A systematic literature review on security and privacy challenges in big data

Achini Dinusha Dissanayaka - @00608177 , Roby Daniel Cheriyan - @ 00606774

School of science engineering and environment, University of Salford

ABSTRACT

The globalization of every nook and corner of the world is pushing us towards a digitalized world, which is now having an exponential growth of data. Nowadays, many firms adopt data mining to extract crucial data that can be utilized for making marketing and sales decisions, monitoring particular behaviours or detecting threats. The computation of these kind of data is inconceivable using traditional tools and techniques, but feasible with the use of a number of techniques called big data analytics. These techniques have tremendous benefits when interacting with massive volume of data which is constantly evolving. The analytics solutions are responsible for storing, monitoring and efficiently evaluating a wide range of data obtained from all feasible and accessible sources. Though big data contains a huge potential for many companies or decision-making individuals, it also poses a significant risk for many people. By adopting powerful protection strategies would allow us to take advantages of big data without jeopardizing security or privacy. In this paper, we emphasize the security and privacy challenges of big data environments. Furthermore, we present some available protection strategies that offer safety against security and privacy of big data context.

Keywords: Big data; security challenges; privacy challenges; confidentiality; integrity; privacy; availability.

1. Introduction

With the innovation of smartphones, computers, Internet of Things (IoT) based devices, and social networks, day-to-day life has led us to a digitalized world. As a result, a massive growth of digital materials can be seen in modern society and these huge amounts of digital materials refers to big data.

Due to the rapid growth of data, storing and retrieving is now becoming a challenge. There are three major qualities of big data: Volume, Variety and Velocity [1]. Cookies, camera records, wish lists, and watched videos are some of the vital data for top level organizations to accomplish their goals by adapting the portfolio to tailor customer expectations. However, with the aid of conventional approaches, analysing huge amounts of diverse data is not that possible. Thus, some new techniques and tools that are perfectly aligned with big data were innovated. These techniques and tools are known as big data analytics (See Table 1).

As a result, the widespread usage of big data in numerous day-to-day operations and outsourcing of sensitive data triggers security and privacy concerns that must be addressed. Therefore, ensuring big data security and privacy during data transferring,

processing, and storing has become a significant requirement today.

Table 1 – Examples for big data analytics tools

Analytics Tools	Description
MapReduce	A model for processing big data with a distributed algorithm.
Hadoop	A framework designed for data computations.
Splunk	Use for monitoring and searching through big data.
MongoDB	A feasible, cross-platform database program
Spark	A distributed processing system used for big data
Cassandra	Used to process large volumes of data.

1.1. Prior Research

According to the previous studies we gathered, confidentiality (authentication, authorization, access control), integrity (data provenance, data leakage, data trustworthiness, data deduplication), monitoring and auditing, data privacy, availability, key management, and computations are the core security and privacy challenges of big data.

In [S1], authors have presented security theories and practices for the confidentiality and integrity issues of big data. Whilst Puthal et al, Nepal et al, Ranjan et al and Chen et al have proposed in [S2], a real time security verification for authentication and integrity challenges. Yang et al, Han et al, Li et al, Zheng et al, Su et al and Shen et al in [S3] have introduced some efficient novel techniques to address access control and data privacy challenges of big data. Azmi et al in [S4] have proposed a set of possible challenges that could be faced related to data integrity and data provenance. In [S5], Bertino et al has presented data trustworthiness issues of big data. The authors in [S6] have introduced a framework for investigating data leakage issues. The authors, Jeong et al and Shin et al have suggested an efficient approach based on data ownership to deal with data duplication and key management in [S7]. In [S8], the authors have proposed a top down levelled multi replica Merkle hash tree based secure public auditing scheme to alleviate monitoring and auditing issue in big data. Rahman et al, Ahamed et al, Yang et al and Wang et al in [S9] have presented data privacy challenges in healthcare sector and designed a generic framework to preserve the privacy of healthcare data in the cloud. The author in [S10], have demonstrated the issues related to the data confidentiality (authentication, authorization, access control), integrity, availability and privacy of big data and dispensed a systematic approach to overcome. In [S11], Moura et al and Serrão et al revealed the challenges related to data computation.

1.2. Research Goals

The scope of this paper is to analyze existing studies and their findings and to summarize the efforts of research in security and privacy challenges of big data. To focus the effort, we generated three research questions, which are displayed in Table 2.

1.3. Contribution and layout

Table 2 – Research questions.

Research questions		Discussion
1.	**What are the possible security and privacy challenges of big data?**	A review on security and privacy issues of big data will assist to understand the impact towards the technology.

2.	**What are the available solutions to manage security and privacy challenges of big data?**	An overview of available solutions for the above identified challenges will provide an insight of the procedures used to implement to ensure the protection of big data.

The remainder of this paper is arranged in the following manner. In section II, we have highlighted the security and privacy challenges that big data poses. We have demonstrated some potential solutions for ensuring security and privacy in the context of big data in section III. Finally, section IV will conclude the research by outlining some potential future possibilities for a safe big data environment.

2. Research Methodology

We performed the Systematic Literature Review (SLR) in line with the guidance introduced by Kitchenham and Charters [3] in order to meet the goals of addressing the research questions.

2.1. Selection of primary studies

Identified primary studies by applying the keywords in the search function of each publisher's website or a search engine. The keywords were chosen accordingly to gather answers to the research questions. AND and OR were the Boolean operators that allowed. ("security and privacy" OR "security challenges" OR "privacy challenges") AND ("big data" OR "big data analytics") were the keywords that were used to search.

The publisher platforms utilised for searching were as follows.

- IEEE Xplore Digital Library
- Google Scholar
- International Journal Of Engineering Research & Technology (IJERT)
- ResearchGate
- SpringerLink
- SpringerOpen

Relying on the search platforms, the searches were performed against the titles, keywords, or abstracts. On 25th March 2022, we started searching and gathered all papers that were published prior to that moment. The outcomes of these searches were sorted using the inclusion/exclusion criteria described in section 2.2. Then, Wohlin's [2] forward and backward snowballing approach was performed to previously generated outcomes, until no more papers fulfilled the inclusion criteria.

2.2. *Inclusion and exclusion criteria*

Table 3 - Inclusion and exclusion criteria for the primary studies.

Criteria for inclusion	**Criteria for exclusion**
Must comprise details about security and privacy challenges of big data.	Focus on other aspects such as, financial challenges of big data.
Must consist of solutions on the enhancement of existing big data security mechanisms.	Not written in English
Should be peer-reviewed paper published in a journal or conference.	Government documents, blogs, website articles.

Studies included in this research paper should be on big data, security and privacy concerns of big data environments with their findings, and suggestions on the enhancement of existing big data security mechanisms. They should be peer-reviewed and written in the English Language. Since Google Scholar may return lower-quality materials, all Google Scholar outcomes will be reviewed to determine whether the criteria were met. This paper will only contain the most recent version of a study. Table 3 demonstrates the major inclusion and exclusion criteria.

2.3. Selection results

The initial keyword searches on the specified platforms resulted in a total of 613 studies. After eliminating duplicate studies, this was dropped to 537. After reviewing the research using the inclusion/exclusion criteria, the number of papers left for reading was 57. The 57 papers were read entirely, and re- applied the inclusion/exclusion criteria, and filtered 33 papers. Forward and backward snowballing revealed an additional 2 and 4 papers, bringing the total number of publications to be included in this paper to 39.

2.4. *Quality assessment*

Using the guidelines established by Kitchenham and Charters [3], the quality of primary studies was evaluated based on their relevance to the research questions and valid experimental data. To determine the effectiveness of five randomly selected papers, we used the quality assessment process introduced by Hosseini et al [4]. We have created a checklist for quality assessment as illustrated in Table 4. Later, we applied this process to all other studies as well.

Table 4 - The checklist for quality assessment

Stage	Quality	Description
1	Big data.	The paper must be on aimed on big data technology.
2	Context	There should be enough context for the objectives and findings.
3	Security and privacy challenges / issues of big data.	There should be details in the paper about security and privacy challenges of big data, to assist in answering RQ1.
4	Recommended solutions for the challenges.	The paper must provide solutions for the challenges to assist in answering RQ2.
5	Data acquisition	There must be details about the way data acquired

2.5. Data extraction

The articles selected from the quality assessment were tested for the accuracy of their content. The five initial documents that underwent quality assessment were selected first to perform the data extraction procedure. With the aid of this process, we categorized the data into three parts as follows and stored them in a spreadsheet.

- **Context data** - Purpose of the research paper
- **Qualitative data** - Findings and conclusions
- **Quantitative data** - Experimental observations

2.6. *Data analysis*

We compiled the data collected underneath the categories, qualitative and quantitative for addressing the study questions. Also, we performed a meta-analysis on the studies that were submitted to the ultimate data extraction procedure.

2.6.1. Significant keyword counts

An examination of keywords was conducted throughout all 39 primary studies to identify their common topics.

3. Findings

All the primary research papers were read entirety, and pertinent qualitative and quantitative data were retrieved. All these primary research papers had focussed on addressing some recent security and privacy challenges that faced by big data as illustrated below in Table 5.

All these addressed research areas were classified into further categories for a better understanding of security and privacy issue of big data. Authentication, authorization, and access control were grouped into confidentiality, whilst data trustworthiness, data leakage (data loss), data provenance (data lineage) and data deduplication were grouped into integrity.

The pie chart in Figure 2 shows the percentages of various topics in the 39 main studies that were included in the data analysis.

The topics highlighted in the primary studies indicate that data confidentiality is a concern for approximately half (51%) of all studies on big data security and privacy issues. With a percentage of 20%, data privacy is the second most popular topic.

Data integrity related challenges are the third most common topic with a 19%. The fourth common topic was computation with a proportion of 4%. All the other topics are having 2% of popularity.

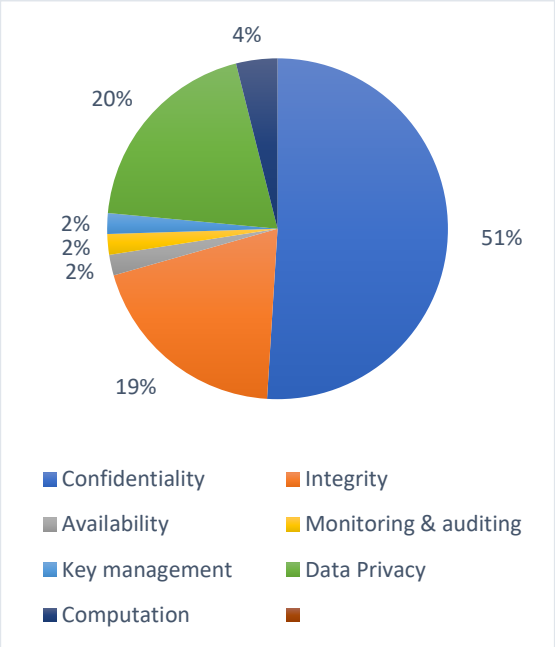


Figure 2 – Chart that denotes the topics of primary studies.

4. Discussion

Table 5 - Research areas of the primary studies

Primary study	Addressed security and privacy challenges
[S1] – [S26]	Confidentiality
[S1] - [S2], [S4] – [S7], [S10], [S27] – [S30]	Integrity
[S10]	Availability
[S8]	Monitoring and auditing
[S7]	Key management
[S3], [S10], [S31] – [S38]	Data privacy
[S11], [S39]	Computation

4.1. RQ1: What are the possible security and privacy challenges of big data?

Identified security and privacy challenges of big data were tabulated with their details as illustrated in Table 6.

Table 6 - Security and privacy challenges of big data environments

Challenges	Description
Confidentiality	Confidentiality means applying rules and limitations that prevent the data from being shared and accessed illegally. Different cryptographic approaches and controlling data access are frequently used to ensure confidentiality. Authentication, Authorization, and Access control (AAA) are crucial to ensure the confidentiality of big data [S10].
Integrity	The process of protecting against illegally modifying of data by an unauthorized user. Data integrity related challenges are arising due to user or hardware errors and intruders [S10]. Some of the most popular attacks against data integrity are data diddling attacks, man in the middle attacks, Salami attacks, session hijacking attacks, trust relationship attacks etc [S12].
	Data trustworthiness, data leakage, data provenance and data deduplication are vital to maintain the integrity of big data.
Data trustworthiness	Decision making is a key aspect of big data applications. To ensure the accuracy of predictions, decisions or actions, data trustworthiness is more significant. Due to lack of error free data, it is complicated to achieve data trustworthiness. Using data correlation and source correlation techniques, data trustworthiness can be ensured.
Data leakage / Data loss	Data leakage refers to the process of transferring or storing of personal, sensitive data illegally. Once the data is being copied, they can be altered and disseminated causing confusion and integrity difficulties.
Data deduplication	Data deduplication is a data compression approach that removes duplicate data from the storage. Deduplication and encryption are incompatible to a significant degree [20]. As encrypted data is distributed randomly all the time, same plaintext, encrypted with arbitrarily generated cryptographic keys will generate separate

	ciphertexts that should not be deduplicated [21].
Availability	Data availability in a big data means that how often the data is availability to be used and in what method the data could be used and is the state of guaranteed access to data without fail. Some of the most well-known attacks to breach data availability are Denial of Service (DoS) attack, Distributed Denial of Service (DDoS) attack and SYN attack that break the high uptime of server [18].
Monitoring and auditing	Traditional method of data blocking on servers and authenticating data on client for data auditing is not practical for the big data technology, due to massive amounts of data.
Key management	Managing (sharing, creating, storing, utilizing, and replacing) of cryptographic keys within a cryptosystem refers to Key Management [19]. Sharing the keys between the users, servers or data centres is not secured [18].
Data privacy	Data aggregation and analytics would lead to a user privacy breach as more data is collected. If the data analytics are outsourced, an unreliable third – party person will be able to infer private information about customers.
Computation	For the processing of big data, distributed programming frameworks such as MapReduce are required. But these distributed frameworks are not secure enough as the mapper which is used to process data has no additional security layer.

4.2. RQ2: What are the available solutions to manage security and privacy challenges of big data?

1. Confidentiality

Confidentiality is the most significant requirement of big data. As the user belongs all the sources, the user is totally reliable in the context of big data. The major attacks would originate from the outside or inside attacker (See Figure 3).

In the research [S1], it covers many encryption strategies for data confidentiality protection for special applications, general purposes, searching, sharing, numerical comparison, inquiry, and confidentiality protection using hardware.

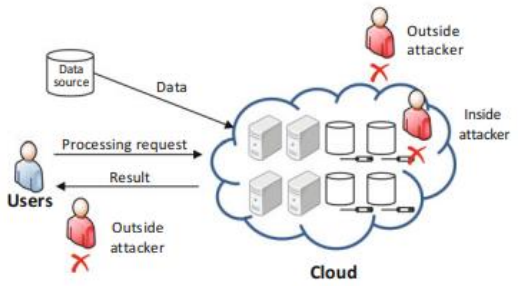


Figure 3 - Big data processing overview with cloud [S1]

1.1. Confidentiality for general purpose

There are two primary tools that are used to enhance confidentiality protect for general purposes of big data.

1.1.1. Fully homomorphic encryption (FHE)

The adoption of an encryption method to encrypt data prior to sending it to the cloud server, is a basic technique to ensure confidentiality of data. However, traditional encryption systems such as the symmetric encryption scheme AES, the asymmetric encryption scheme RSA are ineffective because they hinder the cloud from managing the data. To address this issue, the homomorphic encryption technology was created, which provides both confidentiality protection as well as data processing simultaneously [S1]. FHE comprises four algorithms as illustrated in Table 7.

Table 7 – Four algorithms of FHE

Algorithm	**Input**	**Output**
KeyGen	Security and privacy parameters	A public or private key pair (pk, sk) and an assessment public key evk (pk and evk are public whilst sk is secret)
Encryption	A message m_i and public key pk	A cipher-text a
Evaluation	a set of cipher-texts C = {c_1, c_2, … , c_n}, an evaluation function Evl(), and an evaluation public key evl	a cipher-text c, a result of encrypted outcome of Evl(m_1, m_2, … , m_n)
Decryption	A cipher-text c and the private key sk	The plain text m

1.1.2. Secure multi-party computation

Secure Multi-Party Computation (SMPC) notion is strongly associated to the concept of zero knowledge [S13, S14].

1.2. Confidentiality for special big data applications

It is possible to design more effective confidentiality protection schemes for specific big data applications compared to general purposes.

1.2.1. Searching

This is one of the most well-known operations where two types of encryption approaches are used to assist searching on cipher test.

1.2.1.1. Public Key Searchable Encryption

A public key searchable encryption approach which comprises of four algorithms was designed by Boneh et al. [S15] as in Table 8.

Table 8 - Four algorithms of public key searchable encryption

Algorithm	**Input**	**Output**
KeyGen	security parameter	public / private key pair (pk,sk)

PEKS	public key pk and a word W	searchable encryption W
Trapdoor	private key sk, a word W	trapdoor T_w
TEST	the private key sk, a searchable encryption S = PEKS (pk, W') and a trapdoor T_w = Trapdoor (sk, W)	If W = W' it outputs 'yes', otherwise 'no'

1.2.1.2. Symmetric Key Searchable Encryption

Most of the times, symmetric encryption is swifter than public key encryption.

Goh et al developed Z-IDX scheme which is exploring in the security model IND2-CKA and it consists of four algorithms (See Table 9).

Table 9 - Four algorithms of Z-IDX scheme.

Algorithm	**Input**	**Output**
KeyGen	security parameter s	master private key K_{priv}
Trapdoor	master key K_{priv} and word w	trapdoor T_w for w
BuildIndex	Document D and the master key K_{priv}	The index I_D
SearchIndex	Trapdoor T_w for word w and index I_D for document D	1 if w ϵ D and 0 otherwise.

1.2.2. Numerical comparison

Agrawel et al. introduced the initial Order Preserving Encryption (OPE) programme that permits the cloud to compare the cipher-texts [S16]. The OPES is a symmetric encryption as it utilizes the same information to encrypt new values or decrypt encoded values.

1.2.3. Special applications with SMPC

Normally, a two party SMPC protocol can be created for any function.

1.2.3.1. Private Set Intersection Protocol (PSI)

This requires the two-parties; client and server to mutually calculate the crossing of private input that results the clients learns the intersection while the server learns nothing at the end.

1.2.3.2. Computing ID3

This is a primary algorithm for creating decision tree, which is a technique for resolving the categorization issue in machine learning and data mining.

1.2.4. Sharing

Sharing is a primary product in cloud-based data management systems. At this situation, the owner uses the cloud for sharing data with other users, which is required to ensure that only the selected users be able to have the data while other parties cannot.

1.2.4.1. Classical public key approach

With this classical public key approach, the confidentiality of shared data can be maintained because the cloud can only access the encrypted data of sharing process.

1) **System Setup –** Each user has a public / private key pair and a reliable Certificate Authority (CA) that creates certificates for all public keys.
2) **Data uploading –** Encrypting data using symmetric encryption (with key dek) and upload them to cloud.
3) **Data sharing –** For sharing data with a group, the owner should retrieve certificates of the group members and then the data owner will encrypt dek with all of the certificates and upload the cipher texts to the cloud.
4) **Data retrieving –** To get the encrypted data and subsequent cipher-text of dek, retrievers have to communicate with the cloud. The private key that used to decrypt data should be used by the user to decrypt and get dek.

1.2.4.2. Attribute Based Encryption (ABE) approach

This is an expansion of identity based and fuzzy identity-based encryptions. There are two main types of ABE; Cipher-text Policy Attribute Based Encryption (CP-ABE), and Key Policy Attribute Based Encryption (KP-ABE). KP-ABE consists of four algorithms (See Table 10).

Table 10 - Four algorithms of KP-ABE

Algorithm	**Input**	**Output**
Setup	security parameter	public parameters (PK) and a master key (MK).
Encryption	input message m, a set of attributes γ, public parameters (PK)	the cipher text c_γ
KeyGen	an access structure A, master key (MK) and public parameters (PK)	the decryption key d_A
Decryption	The cipher-text c_γ and decryption key d_A, decryption	the message m if γ satisfies A

With the aid of a tree in which the leaves are characteristics and the internal nodes are logical operations, the access structure A can be defined (See Figure 4).

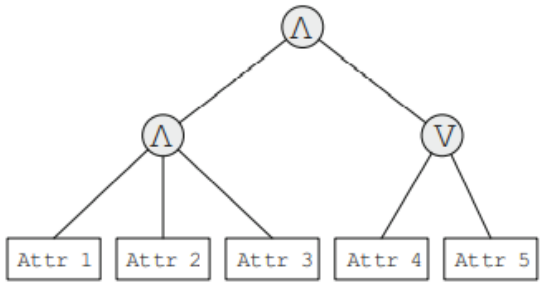


Figure 4 – An example of access structure [S1].

1.2.4.3. Proxy Re-encryption Based approach

This method offers a path to transform a cipher-text encrypted utilizing k_1 to a cipher-text of k_2 without leaking a plain text. This consists of five algorithms as in Table 11.

Table 11 - Five algorithms of proxy re-encryption based scheme

Algorithm	**Input**	**Output**

KeyGen	security and public parameters	public / private key pair (pk, sk) for each user.
ReKeyGen	two key pair (pk_a, sk_a) and (pk_b, sk_b)	re-encryption key $rk_{a \to b}$
Encryption	public key pk_a and a message m	cipher-text c_a
ReEncryption	cipher-text c_a and re-encryption key $rk_{a \to b}$	cipher-text c_b, that can be decrypted using sk_b
Decryption	cipher-text c_a (c_b) and private key sk_a (sk_b)	the message m

1.3. Confidentiality for query

Database as a service (DaaS) is one of the most compelling products of cloud-based big data, which is also able to assist in SQL queries [S1] . A user who controls the data source, transmits the data to the cloud. All the users can send query requests to the cloud and receive subsequent outcomes, but it must ensure that the cloud is unable to discover the contents of the database throughout this procedure.

1.3.1. Bucket technique

The author Hacigümü presented a solution to the problem of confidentiality preserving queries [S17]. An overview of their methodology is illustrated in Figure 5, which includes the introduction of a secure proxy to assist the user in querying the encrypted database controlled by cloud.

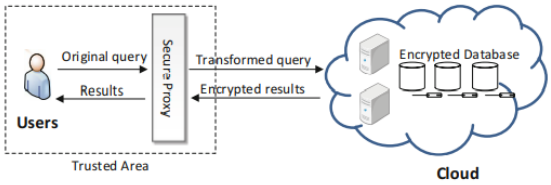


Figure 5 – Overview of the query [S1]

The methodology of this proposed approach is as shown in Table 12.

Table 12 – Bucket technique methodology

Steps	Description
1	• The range is divided into buckets (sub-ranges) for a given attribute. • A random token is generated for each bucket (the original tuple is encrypted and kept together with these tokens).
2	• Once a user queries a database, the query request is converted through a secure proxy. All query values are substituted by subsequent bucket tokens.
3	• Based on the tokens, the cloud executes the converted queries and generate the encrypted outcome to secure proxy.
4	• The generated outcome is then decrypted by secure proxy and sorted out unwanted information that contains due to the use of bucket tokens as a substitute of specific values. • The filtered outcomes are then sent to the end user.

According to the Figure 6, the original tuples are in the left table, while the protected tuples are in the right. The right table's first column has the

encryption of the initial tuples, while the other columns include the bucket IDs of the elements.

Original Data				
eid	ename	salary	addr	did
23	Tom	70K	Maple	40
860	Mary	60K	Main	80
320	John	50K	River	50
875	Jerry	55K	Hopewell	110

Transformed Data				
etuple	eid^5	$ename^1$	$salary^3$	$addr^4$	did^5
1100110011100010...	2	19	81	18	2
1000000000111101...	4	31	59	41	4
1111010000100001...	7	7	7	22	2
1010101010111110...	4	71	49	22	4

Figure 6 – An example of bucket technique [S1]

1.3.2. Data decomposition

To achieve confidentiality preserving queries, the author Aggarwal recommended dividing a database into two sections and managing each of them using two clouds [S18]. The system overview of this approach is depicted in Figure 7.

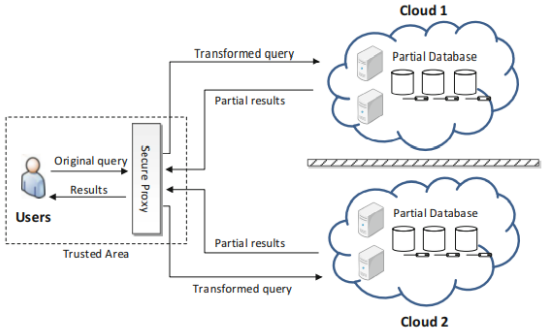


Figure 7 – Outline of the system which uses data allocation for confidentiality protecting query [S1].

1.4. Confidentiality with hardware

For the protection of big data confidentiality, there are high security models with apparent security characteristics. But these solutions are exorbitantly expensive, thus limiting their implementations. People introduced hardware-based methodologies to safeguard big data confidentiality.

1.4.1. TPM based secure HDFS

The Trusted Platform Module (TPM) [5] is a commonly applied hardware-based security protection mechanism that offers a safe environment for storing of secrets and critical processes. (as shown in Figure 8)

Cohen and Acharya created a reliable HDFS storage system that uses TPM to safeguard data confidentiality [S19]. Numerous big data managing frameworks adopt HDFS for data storage. The reliable HDFS storage system manages the encryption key and performs the encryption/decryption tasks using TPM as the basis of trust. To access the decryption engine, programmes must pass the TPM authentication, which makes it difficult for a cybercriminal to violate the confidentiality.

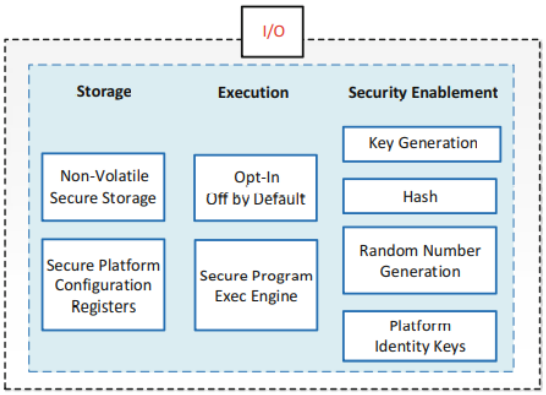


Figure 8 – Basic components of TPM [6]

As TPM is not built to process large amounts of data, the operation turns into a choke point. The researchers advocated to adopt Intel AES instructions for encryption and decryption.

1.4.2. FPGA based MapReduce

FPGAs are configured with bitstreams that entirely define the device's operation. The he bitstream is normally stored outside of the FPGA in a separate

configuration memory. [Figure 9] So, the use of an encrypted and decrypted bitstream every time once it is stacked into the FPGA is required to safeguard the important intellectual properties of FPGA. Then, on every power-up or reset, it is loaded into the FPGA.

Xu et al. recommended using the bitstream encryption technology and tamper resistance capability of FPGA to protect MapReduce secrecy. [S20, S21]

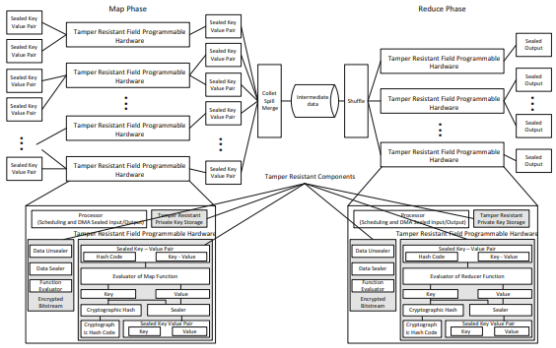


Figure 9 - Overview of FPGA base MapReduce [S1]

1.4.3. TrustZone based solution

TrustZone is a hardware-based approach which is deployed inside the cloud to enhance the confidentiality of big data. [7]

TrustZone is tightly intertwined into ARM processors, and the secure state is extended all over the system via specific TrustZone System IP blocks. (See Figure 10)

Nowadays, ARM processors are employed in data centres used for big data managing as well as mobile platforms. [S22 - S24].

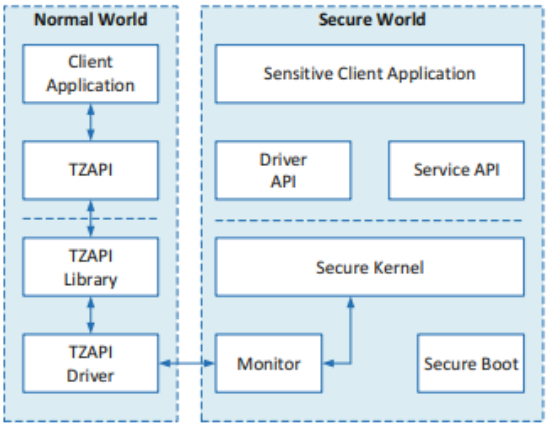


Figure 10 – Overview of TrustZone [S1]

1.4.4. Intel Software Guard Extension (SGX) based MapReduce

This is a collection of recent CPU guidelines that applications be able to utilise to set away confidential data [S25]. SGX facilitates a programme to create a secured container known as an enclave. An enclave is a secured sector in the application's address space that offers confidentiality and integrity though there is an existence of privileged malware (see Figure 11). Any software that is not resident in the enclave is barred from accessing the enclave memory space [8].

Intel SGX can be used in the clouds to provide general-purpose confidentiality protection. Schuster et al. introduce a strong SGX-based secure MapReduce solution [S26].

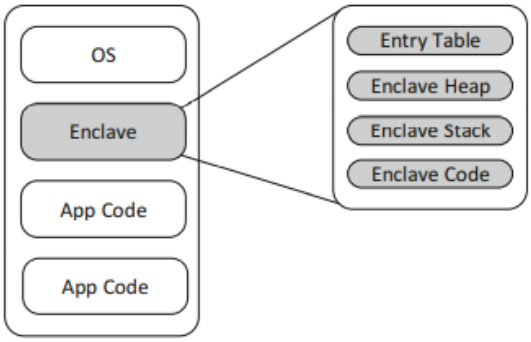


Figure 11 – An enclave inside application's virtual address space. [S1]

2. Integrity

Different methodologies such as data integrity protection with digital signatures, query and storage integrity protection and integrity protection with hardware have been proposed by Xu et al. [S1].

Data provenance (data lineage), data trustworthiness, data loss (data leakage) and data deduplication together can be used to ensure data integrity [18].

2.1. Integrity protection with digital signature

The most well-known methodology for the safeguard of data integrity is digital signature.

2.1.1. Classical digital signature and MAC

For classical digital signature, it needs a PKI system. There are three algorithms in a digital signature scheme (See Table).

Table 13 – Three algorithms of digital signature scheme

Algorithm	**Input**	**Outcome**
Initialization	A public / private key pair (pk, sk).	Here sk is kept as a secret and pk is published as a certificate
Signature generation	A message m	The signer uses sk to calculate a signature sig for m
Signature verification	a message m, a signature sk, a certificate	the verifier can verify whether sig is created using m and subsequent private key sk
	comprises public key pk	

All the parties involved in computation is receiving a pair of public or private key. PKI provides a certificate for all public keys.

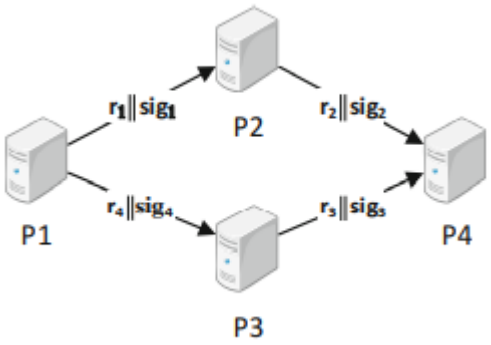


Figure 12 – Overview of classical digital signature [S1]

Once a party completes processing of data, it will use its private key to create a digital signature and append the signature to the outcome. To authenticate integrity of the outcome, the recipient of the outcome could utilize the subsequent certificate that contains the public key. To avoid a cybercriminal from deleting or altering intermediate outcomes, the sending party can encode a sequence number to all outcomes before producing the signature. This is difficult for an attacker to interfere with big data managing system, until the private keys are not shared to the public. Figure 12 demonstrates the integrity protection of big data using classical digital signature.

2.1.2. Homomorphic signature

This is a very powerful tool for addressing the integrity protection of big data.

The process of integrity protection scheme using homomorphic signature is as follows.

1) Key generation

The user who maintains the data source (the owner), creates a public or private key pair and the public key is shared with the cloud server.

2) Data initialization

The data is divided into units and generated signatures for all units by the owner. Both the signatures and data are then submitted to the cloud.

3) Processing

The cloud interacts with data the data and subsequent signature at each stage. The outcome with its newly produced signature is forwarded to the next phase. (Only the public key is necessary for this operation)

4) Verification

The user and the cloud can authenticate the signatures at any phase of data processing to ensure the integrity of the outcome.

2.2. Query integrity protection

One of the most prevalent big data products is database query. Once the cloud administers the database and responds to user queries the set of data, a system must be in place to ensure that complete outcomes are given rather than partial outcomes or outcomes with some items altered.

2.2.1. Dual encryption

To ensure the integrity of the delivered findings, Wang et al. devised a dual encryption approach [S27]. Before transmitting the database to the cloud, it is encrypted at the tuple level using a primary key k, and a small portion of database is encrypted with some other key k0 (here, the tuples operate as checkpoints). A secure proxy transforms user queries prior to sending them to cloud for managing. The secure proxy is also accountable for ensuring accuracy of retrieved outcomes.

A batch of queries $Q = <q_1, q_2, \dots, q_m>$

Each query is encrypted by the secure proxy using the primary key k and k'. Q^k, and $Q^{k'}$ both are sent to the cloud respectively.

$Q^k = <q_1^k, q_2^k, \dots, q_m^k>$

$Q^{k'} = <q_1^{k'}, q_2^{k'}, \dots, q_m^{k'}>$

After receiving query results for both Q^k, and $Q^{k'}$, the secure proxy will examine them to check if they are consistent.

After receiving query results for both and the secure proxy will examine them to check if they are consistent. An attacker is unable to detect if a tuple in database is encrypted with k or k', if the encryption algorithm used is semantically secure. As a result, the proposed methodology is able to diagnose integrity tampering with a high degree of certainty [S1].

2.2.2. Extra records

To ensure query integrity, Xie et al. advocated inserting a limited number of entries into the outsourced set of data [S28]. The core idea is to share a collection of injected tuples on the user and cloud sides, with cloud unable to differentiate from one another.

2.2.3. In dynamic environment

Li et al. address the circumstance in which the user may modify the data which is an additional component of integrity [S29].

Once a user modifies database, cloud still may maintain an older edition and respond to the query with outdated data.

To ensure the freshness of the query outcomes, the author introduced an embedded Merkle tree data structure and used the signature aggregation technique [9].

2.3. Storage integrity protection

The concept of Proofs of retrievability (POR) was proposed by Juels and Kaliski [S13].

The verifier in their approach retains only a single cryptographic key, regardless of the size and number of files. The POR protocol functions as shown below.

User encrypts file F and embeds a set of randomly generated inspection clocks known as sentinels. The utilization of encryption ensures that the sentinels cannot be distinguished from regular file blocks [S1].

The user tests cloud through defining the locations of a set of sentinels and asks the cloud to provide the sentinel values associated with those positions. If the cloud has changed or deleted a significant chunk of F, it is very likely that it has also suppressed a lot of sentinels [S1].

2.4. Integrity protection with hardware

2.4.1. TPM based integrity protection through software protection

TPM's primary job is to provide authentication service, which is a means for software to verify the identity [63]. the purpose of authentication is to demonstrate to a third party that a software is complete and trustworthy. Ruan and Martin used TPM to build a reliable MapReduce infrastructure TMR [S30]. This solution meets the purpose of securing data integrity by safeguarding the MapReduce application stack's integrity. (See Figure 13)

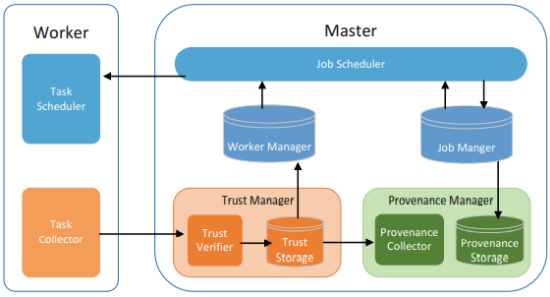


Figure 13 – TMR overview [S1].

3. Availability

3.1. Data loss prevention tool (DLP)

Data loss prevention tools assist in the following areas: Maintaining durable data records that adhere to all applicable compliance standards. It helps to reduce the danger of a data breach by ensuring data security and privacy.

ARCSERVE UDP

- End-to-end encryption, role-based access management, backup infrastructure isolation, as well as full reporting, alerting, and activity logging capability are all included in the security features [10].
- Arcserve UDP encrypts your sensitive data before it leaves your production system and while it is being transferred to local, tape, distant, or cloud disaster recovery storage. Advanced AES-256, AES-192, and AES-128 encryption may be used to encrypt your backups and data at rest on:
- Shared folders and local drives

- On the Recovery Point Server, deduplicated datastores (RPS)

AAR (Authentication, authorization and role-based) ACCESS CONTROL SYSTEM

Arcserve UDP has expanded default and customised setup and features to guarantee that only authorised users have access to data backups and your data protection infrastructure, preventing unauthorised access and data breaches. To make user administration easier, Arcserve UDP may employ built-in local users or link with an organization's Active Directory [10].

4. Monitoring and auditing

Big data refers to a massive amount of data and datasets that can be found in a variety of formats and from a variety of sources. Many businesses have realised the benefits of gathering as much information as possible. But collecting and storing huge data isn't enough; you also need to use it.

4.1. Data analytics and its security aspects

This is the process of identifying trends, patterns, and correlations in huge amounts of raw data to aid in the making of data-driven decisions. There is software which made ease in handling large amounts of unstructured data using various analytics tools.

4.2. Tableau (analytics tool)

This is a powerful tool which helps in data analytics which helps an organization prepare, analyse, collaborate, and share big data insights. Tableau is a leader in self-service visual analysis, allowing users to ask new questions of managed big data and quickly share their findings across the company [11].

Data analytics providers are mobilising to help citizens, scientists, governments, and corporations comprehend the nature and scale of the COVID-19 epidemic, which has impacted practically every facet of life around the world. Tableau is a useful tool in this situation. Tableau is a data analysis and visualisation tool that can easily connect to a variety of data sources. Tableau's ability to create interactive dashboards is a significant benefit. Through a visual straightforward drag and drop interface, these dashboards may be produced without much coding experience. This makes continuous use of application integration advancements such as JavaScript APIs and single sign-on apps to integrate Tableau analytics into fundamental business applications. Tableau can produce a wide variety of visualisations to display data and reveal insights in an interactive way.

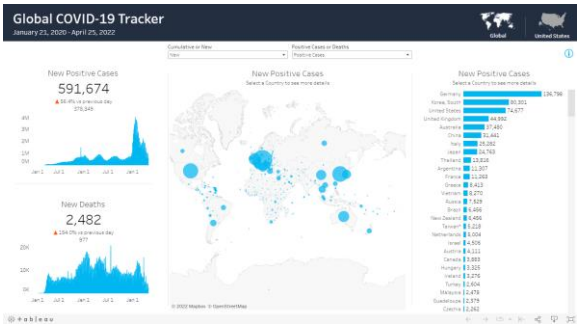


Figure 14 - The Covid-19 tracker dashboard [12]

Tableau can easily handle data with millions of rows. The enormous amount of data can be used to generate different sorts of visualisations without affecting the dashboards' speed [S31].

4.3. Benefits and security enhancements in the tool

- Authentication and authorizations are handled by the application server.
- It oversees online and mobile interface administration and permissions.

- By recording each session id on Tableau server, it ensures security.
- The server's default session timeout can be configured by the administrator.

4.4. Data security when hosted in cloud

Tableau Server could be hosted in cloud to use AWS as the KMS for extract encryption if you're using Data Extract Encryption at Rest. The server must be deployed in AWS EC2 to enable AWS KMS. Tableau Server generates an AWS data key using the AWS KMS customer master key (CMK) in the AWS scenario. For all encrypted extracts, Tableau Server uses the AWS data key as the root master key. The native Java keystore and local KMS are still used for secure storing of secrets on Tableau Server, even when set for AWS KMS. For encrypted extracts, the AWS KMS is solely utilised to encrypt the root master key [13].

5. Key management

A secure key management scheme is introduced to make sure there is adequate security and privacy of big data networking. We upload the data encrypted to protect the security of the big data network in the cloud. The top key encrypts the middle key, and the middle key encrypts the lower key in this suggested system of key hierarchical management. The data is encrypted in the client, and the ciphertext data is sent to the server in the cloud.

Three layered structures in key management:

- The password key is on the first layer, followed by the main key and user public-private key pairs on the second tier, and finally the file encryption key on the third layer.
- The password key encrypts the main key and user private key; the primary key encrypts file encryption keys; the higher key encrypts the lower key, and the middle key encrypts the lower key, ensuring that all keys are secure. The login password is the only thing that users must remember.
- The main key, private key, and file encryption key are all saved as ciphertext on the cloud. With ciphertext, a user can share his file with others [14].

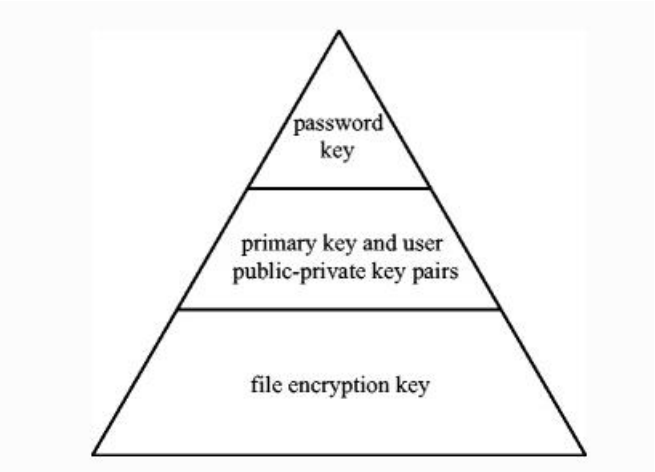


Figure 15 - The key management framework [14]

- The password key makes up the first tier.
- The primary key as well as user public-private key pairs make up the second layer.
- File encryption key is the third layer.

The primary key and the user private key are both encrypted by the password key. Encrypting file encryption keys is done with the primary key. The upper key encrypts the lower keys, ensuring that all the keys remain secure. The login password is the only thing that users must remember. The cloud stores the primary key, private key, and file encryption key as cypher text.

It is important to understand that digital signatures are an important aspect on how a communication could be secured in the big data environment. There are a few security schemes such as RSA and DSA which rely on certain security algorithms and the security complexity depends on how large the data is.

5.1. The Merkle signature scheme

This method provides an alternative signature solution where the security is based solely on a secure hash function and a secure one-time signature. To make the signature scheme work, efficient ways for solving this challenge are needed. The key management is the most serious issue with One-Time Signature Schemes. The process of exchanging a public key is quite difficult. We must ensure that verifying the public key belongs to the intended communication partner and has not been tampered with. As a result, just a few public keys should be utilised, and they should be somewhat short. In contrast to previous signature schemes, One-Time Signature Schemes employ a fresh public key for each signature, thus the public key is relatively large. To make One-Time Signature Schemes practical, effective key management is required, which decreases the number of public keys and their size [S32].

5.2. Hash function

A hash function transforms a collection of inputs into a table or data structure. The hash function produces a unique result for each input. This aids data fingerprinting. The hash functions differ depending on the application (See Figure 16).

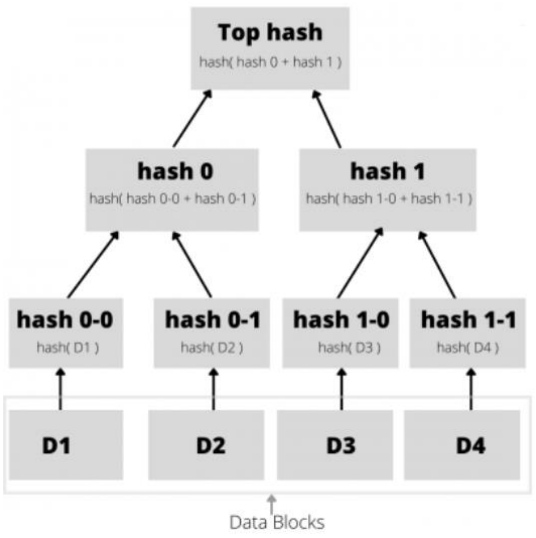


Figure 16 - The Merkle hash tree [14]

How does this work and how efficient are they?

The data is divided into blocks D1, D2, D3, and D4, as seen in the figure above. Using hash functions, these blocks are hashed. Then, until we get to the root node, each pair of nodes is hashed recursively. D1's hash is 0-0, while D2's hash is 0-1. (hash(hash 0-0 + hash 0-1)) Hash 0 holds the sum of its children's hashes.

5.3. Benefits of Merkle tree:

- It ensures data integrity as well as authenticity.
- The amount of memory required to hold the data is drastically decreased.
- It simply requires a tiny quantity of data to be sent across networks.
- It allows to validate transactions in a block without having to download the complete block.

The process on how data verification is performed in Merkle tree:

- We use the untrusted network to download a large amount of data.

- We request verification that this piece is in the tree from the server.
- The proper hashes are returned from the server.
- Then compute the root hash using this information and compare it to the root hash you used to open the file [15].

6. Data privacy

The term privacy specifically means on how the data is used and who all could access them. It also means the methods in which the data could be secured. Data privacy is closely correlated with data handling [S33].

We are about to discuss how privacy plays an important role considering the complexity of big data and its management and the advancements done in the world of big data privacy. The advantage of having some control over how personal data is gathered and utilised is known as information privacy. Information privacy refers to a person's or a group's ability to keep personal information from being shared with persons who are not related to them. Big data privacy is defined to be the policies, procedures and tools which necessitates to protect the sensitive data such as the Personally Identifiable Information (PII) and the INTELECTUAL PROPERTY (IP). This is defined to be the privacy compliance which helps in reducing the risk of a potential privacy risk.

Below are a few mechanisms which discuss on how data privacy is being protected by emerging techniques in the big data technologies.

A. Differential privacy

In this mechanism the personal identity of the individual is not disclosed even though they are stored. This allows the researchers and database analysts to extract meaningful information from databases containing personal data without disclosing the individual's identity. A minimum noise is introduced to the information which would engage an anonymity.

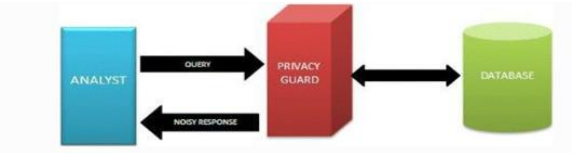


Figure 17 - Differential privacy mechanism [S34]

How does this work?

- A request is sent to the privacy guard by the analyst as show above.
- Now the impact of the query is being evaluated by the guard using an algorithm.
- The request is now put across the database to receive in return with a clean and non-distorted answer.
- The privacy guard now introduces additional noise in a proper scale which adds a proper privacy impact.
- This gives and additional privacy to the original answer which protects the answer from being judged making sure a proper confidentiality of the information in database.
- Now the modified response is sent back to the analyst to process [S34].

6.1. Impact of the noise introduced

The impact is very less in this methodology, since the noise would be less enough to help in preserving the information and very large enough to protect the privacy and security of the information in parallel.

B. Identity-based anonymization

A greater approach on the data anonymization is that this is not considered as personal data anymore which has a greater impact on the GDPR (General Data protection Regulation). Data perturbation and data swapping techniques to prevent the individuals from associating from the critical information is an effective method to protect the privacy.

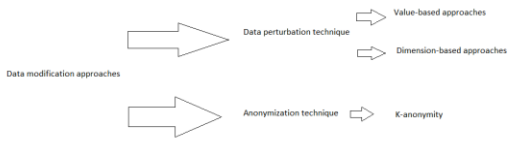


Figure 18 - Self data modification approaches [16]

K- anonymization is one of the efficient methods in protecting the privacy and is one of the efficient ways to prevent the re-identification of data by hiding the identity. As an example, in the medical environment there has been vast research performed to implement diverse technologies and techniques to ensure privacy of the medical data. There are three main methods used which are [S35]:

- k-anonymity
- Random perturbation technique
- Secure Multiparty computation (SMC)

1) The k-anonymization technique

This is a type of algorithm based on quasi-identifier attributes. This is defined using a privacy factor or a parameter named k of k-anonymity which is already known beforehand to the application of the algorithm. This technique is helpful in masking the data for the attacker to be able to detect the identity of the individual.

If an attribute can assist re-identify the subject associated with a record, it is categorised as a quasi-identifier attribute. Age, gender, ZIP code, and other socio-demographic data are common quasi-identifier features. Any characteristic associated with a distinct subject that is known to be externally available should be categorised as a quasi-identifier attribute [S36]. It is well important that the medical data in the health care industry should remain safe and secure according to HIPAA (Health insurance portability and accountability act).

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

Figure 19 - K-anonymity attributes and parameters [16]

Identifying data (name, zip code, gender, etc.) may coexist with sensitive data for a specific person (health records, prescriptions, financial information, passwords, etc.). Identifying data and sensitive data might be used in the wrong hands to re-identify a person and jeopardise their privacy. The goal of k-anonymity is to prevent the two types of data from being linked together.

2) The random perturbation technique

Data perturbation is a form of privacy safe-guard method which is much preferred due to the simplicity and efficiency in operation. This method is set to operate when the computational complexity is very low which helps in preserving the privacy of the data. There is a confidentiality and integrity of the records which are maintained and preserved with

systemically modifying the original data set elements. This method could b classified into two categories which are Value based methods and dimension-based methods [S37].

- *Value-based approach*

In this approach, random noise addition is the common method in which the data perturbation is being performed. The is also known to be a method of value-distortion.

Below shown is how we calculate random noise addition:

X=Xi+r

Xi=The original data value of one-dimensional distribution

r=Random value which is taken from a particular distribution.

In this method the original value is being distorted by the addition of random values which is equivalent to a random noise. Then the processed value has been returned. Uniform or Gaussian distribution is used in this method.

- *Dimension-based approach*

This method is introduced to rectify the cons in the value-based method. It is evident that in the real world the data sets are usually multi-dimensional which increases the difficulty of the data-mining procedure. The majority of value-based perturbation algorithms are concerned mainly with maintaining the distribution information of a single data dimension. As a result, they have a built-in disadvantage when it comes to providing reliable mining results in a data-mining operation that requires data from several associated data dimensions.

Random Rotation Transformation and Random Projection are the most frequent dimension-based data collecting methods.

Random Rotation Transformation: This is one of the privacy preserving methods which is accomplished by multiplying a rotation matrix first to a data set obtained:

g(X)=RX,

R=Rotation Matrix

X=Original data set

The rotation is done in such a way that the original data set's multi-dimensional geometric features, such as Euclidean distance and inner product, are preserved. Because the rotation does not affect all data points equally, and the ones closest to the rotation centre may see minor changes, privacy protection over these points may be compromised [S37].

3) Secure multiparty computation (SMC)

This is a generic cryptographic primitive that enables distributed parties to jointly compute an arbitrary functionality without revealing their own private inputs and outputs. SMPC offers a distinct edge when it comes to resolving security and privacy concerns. Within a distributed computing situation, SMPC handles the challenge of cooperative computation done on private data from several users in a safe manner. Informally, two or more parties with private inputs seek to calculate some shared functionality utilising these inputs in the SMPC scenario. Maintaining security in this assignment necessitates each participant obtaining just their own objective output [S38]. Benefits are as follows:

- It is available commercially

- There are no trusted 3rd parties to observe the data
- Trade-off between data usability and data privacy
- GDPR and data privacy compliance
- Precision and accuracy
- Quantum safe

Table 14 - The data generation phase of data privacy

Active Data Generation	Passive Data Generation
In this method the data is given to a third party by the data owner [S34].	In this method the data is generated by the online actions of the data owner such as web browsing. However, in this method the data owner is unaware of the data gathering procedure done by the third party.

Minimization of the risk for surrounded by the above data generation scheme could be avoided by the following sub-techniques:

A. Access restriction

- This term refers to as refusing to share or uncover sensitive information which is not intended to be shared. Access restriction can be also managed by different levels of Access control mechanisms controlled by the data owner [S34].

B. Data falsification

- In this method the data is falsified using certain tools to counteract the access of sensitive data. Data could be effectively distorted using certain tools before the data could be leaked.

7. Computation

We already discussed on the challenges in the data computational procedure in the previous section. It is well evident in a distributed programming framework where large amounts of data are being dealt, the MapReduce distributed framework does not have efficient security protections.

Let's say, in MapReduce the bulk input data is being split into independent chunks which are then aggregated in map schedules to complete the task in a parallel method. This is then allocated a storage. Eventually if someone could tamper the mapper settings in such a way as to manipulate the vital data that is being processed since MR(MapReduce) does not have added security. It is also difficult to find out untrusted mappers [17].

It is indeed vital to handle and mitigate or even come up with a solution to this security challenge in data computation when dealing with a distributed programming environment while parallelly ensuring the integrity of the data within.

When enormous volumes of information are used, privacy is a big problem. Data mining and predictive analytics are examples of processes that may be used to identify or deduce information links. Transparency and permitting feedback from the data provider can be used to establish data protection control. User input allows a person to express their preferences for the use of their personal information. When dealing with a big number of mappers and reducers, like Map Reduce frequently does, this may be quite tough. Data privacy, integrity, and security may be challenging to control because to the amount

of the data and the complexity of the analytics conducted during a MapReduce. For data integrity and security, the Map and Reduce phases in Map Reduce should be addressed [S39].

MapReduce eases the handling and processing of massive data set computation by the following benefits [S39]:

a) Scalable
b) Fault-Tolerance
c) Simple
d) Independent

7.1. Obstacles

Big Data presents implementation and management problems for corporate enterprises. The Apache team is still working on Hadoop's fundamental standards, which do not yet fully handle enterprise requirements for more robust privacy and security, policy enforcement, and regulatory compliance [S34].

Secure MapReduce (SMR)
This is a system to reduce the fall-backs occurred in the legend MR(MapReduce) architectural system. Once big data is distributed and mapped to a data centre, it must be secure for data analytics to be accurate after it has been reduced. The Secure MR works on a light-weight encryption procedure which accomplishes the big data security and privacy challenges with efficient timing requirement.

7.2. Benefits of lightweight encryption

- Provides lower power designs due to reduced requirements of resources.
- Minimal encryption timing.
- Provides better security than the existing heavyweight algorithms such as AES, RSA, PGP and RC6.
- Delivers key encryption along with multi-level light weighted encryption which reduces the chances of security and threat infiltration and attacker threats [S34].

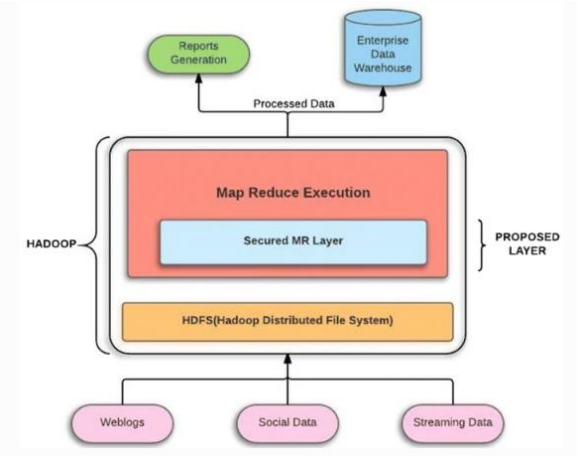


Figure 20 - Illustration SMR proposed scheme [S34]

7.3. How does it work and what improvement has it been of?

The suggested Secured Map Reduce Model provides a strong method for safeguarding distributed computing systems in the company by filling the privacy and security holes that present in all open-source Hadoop deployments. The greater benefit of this model is to have a very minimal information loss. The newly designed algorithm is made on a current “privacy-preserving data technique”.

As the data passes through the map-reduce phase, this new layer applies the security algorithms to each individual piece of data. The security method should use light-weight encryption approaches so that the overhead of new algorithms does not interfere with the Big Data's core functioning.

7.4. Functioning:

- Data is initially collected from weblogs, social data, steaming data and these collected data is then sent to HDFS (Hadoop distributed file system),

- The complete encryption procedure starts in the MapReduce layer.

- The encryption starts once the data enters the MapReduce Layer.

- The encryption consists of two stages [S34].
  - Converting text data to number: Each text is converted to tokens. A method called KVP (key value pair) is introduced to count the repetition of the unique word in the given data where Key= unique word and Value= The number of times the word is repeated

  - Converting number data with a randomization procedure (Improves the privacy level).

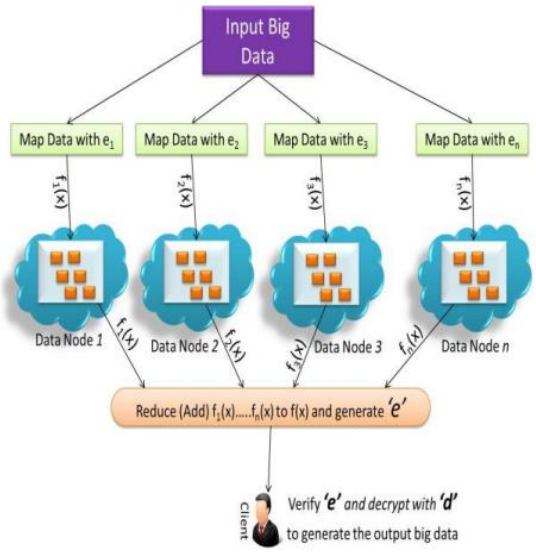


Figure 21 - SMR proposed scheme [17]

Data is encrypted at the Mapping step and decrypted at the Reduce stage in the proposed architecture before it is transmitted to clusters.

The proposed approach is summarised as follows:

1. A secure key pair is created by the data centre's trusted authority and transmitted to the client via secure channel. (Encryption key is represented as e, decryption key is represented as d)

2. Public key e is used to encrypt the data that will be mapped.

3. Polynomials f1(x), f2(x)..... fn(x) are constructed with e1, e2..... en as constant terms in f1(x), f2(x)..... fn(x) as we could see from the above figure. From the image above we could notice (ei is the co-efficient of x0 in fi(x), i=1,2.....n), where I the message digest for the polynomial, is created and added to the encrypted data and stored in the data centre.

4. To acquire f1(x), f2(x), ... fn(x), the encrypted data from each data centre is recovered and reduced (added) to yield f(x) and e.

5. The obtained e is compared to the client's e. If it matches, the client uses d to decode the encrypted data. If it doesn't match, the client informs the data center's trusted authority.

The above method shows how secure the encryption and decryption of the data is being performed [S39].

7.5. Benefits of Secure MapReduce Scheme:

- The data is stored in encrypted format in the data nodes which prevents the modification of

the data during the time of the storage or during the transit of the data by unauthorized users.

- This improves the integrity of big data and integrity of the data is assured when the message “e” is verified by the end user.

- A slight modification of the data at the client end is verified and successively notified to the trusted authority, and hence accurate data is requested so as to proceed with the decode of the original data at the final reduce step [S39].

It is important to protecting the computations of the big data environment as much as how important it is in protecting the storage. It is uncertain to calculate, predict or even to know how and were big data processing is performed, and it is important to accept and adapt some control techniques weather if it’s going to deny the node from accessing the processing results or to calculate the credibility.

5. Conclusion and future work

This research has given a greatest insight on the security and privacy challenges and available recent research on the solutions that address these challenges on big data technology. Big data has tremendous promise for achieving very acquisition of skills. They generate new areas of research and ideas, take use of cutting-edge data gathering and processing technology, and eventually become a standard methodological approach. We've highlighted the key issues and developing trends in big data, as well as the possibilities that have arisen.

Big data may be used in a variety of ways. Indeed, it has the potential to be a gold mine for both business and personal reasons. Cloud computing provided limitless resources for analysing, storing, and managing diverse and large amounts of data. As the demand for cloud services grows throughout the world, cloud servers are facing massive information and data spillage and data leak risks. Data security has been a greater area of discussion in a communication perspective as well as when data is being stored. The work we have showcased gives light on to the greatest data security threats and the current challenging issues which Big Data encounters in cloud servers. The paper also reviews and analyses the works performed previously in different research papers. We also strongly believe that a development should be done in the security and privacy framework in almost every aspect of the stages of big data technology depending on the emerging threats and attacks which we hear in this era.

We started in the discussion of the security and privacy challenges, in the perspective of a researcher. Our analysis was based on a methodological study starting from the discussion of the 3 V’s (Volume, Velocity, and Variety). These are said to be the main attributes which also has a close or a distant relation to the security and privacy challenges in big data. We also picked 7 major core study areas for our research which comprises of the (Confidentiality, Integrity, Availability, Monitoring and Auditing, Key management, Data Privacy and Computation) on which this paper is based on. The solutions were specifically collected for each section of the core area and carefully summarised in co-relation with the emerging technologies along with the fallbacks if any. In total, this paper exhumed lots of challenges pertaining to the area of study and taken in various methods and strategies in the big data world. We believe the recent undiscovered technological advancement along with the solutions

discussed in this paper would create a greater footprint in the era of big data security and privacy sector.

Declarations of interest

None.

REFERENCES

[1] *What is 3Vs (volume, variety and velocity) ? - Definition from WhatIs.com*. WhatIs.com. (2022). Retrieved 19 March 2022, from https://whatis.techtarget.com/definition/3Vs#:~:text=The%203Vs%20(volume%2C%20variety%20and,the%20speed%20of%20data%20processing.

[2] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proc. 18th Int. Conf. Eval. Assess. Softw. Eng. - EASE 14, 2014, p. 110 [4] E. Bertino, “Big Data – Security and Privacy”, 2015 IEEE International Congress on Big Data, 2015, pp. 757-761.

[3] B. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering, in: Engineering, vol. 2, 2007, p. 1051.

[4] S. Hosseini, B. Turhan, D. Gunarathna, A systematic literature review and metaanalysis on cross project defect prediction, IEEE Trans. Softw. Eng. 45 (2) (1 Feb 2019) 111–147.

[5] TCG (2011) TPM Main Specification. http://www.trustedcomputinggroup.org/resources/tpm_ main_specification

[6] Trusted Platform Module (TPM) Summary (2008). Technical Report, Trusted Computing Group

[7] ARM (2009) ARM security technology building a secure system using TrustZone technology

[8] Intel Software Guard Extensions Programming Reference (2014). https://software.intel.com/sites/default/files/managed/48/88/329298-002.pdf

[9] Boneh D, Gentry C, Lynn B, Shacham H et al (2003) A survey of two signature aggregation techniques. RSA Cryptobytes 6(2):1–10

[10] Arcserve, "ASSURING DATA BACKUP SECURITY WITH ARCSERVE UNIFIED DATA PROTECTION," 2022.

[11] Tableau, "Big Data Analytics: What It Is, How It Works, Benefits, And Challenges".

[12] TableauPublic, "GLOBAL COVID-19 TRACKER," 21 01 2020. [Online]. Available: https://public.tableau.com/app/profile/covid.19.data.resource.hub/viz/COVID-19Cases_15840488375320/COVID-19GlobalView. [Accessed 26 04 2022].

[13] Tableau, "Tableau Blueprint," 2022. [Online]. Available: https://help.tableau.com/current/offline/en-us/tableau_blueprint.pdf.

[14] R. A. C, "OpenGenus IQ: Computing Expertise & Legacy," *Merkle Tree [Explained]*, 2022.

[15] S. Kansal, "Merkle Trees: What They Are and the Problems They Solve," Codementor, 22 01 2020. [Online]. Available: https://www.codementor.io/blog/merkle-trees-5h9arzd3n8.

[16] oananiculaescu, "k-Anonymity and cluster based methods for privacy," 22 05 2017. [Online]. Available:

https://elf11.github.io/2017/05/22/kanonymity.html
.

[17] "Big Data Security | Privacy Issues", Encryption Consulting | Encryption Consulting, 2022. [Online]. Available: https://www.encryptionconsulting.com/big-data-security-and-privacy-issues/. [Accessed: 20- Apr- 2022].

[18] "Big Data Security and Privacy Issues – A Survey," in International Conference on Innovations in Power and Advanced Computing Technologies, Chennai, 2017.

[19] D. M. Turner, "What is Key Management? a CISO Perspective," 21 02 2016. [Online]. Available: https://www.cryptomathic.com/news-events/blog/what-is-key-management-a-ciso-perspective.

[20] Li J, Chen X, Li M, Li J, Lee PP, Lou W (2014) Secure deduplication with efficient and reliable convergent key management. IEEE Trans Parallel Distrib Syst 25(6):1615–1625

[21] M. Wen, S. Yu, J. Li, H. Li and Kejie , "Big Data Storage Security," in *Springer International Publishing Switzerland 2016*, 2016.

PRIMARY STUDIES

[S1] L. Xu and W. Shi, "Security Theories and Practices for Big Data", Big Data Concepts, Theories, and Applications, 2016, pp. 157-192.

[S2] D. Puthal, S. Nepal, R. Ranjan and J. Chen, "A Dynamic Key Length Based Approach for Real-Time Security Verification of Big Sensing Data Stream", Lecture Notes in Computer Science, 2015, pp. 93-108.

[S3] K. Yang, Q. Han, H. Li, K. Zheng, Z. Su and X. Shen, "An Efficient and Fine-grained Big Data Access Control Scheme with Privacy-preserving Policy", IEEE Internet of Things Journal, 2016, pp. 1-8.

[S4] Z. Azmi, "Opportunities and Security Challenges of Big Data", Current and Emerging Trends in Cyber Operations, 2015, pp. 181-197.

[S5] E. Bertino, "Big Data – Security and Privacy", 2015 IEEE International Congress on Big Data, 2015, pp. 757-761.

[S6] Y. Gao, X. Fu, B. Luo, X. Du and M. Guizani, "Haddle: A Framework for Investigating Data Leakage Attacks in Hadoop," 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, 2015, pp. 1-6.

[S7] Y. Jeong and S. Shin, "An Efficient Authentication Scheme to Protect User Privacy in Seamless Big Data Services", Wireless Personal Communications, vol. 86, no. 1, 2015, pp. 7-19.

[S8] C. Liu, R. Ranjan, C. Yang, X. Zhang, L. Wang and J. Chen, "MuR-DPA: Top-Down Levelled Multi-Replica Merkle Hash Tree Based Secure Public Auditing for Dynamic Big Data Storage on Cloud," in IEEE Transactions on Computers, vol. 64, no. 9, 2015, pp. 2609-2622.

[S9] F. Rahman, S. Ahamed, J. Yang and Q. Wang, "PriGen: A Generic Framework to Preserve Privacy of Healthcare Data in the Cloud", Inclusive Society: Health and Wellbeing in the Community, and Care at Home, 2013, pp.77-85.

[S10] S. Sudarsan, R. Jetley and S. Ramaswamy, "Security and Privacy of Big Data", Studies in Big Data, 2015, pp. 121-136.

[S11] J. Moura and C. Serrão, "Security and Privacy Issues of Big Data," Lisboa, Portugal.

[S12] “Types of Network Attacks against Confidentiality, Integrity and Avilability”, Omnisecu.com, 2017. [Online]. Available: http://www.omnisecu.com/ccna-security/types-of-network-attacks.php. [Accessed: 23- Jan- 2017].

[S13] Goldwasser S, Micali S, Rackoff C (1985) The knowledge complexity of interactive proofsystems. In: Proceedings of the 7th annual ACM symposium on theory of computing - STOC 1985. ACM, New York, pp 291–304

[S14] Quisquater JJ, Quisquater M, Quisquater M, Quisquater M, Guillou L, Guillou MA, Guillou G, Guillou A, Guillou G, Guillou S (1990) How to explain zero-knowledge protocols to your children. In: Menezes A, Vanstone SA (eds) Advances in cryptology – CRYPTO89 Proceedings. Lecture notes in computer science, vol 537. Springer, Berlin, pp 628–631

[S15] Boneh D, Crescenzo GD, Ostrovsky R, Persiano G (2004) Public key encryption with keyword search. In: Advances in cryptology - EUROCRYPT 2004. Lecture notes in computer science, vol 3027. Springer, Berlin, pp 506–522

[S16] Agrawal R, Kiernan J, Srikant R, Xu Y (2004) Order preserving encryption for numeric data. In: ACM international conference on management of data - SIGMOD 2004. ACM, New York, pp 563–574

[S17] Hacigümü¸s H, Iyer BR, Li C, Mehrotra S (2002) Executing SQL over encrypted data in the database-service-provider model. In: Franklin MJ, Moon B, Ailamaki A (eds) Proceedings of the ACM international conference on management of data - SIGMOD 2002. ACM, New York, pp 216–227

[S18] Aggarwal G, Bawa M, Ganesan P, Garcia-Molina H, Kenthapadi K, Motwani R, Srivastava U, Thomas D, Xu Y (2005) Two can keep a secret: a distributed architecture for secure database services. In: Second biennial conference on innovative data systems research - CIDR 2005, pp 186–199

[S19] Cohen JC, Acharya S (2014) Towards a trusted HDFS storage platform: mitigating threats to hadoop infrastructures using hardware-accelerated encryption with TPM-rooted key protection. J Inf Secur Appl 19(3):224–244

[S20] Xu L, Pham KD, Kim H, Shi W, Suh T (2014) End-to-end big data processing protection in cloud environment using black-box: an FPGA approach. Int J Cloud Comput

[S21] Xu L, Shi W, Suh T (2014) PFC: privacy preserving FPGA cloud - a case study of MapReduce. In: 7th IEEE international conference on cloud computing

[S22] Delplace V, Manneback P, Pinel F, Varette S, Bouvry P (2013) Comparing the performance and power usage of GPU and ARM clusters for MapReduce. In: Third international conference on cloud and green computing - CGC 2013. IEEE, New York, pp 199–200

[S23] Goodacre J, Cambridge A (2013) The evolution of the ARM architecture towards big data and the data-centre. In: Proceedings of the 8th workshop on virtualization in high-performance cloud computing - VHPC 2013. ACM, New York, p 4

[S24] Ou Z, Pang B, Deng Y, Nurminen JK, Yla-Jaaski A, Hui P (2012) Energy-and cost-efficiency analysis of ARM-based clusters. In: 12th IEEE/ACM international symposium on cluster, cloud and grid computing - CCGrid 2012. IEEE, New York, pp 115–123

[S25] Hoekstra M, Lal R, Pappachan P, Phegade V, Del Cuvillo J (2013) Using innovative instructions to create trustworthy software solutions. In: Proceedings of the 2nd international workshop on hardware and architectural support for security and privacy - HASP 2013. ACM, New York

[S26] Schuster F, Costa M, Fournet C, Gkantsidis C, Peinado M, Mainar-Ruiz G, Russinovich M (2015) VC3: trustworthy data analytics in the cloud using SGX. In: 36th IEEE symposium on security and privacy - S&P 2015. IEEE, New York

[S27] Wang H, Yin J, Perng CS, Yu PS (2008) Dual encryption for query integrity assurance. In: Proceedings of the 17th ACM conference on information and knowledge management. ACM, New York, pp 863–872

[S28] Xie M, Wang H, Yin J, Meng X (2007) Integrity auditing of outsourced data. In: Koch C, Gehrke J, Garofalakis MN, Srivastava D, Aberer K, Deshpande A, Florescu D, Chan CY, Ganti V, Kanne CC, Klas W, Neuhold EJ (eds) Proceedings of the 33rd international conference on Very large data bases - VLDB 2007, VLDB Endowment, pp 782–793

[S29] Li F, Hadjieleftheriou M, Kollios G, Reyzin L (2006) Dynamic authenticated index structures for outsourced databases. In: Proceedings of the 2006 ACM international conference on management of data - SIGMOD 2006. ACM, New York, pp 121–132

[S30] Ruan A, Martin A (2012) TMR: towards a trusted MapReduce infrastructure. In: IEEE eighth world congress on services - SERVICES 2012. IEEE, New York, pp 141–148

[S31] N. Akhtar, N. Tabassum, A. Perwej and Y. Perwej, "Data analytics and visualization using Tableau utilitarian for COVID-19 (Coronavirus)," Global Journal of Engineering and Technology Advances, 2020.

[S32] G. Becker, "Merkle Signature Schemes, Merkle," Ruhr-Universit¨at Bochum, 2018.

[S33] K. K. Pandey, "Security and Privacy Challenges in Big Data," Bilaspur, 2018.

[S34] P. Jain, M. Gyanchandani and N. Khare, "Big data privacy: a technological," Journal of Big Data, pp. 1-25, 2016.

[S35] Z. El Ouazzani and H. El Bakkali, "A new technique ensuring privacy in big data: K - anonymity without prior value of the threshold k", 2022.

[S36] Domingo-Ferrer and J. Soria-Comas, "Anonymization in the Time of Big Data", 2022.

[S37] X. Li, Z. Yan and P. Zhang, "A Review on Privacy-Preserving Data Mining," 2014.

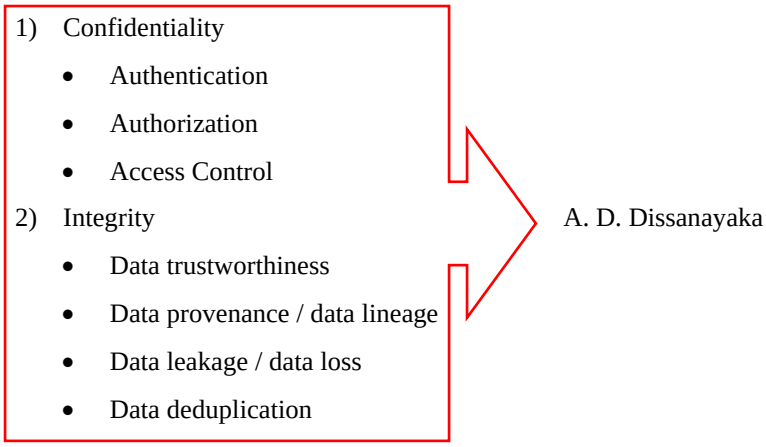
[S38] C. Zhao et al., "Secure Multi-Party Computation: Theory, practice and applications", 2022.

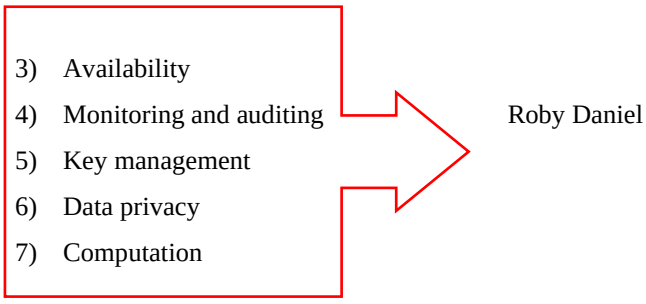
[S39] G.B. Aswani Kumar, Dr.K. Venkataramana, 2015, A Secure MapReduce Scheme for Big Data, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NCACI – 2015 (Volume 3 – Issue 18),

REFLECTION

It was a great learning experience of using the knowledge which I acquired through lectures, practical sessions, and self-studies, and how to work as a team in an effective and successful way. We were able to successfully identify the security and privacy challenges of big data and their impact in real world applications by utilizing the skills and knowledge of each group member, while maintaining professional and effective communication.

In this assignment, me and my group member discussed together and finalized the topic "Systematic Literature Review on Security and Privacy Challenges Of Big Data". Then we gathered information related to our topic. After that, we came up with a plan to succeed the project while meeting the deadline. We shared the major subtopics we selected for this project as follows.





As the two topics confidentiality and integrity are the most significant challenges of big data, we decided to cover up them by one person of the group and the rest of the topics by the other group member as shown in the above figures. Apart from these topics, I wrote down the abstract, introduction, research methodology, and findings of the research while my assignment partner focusing on the discussion and conclusion of the paper.

I learnt about the existing security and privacy challenges of big data and existing countermeasures proposed by various authors. Then I was able to draw conclusions based on them. As we did compare a large number of security mechanisms for our study, now I have a good exposure to the security mechanisms which are used in the industry. One of the most essential things I have learnt is the importance of working as a group. After completing my master's program, eventually, I will work for some organisation in the industry, and I will be working with a team. To be successful as a team we need to be able to adapt, communicate effectively and professionally, respect others' opinions, manage time effectively and find the best way to utilize the skills and knowledge of the team members. These aspects usually do not come from theoretical lessons, by completing this assignment I have gained a firm grip on how to succeed as a team.

One of the difficulties that we have faced is selecting suitable previous work for our case study. But we both shared our knowledge on various platforms to download published research papers and finally we achieved the objective with a number of research papers related to our topic.

Working as a group also comes with extra challenges such as time management, writing skills, agreeing on things, meeting deadlines, following guidelines (follow up of base paper and other guidelines of the assignment) etc. To manage the time effectively we assigned specific tasks to each team member, and we set a time to get together to share the experience. When agreeing on things we would arrange a meeting to discuss the pros and cons and draw a conclusion to take the decision. When we were not able to arrange a meeting, for an example late in the evening, we would use online channels such as Microsoft Teams to conduct meetings.

The draft submission method is very effective and important for us, as we were able to correct many mistakes with that. Also, I would like to recommend that it would be better if you could provide the feedback of the draft a little earlier to the final deadline. To conclude things, I would like to mention that this was a great opportunity for me, not just to use my knowledge and skills on a practical scenario but also to work as a team in an effective and successful way to meet a deadline.