SAVONIA
University of Applied Sciences

THESIS – BACHELOR'S DEGREE PROGRAMME
TECHNOLOGY, COMMUNICATION AND TRANSPORT

HOW IS BIG DATA RELATED TO INFORMATION SECURITY

A literature review

AUTHOR/S Guyangyang Yao

SAVONIA UNIVERSITY OF APPLIED SCIENCES

THESIS
Abstract,

Field of Study Technology, Communication and Transport	
Degree Programme Degree Programme in Information Technology, Internet of Things	
Author(s) Guyangyang Yao	
Title of Thesis How is Big Data related to information security - A literature review	
Date 4 May 2023	Pages/Number of appendices 41/1
Client Organisation /Partners	

Abstract

The purpose of this thesis was to explore the relationship between Big Data and information security and their impact on each other.

By reviewing related literature, the thesis introduces today's popular Big Data technologies and information security problems. In addition, it also explored how Big Data can be made more secure with these Big Data technologies and how Big Data can be utilized to solve information security problems.

This thesis’s main finding is that there is a relationship between Big Data and information security, which means they influence each other. The development of Big Data has led to many information security problems. This has led to the development of more secure Big Data technologies at the same time. On the other hand, Big Data has been utilized to create models which enhance information security.

Keywords
Bid data, Information security, Hadoop, Spark, NoSQL, Cloud computing

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

In today’s digital age, Big Data and information security are increasingly inextricably linked. Due to the popularity of the Internet and the rapid development of computer technology, the application area of Big Data has been expanding and has become an important support for today’s digital economy and social development. Big Data brings convenience to our lives, and people can use Big Data to improve customer service and make better decisions.

However, the application of Big Data has brought new challenges to information security. As the use of Big Data continues to grow, the scale and complexity of the data also increase, resulting in more and more information security problems. Therefore, it is crucial to protect the security of Big Data.

The purpose of this thesis is to explore the relationship between Big Data and information security and how they affect each other by reviewing various resources such as Internet publications and e-books to support this.

This thesis is divided into seven main chapters. The first and seventh chapters of this thesis are the introduction and conclusion. The second chapter provides the necessary background information on the topic by defining Big Data and information security, and discussing the features, advantages, and disadvantages of Big Data.

The third chapter of this thesis reveals the prevalence of information security problems through a survey performed by the Enterprise Strategy Group and Splunk, leading to information security problems that frequently arise today, including personal information leakage, virus invasion, spam, hacking attacks, and ransomware.

Chapter Four introduces popular Big Data technologies. Chapter Five delves into ways of making Big Data more secure within these technologies.

Chapter Six is the highlight of this thesis, as it explores how Big Data can be utilized to enhance information security. Using Big Data and artificial intelligence enables the formation of predictive models, intelligent risk management, threat visualization, and incident response to improve information security.

2 BACKGROUND

2.1 Big Data

Big Data typically means data sets, including structured and unstructured data. Due to its volume and complexity, it cannot be processed by traditional techniques or algorithms and requires specialized tools and methods to manage and analyze. (Sakyi 2016)

Big Data is currently being used in a variety of applications such as healthcare, finance, and a variety of other sectors. E.g., Figure 1 shows the percentage of Big Data used in different industries, with the telecommunication sector using 87% of Big Data, the financial services sector using 76% of big data, and the healthcare sector using 60% of Big Data. Big data is a valuable resource that can be utilized to make informed business decisions, enhance customer experiences, foster innovation, and so on. (Joshi 2022)

In addition, the rapid growth of social media, mobile devices, artificial intelligence (AI), and the Internet of Things (IoT) has increased the complexity of data sources relative to traditional data. Much of the data is real-time, generated at scale. (IBM s.a.)

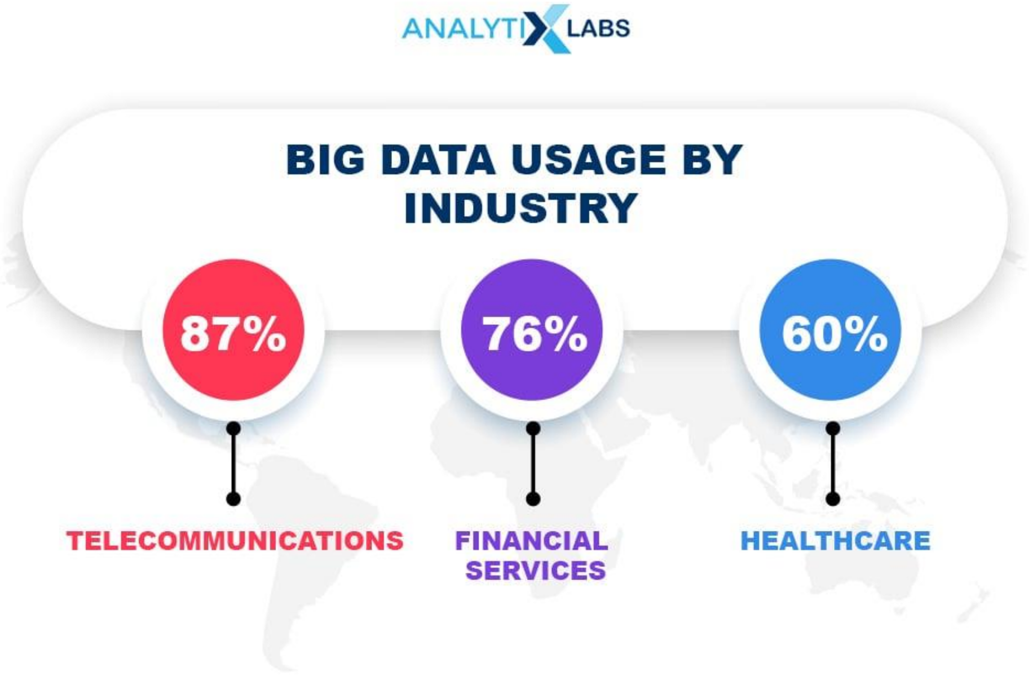


Figure 1. Big Data usage by industry (Joshi 2022)

2.2 The features of big Data

As shown in Figure 2, big data is typically described by five essential characteristics: volume, value, variety, velocity, and veracity.

- Volume: Huge amount of data collected, calculated, and stored. For example, Walmart needs to load more than 2.5 petabytes of data into its database each hour to handle more than one million customer transactions. That is around 167 times as much data as there are books in the whole Library of Congress collection. (Griffith University s.a.)

- Value: Value is the extent to which the data is useful in decision-making. It is the most important feature. Data should be valuable and reliable. If some insight cannot be gained from it, then the data is worthless. (Gutta 2020)

- Variety: Diversification of data types and sources. This includes unstructured data such as videos, semi-structured data like CSV files, sensor data, and structured data such as first name, phone number. (Gutta 2020; Lahn 2019)

- Velocity: Data growth, processing, and collection occur at a rapid pace. As an illustration, on an average day, Google processes more than 63,000 searches every second. (Lahn 2019)

- Veracity: The quality of data. Accuracy and reliability of data. Because of the variety of data, one needs to check the accuracy of the data before using it. (Gutta 2020)

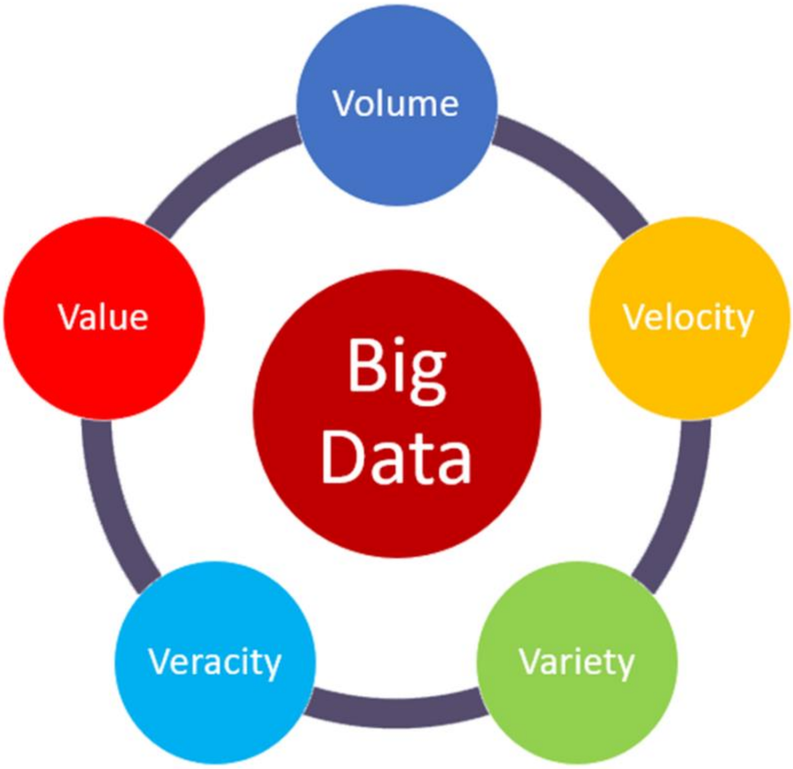


Figure 2. The 5V's of Big Data (Gutta 2020)

2.3 Disadvantages and advantages of Big Data

2.3.1 Advantages of Big Data

There are many advantages in Big Data and a few of them are listed below:

- Improve customer service: Businesses can leverage Big Data to predict customer preferences, and their behavioral tendencies to improve their products and provide better service. (Javatpoints s.a.; Rawat 2022)

- Better decision-making: Big Data yields leading-edge analytical insights and business intelligence, and a company can collect more customer data to gain a deeper understanding of its target market and make more appropriate decisions. (Javatpoints s.a.; Rawat 2022)

- Instant results and feedback: With the real-time fluidity of information and automated Big Data analysis process, analysis results, and feedback can be obtained quickly and timely. (Tenorio 2021)

- Fraud detection: Big Data provides hints about potential fraud. Many financial institutions use Big Data to uncover fraud and detect anomalies and transactional trends by using artificial intelligence and machine learning algorithms in Big Data. (Javatpoints s.a.; Rawat 2022)

2.3.2 Disadvantages of Big Data

Meanwhile, Big Data also has many disadvantages and some of them are listed below:

- High costs: Although numerous Big Data tools are built on open-source technologies, they still come with significant costs associated with hardware, maintenance, and related services. (Harvey 2018)

- Increased risk of information leakage: Access to large data sets may gain unwanted attention from hackers, making users the target of cyber-attacks and causing leaks of sensitive information. (Editorial Team 2021)

- Data quality is problematic: Due to the diversity and volume of data, it is impossible to guarantee the quality of the data and much of it is worthless. (Harvey 2018)

- Lack of talents: There is lack of Big Data specialists. Big data involves processing and analyzing large and complex data sets, often using advanced technologies and tools therefore there is a huge demand for professionals with strong backgrounds in related fields. However, there is a shortage of qualified individuals who possess these skills and can handle big data effectively. (Javatpoints s.a.)

2.4 Information security definition

InfoSec is a common abbreviation for information security. It describes the procedures and devices created and used to safeguard private information or sensitive corporate data against change, examination, and erasure. (Cisco s.a.) There are six types of InfoSec:

- Application security: It is a crucial part of the perimeter defense for information security. It addresses software flaws in application programming interfaces (APIs), and mobile and online apps. (Cisco s.a.; Loshin 2022)

- Cloud security: It refers to a set of security rules that protect cloud-based data, apps, and infrastructure. These approaches aid regulatory data compliance by guaranteeing user and device authentication, data and resource access control, and data privacy protection. (Box s.a.; Froehlich, Shea & Cole 2021)

- Cryptography: The importance of encryption and cryptography is rising. Data confidentiality and integrity are aided by the encryption of both data at rest and data in transit. In cryptography, digital signatures are frequently used to confirm the validity of data. (Cisco s.a.; Richards 2021)

- Infrastructure security: It involves the process of defending vital infrastructure from both physical and virtual attacks. In addition to protecting internal and external networks, workstations, servers, data centers, and mobile devices are also protected. (Cisco s.a.; Hewlett Packard Enterprise s.a.)

- Incident response: It is the function of monitoring and investigating data breaches or cyberattacks. Having an incident response plan allows an organization to quickly detect and stop attacks, minimize damage, and prevent future attacks of the same type. (Cynet s.a.; Cisco s.a.)

- Vulnerability management: It is a method of looking for flaws in the environment to uncover possible risks and vulnerabilities and using this information to prioritize corrective measures. (Cisco s.a.; Cavalancia 2020)

3 INFORMATION SECURITY PROBLEMS TODAY

3.1 Overview

As technology continues to evolve swiftly, Big Data technology has become increasingly important, making our lives more convenient. However, it has also brought increased information security risks. The State of Security 2022 report (Splunk 2022), which surveyed over 1,200 security leaders worldwide and was performed by the Enterprise Strategy Group and Splunk, revealed that there has been a sharp increase in data leakage incidents, with 49% of organizations reporting such incidents in the past two years, up from the previous year. In addition, the number of compromised corporate emails has also risen, 51% of companies reported business email compromise. Furthermore, ransomware attacks have become more prevalent, 79% reported having experienced ransomware attacks, with 35% admitting that one or more of those assaults caused them to lose access to data and systems. (Splunk 2022). This section will describe the information security issues that people have faced in recent years.

3.1.1 Personal information leakage

Big Data presents significant challenges to the protection of personal information. This type of information can include an individual's name, address, sensitive details, credit information, and more, and it is of immense value and importance. (Bitdefender s.a.)

When using software or browsing websites, users may notice that their activities and behaviors are monitored and analyzed by algorithms, which are then used to make product or content recommendations. Unfortunately, some unscrupulous companies use this information to explore potential customers or to sell their personal data for profit. Additionally, malicious actors may exploit this data to engage in fraud, extortion, and other illegal activities. (Ji, Sun, Zhang, & Li 2022)

3.1.2 Virus invasion

A computer virus is a program capable of self-replication and spreading to other computers, disrupting, or destroying its regular use. Computer viruses have become a significant problem in academic computing, causing millions of dollars of damage each year and impeding the free exchange of information vital to education. These viruses can potentially modify the behaviors of their host computers, destroying research and teaching data and damaging computer equipment. Computer viruses seriously threaten academic institutions and can potentially cause widespread harm. (Schneider 1989, 334).

The Trojan horse virus is a type of network attack program. It can manipulate a user's computer by remotely stealing or maliciously modifying files, snooping on system information, stealing various commands and passwords, and even formatting the user's hardware. In addition, the Trojan can memorize keystrokes by keylogging them to obtain accounts and passwords for electronic banking and thus steal the user's wealth. Trojan horses also increase the risk of the machine being invaded by other malicious viruses. (Zhu 2015, 95)

3.1.3 Spam

Spam is unsolicited and unwanted information, usually sent from a computer to a mobile phone via an email address or instant messaging account (Malwarebytes 2023). This not only consumes the user's time and resources but also poses a risk to data privacy and security. These messages are often deceptive and may contain untrue offers or requests such as "You have won a gift card that needs to be redeemed," "Your credit report contains negative information and can be removed for a fee," or "Your account has been deactivated for security reasons, and you need to take steps to reactivate it."

This is also a problem that affects certain Apple products. Whenever you are using the Internet or Bluetooth, other people may be able to share or spread information via AirDrop. Even if you turn off this feature, they may still send spam messages via the calendar. (Bohon 2022; Miller 2021)

3.1.4 Hacking attacks

Hacking is the unauthorized exploitation of computer and system assets. Computer hackers change computer equipment and programs to achieve goals other than the maker's unique reasons. People involved in computer hacking activities are often referred to as hackers. (Kumar & Agarwal 2018, 2253).

Nowadays, the smart home is a popular trend. People will buy smart cameras and control or view their homes in real time remotely from their phones or tablets. Hackers can hack into these cameras' programs to steal user information and video or remotely monitor and listen to the users' homes.

In January 2023, Finnish stores sold 1,300 webcams with dangerous backdoors. These backdoors are openings left in devices or programs that can steal control and data from the outside. The webcams have a root user account with all access rights to the device. All cameras of this model use the same password, so anyone who knows the password can access the cameras and can access the primary users' account by sending specific text commands from the outside, taking over the camera for eavesdropping and subsequent attacks. If the camera is directly connected to the public network, hackers can continue to attack the victim's internal network or view the images sent by the cameras through the cameras. (Kärkkäinen 2023)

3.1.5 Ransomware

Ransomware is a kind of malicious software that is created with the intention of blocking access to a computer system or critical data until a ransom is paid by the victim. Ransomware attacks usually take the form of a Trojan horse that disguises its malicious payload as a legitimate file or software. Once ransomware has infected a system, it may restrict access to critical data by encrypting essential files on the system or the entire hard drive or even threaten to redistribute compromised sensitive information unless the victim pays a certain amount of money. (Popoola & Isaiah 2017)

LockBit 2.0 is a kind of ransomware, it uses dual ransom techniques as part of an attack to force victims to pay a ransom (Elsad, Gumarin & Barr 2022). In 2022, Savonia University of Applied Sciences was the target of a ransomware attack and suffered a massive security breach. Using

Russian LockBit 2.0 encryption software installed on the university’s computers, the criminals encrypted and stole data and rendered all files inaccessible to teachers, requiring a payment of Bitcoin to unlock them. (Kärkkäinen & Linnake 2022; Meri & Salokangas 2022)

4 BIG DATA TECHNOLOGY

4.1 Hadoop

Hadoop is a fundamental tool that is frequently utilized by organizations and researchers to handle and analyze Big Data. It is a distributed computing framework capable of handling large data sets. Here Figure 3 shows the Hadoop core, which is composed of two basic components: Hadoop Distributed File System (HDFS) and MapReduce. (Beakta 2015, 14)

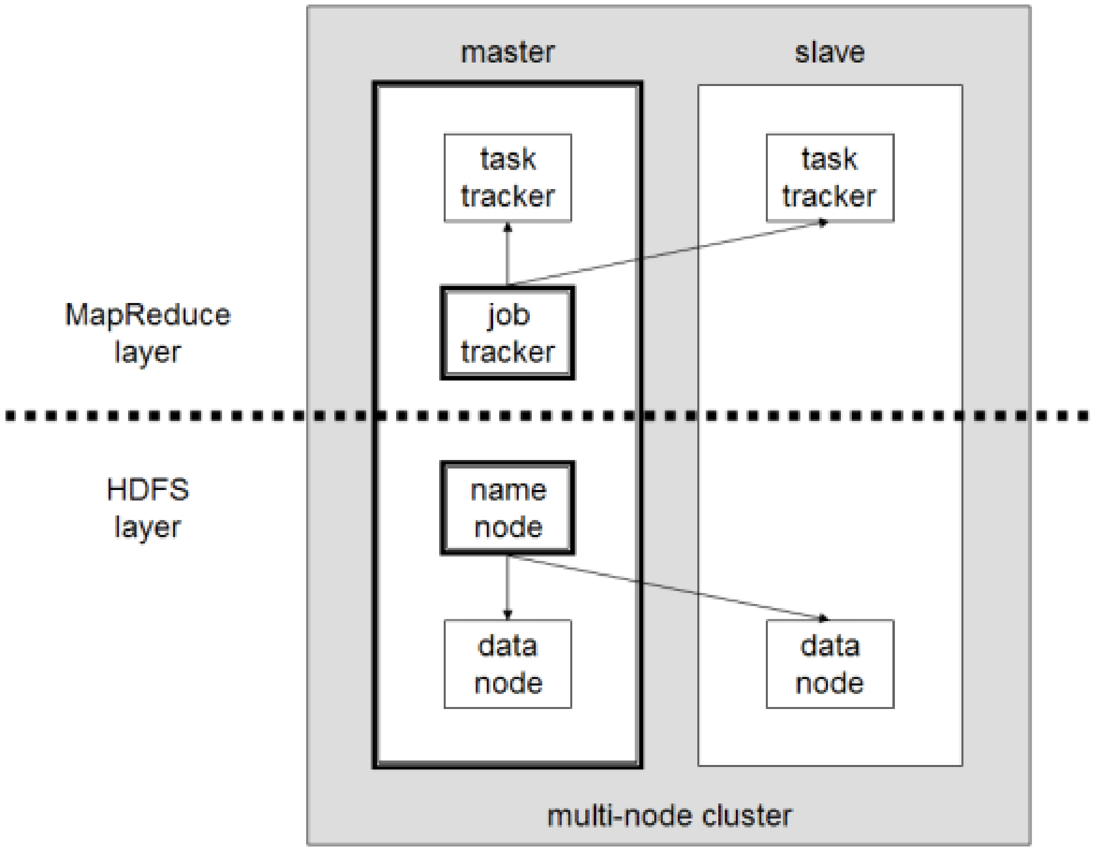


Figure 3. Hadoop Architecture (Agrawal 2014)

4.1.1 Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) provides a distributed file system. It can run on applications with big data volumes and give high throughput accessibility for application data. With the Hadoop Distributed File System (HDFS), data can span thousands of servers. When one or more of these servers may not be working properly, the Hadoop Distributed File System (HDFS) can detect the failure and recover quickly and automatically without loss or interruption of work. (Antony, Boudnik, Adams, Shao, Lee & Sasaki 2016, 2-3).

The Hadoop Distributed File System (HDFS) is built on a master/slave model. A Hadoop Distributed File System cluster is made up of a NameNode and many DataNodes. NameNode and DataNode are both software applications that run on commodity computers. The NameNode is the master server that handles the file system namespace and controls client file access. It is the arbitrator and repository of all Hadoop Distributed File System (HDFS) information. The DataNode is in charge of managing the storage associated to the nodes on which they are running, which is normally one per node in the cluster. (Apache 2023)

Hadoop Distributed File System (HDFS) supplies a file system namespace that allows users to store data in files. On the inside, a file is broken into one or more blocks, which are then distributed for storage over a collection of DataNodes. The DataNodes handle file system client read and write requests, or they create, delete, and copy blocks as instructed by the NameNode, which handles file system namespace activities including creating, closing, and renaming files and directories. (see Figure 4) (Apache 2023)

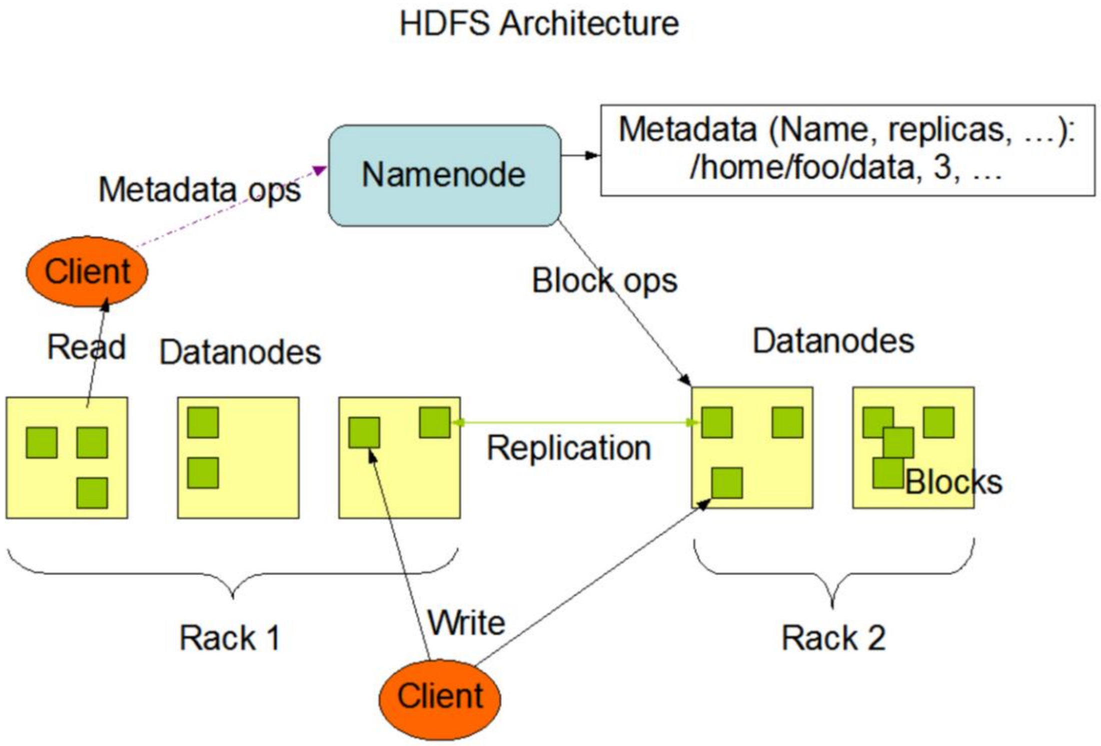


Figure 4. HDFS Architecture (Apache 2023)

4.1.2 MapReduce

MapReduce is a programming model that is a core component of the Big Data processing framework Hadoop. MapReduce is intended to process and read huge volumes of any type of data stored in a Hadoop Distributed File System (HDFS) in parallel in order to deliver the desired results. (Antony, Boudnik, Adams, Shao, Lee & Sasaki 2016, 3-4). Mapper and Reducer interfaces in MapReduce are implemented to provide Map and Reduce methods. The Mapper function converts input key/value pairs to a set of intermediate key/value pairs. The total number of blocks in the input file generally determines the number of maps. Reducer condenses a group of intermediate values that all have the same key into a smaller set of values. The user can choose the number of reductions for the job. (Apache 2023).

In a MapReduce system, a client sends a job of a specific size to the Hadoop MapReduce master. The Master then divides this job into equivalent parts and makes them available for use by the Map and Reduce functions. The input data is sent to the Map function, which returns intermediate key/value pairs. The intermediate outputs are then subjected to the Reduce function. The final result is saved on the Hadoop Distributed File System (HDFS). (see Figure 5) (Dikshantmalidev 2020).

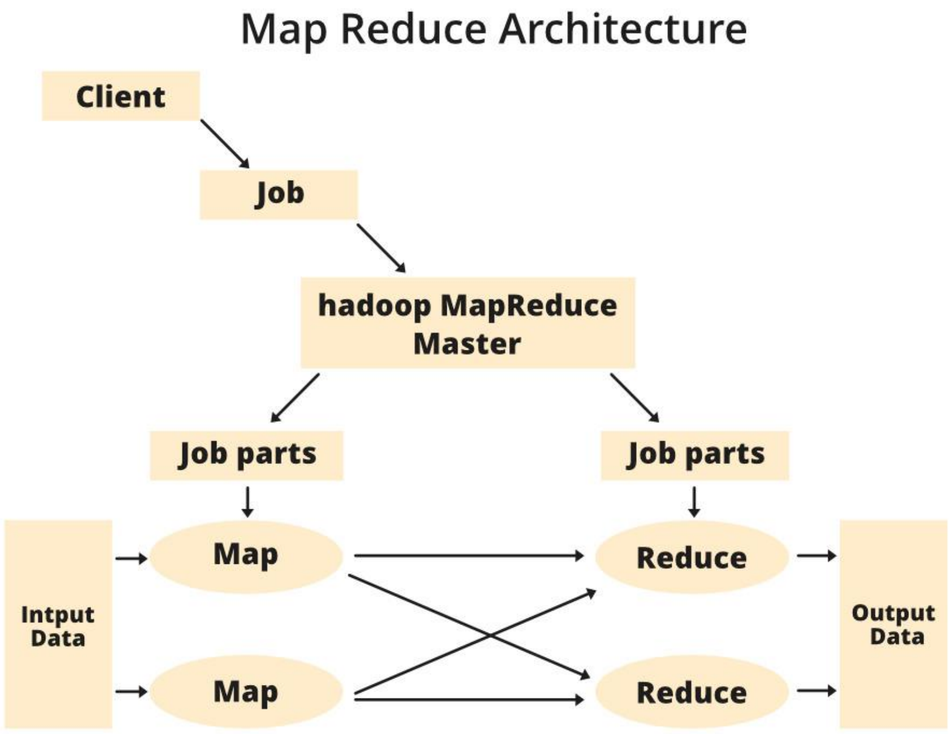


Figure 5. MapReduce Architecture (Dikshantmalidev 2020)

4.2 Cloud computing

Cloud computing is a contemporary computing paradigm that includes a diverse set of IT resources such as hardware, software, networks, services, and development processes. Cloud computing provides a flexible and scalable computing environment that enables users to develop and deliver services over the Internet or private networks. (Winkler, Speake & Foxhoven 2011, 2).

The front-end platform and the back-end platform are the two fundamental components of cloud computing architecture. The front-end platform consists primarily of the software, user interface, client devices, or networks that communicate with the back-end via a network or internet connection. It also forms an important part of how end-users connect to the cloud computing infrastructure. The front-end platform includes the graphics card and operating system, providing access to the vendor’s bespoke programs as well. Figure 6 shows that the back-end platform has seven components: applications, services, runtime cloud, storage, infrastructure, management, and security. The back-end platform drives the front-end platform and protects essential data from the demands of customer-facing technologies. (InterviewBit 2022)

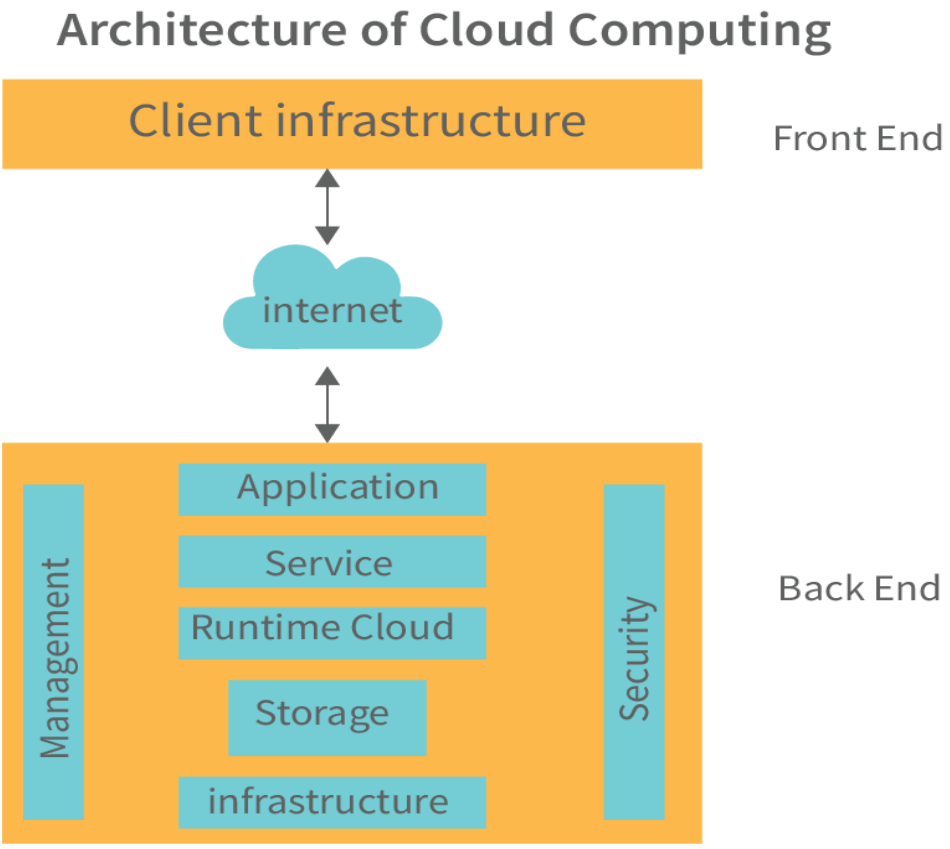


Figure 6. Cloud Computing Architecture (InterviewBit 2022)

4.3 Spark

Spark is an open-source framework for efficiently and quickly processing massive data sets stored in disparate data storage. Spark processes data in random access memory (RAM) unlike MapReduce which keeps the data on disk after each Map or Reduce operation. This is the reason why Spark executes faster than MapReduce (Madhugiri 2022). The framework is written primarily using the Scala and Java programming languages. Spark provides interfaces to a wide range of distributed and non-distributed data storage. Furthermore, Spark provides a number of language APIs for evaluating data stored in these data stores. (Mehrotra & Grade 2019, 5-8)

The architecture of Spark has a driver program and executors. A driver program is the place where users write Spark code. It is in charge of starting the cluster’s different simultaneous processes. An executor is a Java virtual machine that operates on one of the cluster's working nodes. It manages the hardware resources required to complete the activities initiated by the driver. (Mehrotra & Grade 2019, 5-8)

Figure 7 shows the relationship among driver, executors, and works. On each worker node, the driver process parses the user code and spawns numerous executors. The driver process not only forks the executors on the worker machine but also assigns tasks to these executors, allowing the entire program to operate in parallel. When the computation is complete, the output data is either returned to the driver software or written to the file system. (Mehrotra & Grade 2019, 5-8).

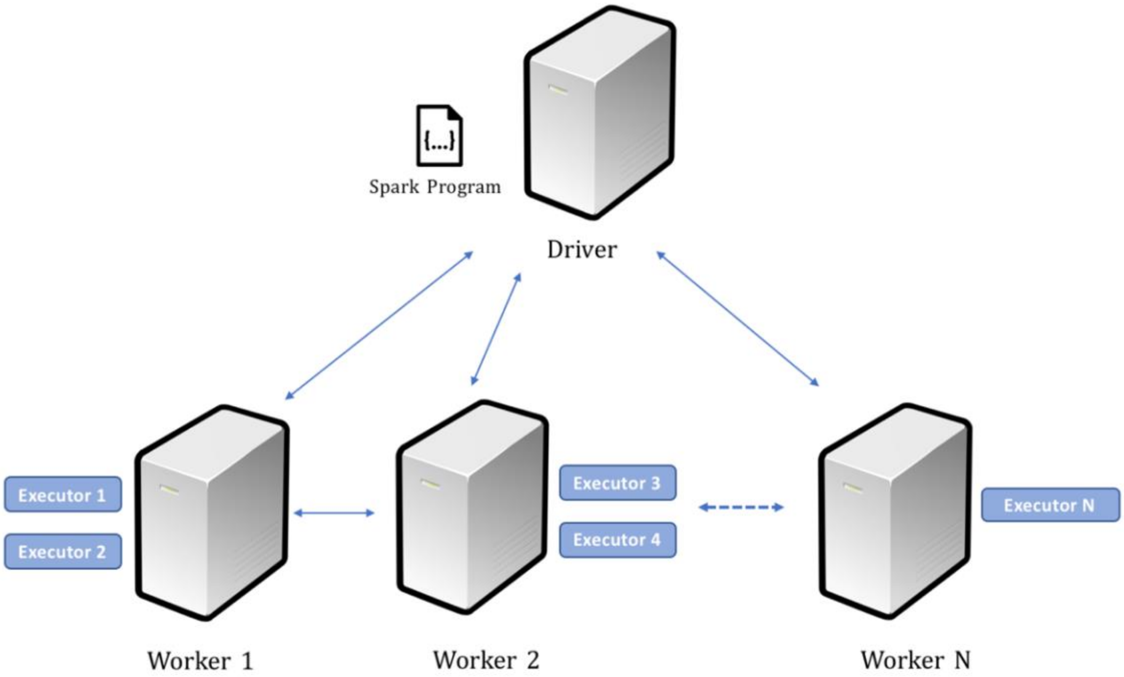


Figure 7. Relationships between Spark (Mehrotra & Grade 2019)

4.4 NoSQL

NoSQL is a non-relational database that provides a flexible mechanism for storing and retrieving data. Hbase is a NoSQL database that serves as a distributed and scalable Big Data storage on top of Hadoop. (Informit 2014).

Unlike traditional SQL databases, NoSQL does not require a fixed table schema or metadata. This makes it a more scalable and adaptable option for managing large and complex data sets. One of NoSQL's most important features is its capacity to expand horizontally, which means it can spread data. across multiple servers to improve performance and availability. It also has a dynamic schema that allows for more flexible data modeling and reduces the need for time-consuming migrations. However, NoSQL is unsuitable for complex queries requiring advanced data processing and aggregating. As such, it may not be the best choice for applications that rely heavily on complex relational queries or where data consistency is critical. (see Table 1) (Ayusharma0698 2022)

TABLE 1. Key difference between SQL and NoSQL (Ayusharma0698 2022)

SQL	NoSQL
Relational database	Non-relational database
Has fixed or static or predefined schema	Has dynamic schema
Not suited for hierarchical data storage	Best suited for hierarchical data storage

Vertically Scalable	Horizontally scalable
Best suited for complex queries	Not suitable for complicated queries.
Follows ACID (atomicity, consistency, Isolation, and durability)	Follows CAP (consistency, availability, partition tolerance)

5 HOW TO MAKE BIG DATA MORE SECURE

5.1 Securing data in Hadoop

5.1.1 Data Classification

Before data is stored, data classification is required to be conducted. The classification of certain data sets aids in determining how to transmit data across a Hadoop cluster, restrict access to data while it is kept in the cluster, and safeguard data during processing. The data can be divided into four categories as follows: (Antony, Boudnik, Adams, Shao, Lee & Sasaki 2016, 125)

- Public: Since this is public information, there are no restrictions on accessing it.

- Limited or private: Information that should not be made available to the general public. Even though this material has no sensitive aspects, it must be kept secret.

- Confidential: A dataset containing information that should be kept private, such as email addresses and phone numbers. This dataset may be restricted in access, and sensitive data pieces must be encrypted or disguised to guarantee their confidentiality.

- Restricted: Access to this dataset should be limited to a specific group of authorized users only and should not be accessible to anyone else. Additionally, some data elements may require encryption so that they can only be read by approved users who possess a secret key.

However, some users may neglect to classify their data before storing it on Hadoop Distributed File System (HDFS). In such cases, administrators need to use specialized tools to review the data schema and employ YARN's framework to determine the appropriate classification. (Antony, Boudnik, Adams, Shao, Lee & Sasaki 2016, 125)

5.1.2 Data encryption

5.1.2.1 Hadoop KMS

Hadoop KMS is a server for managing cryptographic keys that utilizes Hadoop's KeyProvider API as its foundation. It provides encryption and decryption algorithms and key management services. (Apache 2023)

At first, users can use the key generator keytool to generate keys and set permissions to form a keystore as a key storage database for the Hadoop KMS. The keystore interacts with the Hadoop KMS via the key provider API. Hadoop Distributed File System (HDFS) acts as a client. When a user requests data from the client, the client detects the user’s permission. If the user has permission and the file is not encrypted, the client opens the file directly and returns it to the user. If the file is encrypted, the client requests a key from the Hadoop KMS based on the permissions and data. Hadoop KMS determines the user’s permission and data and returns the appropriate key if it matches the permissions set in the keystore. Finally, the client decrypts the file with the key and returns it to the user. (see Figure 8) (Faxiancunzai 2017)

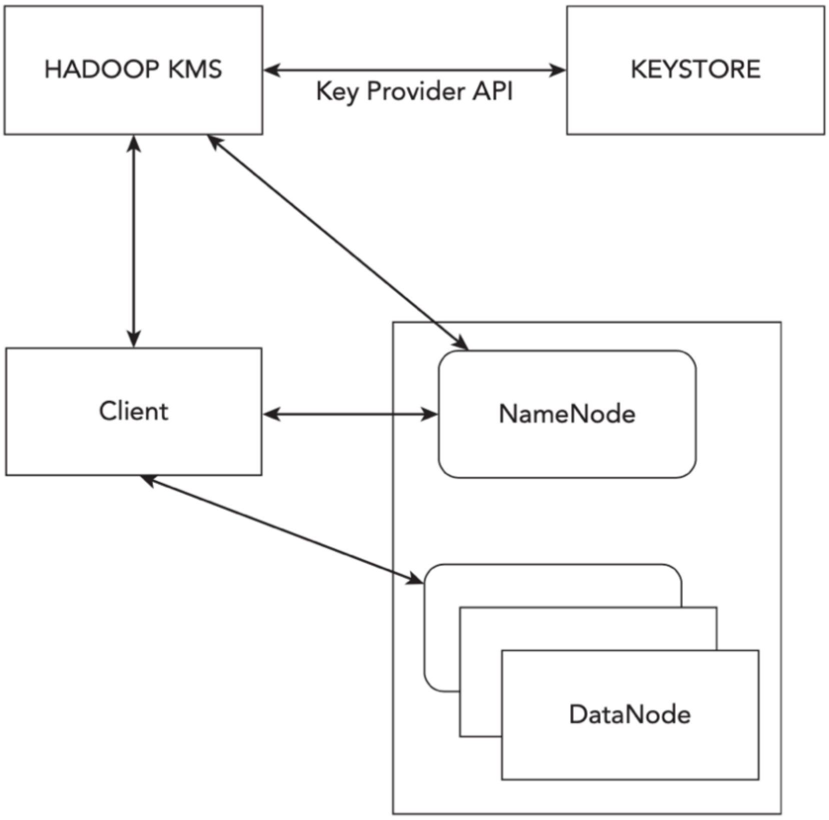


Figure 8. Key management through Hadoop KMS (Antony, Boudnik, Adams, Shao, Lee & Sasaki 2016)

5.1.2.2 Encryption support in MapReduce

There is no direct support for data encryption in Hadoop, but support for encryption can be built into a custom compression codec by taking advantage that compressing files saves disk space and speeds up the transfer of data over the network (White s.a.). Figure 9 shows a method to implement custom data encryption in Hadoop. First, extend the compression codec class in the Hadoop API. Creating a custom compression codec and adding it to the Hadoop classpath and core-site.xml configuration file. To enable compression, it needs to set the compression details in the MapReduce driver class. Then, set up the map output compression class to compress the intermediate output of the mapper, combiner, and partitioner. Finally, use the job configuration to pass the encryption key to the job to enable encryption. (Narayanan 2013, 78)

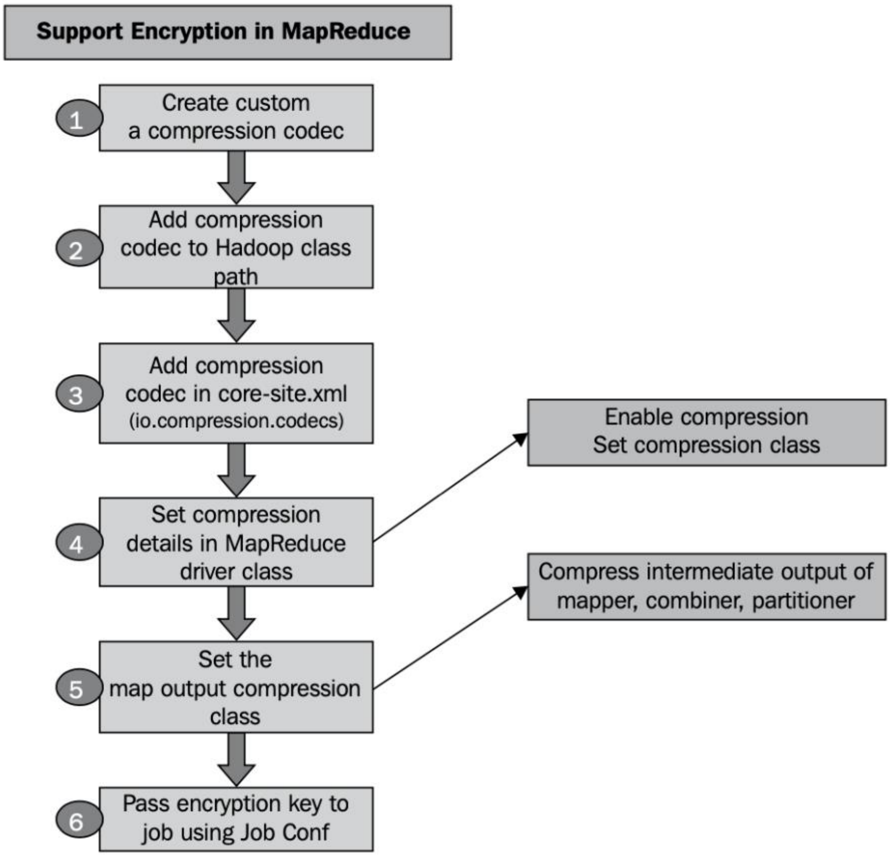

Support Encryption in MapReduce
1
Create custom a compression codec
2
Add compression codec to Hadoop class path
3
Add compression codec in core-site.xml (io.compression.codecs)
4
Set compression details in MapReduce driver class
5
Set the map output compression class
6
Pass encryption key to job using Job Conf
Enable compression Set compression class
Compress intermediate output of mapper, combiner, partitioner

Figure 9.Support Encryption in MapReduce (Narayanan 2013)

5.1.3 File and directory permission

To secure the data, file and directory permissions can also be used. They are sufficient for most data protection requirements. To authorize users, Hadoop HDFS checks file and directory permissions after the user has been authenticated. There is a file and directory permissions model in Hadoop Distributed File System (HDFS). Each file and directory in Hadoop Distributed File System (HDFS) has an owner and a group. The owner, the group, and all other users can be given rwx rights. The w permission is used to write to or modify a file. The r permission is used to list the contents of a directory, the w permission is used to create or remove a directory, and the x permission is used to access a subfolder of a directory. However, file and directory permissions are limited, as a file or directory can only have one user and one group. (Antony, Boudnik, Adams, Shao, Lee & Sasaki 2016, 129-130)

To deal with the limitations of file and directory permission, we can use Access Control Lists (ACLs). It can allow one to establish alternative permissions for a specific named user or named group, rather than merely the file's owner and group. By default, support for Access Control Lists (ACLs) is enabled and the NameNode allows Access Control Lists (ACLs) to be created. If not, it can be set (dfs.namenode.acls.enabled) to true in the NameNode configuration. Also setting Access Control Lists (ACLs) of files and directories can be used by commands (setfacl). (Antony, Boudnik, Shao, Lee & Sasaki 2016, 129-130; Apache 2023)

5.2 Securing data in Spark

5.2.1 Security Manager

The majority of security-related tasks in Spark's source code are assigned to the SecurityManager, which can check the exact implementation of your overall architecture. It is initiated via SparkContext and accessed from all drivers, workers, and masters. In addition, the Security Manager has access to the configuration and almost all configurations are passed to this class. (Ganelin, Orhian, Sasaki & York 2016)

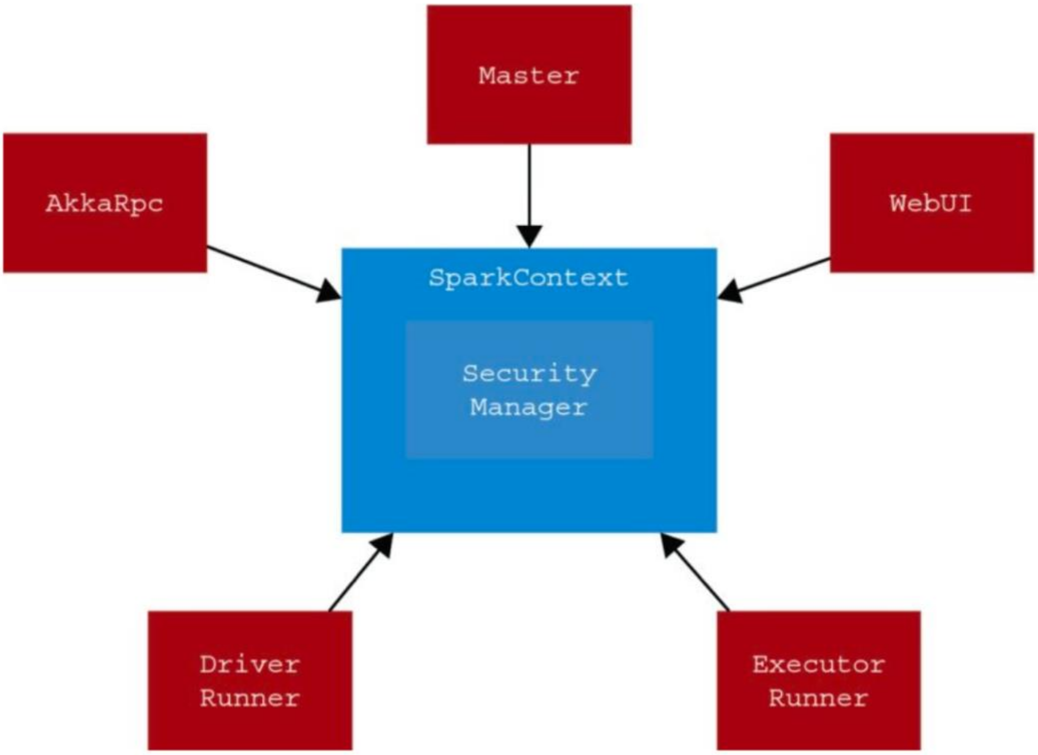


Figure 10.Security Manager (Ganelin, Orhian, Sasaki & York, 2016)

5.2.2 Data encryption

Using Spark to enable SSL/TLS encrypted connection allows encryption of the transmitted data and ensures that it is sent securely and not intercepted or read by unauthorized third parties. Before configuring SSL/TLS encrypted communication need to do the following: (Ganelin, Orhian, Sasaki & York 2016)

❿ Generate a private key for the server

❿ Make a Request for Certificate Design

❿ Obtain the signed certificate file from the Certification Authority or check the usage of SSL/TLS on a Spark cluster using a self-signed certificate

❿ Upload the signed certificate file to a reliable storage location

5.2.3 Access Control Lists (ACLs)

As mentioned in the previous chapter, Access Control Lists (ACLs) were used to handle restrictions on file and directory permissions in Hadoop. Access Control Lists (ACLs) can also be used in Spark. Spark manages Access Control Lists (ACLs) and the corresponding authentication by using the shared secret mechanism, which can help secure a Spark cluster. Only users with shared secrets can access restricted resources, thus increasing the security of data within Spark. (Ganelin, Orhian, Sasaki & York 2016)

5.3 Securing data in Cloud computing

5.3.1 Cryptographic techniques

Cloud computing security often uses cryptography, which is a technology that enables secure communication and data protection. Cryptography works by using an encryption key to transform plaintext into ciphertext, which can only be decrypted with the corresponding decryption key. (Winkler, Speake & Foxhoven 2011, 133-134)

In symmetric cryptography (see Figure 11), these keys are identical. Symmetric cryptography has wide applicability, but as each pair of communicators has to share a unique secret key, this makes key management more difficult. When a secure channel is absent, it is also highly challenging to establish a secret key between two communicating parties to securely exchange a shared secret key. (Winkler, Speake & Foxhoven 2011, 133-134)

In asymmetric cryptography (see Figure 12), the encryption key is different from the decryption key but is mathematically related. Only the public key is permitted to be released, and it is not necessary to be kept hidden. Pretty Good Privacy (PGP) is a good example. Every user has a Private Key as well as a Public Key. The data is encrypted using a public key, which can only be decoded with its matching private key. (Winkler, Speake & Foxhoven 2011, 133-134; Raicea 2017)

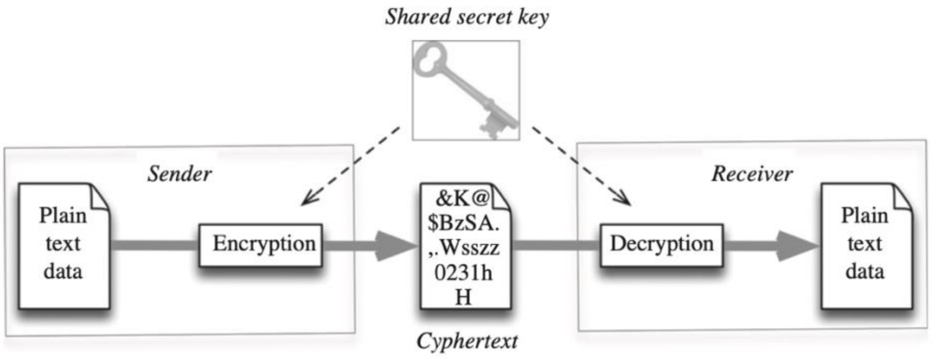


Figure 11. Symmetric encryption (Winkler, Speake & Foxhoven 2011)

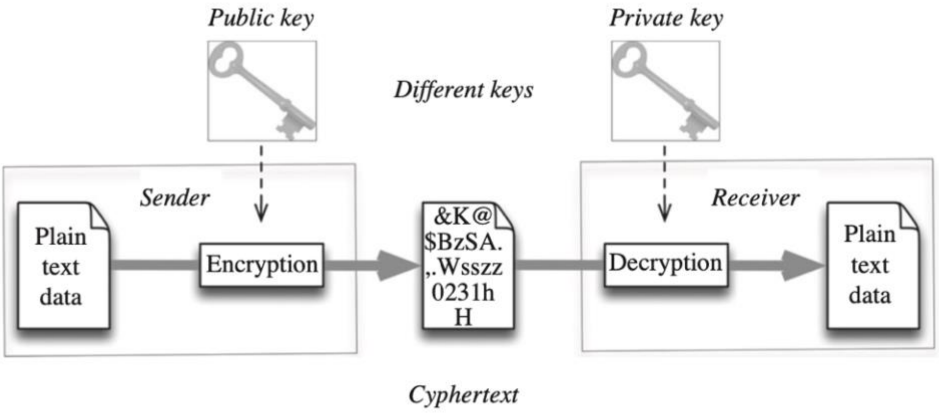
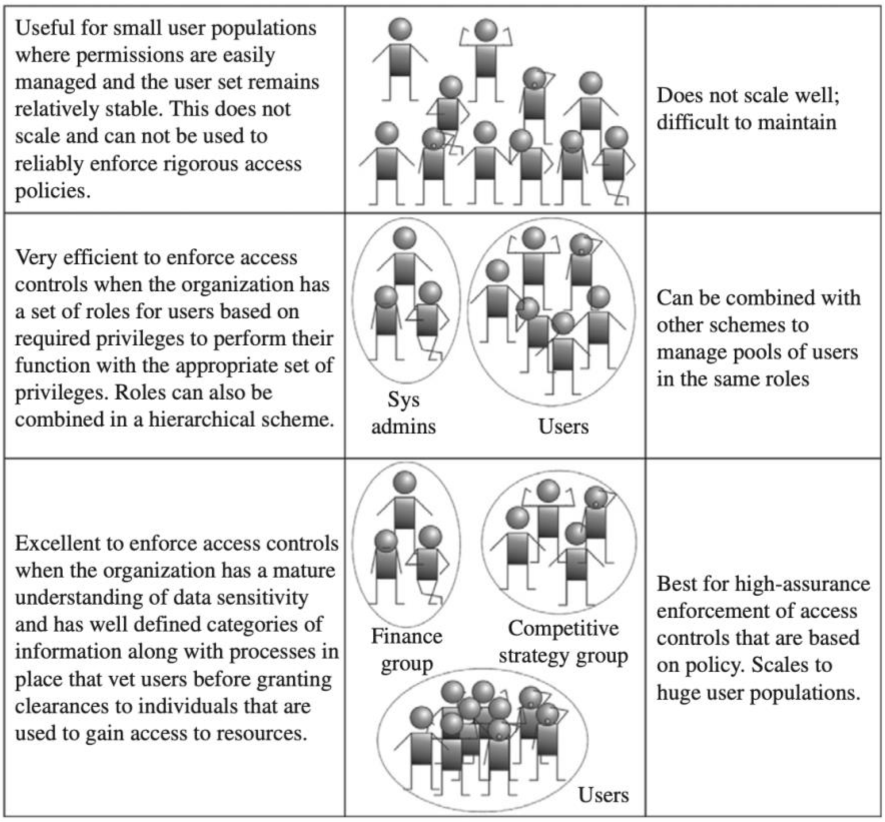


Figure 12 . Asymmetric encryption (Winkler, Speake & Foxhoven 2011)

5.3.2 Access control

Access control can also be used in the cloud. It is classified into three categories. There are three types of access control: Discretionary Access Control (DAC), Mandatory Access Control (MAC), and Role Based Access Control (RBAC). The concept of Discretionary Access Control (DAC) refers to the limitation of access to things depending on the user’s or group's identity. It maintains a relatively stable user base, but it is more suited to small user sizes and is somewhat difficult to maintain for rights management. Mandatory Access Control (MAC) is a mechanism for restricting access to resources based on their sensitivity and the level of authorization of the user. It compensates for the shortcomings of Discretionary Access Control (DAC) by being able to scale users while also having high assurance in the process of enforcing access control. Role Based Access Control (RBAC) is an approach to restrict network access based on the responsibilities of individual users inside an organization. Compared to Mandatory Access Control (MAC) in that although their access policies are both system-determined, access in Mandatory Access Control (MAC) depends on the subject's level of trust, whereas it is related to the subject’s role in Role Based Access Control (RBAC). (see Figure 13) (Winkler, Speake & Foxhoven 2011, 138-140; IBM 2021; Frontegg s.a.)

Overall, when using access control, the advantages and disadvantages of these three types can be combined to form different combinations of approaches to achieve good results. (Winkler, Speake, & Foxhoven 2011, 138-140)



DAC	Useful for small user populations where permissions are easily managed and the user set remains relatively stable. This does not scale and can not be used to reliably enforce rigorous access policies.		Does not scale well; difficult to maintain
RBAC	Very efficient to enforce access controls when the organization has a set of roles for users based on required privileges to perform their function with the appropriate set of privileges. Roles can also be combined in a hierarchical scheme.	Sys admins Users	Can be combined with other schemes to manage pools of users in the same roles
MAC	Excellent to enforce access controls when the organization has a mature understanding of data sensitivity and has well defined categories of information along with processes in place that vet users before granting clearances to individuals that are used to gain access to resources.	Finance group Competitive strategy group Users	Best for high-assurance enforcement of access controls that are based on policy. Scales to huge user populations.

Figure 13. The difference among MAC, DAC & RBAC (Winkler, Speake & Foxhoven 2011)

5.3.3 Data masking

Data masking is a technique for removing any identifying and customized qualities from data by anonymizing the changed data, lowering the danger of disclosing sensitive information while not changing the real values. One well-known data masking approach is to substitute keys from an external lookup table containing the genuine data values for the actual data values. Such disguised data values can be handled in operation with fewer controls than unmasked data. (Winkler, Speake & Foxhoven 2011, 144)

However, regardless of the data masking method utilized, each process must be performed with care to guarantee that the structure and relationships built between database rows, columns, and tables are preserved. If the masking is not performed properly, the output data can still reveal sensitive information. (Winkler, Speake & Foxhoven 2011, 144)

5.4 Securing data in NoSQL

5.4.1 Security model

Security models are the basic theoretical tools used in the development of security systems in databases (Papisetty 2020). The model is ideal for starting with a single application and a single data collection, which classifies users according to their access type and role in the organization. The concentric circle model is one of the security models. (see Figure 14) It consists of the general public which is outermost, intranet users, authenticated users, and database administrators who have granted all rights within the system. (Mccreary & Kelly 2014, 233-234)

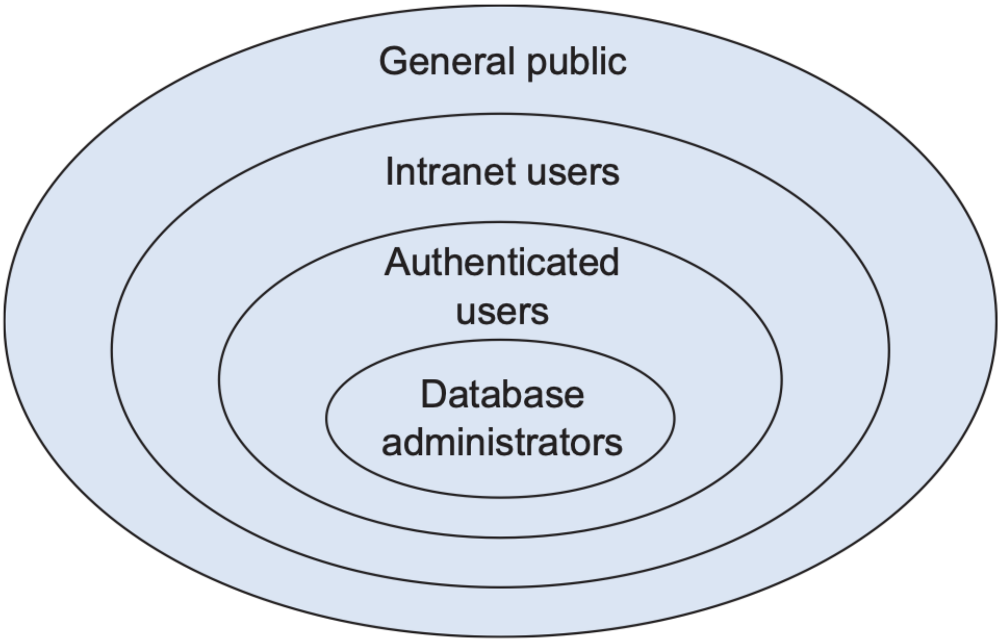


Figure 14. The concentric circle model (Mccreary & Kelly 2014)

5.4.2 Using services to mitigate the need for in-database security

Splitting an application into a series of reusable data services such as request-response services, and lookup services can avoid the more time-consuming and costly situation of moving a standalone application running on a standalone database with its own security model to a centralized, enterprise-class database running on a different security model. However, as the volume of data and synchronization complexity increases, the strategy of using services may fail. To meet more complex data requirements, data dumps as well as incremental updates must be provided and may require the use of specialized reporting tools to access these services directly. (Mccreary & Kelly 2014, 235)

5.4.3 Using data warehouses and OLAP to mitigate the need for in-database security

For database security, one can use data warehouses and OnLine Analysis Processing (OLAP) tools. Data Warehouses are centralized data storage systems used by an enterprise or organization to integrate, manage, and analyze large volumes of data. It integrates data from different data sources into a unified data model after cleansing, integration, and transformation processes, which not only ensures data quality and consistency, but also enables users to access and analyze this data flexibly and efficiently. (Mccreary & Kelly 2014, 235-236; IBM s.a.; Javatpoint s.a.)

OnLine Analysis Processing (OLAP) is a major application in data warehousing and is a core component of the implementation of a data warehouse, which allows for high-speed multidimensional analysis of the large amount of data stored in the data warehouse. It is possible to move data from the NoSQL system to OLAP while employing the option of securing the data in the OLAP tool to increase data security. Setting a policy is one of the options that enables reports to be generated only when there is a minimum number of responses so that individuals cannot be identified or their private data viewed. (Mccreary & Kelly 2014, 235-236; IBM s.a.)

6 HOW TO USE BIG DATA TO IMPROVE INFORMATION SECURITY

6.1 Predictive models

Predictive modeling is a statistical technique for developing predictive algorithms or creating predictive models to predict and analyze potential future behavior, trends, and outcomes by training artificial intelligence models using big data and machine learning algorithms. (Ali 2020)

To ensure network and information security, one can use a predictive model to identify potential vulnerabilities, hacking attacks, threats, etc. In the report A Survey on Internet of Things and Cloud Computing for Healthcare (Dang, Piran, Han, Min & Moon 2019), they created a security model (see Figure 15) to guard against any weaknesses, vulnerabilities, and assaults that may happen in the IoT in healthcare model using this method. In the security model, the security experts first examine potential threats or attacks and choose a range of properties for unknown attacks to train the Al model. When attacks happen, the Al model will recognize the attacks at first. If it is a known attack, the Al model will select the appropriate approach to mitigate the damage. If it is an unknown attack, it will be eliminated.

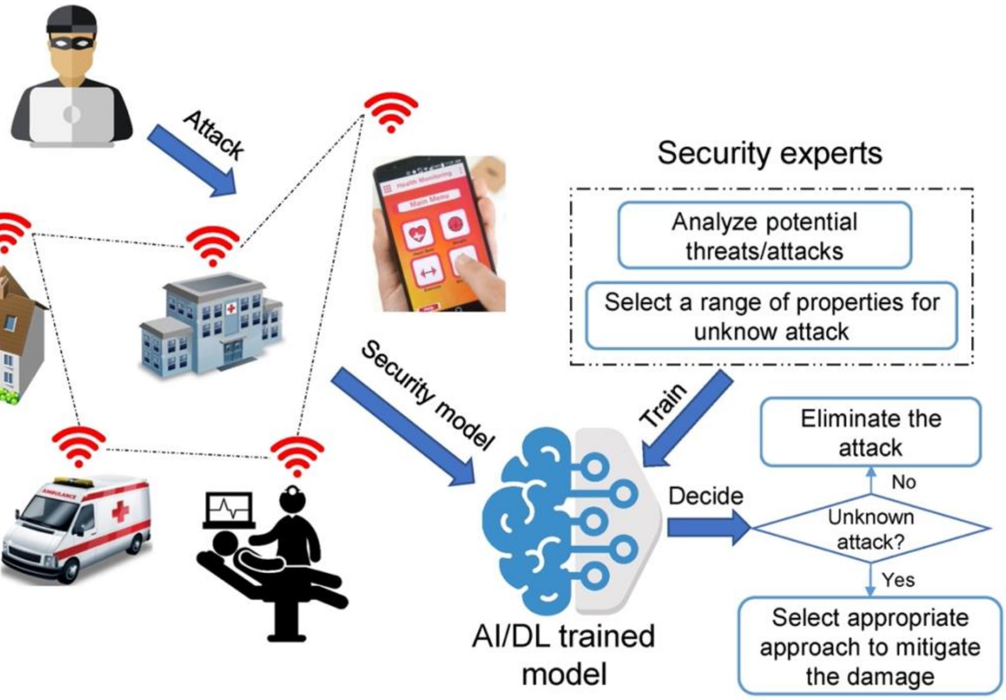

Attack
Security experts
Analyze potential threats/attacks
Select a range of properties for unknow attack
Train
Security model
Eliminate the attack
No
Decide
Unknown attack?
Yes
Select appropriate approach to mitigate the damage
AI/DL trained model

Figure 15. Security model (Dang, Piran, Han, Min & Moon 2019)

In the report from Towards Developing a Robust Intrusion Detection Model Using Hadoop–Spark and Data Augmentation for IoT Networks (Alejandro, Sanchez, Zaman, Goel, Naik & Joshi 2022), a predictive model was also used. The authors used the entire BoT-IoT dataset which is botnet attack traffic in IoT networks for detecting incoming intrusions for cyber-attacks on IoT devices detection to train machine learning models to detect anomalies in the IoT network (Leevy, Hancock, Khoshgoftaar & Peterson 2021). They created two systems. The first one will identify whether a data sample is malicious or normal. If the data sample is malicious, it will go to the second system, in which the trained machine learning model will perform intrusion detection to determine the type of attack. The attacks will be categorized as theft, reconnaissance, distributed denial of service (DDoS), and denial of service (DoS). (see Figure 16)

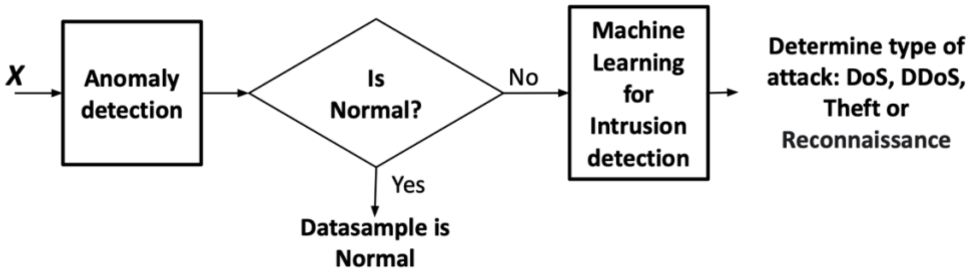

X
Anomaly detection
Is Normal?
No
Yes
Datasample is Normal
Machine Learning for Intrusion detection
Determine type of attack: DoS, DDoS, Theft or Reconnaissance

Figure 16.Two systems (Alejandro, Sanchez, Zaman, Goel, Naik & Joshi 2022)

6.2 Incident response

Incident response refers to a collection of information security policies and processes that employ Big Data to identify, control, and eradicate cyber-attacks. It lets enterprises to promptly detect and halt assaults, limit damage, and avoid similar attacks in the future. (Cynet s.a.)

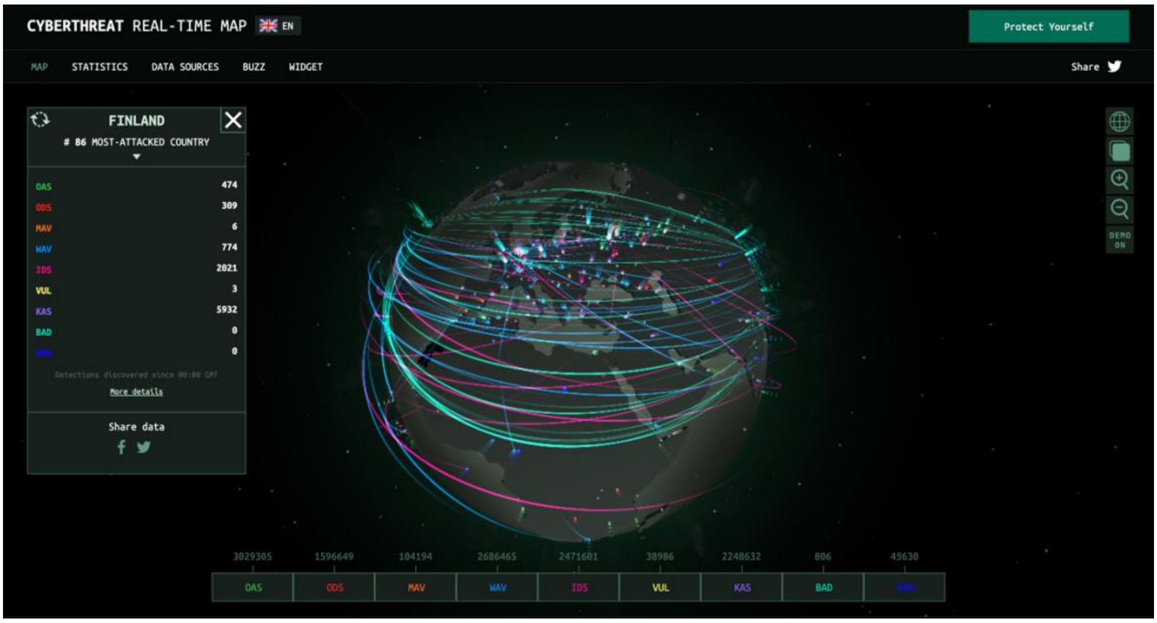
However, many of today’s security operations teams are always understaffed to handle increasing amounts of attacks. That is why many functions of incident response must be handled by automation. One can train artificial intelligence models using big data and machine learning algorithms to automatically analyze and correlate data from disparate sources to identify and classify incidents that threaten an organization’s cybersecurity, as well as automate routine, standardized tasks to speed up the incident response process and improve the efficiency and effectiveness of security operations teams. (Irei 2023)

For example, Fully Integrated Defense Operation (FIDO) is a Netflix orchestration layer that automates incident response by detecting, reviewing, and responding to malware and other detected threats. FIDO receives events through APIs, SQL databases, log files and emails, collects data, scores it for analysis, and decides whether to send emails to the security team or disable accounts, network ports and other measures based on the score. (Fry, Evans & Chan 2015)

6.3 Threat visualization

One can use data analysis techniques to create threat visualizations. Security experts or data analysts study cyber threat intelligence using analysis tools to reveal trends, patterns, outliers, and anomalies. The data is visualized using charts, bar graphs, 3D charts, etc. to monitor and quickly identify potential threats and attacks in real time so they can develop appropriate security strategies to minimize damage. (McGuigan s.a.)

Figure 17 is a real-time cyber threat map made by Kaspersky. This is a good example of threat visualization where users can understand the cyber threats occurring in real-time in different countries and regions by using different color lines and different patterns. Like Finland, which is ranked 86th among the most attacked countries. Figure 18 also shows what kind of virus they scanned per week and per country. Trojan horses account for a large percentage of Finland.



FINLAND

86 MOST-ATTACKED COUNTRY

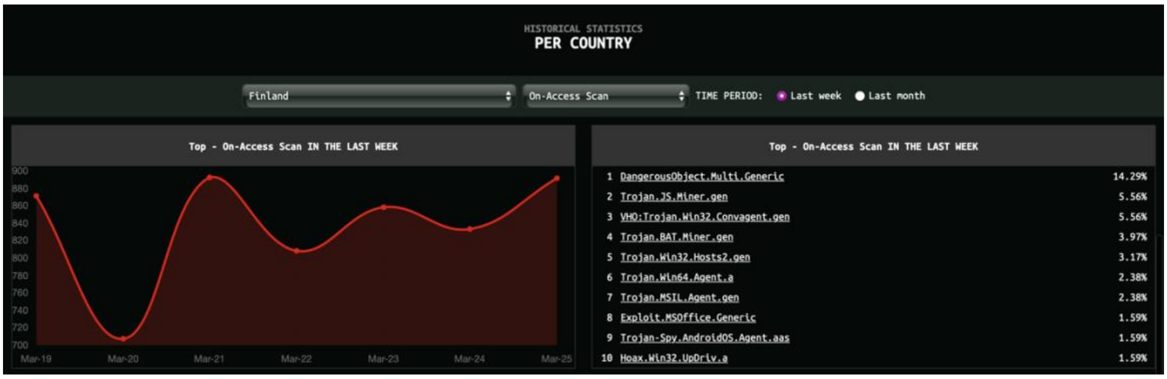
OAS	474
ODS	309
MAV	6
WAV	774
IDS	2021
VUL	3
KAS	5932
BAD	0
[illegible]	0

Detections discovered since 00:00 GMT

More details

OAS	ODS	MAV	WAV	IDS	VUL	KAS	BAD	[illegible]
3029305	1596649	104194	2686465	2471601	38986	2248632	806	45630

Figure 17. The real-time cyber threat map (Kaspersky s.a.)



HISTORICAL STATISTICS

PER COUNTRY

Finland

On-Access Scan

TIME PERIOD: ● Last week ○ Last month

Top - On-Access Scan IN THE LAST WEEK

Top - On-Access Scan IN THE LAST WEEK

1	DangerousObject.Multi.Generic	14.29%
2	Trojan.JS.Miner.gen	5.56%
3	VHO:Trojan.Win32.Convagent.gen	5.56%
4	Trojan.BAT.Miner.gen	3.97%
5	Trojan.Win32.Hosts2.gen	3.17%
6	Trojan.Win64.Agent.a	2.38%
7	Trojan.MSIL.Agent.gen	2.38%
8	Exploit.MSOffice.Generic	1.59%
9	Trojan-Spy.AndroidOS.Agent.aas	1.59%
10	Hoax.Win32.UpDriv.a	1.59%

Figure 18. Historical statistics per country (Kaspersky s.a.)

6.4 Intelligent risk management

Using Big Data for risk assessment, a business or organization can gain real-time insight and reduce threats and risks as a way to make better decisions and improve risk management, while protecting customer data and providing better service.

Intelligent risk management can prevent fraud. For example, in the financial industry, Big Data is deeply integrated into payment processing systems so that when suspicious activity occurs, detection models built using Big Data can be utilized to identify it in real time and stop potentially fraudulent money transactions in a timely manner. Intelligent risk management can also identify customer and employee churn. Customer-wise, Big Data is used to assess customer loyalty and determine which types of customers are most at risk, allowing companies to accelerate actions to reduce churn and prevent customer defections. For employees, HR-focused analysis of why employees leave can be used to improve employee treatment, reduce turnover and the time and money costs of training new employees. (Reciprocity 2021; American Express 2020)

6.5 Big Data Technologies Vs Secure Methodologies

Table 2 visualizes several big data technologies and the methods of making Big Data secure.

TABLE 2. Big Data Technologies Vs Secure Methodologies

Big Data Technologies	Secure Methodologies			
	Security model	Access control	Cryptographic techniques	Data encryption
Hadoop		✓		✓
Cloud Computing		✓	✓	
Spark		✓		✓
NoSQL	✓			

7 CONCLUSION

The research in this thesis shows that Big Data and information security have a close relationship. Widespread use of Big Data has increased the size and complexity of data, thus increasing the risk of information security problems including personal information leakage, virus invasion, spam, hacking attacks, and ransomware.

The development of Big Data has led to the development of more secure Big Data technologies as well. More and more Big Data technologies such as Hadoop, NoSQL, cloud computing, and more are being developed. One can learn these tools and use encryption and decryption techniques, access control, and other methods to make Big Data more secure.

In addition, by using Big Data combined with artificial intelligence to form predictive models, intelligent risk management, emergency response, and threat visualization, one can identify potential risks in advance, reduce security problems, and prevent similar situations from occurring in the future, thus protecting the security of information.

In my opinion, in the future, the application of Big Data technology will be more extensive, for example, smart homes, artificial intelligence, Internet of Things, etc., these areas will involve a large amount of data, so there will be more and more challenges in information security, the protection of data security will become a vital issue.

In order to solve these challenges, we can first keep learning more about Big Data technology. Different Big Data technologies have different characteristics. When using them, they can be used in conjunction with each other by combining their respective strengths to provide better solutions.

Secondly, using big data technologies, and artificial intelligence flexibly. Combining Big Data technology and artificial intelligence, machine learning algorithms are used to quickly identify abnormal behavior and potential threats and dispose of them in a timely manner, minimizing the damage caused by security issues while preventing similar problems in the future.

Thirdly, with the rapid growth of Big Data technology, the demand for related personnel is also increasing, but the problem of talent shortage is also becoming obvious. More related training institutions can be opened in the future, and universities can set up more professional courses or interest classes.

Overall, information security problems will be more serious in the future, and we can continue to study in depth how to use big data combined with specific situations to develop better technologies or systems and formulate better solutions, so as to continuously improve the level of information security.

REFERENCES

Alissa Irei 2023. Incident response automation: What it is and how it works. Internet publication. Techtarget. https://www.techtarget.com/searchsecurity/tip/Incident-response-automation-What-it-is-and-how-it-works. Accessed 02.04.2023.

Amer Elsad, JR Gumarin & Abigail Barr 2022. LockBit 2.0: How This RaaS Operates and How to Protect Against It. Internet publication. Unit42. Updated 09.06.2022. https://unit42.paloaltonetworks.com/lockbit-2-ransomware/. Accessed 05.03.2023.

American Express 2020. Benefits of Using Big Data in Risk Management. Internet publication. Updated 11.09.2020. https://www.americanexpress.com/en-ca/business/trends-and-insights/articles/benefits-of-using-big-data-in-risk-management/

Andrew Froehlich, Sharon Shea, & Ben Cole 2021. Cloud security. Internet publication. Techtarget. Updated 02.2021. https://www.techtarget.com/searchsecurity/definition/cloud-security. Accessed 10.04.2023.

Anil Papisetty 2020. Database security methodologies of SQL Server. Internet publication. Updated 03.08.2020. https://infohub.delltechnologies.com/p/database-security-methodologies-of-sql-server/. Accessed 27.04.2023.

Apache 2023. Hadoop Key Management Server (KMS) - Documentation Sets. Internet publication. Updated 15.03.2023. https://hadoop.apache.org/docs/stable/hadoop-kms/index.html#KeyProvider. Accessed 08.03.2023.

Apache 2023. HDFS Architecture. Internet publication. Updated 15.03.2023. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html. Accessed 19.02.2023.

Apache 2023. HDFS Permissions Guide. Internet publication. Updated 15.03.2023. https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsPermissionsGuide.html. Accessed 08.03.2023.

Apache 2023. MapReduce Tutorial. Internet publication. Updated 15.03.2023. https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html. Accessed 19.02.2023.

Ayush Singh Rawat 2022. 10 Advantages of Big Data. Internet publication. Analyticssteps. Updated 03.01.2022. https://www.analyticssteps.com/blogs/advantages-big-data. Accessed 10.04.2023.

Ayusharma0698 2022. Difference between SQL and NoSQL. Internet publication. Geeksforgeeks. Updated 15.11.2022. https://www.geeksforgeeks.org/difference-between-sql-and-nosql/. Accessed 19.02.2023.

Benoy Antony, Konstantin Boudnik, Cheryl Adams, Branky Shao, Cazen Lee, & Kai Sasaki 2016. Professional Hadoop,2-3. E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=4519258. Accessed 19.02.2023.

Benoy Antony, Konstantin Boudnik, Cheryl Adams, Branky Shao, Cazen Lee, & Kai Sasaki 2016, Professional Hadoop,3-4. E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=4519258. Accessed 19.02.2023.

Benoy Antony, Konstantin Boudnik, Cheryl Adams, Branky Shao, Cazen Lee, & Kai Sasaki 2016. Professional Hadoop,125. E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=4519258. Accessed 19.02.2023.

Benoy Antony, Konstantin Boudnik, Cheryl Adams, Branky Shao, Cazen Lee, & Kai Sasaki 2016. Professional Hadoop,129-130. E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=4519258. Accessed 19.02.2023.

Bitdefender s.a. What are Private Data Leaks? Internet publication. https://www.bitdefender.com/cyberpedia/what-are-private-data-leaks/. Accessed 05.03.2023.

Box s.a. What is cloud security? Internet publication. https://www.box.com/resources/what-is-cloud-security. Accessed 05.03.2023.

Casey McGuigan s.a. How to Stay Ahead of Cyberattacks Through Data Visualizations. Internet publication. Theceoviews. https://theceoviews.com/how-to-stay-ahead-of-cyberattacks-through-data-visualizations/Accessed 26.03.2023.

Chance Miller 2021. iCloud Calendar spam continues to impact users, despite Apple's multiple fixes. Internet publication. 9to5mac. Updated 21.06.2021. https://9to5mac.com/2021/06/21/icloud-calendar-spam-problems/. Accessed 05.03.2023.

Cisco s.a. What Is Information Security? Internet publication. https://www.cisco.com/c/en/us/products/security/what-is-information-security-infosec.html. Accessed 19.02.2023.

Cory Bohon 2022. How to limit AirDrop spam in iOS 16.2. Internet publication. TechRepublic. Updated 23.12.2022. https://www.techrepublic.com/article/how-to-limit-airdrop-spam-ios-16-2/. Accessed 05.03.2023.

Cynet s.a. Incident Response. Internet publication. https://www.cynet.com/incident-response/. Accessed 05.03.2023.

Cynthia Harvey 2018. Big Data Pros and Cons. Internet publication. Datamation. Updated 09.08.2018. https://www.datamation.com/big-data/big-data-pros-and-cons/. Accessed 04.03.2023.

Dan Mccreary & Ann Kelly 2014. Making Sense of NoSQL,233-234. Internet publication. http://www.bigdata.ir/wp-content/uploads/2016/08/5FB45AB6A5AEEC2E405B214983F9A04B.pdf. Accessed 19.02.2023.

Dan Mccreary & Ann Kelly 2014. Making Sense of NoSQL,235. Internet publication. http://www.bigdata.ir/wp-content/uploads/2016/08/5FB45AB6A5AEEC2E405B214983F9A04B.pdf. Accessed 19.02.2023.

Dan Mccreary & Ann Kelly 2014. Making Sense of NoSQL,235-236. Internet publication. http://www.bigdata.ir/wp-content/uploads/2016/08/5FB45AB6A5AEEC2E405B214983F9A04B.pdf. Accessed 19.02.2023.

Devashree Madhugiri 2022. Apache Spark Vs. Hadoop MapReduce – Top 7 Differences. Internet publication. Analyticsvidhya. Updated 22.06.2022. https://www.analyticsvidhya.com/blog/2022/06/apache-spark-vs-hadoop-mapreduce-top-7-differences/. Accessed 07.03.2023.

Dikshantmalidev 2020. MapReduce Architecture. Internet publication. Geeksforgeeks. Updated 10.09.2020. https://www.geeksforgeeks.org/mapreduce-architecture/. Accessed 19.02.2023.

Edgar Mondragón Tenorio 2021. Advantages and disadvantages of Big Data. Internet publication. Bbva. Updated 26.05.2021. https://www.bbva.ch/en/news/advantages-and-disadvantages-of-big-data/. Accessed 04.03.2023.

Editorial Team 2021. Advantages & Disadvantages of Big Data. Internet publication. Towards AI. Updated 05.08.2021. https://towardsai.net/p/l/advantages-disadvantages-of-big-data. Accessed 04.03.2023.

Faxiancunzai 2017. Hadoop-kms zongjie. Internet publication. CSDN. Updated 31.07.2017. https://blog.csdn.net/yunduanyou/article/details/76461223?spm=1001.2101.3001.6650.8&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7EESLANDING%7Edefault-8-76461223-blog-123094930.pc_relevant_landingrelevant&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7EESLANDING%7Edefault-8-76461223-blog-123094930.pc_relevant_landingrelevant&utm_relevant_index=11. Accessed 08.03.2023.

Frontegg s.a. What Is Role-Based Access Control (RBAC)? A Complete Guide. Internet publication. https://frontegg.com/guides/rbac. Accessed 09.03.2023.

Griffith University s.a. What Does Big Data Mean? Internet publication. FutureLearn. https://www.futurelearn.com/info/courses/applied-big-data-analytics/0/steps/52404. Accessed 02.03.2023.

Henrik Kärkkäinen & Tuomas Linnake 2022. Savonia-ammattikorkeakouluun tehty massiivinen tietoturvahyökkäys – kiristysohjelma lukinnut tietoja. Internet publication. Is. Updated 12.02.2022. https://www.is.fi/digitoday/tietoturva/art-2000008607041.html. Accessed 05.03.2023.

Henrik Kärkkäinen 2023. Suomessa myyty 1 300 vaarallisen takaportin sisältänyttä nettikameraa – myynnissä myös Prismoissa. Internet publication. Is. Updated 20.01.2023. https://www.is.fi/digitoday/tietoturva/art-2000009324472.html. Accessed 19.02.2023.

Hewlett Packard Enterprise s.a. Infrastructure Security. Internet publication. https://www.hpe.com/us/en/what-is/infrastructure-security.html. Accessed 05.03.2023.

IBM 2021. Discretionary access control (DAC). Internet publication. Updated 22.03.2021. https://www.ibm.com/docs/en/zos/2.2.0?topic=controls-discretionary-access-control-dac. Accessed 09.03.2023.

IBM 2021. Mandatory access control (MAC). Internet publication. Updated 03.03.2021. https://www.ibm.com/docs/en/zos/2.4.0?topic=environment-mandatory-access-control-mac. Accessed 09.03.2023.

IBM s.a. Big data analytics. Internet publication. https://www.ibm.com/analytics/big-data-analytics. Accessed 19.02.2023.

IBM s.a. Data Warehouse. Internet publication. https://www.ibm.com/uk-en/topics/data-warehouse. Accessed 27.02.2023.

IBM s.a. OLAP. Internet publication. https://www.ibm.com/uk-en/topics/olap. Accessed 27.02.2023.

Ilya Ganelin, Ema Orhian, Kai Sasaki & Brennon York 2016. Spark : Big Data Cluster Computing in Production, chapter 4.E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=4451522. Accessed 19.02.2023.

Informit 2014. Introduction to HBase, the NoSQL Database for Hadoop. Internet publication. Updated 27.10.2014. https://www.informit.com/articles/article.aspx?p=2253412. Accessed 07.03.2023.

InterviewBit 2022. Cloud Computing Architecture – Detailed Explanation. Internet publication. Updated 10.06.2022. https://www.interviewbit.com/blog/cloud-computing-architecture/. Accessed 19.02.2023.

Javatpoint s.a. Data Warehouse Tutorial. Internet publication. https://www.javatpoint.com/data-warehouse. Accessed 18.03.2023.

Javatpoints s.a. Advantages & Disadvantages of Big Data. Internet publication. https://www.javatpoint.com/advantages-and-disadvantages-of-big-data. Accessed 04.03.2023.

Joffrey L. Leevy, John Hancock, Taghi M. Khoshgoftaar & Jared M. Peterson 2021. An Easy-to-Classify Approach for the Bot-IoT Dataset. Internet publication. https://www.computer.org/csdl/proceedings-article/cogmi/2021/162100a172/1CxzT10zaAo. Accessed 27.04.2023.

Kaspersky s.a. CYBERTHREAT REAL-TIME MAP. Internet publication. https://cybermap.kaspersky.com. Accessed 03.04.2023.

Kathleen Richards 2021. Internet publication. Cryptography. Techtarget. Updated 09.2021. https://www.techtarget.com/searchsecurity/definition/cryptography. Accessed 10.04.2023.

Kevin Taylor-Sakyi 2016. Big data: Understanding Big Data. Internet publication. arXiv. Updated 15.01.2016. https://arxiv.org/abs/1601.04602. Accessed 19.02.2023.

L. Minh Dang, Md. Jalil Piran, Dongil Han, Kyungbok Min & Hyeonjoon Moon 2019. A Survey on Internet of Things and Cloud Computing for Healthcare. Internet publication. https://www.mdpi.com/2079-9292/8/7/768. Accessed 24.03.2023.

Malwarebytes s.a. What is spam? Internet publication. https://www.malwarebytes.com/spam. Accessed 05.03.2023.

Mark Lahn 2019. The 5 V's of Big Data: Velocity, Volume, Value, Variety, and Veracity. Internet publication. Servermania. Updated 06.09.2019. https://www.servermania.com/kb/articles/vs-of-big-data/. Accessed 04.03.2023.

Meri Remes & Keijo Salokangas 2022. Savonia-ammattikorkeakouluun tehty massiivinen tietoturvahyökkäys – kiristysohjelma lukinnut tietoja. Internet publication. Yle. Updated 04.02.2022. https://yle.fi/a/3-12302764. Accessed 05.03.2023.

Nick Cavalancia 2020. Vulnerability management explained. Internet publication. AT&T. Updated 02.07.2020. https://cybersecurity.att.com/blogs/security-essentials/vulnerability-management-explained. Accessed 10.04.2023.

Peter Loshin 2022. Application security. Internet publication. Techtarget. Updated 01.2022. https://www.techtarget.com/searchsoftwarequality/definition/application-security. Accessed 10.04.2023.

Popoola & Segun Isaiah 2017. RANSOMWARE: Most Recent Threat to Computer Network Security. Internet publication. Researchgate. Updated 02.2017. https://www.researchgate.net/publication/313905642_RANSOMWARE_Most_Recent_Threat_to_Computer_Network_Security. Accessed 19.02.2023.

Radu Raicea 2017. How Pretty Good Privacy works, and how you can use it for secure communication. Internet publication. Freecodecamp. Updated 08.10.2017. https://www.freecodecamp.org/news/how-does-pretty-good-privacy-work-3f5f75ecea97/. Accessed 09.03.2023.

Rahul Beakta 2015. Big Data And Hadoop: A Review Paper,14. Internet publication. Researchgate. Updated 01.2015. https://www.researchgate.net/publication/281403776_Big_Data_And_Hadoop_A_Review_Paper. Accessed 19.02.2023.

Rami Ali 2020. Predictive Modeling: Types, Benefits, and Algorithms. Internet publication. Netsuite. Updated 23.09.2020. https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml. Accessed 24.03.2023.

Reciprocity 2021. Applying Big Data to Risk Management. Internet publication. Updated 01.06.2021. https://reciprocity.com/blog/applying-big-data-to-risk-management/. Accessed 03.04.2023.

Ricardo Alejandro, Manzano Sanchez, Marzia Zaman, Nishith Goel, Kshirasagar Naik & Rohit Joshi 2022. Towards Developing a Robust Intrusion Detection Model Using Hadoop–Spark and Data Augmentation for IoT Networks. Internet publication. Proquest https://www.proquest.com/docview/2728531833/fulltextPDF/1843FA3F9BB44D94PQ/1?accountid=27296. Accessed 19.02.2023.

Rob Fry, Brooks Evans, & Jason Chan 2015. Introducing FIDO: Automated Security Incident Response. Internet publication. Netflix TechBlog. Updated 04.05.2015. https://netflixtechblog.com/introducing-fido-automated-security-incident-response-1961f34f7da3. Accessed 03.04.2023.

Shrey Mehrotra & Akash Grade 2019.Apache Spark Quick Start Guide: Quickly Learn the Art of Writing Efficient Big Data Applications with Apache Spark,5-8.E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=5675596. Accessed 19.02.2023.

Shreya Joshi 2022. 12 Reasons to Consider a Data Analyst Career Path. Internet publication. Analytixlabs. Updated 22.07.2022. https://www.analytixlabs.co.in/blog/data-analyst-career-path/. Accessed 19.02.2023.

Sudheesh Narayanan 2013. Securing Hadoop,78. E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=1561454. Accessed 19.02.2023.

Sunil Kumar & Dilip Agarwal 2018. Hacking Attacks, Methods, Techniques And Their Protection Measures,2253. Internet publication. Researchgate. Updated 05.2018. https://www.researchgate.net/profile/Sunil-Kumar-310/publication/324860675_Hacking_Attacks_Methods_Techniques_And_Their_Protection_Measures/links/5ae7ea5ca6fdcc03cd8dbf8f/Hacking-Attacks-Methods-Techniques-And-Their-Protection-Measures.pdf. Accessed 19.02.2023.

Surya Gutta 2020. Data Science: The 5 V’s of Big Data. Internet publication. Analytics Vidhya. Updated 04.05.2020. https://medium.com/analytics-vidhya/the-5-vs-of-big-data-2758bfcc51d. Accessed 04.03.2023.

The State of Security 2022. Global research: Security leaders’ priorities for cloud integrity, the talent. gap and the most urgent attack vectors. Internet publication. Splunk. Updated 12.04.2022. https://www.splunk.com/en_us/pdfs/gated/ebooks/state-of-security-2022.pdf. Accessed 19.02.2023.

Tom White s.a. Hadoop: The Definitive Guide. Internet publication. Oreill. https://www.oreilly.com/library/view/hadoop-the-definitive/9780596521974/ch04.html. Accessed 09.03.2023.

Vic Winkler, Graham Speake & Patrick Foxhoven 2011. Securing the Cloud : Cloud Computer Security Techniques and Tactics, 2.E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=686827. Accessed 19.02.2023.

Vic Winkler, Graham Speake & Patrick Foxhoven 2011. Securing the Cloud : Cloud Computer Security Techniques and Tactics, 133-134. E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=686827. Accessed 19.02.2023.

Vic Winkler, Graham Speake & Patrick Foxhoven 2011. Securing the Cloud : Cloud Computer Security Techniques and Tactics, 138-140. E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=686827. Accessed 19.02.2023.

Vic Winkler, Graham Speake & Patrick Foxhoven 2011. Securing the Cloud : Cloud Computer Security Techniques and Tactics, 144. E-book. https://ebookcentral-proquest-com.ezproxy.savonia.fi/lib/savoniafi/detail.action?docID=686827. Accessed 19.02.2023.

Walter Schneider 1989. Computer viruses: What they are, how they work, how they might get you, and how to control them in academic institutions, 334. Internet publication. Springer. Updated 03.1989. Accessed 19.02.2023.

Yitong Ji, Aixin Sun, Jie Zhang, & Chenliang Li 2022. A Critical Study on Data Leakage in Recommender System Offline Evaluation. Internet publication. arXiv. Updated 21.10.2022. https://arxiv.org/abs/2010.11060. Accessed 05.03.2023.

ZHU Zhenfang 2015. Study on Computer Trojan Horse Virus and Its Prevention, 95. Internet publication. International Journal of Engineering and Applied Sciences. Updated 08.2015. https://www.ijeas.org/download_data/IJEAS0208024.pdf. Accessed 19.02.2023.

APPENDIX 1: CREATIVE COMMONS LICENCES – IMAGE COPYRIGHT

Image copyright can be ensured by setting search criteria in the image search that allow the image to be used and edited. An image taken from a source can be edited, but you need to check with the copyright whether this can be done. For example, cropping an image or annotating an image is editing. If a finished map is used to mark off an area, it is regarded as editing.

This is how to check image copyright: If you search for an image using Google's Image Search tool, then use the Searching Tools and select Access. Select one of the following options: A derivative work means that you have edited the image in question.

Attribution (BY) The work may be copied, distributed, shown and performed in public and derivative works may be created, provided that the name of the author or copyright holder is duly mentioned.

Non-Commercial (NC) The work may be copied, distributed, shown and performed in public and derivative works may be created only when they are not used for commercial purposes.

No Derivative Works (ND) The work may be copied, distributed, shown and performed in public, but no derivative works may be created from it.

Share Alike (SA) Derivative works may only be distributed under the same license as the original work.

Indicating CC licenses in combinations

The licenses are written with the letter combination CC first. This is followed by a space and a dashed list of abbreviations for license terms. The first of the conditions is always BY, followed by a possible NC and followed by a possible third condition.

The licenses obtained by combining the terms are:

- Attribution (CC BY)
- Attribution – Share-alike (CC BY-SA)
- Attribution – No Derivative Works (CC BY-ND)
- Attribution – Non-Commercial (CC BY-NC)
- Attribution – Non-Commercial – Share-alike (CC BY-NC-SA)
- Attribution –Non-commercial – No Derivative Works (CC BY-NC-ND)

In addition, there is a special CC0 license, which allows the author to waive all rights to the work to the extent permitted by law. In Finland, the author cannot waive his moral rights, thus the name of the author / photographer must always be mentioned.

Example: FIGURE 1. A detail of the Colosseum in Rome (Laamanen 2015, CC BY-SA)

The image may be edited and shared, but the original creator must be mentioned. A modified image may only be shared under the same license as the original image.