656264/iban-regex-design

IBM. (2015, February 13). *What is big data?* IBM; IBM Corporation. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Jabbar, A., Akhtar, P., & Dani, S. (2020). Real-time big data processing for instantaneous marketing decisions: A problematization approach. *Industrial Marketing Management*, *90*, 558–569. https://doi.org/10.1016/J.INDMARMAN.2019.09.001

Jacobs, A. (2009). The Pathologies of Big Data. *Communications of the ACM*, *52*(8), 36–44. https://doi.org/10.1145/1563821.1563874

Jain, H., & Gosain, A. (2012). A Comprehensive Study of View Maintenance Approaches in Data Warehousing Evolution. *ACM SIGSOFT Software Engineering Notes*, *37*(5), 1–8. https://doi.org/10.1145/2347696.2347705

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *SpringerOpen - Journal of Big Data*, *3*(1). https://doi.org/10.1186/s40537-016-0059-y

Jensen, M. H., Nielsen, P. A., & Persson, J. S. (2021). *Improving the impact of Big Data analytics projects with benefits dependency networks*. 2th Scandinavian Conference on Information Systems. https://aisel.aisnet.org/scis2021/2

Johnson, B. D. (2012). The Secret Life of Data in the Year 2020. *The Futurist - World Future Society*, *46*(4), 20–23. http://www.wfs.org/futurist/july-august-2012-vol-46-no-4/secret-life-data-year-2020

Jose, J., Subramoni, H., Luo, M., Zhang, M., Huang, J., Wasi-Ur-Rahman, M., Islam, N. S., Ouyang, X., Wang, H., Sur, S., & Panda, D. K. (2011). Memcached Design on

High Performance RDMA Capable Interconnects. *International Conference on Parallel Processing*, 743–752. https://doi.org/10.1109/ICPP.2011.37

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Sciences*, 995–1004. https://doi.org/10.1109/HICSS.2013.645

Kakish, K., & Kraft, T. a. (2012). ETL Evolution for Real-Time Data Warehousing. *Conference on Information Systems Applied Research*, 1–12. www.aitp-edsig.org

Kalambe, Y. S., Pratiba, D., & Shah, P. (2015). Big Data Mining Tools for Unstructured Data: A Review. *INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH*, *3*(2), 2012–2017. http://www.ijitr.com/index.php/ojs/article/view/610

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561–2573. https://doi.org/10.1016/j.jpdc.2014.01.003

Kamble, S. S., Belhadi, A., Gunasekaran, A., Ganapathy, L., & Verma, S. (2021). A large multi-group decision-making technique for prioritizing the big data-driven circular economy practices in the automobile component manufacturing industry. *Technological Forecasting and Social Change*, *165*. https://doi.org/10.1016/J.TECHFORE.2020.120567

Karpathiotakis, M., Alagiannis, I., & Ailamaki, A. (2016). Fast Queries Over Heterogeneous Data Through Engine Customization. *2016 VLDB Endowment*, *9*(12), 972–983. https://doi.org/10.14778/2994509.2994516

Kataria, M., & Mittal, M. P. (2014). BIG DATA: A Review. *International Journal of Computer Science and Mobile Computing*, *3*(7), 106–110.

Kaur, N., & Sood, S. K. (2017). Efficient Resource Management System Based on 4Vs of Big Data Streams. *Elsevier - Big Data Research*. https://doi.org/10.1016/j.bdr.2017.02.002

Kaur Sandhu. Amanpreet. (2022). Big Data with Cloud Computing: Discussions and Challenges. *BIG DATA MINING AND ANALYTICS (IEEE)*, *5*(1), 32–40. https://doi.org/10.26599/BDMA.2021.9020016

Kemp, R. (2014). Legal aspects of managing Big Data. *Computer Law & Security Review*, *30*(5), 482–491. https://doi.org/10.1016/j.clsr.2014.07.006

Kimura, C. (2014). *Beyond the "Big": Solving for Data Variety Requires New Thinking - ClearStory Data*. CrearStory Data. http://www.clearstorydata.com/2014/12/beyond-big-solving-data-variety-requires-new-thinking/

Kleinman, Z. (2018). Cambridge Analytica: The story so far. *BBC*. http://www.bbc.com/news/technology-43465968

Koenig, S. (2019). *Can DLP protect credit card numbers without burying you in false positives?* ZScaler. https://www.zscaler.com/blogs/product-insights/can-dlp-protect-credit-card-numbers-without-burying-you-false-positives

Komar, M., Savenko, O., Sachenko, A., Lendiuk, T., Lipianina-Honcharenko, K., Hladiy, G., & Vasylkiv, N. (2022). Evaluation the Efficiency of Information Technology of Big Data Intelligence Analysis and Processing. *6th International Conference on Computational Linguistics and Intelligent Systems*.

Koo, J., Kang, G., & Kim, Y. G. (2020). Security and privacy in big data life cycle: A survey and open challenges. *Sustainability (Switzerland)*, *12*(24), 1–32. https://doi.org/10.3390/su122410571

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *National Academy of Sciences* *of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Krawczyk, B., Stefanowski, J., & Wozniak, M. (2015). Data stream classification and big data analytics. *Neurocomputing*, *150*, 238–239. https://doi.org/10.1016/j.neucom.2014.10.025

Kshetri, N. (2014). Big data's impact on privacy, security and consumer welfare. *Telecommunications Policy*, *38*, 1134–1145. https://doi.org/10.1016/j.telpol.2014.10.002

Kumar, M. (2013). *Variety Is the Unsolved Problem in Big Data | SmartData Collective*. Smart Data Collective. http://www.smartdatacollective.com/maheshkumar1/156256/why-variety-unsolved-problem-big-data

Kumar, M. (2014). *Today's Big Data Challenge Stems From Variety, Not Volume or Velocity*. Technopedia. https://www.techopedia.com/2/29109/trends/big-data/todays-big-data-challenge-stems-from-variety-not-volume-or-velocity

Lamberti, H. (2013). *Delusion in Organizational Excellence*. McGraw Hill Education.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies*, *949*, 4.

Lassoued, R., Macall, D. M., Smyth, S. J., Phillips, P. W. B., Hesseln, H., Canavari, M., Hingley, M., & Luis Vilalta-Perdomo, E. (2021). Expert Insights on the Impacts of, and Potential for, Agricultural Big Data. *MDPI*. https://doi.org/10.3390/su13052521

Leavitt, N. (2010). Will NoSQL Databases Live Up to Their Promise? *IEEE Computer*, 12–14. https://doi.org/10.1109/MC.2010.58

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Elsevier - Business Horizons*, *60*, 293–303. https://doi.org/10.1016/j.bushor.2017.01.004

Lennard. (2014). *Data Variety - The Ugly Duckling of Big Data*. DataShaka. http://www.datashaka.com/blog/non-techie/2014/01/data-variety-ugly-duckling-big-data

Livingstone, D. (2008). *Artificial Neural Networks*. https://doi.org/10:1007/978-1-60327-101-1

Lounici, S., Rosa, M., Negri, C. M., Trabelsi, S., & Önen, M. (2021). Optimizing leak detection in open-source platforms with machine learning techniques. *ICISSP 2021 - 7th International Conference on Information Systems Security and Privacy*, *Icissp*, 145–159. https://doi.org/10.5220/0010238101450159

Lovejoy, B. (2019). *GDPR fines total €56M in first year as Facebook faces 11 investigations*. 9To5Mac. https://9to5mac.com/2019/05/28/gdpr-fines/

Lucid Software Inc. (2015a). *ER Diagram Symbols and Meaning | Lucidchart*. Lucid Software Inc. https://www.lucidchart.com/pages/ER-diagram-symbols-and-meaning

Lucid Software Inc. (2015b). *What is ERD (Entity Relationship Diagram)? | Lucidchart*. Lucid Software Inc. https://www.lucidchart.com/pages/what-is-ERD

Luo, J., Dang, A.-R., & Mao, Q.-Z. (2008). The Study of Integration of Multi-Sources Heterogeneous Data Based on the Ontology. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XXXVII*.

Mahmoud, H., Hegazy, A., & Khafagy, M. H. (2018). An approach for Big Data Security based on Hadoop Distributed File system. *2018 International Conference on Innovative Trends in Computer Engineering*.

Majid, M. A. A., Othman, M., Mohamad, S. F., Lim, S. A. H., & Yusof, A. (2017). Piloting for Interviews in Qualitative Research: Operationalization and Lessons Learnt. *International Journal of Academic Research in Business and Social*

Sciences, *7*(4), 1073–1080. https://doi.org/10.6007/ijarbss/v7-i4/2916

Malhotra, N., & Peterson, M. (2006). *Basic Marketing Research: A Decision-Making Approach* (2nd ed.). Upper Saddle River, N.J. : Pearson/Prentice Hall.

Manning, K. (2020). *2 Eyes- 4 Eyes- 6 Eyes Principle | ProcessMaker*. ProcessMaker. https://www.processmaker.com/blog/2-eyes-4-eyes-6-eyes-principle/

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hungg Bayers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, *June*, 156. https://doi.org/10.1080/01443610903114527

Markov, I. L., Liu, J., & Vagner, A. (2021). Regular Expressions for Fast-response COVID-19 Text Classification. *ArXiv Cornell University*, *1*(1). http://arxiv.org/abs/2102.09507

McAbee, S. T., Landis, R. S., & Burke, M. I. (2017). Inductive reasoning: The promise of big data. *Human Resource Management Review*. https://doi.org/10.1016/j.hrmr.2016.08.005

Mcdaniel, P., Cárdenas, A. A., & Rajan, S. P. (2013). Big Data Analytics for Security. *IEEE Computer and Reliability Societies*. www.computer.org/security

McElhenny, J. (2014). Leaving Data on the Table : New Survey Shows Variety , Not Volume , is the Bigger Challenge of Analyzing Big Data. *InkHouse (for Paradigm4)*, 3–4.

McKinsey & Company. (2013). *How advanced analytics are redefining banking*. McKinsey & Company. http://www.mckinsey.com/insights/business_technology/how_advanced_analytics_are_redefining_banking

Michel, P., Dmitriyev, V., Abilov, M., & Marx, J. (2014). ELTA : New Approach in Designing Business Intelligence Solutions in Era of Big Data. *Elsevier - Procedia Technology*, *16*(1), 667–674. https://doi.org/10.1016/j.protcy.2014.10.015

Miettinen, J., & Tergujeff, R. (2021). Conclusions and Outlook—Summary of Big Data in Forestry. In *Big Data in Bioeconomy* (pp. 363–367). Springer, Cham.

Mikalef, P., Pappas, I., Krogstie, J., & Pavlou, P. A. (2020). Big data and business analytics: A research agenda for realizing business value. *Information & Management - Elsevier*.

Mohamed, I., & Noordin, M. F. (2011). Business metadata with the STA data modelling technique. *2011 International Conference on Electrical Engineering and Informatics*, *July*, 1–4. https://doi.org/10.1109/ICEEI.2011.6021607

MongoDB, I. (2013). *Introduction to MongoDB*. MongoDB Inc. http://www.mongodb.org/about/introduction/

Mulherrin, E. a, & Abdul-hamid, H. (2009). *The evolution of ETL* (Vol. 3). https://doi.org/10.1016/S1361-3723(14)70541-X Feature

Munawar, H. S., Qayyum, S., Ullah, F., & Sepasgozar, S. (2020). Big Data and Its Applications in Smart Real Estate and the Disaster Management Life Cycle: A Systematic Analysis. *Big Data and Cognitive Computing*.

Mylka, A., Mylka, A., Kryza, B., & Kitowski, J. (2012). Integration of Heterogeneous Data in an Ontological Knowledge dase. *Computing and Informatics*, *31*, 189–223.

Narayanan, A., & Shmatikov, V. (2009). De-anonymizing Social Networks. *2009 30th IEEE Symposium on Security and Privacy*, 173–187. https://doi.org/10.1109/SP.2009.22

Nazabal, A., Olmos, P. M., Ghahramani, Z., & Valera, I. (2020). Handling Incomplete Heterogeneous Data using VAEs. *Pattern Recognition*.

NcCaffrey, J. (2015). *Neural Network Train-Validate-Test Stopping -- Visual Studio Magazine*. Visual Studio Magazine.

https://visualstudiomagazine.com/articles/2015/05/01/train-validate-test-stopping.aspx

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification.' *The Journal of Strategic Information Systems*, 1–12. https://doi.org/10.1016/j.jsis.2015.02.001

Ninghui, L., Tiancheng, L., & Suresh, V. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. *2007 IEEE 23rd International Conference on Data Engineering*.

O'Driscoll, A., Daugelaite, J., & Sleator, R. D. (2013). "Big data", Hadoop and cloud computing in genomics. *Journal of Biomedical Informatics*, *46*, 774–781. https://doi.org/10.1016/j.jbi.2013.07.001

Oesterreich, T. D., Anton, E., & Teuteberg, F. (2022). What translates big data into business value? A meta-analysis of the impacts of business analytics on firm performance. *Information and Management*, *59*(6). https://doi.org/10.1016/J.IM.2022.103685

Official PCI Security Standards Council Site - Verify PCI Compliance, Download Data Security and Credit Card Security Standards. (2006). https://www.pcisecuritystandards.org/

Oganian, A., & Domingo-Ferrer, J. (2001). On the Complexity of Optimal Microaggregation for Statistical Disclosure Control. *Statistical Journal of the United Nations Economic Commission for Europe*, *18*(4). https://doi.org/10.3233/SJU-2001-18409

Oguntimilehin, A., & Ademola, E. (2014). A Review of Big Data Management , Benefits and Challenges. *Journal of Emerging Trends in Computing and Information Sciences*, *5*(6), 433–438.

OKeefe, C. (2017). Privacy and Confidentiality in Service Science and BigData Analytics. *Springer*, 978–981. https://doi.org/10.1007/978-3-319-18621-4_5ï

Olaru, O. (2014). *Heterogeneous Data Warehouse Analysis and Dimensional Integration*. UNIVERSITYOFMODENA AND REGGIO EMILIA.

Padhy, R. P., Patra, M. R., & Satapathy, S. C. (2011). RDBMS to NoSQL: Reviewing Some Next-Generation Non-Relational Database's. *International Journal of Advanced Engineering Sciences and Technologies*, *11*(1), 15–30.

Palantir. (2015). *Palantir Gotham | Palantir*. Palantir. https://www.palantir.com/palantir-gotham/

Papadakis, G., Ioannou, E., Niederée, C., Palpanas, T., & Nejdl, W. (2012). Beyond 100 million entities: Large-scale Blocking-based Resolution for Heterogeneous Data. *Fifth ACM International Conference on Web Search and Data Mining - WSDM '12*. https://doi.org/10.1145/2124295.2124305

Pijanowski, B. C., Tayyebi, A., Doucette, J., Pekin, B. K., Braun, D., & Plourde, J. (2014). A big data urban growth simulation at a national scale: Configuring the GIS and neural network based Land Transformation Model to run in a High Performance Computing (HPC) environment. *Environmental Modelling and Software*, *51*, 250–268. https://doi.org/10.1016/j.envsoft.2013.09.015

Power, D. J. (2014). Using 'Big Data' for analytics and decision support. *Taylor & Francis - Journal of Decision Systems*, *23*(2), 222–228. https://doi.org/10.1080/12460125.2014.888848

Rai, S., & Sharma, A. (2020). Research Perspective on Security Based Algorithm in Big Data Concepts. *International Journal of Engineering and Advanced Technology*, *9*(3), 2138–2143. https://doi.org/10.35940/ijeat.c5407.029320

Rake, R., & Kumar, V. (2020). Big data as a service Market Statistics: 2026. In *Allied Market Research*. https://www.alliedmarketresearch.com/big-data-as-a-service-market

Rawat, D. B., Doku, R., & Garuba, M. (2019). Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security. *IEEE*.

regex - Algorithms for detecting Credit Card Number reducing false positives/negatives - Stack Overflow. (2013). StackOverflow. https://stackoverflow.com/questions/18842081/algorithms-for-detecting-credit-card-number-reducing-false-positives-negatives

RegExp Library Formats. (2019). https://searchcode.com/codesearch/view/54908594/

Reggio, G., & Astesiano, E. (2020). Big-Data/Analytics Projects Failure: A Literature Review; Big-Data/Analytics Projects Failure: A Literature Review. *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. https://doi.org/10.1109/SEAA51224.2020.00050

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da, (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj

Rehman, F., Khalid, O., Haq, N. ul, Maqsood, T., Ali, M., Ahmad, R. W., Shuja, J., Sarwar, S., & Madani, S. A. (2020). A scalable model for real-time venue recommendations using MapReduce. *Concurrency and Computation: Practice and Experience*, *32*(21). https://doi.org/10.1002/cpe.5597

Reinsel, D., Gantz, J., & Rydning, J. (2018). The Digitization of the World From Edge to Core. *IDC*.

Richards, N., & King, J. (2013). Three Paradoxes of Big Data. *Stanford Law Review*, *66*(41), 41–46. http://www.stanfordlawreview.org/online/privacy-and-big-data/three-paradoxes-big-data%5Cnhttp://www.stanfordlawreview.org/sites/default/files/online/topics/66_StanLRevOnline_41_RichardsKing.pdf

Rimol, M. (2022, June). *Gartner Forecasts Worldwide IT Spending to Reach $4.4 Trillion in 2022*. Gartner. https://www.gartner.com/en/newsroom/press-releases/2022-04-06-gartner-forecasts-worldwide-it-spending-to-reach-4-point-four-trillion-in-2022

Rui, M., Honglong, X., Wenbo, W., Jianqiang, L., Yan, L., & Minhua, L. (2015). Overcoming the Challenge of Variety : Big Data Abstraction , the Next Evolution of Data Management for AAL Communication Systems. *IEEE Communications Magazine*, *January*, 42–47.

Rumbold, J., & Pierscionek, B. (2018). Contextual Anonymization for Secondary Use of Big Data in Biomedical Research: Proposal for an Anonymization Matrix. *JMIR Medical Informatics*, *6*(4), e47. https://doi.org/10.2196/medinform.7096

Rydning, J. (2022, May). *Worldwide IDC Global DataSphere Forecast, 2022–2026: Enterprise Organizations Driving Most of the Data Growth*. IDC. https://www.idc.com/getdoc.jsp?containerId=US49018922

Sabelfeld, A., & Myers, A. C. (2003). Language-Based Information-Flow Security. *IEEE Journal on Selected Areas in Communications*, *21*(1), 5–19. https://doi.org/10.1109/JSAC.2002.806121

Saeed, N., & Husamaldin, L. (2021). Big Data Characteristics (V's) in Industry. *Iraqi Journal of Industrial Research*, *8*(1). https://doi.org/10.53523/ijoirVol8I1ID52

Saha, A., Denning, T., Srikumar, V., & Kasera, S. K. (2020). Secrets in Source Code: Reducing False Positives using Machine Learning. *2020 International Conference on COMmunication Systems and NETworkS, COMSNETS 2020*, 168–175.

https://doi.org/10.1109/COMSNETS48256.2020.9027350

Sakharkar, V. S., Jeeri Dande, M., & Mate, S. (2017). Cloud and Big Data: A Compelling Combination. *International Journal of Engineering Science and Computing*. http://ijesc.org/

Santiago, R. (2005). *Regular Expression Library*. RegExLib.Com. https://regexlib.com/(X(1)A(1No-5dalAoSIC8mpU-0wtp7B9cY7gTSTHVOaIrtzqHDEK9roERzQ2Qro29iJ7wrPJbrhL8nohAyR_1PpvX2SoTm4pa4WPt95rUUSi-P_9scEQpMFdfhWIgH_Rad6LjTZ9_gaGMBlTOllz7DwBix9fkQTZHNPpix3sG1xigIlVPjkvdyXSRSwkAvKCxN87A5j0))/REDetails.aspx?regexp_id=993&AspxAutoDetectCookieSupport=1

Schmidt, B., & Hildebrandt, A. (2017). Next-generation sequencing: big data meets high performance computing. In *Elsevier - Drug Discovery Today* (Vol. 22, Issue 4, pp. 712–717). https://doi.org/10.1016/j.drudis.2017.01.014

Schmidt, M. (2015). *Return on Investment ROI Defined Explained Calculated Compared*. Solution Matrix Limited. https://www.business-case-analysis.com/return-on-investment.html

Schneider, R. (2012). *Hadoop for Dummies*.

Schryen, G., Wagner, G., & Benlian, A. (2015). Theory of Knowledge for Literature Reviews: An Epistemological Model, Taxonomy and Empirical Analysis of IS Literature. *Thirty Sixth International Conference on Information Systems*.

Segura, D. C. M., Oliveira, M. D. C., Okada, T. K., Lobato, R. S., Manacero, A., Carvalho, L. R., & Spolon, R. (2015). Availability in the Flexible and Adaptable Distributed File System. *IEEE 14th International Symposium on Parallel and Distributed Computing, ISPDC 2015*. https://doi.org/10.1109/ISPDC.2015.24

Sei, Y., Okumura, H., Takenouchi, T., & Ohsuga, A. (2019). Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness. *IEEE Transactions on Dependable and Secure Computing*, *16*(4), 580–593. https://doi.org/10.1109/TDSC.2017.2698472

Selcuk, A., Orencik, C., & Savas, E. (2015). Private Search Over Big Data Leveraging Distributed File System and Parallel Processing. *The Sixth International Conference on Cloud Computing*.

Seliya, N., Abdollah Zadeh, A., & Khoshgoftaar, T. M. (2021). A literature review on one-class classification and its potential applications in big data. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/S40537-021-00514-X

Sen, A. (2004). Metadata management: past, present and future. *Elsevier - Decision Support Systems*, *37*, 151–173. https://doi.org/10.1016/S0167-9236(02)00208-7

Shacklett, M. (2014, September 4). *How to cope with the big data variety problem*. TechRepublic. http://www.techrepublic.com/article/how-to-cope-with-the-big-data-variety-problem/

Shan, T. C., & Hua, W. W. (2006). Taxonomy of Java Web Application Frameworks. *IEEE International Conference on E-Business Engineering, ICEBE 2006*, 378–385. https://doi.org/10.1109/ICEBE.2006.98

Shaw, N., Eschenbrenner, B., & Brand, B. M. (2022). Towards a Mobile App Diffusion of Innovations model: A multinational study of mobile wallet adoption. *Elsevier - Journal of Retailing and Consumer Services*, *64*. https://doi.org/10.1016/j.jretconser.2021.102768

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Elsevier - Journal of Business Research*, *70*. https://doi.org/10.1016/j.jbusres.2016.08.001

smartdraw. (1994a). *Data Flow Diagram (DFD)*. Smartdraw. http://www.smartdraw.com/data-flow-diagram/

smartdraw. (1994b). *Entity Relationship Diagram (ERD)*. Smartdraw. http://www.smartdraw.com/entity-relationship-diagram/

Smitha, F. (1998). *Civilization in Mesopotamia*. MACROHISTORY : WORLD HISTORY. http://www.fsmitha.com/h1/ch01.htm

Solbers, R. (2012). Data Classification Tips: Finding Credit Card Numbers. *Varonis*. https://www.varonis.com/blog/data-classification-tips-finding-credit-card-numbers

Speegle, G., & Baker, E. (2014). Integration of Big Data Components for NoSQL Problems. *The 2014 International Conference on Advances in Big Data*.

Spishak, E., Dietl, W., & Ernst, M. D. (2012). A type system for regular expressions. *FTfJP 2012: The 14th Workshop on Formal Techniques for Java-Like Programs*, 20–26. https://doi.org/10.1145/2318202.2318207

Srivastava, D., & Dong, X. (2013). Big Data Integration. *Data Engineering (ICDE), 2013 IEEE 29th International Conference On*, 1245–1248. https://doi.org/10.1109/ICDE.2013.6544914

Sterner, B., & Franz, N. M. (2017). Taxonomy for Humans or Computers? Cognitive Pragmatics for Big Data. *Springer - Biological Theory*. https://doi.org/10.1007/s13752-017-0259-5

Storey, V. C., & Song, I. Y. (2017). Big data technologies and Management: What conceptual modeling can do. *Elsevier - Data and Knowledge Engineering*, *108*, 50–67. https://doi.org/10.1016/j.datak.2017.01.001

Studytonight. (2014). *1NF, 2NF, 3NF and BCNF in Database Normalization*. Studytonight. http://www.studytonight.com/dbms/database-normalization.php

Swanborn, P. (2010). *What is a Case Study?* SAGE Publications, Inc.

Tankard, C. (2012). Big data security. *Network Security*, *2012*(7), 5–8. https://doi.org/10.1016/S1353-4858(12)70063-6

Tannahill, B. K., & Jamshidi, M. (2014). System of Systems and Big Data analytics – Bridging the gap. *Elsevier - Computers & Electrical Engineering*, *40*(1), 2–15. https://doi.org/10.1016/j.compeleceng.2013.11.016

Taylor, D. G., Voelker, T. A., & Pentina, I. (2011). *Mobile Application Adoption by Young Adults: A Social Network Perspective* (Vol. 6, Issue 2). http://digitalcommons.sacredheart.edu/wcob_fac/1

The High Court of Justice - Chancery Division. (2007). *Approved Judgment*. *3053*(November), 12–14.

The World Bank. (2015). *GDP (current US$) | Data | Table*. The World Bank. http://data.worldbank.org/indicator/NY.GDP.MKTP.CD

Top 100 Big Data Companies of 2022. (2022, May). The Manifest. https://themanifest.com/big-data/companies

Top 13 Best Big Data Companies of 2022. (2022, April 3). SoftwareTestingHelp . https://www.softwaretestinghelp.com/big-data-companies/

Trader, T. (2014). *Big Data Future Hinges on Variety*. Datanami. http://www.datanami.com/2014/02/24/big_data_future_hinges_on_variety/

train test validation split python. (2020). IIIT-Delhi Blog. https://blog.iiitd.ac.in/wp-content/uploads/usborne-train-vihc/train-test-validation-split-python-943e13

Trifu, M. R., & Ivan, M. L. (2014). Big Data: present and future. *Database Systems Journal*, *V*(1), 32–41.

Turner, D., Schroeck, M., & Shockley, R. (2013). Analytics : The real-world use of big data in financial services. In *IBM Global Services*.

tutorialspoint.com. (2015). *Dbms - normalization*. Tutorialspoint.Com.

http://www.tutorialspoint.com/dbms/database_normalization.htm

U.S. Department of Justice - Office of Privacy and Civil Liberties. (2020). *United States Department of Justice Overview of the Act of 1974 - 2020 Edition Preface*. https://www.justice.gov/opcl/overview-privacy-act-1974-2020-edition.

US Bureau of Economic Analysis. (2022). *Gross Domestic Product, First Quarter 2022 (Advance Estimate) | U.S. Bureau of Economic Analysis (BEA)*. US Bureau of Economic Analysis . https://www.bea.gov/news/2022/gross-domestic-product-first-quarter-2022-advance-estimate

Vangie, B. (2015). *What is Normalization?* ITBusinessEdge. http://www.webopedia.com/TERM/N/normalization.html

Villars, R. L., Olofson, C. W., & Eastwood, M. (2011). *Big Data: What It Is and Why You Should Care*.

Vranopoulos, G., Trianatafylidis, A., & Yiannopoulos, M. (2016). Putting ATM Cash Requirements into Context, ANN relation to Socioeconomic Events and Variety. *BEFB 2016 - International Congress on Banking, Finance and Business 2016* , 526–545. https://doi.org/212-4044

Wang, C.-N., Dang, T.-T., Ai, N., Nguyen, T., Sakas, D. P., Reklitis, D. P., Trivellas, P., Vassilakis, C., & Terzi, M. C. (2022). The Effects of Logistics Websites' Technical Factors on the Optimization of Digital Marketing Strategies and Corporate Brand Name. *MDPI*. https://doi.org/10.3390/pr10050892

Wang, D., Muller, M., Liao, Q. V., Zhang, Y., Khurana, U., Samulowitz, H., Park, S., & Amini, L. 2021. (2021). How Much Automation Does a Data Scientist Want? *Association for Computing Machinery*.

Wang, L. (2017). Heterogeneous Data and Big Data Analytics. *Automatic Control and Information Sciences*, *3*(1), 8–15.

Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Toward quality data: An attribute-based approach. *Elsevier - Decision Support Systems*, *13*(3–4), 349–372. https://doi.org/10.1016/0167-9236(93)E0050-N

Wang, T., Yao, S., Xu, Z., Xiong, L., Gu, X., & Yang, X. (2015). An effective strategy for improving small file problem in distributed file system. *2015 2nd International Conference on Information Science and Control Engineering, ICISCE 2015*. https://doi.org/10.1109/ICISCE.2015.35

Warden, P. (2011). *Big Data Glossary*.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii. http://www.misq.org/misreview/announce.html

Whiting, R. (2015). *The 10 Coolest Big Data Startups Of 2015 (So Far)*. CRN. http://www.crn.com/slide-shows/data-center/300077457/the-10-coolest-big-data-startups-of-2015-so-far.htm

Wu, J., Li, H., Liu, L., & Zheng, H. (2017). Adoption of big data and analytics in mobile healthcare market: An economic perspective. *Elsevier - Electronic Commerce Research and Applications*, *22*, 24–41. https://doi.org/10.1016/j.elerap.2017.02.002

Y'barbo, D. (2012). *machine learning - multi-layer perceptron (MLP) architecture: criteria for choosing number of hidden layers and size of the hidden layer?* Stack Overflow. http://stackoverflow.com/questions/10565868/multi-layer-perceptron-mlp-architecture-criteria-for-choosing-number-of-hidde

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges. In *Taylor & Francis - International Journal of Digital Earth*. https://doi.org/10.1080/17538947.2016.1239771

Yang, J. (2021, October). *Healthcare expenditure as share of GDP UK 2020*. Statista.

https://www.statista.com/statistics/317708/healthcare-expenditure-as-a-share-of-gdp-in-the-united-kingdom/

Yang, J. (2022, January). *U.S. health spending as share of GDP 1960-2020*. Statista. https://www.statista.com/statistics/184968/us-health-expenditure-as-percent-of-gdp-since-1960/

Young, S., Schroeder, A., Garikapati, V., Fish, J., & Blumenthal, M. (2021). The Role of Mobility Data Hubs in an Integrated Decarbonized Transportation Future; The Role of Mobility Data Hubs in an Integrated Decarbonized Transportation Future. *IEEE Green Technologies Conference (GreenTech)*. https://doi.org/10.1109/GreenTech48523.2021.00100

Yourdon, E., & Constantine, L. (1979). *Fundamentals of a Discipline of Computer Program and System Design*. Prentice-Hall.

Yu, S., Muller, P., & Zomaya, A. (2017). Editorial: special issue on ``big data security and privacy''. *Elsevier*. https://doi.org/10.1016/j.dcan.2017.10.004

Yue, Z., Songzheng, Z., & Tianshi, L. (2011). Bayesian regularization BP Neural Network model for predicting oil-gas drilling cost. *2011 International Conference on Business Management and Electronic Information*, *2*, 483–487. https://doi.org/10.1109/ICBMEI.2011.5917952

Zaslow, J. (2002). *If TiVo Thinks You Are Gay, Here's How to Set It Straight*. The Wall Street Journal. http://www.wsj.com/articles/SB1038261936872356908

Zhang, L. (2014). A Framework to Model Big Data Driven Complex Cyber Physical Control Systems. *20th International Conference on Automation & Computing*.

Zhang, X., Gaddam, S., & Chronopoulos, A. T. (2016). Ceph Distributed File System Benchmarks on an Openstack Cloud. *2015 IEEE International Conference on Cloud Computing in Emerging Markets, CCEM 2015*. https://doi.org/10.1109/CCEM.2015.12

Zhaowei, L., Yunlong, Y., Jintao, M., Zhaocong, W., & Junmin, W. (2017). Performance Optimization of In-Memory File System in Distributed Storage System. *2017 IEEE International Conference on Networking, Architecture, and Storage, NAS 2017 - Proceedings*. https://doi.org/10.1109/NAS.2017.8026870

Zheng, L. (2006). Dynamic Security Labels And Noninterference. *Formal Aspects of Security and Trust*.

Zhou, B., & Pei, J. (2010). The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Springer - Knowledge and Information Systems*, *28*(1), 47–77. https://doi.org/10.1007/s10115-010-0311-2

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and Challenges. *Elsevier - Neurocomputing*, *237*, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026

Zimmer, M. (2008). *More On the "Anonymity" of the Facebook Dataset – It's Harvard College (Updated) | MichaelZimmer.org*. Internet Research Ethics, Privacy, Social Media. http://www.michaelzimmer.org/2008/10/03/more-on-the-anonymity-of-the-facebook-dataset-its-harvard-college/

APPENDIX

I. Publications

George E. Vranopoulos, Athanasios A. Triantafyllidis, & Konstantinos Lefteriotis. (2020). Big Data Variety, “Where Do we Stand”, An Overview of Big Data and the Variety Challenge. International Journal of Management and Applied Science, 6(3).

George E. Vranopoulos and A. Triantafyllidis, “Managerial & Technical Insight in Knowledge Management Challenges in Governing Corporate Knowledge,” 2017, doi: 2412-4044.

George E. Vranopoulos, A. Triantafyllidis, K. Chalkiadaki, and M. Tsoli, “Knowledge Management as an IT Policy,” 2017, doi: RW.26092017.5158.

George E. Vranopoulos, A. Trianatafylidis, and M. Yiannopoulos, “Putting ATM Cash Requirements into Context, ANN relation to Socioeconomic Events and Variety,” in BEFB 2016 - International Congress on Banking, Finance and Business 2016, 2016, pp. 526–545, doi: 212-4044.

II. IT Milestones over the Ages

Embarking on your trip towards today's IT ecosystem, the following milestones can be identified.

Antiquity: People started accumulating and storing information as soon as the first writing systems were invented. In today's RDMBS terms, orders master-detail records, records of shipments dating back to ancient Mesopotamian (Sumerians[11]) civilizations around 4000BC.

1950's: The first commercial computer is installed. IBM introduces the first device that can read data in a non-sequential manner.

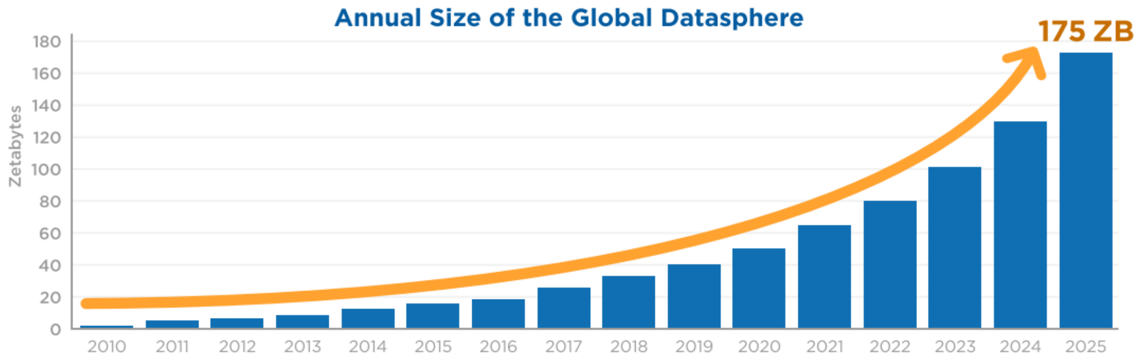
1960's: The first database systems using hierarchical (IMS) and network (CODASYL) models emerged.

1970's: The first relation databases are implemented. The ER entity-relationship model is developed.

1980's: SQL, Structured Query Language, becomes an industry standard.

1990's: Client tools (like Oracle Developer, PowerBuilder, VB) and productivity tools (like ODBC) are developed and released. The advancement and spread of the Internet boosted the usage of RDMBSs. Online business, the need for internet database connections resulted in creating open-source databases (like MySQL).

[11] Sumerian writing is the oldest full-fledged writing ever to be discovered by the archaeologists. It was primarily used for record keeping and had a decimal numeric system. (Smitha, 1998)



Source: Reinsel et al., 2018

Figure 50. IDC's Digital Universe Study

2000's: Although the Internet experienced a decline PDAs, mobile computing boosted the boundaries of RDBMSs even further. At the time, prevailing systems are provided by Oracle, Microsoft and IBM. For the first time in history, in 2007, the amount of data created exceeded the world's storage capacity (Manyika et al., 2011). As shown in Figure 50, data in the digital universe are exponentially increasing in the years to come (B. J. Gantz et al., 2012), and 2012 estimates are upwards scaled, reaching 175 zettabytes in the more recent report as of November 2018 (Reinsel et al., 2018).

The above periods can also be classified into four eras (Manyika et al., 2011), depicting the different adoption rates and respective impacts on productivity growth:

1959-1973: The "Mainframe" era. The productivity growth rate was high.

1973-1995: The "Minicomputers and PC's" era. Although there is a decline in productivity growth, companies have begun to spend IT budgets on distributed computing.

1995-2000: The "Internet and Web 1.0" era. Productivity growth regained its high rates, underpinned by a substantial increase in IT expenditure. Since there is a lag between IT investment and productivity, the most considerable portion of growth for this era is attributed to the previous era's investment.

2000-2006: The “Mobile Devices and Web 2.0” era. Due to the economic recession, IT budgets tend to decline; nonetheless, the investment of the past years has created momentum; thus, productivity growth rates remain high.

III. Big Data Programming Environments

There are several user interfaces, , and programming environments available to help developers utilise Big Data platforms. A non-exhaustive list is presented below.

- **Hive**: add the capability of writing SQL into Hadoop.
- **Pig**: a procedural data processing language for Hadoop.
- **R**: programming language/environment for statistical computing.
- **Cascading**: a workflow engine to build a series of Hadoop processing steps.
- **mrjob**: a framework that allows for rapid prototyping by allowing to write code for data processing and then transparently executing it locally, on Elastic MapReduce or a Hadoop cluster.
- **S4**: initially used by Yahoo! In ads placement was identified to be very useful in processing arbitrary stream of events.
- **Flume**: designed to make the data gathering process easy and scalable by running agents on client systems that pass data to aggregation collectors.
- **Azkaban**: is an open project from LinkedIn that workflows jobs and their independent steps, keeping track of logs, errors outputs etc.
- **Oozie**: is a job control system like Azkaban exclusively focused on Hadoop.
- **SOLR / ElasticSearch**: a library that handles indexing and searching of large collections of documents. Solr is mainly intended for the corporate environment for non-technical users, while ElasticSearch is primarily used by people in the “web world” with sufficient technical skills.
- **Datameer**: offers a simplified programming environment for users to define what they want and then utilises Hadoop infrastructure to convert it into MapReduce jobs.
- **BigSheets**: is an IBM implementation that allows non-technical users to gather unstructured data, and it uses Hadoop behind the scenes.

- **WEKA**: is a Java-based framework and Graphical User Interphase (GUI) for machine learning.
- **Mahout**: is an open-source framework for machine learning that can execute common algorithms on massive data sets, primarily by utilising Hadoop.
- **Sqoop**: is an open-source project for transferring data between relational databases and Hadoop.
- **scikts.learn**: is an easy-to-use, well-documented Python package offering high-level interphase to common machine learning techniques.

IV. DFD, ERD, ETL and NF

Data Flow Diagram (DFD)

The DFD illustrates the processing of data within a systems in terms of inputs and outputs and as its name indicates the primary focus is on the flow of information (smartdraw, 1994a). In 1993 SSADM[12] was accepted as the national standard for information systems development in Great Britain (CS Odessa Corp., 1993). An integral part of SSADM is DFD's in order to involve business users in a friendly and understandable way.

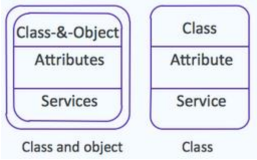
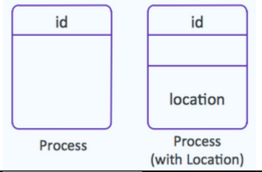
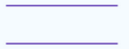
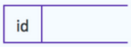
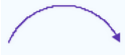
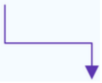
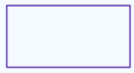
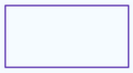
Data Flow Diagrams where fist described in 1979 by Larry Constantine and Ed Yourdon (Yourdon & Constantine, 1979). Their basic ingredients are outlined below (smartdraw, 1994a):

- **Process Nodes**: In these nodes a certain process is identified, that being a calculation, reform Etc. and is associated with data flows representing the required data for input and the outcome.
- **Datastore Nodes**: These are the data repositories of the system. The model was designed based on the concept of files and evolved in representing tables in the DBMS's. A further evolvement could be easily made into representing data sets in a Big Data environment.
- **Dataflows**: The pipeline through which information flow within the system.
- **External Entities**: Here systems outside the one under analysis are representation. These entities provide links with other systems and flows. With this concept, references are made to other "environment" / systems in an attempt to identify sources and destination and complete the picture of the system.

There are two different types of notations as far as visualisation of DFD's is concerned. The "Yourdon & Coad" which is mainly used for systems analysis and design and the "Gane & Sarson" which is most common in information system visualisation (smartdraw, 1994a). In Table 37 the respective symbols are outlined.

[12] Stands for Structured Systems Analysis and Design Method.

Table 37. DFD Notations

	Yourdon & Coad	**Gane & Sarson**
Process	 Class-&-Object Attributes Services Class Attribute Service Class and object Class	
Datastore		
Dataflow		
External Entity		

Source: CS Odessa Corp., 1993a

DFD's do not incorporate notions like parallelism or timing, but do provide with valuable information and most importantly with visual representation of the processes that data undergo within a system. At first glance there seems to be no evident link with Variety but if investigated closely there is. Variety is underlined by credibility of data and understanding of their origin. In a Big Data environment there are several transformations and calculations that take place which if documented properly could diminish the uncertainty of data. By utilising methodologies like the DFD it would be easy for Data Scientists to know how the data were manipulated, filtered, adjusted Etc . in order to decide if that is the best data set they could use for their analysis based on the credibility of the process and its outputs respectively.

Integrating information about the origin of data, their processing and results, could lift the burden from Data Scientists in validating and revalidating already validated result sets. Dealing with Variety, in respects to processing and methods used to validate, extrapolate, find outliers Etc., could be aided by documenting these processed in a visual manner such as the one outlined by the DFD's. A graphical representation can be easily understood by both business and IT users thus minimising confusion and ambiguity.

Entity Relationship Diagram (ERD)

ERD and its variations are frequently used in the conceptualisation of database tools whilst their concepts are incorporated in several design tools (Mohamed & Noordin, 2011). Tools like ErWin or Vision utilise this concept in depicting the DW design and eventually exposing its metadata in an XML format that DBMS can utilise (Sen, 2004). The word "Relationship" in this methodology is of importance since within such diagrams it is possible to identify and depict the relations between data. In relation to Variety such visualisation and metadata accompanying it, would aid in "putting data sets into context" and could be utilised in identifying conflicting or erroneous data.

ER Model was proposed by Peter Chan in 1976 (Chen, 1976) and since then refinements were applied by Charles Bachman and James Martin (smartdraw, 1994b). There are three basic ingredients in a standard ER Model (Lucid Software Inc., 2015b):

- **Entities**: Represent components of the structure like people, places, items, events or concepts. A certain degree of generalisation is required in identifying entities since instances, depending on the context, should not be represented as entities. For example, Teachers could be an entity but English Teachers, CIS Teachers Etc. and Mrs. X, Mr. Y so on so forth would be instances of the entity Teacher. In another case although Employees and Customers are people it's prudent to have two instances instead of aggregating into one People entity, mainly because of their different roles in the system and the difference in characteristics (attributes) that should be recorded.
- **Attributes**: Each and every entity has characteristics and qualities that should be represented in the system. These are collectively referred to as attributes. They are also known as data elements and are the data needed for the system to operate on. Name, date of birth, price, quantity and so on are examples of data that are stored in a system to facilitate operations and processing. When analysing entities and identifying their attributes it is important to include attributes that are related to the under analysis system or else data that are impossible to gather might be included. An example would be in a Student Course Registration system to have attributes of the Student entity related to their retail purchasing habits.
- **Relationships**: Nothing more than links, so simple but of outmost importance. Relationships represent the way two entities are interrelated. In understanding the

world and putting it into context relations play a very important role in understanding the whole picture.

As identified previously, of importance is the graphical representation of relationships. The concept behind relationships is cardinality which defines the relationship in terms of numbers. There are three main cardinal relationships (Biscobing, 2019):

- **One-to-one (1:1)**, which specifies that each instance of an entity is related with only one instance of another entity. Usually in the RDBMS world such relationships are "against" normalization and are usually merged into one structure. Of course this is not always the case since attributes of an entity could be quite many and are grouped together in entities with 1:1 relationships.
- **One-to-Many (1:∞)**, is the most common relationship were one instance of an entity is related to many instances of another entity. This relationship is also known as master-detail relationship, a very common example is the "Sales Order" in which the order has many products associated with it.
- **Many-to-Many (∞:∞)**, is also quite common in the business world but impossible to implement in an RDBMS. As a result when ∞:∞ relationships are identified, they are "broken" into two 1:∞ relationships with the introduction of an intermediate entity. An example would be the case of teachers and students where a teacher would have many students in a class and a student would attend several classes thus having many teachers.

The identifying components of an ERD are symbolised since graphical representation is enhancing understanding by visualisation of the system. The logic is represented in symbols that enable everybody on the team, Business and IT people, to see and understand the same things. Before presenting the common symbols of ERD's it is important to distinguish amongst the three variations of ERD's (Lucid Software Inc., 2015a):

- **Conceptual Data Model**, includes important entities and their relationships. It is the first level of analysis, quite aggregated and represent a broad view of what should be included in the system.
- **Logical Data Model**, is a step further in the analysis where more detail is added to the Conceptual Model. Attributes and their use in relationships are identified

by defining Primary Keys and Foreign Keys. In this step the rules of Normalization are also applied.

- **Physical Data Model**, the final step in which the Logical Model is tailored for deployment in a specific RDBMS environment. Structures like tables, columns, primary keys Etc. are depicted in an attempt to define the actual physical layer of the database implementation.

In Table 38, the processes / features identified with each step / model are summarised (Lucid Software Inc., 2015a).

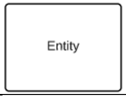
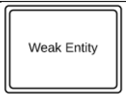
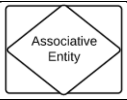
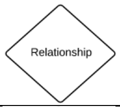
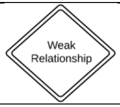
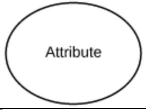
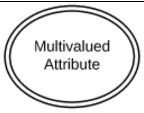
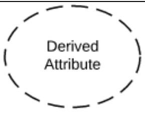
Table 38. Features in Data Models

Feature / Process	Conceptual	Logical	Physical
Entity Names	✓	✓	
Entity Relationships	✓	✓	
Attributes		✓	
Primary Keys		✓	✓
Foreign Keys		✓	✓
Table Names			✓
Column Names			✓
Column Datatypes			✓

Source: Lucid Software Inc. 2015a

In Big Data environments it is expected that the last step will be omitted or redefined since the structure is not so fidget and in many cases totally "hidden" from the user. The other two steps should be customised / enhanced in such a way that they can represent semi-structured and unstructured sets. Further advancement, could be the introduction of time in such models. In identifying Velocity and visually depicting it, it could provide essential information (in the form of metadata) against Variety. The first step in solving a problem is identifying it and if a new visualisation could present all the aspects of the Big Data data sets structure, it would be easier to identify and address them. Visualisation and graphical representation have been frequently referenced and stressed but no graphics have been presented. In Table 39 and Table 40 the most common graphical symbols of the methodology are presented (Lucid Software Inc., 2015a; smartdraw, 1994b).

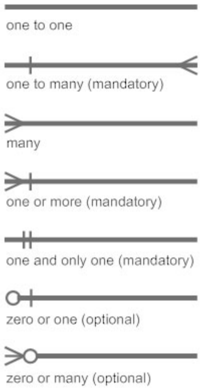
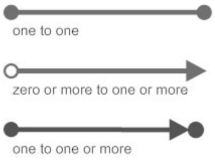
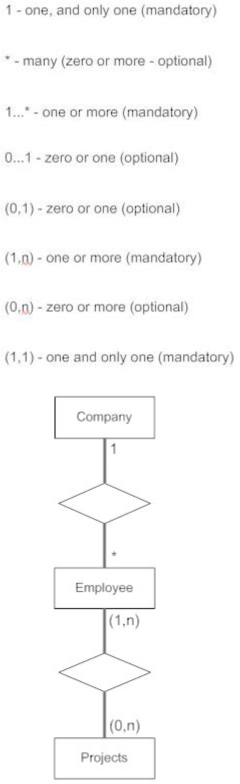
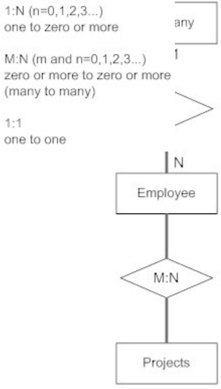
Table 39. ERD Notations

Entity Symbols	
	Strong Entities are the ones that can be uniquely identified by their attributes.
	Weak Entities are the ones that cannot be uniquely identified on their own and require the reference to attributes of another entity.
	Associative entities are not very commonly used and resemble the concept of inheritance. They are entities that associate the instance of one or more entity types.
Relationship Symbols	
	Strong Relationships that represent meaningful associations between entities.
	Weak Relationships, also known as Identifying Relationships, represent the connection between a weak entity and its owner.
Attributes Symbols	
	Attributes represent characteristics of an entity or a relation.
	Multivalued Attributes, are attributes capable of taking more than one values in any one instance of an entity.
	Derived Attributes are attributes that can be calculated based on other attributes. The most common example would be age which is calculated based on date of birth.

Source: smartdraw, 1994b

There are several styles of connector links in representing cardinality with the most common outlined in Table 40:

Table 40. Most Common ERD Cardinality Links Representations

Information Engineering Style (Crows Foot)	Buchman Style	Martin Style	Chen Style
 one to one one to many (mandatory) many one or more (mandatory) one and only one (mandatory) zero or one (optional) zero or many (optional)	 one to one zero or more to one or more one to one or more	 1 - one, and only one (mandatory) * - many (zero or more - optional) 1...* - one or more (mandatory) 0...1 - zero or one (optional) (0,1) - zero or one (optional) (1,n) - one or more (mandatory) (0,n) - zero or more (optional) (1,1) - one and only one (mandatory) Company 1 * Employee (1,n) (0,n) Projects	 1:N (n=0,1,2,3...) one to zero or more any 1 M:N (m and n=0,1,2,3...) zero or more to zero or more (many to many) 1:1 one to one N Employee M:N Projects

In tackling Variety, "Relationship" is important in identifying context and data variations. Although ERD's current format and visualisation does not effectively depict semi-structured and unstructured data sets, via adjustments in methodology, the concept of relationships identification and visualisation can help in understanding the Big Data lakes which otherwise would be "black boxes" for Data Scientists.

Normalization

Normalization was initially described by Codd in 1970, in his attempt to provide with a reliable method to maximise data independence and minimise data inconsistency (Codd, 1970).

The normalization process has 4 stages in which we progress sequentially. In order to continue to the next NF the schema must already satisfy all requirements of previous NF's. The following definitions are offered in understanding the steps leading to a normalized DB (Chapple, 2022; Codd, 1974; Studytonight, 2014; tutorialspoint.com, 2015).

1st Normal Form: All attributes in a relation must have atomic domains - that is - values in an atomic domain are indivisible units.

2nd Normal Form: None-prime[13] attributes must be functionally dependent on prime[14] key attributes.

3rd Normal Form: No none-prime attribute is transitively dependant on prime key attribute. Boyce-Codd Normal form (BCNF) is an extension to the 3rd NF, sometimes referred to as 3½NF, in which there shouldn't be any overlapping prime-key attributes.

4th Normal Form: Was introduced by Ronal Fagin in 1977 and states that all multivalued dependencies should be eliminated. That is, elimination of the presence of one or more instances of an entity that imply the presence of one or more instances of the same entity.

5th Normal Form: Was introduced by Ronal Fagin in 1979 and states that every non-trivial joint dependency is implied by the candidate keys.

[13] None-prime attributes are attributes that do not uniquely identify the instance of an entity. In DBMS terms none-prime attributes are the ones not included in the primary key.

[14] Prime key attributes are attributes that uniquely identify the instance of an entity. In DBMS terms prime attributes are the ones included in the primary key.

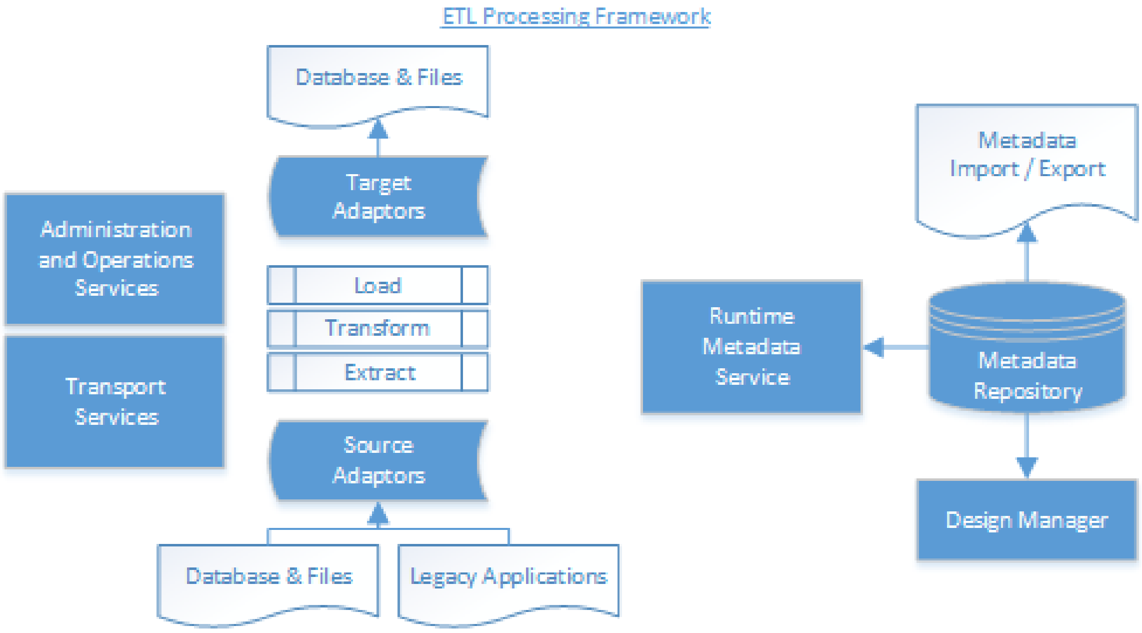
6th Normal Form: Was introduced by Date Christopher and its definition was revised in 2014. It states that every join dependency must be trivial – where a join dependency is trivial if and only if one of its components is equal to the pertinent heading in its entirety.

From the mentioned NF's the most common stage for DBMS's is the 3rd NF or the BCNF, the rest are seldom used since they pose elaborate and in many cases unnecessary fragmentation of the data.

While normalization makes databases maintenance easy, complexity is added since data are "spread out" in many structures (Vangie, 2015). Experienced analysts and IT professionals tend trade-off fully normalized databases for simplicity and performance, database design is an art (Agrawal et al., 2011), thus "forgiving" deviations from the rules. In many cases data duplication and deformalized structures are utilised, of course with caution not to make the database inconsistent, in grouping related data and storing aggregations.

A Big Data infrastructure is by definition denormalized since the data are usually stored in the way they are acquired from the respective data sources. Although structuring such a repository by applying the Normalization process would be unpractical if not impossible, there are lessons to be learned. Normalization is not only the way data are eventually structured. It is a way of thinking. Decomposition of data streams, investigation of dependencies and relations are techniques utilised while navigating though the NF's. The same techniques could be utilised in the Big Data environments not in an attempt to define structure but in view of understanding them and minimise the effects of Variety.

With the use of the "logic behind Normalization" data sets could be documented in identifying dependencies which can be possible sources for inconsistencies and erroneous data. Also dependencies can lead to identifying possible sources of data completion since transitively depended data might exist in one data set that could be complementary to another. By identifying all possible dependencies, context and data interrelations, the "lake" can be mapped and thus Variety issues posed by such challenges could be addressed. A "bigger picture" of the data lakes and their associations would assist Data Scientists in utilising and integrating different set in their analysis.



Mulherrin & Abdul-hamid 2009

Figure 51. ETL Processing Framework

<u>The ETL Framework</u>

The ETL Framework (shown in Figure 51) is composed of several components of which the most important are the following (Mulherrin & Abdul-hamid, 2009):

- **Extraction**: The process of “getting” the data from the sources systems, usually the OLTP company systems.
- **Transformation**: The process where extracted data are pre-processed in order to be integrated into the DW. This process is the most complicated since it entails functions like data validation, schema conversion, business rules application, data accuracy validation, data-type conversion Etc.
- **Load**: The process where data are actually imported in to the DW. In this process simple validations are applied in order to cater for transformation efficiency.
- **Metadata**: Information about the ETL process is stored in order to be available during run-time. The “data about data” actually describes what has to be extracted and from where, how should it be transformed and where should it be deposited.
- **Admin and Transport service**: Scheduling and underling technology (eg network infrastructure, File Transfer Protocols Etc.) are utilised though out the ETL process in order to govern execution and monitoring.

V. “Booster Metrics” Configuration sample XML

Please note any BIN provided below is a random set of numbers

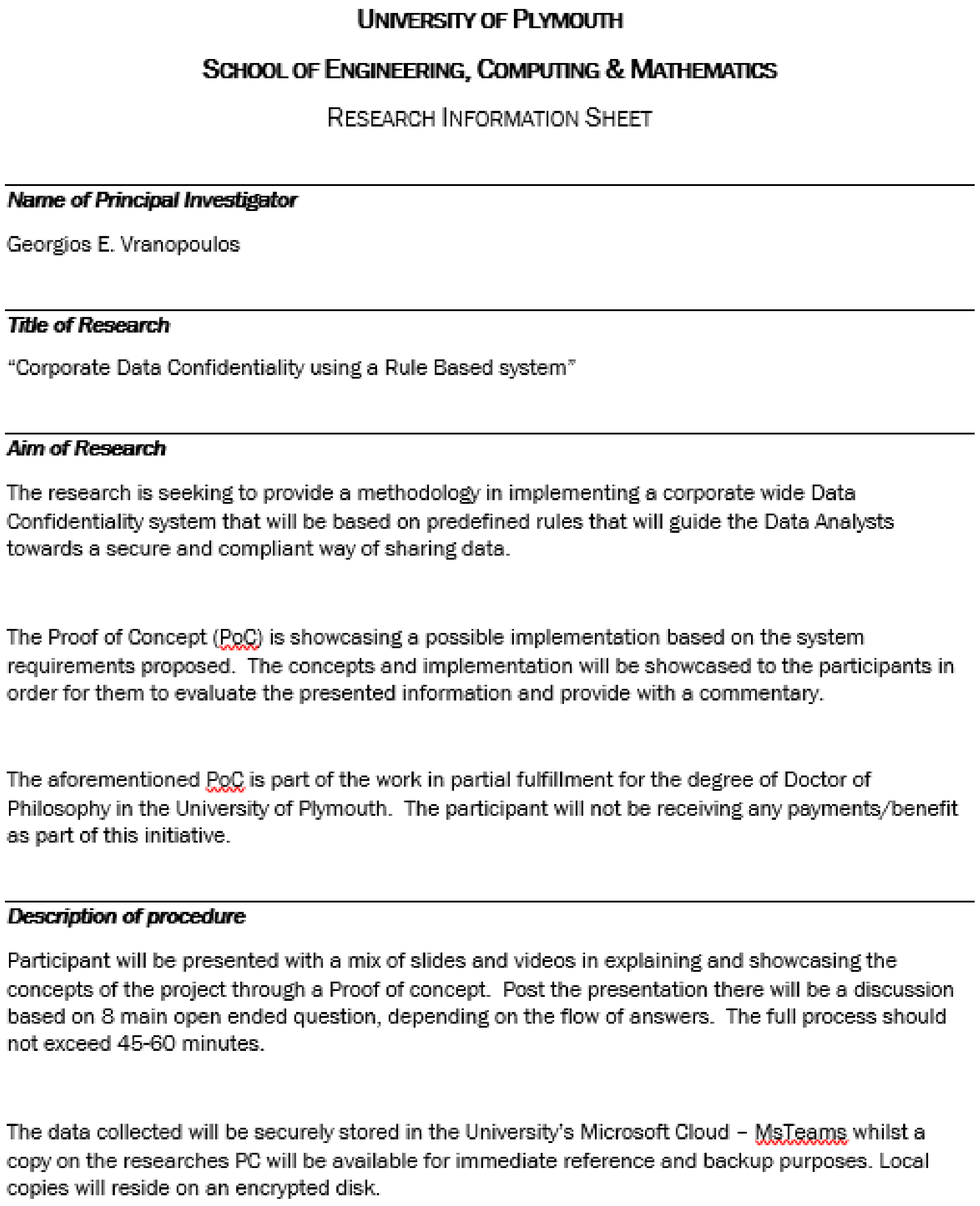
```
<?xml version="1.0"?>

<ConfidentialityConfiguration>
    <DebugMode>false</DebugMode>
    <DefaultEncoding>ISO-8859-7</DefaultEncoding>
    <OutputFileName>ConfidentialityResults.csv</OutputFileName>
    <DefaultMaskingCharacter>*</DefaultMaskingCharacter>
    <EncryptionKey>PhD-GV-PoC-12345</EncryptionKey>
    <DisplayPlainInfo>true</DisplayPlainInfo>
    <DisplayMaskedInfo>true</DisplayMaskedInfo>
    <DisplayHashedInfo>false</DisplayHashedInfo>
    <DisplayReplacedInfo> false</DisplayReplacedInfo>
    <DisplayEncryptedInfo>false</DisplayEncryptedInfo>
    <ReplaceFileExtension>.rpl</ReplaceFileExtension>
    <TextProximity>25</TextProximity>
    <BlockMaxLines>10</BlockMaxLines>

    <Config description = "Credit Cards" classification = "Card">
        <RegExs>
            <Entry>
                <Dscr>MasterCard CC</Dscr>
                <regEx>(?:5[1-5][0-9]{2}|222[1-9]|22[3-9][0-9]|2[3-6][0-9]{2}|27[01][0-9]|2720)[0-9]{12}</regEx>
            </Entry>
            <Entry>
                <Dscr>Visa CC</Dscr>
                <regEx>4[0-9]{12}(?:[0-9]{3})?</regEx>
            </Entry>
            <Entry>
                <Dscr>American Express CC</Dscr>
                <regEx>3[47][0-9]{13}</regEx>
            </Entry>
            <Entry>
                <Dscr>Diners Club CC</Dscr>
                <regEx>3(?:0[0-5]|[68][0-9])[0-9]{11}</regEx>
            </Entry>
            <Entry>
                <Dscr>Discover CC</Dscr>
                <regEx>6(?:011|5[0-9]{2})[0-9]{12}</regEx>
            </Entry>
            <Entry>
                <Dscr>JCB CC</Dscr>
                <regEx>(?:2131|1800|35\d{3})\d{11}</regEx>
            </Entry>
        </RegExs>
    <MyBins>
        <Entry>42345</Entry>
        <Entry>34536</Entry>
    </MyBins>
    <Masking>
        <ShowFromStart>6</ShowFromStart>
        <ShowFromEnd>3</ShowFromEnd>
    </Masking>
    <!-- Replacement Options : Truncate (Default) - Mask - Hash - Encrypt -->
    <ReplacementOption>Mask</ReplacementOption>
    <ReplacementStrengthBoarder>50</ReplacementStrengthBoarder>
    <AlgorithmicStrength>
        <BoundTextPercent>50</BoundTextPercent>
        <UnoundTextPercent>40</UnoundTextPercent>
        <LuhnAddOn>40</LuhnAddOn>
        <inMyBinsAddon>5</inMyBinsAddon>
    </AlgorithmicStrength>
</Config>
<Config description = "Debit Cards" classification = "Card">
    <RegExs>
        <Entry>
            <Dscr>Debit Card</Dscr>
            <regEx>[4,5]{1}[0-9]{15}</regEx>
        </Entry>
    </RegExs>
    <MyBins>
        <Entry>465412</Entry>
        <Entry>519260</Entry>
    </MyBins>
    <Masking>
        <ShowFromStart>6</ShowFromStart>
        <ShowFromEnd>3</ShowFromEnd>
    </Masking>
    <!-- Replacement Options : Truncate (Default) - Mask - Hash - Encrypt -->
    <ReplacementOption>Mask</ReplacementOption>
    <ReplacementStrengthBoarder>50</ReplacementStrengthBoarder>
    <AlgorithmicStrength>
        <BoundTextPercent>50</BoundTextPercent>
        <UnoundTextPercent>40</UnoundTextPercent>
        <LuhnAddOn>40</LuhnAddOn>
        <inMyBinsAddon>5</inMyBinsAddon>
    </AlgorithmicStrength>
</Config>

<Config description = "IDs" classification = "List">
    <RegExs>
        <Entry>
            <Dscr>Kuwaiti Civil Id</Dscr>
```

```
                <regEx>(1|2|3)((\d{2}((0[13578]|1[02])(0[1-
9]|[12]\d|3[01])|(0[13456789]|1[012])(0[1-9]|[12]\d|30)|02(0[1-9]|1\d|2[0-
8])))|([02468][048]|[13579][26])0229)(\d{5})</regEx>
            </Entry>
         -->
            <Entry>
                <Dscr>Gulf Civil ID</Dscr>
                <regEx>\d{1} (?!00)\d{2} (?!00)\d{2} (?!00)\d{2}
(?!0000)\d{4}</regEx>
            </Entry>
            <Entry>
                <Dscr>Danish Civil ID</Dscr>
                <regEx>[0-3][0-9][0-1]\d{3}-\d{4}</regEx>
            </Entry>
            <Entry>
                <Dscr>UK Passport</Dscr>
                <regEx>[0-9]{10}GBR[0-9]{7}[U,M,F]{1}[0-9]{9}</regEx>
            </Entry>
            <Entry>
                <Dscr>International Passport</Dscr>
                <regEx>[A-Z0-9&lt;]{9}[0-9]{1}[A-Z]{3}[0-9]{7}[A-Z]{1}[0-
9]{7}[A-Z0-9&lt;]{14}[0-9]{2}</regEx>
            </Entry>
            <Entry>
                <Dscr>Indian Passport</Dscr>
                <regEx>[A-Z]{1}-[0-9]{7}</regEx>
            </Entry>
            <Entry>
                <Dscr>Greek Civil ID</Dscr>
                <regEx>[A-Ω]{1,2}[0-9]{6}</regEx>
            </Entry>
            <Entry>
                <Dscr>AD User Name</Dscr>
                <regEx>(r|sd|op)[1-9]{1}[0-9]{1,4}</regEx>
            </Entry>
            <Entry>
                <Dscr>Social Security Number</Dscr>
                <regEx>(\d{3}-\d{2}-\d{4})|(\d{3}\d{2}\d{4})$</regEx>
            </Entry>
        </RegExs>
        <TextProximityEnhancers>
            <Entry>Passport</Entry>
            <Entry>Civil</Entry>
            <Entry>Identity</Entry>
            <Entry>User</Entry>
            <Entry>Seller</Entry>
            <Entry>Party</Entry>
        </TextProximityEnhancers>
        <Masking>
            <ShowFromStart>2</ShowFromStart>
            <ShowFromEnd>1</ShowFromEnd>
        </Masking>
        <!-- Replacement Options : Truncate (Default) - Mask - Hash - Encrypt -->
        <ReplacementOption>Mask</ReplacementOption>
        <ReplacementStrengthBoarder>50</ReplacementStrengthBoarder>
        <AlgorithmicStrength>
            <BoundTextPercent>50</BoundTextPercent>
            <UnoundTextPercent>40</UnoundTextPercent>
            <ProximityAddOn>10</ProximityAddOn>
        </AlgorithmicStrength>
    </Config>
    <Config description = "Accounts" classification = "List">
        <RegExs>
            <Entry>
                <Dscr>IBAN</Dscr>
                <regEx>[a-zA-Z]{2}[0-9]{2}[a-zA-Z0-9]{4}[0-9]{7}([a-zA-Z0-
9]?){0,16}</regEx>
            </Entry>
        </RegExs>
        <TextProximityEnhancers>
            <Entry>IBAN</Entry>
            <Entry>Product</Entry>
        </TextProximityEnhancers>
        <Masking>
            <ShowFromStart>2</ShowFromStart>
            <ShowFromEnd>4</ShowFromEnd>
        </Masking>
        <!-- Replacement Options : Truncate (Default) - Mask - Hash - Encrypt -->
        <ReplacementOption>Mask</ReplacementOption>
        <ReplacementStrengthBoarder>50</ReplacementStrengthBoarder>
        <AlgorithmicStrength>
            <BoundTextPercent>50</BoundTextPercent>
            <UnoundTextPercent>40</UnoundTextPercent>
            <ProximityAddOn>10</ProximityAddOn>
        </AlgorithmicStrength>
    </Config>
    <Config description = "IPs" classification = "List">
        <RegExs>
            <Entry>
                <Dscr>IPv4</Dscr>
                <regEx>([01]?[0-9]{1,2}|2[0-4][0-9]|25[0-5])\.([01]?[0-9]{1,2}|2[0-
4][0-9]|25[0-5])\.([01]?[0-9]{1,2}|2[0-4][0-9]|25[0-5])\.([01]?[0-9]{1,2}|2[0-4][0-
9]|25[0-5])</regEx>
            </Entry>
            <Entry>
                <Dscr>IPv6</Dscr>
                <regEx>(([0-9a-fA-F]{1,4}:){7,7}[0-9a-fA-F]{1,4}|([0-9a-fA-
F]{1,4}:){1,7}:|([0-9a-fA-F]{1,4}:){1,6}:[0-9a-fA-F]{1,4}|([0-9a-fA-
F]{1,4}:){1,5}(:[0-9a-fA-F]{1,4}){1,2}|([0-9a-fA-F]{1,4}:){1,4}(:[0-9a-fA-
F]{1,4}){1,3}|([0-9a-fA-F]{1,4}:){1,3}(:[0-9a-fA-F]{1,4}){1,4}|([0-9a-fA-
F]{1,4}:){1,2}(:[0-9a-fA-F]{1,4}){1,5}|[0-9a-fA-F]{1,4}:((:[0-9a-fA-
F]{1,4}){1,6})|:((:[0-9a-fA-F]{1,4}){1,7}|:)|fe80:(:[0-9a-fA-F]{0,4}){0,4}%[0-9a-zA-
Z]{1,}|::(ffff(:0{1,4}){0,1}:){0,1}((25[0-5]|(2[0-4]|1{0,1}[0-9]){0,1}[0-
9])\.){3,3}(25[0-5]|(2[0-4]|1{0,1}[0-9]){0,1}[0-9])|([0-9a-fA-F]{1,4}:){1,4}:((25[0-
5]|(2[0-4]|1{0,1}[0-9]){0,1}[0-9])\.){3,3}(25[0-5]|(2[0-4]|1{0,1}[0-9]){0,1}[0-
9]))</regEx>
            </Entry>
            <Entry>
                <Dscr>MAC IEEE802</Dscr>
                <regEx>([0-9A-Fa-f]{2}[:-]){5}([0-9A-Fa-f]{2})</regEx>
            </Entry>
        </RegExs>
        <TextProximityEnhancers>
            <Entry>Client</Entry>
            <Entry>Server</Entry>
            <Entry>IP</Entry>
            <Entry>Mac</Entry>
        </TextProximityEnhancers>
        <Masking>
            <ShowFromStart>2</ShowFromStart>
            <ShowFromEnd>0</ShowFromEnd>
        </Masking>
        <!-- Replacement Options : Truncate (Default) - Mask - Hash - Encrypt -->
        <ReplacementOption>Mask</ReplacementOption>
        <ReplacementStrengthBoarder>50</ReplacementStrengthBoarder>
        <AlgorithmicStrength>
            <BoundTextPercent>60</BoundTextPercent>
            <UnoundTextPercent>50</UnoundTextPercent>
            <ProximityAddOn>30</ProximityAddOn>
        </AlgorithmicStrength>
    </Config>
    <Config description = "Personal Data" classification = "List">
        <RegExs>
            <Entry>
                <Dscr>eMail</Dscr>
                <regEx>(?:[a-z0-9!#$%&amp;'*+/=?^_`{|}~-]+(?:\.[a-z0-
9!#$%&amp;'*+/=?^_`{|}~-]+)*|\"(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-
\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])*\")@(?:(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-
z0-9](?:[a-z0-9-]*[a-z0-9])?|\[(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\.){3}(?:25[0-
5]|2[0-4][0-9]|[01]?[0-9][0-9]?|[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-
\x5a\x53-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])</regEx>
            </Entry>
            <Entry>
                <Dscr>US Address</Dscr>

    <regEx>(\d{2,5}\s+)(?![a|p]m\b)(NW|NE|SW|SE|north|south|west|east|n|e|s|w)?([\s|
\,|\.]+)?(([a-zA-Z|\s+]{1,30}){1,4})([0-
9|\s+]{1,3})?(court|ct|street|st|drive|dr|lane|ln|road|rd|blvd|ave|ter)(\,| )*[a-z A-Z]*(\,|
)*((AK|AL|AR|AZ|CA|CO|CT|DC|DE|FL|GA|GU|HI|IA|ID|IL|IN|KS|KY|LA|MA|MD|
ME|MI|MN|MO|MS|MT|NC|ND|NE|NH|NJ|NM|NV|NY|OH|OK|OR|PA|RI|SC|SD|TN|
TX|UT|VA|VI|VT|WA|WI|WV|WY)|[a-zA-Z])*(\,| )(\d{2,5}\s*)</regEx>
            </Entry>
            <Entry>
                <Dscr>Address</Dscr>
                <regEx>(\d{1,}) [a-zA-Z0-9\s]+(\.)? [a-zA-Z]+(\,)? [A-Z]{2} [0-
9]{5,6}</regEx>
            </Entry>
            <Entry>
                <Dscr>Street Address</Dscr>
                <regEx>\d+[ ]{1}(?:[A-Za-z0-9.-]+[
]{1}?)+(?:Avenue|Lane|Road|Boulevard|Drive|Street|Ave|Dr|Rd|Blvd|Ln|St)\.?</regEx>
            </Entry>
        </RegExs>
        <TextProximityEnhancers>
            <Entry>Mail</Entry>
            <Entry>Address</Entry>
            <Entry>Street</Entry>
            <Entry>Zip</Entry>
            <Entry>Block</Entry>
            <Entry>Plot</Entry>
            <Entry>Addr</Entry>
            <Entry>City</Entry>
        </TextProximityEnhancers>
        <Masking>
            <ShowFromStart>0</ShowFromStart>
            <ShowFromEnd>0</ShowFromEnd>
        </Masking>
        <!-- Replacement Options : Truncate (Default) - Mask - Hash - Encrypt -->
        <ReplacementOption>Truncate</ReplacementOption>
        <ReplacementStrengthBoarder>50</ReplacementStrengthBoarder>
        <AlgorithmicStrength>
            <BoundTextPercent>50</BoundTextPercent>
            <UnoundTextPercent>40</UnoundTextPercent>
            <ProximityAddOn>10</ProximityAddOn>
        </AlgorithmicStrength>
    </Config>
    <Config description = "Security Data" classification = "List">
        <RegExs>
            <Entry>
                <Dscr>Password</Dscr>
                <regEx></regEx>
            </Entry>
            <Entry>
                <Dscr>PIN</Dscr>
                <regEx></regEx>
            </Entry>
            <Entry>
                <Dscr>RSA</Dscr>
                <regEx></regEx>
            </Entry>
        </RegExs>
        <TextProximityEnhancers>
            <Entry>Password</Entry>
            <Entry>PIN</Entry>
            <Entry>pwd</Entry>
            <Entry>RSA</Entry>
            <Entry>Security</Entry>
        </TextProximityEnhancers>
        <Masking>
            <ShowFromStart>2</ShowFromStart>
            <ShowFromEnd>0</ShowFromEnd>
        </Masking>
        <!-- Replacement Options : Truncate (Default) - Mask - Hash - Encrypt -->
        <ReplacementOption>Truncate</ReplacementOption>
        <ReplacementStrengthBoarder>40</ReplacementStrengthBoarder>
        <AlgorithmicStrength>
            <BoundTextPercent>50</BoundTextPercent>
            <UnoundTextPercent>40</UnoundTextPercent>
            <ProximityAddOn>10</ProximityAddOn>
        </AlgorithmicStrength>
    </Config>

    <XMLConfig description = "Absolute XML tags" classification =
"XMLAbsolute">
        <ReplacementStrengthBoarder>50</ReplacementStrengthBoarder>
        <!-- Replacement Options : Truncate (Default) - Mask - Hash - Encrypt -
Replace -->
        <Tags>
            <Entry>
                <Dscr>Customer FullName</Dscr>
                <Tag>FullName</Tag>
                <RelOpt>Replace</RelOpt>
                <ShowFromStart>0</ShowFromStart>
                <ShowFromEnd>0</ShowFromEnd>
                <Replacement>Customer Name</Replacement>
                <RegExCandidate>false</RegExCandidate>
            </Entry>
            <Entry>
                <Dscr>Customer ShortName</Dscr>
                <Tag>ShortName</Tag>
                <RelOpt>Mask</RelOpt>
                <ShowFromStart>2</ShowFromStart>
                <ShowFromEnd>2</ShowFromEnd>
                <Replacement>Customer Name</Replacement>
                <RegExCandidate>false</RegExCandidate>
            </Entry>
            <Entry>
                <Dscr>Civil Id</Dscr>
                <Tag>CIVIL_ID</Tag>
                <RelOpt>Mask</RelOpt>
                <ShowFromStart>2</ShowFromStart>
                <ShowFromEnd>2</ShowFromEnd>
                <Replacement>Customer ID</Replacement>
                <RegExCandidate>true</RegExCandidate>
            </Entry>
        </Tags>
    </XMLConfig>
    <XMLConfig description = "Relative XML tags" classification = "XMLRelative">
        <ReplacementStrengthBoarder>50</ReplacementStrengthBoarder>
        <Tags>
            <Entry>
                <Dscr>General Names</Dscr>
                <Tag>Name</Tag>
                <RelOpt>Replace</RelOpt>
                <ShowFromStart>0</ShowFromStart>
                <ShowFromEnd>0</ShowFromEnd>
                <Replacement></Replacement>
                <RegExCandidate>false</RegExCandidate>
            </Entry>
            <Entry>
                <Dscr>Passport Number</Dscr>
                <Tag>Passport</Tag>
                <RelOpt>Mask</RelOpt>
                <ShowFromStart>1</ShowFromStart>
                <ShowFromEnd>3</ShowFromEnd>
                <Replacement></Replacement>
                <RegExCandidate>true</RegExCandidate>
            </Entry>
        </Tags>
        <AlgorithmicStrength>
            <ExactMatch>100</ExactMatch>
            <ApproximateMatch>50</ApproximateMatch>
        </AlgorithmicStrength>
    </XMLConfig>

    <BlockConfig description = "WireShark Data Block" classification = "TXT">
        <BlockFromColumn>56</BlockFromColumn>
        <BlockWidth>16</BlockWidth>
        <BlockTerminator></BlockTerminator>
        <BlockWindoeLines>10</BlockWindoeLines>
    </BlockConfig>
    <BlockConfig description = "WireShark Data Block Hex" classification = "HEX">
        <BlockFromColumn>6</BlockFromColumn>
        <BlockWidth>48</BlockWidth>
        <BlockTerminator></BlockTerminator>
        <BlockWindoeLines>10</BlockWindoeLines>
    </BlockConfig>
</ConfidentialityConfiguration>
```

VI. Dataset Characterisation sample JSON configuration

```
{ "DebugMode" : false
, "DefaultEncoding" : "ISO-8859-1"
, "delimiters" :  [ { "regEx" : ";", "Dscr" : "Semicolon" }
                  , { "regEx" : ",", "Dscr" : "Comma" }
                  , { "regEx" : "\~\|\~", "Dscr" : "TildePipeTilde" }
                  , { "regEx" : "\~\|\|\~", "Dscr" : "TildePipePipeTilde" }
                  , { "regEx" : "\~", "Dscr" : "Tilde" }
                  , { "regEx" : "\t", "Dscr" : "Tab" }
                  ]
, "LinesToReadFromStartOfFile" : 100
, "LinesOutputFileName" : "DelimiterLineResults"
, "StatsOutputFileName" : "DelimiterStatResults"
, "FileInfoOutputFileName" : "DelimiterFileInfo"
, "OutputFileExtension" : ".txt"
, "FileSizeLimit" : 104857600
}
```

VII. Information Form



University of Plymouth

School of Engineering, Computing & Mathematics

Research Information Sheet

Name of Principal Investigator

Georgios E. Vranopoulos

Title of Research

"Corporate Data Confidentiality using a Rule Based system"

Aim of Research

The research is seeking to provide a methodology in implementing a corporate wide Data Confidentiality system that will be based on predefined rules that will guide the Data Analysts towards a secure and compliant way of sharing data.

The Proof of Concept (PoC) is showcasing a possible implementation based on the system requirements proposed. The concepts and implementation will be showcased to the participants in order for them to evaluate the presented information and provide with a commentary.

The aforementioned PoC is part of the work in partial fulfillment for the degree of Doctor of Philosophy in the University of Plymouth. The participant will not be receiving any payments/benefit as part of this initiative.

Description of procedure

Participant will be presented with a mix of slides and videos in explaining and showcasing the concepts of the project through a Proof of concept. Post the presentation there will be a discussion based on 8 main open ended question, depending on the flow of answers. The full process should not exceed 45-60 minutes.

The data collected will be securely stored in the University's Microsoft Cloud – MsTeams whilst a copy on the researches PC will be available for immediate reference and backup purposes. Local copies will reside on an encrypted disk.

Description of risk

The confidentiality and privacy of the data collected will be safeguarded since interviews will be contacted in person. Recording will happen only with the explicit consent of the interviewee which will be affirmed with an explicit statement in the beginning of the discussion after the presentation. In case of the no consent given, only the refusal will be recorded and the interview will be terminated while the interviewee will be excluded from the project results The data collected will only be used for the project and therefore will not be shared with anyone outside the project. The recordings will be destroyed a quarter post project completion which is expected to be in early 2023.

Benefits of proposed research

The approach is seeking to standardize the functions of Data Confidentiality along with a proposition of a system to automate the procedures. In this way the implementer of the methodology will be able to centrally handle Data Confidentiality thus minimizing any Risks of Disclosure and centrally manage compliance to any regulatory requirements.

Right to withdraw

All participant are volunteers; therefore you are free to withdraw in the first stage of the project, namely the data collection when the interviews are being conducted. The timeframe of any interviewee to withdraw, with the use of an official email in the bellow mentioned contact, will be acceptable for two (2) weeks from the date of the scheduled interview or the signing of the consent for; whichever is later.

If you are dissatisfied with the way the research is conducted, please contact the principal investigator in the first instance: e-mail: Georgios.vranopoulos@postgrad.plymouth.ac.uk. If you feel the problem has not been resolved please contact the secretary to the Faculty of Science and Engineering Research Ethics & Integrity Committee:

scienghumanethucs@plymouth.ac.uk

The objectives of this research have been explained to me.

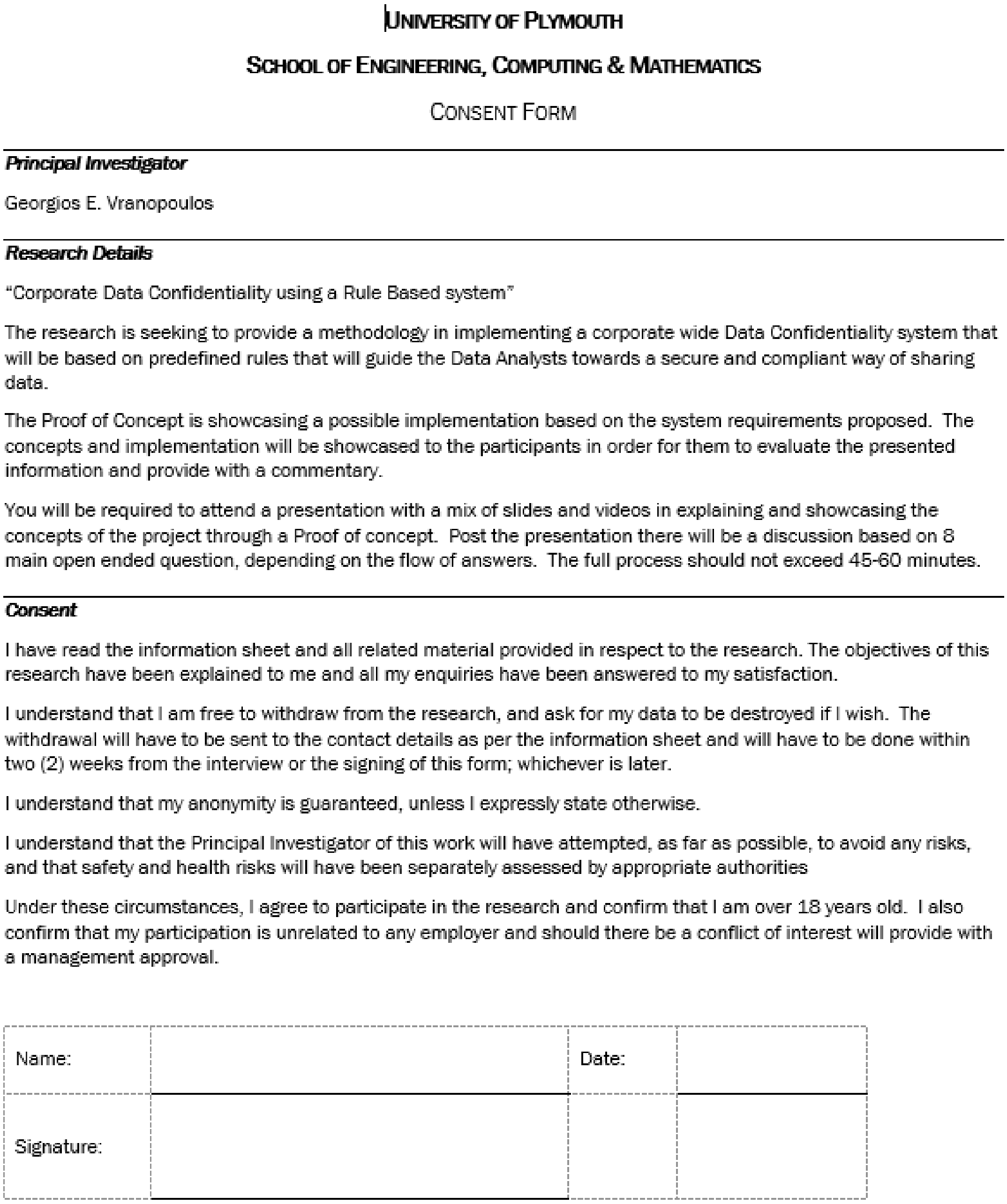
I understand that I am free to withdraw from the research at any stage, and ask for my data to be destroyed if I wish.

I understand that my anonymity is guaranteed, unless I expressly state otherwise.

I understand that the Principal Investigator of this work will have attempted, as far as possible, to avoid any risks, and that safety and health risks will have been separately assessed by appropriate authorities

Under these circumstances, I agree to participate in the research and confirm that I am over 18 years old.

VIII. Consent Form



UNIVERSITY OF PLYMOUTH

SCHOOL OF ENGINEERING, COMPUTING & MATHEMATICS

CONSENT FORM

Principal Investigator

Georgios E. Vranopoulos

Research Details

“Corporate Data Confidentiality using a Rule Based system”

The research is seeking to provide a methodology in implementing a corporate wide Data Confidentiality system that will be based on predefined rules that will guide the Data Analysts towards a secure and compliant way of sharing data.

The Proof of Concept is showcasing a possible implementation based on the system requirements proposed. The concepts and implementation will be showcased to the participants in order for them to evaluate the presented information and provide with a commentary.

You will be required to attend a presentation with a mix of slides and videos in explaining and showcasing the concepts of the project through a Proof of concept. Post the presentation there will be a discussion based on 8 main open ended question, depending on the flow of answers. The full process should not exceed 45-60 minutes.

Consent

I have read the information sheet and all related material provided in respect to the research. The objectives of this research have been explained to me and all my enquiries have been answered to my satisfaction.

I understand that I am free to withdraw from the research, and ask for my data to be destroyed if I wish. The withdrawal will have to be sent to the contact details as per the information sheet and will have to be done within two (2) weeks from the interview or the signing of this form; whichever is later.

I understand that my anonymity is guaranteed, unless I expressly state otherwise.

I understand that the Principal Investigator of this work will have attempted, as far as possible, to avoid any risks, and that safety and health risks will have been separately assessed by appropriate authorities

Under these circumstances, I agree to participate in the research and confirm that I am over 18 years old. I also confirm that my participation is unrelated to any employer and should there be a conflict of interest will provide with a management approval.

Name:		Date:	
Signature:			

IX. BD-CPS PoC Entities Data Samples

Data Set

DataSetDscr
Credit Card Transactions
EOD Account Balances
Customer Master
Mobile Clickstream

Data Set Fields

DataSet	FieldID	FieldName	Field Entity
Credit Card Transactions	00001	Card Number	Credit Card Number
Credit Card Transactions	00002	Merchant Name	Party Name
Credit Card Transactions	00003	Transaction Amount	Transaction Amount
Credit Card Transactions	00004	Terminal Number	Terminal Number
Credit Card Transactions	00005	Terminal Geolocation	Geolocation
Credit Card Transactions	00006	CID Magnetic Block	Magnetic Stripe Data
EOD Account Balances	00007	Account Number	IBAN Account Number
EOD Account Balances	00008	Balance	Balance
Customer Master	00009	Customer Name	Party Name
Customer Master	00010	Birth Date	Party Date of Birth
Customer Master	00011	Nationality	Party Nationality
Customer Master	00012	Home Address	Party Home Address Street
Customer Master	00013	Mobile Number	Party Telephone
Customer Master	00014	email	Party eMail
Customer Master	00015	Work Address	Party Work Address Street
Customer Master	00016	Home Telephone Number	Party Telephone
Customer Master	00017	Work Telephone Number	Party Telephone
Customer Master	00018	Identification Type	Party ID Type
Customer Master	00019	Identification Number	Party ID Number
Mobile Clickstream	00020	Customer ID	Party System ID
Mobile Clickstream	00021	IP Address	Network Information
Mobile Clickstream	00022	Form	Application Log Elements
Mobile Clickstream	00023	Action	Application Log Elements
Mobile Clickstream	00024	Account Number	IBAN Account Number
Mobile Clickstream	00025	Account Balance	Balance
Mobile Clickstream	00026	Credit Card Number	Credit Card Number
Mobile Clickstream	00027	Credit Card Balance	Balance
Mobile Clickstream	00028	Timestamp	Application Log Elements

Data Set Releases

ReleaseName	DataSet
Test Card System Migration	Credit Card Transactions
Core Banking System Migration - 3	EOD Account Balances
Core Banking System Migration - 1	EOD Account Balances
Core Banking System Migration - 2	EOD Account Balances
Fraud System Migration	EOD Account Balances
Customer OnBoarding	Customer Master
Mobile COVID19 Donation	Mobile Clickstream
Mobile Geofencing	Mobile Clickstream

Purpose	Requestor	Approval	IsApproded	Create Date
Production Support Environment	Georgios V.	Data Analyst	N	17-Feb-2021
Production Support Environment	Georgios V.	CDO	N	07-Feb-2021
Test Environment	Georgios V.	CDO Rejected	N	07-Feb-2021
Isolated Test Environment	Georgios V.	ISO Rejected	N	07-Feb-2021
Production Support Environment	Georgios V.	CDO	N	07-Feb-2021
Unit testing	Georgios V.	ISO	N	27-Feb-2021
Unit testing	Nathan C.		Y	27-Feb-2020
Unit testing	Shirley A.		Y	15-May-2020

Create Time	ReleaseStage	Anonymization Score	Rejection Justification
11:00:00	1		
9:30:00	1		
7:30:00	1		Transactional data cannot be replicated to test
8:30:00	1		Test is not isolated
9:30:00	1		
9:40:00	1		
17:00	2	70%	
17:00	3	95%	

Data Sets Release Fields

ReleaseName	Field	Field Entity	Action
Test Card System Migration	Card Number	Credit Card Number	Mask
Test Card System Migration	Merchant Name	Party Name	Micro-Aggregate
Test Card System Migration	Transaction Amount	Transaction Amount	Pain Text
Test Card System Migration	Terminal Number	Terminal Number	Pain Text
Test Card System Migration	Terminal Geolocation	Geolocation	Pain Text
Test Card System Migration	CID Magnetic Block	Magnetic Stripe Data	Encrypt

Requestor	RequiresApproval	Create Date	Create Time
Georgios V.	Y	20-Jan-2020	15:30:00
Georgios V.	Y	21-Jan-2020	11:30:00
Georgios V.	Y	20-Jan-2020	15:30:00
Georgios V.	N	20-Jan-2020	15:30:00
Georgios V.	Y	20-Jan-2020	15:30:00
Georgios V.	Y	20-Jan-2020	15:30:00

Classification Actions

Action	Business Classification	Regulatory Classification	Anonymization Risk Identifiers	Strength
Pain Text	Public	Non-Regulated	Non-Sensitive Identifier Attribute	10
Mask	Internal Use	PCI	Sensitive Identifier Attribute	20
Encrypt	Internal Use	PCI	Sensitive Identifier Attribute	20
Hash	Confidential	PCI	Sensitive Identifier Attribute	30
Micro-Aggregate	Confidential	PCI	Quassi Identifier Attribute	30
Delete / Static Value	Restricted	PII	Identifier / Key Attribute	40
Delete / Random Data	Restricted	PII	Identifier / Key Attribute	40

Actions

Actions
Pain Text
Mask
Encrypt
Hash
Micro-Aggregate
Delete / Static Value
Delete / Random Data

Exceptions

ReleaseName	Field	Action	Justification	Requestor
Test Card System Migration	Card Number	Mask	Card Number is required to verify correctness of the Migration.	Nathan C.
Test Card System Migration	Transaction Amount	Pain Text	Amount is required for reconciliation	Georgios V.
Test Card System Migration	Merchant Name	Pain Text	Data is required to verify correctness of the Migration	Georgios V.
Test Card System Migration	Merchant Name	Encrypt	Data is required to verify correctness of the Migration	Georgios V.
Test Card System Migration	Merchant Name	Micro-Aggregate	Merchant name is required to verify correctness of the Migration	Georgios V.

Approval	IsApproded	Create Date	Create Time	Rejection Justification
CDO	N	20-Jan-2020	15:30:00	
	Y	20-Jan-2020	15:30:00	
CDO Rejected	N	21-Jan-2020	10:00:00	Production Data cannot be moved
ISO Rejected	N	21-Jan-2020	11:00:00	Masking is not efficient
ISO	N	21-Jan-2020	11:30:00	

Business Entities

Business Entity	Business Classification	Regulatory Classification	Anonymization Risk Identifier
Party Name	Confidential	PII	Identifier / Key Attribute
Party Gender	Internal Use	Sensitive PII	Sensitive Identifier Attribute
Party Date of Birth	Internal Use	PII	Identifier / Key Attribute
Party System ID	Internal Use	Non-Regulated	Identifier / Key Attribute
Party Nationality	Internal Use	Sensitive PII	Non-Sensitive Identifier Attribute
Party Country of Birth	Internal Use	Sensitive PII	Non-Sensitive Identifier Attribute
Party ID Number	Internal Use	PII	Identifier / Key Attribute
Party ID Type	Internal Use	Non-Regulated	Quasi Identifier Attribute
Party ID Country of Issue	Internal Use	Non-Regulated	Quasi Identifier Attribute
Party Salary Amount	Confidential	Non-Regulated	Sensitive Identifier Attribute
Party Private Telephone	Internal Use	Sensitive PII	Identifier / Key Attribute
Party Work Telephone	Internal Use	PII	Identifier / Key Attribute
Party eMail	Internal Use	PII	Identifier / Key Attribute
Party Home Address Street	Internal Use	PII	Identifier / Key Attribute
Party Home Address Country	Internal Use	PII	Non-Sensitive Identifier Attribute
Party Work Address Street	Internal Use	Non-Regulated	Quasi Identifier Attribute
Party Work Address Country	Internal Use	Non-Regulated	Non-Sensitive Identifier Attribute
IBAN Account Number	Public	Non-Regulated	Identifier / Key Attribute
Credit Card Number	Internal Use	PCI	Identifier / Key Attribute
Credit Card Pin	Restricted	PCI	Sensitive Identifier Attribute
Credit Card CVV	Confidential	PCI	Quasi Identifier Attribute
Credit Card Expiry Date	Internal Use	PCI	Non-Sensitive Identifier Attribute
Credit Card Number	Public	Non-Regulated	Non-Sensitive Identifier Attribute
Credit Card Number	Public	PCI	Non-Sensitive Identifier Attribute
Transaction Amount	Internal Use	Non-Regulated	Quasi Identifier Attribute
Terminal Number	Public	Non-Regulated	Non-Sensitive Identifier Attribute
Geolocation	Public	Non-Regulated	Quasi Identifier Attribute
Balance	Internal Use	Non-Regulated	Non-Sensitive Identifier Attribute

Business Entity	Business Classification	Regulatory Classification	Anonymization Risk Identifier
Application Log Elements	Internal Use	Non-Regulated	Non-Sensitive Identifier Attribute
Network Infomration	Internal Use	PII	Identifier / Key Attribute
Magnetic Stripe Data	Public	PCI	Identifier / Key Attribute

Requestor	Approval	IsApproved	Create Date	Create Time	Rejection Justification
Georgios V.		Y	20-Jan-2020	15:30:00	
Georgios V.	CDO	N	21-Jan-2020	10:45:00	
Georgios V.	CDO	N	21-Jan-2020	9:30:00	
Georgios V.		Y			
Georgios V.	ISO	N	21-Jan-2020	10:30:00	
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.	CDO Rejected	N	21-Jan-2020	10:30:00	Credit Card Number is regulated by PCI
Georgios V.	ISO Rejected	N	21-Jan-2020	11:30:00	Credit Card Number cannot be shared
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			
Georgios V.		Y			

Requestor	Approval	IsApproved	Create Date	Create Time	Rejection Justification
Georgios V.		Y			
Georgios V.		Y			

Combined Business Entities

Description	Business Entities	Business Classification	Regulatory Classification	Anonymization Risk Identifiers
Address Combo	Party Home Address Street Party Home Address Country Party Work Address Street Party Work Address Country	Internal Use	PII	Identifier / Key Attribute
Work Addr & IBAN	Party Work Address Street IBAN Account Number	Internal Use	Non-Regulated	Identifier / Key Attribute
Home Addr & DoB	Party Home Address Street Party Date of Birth	Internal Use	PII	Identifier / Key Attribute
Home Addr & DoB	Party Home Address Street Party Date of Birth	Public	Non-Regulated	Non-Sensitive Identifier Attribute
Home Addr & DoB	Party Home Address Street Party Date of Birth	Public	Non-Regulated	Quasi Identifier Attribute

Requestor	Approval	IsApproved	Create Date	Create Time	Rejection Justification
Georgios V.	CDO	N	10-Jun-2020	10:00:00	
Georgios V.		Y	25-Jun-2020	13:05:00	
Georgios V.	ISO	N	25-Jun-2020	13:15:00	
Georgios V.		N			

Requestor	Approval	IsApproved	Create Date	Create Time	Rejection Justification
	CDO Rejected		25-Jun-2020	10:15:00	Date of Birth is Identifier
Georgios V.	ISO Rejected	N	25-Jun-2020	12:15:00	Date of Birth is Identifier / Key Attribute

Business Classifications

Business Classification	Strength
(Empty)	0
Restricted	40
Confidential	30
Internal Use	20
Public	10

Regulatory Classifications

Regulatory Classification	Strength
(Empty)	0
Sensitive PII	40
PII	30
PCI	20
Non-Regulated	10

Anonymization Risk Identifiers

Anonymization Risk Identifiers	Strength
(Empty)	0
Identifier / Key Attribute	40
Quasi Identifier Attribute	30
Sensitive Identifier Attribute	20
Non-Sensitive Identifier Attribute	10

Users

UserName	Password	Role	Name	isActive	Image
gv-CDO	A123456!	CDO	Georgios V.	Y	Y:\Images\GV-ProfilePhoto.jpg
gv-ISO	A123456!	ISO	Georgios V.	Y	Y:\Images\GV-ProfilePhoto.jpg
gv-DA	A123456!	Data Analyst	Georgios V.	Y	Y:\Images\GV-ProfilePhoto.jpg
gv-Admin	A123456!	Administrator	Georgios V.	Y	Y:\Images\GV-ProfilePhoto.jpg

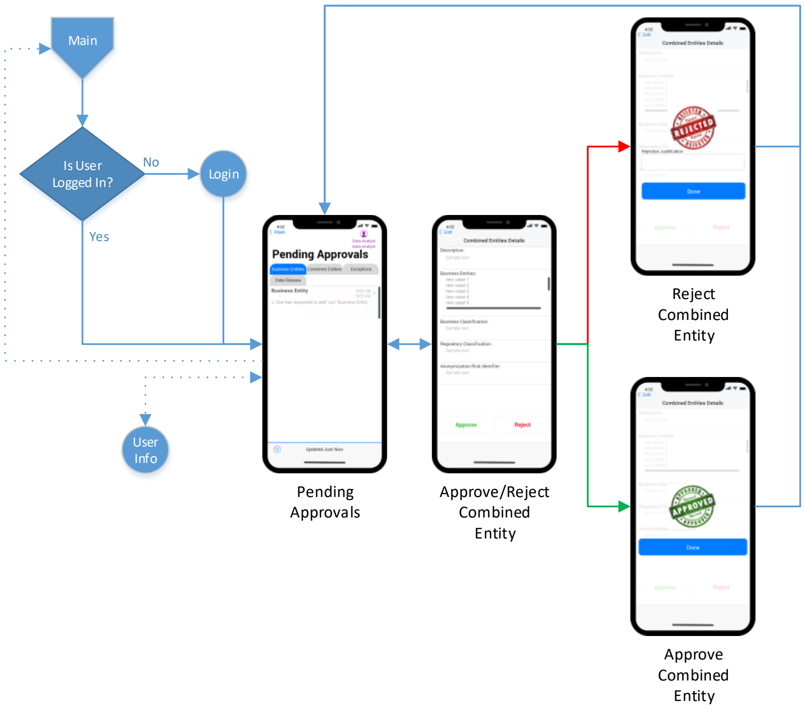
UserName	Password	Role	Name	isActive	Image
nc-CDO	A123456!	CDO	Nathan C.	Y	Y:\Images\Male-ProfilePhoto.jpg
nc-ISO	A123456!	ISO	Nathan C.	Y	Y:\Images\Male-ProfilePhoto.jpg
nc-DA	A123456!	Data Analyst	Nathan C.	Y	Y:\Images\Male-ProfilePhoto.jpg
sa-CDO	A123456!	CDO	Shirley A.	Y	Y:\Images\Female-ProfilePhoto.jpg
sa-ISO	A123456!	ISO	Shirley A.	Y	Y:\Images\Female-ProfilePhoto.jpg
sa-DA	A123456!	Data Analyst	Shirley A.	Y	Y:\Images\Female-ProfilePhoto.jpg

User Roles

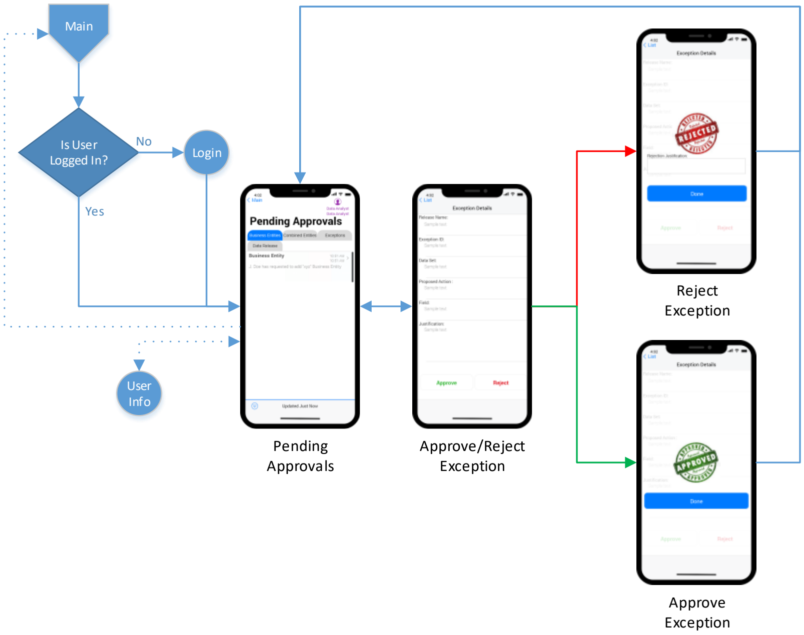
RoleName
ISO
CDO
Data Analyst
Administrator

X. BD-CPS PoC Additional Approval Journeys

Combined Entity Approval

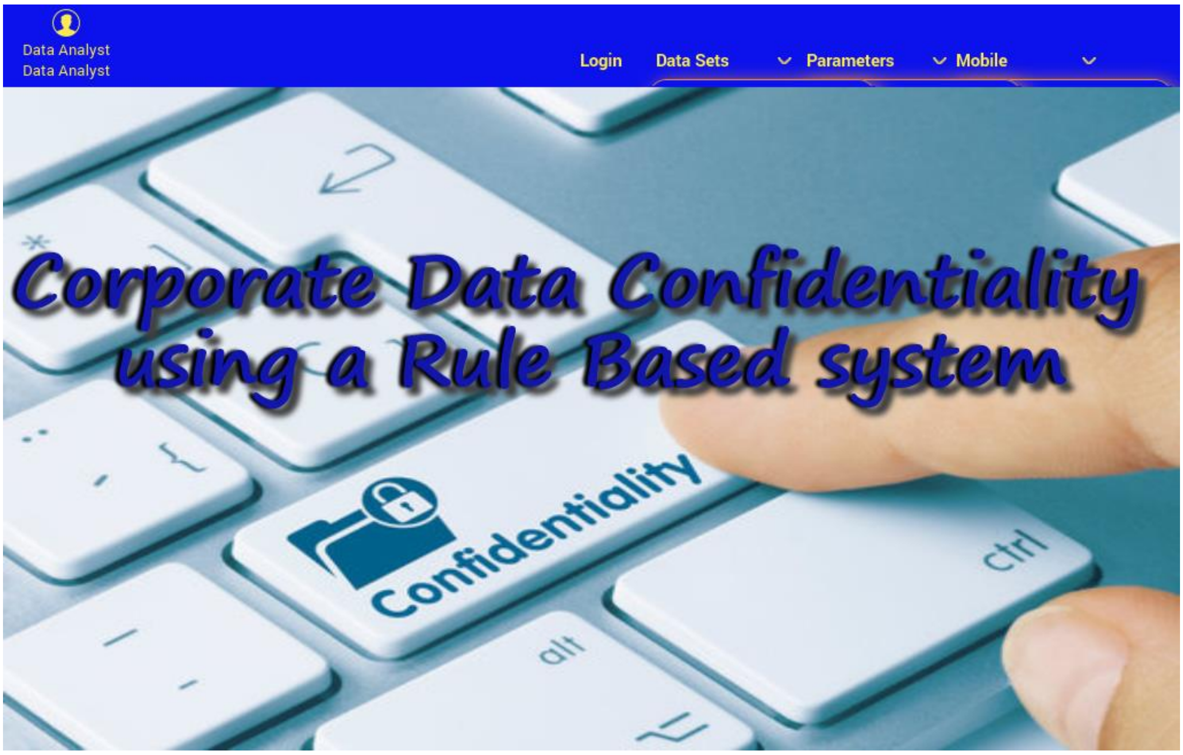


Exception Approval

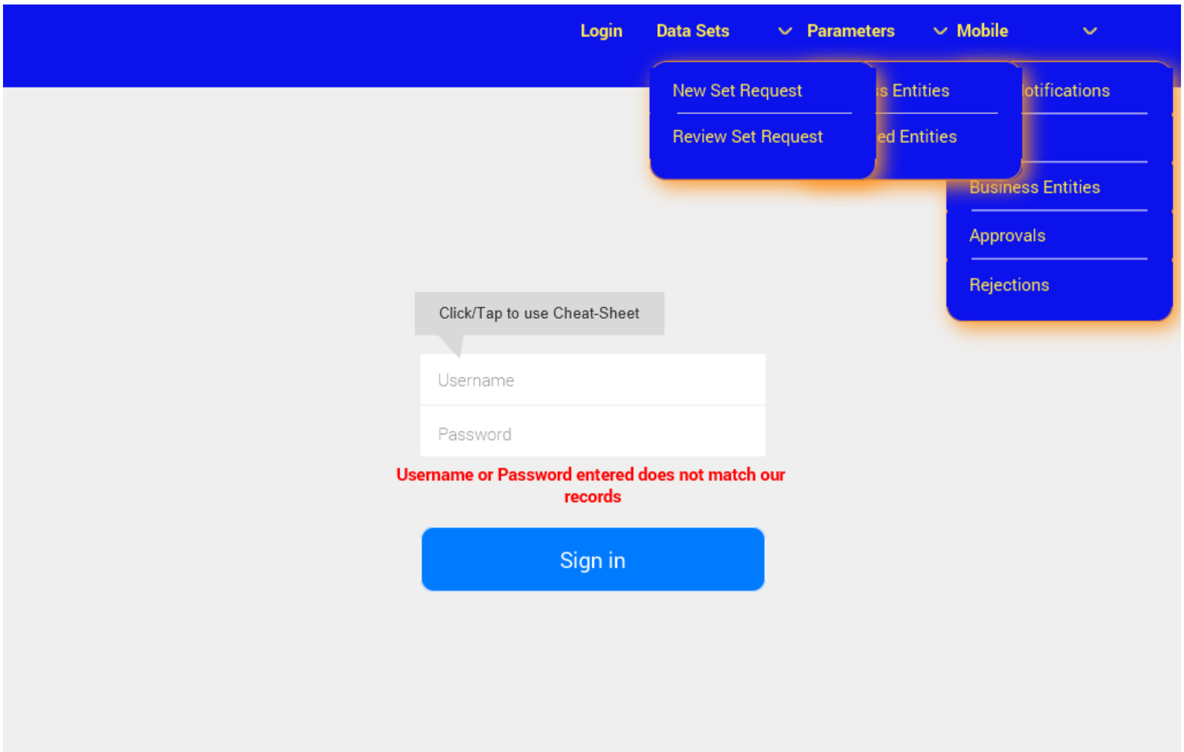


XI. BD-CPS PoC Design UI/UX Elements

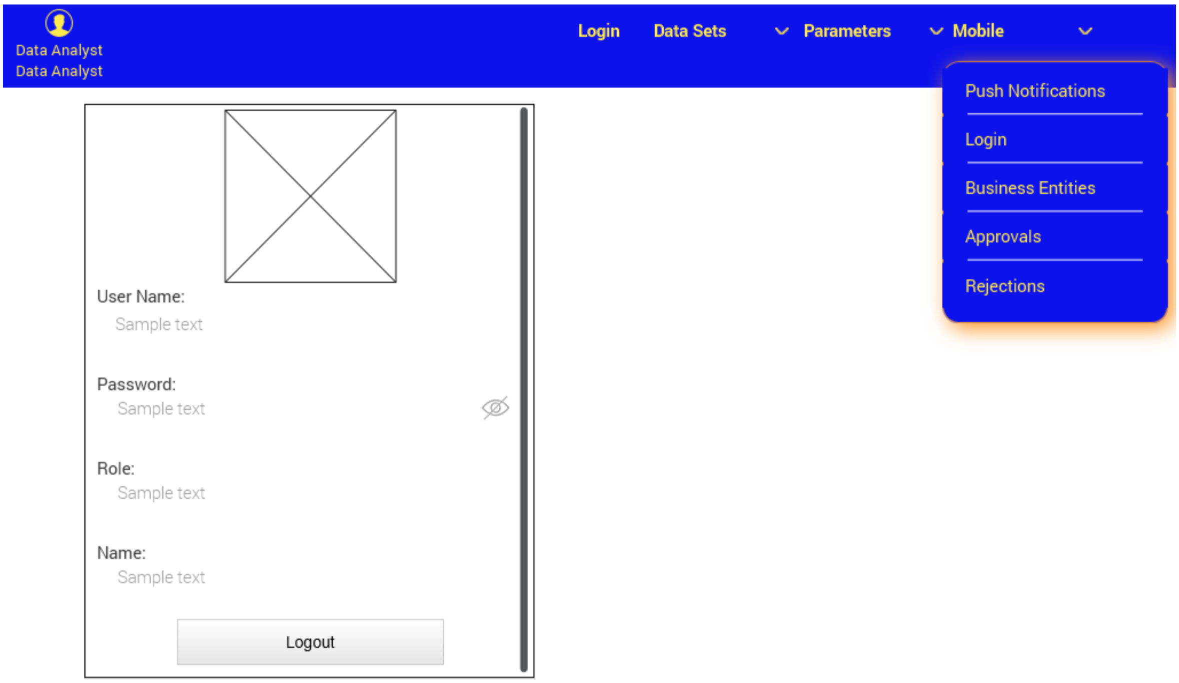
Web Interface



Corporate Data Confidentiality using a Rule Based system



Username or Password entered does not match our records

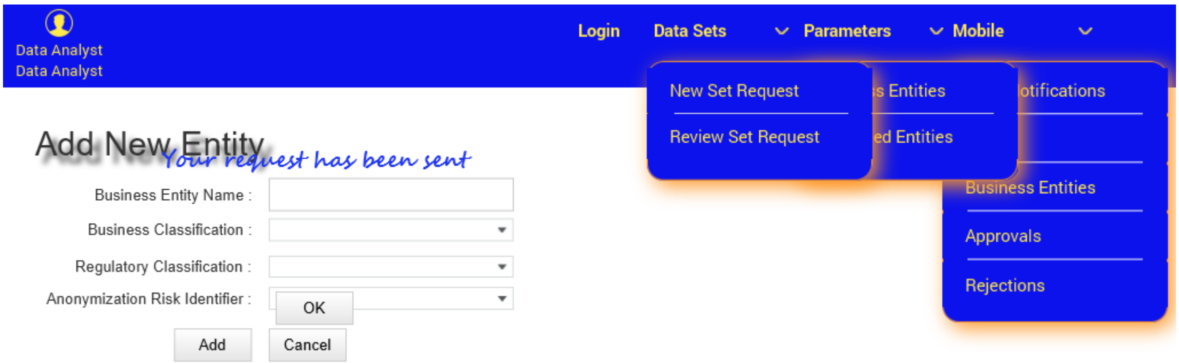


User Name:
Sample text

Password:
Sample text

Role:
Sample text

Name:
Sample text



Add New Entity

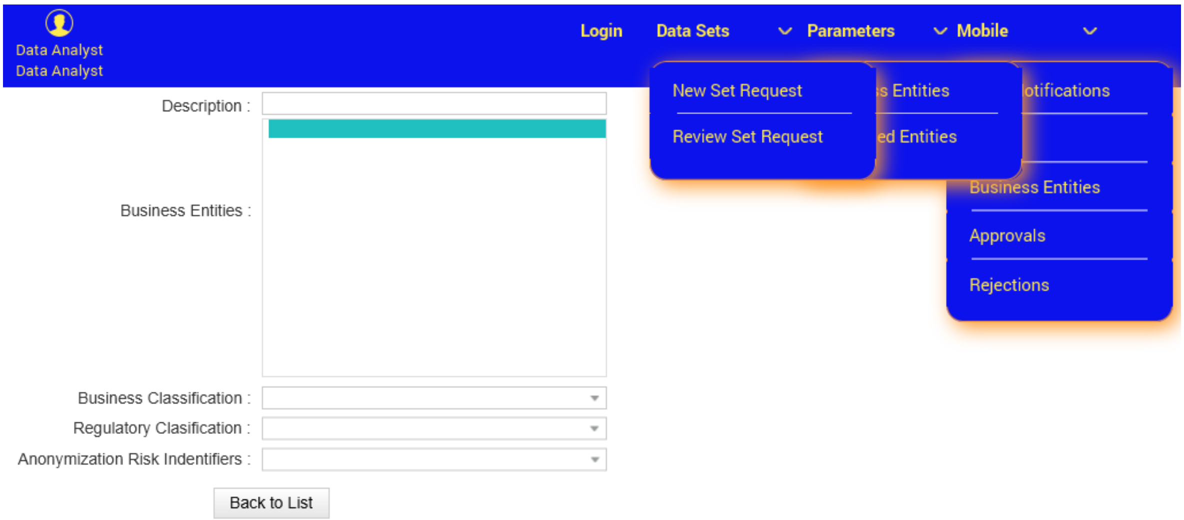
Your request has been sent

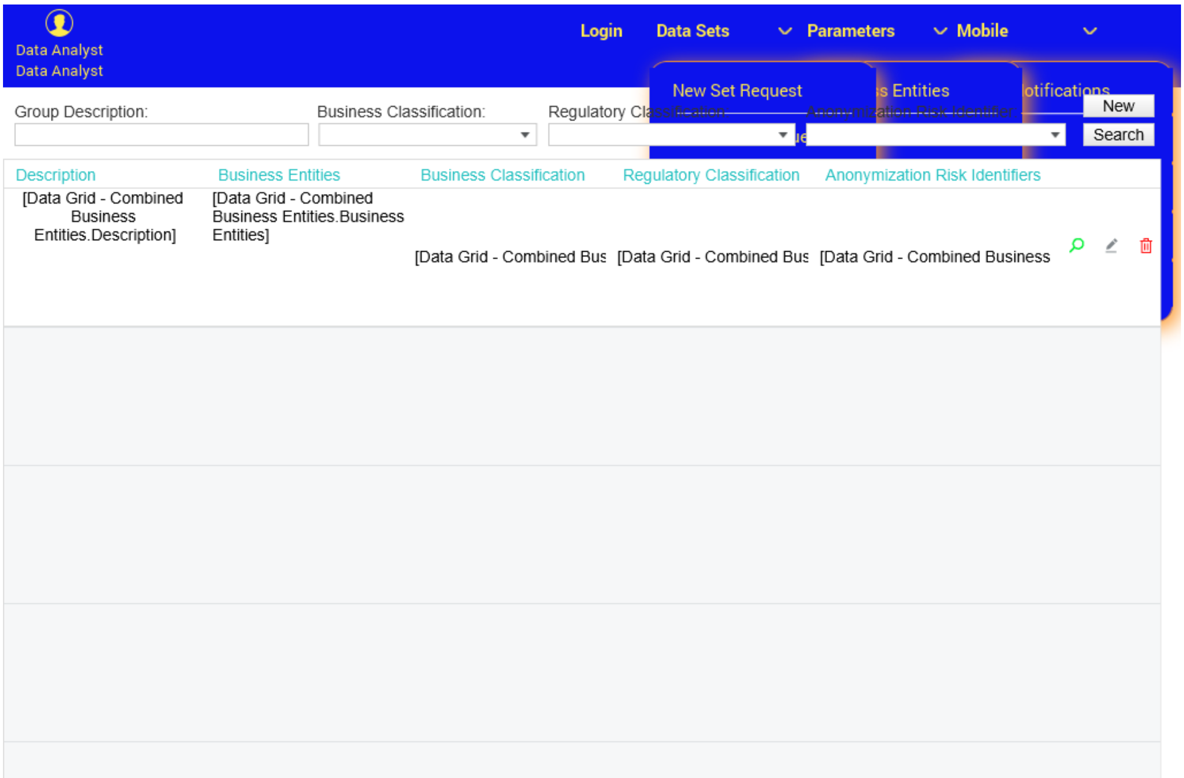
Business Entity Name :

Business Classification :

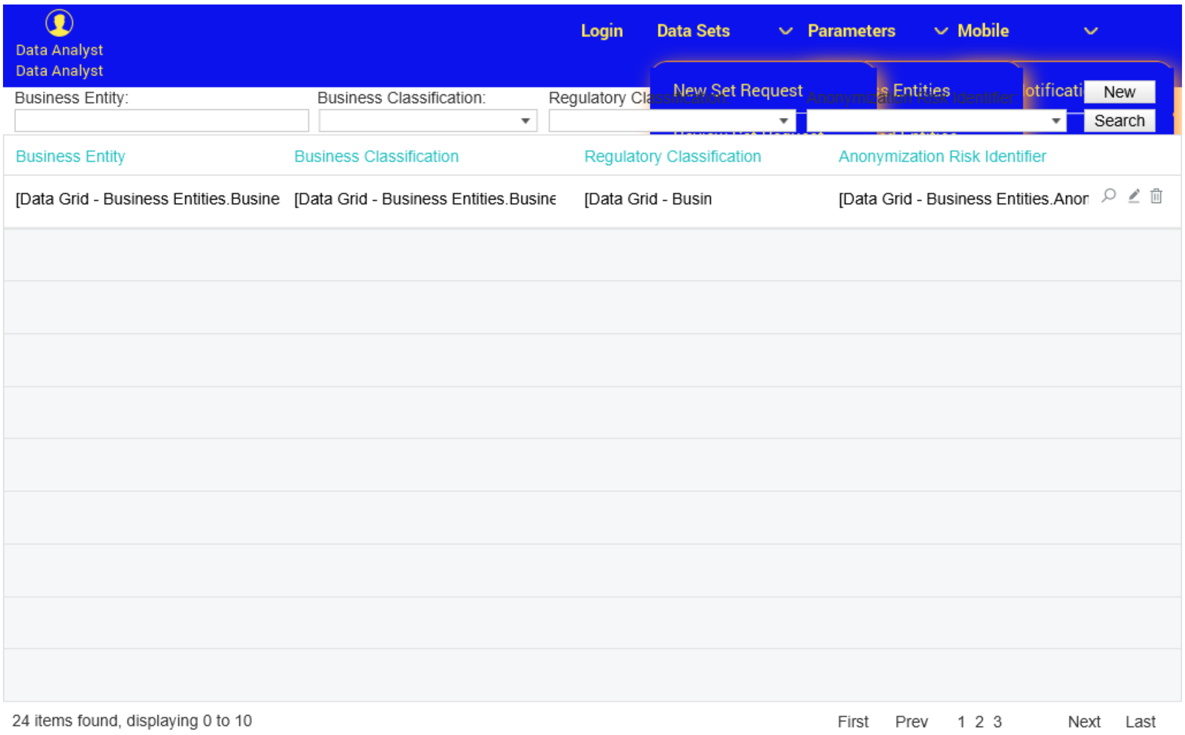
Regulatory Classification :

Anonymization Risk Identifier :





Description	Business Entities	Business Classification	Regulatory Classification	Anonymization Risk Identifiers
[Data Grid - Combined Business Entities.Description]	[Data Grid - Combined Business Entities.Business Entities]	[Data Grid - Combined Bus	[Data Grid - Combined Bus	[Data Grid - Combined Business



Business Entity:

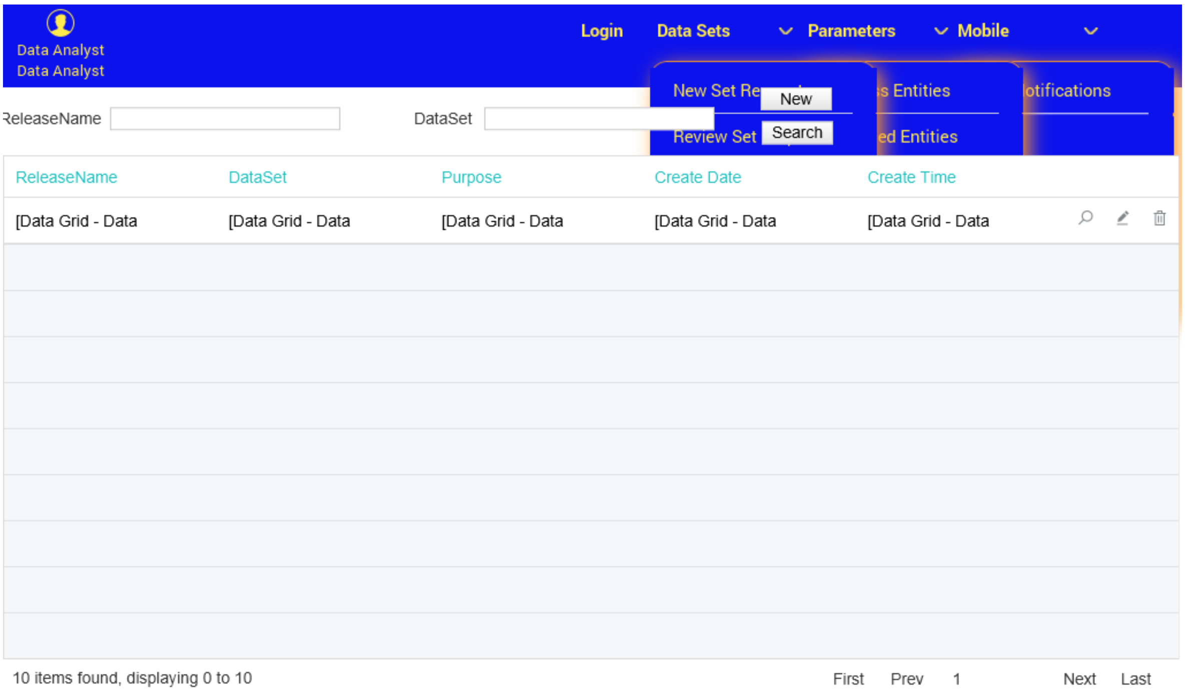
Business Classification:

Regulatory Classification:

Anonymization Risk Identifier:

Business Entity	Business Classification	Regulatory Classification	Anonymization Risk Identifier
[Data Grid - Business Entities.Busine	[Data Grid - Business Entities.Busine	[Data Grid - Busin	[Data Grid - Business Entities.Anor

24 items found, displaying 0 to 10

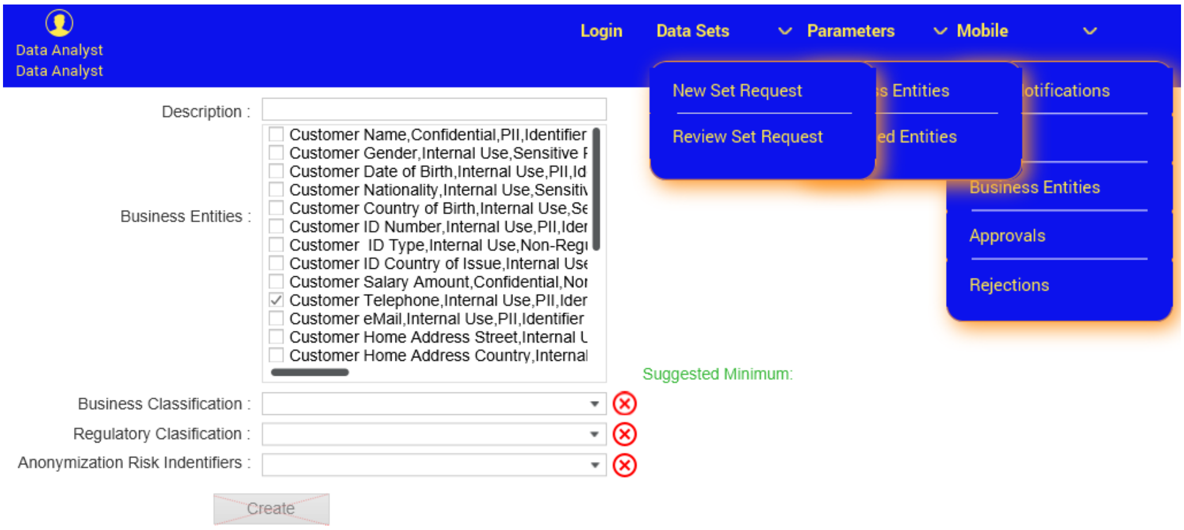


ReleaseName

DataSet

ReleaseName	DataSet	Purpose	Create Date	Create Time
[Data Grid - Data	[Data Grid - Data	[Data Grid - Data	[Data Grid - Data	[Data Grid - Data

10 items found, displaying 0 to 10



Description :

Business Entities :

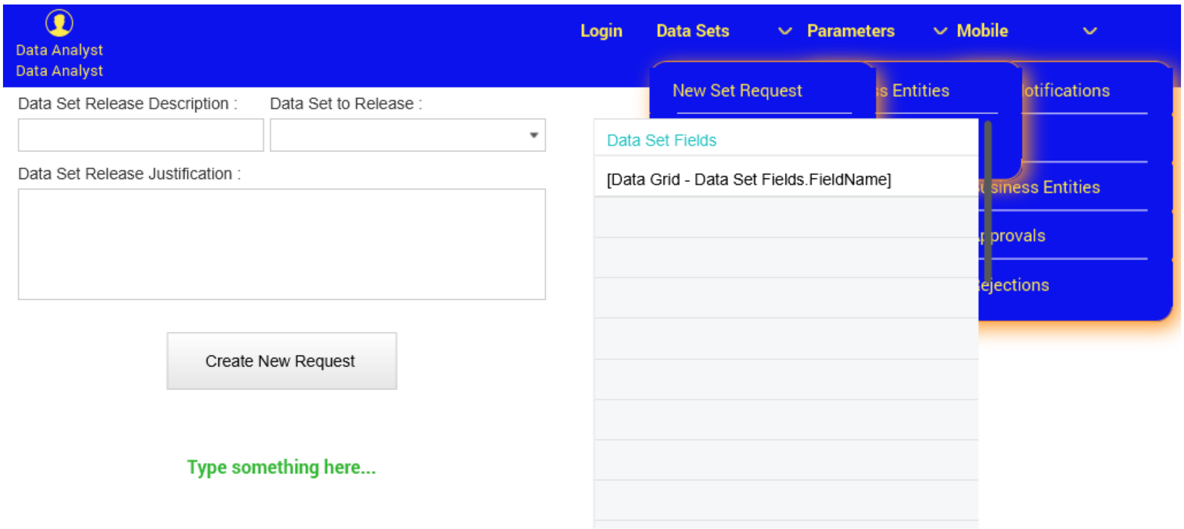
□ Customer Name,Confidential,PII,Identifier
□ Customer Gender,Internal Use,Sensitive F
□ Customer Date of Birth,Internal Use,PII,Id
□ Customer Nationality,Internal Use,Sensitiv
□ Customer Country of Birth,Internal Use,Se
□ Customer ID Number,Internal Use,PII,Ider
□ Customer ID Type,Internal Use,Non-Regı
□ Customer ID Country of Issue,Internal Use
□ Customer Salary Amount,Confidential,Nor
☑ Customer Telephone,Internal Use,PII,Ider
□ Customer eMail,Internal Use,PII,Identifier
□ Customer Home Address Street,Internal U
□ Customer Home Address Country,Internal

Suggested Minimum:

Business Classification : ☒

Regulatory Clasification : ☒

Anonymization Risk Indentifiers : ☒



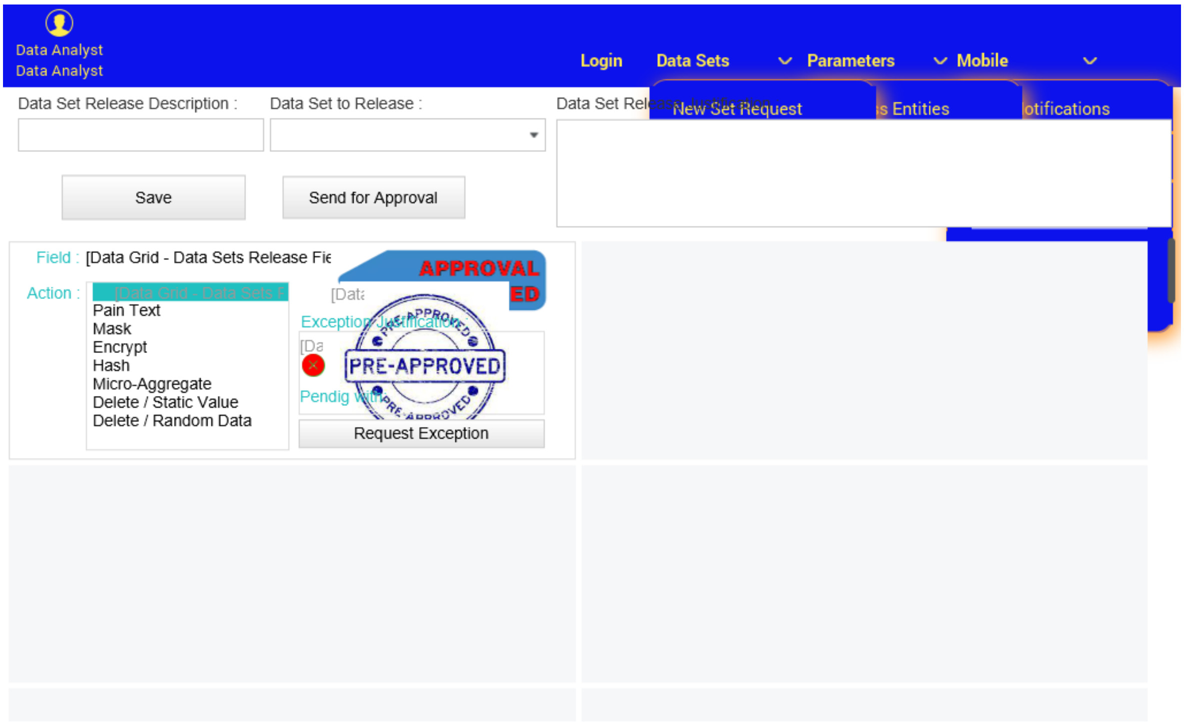
Data Set Release Description :

Data Set to Release :

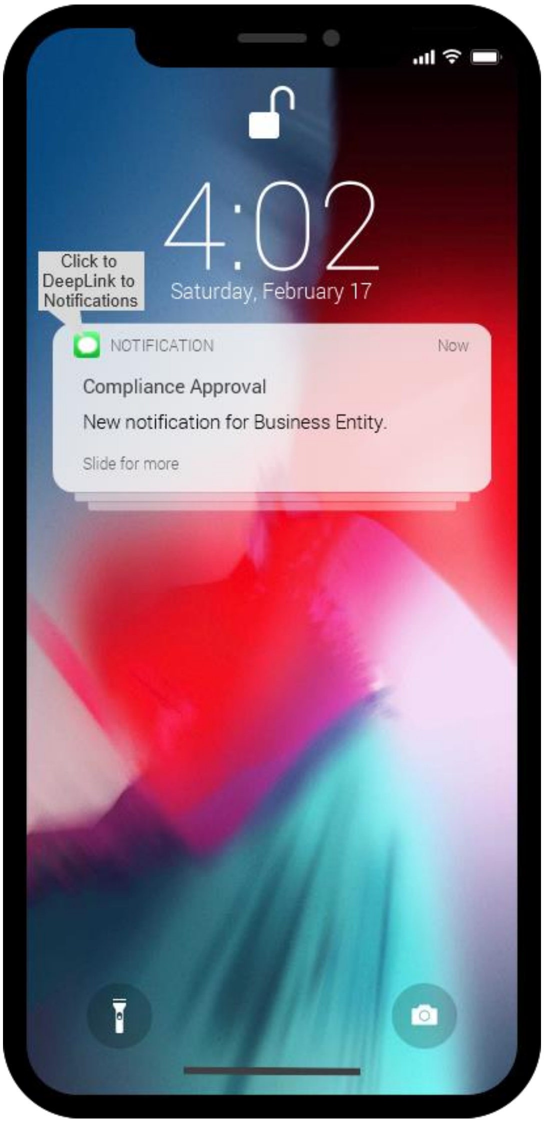
Data Set Release Justification :

Type something here...

Data Set Fields
[Data Grid - Data Set Fields.FieldName]


Data Analyst
Data Analyst
Login
Data Sets
Parameters
Mobile
New Set Request
s Entities
otifications
Data Set Release Description :
Data Set to Release :
Data Set Rel
Save
Send for Approval
Field : [Data Grid - Data Sets Release Fie
Action :
[Data Grid - Data Sets
Pain Text
Mask
Encrypt
Hash
Micro-Aggregate
Delete / Static Value
Delete / Random Data
APPROVAL
ED
[Data
Exception Justification
[Da
Pendig with
PRE-APPROVED
Request Exception

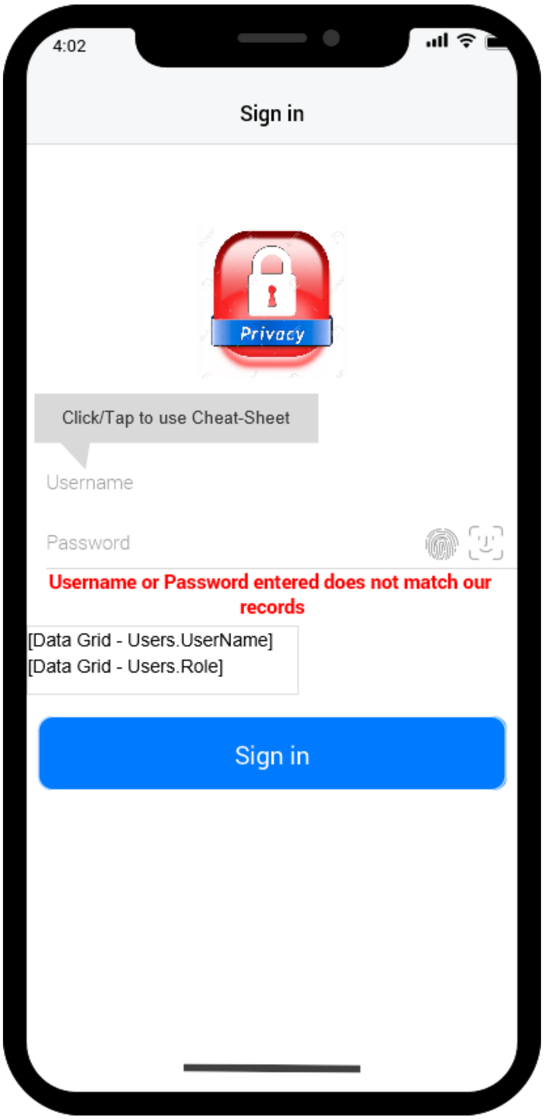
Mobile Interface



Compliance Approval

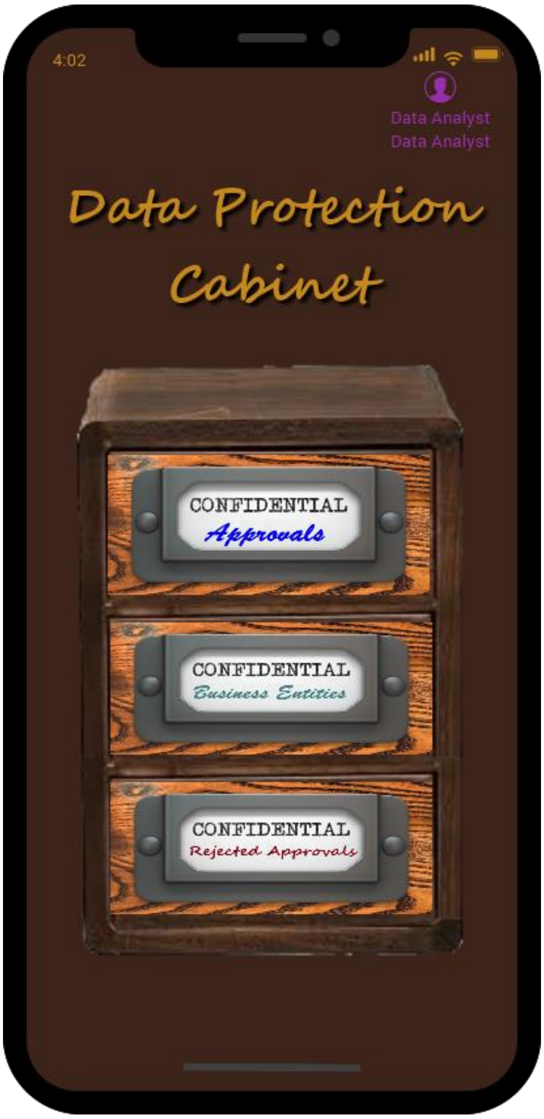
New notification for Business Entity.

Slide for more



Username or Password entered does not match our records

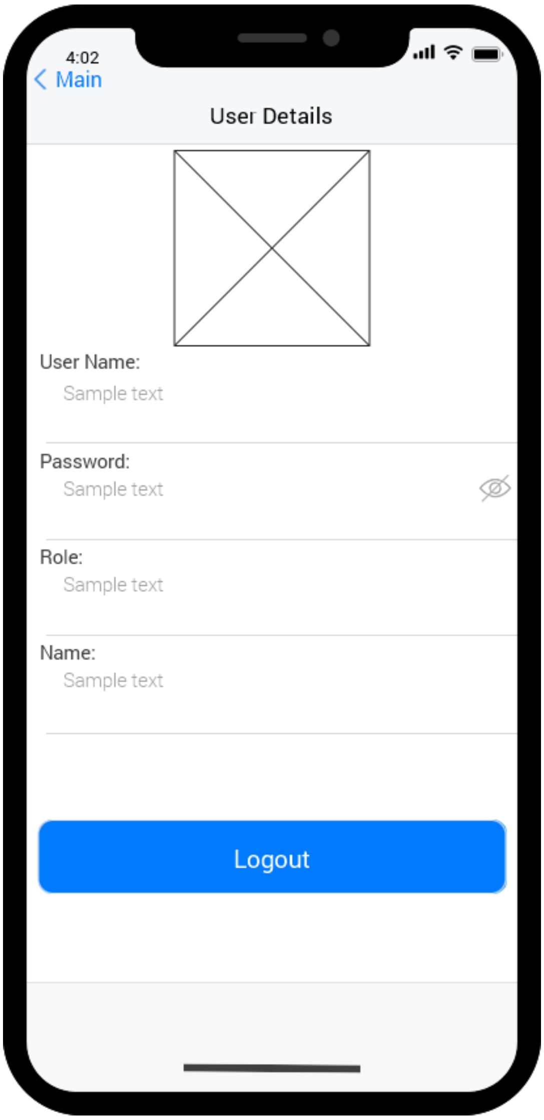
[Data Grid - Users.UserName]
[Data Grid - Users.Role]



4:02

Data Analyst
Data Analyst

Data Protection Cabinet



4:02

< Main

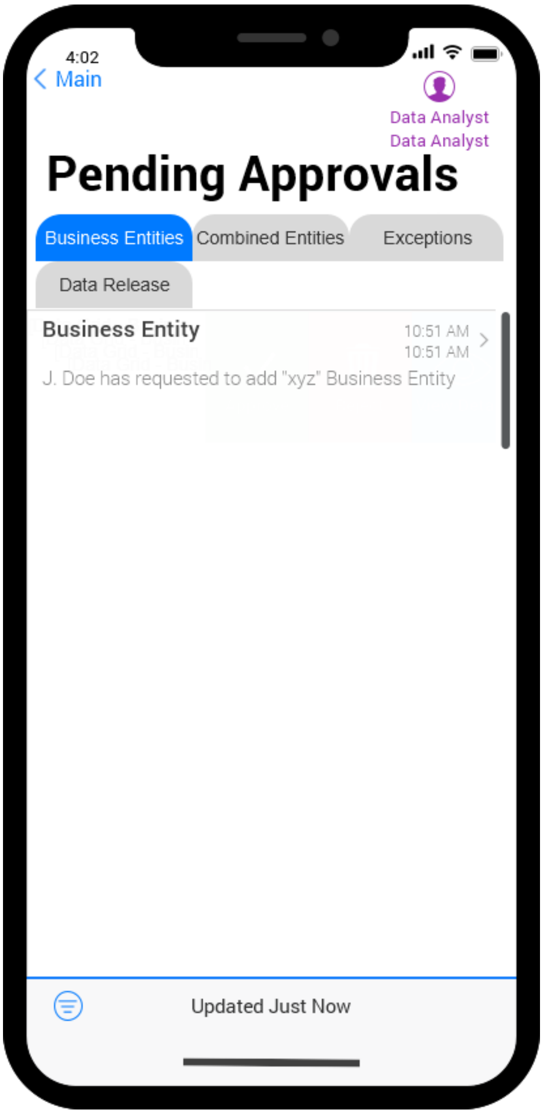
User Details

User Name:
Sample text

Password:
Sample text

Role:
Sample text

Name:
Sample text



4:02

Main

Data Analyst
Data Analyst

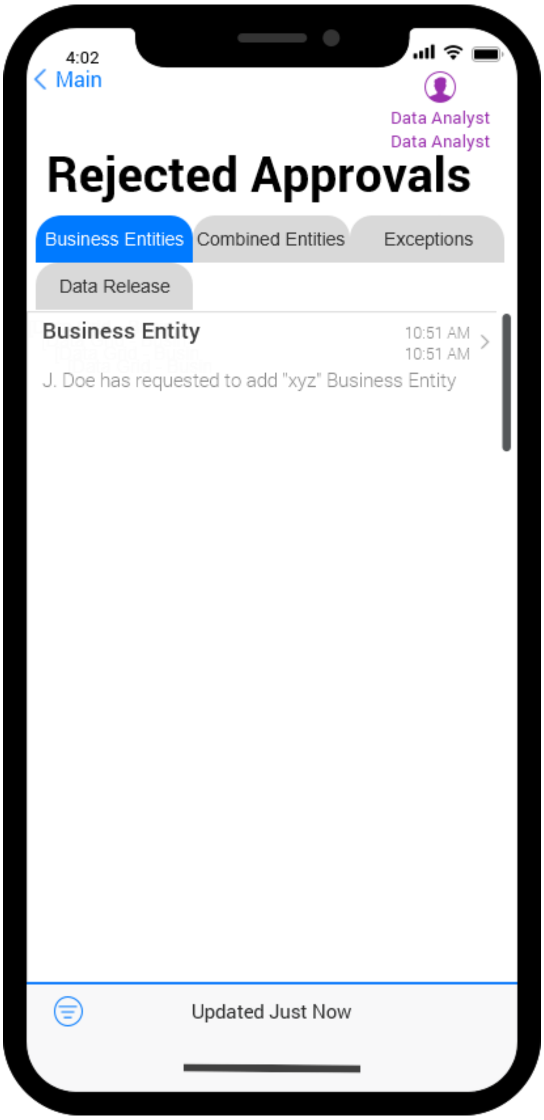
Pending Approvals

Business Entities | Combined Entities | Exceptions | Data Release

Business Entity 10:51 AM 10:51 AM

J. Doe has requested to add "xyz" Business Entity

Updated Just Now



4:02

Main

Data Analyst
Data Analyst

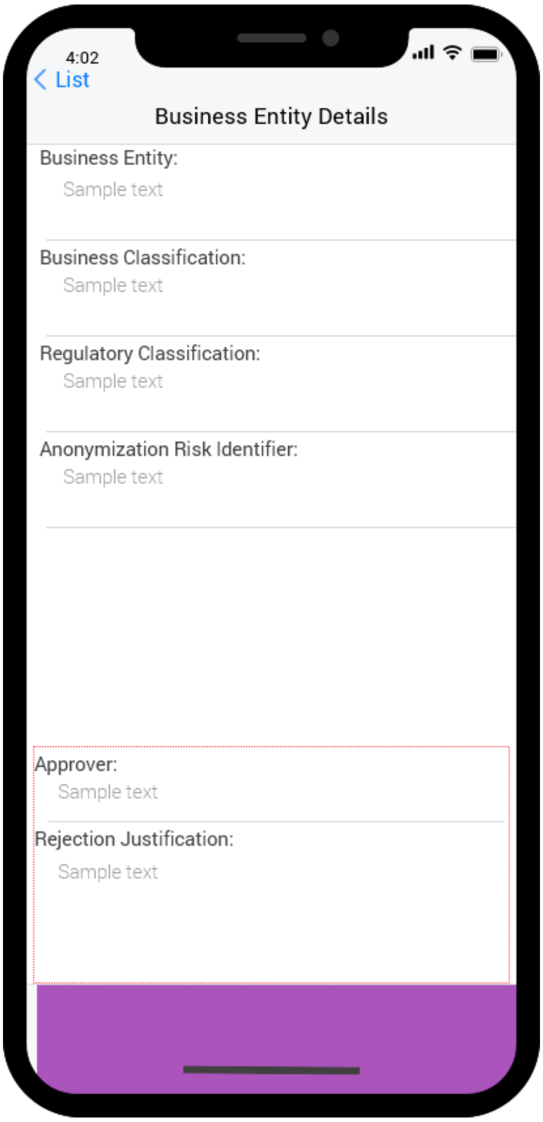
Rejected Approvals

Business Entities | Combined Entities | Exceptions | Data Release

Business Entity 10:51 AM 10:51 AM

J. Doe has requested to add "xyz" Business Entity

Updated Just Now



4:02

List

Business Entity Details

Business Entity:
Sample text

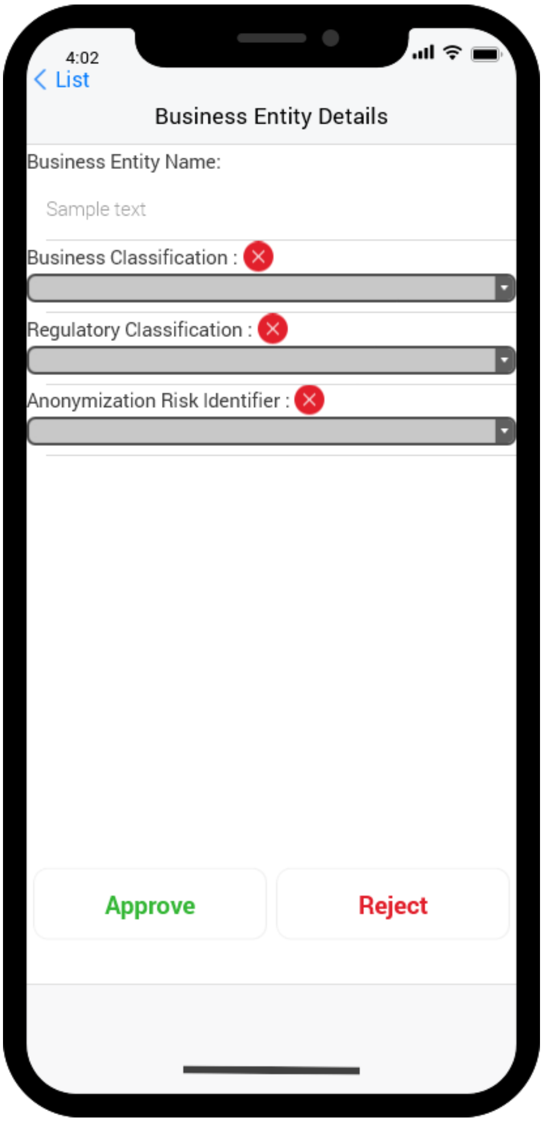
Business Classification:
Sample text

Regulatory Classification:
Sample text

Anonymization Risk Identifier:
Sample text

Approver:
Sample text

Rejection Justification:
Sample text



4:02

List

Business Entity Details

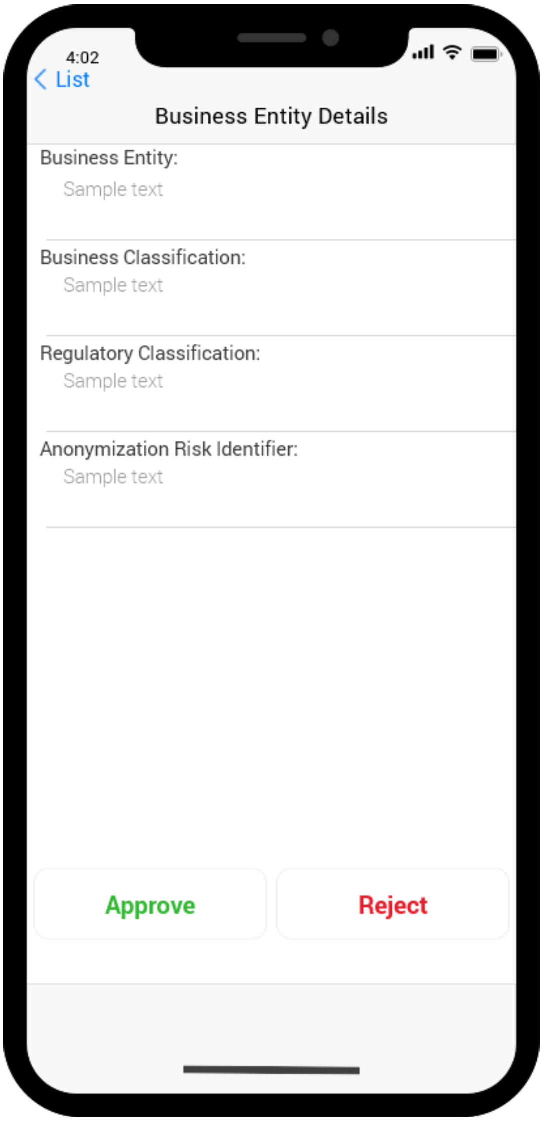
Business Entity Name:

Sample text

Business Classification : ☒

Regulatory Classification : ☒

Anonymization Risk Identifier : ☒

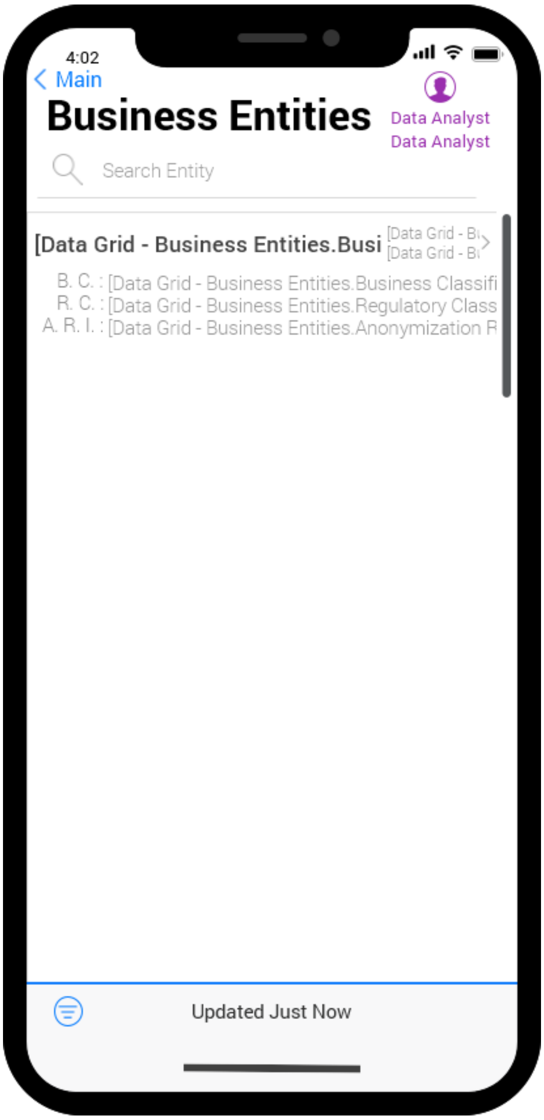


Business Entity:
Sample text

Business Classification:
Sample text

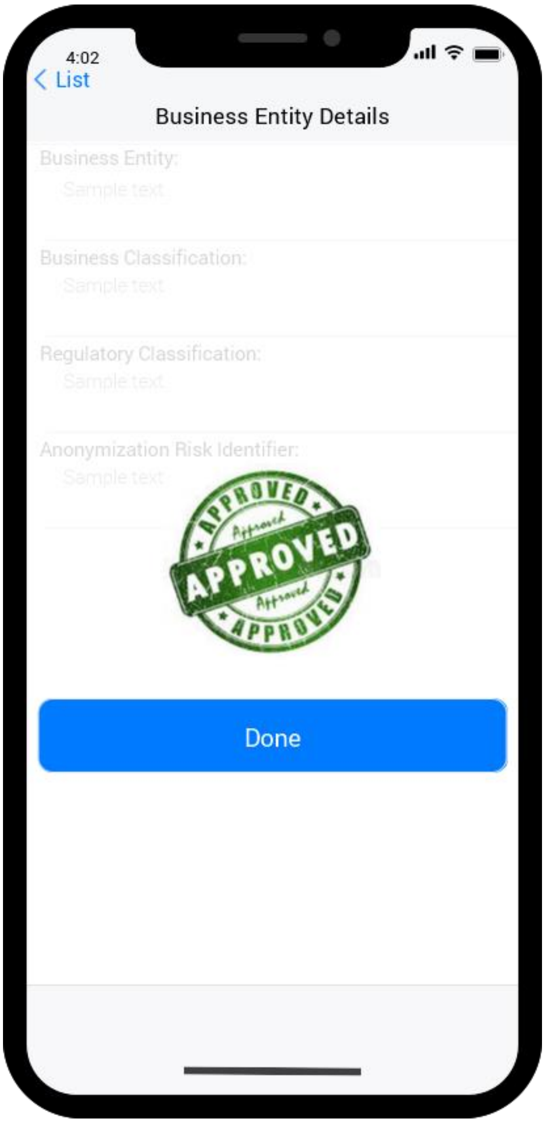
Regulatory Classification:
Sample text

Anonymization Risk Identifier:
Sample text



[Data Grid - Business Entities.Busi [Data Grid - Bi [Data Grid - Bi

B. C. : [Data Grid - Business Entities.Business Classifi
R. C. : [Data Grid - Business Entities.Regulatory Class
A. R. I. : [Data Grid - Business Entities.Anonymization R



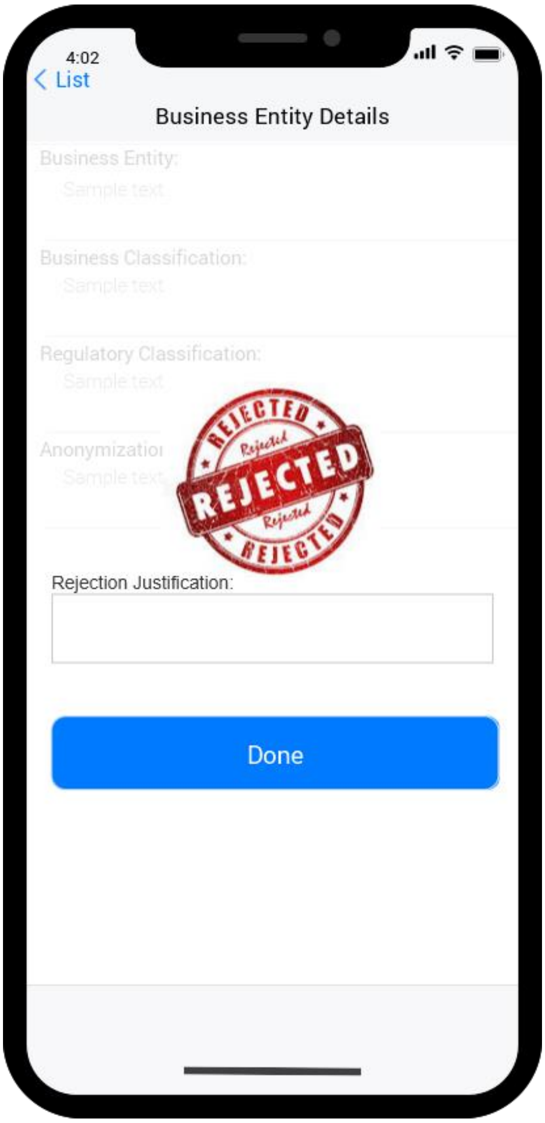
Business Entity Details

Business Entity:
Sample text

Business Classification:
Sample text

Regulatory Classification:
Sample text

Anonymization Risk Identifier:
Sample text



Business Entity Details

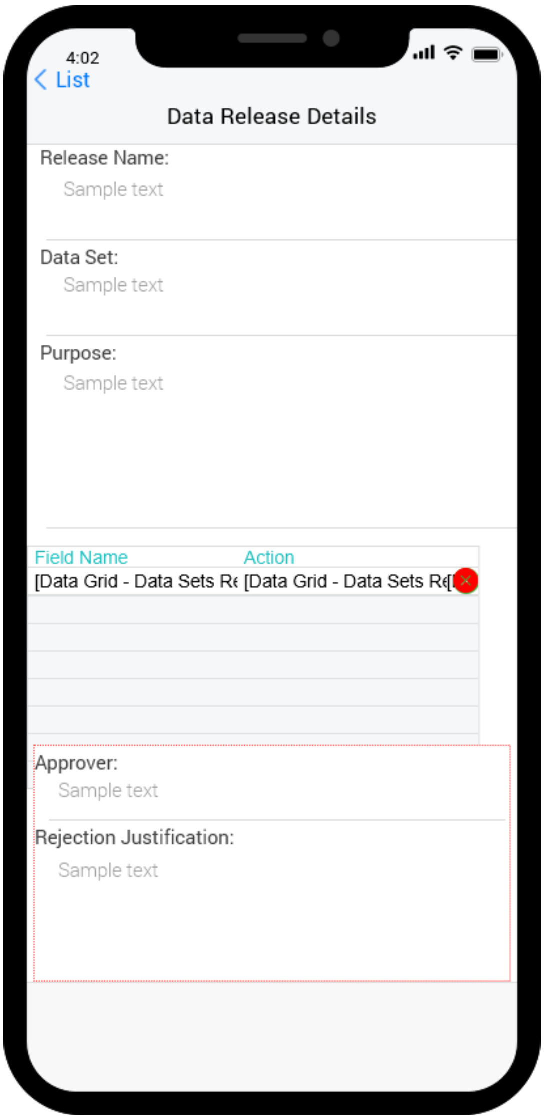
Business Entity:
Sample text

Business Classification:
Sample text

Regulatory Classification:
Sample text

Anonymizatioi
Sample text

Rejection Justification:



Data Release Details

Release Name:
Sample text

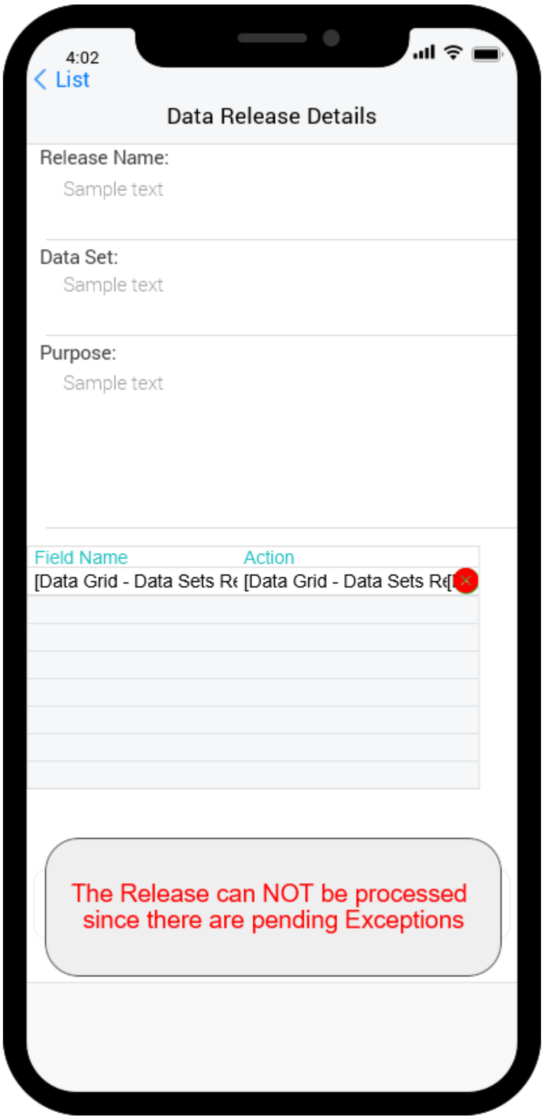
Data Set:
Sample text

Purpose:
Sample text

Field Name	Action
[Data Grid - Data Sets Re	[Data Grid - Data Sets Re[

Approver:
Sample text

Rejection Justification:
Sample text



Data Release Details

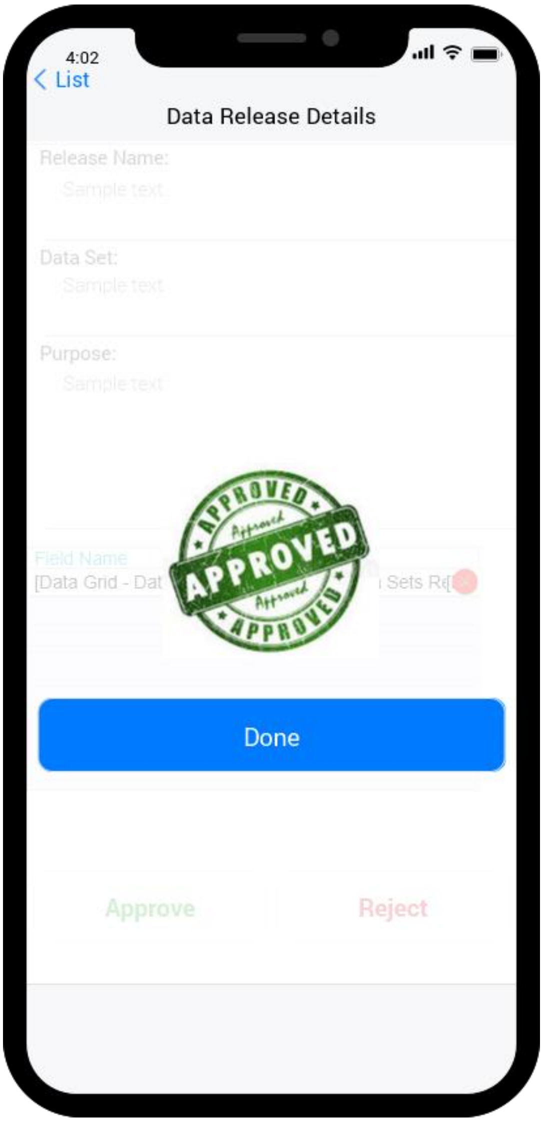
Release Name:
Sample text

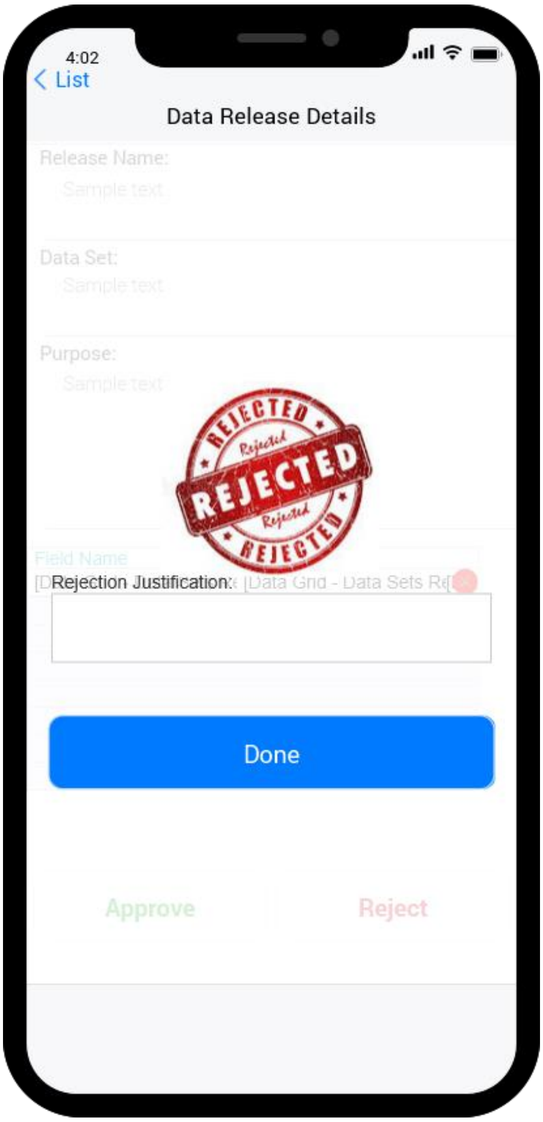
Data Set:
Sample text

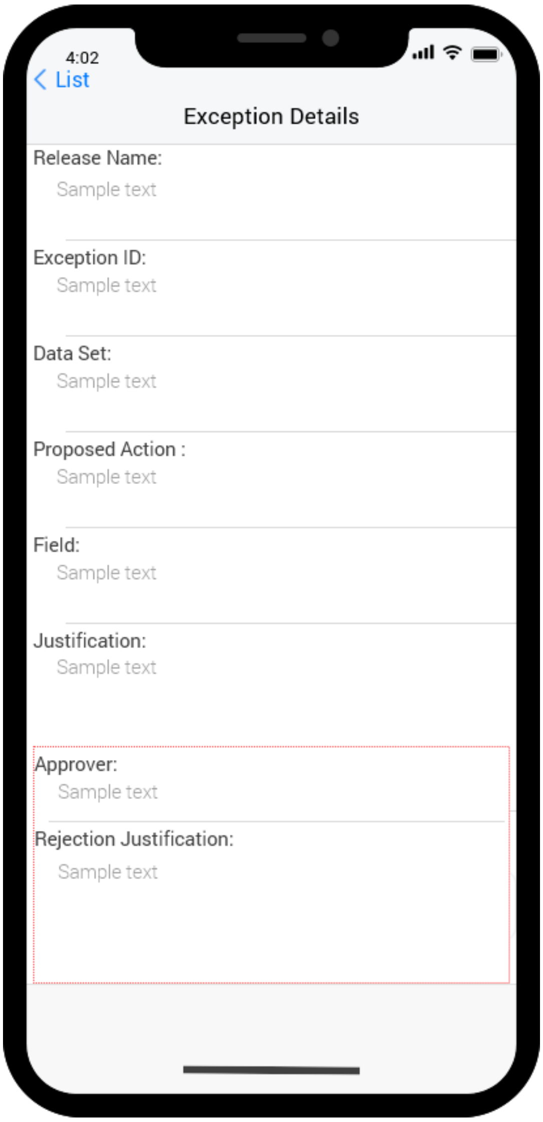
Purpose:
Sample text

Field Name	Action
[Data Grid - Data Sets Re	[Data Grid - Data Sets Re[

The Release can NOT be processed since there are pending Exceptions


4:02
List
Data Release Details
Release Name:
Sample text
Data Set:
Sample text
Purpose:
Sample text
Field Name
[Data Grid - Dat
Sets Re
APPROVED
Approved
APPROVED
Approved
APPROVED
Done
Approve
Reject


4:02
List
Data Release Details
Release Name:
Sample text
Data Set:
Sample text
Purpose:
Sample text
REJECTED
Rejected
REJECTED
Rejected
REJECTED
Field Name
Rejection Justification:
[Data Grid - Data Sets Re
Done
Approve
Reject



4:02

List

Exception Details

Release Name:
Sample text

Exception ID:
Sample text

Data Set:
Sample text

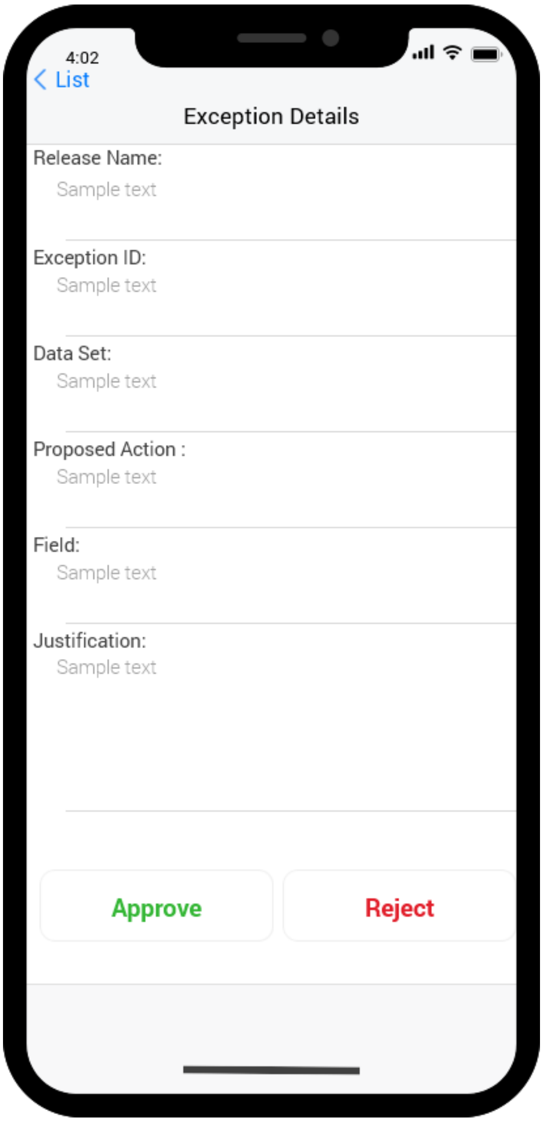
Proposed Action :
Sample text

Field:
Sample text

Justification:
Sample text

Approver:
Sample text

Rejection Justification:
Sample text



4:02

List

Exception Details

Release Name:
Sample text

Exception ID:
Sample text

Data Set:
Sample text

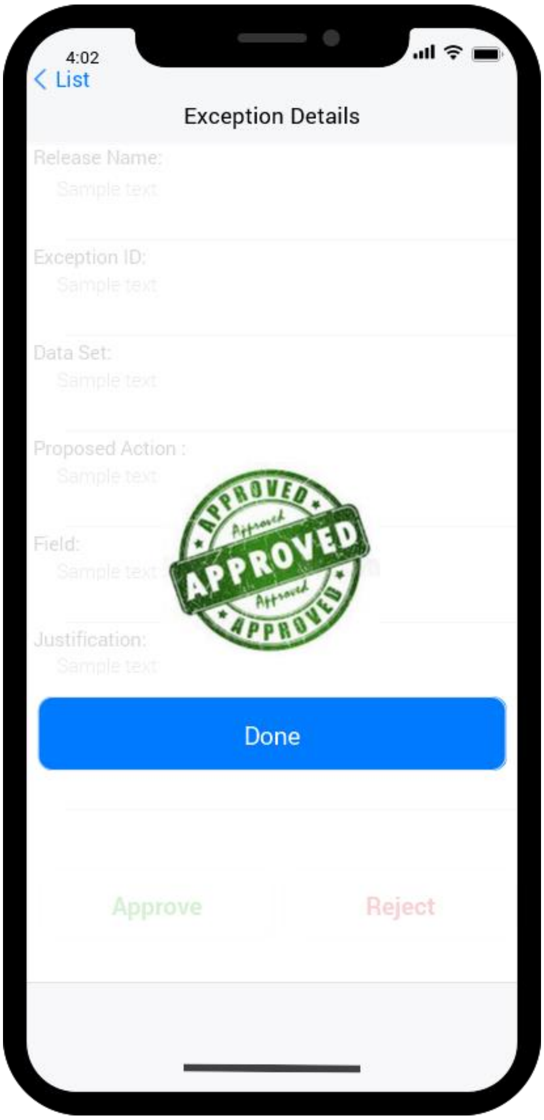
Proposed Action :
Sample text

Field:
Sample text

Justification:
Sample text

Approve

Reject



Release Name:
Sample text

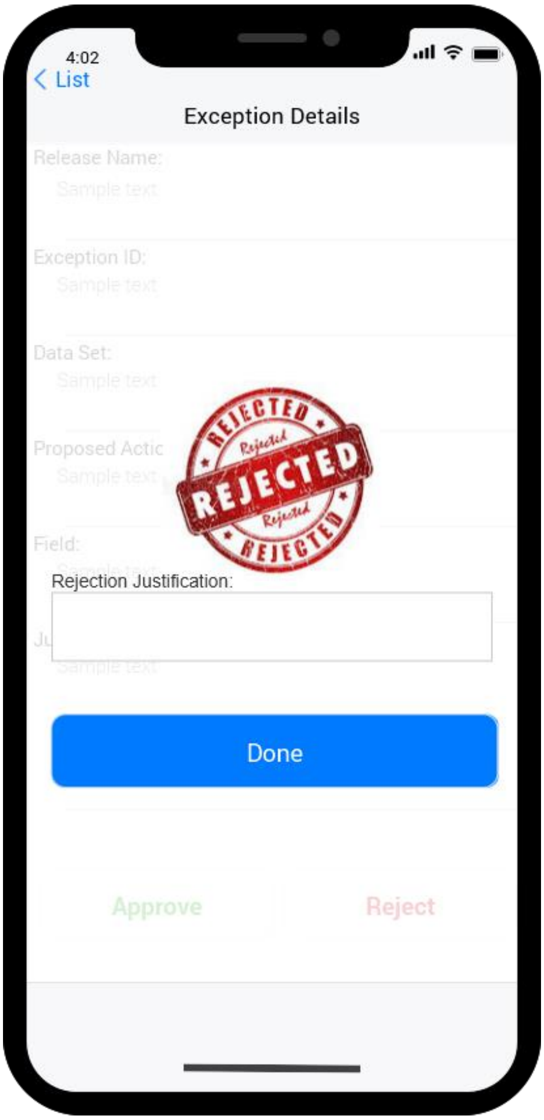
Exception ID:
Sample text

Data Set:
Sample text

Proposed Action :
Sample text

Field:
Sample text

Justification:
Sample text



Release Name:
Sample text

Exception ID:
Sample text

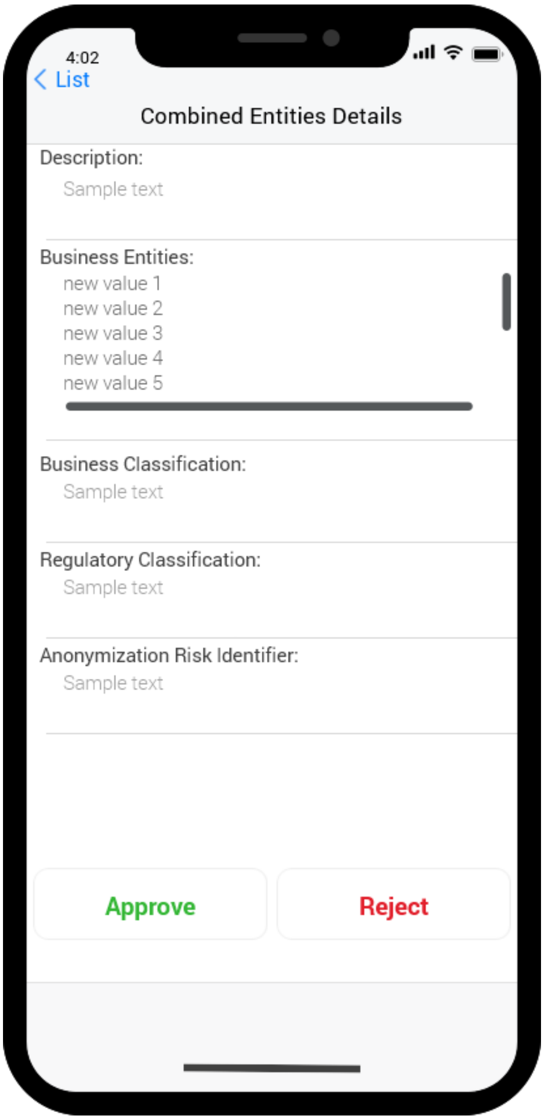
Data Set:
Sample text

Proposed Actic
Sample text

Field:

Rejection Justification:

Sample text



Combined Entities Details

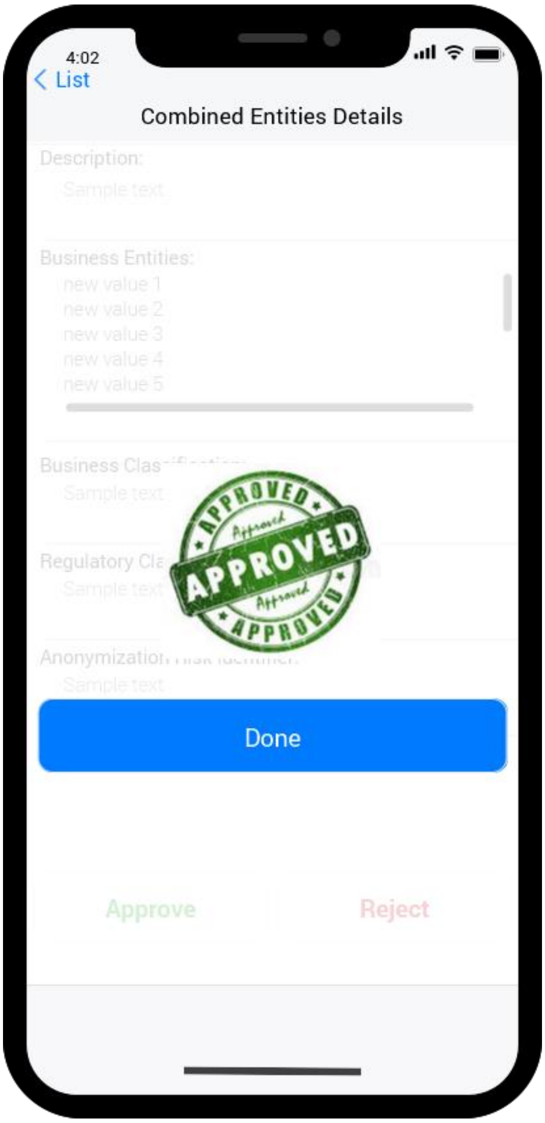
Description:
Sample text

Business Entities:
new value 1
new value 2
new value 3
new value 4
new value 5

Business Classification:
Sample text

Regulatory Classification:
Sample text

Anonymization Risk Identifier:
Sample text



4:02

List

Combined Entities Details

Description:

Sample text

Business Entities:

new value 1
new value 2
new value 3
new value 4
new value 5

Business Clas

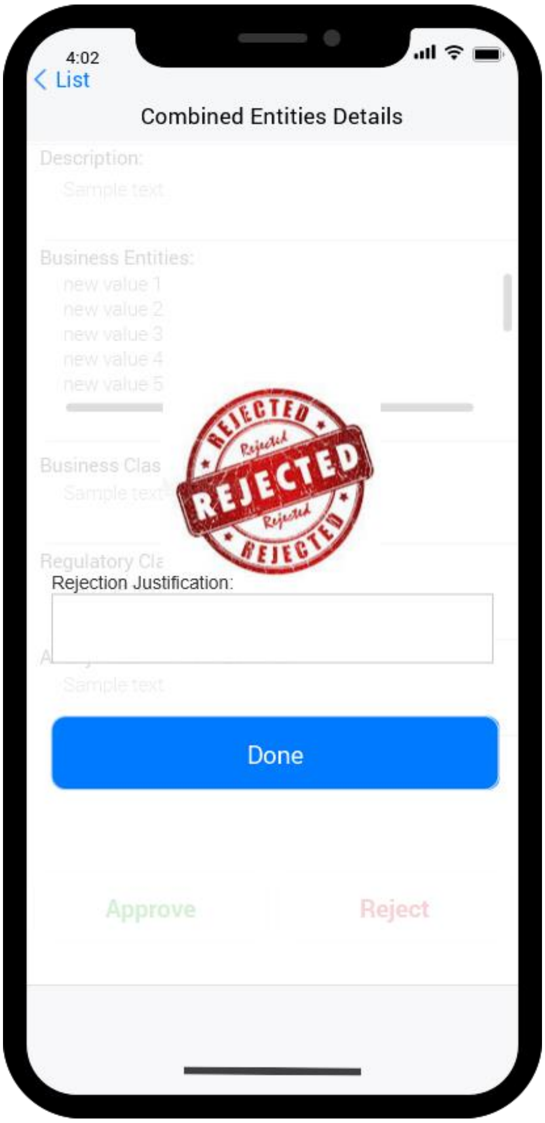
Sample text

Regulatory Cla

Sample text

Anonymization

Sample text



4:02

List

Combined Entities Details

Description:

Sample text

Business Entities:

new value 1
new value 2
new value 3
new value 4
new value 5

Business Clas

Sample text

Regulatory Cla

Rejection Justification:

Sample text