2023-10-31

Big Data Confidentiality: An Approach Toward Corporate Compliance Using a Rule-Based System

Vranopoulos, G

10.1089/big.2022.0201
Big Data
Mary Ann Liebert Inc

Big Data Confidentiality: An Approach toward Corporate Compliance using a Rule-Based System.

Georgios Vranopoulos,[1*] Nathan Clarke,[1] Shirley Atkinson[1]

Abstract – 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, especially for highly regulated sectors like health or finance/banking. Data handlers like Facebook and Amazon have already sustained considerable fines or are under investigation due to violations of data governance. The era of Big Data has further intensified the challenges of minimising the risk of data loss by introducing the dimensions of Volume, Velocity and Variety into confidentiality. Although Volume and Velocity have been extensively researched, Variety, "the ugly duckling" of Big Data, is often neglected and difficult to solve, thus increasing the risk of data exposure and data loss. In mitigating the risk of data exposure and data loss in this paper, a framework is proposed to utilise algorithmic classification and workflow capabilities to provide a consistent approach towards data evaluations across the organisation. 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. The framework includes an exception handling process with appropriate approval for extenuating circumstances. The system was implemented in a Proof of Concept working prototype to showcase the capabilities and provide a hands-on experience. The Information System was evaluated and accredited by a diverse audience of academics and senior business executives in the fields of security and data management. The audience had an average experience of approximately 25 years and amasses a total experience of almost three centuries (294 years). The results confirmed that the 3Vs are of concern and that Variety, with a majority of 90% of the commentators, is the most troubling. In addition to that, with an approximate average of 60%, it was confirmed that appropriate policies, procedure and prerequisites for classification are in place whilst implementation tools are lagging.

Keywords – Big Data, Variety, Data Confidentiality, Data Governance Data Loss Prevention, Data Exposure, Anonymization, Depersonalization, Deanonymization, Reidentification

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.[1] In an approach to adequately protect data, it is imperative to know the data characteristics and understand the aspects/dimensions of Big Data.[2] Laney, often referred to as the father of Big Data, had introduced three dimensions that characterise it.[3] *Volume* is the first and refers to the amount of data created and stored in the digital universe.[4] The second is *Velocity*, which refers to the speed at which data changes in Big Data environments. Last but not least is *Variety*; this characteristic has to do with the nature of the data itself and the manifestations it can pertain to.

Volume and *Velocity* have had a significant research focus during recent years; however, *Variety* revealed a different trend.[5] *Variety* has proven challenging to overcome.[6–8] There are no technological tools available to deal with the proliferation of data from many sources (e.g. internal and external, public and private), thus making *Variety* resilient

[1] School of Engineering, Computing & Mathematics, University of Plymouth

[*] Address correspondence: Georgios Vranopoulos, School of Engineering, Computing & Mathematics, University of Plymouth, Portland Square, Drake Circus, UK – Plymouth PL4 8AA.
eMail: Georgios Vranopoulos, Georgios.Vranopoulos@plymouth.ac.uk
Telephone: +44 1752 585858

to software solutions and heavily relying on human effort.[9,10] The plethora of different formats, standards and notations have also increased the risk of identifying personal information, thus augmenting the risk of disclosure.[2] The significant human dependency and the risk of data loss imposed by *Variety*, alongside the desire of organisations to fully utilise the data for competitive advantage, is a critical challenge. In overcoming the limitation and securing the data towards safe processing, the corporate environment requires a structured approach in minimising the risk with a standardised and automated proposal.

The data confidentiality challenge is driven by regulatory requirements.[11] As it is imperative to govern data custodians activities through legislative frameworks, information is increasingly being regulated in multiple sectors.[11] 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 anonymisation standards like European Medicines Agency Policy 0070 or the Health Insurance Portability and Accountability Act.[12–17] 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.[18,19] 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 business community, leading to the addition of *Value* as one of the essential V's of Big Data.[20,21] 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 important aspect of big data protection.[1] The increasing use of data along with the evolving regulatory frameworks present an area of concern for organizations to effectively manage in avoiding associated risks. Towards this requirement an opportunity for automation and standardization has been identified. Most organizations have embarked on a journey with their Information Security Office (ISO), Chief Data Office and Information Technology departments in establishing the required policies and processes. Having this in place, organizations require a framework to automate the process and minimize human intervention which could lead to errors or impartiality. The corporate cost of not sharing and utilizing the data, as per the Big Data Era adoption has indicated, could limit the businesses competitive advantage.[22] To that end, having identified the need and associated cost, the opportunity of automation was investigated.
The paper proposes a framework towards 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. In the following sections, the challenge posed will be presented, followed by the proposal, including the system proposal, the prototype, and the evaluation of the approach.

2. Background Research

In the context of Big Data, due to real-time processing and the massive quantum of data, new challenges pertaining to security risks and vulnerabilities have come into play.[23] To date, with respect to the issue of confidentiality, most studies have focussed upon the development of anonymization techniques. With much of the research in anonymisation focussed upon static data, rather than unstructured data which is a substantial portion of Big Data Lakes.[24] Taking into consideration Big Data *Variety*, which introduces further data uncertainty, the problem is further exacerbated. Little work has been done in the proliferation of data and research in academia and industry is limited.[7,25] As Big Data analytics thrive, apart of the legal and governance framework (e.g. General Data Protection Regulation), ethical considerations in respect to the privacy and dignity of human beings are of major concern even in cases where consents for personal information usage are collected.[26–28] In addition to these challenges highlighted by researchers, the legal and regulatory framework is constantly changing. The Open Data initiative, along with several other regulatory changes over the last 10 years, have been augmenting researchers access to confidential information. Although everyone agrees that sensitive information should not be proliferated, the extent of the data

protection is covered by different legal guidelines which can be confusing and lead to unintentional exposure.[29] The exposure risk along with the aforementioned elevated access, has created the necessity to design a framework that will enforce strong and transparent governance on access.[30]

Table 1. User Attributes' Types Definitions

Type	Definition / Description
Identifying Identifiers	Attributes that uniquely identify a subject within a set.
Quasi Identifiers	With the use of additional information and correlation can reveal the identity of a subject.
Sensitive Attributes	Attribute/Information related to a subject that should not be disclosed
Non-Sensitive Attributes	All attributes that do not fall on prior types and cannot be used to identify a subject.

Anonymisation is the irreversible process of altering the data in such a manner that it is impossible to identify the person to whom the data refers to, at least in principle.[26] There are 3 types of attributes in any data pertaining to privacy affecting confidentiality as described in Table 1. Pseudonymisation is a reversible anonymisation where data are encrypted and with the use pseudonyms the process can be reversed, Table 2 outlines the various anonymization levels. In attaining the destruction of the link between the data and the person it pertains to, tow technics are most commonly used.[31]

- Compete deletion or alteration of the characteristics by which means the link becomes none identifiable pointing to several subjects.
- Increasing the number of similar subjects in which case the link is no longer unique.

Table 2. Classification Levels for Anonymisation

Classification	Definition
Pseudonymised data	The identifies are replaced by an random code or hashed value.
Pseudo-anonymised data	Any identified is simply deleted from the data set. Other linked data sets can utilised quasi identifies in re-identifying the data.
Anonymised data	All correlation and identifying information are removed making re-identification impossible.

Many studies argue that it is impossible to have 100% certainty a set cannot be de-anonymised.[26] Since the process cannot be 100% safe it is imperative to include in the framework detailed reference information in order to have traceability which can be used in countering the four threats to confidentiality.[32]

- Re-Identification or De-Anonymization, which refers to the capability of an attacker to identify the identity of an information subject although the information shared are supposedly anonymized.
- Reconstruction, is the risk of reconstructing the data set from the partially distributed one, like aggregated statistics.
- Membership Disclosure, is similar to re-identification with the difference that the subject is not identified but rather its membership to a group can be ascertained.
- Cryptanalysis, is mainly related to pseudonymisation, where utilizing frequency analysis the attacker can correlate and decipher encrypted data.

The analysis of literature indicates a growing concern for the requirements of confidentiality. These challenges stem from legal and compliance along with ethical requirements. It is apparent that Big Data era has intensified these challenges where *Variety*, being the least researched V is attributing highly. In taming the risks researched agree there has to be a framework. Scholars have suggested and implemented several techniques to ensure Big Data privacy throughout the Big Data life cycle, which comprises different stages, i.e. generation, collection, storage, processing, analytics, utilisation and destruction.[33,34] In the initial stages of generation and collection, most scholars would target to limit the risk exposure by restricting access or restricting information.[11] The restriction of

information, which is essentially nothing more than the abolishment or falsification of confidential and personal information, is usually suggested by many authors to be achieved with anonymisation.[33,35] In achieving anonymisation or depersonalisation 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 anonymisation or k-Anonymity).[36–41] 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 in reaching the balance between data disclosure and data loss.[36,42–44] 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. Currently, all 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" should be safeguarded has not been referenced in the research. Research and most academic work focus on achieving anonymisation with sophisticated techniques rather than identifying the element to be anonymised. Theoretical and practical implementations regarding data confidentiality and their enforcement are unsatisfactory.[45] Label systems mainly focus on enforcement of security and data access rather than identification and dissemination.[46] 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.[47] The identification of data elements and the decision on what has to be done with/on them is imperative for any organisation since it is the primary measure in countering the incompliance risk.

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

The paper 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:

a) to suggest automation of the process with a software-driven, algorithmic rule-based system.
b) to suggest alternatives in minimising the Data Loss risk for an organisation through standardisation
c) 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 to 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 3. The system will store the definition of all the data elements of the corporate data dictionary in formulating a data confidentiality corporate taxonomy utilising an automated ingestion and identification process.[48]

Table 3. BD-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 with 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 with 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 an application, One-Time-Password delivery number). In some instances, the presence of multiple entities as a group might necessitate the need for 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 in 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 entities review and confirm the ratings for each attribute of every entity. To that end, the system (illustrated in Figure 1) 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 approvals 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.

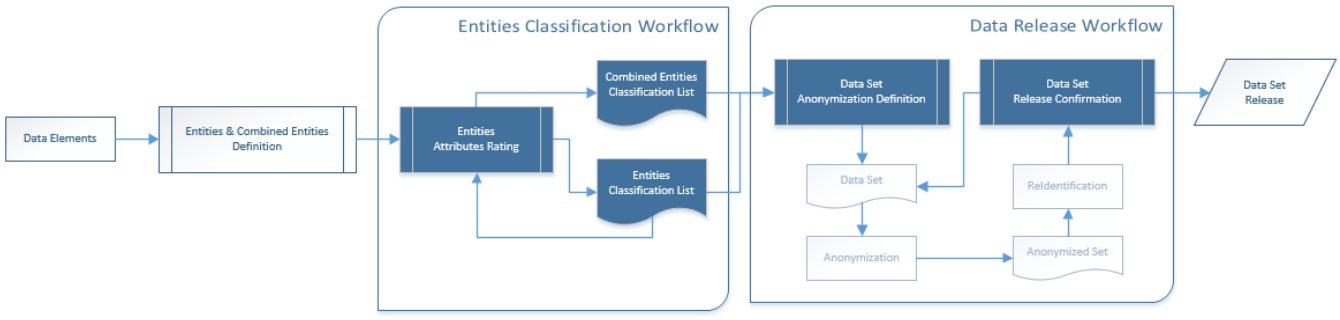

Figure 1

The proposed system provides a configurable environment where the different subject matter experts from business, risk, security and other control disciplines, for example 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 is provided by the data owner indicating the operational use of the data.
- **Regulatory Classification**, which is the classification that is derived from industry, country or international regulatory requirements, acts and laws.
- **Anonymisation Risk**, which will provide with classification relevant to anonymisation and re-identification risks.

The proposed attributes are independent in nature since they are designed to identify different aspects of the data classification 360 view. The owner (business) perspective is addition to the corporate (regulatory & risk) perspective are segregated in depicting the different viewport by which the classification can be viewed upon. Invariably since they classify the same entity there is correlation but only towards the describing entity and not amongst them.

Once the corporate classification policy is enriched with the classification and depicted, the system will employ algorithmic rules in identifying if the proposed anonymisation 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 along with some examples are provided in Table 4.

Table 4. 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.
Anonymisation 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.

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 with the 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, but not overwhelming, leading 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 centred on the concept of segregation of duty, commonly referred to as "duality."[49] 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.[50] In this system, the 6-eyes principle instead of 4-eyes is proposed due to its importance and criticality of the system. Additionally, the workflow will enable the engagement of all related parties into 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 anonymisation 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 in a) aggregating underlined entities ratings for combined entities and b) auto-calculating the baseline for elements anonymisation action as a corporate standard and as a run-time feature for each dataset evaluation.

The first suggested algorithmic implementation of minimum levels for combined entities is proposed in facilitating 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 5. They are calculated based on the maximum of all underlying entities for business and regulatory classification and with the advancement of one level for anonymisation risk. For anonymisation risk, the advancement of one level was implemented since the combination of multiple underlying elements will increase the exposure risk.[51]

Table 5. Combined Entity Minimum Levels Algorithms

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

A worked example of the calculation is presented in Table 6. 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 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 6. 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 anonymisation. Similarly to combined entities, when the anonymisation actions have to be selected, the system will suggest a minimal level restriction based on the configuration.

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 policies through exposing users to the corporate policies on anonymisation via the automated approach. . The users will be able to immediately get feedback on organizationally approved practices and guidelines with respect to the selected data element anonymisation 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 2, 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 2). Based on the underlying numeric equivalent, a weighted average is calculated. The weights for the weighted average calculation can be parameterised in the system. 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 2), 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.

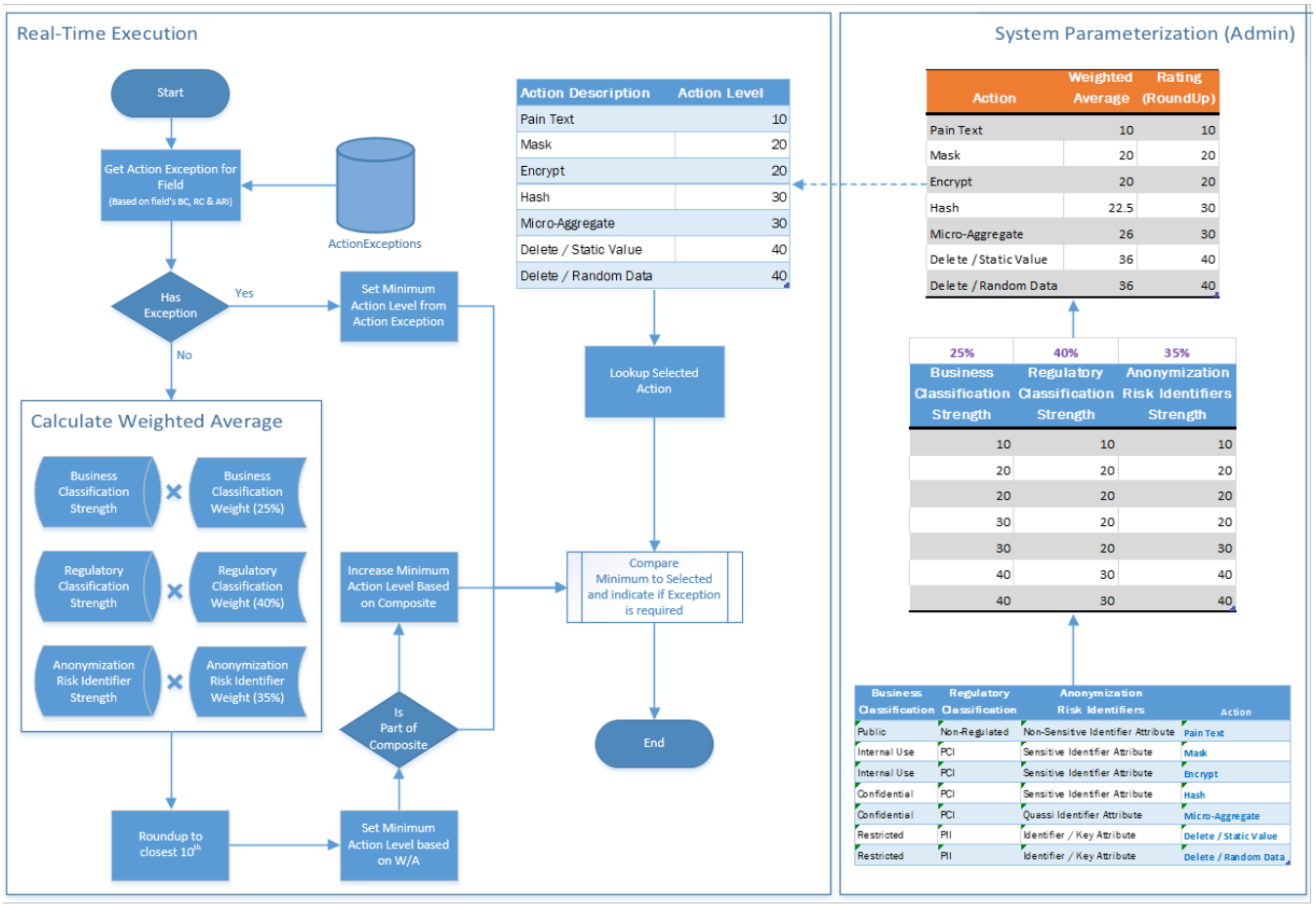

Figure 2

In the system we can identify that there are two parameterized scales, a) the classifications scale and b) the attributes weights. The classification scale is independent for each classification attribute and for ease of use has a literal and an equivalent. In Table 7 a sample of the attributes scales is presented as utilized in the PoC. The scales can be adjusted per organization in respect to the levels (number of options) and respective scales (individual numeric values e.g. scales 0-4 or 0-40 etc.). For simplicity the PoC utilized a uniform scale amongst all attributes. The weights per attribute have been used in order to be able to consolidate all three independent attributes into a consolidated metric that will be used in all calculations for numeric comparison, the "action strength". Similarly the percentages to be utilised for the weighted average are set by the organization in reflecting the importance of each aspect to the business. For instance in a highly regulated environment like banking, regulatory classification would tend to be higher. The risk appetite of the organization or each classification attribute is in essence reflected by these percentages.

Table 7. Attributes Classification Sample

Business Classification	Regulatory Classification	Anonymization Risk Identifiers	Numeric Equivalent
Restricted	Sensitive PII	Identifier / Key Attribute	40
Confidential	PII	Quasi Identifier Attribute	30
Internal Use	PCI	Sensitive Identifier Attribute	20
Public	Non-Regulated	Non-Sensitive Identifier Attribute	10
(Empty)	(Empty)	(Empty)	0

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 cover a set of functionalities, including a) immediate interactive calculations, without going to the backend server for recalculation, b) capability to request or revoke exceptions, c) capability to view at which 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.

In understanding the classification procedure and associated features, an illustrative example will be presented. The example will reference the values outlined in Table 7 and Figure 2. The corporate has identified that the low risk elements can be disseminated as pain text. By low risk, they have identified that the business classification should be "Public", the regulatory classification should be "Non-Regulated" and the anonymization risk identifiers should be "Non-Sensitive Identifier Attribute". The numeric equivalent for all classifications rating is 10. Thus with the use of the weighted average $((0.25 \times 10) + (0.40 \times 10) + (0.35 \times 10) = 10)$ the action strength for "Plain Text" classification would be 10. Table 8 show the definition and the calculation of the action strength for a set of example business entities. Based on the aforementioned definition the only entity that can be shared using "Plain Text" would be "Terminal Number" which has the same action strength. All the other sampled entities will require a higher anonymization depersonalization action since their action strength is higher.

Table 8. Sample Business Entities Classification

Business Entity	Business Classification	Regulatory Classification	Anonymization Risk Identifier	Action Strength
Party Name	Confidential	PII	Identifier / Key Attribute	40
Party Date of Birth	Internal Use	PII	Identifier / Key Attribute	40
Party Salary Amount	Confidential	Non-Regulated	Sensitive Identifier Attribute	20
Party Work Address Street	Internal Use	Non-Regulated	Quasi Identifier Attribute	20
IBAN Account Number	Public	Non-Regulated	Identifier / Key Attribute	30
Terminal Number	Public	Non-Regulated	Non-Sensitive Identifier Attribute	10
Geolocation	Public	Non-Regulated	Quasi Identifier Attribute	20

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 will define the anonymisation or depersonalisation action to be assigned against each field of the data set.
- **Release Anonymisation** is the stage where the anonymisation or depersonalisation actions are implemented. The implementation can be done with corporate proprietary systems or subcontracting to an external trusted entry. Since the data set is not yet anonymised or depersonalised, strict rules and contractual agreements should apply in case of external entities involvement. In this stage, the re-identification or de-anonymisation 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 anonymisation 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 for the parameterisation of the weights between the classifications attributes and the association of the anonymisation/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 in segregating the inputter and authoriser function. In this way, the two-step process will further minimise the risk of erroneous or unlawful alterations regarding access and global BD-CPS parameterisations.

4. The Prototype

The purpose of the experiment and Proof of Concept (PoC) using a working prototype is to evaluate the value a Data Confidentiality framework would bring to an organisation and its users. The organisation utilizing 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. user management module, mobile push notifications or elaborate audit trails were not implemented, 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,[52,53] but the adoption of mobile technology is also very high, and COVID-19 has intensified the adoption.[54,55] 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 3). The **Web Application** targets users who use their PC's, and its User Interface (UI) is elaborate. The application provides capabilities in managing the configurations, data entry for classification, data request and release process. **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.

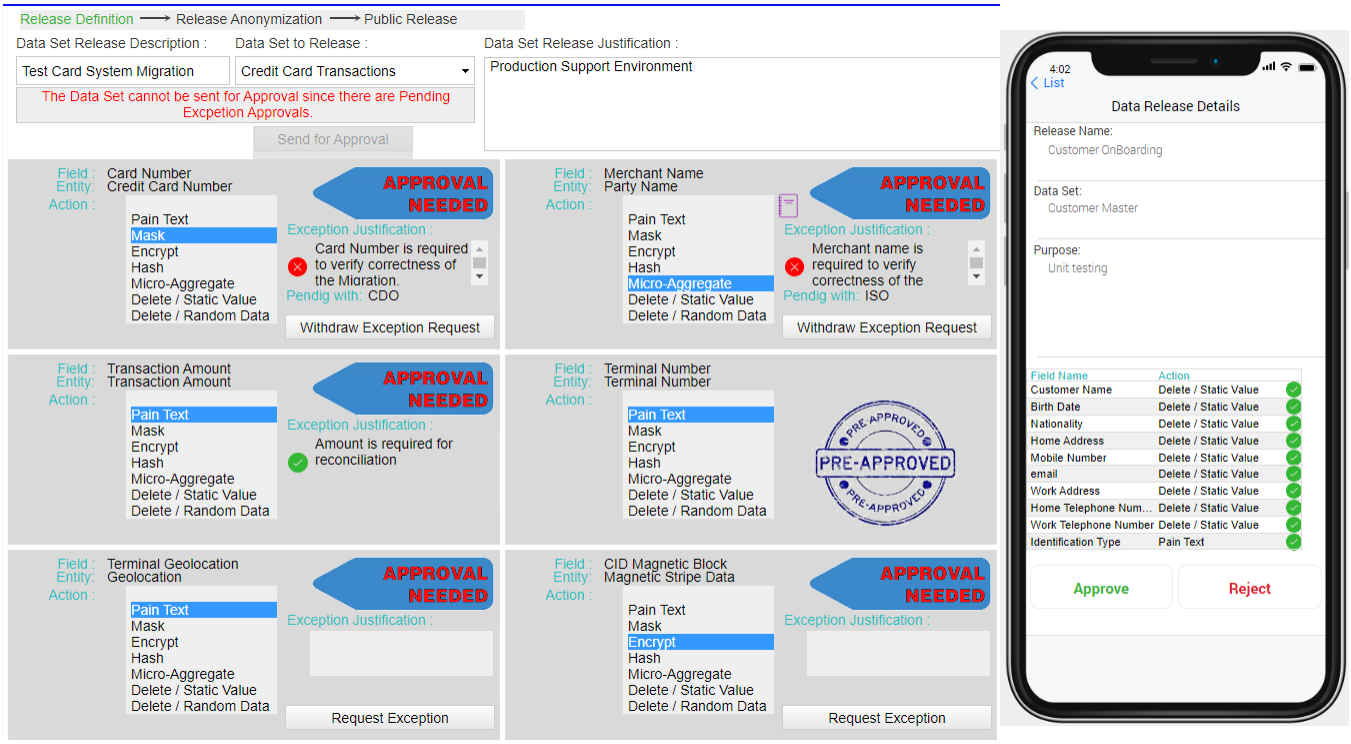

Figure 3

A use case has been utilised to understand better 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 put in place. 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) **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 anonymisation risk. The mix is suggested since the impact of any regulatory or anonymisation risk in that order can have profound legal/compliance implications and penalties.
c) **Minimum levels association** between the actions and the classification attributes. A visualisation from the PoC is provided in Figure 4.

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 4

Based on *Data Origination* of the data ingestion journey, the fields and entities of a dataset can be automatically identified.[48] The respective identified information can become an automated feed to the proposed system. In addition, information available at the corporate in Information Classification Policy can be imported into the system. Having the basic information available, configurations and fields/entities, the next configuration level will have to be performed. If available, the classifications can be imported from the institute's already defined Data Classification policy; otherwise, the evaluation and classification of the entities about the aforementioned *Classification Attributes* will commence manually in the system.

Using the web interface, see flow in Figure 5, the DA will search for any entity and define the Business, Regulatory and Anonymisation Risk. Once the levels have been identified, an approval request will be sent to CDO on the mobile applications 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. In case the approver is not in agreement, there is a rejection option whether the approver can also record the comments and rejection justification. The same approval process will be automatically triggered for the next level of approval, ISO. The system is fully parameterised in implementing the rule-based corporate-wide evaluation of datasets for internal or external dispatching.

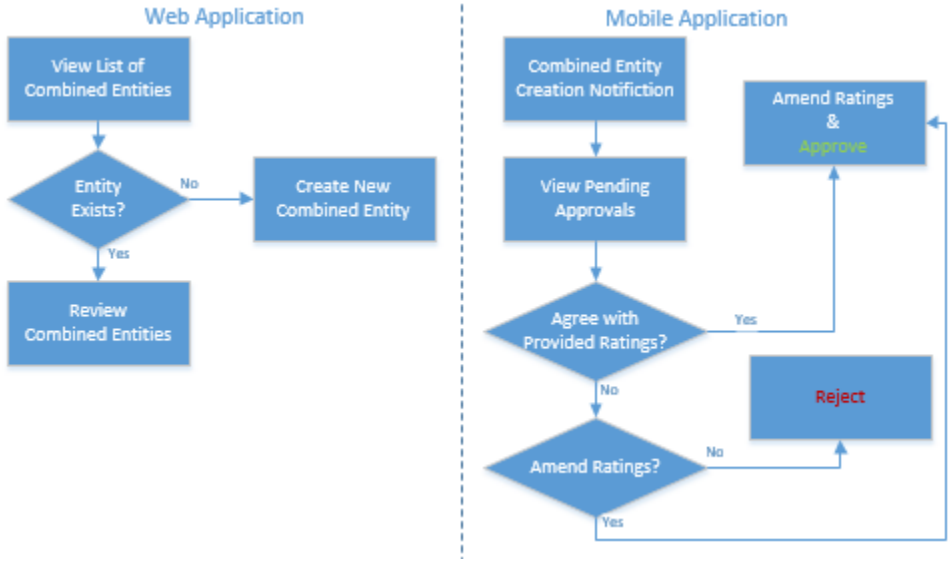

Figure 5

A credit card migration project is utilised to showcase the data dissemination process and prototyped UI, exhibited in Figure 6. In this use case (highlighted in blue parallelogram), the credit card transaction will have to be moved from the production environment to lower environments (development or test environments) to facilitate and verify the migration process's correctness. In authorising the data sharing, the approval level is implemented in three stages, visualised on the form in the red parallelogram. For all the stages of approvals, the mobile application is

used whilst for editing the web application is used. The web application was selected since many calculations are to be implemented, and a substantial amount of information is exhibited. A mobile application interface would be complex and fragmented, leading to user confusion, omissions, or errors. The form, presented in Figure 6., is showcasing (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 d) enquiring for prior exceptions approvals/rejections for the element.

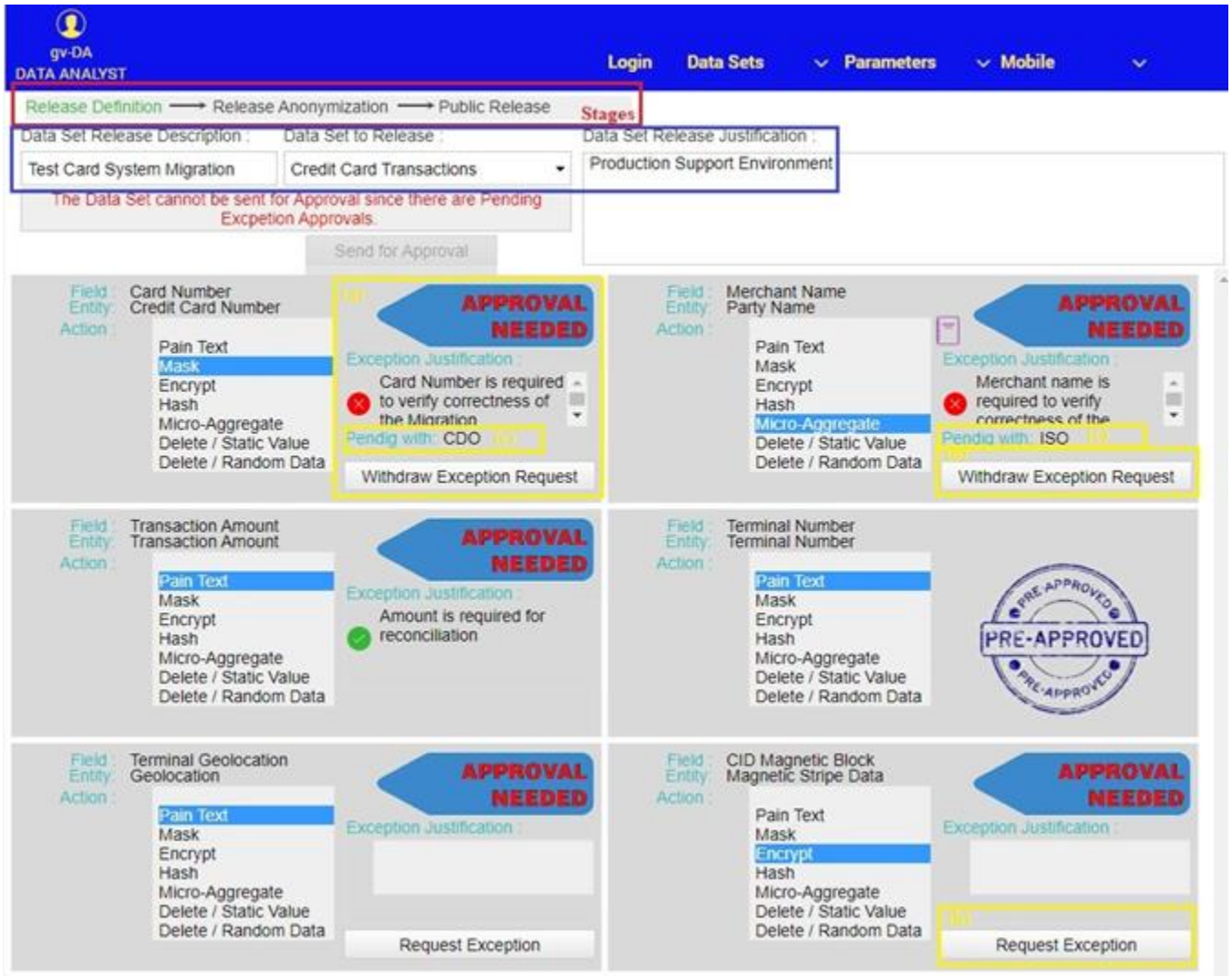

Figure 6

The stage is available on the top of the screen, red parallelogram in Figure 6 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) required adjustment, b) required exception, c) recommendations or d) confirmation to proceed. When all fields are associated with a preapproved 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 mentioned above process for the (Combined) Business Entities. However, the approvers cannot edit the proposed action levels in this case. The reason being that the approvers have already accepted any deviations to the corporate policy by approving the individual exceptions

and now will have to approve the entire set. In case the DA is trying to distribute an already approved release of a dataset for the same purpose, 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. Post the anonymised/depersonalised process, the set will have to be submitted to an external engine, which can be one of the available public service providers, in evaluating the anonymised/depersonalised and calculate the re-identification factor/per cent of the set. The BA 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 anonymised/depersonalised is achieved. The set has reached its final **3rd stage**, where the data set is ready to be distributed.

What was exhibited by know is that the system will provide the reusability of the configurations 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, with 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 on any data release, along with an assurance that the system will preserve the minimal levels already defined in the system.

5. Evaluation

5.1 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 gaged 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 difference between the approaches is in many respects, and 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 from experts a rich response with in-depth retrospect and outlook on the proposed concepts designed.[56–58] In ensuring a holistic review, experts from both academia and industry were invited. The diverse origin and different accumulated experiences from 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 background and experience, the geolocation diversity and cultural background were considered by involving experts working in multiple countries, namely 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 Section 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 set of questions, see Table 10. The audience mix, see Table 9, was decided to be diverse in covering from 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.[59–61]

Table 9. 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 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 outlining the initiative's details and the 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 45min presentation, a set of rating and open-ended questions were discussed.

Table 10. 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 anonymisation 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 "Anonymisation 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 anonymisation/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

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 10. The questionnaire was formulated in facilitating the discussion and acquiring insight from the experts. During the discussion, the focus of the insight was to explain the concepts, acquire feedback on the subject by the experts and drive experts in providing further suggestions of enhancements. A 5-point scale similar to psychometric Likert and Five-Star quality rating grade was selected since it is easy and many people have prior experience with it. The five levels gauge instead of three was used, although it has minimal statistical impact or value since we are interested in individual behaviour.[62–64]

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 not only the response content but also 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.

5.2 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 cases 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.

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 anonymisation, thus often reducing their value. On the other hand, business originating commentators were on the sharing side where the concerns confirmed stemmed from the regulatory and security perspectives.

In acquiring further context, the commentators were asked pointed questions they had to provide a rating, see Table 10. The responses are tabulated in the heat map presented in Figure 7. 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 7

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 towards the middle of the scale. This is of vital importance towards 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 anonymisation 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 on all parties, and the process will be smooth. In addition to that, if there is a high-level strategy for anonymisation and depersonalisation, it would mean that the engaged 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 anonymisation 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 algorithmic 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 be increasing 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 anonymisation 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 the use of 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*. Indicative of the consent is the wording used; *"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 upward (ceiling function) roundup in increasing controls and reducing the risk of the assigned Action Strengths. For the anonymisation 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 3, 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 interests of the organisation 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. Through-out 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) proposal where the organisations engaged can, should they choose to, share information amongst them. In this way, based on a sectoral classification, the system can provide templates and proposed values, percentages, classifications, actions or levels for the participant 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 anonymisation 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;[65] towards that end, the system will have to provide hooks or addins for the most commonly used business applications.

6. Discussion

DLP has been identified as a matter of concern, based on the imposed risk, for all corporates. In mitigating the risk organisations put in place end-point controls. That could include solutions which will scan documents, e-mails etc. and highlight privileged information. Nevertheless there are many limitations in identifying the attributes and the substantial set of false positives hinders the process.[48] In many cases corporate policies will enforce DLP procedures in external communications, e.g. sending mails outside the organization, and block such communication. Such barriers, present a different communication risk since due to false-positives the communication will be falsely quarantined without the initiator's knowledge, and as such important matters can be delayed. Most of these tools depend on personal judgments and in many cases individual classification, e.g. document or e-mail self-classification as confidential, which is not standardised nor corporate policy driven. The risk of misclassification due to lack of understanding in respect to legal and/or business aspects, or even worse, intentional misclassification, is not centrally managed and compliance is not monitored nor safeguarded. Another level of data distribution is related to actual data sets and information proliferated from the organization. These sets could be related to reports that are publicly available, data sets exchanged with other organizations, sets pertaining to data collected on cloud for customer behaviour or even sets that are moved within the organization from the production environment to lower environments like development. For these data exchanges and movements, as confirmed also by the commentators, there is adequate process understanding and policies and procedures are in place but there is limited, if any, set of tools to facilitate and manage the process . This limitation in essence will suggest that the work is done manually in a non-standardised manner with little governance and minimal auditing.

In proliferating any kind of information a common approach would be to proceed with a manual approval process. Based on the exhibited lack of tools, automation and non-standardization, there is reluctance from all control agencies within an organization in taking ownership and approval. Often it can be identified that in the business environment, the request will mostly oscillate between ISO and CDO. Furthermore, in several cases Risk, Compliance, Internal Audit and Human Resources departments will be involved adding to the complexity of the evaluation. The communication will be fragmented in multiple e-mails and by no means can there be any reference to prior cases of already proliferated and anonymized sets. In many cases the final decision will have to be taken by the data owner, in which case being business, could luck the understanding and technical depth to evaluate the request, especially when it comes to correlated or interrelated data sets. When it comes to Big Data and the manifestation of *Variety,* it could be very difficult to have a documented and risk free approach unless it is corporate wide, standardised and process wise safeguarded by involving the appropriate skillset at the appropriate approval level.

Most organization that embark on such a compliance and loss prevention journey will focus on prevention and training at the distributor level by imposing policies. This approach will introduce red tape and focus on prevention measures rather that automating and expediting the process. Inherent risks of impartiality, intentional misguidance and misclassification will not be addressed. Ownership and accountability along with track records will not be recorder and safeguarded in any system thus hindering the process's integrity and reusability. Based on Covid-19 limitations in respect to direct human contact and group meetings, most organization will provide computer assisted training toward data classification, identification and the regulatory framework in an attempt to increase awareness. This educational material, although a very important and essential step, is generic and thus can only serve as the first step of the journey. Additional specialized training will have to be provided in respect to each user's needs, work responsibilities/job description and skills. To that end, an event driven on the job training could be a plausible approach.

The framework presented is utilized in remediating some of the aforementioned risks and process inefficiencies. With the utilization and use of BD-CPS, auditability and tractability are ensured. The right resources with appropriate skills are assigned to the approval levels in safeguarding that approvals are provided by the most suitable practitioner. Standardization and sustainability is another important dimension in which BD-CPS is contributing. With the use of a centralized repository, conformity on the rules is achieved whilst reusability it facilitated. In this way the organization will be much safer and reduce the risk exposures in comparison to the manual and unstructured approval process. Having digitized the journey and recorded all elements, be it attributes or approvals, the organization can achieve compliance. Taking the next step, it is of outmost importance to proceed in assisting the user within the process and streamline it. To that extend DB-CPS, as in detail presented, has introduced several automations. Main focus is to increase task specific user awareness, provide with on the job training and in general assist the users in making the right decision within a minimal timeframe. The use of suggestion lists that are on the fly calculated based on the user selection and predefined policies, guide the user in achieving operational efficiency both in terms of accuracy and timeliness of the decision. Workflows and approval levels along with 4-eyes, or 6-eyes, principle will safeguard the user form accidental mistakes whilst will also serve as a validation mechanism for possible intentional mistakes. But the automations extend beyond only the end user/data distributor. BD-CPS will use algorithmic facilities in assisting the control function representatives to formulate and preserve the corporate policies. Changes can be quickly and easily applied in respect to the weights utilised, thus tuning the system toward the corporate risk appetite and residual risk acceptance. Last but not least prior approved sets and set elements/attributes can easily serve as examples and guidelines for future use. The fact that even not approved decision along with their justification are captured can and will serve as lessons learned for all users at all levels.

BD-CPS is an approach towards a framework but just like any other suggestion can be affected by environmental limiting factors and could also be augmented or enhanced. Since compliance is a regulatory factor but not an income producing initiative, there is probability that senior management sponsorship will be lagging thus delaying implementation. As long as DLP is imposed, even in the form of restrictions, a profit bearing project might be prioritised over such an initiative. When there is reference to awareness and education, which are time consuming processes, the timelines are extensive for such a project and cannot be considered as a quick win which once again could lead to reprioritization over other business critical projects. On the other hand the introduction of new technologies and systems stemming from the Artificial Intelligence (AI) and Machine Learning (ML) can greatly enhance BD-CPS by utilizing them to further guide or even predict the user/approver responses. The very fact that Big Data is involved could retrofit in to the system. In such an enhancement, community modelling can be utilised.

7. Conclusion

While organisations strive to attain a competitive advantage through big data and analytics by utilising data sets to their full extent, data loss is becoming a serious challenge. Organisations in many sectors, especially the highly regulated ones, are unable to utilise the data to their full potential over the fear of data compromise and breach of legislation and regulations. Governmental and international organisations impose standards on data sharing and disclosure, which can entail monetary fines if not adhered to. Based on this imposed regulatory governance,

organisations face the challenge of Data Confidentiality. With the use of Business Classification, Regulatory Classification and Anonymisation Risk attributes, the BD-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.

Future advancements of the systems can be obtained once the system and framework are adopted and implemented within an organisation, preferably in a large one within a highly regulated industry. The adoption will provide insight into the system's realignments and methodology to gain easier adoption. Possible extensions and additional submodules might need to be introduced to serve needs that the adoption will uncover. The respective implementation along with evaluation will span in time since a) planning and implementation along with corporate and legal processes have to be followed (project timelines will exceed one year), b) the adoption process for any large organisation will be slow, c) the critical mass for evaluation will not be readily available since the usage will increase with time extending to multiple departments and functions.

Declarations

Ethics approval and consent to participate

Issued by the University of Plymouth Research Ethics Application Approval - Faculty Research Ethics and Integrity Committee reference number 2862, "Tackling Big Data Variety using Metadata - Corporate Data Confidentiality using a Rule Based system". All participants have signed the required consent form after acquiring the Information form as well as confirming their consent during the presentation & interview session.

Consent for publication

Not applicable

Availability of data and materials

Interviews are not available since they are under confidentiality and personal information disclosure acts. The prototype is available in JustInMind site: https://rb.gy/ekm5j1.

Competing interests

The authors declare that they have no competing interests

Funding

Not applicable

Authors' contributions

GV was responsible for authoring and NC along with SA were responsible for guidance and editing.

Acknowledgements

Not applicable

References

1. Rawat DB, Doku R, Garuba M. Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security. *IEEE*. 2019;14(6).
2. Cuquet M, Vega-Gorgojo G, Lammerant H, Finn R, Hassan U. Societal impacts of big data: challenges and opportunities in Europe. *arXiv - Cornell Univ*. Published online April 11, 2017.
3. Laney D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Appl Deliv Strateg*. 2001;949:4.

4. Ali-Ud-Din Khan M, Uddin MF, Gupta N. Seven V's of Big Data understanding Big Data to extract value. *Proc 2014 Zo 1 Conf Am Soc Eng Educ - "Engineering Educ Ind Involv Interdiscip Trends", ASEE Zo 1 2014*. Published online 2014:1-5. doi:10.1109/ASEEZone1.2014.6820689
5. Vranopoulos G, Triantafyllidis A, Lefteriotis K. Big Data Variety, "Where Do we Stand", An Overview of Big Data and the Variety Challenge. *Int J Manag Appl Sci*. 2020;6(3).
6. Lennard. Data Variety - The Ugly Duckling of Big Data. 2014;1. http://www.datashaka.com/blog/non-techie/2014/01/data-variety-ugly-duckling-big-data
7. Rui M, Honglong X, Wenbo W, Jianqiang L, Yan L, Minhua L. Overcoming the Challenge of Variety : Big Data Abstraction , the Next Evolution of Data Management for AAL Communication Systems. *IEEE Commun Mag*. 2015;(January):42-47.
8. Brown E. Big Data Problems - Variety not Volume. Big-Data Forum. Published 2014. Accessed September 2, 2015. http://www.big-dataforum.com/605/big-data-problems-variety-not-volume
9. Trader T. Big Data Future Hinges on Variety. datanami. Published February 24, 2014. http://www.datanami.com/2014/02/24/big_data_future_hinges_on_variety/
10. Kimura C. Beyond the "Big": Solving for Data Variety Requires New Thinking - ClearStory Data. CrearStory Data. Published 2014. Accessed September 2, 2015. http://www.clearstorydata.com/2014/12/beyond-big-solving-data-variety-requires-new-thinking/
11. OKeefe C. Privacy and Confidentiality in Service Science and BigData Analytics. *Springer*. 2017;457:978-981. doi:10.1007/978-3-319-18621-4_5ï
12. Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC. Published 1996. Accessed December 11, 2021. https://www.cdc.gov/phlp/publications/topic/hipaa.html
13. European Medicines Agency. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. Published online 2017. Accessed December 11, 2021. www.ema.europa.eu/contact
14. Official PCI Security Standards Council Site - Verify PCI Compliance, Download Data Security and Credit Card Security Standards. Published 2006. Accessed December 11, 2021. https://www.pcisecuritystandards.org/
15. General Data Protection Regulation (GDPR) Compliance Guidelines. GDPR.EU. Published 2020. Accessed December 11, 2021. https://gdpr.eu/
16. European Commission. Implementation of the Public Sector Information Directive. Published July 2022. Accessed December 11, 2021. https://digital-strategy.ec.europa.eu/en/policies/public-sector-information-directive
17. U.S. Department of Justice - Office of Privacy and Civil Liberties. United States Department of Justice Overview of the Act of 1974 - 2020 Edition Preface. Published online 2020. Accessed December 11, 2021. https://www.justice.gov/opcl/overview-privacy-act-1974-2020-edition.
18. Cristina Abellan Matamoros. Facebook to pay record $5 billion fine over privacy violations, but are they getting off lightly? Euronews - REUTERS. Published July 24, 2019. https://www.euronews.com/2019/07/24/facebook-to-pay-record-5-billion-fine-over-privacy-violations-but-are-they-getting-off-lig
19. Lovejoy B. GDPR fines total €56M in first year as Facebook faces 11 investigations. 9To5Mac. Published May 28, 2019. https://9to5mac.com/2019/05/28/gdpr-fines/
20. Yang C, Huang Q, Li Z, Liu K, Hu F. Big Data and cloud computing: innovation opportunities and challenges. *Taylor Fr - Int J Digit Earth*. 2017;10(1):13-53. doi:10.1080/17538947.2016.1239771
21. Davenport TH. *The Human Side of Big Data and High-Performance Analytics*.; 2012.
22. Madeline G. Strategies for Implementing Big Data Analytics. Published online 2013.
23. Gopala M, Sriram K. SECURITY CHALLENGES OF BIG DATA COMPUTING. *Int Res J Mod Eng Technol Sci*. 2022;4(1). Accessed February 19, 2023. www.irjmets.com
24. Csányi GM, Nagy D, Vági R, Vadász JP, Orosz T. Challenges and open problems of legal document anonymization. *Symmetry (Basel)*. 2021;13(8). doi:10.3390/sym13081490

25. Georgios Vranopoulos. Tackling Big Data Variety using Metadata. Published online 2022.
26. Rossi A, Arenas MP, Kocyigit E, Hani M. Challenges of protecting confidentiality in social media data and their ethical import. In: *Proceedings - 7th IEEE European Symposium on Security and Privacy Workshops, Euro S and PW 2022*. Institute of Electrical and Electronics Engineers Inc.; 2022:554-561. doi:10.1109/EuroSPW55150.2022.00066
27. Weaver A. Tourism, big data, and a crisis of analysis. *Ann Tour Res*. 2021;88. doi:10.1016/j.annals.2021.103158
28. Spevakov AG, Spevakova S V., Primenko D V. METHOD OF DATA DEPERSONALIZATION IN PROTECTED AUTOMATED INFORMATION SYSTEMS. *Radio Electron Comput Sci Control*. 2020;0(1):162-168. doi:10.15588/1607-3274-2020-1-16
29. Nittari G, Khuman R, Baldoni S, et al. Telemedicine Practice: Review of the Current Ethical and Legal Challenges. *Telemed e-Health*. 2020;26(12):1427-1437. doi:10.1089/TMJ.2019.0158
30. Silberman R. Developing access to confidential data in France: results and new challenges. *J Priv Confidentiality*. 2021;11(2). doi:10.29012/jpc.788
31. Gazizov A, Gazizov E, Gazizova S. Theoretical aspects of the protection of personal data of employees of the enterprise by the method of pseudonymization. In: *E3S Web of Conferences*. Vol 210. EDP Sciences; 2020. doi:10.1051/e3sconf/202021011001
32. Elkoumy G, Fahrenkrog-Petersen SA, Sani MF, et al. Privacy and Confidentiality in Process Mining: Threats and Research Challenges. *ACM Trans Manag Inf Syst*. 2022;13(1):1-17. doi:10.1145/3468877
33. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. *SpringerOpen - J Big Data*. 2016;3(1). doi:10.1186/s40537-016-0059-y
34. Koo J, Kang G, Kim YG. Security and privacy in big data life cycle: A survey and open challenges. *Sustain*. 2020;12(24):1-32. doi:10.3390/su122410571
35. Rai S, Sharma A. Research Perspective on Security Based Algorithm in Big Data Concepts. *Int J Eng Adv Technol*. 2020;9(3):2138-2143. doi:10.35940/ijeat.c5407.029320
36. Rumbold J, Pierscionek B. Contextual Anonymization for Secondary Use of Big Data in Biomedical Research: Proposal for an Anonymization Matrix. *JMIR Med Informatics*. 2018;6(4):e47. doi:10.2196/medinform.7096
37. Ninghui L, Tiancheng L, Suresh V. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: *2007 IEEE 23rd International Conference on Data Engineering*. ; 2007.
38. Zhou B, Pei J. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Springer - Knowl Inf Syst*. 2010;28(1):47-77. doi:10.1007/s10115-010-0311-2
39. Domingo-Ferrer J, Rebollo-Monedero D. Measuring Risk and Utility of Anonymized Data Using Information Theory. In: *2009 EDBT/ICDT Workshops*. ; 2009.
40. Sei Y, Okumura H, Takenouchi T, Ohsuga A. Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness. *IEEE Trans Dependable Secur Comput*. 2019;16(4):580-593. doi:10.1109/TDSC.2017.2698472
41. Ni C, Cang LS, Gope P, Min G. Data anonymization evaluation for big data and IoT environment. *Inf Sci (Ny)*. 2022;605:381-392. doi:10.1016/j.ins.2022.05.040
42. Gouweleeuw JM, Kooiman P, Willenborg LCRJ, De Wolf P-P. Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *J Off Stat*. 1998;14(4):463-478.
43. Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans Knowl Data Eng*. 2002;14(1):189-201. doi:10.1109/69.979982
44. Oganian A, Domingo-Ferrer J. On the Complexity of Optimal Microaggregation for Statistical Disclosure Control. *Stat J UN Econ Comm Eur*. 2001;18(4). doi:10.3233/SJU-2001-18409
45. Sabelfeld A, Myers AC. Language-Based Information-Flow Security. *IEEE J Sel Areas Commun*. 2003;21(1):5-19. doi:10.1109/JSAC.2002.806121
46. Zheng L. Dynamic Security Labels And Noninterference. *Form Asp Secur Trust*. 2006;173.

47. Blažič AJ, Šaljić S. Confidentiality Labeling Using Structured Data Types. *IEEE*. Published online 2010:182-187. doi:10.1109/ICDS.2010.70
48. Vranopoulos G, Clarke N, Atkinson S. Addressing Big Data Variety using an Automated Approach for Data Characterization. *Springer J Big Data*. 2022;9(8). doi:10.1186/s40537-021-00554-3
49. Manning K. 2 Eyes- 4 Eyes- 6 Eyes Principle | ProcessMaker. ProcessMaker. Published 2020. Accessed October 2, 2021. https://www.processmaker.com/blog/2-eyes-4-eyes-6-eyes-principle/
50. Lamberti H. *Delusion in Organizational Excellence*. McGraw Hill Education; 2013.
51. Armando A, Bezzi M, Metoui N, Sabetta A. *Risk-Based Privacy-Aware Information Disclosure*.; 2015.
52. Hieatt E, Mee R. Going faster: Testing the Web application. *IEEE Softw*. 2002;19(2):60-65. doi:10.1109/52.991333
53. Shan TC, Hua WW. Taxonomy of Java Web Application Frameworks. In: *IEEE International Conference on E-Business Engineering, ICEBE 2006*. IEEE Computer Society; 2006:378-385. doi:10.1109/ICEBE.2006.98
54. Shaw N, Eschenbrenner B, Brand BM. Towards a Mobile App Diffusion of Innovations model: A multinational study of mobile wallet adoption. *Elsevier - J Retail Consum Serv*. 2022;64. doi:10.1016/j.jretconser.2021.102768
55. Taylor DG, Voelker TA, Pentina I. *Mobile Application Adoption by Young Adults: A Social Network Perspective*. Vol 6.; 2011. http://digitalcommons.sacredheart.edu/wcob_fac/1
56. Hyett N, Kenny A, Dickson-Swift V. Methodology or method ? A critical review of qualitative case study reports. *Int J Qual Stud Health Well-being*. 2014;1:1-12. doi:10.3402/qhw.v9.23606
57. Swanborn P. *What Is a Case Study?* SAGE Publications, Inc.; 2010.
58. Majid MAA, Othman M, Mohamad SF, Lim SAH, Yusof A. Piloting for Interviews in Qualitative Research: Operationalization and Lessons Learnt. *Int J Acad Res Bus Soc Sci*. 2017;7(4):1073-1080. doi:10.6007/ijarbss/v7-i4/2916
59. Guest G, Bunce A, Johnson L. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field methods*. 2006;18(1):59-82. doi:10.1177/1525822X05279903
60. Creswell J. *Qualitative Inquiry & Research Design*.; 2013.
61. Hwang W, Salvendy G. Number of People Required for Usability Evaluation: The 10±2 Rule. *Commun ACM*. 2010;53(5):130-133. doi:10.1145/1735223.1735255
62. Malhotra N, Peterson M. *Basic Marketing Research: A Decision-Making Approach*. 2nd ed. Upper Saddle River, N.J. : Pearson/Prentice Hall; 2006.
63. Dawes J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *Int J Mark Res*. 2008;50(1).
64. Friedman H, Amoo T. Rating The Rating Scales. *J Mark Manag*. 1999;Winter:114-123.
65. Collins C. History of ODBC. WordPress. Published 2007. Accessed November 6, 2015. https://ccollins.wordpress.com/2007/06/03/history-of-odbc/