Shibakali Gupta, Indradip Banerjee, Siddhartha Bhattacharyya (Eds.)
Big Data Security

De Gruyter Frontiers in Computational Intelligence

Edited by Siddhartha Bhattacharyya

Volume 3

Already published in the series

Big Data Security

Edited by Shibakali Gupta, Indradip Banerjee,
Siddhartha Bhattacharyya

Editors
Dr. Shibakali Gupta
Department of Computer Science & Engineering, University Institute of Technology
The University of Burdwan
Golapbag North
713104 Burdwan, West Bengal, India
skgupta.81@gmail.com

Dr. Indradip Banerjee
Department of Computer Science & Engineering, University Institute of Technology
The University of Burdwan
Golapbag North
713104 Burdwan, West Bengal, India
ibanerjee2001@gmail.com

Prof. (Dr.) Siddhartha Bhattacharyya
RCC Institute of Information Technology
Canal South Road, Beliaghata
700 015 Kolkata, India
dr.siddhartha.bhattacharyya@gmail.com

———

Dr. Shibakali Gupta would like to dedicate this book to his daughter, wife & parents.

Dr. Indradip Banerjee would like to dedicate this book to his son, wife & parents.

Prof. (Dr.) Siddhartha Bhattacharyya would like to dedicate this book to his parents Late Ajit Kumar Bhattacharyya and Late Hashi Bhattacharyya, his beloved wife Rashni, and his youngest sister's parents-in-laws Late Anil Banerjee and Late Sandhya Banerjee.

Preface

With the advent of a range of data-driven avenues and explosion of data, research in the field of big data has become an important thoroughfare. Big data produces exceptional amounts of data points, which give greater insights that determine sensational research, better business decisions, and greater value for customers. To accomplish these endings, establishments need to be able to handle the data while including measures for using sensitive private information efficiently and quickly, and thus the implementation of security issue creates a vigorous role. End-point devices create the main factors for observance of the big data. Processing, storage, and other necessary responsibilities have to be performed with the help of input data, which is generated by the end-points. Therefore, an association should make sure to use an authentic and valid end-point security. Due to large amounts of data generation, it is quite impossible to maintain regular checks by most of the establishments. Therefore, periodic observation and performing security checks can be utmost promising in real time. On the other hand, cloud-based storage has enabled data mining and collection. However, this big data and cloud storage incorporation have introduced concerns for data secrecy and security threats.

This volume intends to deliberate some of the latest research findings regarding the security issues and mechanisms for big data. The volume comprises seven well-versed chapters on the subject.

The introductory chapter provides a brief and concise overview of the subject matter with reference to the characteristics of big data, the inherent security concerns, and mechanisms for ensuring data integrity.

Chapter 2 deals with the motivation for this research that came from lack of practical applications of block chain technology, its history, and the principle of how it functions within the digital identity and importance of EDU certificate transparency and challenges in their sharing. In the theoretical part of the chapter, a comparison of the "classical" identity and digital identity is set out, which is described through examples of personal identity cards and e-citizen systems. Then, following the introduction into block chain technology and describing the method of achieving consensus and transaction logging, the principle of smart contracts is described, which provide the ability to enter code or even complete applications and put them into block chains, enabling automation of a multitude of processes. The chapter also explains common platforms through examples that are described as business models that use block chain as a platform for developing their processes based on digital identity.

Chapter 3 describes the anomaly detection procedure in cloud database metric. Each and every big data source or big database needs a security metric monitoring. The monitoring software collects various metrics with the help of custom codes, plugging, and so on. The chapter describes the approach of modifying the normal metric thresholding to anomaly detection.

With the tangible and exponential growth of big data in various sectors, every day-to-day activities like websites traversed, locations visited, movie timings, and others were stowed by various companies such as Google through Android cell phone. Even bank details are accessible by Google. In such situations, wherein a person's identity can be mentioned almost completely by just a small number of datasets, the security of those datasets is of huge importance especially in terms of situations where human manipulations are involved. Using social engineering to retrieve few sensitive information could lead to completely rip off a person's identity and his/her personal life. Chapter 4 deals with similar facts, that is, social engineering angle of hacking for big data along with other hacking methodologies that can be used for big data and how to secure the systems from the same. This chapter helps the users to visualize major vulnerabilities in data warehousing systems for big data along with an insight of major such hacking in recent past, which lead to disclosure of major private and sensitive data of millions of people.

Chapter 5 describes the information hiding technique as well as consumptions of this one in big data. Global communication has no bounds and more information is being exchanged over the public medium that serves an important role in the communication mode. The rapid growth in the usage of sensitive information exchange through the Internet or any public platform causes a major security concern in these days. More essentially, digital data has given an easy access to communication of its contents that can also be copied without any kind of degradation or loss. Therefore, the urgency of security during global communication is obviously quite tangible nowadays.

Some of the big data security Issues have been discussed in Chapter 6 with some solution mechanisms. Big data is a collection of huge sets of data of different categories, where it could be distinguished as structured and unstructured ways. As are revolutionizing to zeta bytes from Giga/Tera/Peta/Exabytes in this phase of computing, the threats have also increased in parallel. Big data analysis is flattering essential means for automatic determination of astuteness that is concerned in the recurrently stirring outline and secreted convention. This can facilitate companies to obtain an improved resolution, to envisage and recognize revolution, and to categorize new fangled prospects. Dissimilar procedure in support of big data analysis as well as numerical analysis, batch processing, machine learning, data mining, intelligent investigation, cloud computing, quantum computing, and data stream preparing become possibly the most important factor.

Chapter 7 summarizes the main contributions and findings of the previously discussed chapters and offers future research directions. A conclusion has also been derived out on possible scope of extension or future direction. In this book, several security issues have been addressed in big data domain.

The book is targeted to meet the academic and research interests of the big data community. It would come to use to students and faculty members involved in the disciplines of computer science, information science, and communication

engineering. The editors would be more than happy if the readers find it useful in exploring further ideas in this direction.

<div align="right">

Shibakali Gupta
Indradip Banerjee
Siddhartha Bhattacharyya

</div>

October 2019
Kolkata, India

Contents

List of Contributors

Ayan Mukherjee
Cognizant, Kolkata, India
mukherjeeayan16@gmail.com

Goran Fijacko
Algebra University College, Zagreb, Croatia
gfijacko@gmail.com

Goutam Sanyal
National Institute of Technology, Durgapur,
India
nitgsanyal@gmail.com

Indradip Banerjee
Department of Computer Science &
Engineering
University Institute of Technology,
The University of Burdwan
Burdwan, West Bengal, India
ibanerjee2001@gmail.com

Leo Mrsic
Algebra University College, Zagreb, Croatia
leo.mrsic@algebra.hr

Mislav Balkovic
Algebra University College, Zagreb, Croatia
mislav.balkovic@algebra.hr

Santanu Koley
Department of Computer Science and
Engineering
Budge Budge Institute of Technology,
Kolkata, India
santanukoley@yahoo.com

Shibakali Gupta
Department of Computer Science &
Engineering
University Institute of Technology,
The University of Burdwan
Burdwan, West Bengal, India
skgupta.81@gmail.com

Siddhartha Bhattacharyya
RCC Institute of Information Technology,
Kolkata, India
dr.siddhartha.bhattacharyya@gmail.com

Souvik Chowdhury
Oracle India, Bangalore, India
souvikcho@gmail.com

Srilekha Mukherjee
National Institute of Technology, Durgapur,
India
srilekha.mukherjee3@gmail.com

Shibakali Gupta, Indradip Banerjee, Siddhartha Bhattacharyya

1 Introduction

Security is one of the leading accomplishment of awareness in information technology and communication system. In the contemporary communication epoch, digital channels are used to communicate hypermedia content, which governs the field of arts, entertainment, education, commerce, research, and so on. The users of the field of the digital media technology are increasing massively, and they realized that data on web is an extremely important aspect of modern life.

Devising discoursed certain security issues, there exist some chief principles. Privacy principles specify that only sender and the receiver have a duty to be able to access the message from the web. No other unsanctioned creature can access this one. Authentication apparatuses help to launch the proof of identity. The authentication confirms that the origin of a digital message is correctly recognized. When the content of the message is altered after directing by the sender and before obtaining by the receiver, the uprightness of the message is lost. Access control regulates who should be able to admit the system and what. It has two areas: role and rule management.

The digital data content includes audio, video, and image media, which can be easily stored and manipulated. The superficial transmission and manipulation of digital content constitute an authentic threat to multimedia content engenderers and traders.

Big data is a term that is used to explain datasets that are enormous in size against normal database. Big data is becoming more and more popular each day. Big data generally consists of unstructured, semistructured, or structured datasets. Some algorithms as well as tools are used to process these data within the reasonable finite amount of time, but the main prominence is known on the unstructured data [1].

The characteristics of big data mainly depend on 4Vs (volume, velocity, variety, veracity) [2, 3]. Volume is a key characteristic of big data, which decides whether the information is a normal dataset or not, the size of raw data or the data generated is important because the time complexity, specifications cost which depend on it. Velocity is the speed with a direction, which means the throughput or the speed of the data processed. How fast the information can be generated in real time is to meet the requirements. Variety is important in this literature because it stands for the quality and the type of data required in order to process it successfully. Data can be text, audio, video, image, and so on. The quality of data on which the processing will be done is vital, because if the information is corrupted or stolen then anybody can't expect accurate result from it.

To resolve these potential threats, the awareness of "Information Hiding" has been weighed [4, 5]. The idiom Information Hiding is discussed to construct the information undetectable as well as keeping the survival of the information secret. According to the *Oxford English Dictionary* [6], the implication of information is the "formation or molding of the mind or character, training, instruction, teaching." This word is originated in the fourteenth century for English and some other European languages. The theories of cryptography [7] and watermarking [8] were also developed after the birth of the information concept. But elevating computational supremacy of those has been developed with the generation of modern-day cryptographic and watermarking algorithms.

The word "Security" is not identically synonymous what it was in 10 years back, because the research in capsizal engineering techniques has incremented the processing power and the most important race between the study in cryptanalysis [9] and watermarking detection [10]. To solve the above specified problems, the concepts of steganography [11] has been proposed by the researchers. Steganography diverges from cryptography. Cryptography refers to a secure communication, which transmutes the data into a concrete form and for that reason an eavesdropper can't understand it. Steganography techniques can endeavor to obnubilate the subsistence of the message itself, so that an observer or eavesdropper does not know that the information is present or not.

The term big data is used for large and complex dataset, which systematically analyzes and processes data easily with a lesser amount of time ensemble. The key responsibility of big data is data capturing and storage, searching of data through several behaviors, sharing and transfer of information, data analysis, querying like visualization, updating, and so on. Thus the security of information is very much important in this terminology. From these points of view, the big data security is very challenging in this literature.

In the last few decades, researchers, engineers, and scientists have developed new models, techniques, and algorithms for the generation of robust security system and better analysis principle. Nowadays, the researchers used different methodologies for achieving better performance as well as improving the privacy of the hidden information. This book investigates the current state-of-the-art big data security systems.

There are different types of information in today's world, which are in the form of Text information, Digital Image or Video Frames related information and information of Audio signal additionally. This book aims at contributing toward the understanding of big data security in the form of Text and Digital Images through various security principles which addresses both the theoretical parts and practical observations. In this book, a throughout mathematical restorative has been carried out for achieving better security models.

The book has been organized into eight chapters. Following is a brief description of each chapter:

Chapter 2: Digital identity protection using blockchain for academic qualification certificates

This chapter deals with the motivation for this research came from lack of practical applications of block chain technology, its history, and the principle of how it functions within the digital identity and importance of EDU certificate transparency and challenges in their sharing. In the theoretical part of the chapter, a comparison of the "classical" identity and digital identity is set out, which is described through examples of personal identity cards and e-citizen systems. Then, following the introduction into block chain technology and describing the method of achieving consensus and transaction logging, the principle of smart contracts is described, which provides the ability to enter code or even complete applications and put them into block chains, enabling automation of a multitude of processes. This chapter explains common platforms through examples describing business models that use block chain as a platform for developing their processes based on digital identity. Also, traditional models with those based on smart deals have been compared. Through examples of cancelation or delays in air travel, voting, music industry, and tracking of personal health records, it was established that how existing models are actually sluggish, ineffective, and prone to manipulation, and through examples of block chain implementation, they showed that these systems functioned faster, more transparent, and most importantly, safer. The application of technology in several industries, from the Fintech industry to the insurance and real estate industry, is also described in this chapter. Concepts and test solutions are described, which are slowly implemented in the production phase and show excellent results. For this reason, we believe that similar solutions will implement increasing adoption of block chain technology globally. In the last, practical part of the chapter, a survey of existing solutions that offer creation of its own block chain and a multichain platform was selected. By having easy to apply and understand guidelines, it is easier for wider audience to accept and use/reuse sometimes complex digital concepts as part of their solutions and business processes.

Chapter 3: Anomaly detection in cloud big database metric

This chapter describes the anomaly detection procedure in cloud database metric. Each and every big data source or big database needs a security metric monitoring. The monitoring software collects various metrics with the help of custom codes, plugging, and so on. The chapter describes the approach of modifying the normal metric thresholding to anomaly detection. In this concept, system administration

possesses a common problem to deal with some intelligent alarm method, which can produce predictive warnings, that is, the system can detect any anomalies or problems before it occurs. The novel concept detects all the anomalies by analyzing previous metric data continuously. The chapter also deals with the power exponential moving average and exponential moving standard deviation method to produce an effective solution. The work has been tested on CPU utilization and memory utilization of big database servers, which reflects the real-time quality of the solution.

Chapter 4: Use of big data in hacking and social engineering

With the tangible and exponential growth of big data in various sectors, every day-to-day activities like websites traversed, locations visited, movie timings, and so on were stowed by various companies such as Google through Android cell phone. Even bank details are accessible by Google.

In such situation, wherein a person's identity can be mentioned almost completely by just few datasets, the security of those datasets is of huge importance especially in terms of situations where human manipulations are involved. Using social engineering to retrieve few sensitive information could lead to completely rip off a person's identity and his personal life.

This chapter deals with similar facts, that is, social engineering angle of hacking for big data along with other hacking methodologies that can be used for big data and how to secure the systems from the same. This chapter helps users to visualize major vulnerabilities in data warehousing systems for big data along with an insight of such major hacking in recent past, which lead to disclosure of major private and sensitive data of millions of people. The insight provided in this chapter will help single users and corporates to visualize how their data are at stake and what precautions they can take to secure them, let it be phishing type of social engineering attack or Scareware type of attacks.

Chapter 5: Steganography, the widely used name for data hiding

This chapter describes the information hiding technique as well as consumptions of this one in big data. Global communication has no bounds and more information is being exchanged over the public medium, which serves an important role in the communication mode. The rapid growth in the usage of sensitive information exchange through the Internet or any public platform causes a major security concern

these days. More essentially, digital data has given an easy access to communication of its content that can also be copied without any kind of degradation or loss. Therefore, the urgency of security during global communication is obviously quite tangible nowadays. Without the communication medium, the field of technology seems to downfall. But appallingly, these communications often turn out to be fatal in terms of preserving the sensitivity of vulnerable data. Unwanted sources hamper the privacy of the communication and may even annoyance with such data. The importance of security is thus gradually increasing in terms of all aspects of protecting the privacy of sensitive data. Various concepts of data hiding are hence into much progress. Cryptography is one such concept, the others being watermarking, and so on. But to protect the complete data content with some seamlessness, this chapter incorporates concepts of steganography. The realm of steganography ratifies the stated fact to safeguard the privacy of data. Unlike cryptography, steganography brings forth various techniques that strive to hide the existence of any hidden information along with keeping it encrypted. On the other hand, any apparently visible encrypted information is definitely more likely to captivate the interest of some hackers and crackers. Therefore, precisely saying, cryptography is a practice of shielding the very contents of the cryptic messages alone. On the other hand, steganography is seriously bothered with camouflaging the fact that some confidential information is being sent, along with concealing the very contents of the message. Hence, the data hiding in the seemingly unimportant cover medium is perpetuated. The field of big data is quite into fame these days as they deal with complex and large datasets. Steganographic methodologies may be used for the purpose of enhancing security of big data since they also find ways of doing so.

Chapter 6: Big data security issues with challenges and solutions

Some of the big data security issues have been discussed in this chapter with some solution mechanism. Big data is a collection of huge sets of data of different categories, where it could be distinguished as structured and unstructured ways. As are revolutionizing to zeta bytes from Giga/Tera/Peta/Exabytes in this phase of computing, the threats have also increased in parallel. Big data analysis is flattering essential means for automatic determination of astuteness that is concerned in the recurrently stirring outline and secreted convention. This can facilitate companies to obtain an improved resolution, to envisage and recognize revolutionize, and to categorize newfangled prospects. Dissimilar procedure in support of big data analysis as well as numerical analysis, batch processing, machine learning, data mining, intelligent investigation, cloud computing, quantum computing, and data stream preparing become possibly the most important factor. There is a gigantic open door for the big

data industry in addition to plenty of possibilities for research and enhancement. Besides big organizations, cost reduction is the criterion for the use of small- and medium-sized organizations too, thus increasing the security threat. Checking of the streaming data once is not the solution as security breaches cannot be understood. The data stack up within the clouds is not the only preference as big data technology is available for dispensation of both structured and unstructured data. Nowadays an enormous quantity of data is provoked by mobile phones (Smartphone) of equally the symphony form. Big data architecture is comprehend among the mobile cloud designed for supreme consumption by means. The best ever implementation is able to be conked out realistic for the use of a novel data-centric architecture of MapReduce technology, while HDFS also acts immense liability in using data with divergent arrangement. As time approaches the level of information and data engendered from different sources, enhanced and faster execution is the claim for the same. Here in this chapter the aim is to find out big data security vulnerable and also find out the best possible solutions for them. Considering this attempt will dislodge a stride forward along the way to an improved evolution in secure propinquity to opportunity.

Chapter 7: Conclusions

This chapter summarizes the main contributions and findings of the previously discussed chapters and offers future research directions. A conclusion has also been derived out on the possible scope of extension or future direction. In this book, several security issues have been addressed in big data domain. The book covers a wide area of big data security as well as steganography and points to a fairly large number of ideas, where the concepts of this book may be improvised. Design of numerous big data security concept through steganography has been discussed, which can meet different requirements like robustness, security, embedding capacity, and imperceptibility. Experimental studies are carried out to compare the performance of these developments. The comparative study of each method along with the existing method is also established.

References

[1]   Snijders, C., Matzat, U., & Reips, U.-D. "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science, (2012), 7(1)
[2]   Martin, Hilbert. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review". martinhilbert.net. Retrieved 7 October 2015.
[3]   DT&SC 7-3: What is Big Data?. YouTube. 12 August 2015.
[4]   Cheddad, Abbas., Condell, Joan., Curran, Kevin., & Kevitt, Paul Mc. Digital image steganography: Survey and analysis of current methods Signal Processing 90, 2010,pp. 727–752.

[5]   Capurro, Rafael., & Hjørland, Birger. (2003). The concept of information. Annual review of information science and technology (s. 343–411). Medford, N.J.: Information Today. A version retrieved November 6, 2011.

[6]   Anthony Reading. Meaningful Information: The Bridge Between Biology, Brain, and Behavior. Originally published: January 1, 2011.

[7]   Liu, Shuiyin., Hong, Yi., & Viterbo, E. "Unshared secret key cryptography". IEEE Transactions on Wireless Communications, 2014, 13(12), 6670–6683.

[8]   Hsu, Fu-Hau., Min-Hao, Wu., & WANG, Shiuh-Jeng., "Dual-watermarking by QR-code Applications in Image Processing", 2012 9th International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic and Trusted Computing.

[9]   Kaminsky, A., Kurdziel, M., & Radziszowski, S."An overview of cryptanalysis research for the advanced encryption standard", Military Communications Conference, 2010 – MILCOM 2010, San Jose, CA, Oct.31 2010-Nov.3 2010, 1310–1316, ISSN: 2155-7578

[10] Bian, Yong., & Liang, S. "Locally optimal detection of image watermarks in the wavelet domain using Bessel K Form distribution". IEEE Transactions on Image Processing, 2013, 22 (6), 2372–2384.

[11] EL-Emam, N.N. "Hiding a large amount of data with high security using steganography algorithm," Journal of Computer Science, 3(4), 223–232, April 2007.

Leo Mrsic, Goran Fijacko and Mislav Balkovic

2 Digital identity protection using blockchain for academic qualification certificates

Abstract: Although it was always an important issue, digital era increases the importance of both questions "what identity is" and "why is important to manage and protect one." There are many views and definitions of digital identity in the literature; however, this chapter explains the identity related to the identification of an individual, his/her qualification, and his/her status in society. Using modern approach, this chapter focuses on the academic qualification of an individual where blockchain is presented as an efficient concept for publishing, storing, and verifying educational certificates/diplomas. Motivation for this research came from lack of practical applications of blockchain and importance of EDU certificate transparency and challenges in their sharing (policy issues, national standard, etc.). By having easy to apply and understand guidelines, it is easier for wider audience to accept and use/reuse sometimes complex digital concepts as part of their solutions and business processes. As part of institution research lab, explained approach and proof of concept solution was developed in Algebra University College.

Keywords: digital identity, blockchain, smart contracts, academic qualification certificates, certificate Mrsic, Fijacko and Balkovic

2.1 Introduction

If we talk about a classical identity of an individual, we can think of personal identity card, a birth certificate, a driving license but also a university certificate/diploma or other EDU certificate. In terms of the digital identity of an individual, we can talk about an e-personal ID card, e-birth certificate, e-homepages, e-driver's license or e-diplomas [1]. The e-name tag is electronic, which means these documents also have a digital component. This digital component can be, for example, an electronic data carrier (chip) that stores certain data or certificates that are readily uploaded to the computer by a reader if needed. The data being displayed are centralized and are guaranteed by and responsible to the institution issuing them, where the data is stored [2].

Digital identity does not necessarily have to be a physical document or document [3]. It also includes our email addresses as well as various user accounts and

Leo Mrsic, Goran Fijacko and Mislav Balkovic, Algebra University College, Croatia, Europe.

profiles on the Internet, such as an eCitizen account, Facebook profile, and email (Figure 2.1).
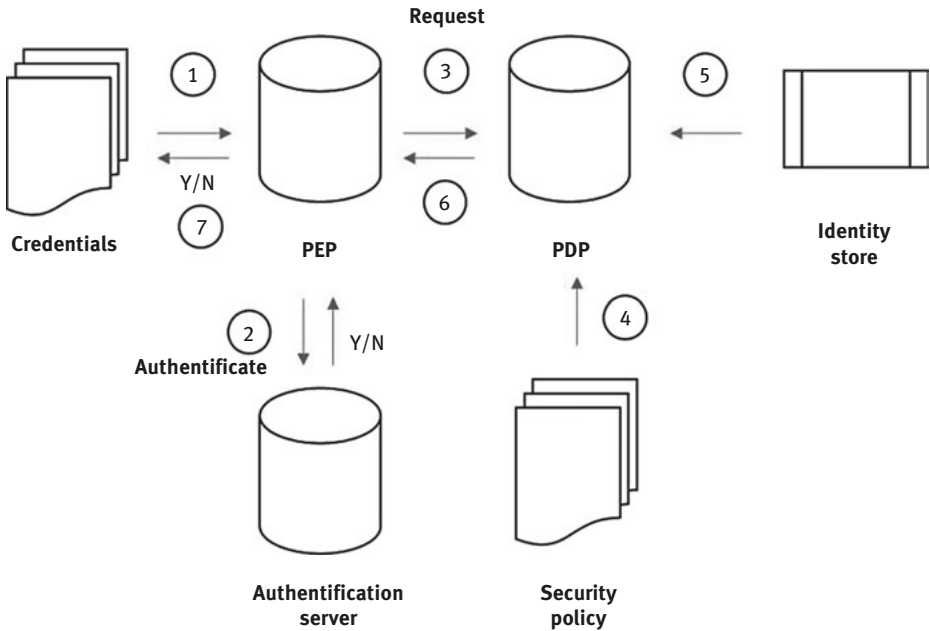
Figure 2.1: Authentication process (adopted from Phil Windley, Digital identity, p. 37).

Blockchain is one of the disruptive technologies that are often called technology, which will change the world and enable a new revolution. Blockchain represents a decentralized database that is publicly available to everyone via the Internet. Take for example databases and registers owned by the state and its institutions such as ministries, banks, and mobile operators, and all listed registers and data are published publicly by placing them in the blockchain. It allows us to access all data concerning our own, with authorization and access to the Internet. Likewise, we can present the same information to the other side as soon as we need to prove their identity, valid information or some information.

2.2 Smart contracts

With the help of smart contracts in the blockchain you can enter a program code or even entire applications. Smart deals define the relationships and behavior of two or more sides in the blockchain that have some cryptocurrency or some other information

or value. With this type of blockchain technology, there is a possibility that in the future there will be no standard services of attorneys, commercial courts, public notaries, and the like. A good part of the services they currently offer is likely to be easily replaced by smart deals because the relationship between the users and the providers of the above-mentioned services can be precisely defined through the code and entered into smart deals that are later realized by meeting all the conditions within the transaction. Realization of the services is automated and extremely fast, which in today's form is not. The use of smart deals allows the exclusion of a whole range of mediators in different processes and thus enables faster and easier to perform activities, private, and more importantly, business. They can be used, for example, in the following:

- **Under insurance**: If authorized agents in the blockchain register that the conditions for the payment of the insurance are met, the payment will be automatically made.
- **Medical insurance**: If a doctor finds that a patient is ill and unable to complete his/her business obligations, blockchain inserts that information, and the patient automatically starts to pay the sickness benefit.
- **Pension insurance**: If an authorized person or a state body certifies that a person has fulfilled the retirement conditions, a person will automatically be paid a pension.
- **Audio and video industry**: If a user pays for viewing or listening to a certain material, he/she automatically gets access to and rights to the purchased material.
- **Gambling industry**: The user who makes a bet is paid into the account of a smart contract. After the event is complete, the authorized party registers the data on the winner in the blockchain and those who successfully hit the results automatically receive payments.

2.3 Digital identity protection

Identity is very valuable to us, and not to institutions; we are not behaving accordingly [4]. The lack of awareness and education about importance of identity protection in digital world, powered with centralization of databases that store data about identities in general, represents unavoidable weakening trend that undermine the systematic value of our personal data. Centralized systems are a good booty for attackers with bad intentions because, if they break into the system, they can easily steal (copy) large amounts of data stored in that system. We have witnessed a lot of attacks on centralized systems, not small business systems, but big and globally influential companies such as Yahoo, eBay, Adobe, JP Morgan Chase and Sony.

Blockchain technology offers the solution to this problem that is becoming more and more constant due to constant needs, increased demand, and the use of digital identity. But, as we mentioned earlier, this is a new technology and is just in

the early stages of the project and we are still investigating all the possibilities and the application of this technology [5].

With the need to prove our identity, we meet each day and in different places: at work, in a bank, in a shop, in travel, in state institutions, and in many different places [6].

Currently, there are many new and prospective projects and young companies dealing with this problem and are trying to find their place in the market. In this part, we will mention some of them and more specifically explain their business models.

2.3.1 Civic

Civic is a company that develops an identification system that allows users to selectively share identifying information with companies. Their platform has a mobile application, where users enter their personal information and then store them in encrypted format. The company's goal is to establish partnerships with state governments and banks, that is, all those who can validate user identity data, and then leave a verification stamp in blockchain. The system encrypts the hash of all verified data and stores it in the blockchain and deletes all personal information of the user from their own servers.

As the company has written in its White Paper, the Civic Ecosystem is designed to encourage the participation of trusted authentication bodies called "validators." "Validators" can be the aforementioned state governments, banks, various financial institutions, and others. As Civic currently validates user identity information through its application, "validators" have the ability to verify the identity of an individual or a company that is "user" of the application. They then affix the certificate and place it in a blockchain in the form of a record known as attestation. This "verification" is actually a user's hash of personal information. Parties known as 'service providers wanting to verify the same user identity data should no longer be able to independently verify that information but rather use the verified information valid for those validators of that information. The goal is to remain a "ruler" of your identity and to have full control over personal information so that it must give prior consent to each transaction of information about its identity between the validator and the service provider. By smart deals, validators have the ability to sell their approvals to service providers, but also to service providers to see at what prices different validators offer their approvals. Each validator can declare the price it is willing to sell personal user information. After the user, validator and provider confirm the transactions through the smart deal system, the service provider pays the validator the required amount in the form of CVC tokens (utility tokens that support decentralized identity ecosystem supported by Civic's model which allow on-demand, secure and lower cost access to identity verification via blockchain). After that, a clever contract will allocate CVC tokens and the user will get their share of the participation. The user can use their tokens to purchase products and services on

the Civic platform. As we mentioned, the user is the one who is responsible for their data and stores them on some of their personal devices using the Civic app, and it is also recommended to back up a personal account on the cloud system. Since user identity data is not centralized, that is, not on Civic servers, there is no possibility of massive identity theft since the data of each user is actually on their devices and that data will be stolen, so it is necessary to break it into each device separately. This information largely helps to suppress the black market for personal information, for example, Black credit card market is quite widespread because transactions can only be done by knowing these data without the knowledge of the user. If a credit card number needs to go through the blockchain mechanism of proofing where the user's consent for each transaction should be, then the black market of such data slowly loses its meaning and value (Figure 2.2) [7].
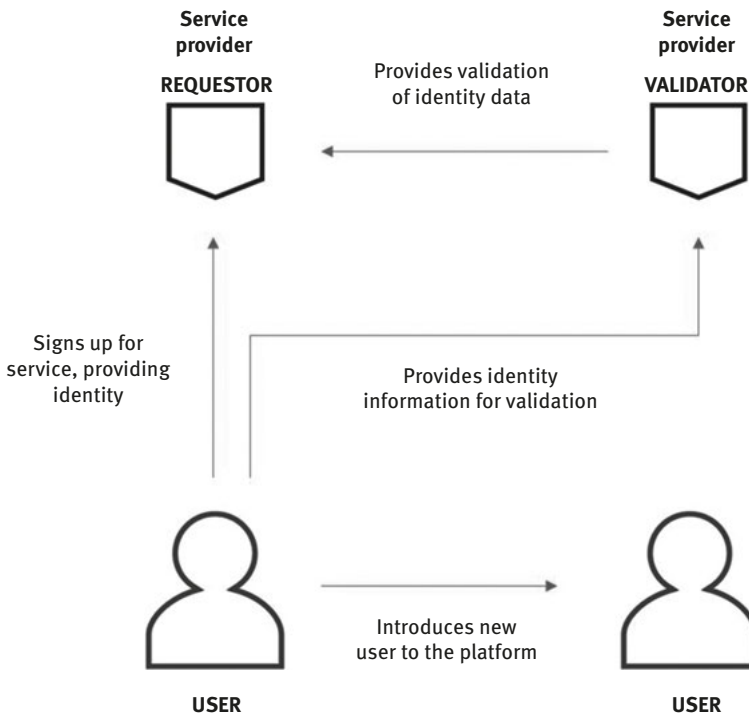

Figure 2.2: Civic concept.

2.3.2 HYPR

HYPR is a young company, founded in 2014. Their business model is based on merging biometric identification methods and blockchain technology. Biometric identification can replace a classic identification with a username and password, which is faster and

safer. Biometrics can recognize different parts of the human body such as palm geometry, fingerprint, eye iris, scent, face, and many long physiological elements unique to the individual. Biometrics is a very good way of verifying an individual's identity because it is very difficult or impossible to forge it.

HYPR therefore offers a password-free authentication platform with biometric encryption. The company does not deal with the development and production of identification devices, but develops a distributed security system. As mentioned earlier, every digital data can be used to insert some of the cryptographic algorithms and get their hash. This hash can be used to validate these digital data without the need for a validator to have a copy of that data. For example, we read our finger on a fingerprint reader on a mobile phone, and a company that has access to the hash of our fingerprint in digital form can confirm our identity, without the possibility of being false as we do. Digital print is just a part of the offering that is offered. HYPR supports many types of biometric data, from simple authentication algorithms to face and speech algorithms to much more complex algorithms such as keyboard typing, rhythm writing on mobile devices, or the way we walk. With blockchain and data decentralization, authentication becomes much faster and simpler. Each user is responsible for their biometric data, such as on his mobile device. This avoids massive data theft, while individual theft may still be possible if the user is not careful enough to protect their data and devices. Such a system based on blockchain technology is resistant to denial of service (DoS), which is a better centralized system. DoS attacks are attacks on some computer service in order to disable its use. In this case, instead of attacking a single server used to authenticate data, DoS attackers should identify and attack all blockchain nodes in that system. The company emphasizes that protecting against DoS attacks is equally important and the interoperability of business processes. There is currently no possibility of authentication between two different corporate entities such as a bank and an insurance company. Each company has a different identity database and they are not interoperable. Using blockchain technology, we can have an interoperable distributed mainstream identity book between multiple entities without the need for complex and expensive infrastructure. Thus, the insurance company can prove our identity to the bank through biometric data [8–10].

2.3.3 Blockverify

The problem of proof of identity does not only appear in people. It may also be present in various products such as medicines, luxury products, diamonds, electronics, music, and software. These products are often counterfeit, causing damage to manufacturers in billions of dollars.

People behind the Blockverify project want to reduce the number of counterfeit products on the market by preventing duplicate appearances. Different companies

from different industries can register and track their products using Blockverify and blockchain technology.

The company believes that improvement in counterfeit products can only be achieved by using decentralized, scalable, and safe solution attacks. Blockverify has its own private blockchain, but it also uses Bitcoin's blockchain to record important changes in its chain. Their chain is highly scalable and transparent so that each manufactured product can enter into it as an asset. After that, each of these assets will be added to the blockchain and assigned a unique hash. Anyone with that hash can access blockchain and check whether the product is valid or not. The primary goal of the company is to address the problem of counterfeit medicines, which is first on the scale of counterfeit products, but also one of the more dangerous counterfeit products because it directly affects people's health and causes millions of deaths per year. Another problem that a company wants to solve is the problem of verification of ownership. Thanks to blockchain technology, ownership changes can be easily recorded permanently. By this mode, individuals are prevented from making duplicate records and unauthorized changes.

2.4 Digital identity and blockchain

The main features of blockchain are transparency and decentralization, which today's systems cannot boast. Digital identity combined with blockchain technology will enable people to perform tasks that are faster, simpler, and safer, including proof of identity, facts, status, and data. Incredibly, the fact is that searching for new employees, checking candidate data, and job application itself could be a process that would take just a couple of mouse clicks on the computer, with the utmost certainty of the data being obtained [11]. But blockchain is just offering it. By placing all the information on our identity on it, with cryptography that makes the whole thing safe and transparent and always accessible through the Internet, we spend all the time spent on proving identity, data, facts, and state of affairs on the most important things. Imagine that we can also enclose three cryptographic keys with the application for business, so that the employer can easily check with the absolute certainty that we have actually completed the college we have stated in his CV, whether we are unhappy and whether we are at all a person who claims to be [12]. This process would take about a few minutes, while the same process lasts for several days, if not weeks, as the data verification is done by writing queries in each of these systems from which data comes [13, 14].

Blockchain got its name by the way it stores the transaction data that is happening. It stores them in blocks that together form a chain.

By increasing the number of transactions being made, the size of the chain in which they arise increases (Figure 2.3). The blocks record the sequence and time of
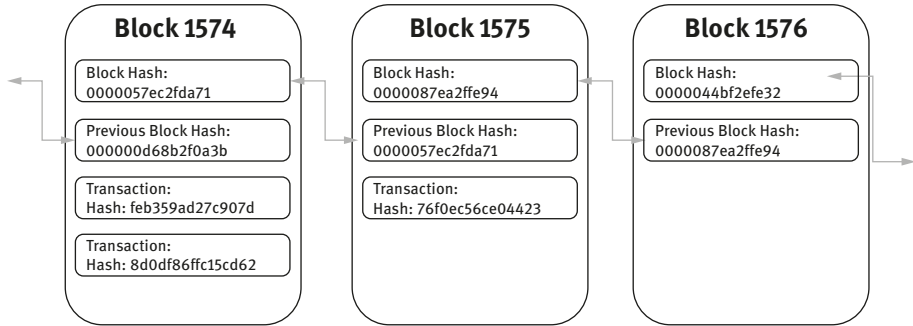
Figure 2.3: Showing transactions stored in blocks that connect each other to a chain
(Source: Gupta, M., Blockchain for dummies, 2nd IBM Limited Edition, 2018, p. 14th).

the transactions that are then recorded in the network chain according to certain security rules agreed between the participants. Each block contains a hash, that is, a digital imprint or unique identifier, then time-tagged valid transactions and hash of the previous block. Hash of the previous block mathematically links the blocks to the chain and prevents any change of data and information in the previous blocks or inserting new blocks between the existing ones. Thus, each of the following blocks increases the security of the entire chain and reduces the already small chance of manipulation and change of value or data in the chain.

There are several types of blockchains. In this chapter, we will mention the two most common types:

A public blockchain, such as Bitcoin blockchain (the first and most known cryptovalue based on this technology), is a large distributed network that runs with the release of a native token. The public blockchain is visible and open to everyone to use at all levels. An open code is maintained by the developer community.

Private blockchain is smaller in volume and usually does not run with token issuance. Membership in this type of blockchain is highly controlled and is often used by organizations that have confidential members or traded with confidential information.

All types of blockchains use cryptography to enable each participant to use the network in a safe manner, and most importantly, without the need for a central authority that applies the rules. Because of this, blockchain is considered revolutionary because it is the first way to gain confidence in sending and writing digital data.

An example, my name text, Goran, passing through the SHA-256 algorithm gives the result

dbe08c149b95e2b97bfcfc4b593652adbf8586c6759bdff47b533cb4451287fb

The word Goran will always result in an identical hash value. Adding any character or letter at the input changes the complete hash appearance, but of course, the mentioned 32-character length always remains the same. An example, word Gordan, gives the result

dbe08c149b95e2b97bfcfc4b593652adbf8586c6759bdff47b533cb4451287fb

In addition to the mentioned blocks and chains that are interconnected, there is another very important segment, which is a network. The network consists of nodes and full nodes. The device that connects and uses a blockchain network becomes a node, but if this device becomes a complete node, it must retrieve a complete record of all transactions from the very beginning of the creation of that chain and adhere to the security rules that define the chain. A complete node can lead anyone and anywhere, only the computer and the Internet are needed. But that's not so simple as it sounds.

Many people mix Bitcoin and Blockchain concepts or misuse them. Those are two different things. Blockchain technology was introduced in 2008, but it was only one year later launched in the form of cryptocurrency Bitcoin. Bitcoin is therefore a cryptocurrency that has its blockchain. This blockchain is a protocol that enables secure transmission and monitoring of cryptocurrency Bitcoin, all from the emergence of its first block (genesis block) and the first transaction. Bitcoin is designed solely as a criterion of vision that one day completely replaces fiat (paper) money and crushes the money transfer barriers that are present today. Through the years that passed, the community found that blockchain is more powerful than it originally thought, so if Bitcoin as a cryptocurrency does not live globally in everyday life, it will leave behind a revolutionary invention that potentially can change the technological world we are currently familiar with.

Blockchain through its mechanism of consensus eliminates the central authorities that we know today and which are based on today's technology.

2.5 Platform: issuing EDU certificate

MultiChain was selected as the platform for creating an application concept for entering, issuing, and verifying educational certificates. MultiChain is an open-source platform that allows you to create or block your own blockchain, and manage its capabilities. It is optimized for creating licensed chains (permissioned blockchains). MultiChain is compatible with Linux, Windows, and Mac operating systems. Currently optimized for Linux operating systems, here referred to as Linux's 64-bit Ubuntu 18.04.1 operating system with virtual machine on a single physical server using Oracle VM VirtualBox.

Virtualization of operating systems enables us to have multiple operating systems on one server, workstation, or computer, and we use them at the same time. All operating systems share computer resources. The number of virtual machines is unlimited, that is, it depends on the amount of disk space and the memory of the computer hosted on.

In the Oracle VM VirtualBox workstation there is a File, Machine, Help, icons for the most important activities on virtual machines (New, Settings, Discard, Show), and below them there is a window where show installed virtual machines, including Ubuntu's virtual machine, with assigned 100 GB of disk space and 3 GB of work memory. Before installing the operating system, it was necessary to create a virtual disk of the specified size that the machine would use.

The complete work console and virtual machine installation is very intuitive and simple. VirtualBox offers the ability of detachable virtual machine start-ups, enabling the entire process to run without open windows and a graphical user interface.

After installing and configuring a virtual machine, we must download and install the MultiChain application. Download, install, and all other actions within MultiChain are performed through the Terminal of Ubuntu Operating System Interface. The following commands are used to download and install:

wget https://www.multichain.com/download/multichain-1.0.6.tar.gz
tar -xvzf multichain-1.0.6.tar.gz
cd multichain-1.0.6
mv multichaind multichain-cli multichain-util /usr/local/bin

The last command line transfers the most important files to the bin folder for easier calling through the commands in the next steps.

After installing the MultiChain application, the first step is to create your own chain. Since the goal of the application is to enter, issue, and validate educational certificates for the purpose of this project, we called the chain BlockchainCertificate. This is done by performing the following function:

multichain-util create BlockchainCertificate

Using this command, we create the chain of this name with the default settings. After that, you must launch the created chain using the following command:

multichaind BlockchainCertificate -daemon

The chain was launched and its first block (genesis block) was created. After launch, the newly created chain gets its IP address and port through which it can be accessed from another device. The device on which the chain is created becomes

the first node, and each subsequent computer that connects to that chain over its IP address and the default port receives the complete chain data and also becomes a node. For the purpose of this chapter, only one node has been used, but in production it is not recommended to use only one node, of course, for safety reasons mentioned earlier in the work.

If the other computer joins this chain, it must also have the installed MultiChain application and must run the command:

multichaind BlockchainDiploma@[ip-adress]:[port]

After merging other computers/nodes, the first node only has the authority to assign certain rights, such as read and write rights, to other nodes.

Among other things, MultiChain has the ability to store data in a blockchain using the so-called stream. With storage, it also offers the ability to extract data. This functionality is most important for the concept of the application shown here. So, at the main node you need to create a new stream, which we will call certificate/diplomas in this example. The above statement is executed:

create stream certificate false

The false statement in the command means that only those explicitly licensed addresses can be written in that stream. Since in this example we have only one node that created this stream, it is not necessary to assign special rights. If there is a second node and some other address from which you want to write something on that same stream, you need to assign the rights for each address from the first node by a special grant command.

The next step is to store the data in the created stream certificate/diploma. Data is stored in hexadecimal form. In this example, we will store name and last name and OIBID (personal identification number) with the command publish certificate key1

476f72616e2046696a61636b6f2036383831363937343935035
Hexadecimal number

After issuing a command, we can obtain data record from stream using the simple query. The following command gives us all the information recorded in the stream certificate/diploma (Figure 2.4) liststreamkeys certificate

2.5.1 Application modules

This type of application is intended for private blockchain. This means that each educational institution should have its own stream that only the people in the

Figure 2.4: Displaying a textual (CLI) interface where chain creation, chain startup, and creation of a graduate stream are shown.

institution have the authority to store the certificate. All streams are stored in the main book that is distributed to all nodes, that is, educational institutions in this example. The more nodes in the chain, the better, because the chain becomes ever stronger and safer.

The application consists of three modules:

1. Module for certificate/diploma input
2. Certificate/diploma check module
3. Certificate/diploma print module

The first module is for entering a certificate/diploma. It switches the entered data into a hexadecimal form and stores them in the chain and returns the transaction ID (txid) back. Transaction ID is a private key that is awarded to a graduate student because it can be used to check the certificate/diploma data in the chain.
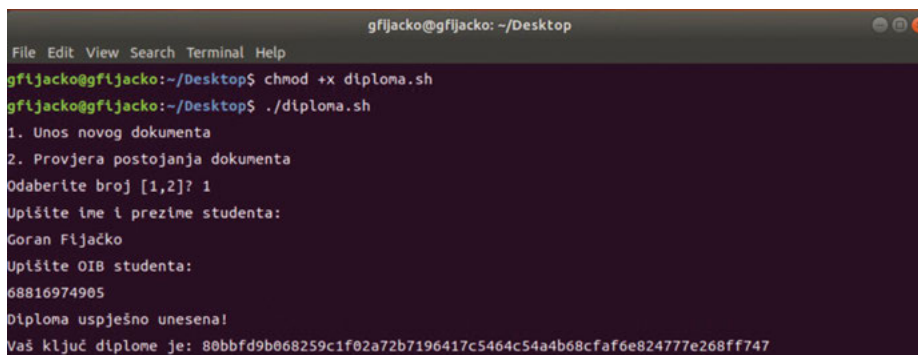
The Certificate/Diploma Check Module, combined with the OIB and Transaction ID, sends a query to the chain and verifies whether there is a record in the chain. Thereafter, it gives a positive or negative answer, depending on whether there is a really required degree in the chain and whether it complies with the OIB entered.

The certificate/diploma printing module prints a certificate/diploma on the screen in PDF format.

All of the modules listed in this example are displayed in the command-line text interface, that is, in the Ubuntu Operating System Terminal. They can also be programmed into a web application and used in WEB browsers.

2.5.2 User roles

Once the student successfully completes the faculty and defends his graduate thesis, the faculty system reports that the student has graduated. With this application and the module for entering the certificate/diploma, an authorized person at the university will enter the name, last name, and OIBID graduate student and this information will be stored in the chain. As a feedback, he receives a Transaction ID, which gives the student and enrolls on the original print certificate/diploma. It can also be printed in the form of a bar code whose scan is the value of the Transaction ID (Figure 2.5).

Figure 2.5: Certificate input.

The student gets his certificate/diploma and his private key certificate/diploma, which in this case is

80bbfd9b068259c1f02a72b7196417c5464c54a4b68cfaf6e824777e268ff747.

He then reports for a job and after a call from the employer goes to the job interview. The employer asks for a degree to check his qualifications. The procedure is currently being conducted so that the employer contacts the educational institution to verify the validity of the certificate/diploma, most often in writing. This process is long-lasting and consumes a lot of resources. But in this case, the

employer gets a certificate/diploma with a private key. The employer then appoints the OIBID of a person applying for a job and the public key in the application. This way in a fraction of a second returns the information on the validity of the diploma certificate.

After the application's confirmation is answered, the screen prints (Figure 2.6). The name and surname of the student, educational institution, orientation, date and place of graduation are written in the print. The employer eventually has the option of printing a copy of the certificate/diploma for his own archive. If you choose a print option, the certificate/diploma will be generated and opened in PDF format.

For ease of use of the application after release to production, it is a better choice to use it as a WEB application. This means that everything shown will be moved to a web server and the application will access the https protocol (e.g., via URL https://www.diplome.hr) in web browsers. This means that users only need an Internet connection and an account in the application to quickly and securely check the validity of the certificate/diploma.


```
Molim Vas odaberite opciju:
1. Unos novog dokumenta
2. Provjera postojanja dokumenta
Odaberite broj [1,2]? 2
OIB:
68816974905
Ključ:
80bbfd9b068259c1f02a72b7196417c5464c54a4b68cfaf6e824777e268ff747
Goran Fijačko diplomirao na Visokom učilištu Algebra, smjer Multimedija, 15.10.2018. u Zagrebu.
Prikazati diplomu?
1. Da
2. Ne
Odaberite broj [1,2]? 1
Diploma će se prikazati u PDF-u!
```

Figure 2.6: Certificate verification module.

2.6 Conclusion

This chapter presents blockchain technology, its history, and the principle of how it functions within the digital identity [15]. In the theoretical part of the chapter, a comparison of the "classical" identity and digital identity is set out, which is described through examples of personal identity cards and e-citizen systems [16]. Then, following the introduction into blockchain technology and describing the method of achieving consensus and transaction logging, the principle of smart contracts is described, which provide the ability to enter code or even complete applications and put them into blockchains, enabling automation of a multitude of processes [17].

This chapter explains common platforms through three examples (Civic, HYPR, Blockverify), and describes business models that use blockchain as a platform for developing their processes based on digital identity. Also, traditional models with those based on smart deals have been compared. Through examples of cancellation or delays in air travel, voting, music industry, and tracking of personal health records, it was established how existing models are actually sluggish, ineffective, and prone to manipulation, and through examples of blockchain implementation, they showed that these systems functioned faster, more transparent, and most importantly, safer.

The middle part of this chapter describes the application of technology in several industries, from the Fintech industry to the insurance and real estate industry. Concepts and test solutions are described, which are slowly implemented in the production phase and show excellent results. For this reason, we believe that similar solutions will be implemented, increasing adoption of blockchain technology globally.

In the last, practical part of the chapter, a survey of existing solutions that offer creation of its own blockchain and a MultiChain platform was selected. For the purpose of this work, it was necessary to create an Ubuntu virtual machine in the Oracle VM VirtualBox, on which we then installed the MultiChain. Through the Ubuntu Terminal, the application concept for entering, issuing, and verifying university certificate/diplomas is presented; and all functionalities and user roles are described in this process.

Motivation for this research came from lack of practical applications of blockchain and importance of EDU certificate transparency and challenges in their sharing (policy issues, national standard, etc.). By having easy to apply and understand guidelines, it is easier for wider audience to accept and use/reuse sometimes complex digital concepts as part of their solutions and business processes. Taking place in novelty approach, we believe this chapter will contribute and be valuable information for future researchers looking to implement blockchain but moreover ones who are looking to improve exchange, storing, and harmonization of often heterogeneous EDU certificates (in forms, acceptability, and content). As part of institutional research lab, explained approach and proof of concept solution was developed in Algebra University College.

References

[1]   Ashworth, Andrew. Principles of Criminal Law (5th ed, 2006).
[2]   Avery, Lisa. 'A Return to Life: The Right to Identity and the Right to Identify Argentina's "Living Disappeared"(2004) 27 Harvard Women's Law Journal 235.
[3]   Conte, Frances. `Sink or Swim Together: Citizenship, Sovereignty, and Free Movement in the European Union and the United States' (2007) 61 University of Miami Law Review 331.
[4]   Buergenthal, Thomas. 'International Human Rights Law and Institutions: Accomplishments and Prospects' (1988) 63 Washington Law Review 1.

[5] Klepac, G., Kopal, R., & Mršić, L. (2015). Developing Churn Models Using Data Mining Techniques and Social Network Analysis (pp. 1–361). Hershey, PA: IGI Global. doi:10.4018/ 978-1-4666-6288-9

[6] Mohr, Richard. 'Identity Crisis: Judgment and the Hollow legal Subject,' 2007 11 Passages – Law, Aesthetics, Politics 106.

[7] Bromby, Michael., & Ness, Haley. 'Over-observed? What is the Quality of this New Digital World?'7 Paper presented at 20th Annual Conference of British and Irish Law, Education and Technology Association, Queens University, Belfast, April 2005.

[8] Nekam, Alexander. The Personality Conception of the Legal Entity (1938).

[9] Palmer, Stephanie. `Public, Private and the Human Rights Act 1988: An Ideological Divide'. Cambridge Law Journal, 559, 2007.

[10] Solove, Daniel. The Digital Person, Technology and Privacy in the Information Age (2004).

[11] Third, A., Quick K., Bachler M., Domingue J. (2018), Government services and digital identity, Knowledge Media Institute of the Open University.

[12] Davies, Margaret., & Naffine, Ngaire. Are Persons Property? Legal Debates About Property and Personality (2001).

[13] UNWTO. UNWTO World Tourism Barometer: Advance Release January 2017. UNWTO. [Online] January 2017.

[14] World Economic Forum. Digital Transformation Initiative: Aviation, Travel and Tourism Industry. Geneva: World Economic Forum, 2017. REF 060117.

[15] Derham, David. 'Theories of Legal Personality' in Leicester Webb (ed), Legal Personality and Political Pluralism (1958) 1.

[16] Naffine, Ngaire. 'Who are Law's Persons? From Cheshire Cats to Responsible Subjects'(2003) May Modern Law Review 346.

[17] Stacey, Robert. 'A Report on the Erroneous Fingerprint Individualization in the Madrid The Council Of Europe's Convention On Cybercrime. (2001). European Treaty Series 185. Retrieved from http://www.europarl.europa.eu/meetdocs/2014_2019/documents/libe/dv/7_conv_ budapest_/7_conv_budapest_en.pdf.

Souvik Chowdhury and Shibakali Gupta

3 Anomaly detection in cloud big database metric

Abstract: After cloudification of various big data sources or big databases, there is a need to monitor health and security metrics of these big databases. Now there is already a monitoring setup provided by various monitoring suites. The monitoring software collects various metrics with the help of custom codes, plugins, and so on. Here we are proposing a novel approach of modifying the normal metric thresholding to anomaly detection. Every system administrator has a common problem to deal with some intelligent alarm methods, which can produce predictive warnings, that is, the system can detect any anomalies or problem before it occurs.

Here we are proposing an approach, which is basically a modification on standard monitoring where it will detect any anomalies by analyzing previous metric data and indicate any problem. We are planning to harness the power exponential moving average and exponential moving standard deviation method to implement the solution.

Keywords: cloud, big data, statistics, monitoring, moving average, anomaly

3.1 Introduction

Anomaly occurs when a system deviates from its normal running operation. Normally in any anomaly situations, except system overload, system malfunction, brutal network attack, defect, or error during a program arises [1]. Anomaly is also being termed as malware, intrusion, or outlier sometimes [2]. Any system, for example, monitoring software generates data continuously, that is, in time series manner. The volume of data is huge and due to its time series data, this is constantly changing; hence, processing of these data and tracking any kind of problem are very difficult by using a traditional threshold-based monitoring approach.

Let us start with an example. Let's consider a big database hosted in a server, which is running various complex queries, data loading activities, and others. Now we put a threshold of 70% for CPU utilization in that server. This may happen during weekends, month ends, quarter ends as per business needs. A havoc of data loading activity happens and due to that we saw a spike in CPU utilization till 99%. Does that

Souvik Chowdhury, Oracle India, Bangalore, India.
Shibakali Gupta, Department of Computer Science, University Institute of Technology, The University of Burdwan, Burdwan, West Bengal, India.

mean there is a problem in CPU that time? No, the CPU spike happened due to extra load. Now all the traditional monitoring software collects metrics, projects them in graph, and triggers alarm in case if it breaches the predefined threshold. The anomalies in various health and security metric could be due to bug in latest patch/upgrade happened or random hardware failures. Simple metrics can support majority cases. But this is not a good approach since it does not indicate any problem.

This is becoming tougher when the data is complex in nature. Because complex data has various features to monitor. Anomaly detection for such a system is a tough job [3]. That is why it is so important to do anomaly detection [4]. In real-time anomaly detection setup, it constantly monitors the time series data and automatically detects any anomaly and executes corrective actions based on the situation. Hence, this kind of proactive prevention not only saves from a drastic system crash but also saves lot of time and money.

Infrastructure as service is a private cloud-based infrastructure service available at a data center. End users can deploy their workloads using this secure on-demand system. Resource management for information centers needs examining user necessities and, it's on the market resources to satisfy its client demands [5]. The system should have dynamic scalability in terms of infrastructure resource because of varying user demand to stay the pace of all these, an information center should have a period observance system.

The system tracks the resource performance info that comes ceaselessly and detects the abnormal behavior. It then reschedules its resources to remain the system in balance. As an example, an information center allocates a tough and quick amount of CPU and memory for sort of users [6]. After a certain time, if the users run extra computationally expensive programs, the electronic equipment and memory usage are high, which they have extra resources to try and do their job [7]. The amount the system monitors the data and detects these abnormal resource statistics then allocates extra resources dynamically.

Stream process could also be a time-sensitive operation. It needs info preprocessing before doing any analysis. It converts continuous high-volume, high-rate, structured, and unstructured Brobdingnagian info into amount price. Moreover, these Brobdingnagian info volumes must be compelled to be processed with ease, and delivered with low latency, even once info rate is high. To create such a system, Brobdingnagian's ascendible amount operational capability is needed. Recently, we have got seen several massive info frameworks (e.g., Hadoop, Map Reduce, HBase, Mahout, and Google Bigtable) that address the quantifiability issue [8]. However, they surpass batch-based process. The amount the system demands put together the streaming process. Apache Storm could also be a distributed framework that gives streaming integration with time-based in-memory analytics from the live machine info as they're accessible in stream. It generates low latency, amount results. It has instant productivity and no hidden obstacles. When put together contains an easy cluster setup procedure [9].

In statistics, smoothing of information set is to make necessary approximation which helps to capture important patterns within the data and removal of noise or alternative fine-scale structures/rapid phenomena. In smoothing is basically indication of change in square measure of info points i.e. individual points (mostly due to noise) square measure reduced and points that square measure not up to the adjacent points' square measure multiplied resulting in a Sander signal [10]. Smoothing is also utilized in two vital ways, which aids in information analysis:

1. The setup must be able to get more information out of the data, provided the smoothing assumption is reasonable enough.
2. This should provide both robust and flexible analyses. A variety of algorithms are applied in smoothing technique.

- Smoothing technique can be differentiated from the partially and related overlapping concept of curve fitting in the below ways:
- Many times, an explicit function is being used in curve fitting to produce result and if a function form is ever being used, then the immediate results from smoothing are basically the smoothed values and which have no further user.
- Smoothing mainly generalizes the idea of relative slow movement of values, where a small attention is being paid to the close matching of data, whereas curve fitting focuses on matching to the nearest possible data values.
- A tuning parameter is often available with the smoothing method to control the behavior of smoothing. To get the "best fit," the data curve fitting can modify any number of parameters.

Exponential smoothing may be a rule-of-thumb technique for smoothing statistic knowledge victimization the exponential window operates. Whereas within the easy moving average, the historical observations assigned equal weights, exponential functions would not allocate exponentially decreasing weights over time. Its associate in nursing simply learned and applied the procedure for creating some determination that supported previous assumptions by the user, like seasonality. Exponential smoothing is usually used for the analysis of time series knowledge [11].

3.2 Time series modeling

Before understanding the smoothing technique, we will have a quick look into time series data and time series modeling to understand various smoothing techniques well.

A time series could be an assortment of observations created consecutively over an amount. In different words, the information on any characteristic collected with relation to time over a span of your time is named as statistic. Normally, we tend to assume that observations square measure offered at equal intervals of your time,

for example, on associate hourly, daily, quarterly, or yearly basis. The ways of ana-lyzing statistic represent a crucial space of study in statistics [12]. However, before we tend to discuss the statistical analysis, we would prefer to show the plots of its slow series from completely different fields. Within the next three sections, we glance at the plots of three time series, namely, statistic with trend impact, statistic with seasonal impact, and statistic with cyclic impact. These plots of square mea-sure are known as time plots.

3.2.1 Trend effect

A trend could be a future swish variation (increase or decrease) within the statistic. Once values in a very statistic area unit aforethought in a very graph and, on a me-dian, these values show associate degree increasing or decreasing trend over an ex-tended amount of your time, the statistic is named as the statistic with trend impact. We should always note that incomparable series don't show associate degree increas-ing or decreasing trend [13]. In some cases, the values of the statistic fluctuate around a continuing reading and don't show any trend about time. We should always addi-tionally remember that an increase or decrease might not essentially be within the same direction throughout the given amount. Statistic could show associate degree up-ward trend, a downward trend, or haven't any trend in any respect, which allow us to make a case for all three cases with the assistance of examples [14].

3.2.1.1 Time series with upward trend

If a time series exhibits upward trend, that is, the metric values that get increased as time progresses is termed as time series with upward trend [15]. Figure 3.1 shows an upward trend, which depicts the profit of a company plotted for the period 1981–2012.
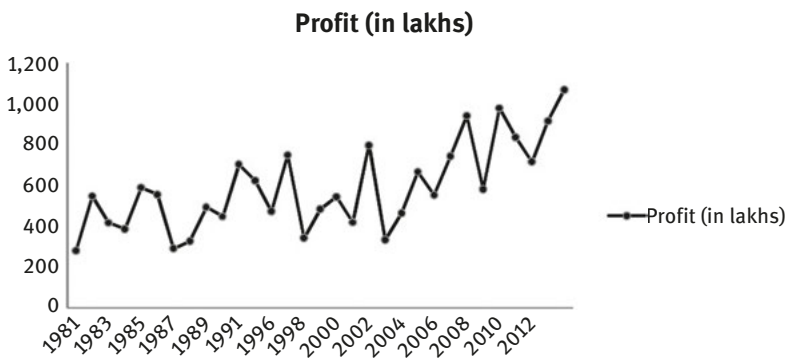
Figure 3.1: Profit of a company from 1981 to 2012.

3.2.1.2 Time series with downward trend

Metric values of a time series when plotted in a graph with respect to time showing a downward trend are termed as time series with downward trend. Figure 3.2 shows a downward trend in the time plot, which describes the values of mortality rate for a developing country from 1991 to 2009.
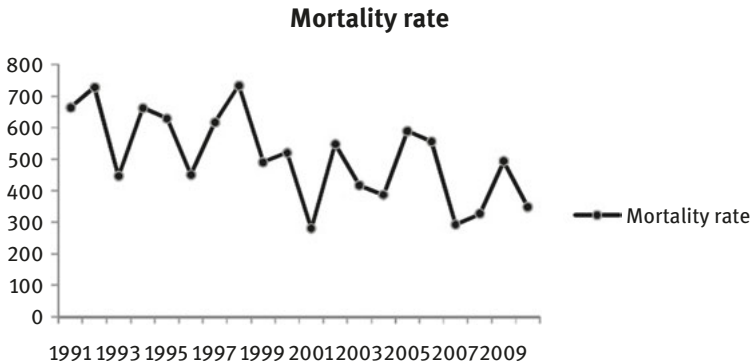
Mortality rate

Figure 3.2: Mortality rates of a developing country from 1991 to 2009.

3.2.1.3 Time series with no trend

Metric values in a time series when plotted in a graph with respect to time if it does not show any trend or to be precise random behavior, thus not showing upward or downward trend then the time series is described as time series with no trend. Figure 3.3 shows the time plot of the commodity production in tons of a factory from 1988 to 2012 where the time series shows no trend.
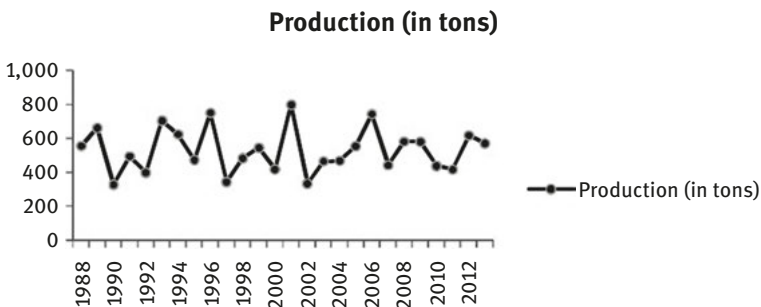
Production (in tons)

Figure 3.3: Commodity production in tons from 1988 to 2012.

3.2.2 Time series with seasonal effect

Metric values in a time series when plotted at a graph with respect to time if it reflects variation seasonally with respect to any period such as half yearly, quarterly, monthly, or a yearly, then the time series is termed as time series with seasonal effect. Figure 3.4 describes time series with seasonal effect, which plots the data of weekly sales of air.

Weekly sales of air coolers
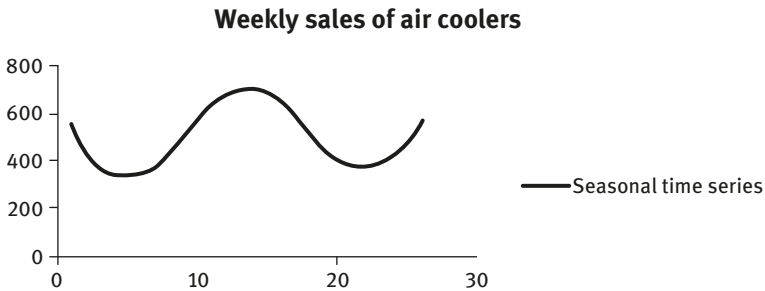
Figure 3.4: Weekly sales of air coolers.

3.2.3 Time series with cyclic effect

Metric values in a time series when plotted in a graph with respect of time if shows a cyclic trend then the time series is termed as time series with cyclic effect. Figure 3.5 shows an example of time series with cyclic effect, which gives employees attrition rate in software industries for last 25 years.

Number of employees in software industries (1988–2012)
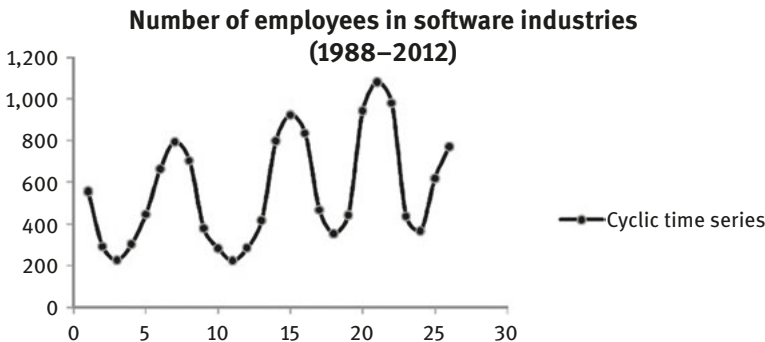
Figure 3.5: Employees attrition rate in software industries for the last 25 years.

3.2.4 Time series components

The variations within the time series statistic values square measure of various varieties arise because of a spread of things. These differing kinds of variations within the values of the information in a very statistic are referred to as elements of the statistic [16].

Various components of time series that involved in variations are (i) trend, (ii) seasonal, (iii) cyclic, and (iv) remaining variations attributed to random fluctuations.

3.2.4.1 Trend component

Usually statistic information shows random variations; however, over a protracted amount of your time, there is also a steady shift within the mean level to the next or a lower level. This steady shift within the level of your time series is thought because of the trend. In other words, the final tendency of values of the information to extend or decrease during a protracted amount of your time is named as the trend.

3.2.4.2 Seasonal component

In a time series, variations that occur because of regular or natural forces/factors and operate in an exceedingly regular and periodic manner over a span of bus or adequate 1-year area unit termed as differences due to the season [17]. Though we tend to typically think about seasonal movement in statistic as occurring over one year, it may also represent any frequent continuance pattern that's but one year in period. As an example, daily traffic volume knowledge shows seasonal behavior at intervals a similar day, with peak level occurring throughout rush hours, moderate flow throughout the remainder of the day, and lightweight ensue hour to early morning. Thus, in an exceedingly statistic, differences due to the season could exist if knowledge area unit recorded on a yearly, monthly, daily, or hourly basis.

3.2.4.3 Cyclic component

Sometimes time series show variation for a fixed period due to certain physical reasons and this is not part of seasonal effects. For example, sometimes economic data are prone to be affected by business cycles with a period varying from few to several years (see Figure 3.5). A period of moderate inflation followed by high inflation is a primary reason for these cyclic variations. Hence, the existence of these business cycles causes some intermittent bias about various cyclic, trend, and seasonal effects [18]. To overcome this problem, we will consider a cyclic pattern or trend in time series only when the duration is more than a year.

3.2.4.4 Irregular component

The long variations, that is, the trend part, and short-run variations, that is, the seasonal and cyclic parts, are referred to as regular variations. Except for these regular variations, random or irregular variations, that don't seem to be accounted for by trend, seasonal, or cyclic parts, exist in virtually incomparable series [19].

3.3 Smoothing of time series

Smoothing of time series is basically filtering out the effect of irregular fluctuations from any time series so that the trend and seasonal effect could be estimated easily and error freely for any time series [20].

Smoothing can be done using three methods: simple moving average method (equal weight), the weighted moving average (unequal) method, and the exponential moving average (EMA) method.

3.3.1 Simple moving average method (equal weight)

Here we calculate the simple moving average of time series data over n periods of time called n-period moving averages. Below is the step to calculate simple moving average.
1. The average of the first m values of the time series is calculated.
2. First value has been discarded, and the average of the next n values has been measured again.
3. Steps 1 and 2 are repeated till all data are used.

The above steps generate a new time series of n-period moving averages [21].

3.3.2 Weighted moving average method (unequal)

The simple moving average methodology delineated in Section 4.1 isn't typically well compatible for activity trend though we are able to take away seasonal variation victimization. It should additionally not lie about to the most recent values. Therefore, the weighted (unequal) moving average methodology is employed. During this methodology, rather than giving equal weights to any or all values, unequal weights are given in such a way that each one of the weights is positive, and there add is capable one. If w_i depicts ith observation weight, the weighted moving average value y_t can be given as follows:

$$Y_t = \sum_{i=-q}^{q} W \ X_i \ t_i + , \ \ w_i \geq 0, \ \ \sum_{i=-q}^{q} W_i = 1 \qquad\qquad (3.1)$$

where x_t is the actual series.

Simple moving average is nothing but a weighted moving average for

$$m \ = \ 2q + 1 \ \text{and} \ w_i = \frac{1}{(2q+1)}$$

3.3.3 Exponential smoothing

Exponential smoothing could be a golden rule technique for smoothing statistic information victimization the exponential window operates. Whereas within the easy moving average the historical observations square measure weighted equally, exponential functions square measure would not assign exponentially decreasing weights over time. Its associate in nursing simply learned and applied the procedure for creating some determination that supported previous assumptions by the user, like seasonality. Exponential smoothing is usually used for the analysis of time series information.

3.3.3.1 Exponential moving average

Below are the steps to calculate exponential moving average (EMA) method.
1. If $y_1, y_2, \ldots, y_t$ is a time series, then the first element of EMA is equal to first element of time series, that is, $y_{1'} = y_1$.
2. The second smoothed value based on EMA, $y_{2'} = w^\star y_2 + (1 - w)^\star y_{1'}$, where $y_{1'}$ is the first element of smoothed series and y_2 is the second element of original time series and w is the smoothing factor and it ranges from $0 < w < 1$.

In this way, tth smoothed value can be calculated by $y_{t'} = w^\star yt + (1 - w)^\star y(t - 1)'$.

3.3.3.2 Exponential moving standard deviate

We have introduced the modified concept of exponential moving standard deviation (EMSD), which is basically refined as smoothing technique over EMA method.

Steps to get EMSD as follows:
1. First element of EMSD is calculated as follows: $y_{1'} = \text{sqrt}((1 - w) \star (y_1 - y_{1'})^2)$ where $y_{1'}$ is the first element of EMA.
2. Second element of EMSD is calculated by $y_{2'} = \text{sqrt}(w^\star y_{1'} + ((1 - w) \star (y_2 - y_{2'})^2)$.

3.4 Working methods

Let's check out the following graph for normal CPU utilization in a server based on hourly aggregation. The graph suggests normal graphical representation of CPU utilization metric being collected with normal thresholding set.

In Figure 3.6, the fixed threshold mechanism detects a problem well; however, let us think an example where the server is very least loaded and normally during weekends runs some batch jobs, which adds load to the server. In these cases, fixed threshold mechanism don't hold good because it will detect the extra load as a problem.
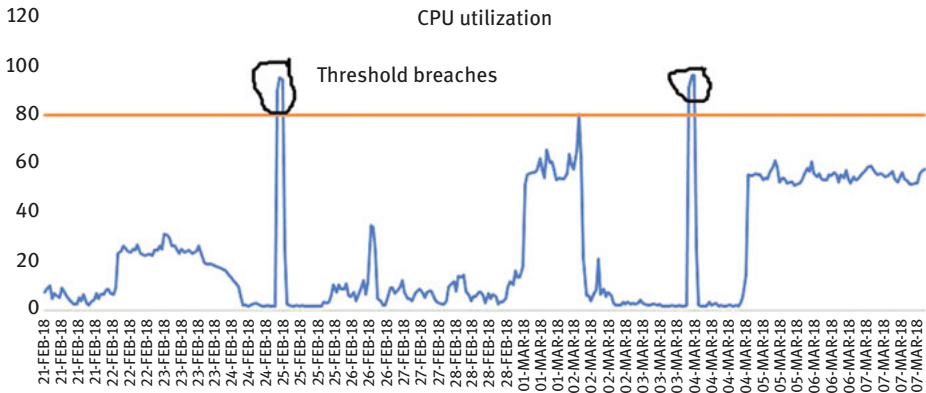
Figure 3.6: Fixed threshold-based mechanism with random spikes.

From Figure 3.7, it is evident that during weekends there is a high growth of CPU, which could be due to some batch job being run on weekends. Hence, this fixed threshold mechanism does not hold good in these cases.
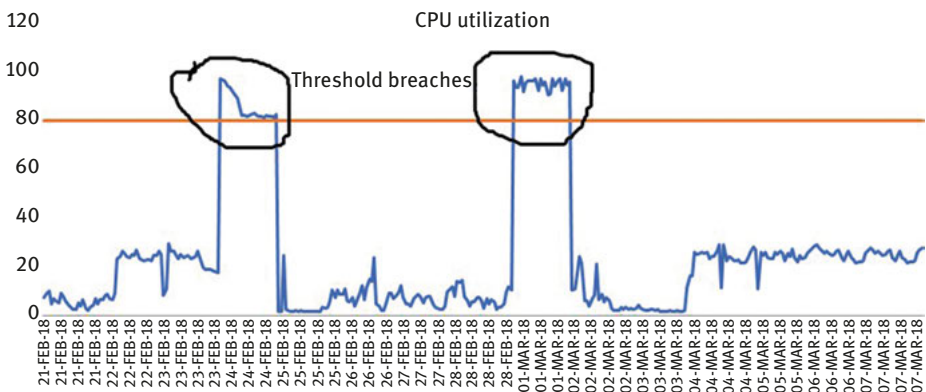
Figure 3.7: Fixed threshold-based mechanism with steady spikes.

Hence, anomaly detection with its adaptive capabilities also can be called as adaptive thresholding, which is more intelligent than the traditional fixed threshold approach. We will see them in Figure 3.7.

In the following example (Figure 3.8), we will see a mix of both unusual growth or spikes (which turn out to be a problem) and steady growth (which could be due to some job) and hence not a problem is successfully being distinguished by this new anomaly detection-based mechanism.
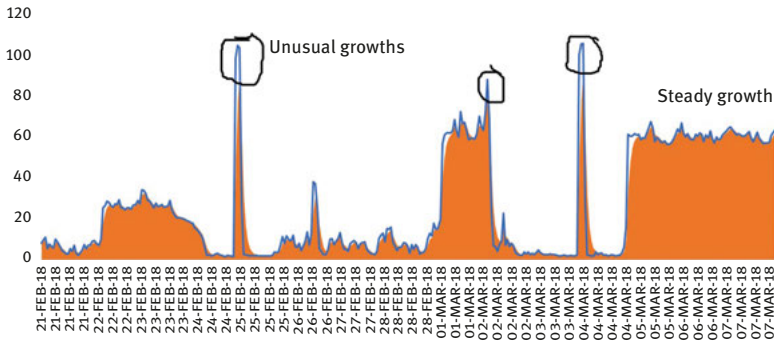
Figure 3.8: Anomaly-based threshold mechanism for unusual spikes and steady spikes.

We will see some more graphs, for example, how the anomaly detection-based mechanism works for memory utilization, where it shows a cyclic pattern as demonstrated early.

From Figure 3.9, we could see there has been a cyclic pattern and anomaly detection-based method that successfully smoothed the raw metric data.
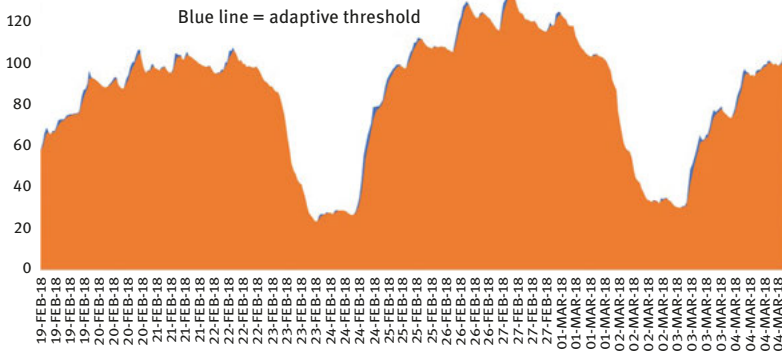
Figure 3.9: Anomaly-based threshold mechanism for cyclic pattern (memory utilization in big database servers).

Now we will see how this anomaly detection-based mechanism works good for cyclic spikes as shown in Figure 3.10.
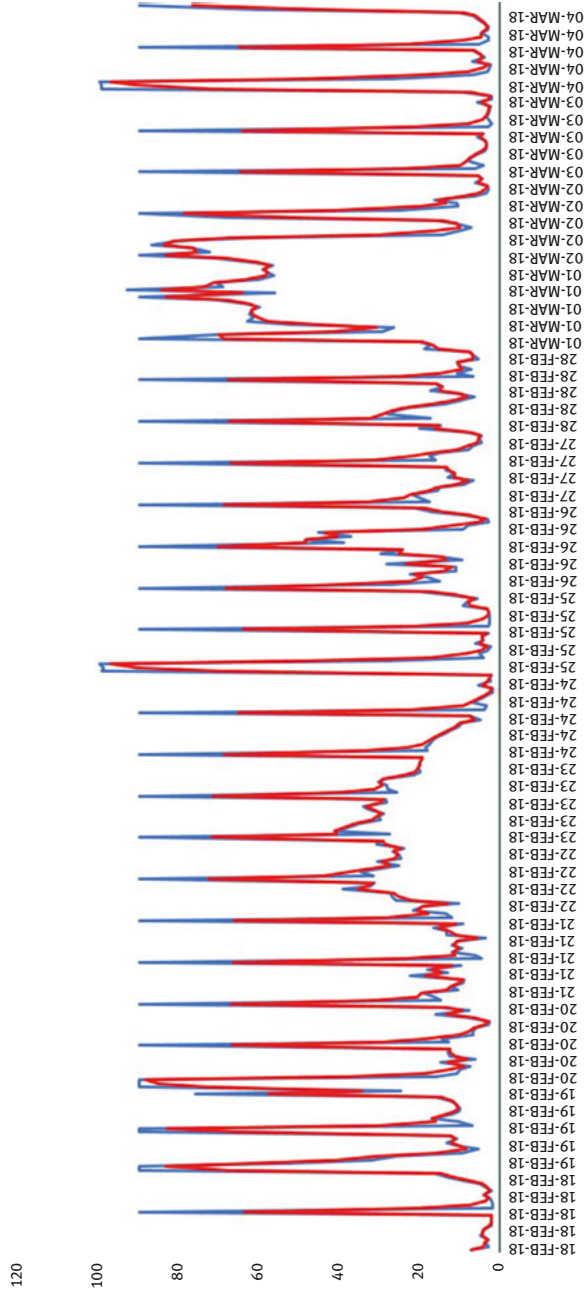
Figure 3.10: Distribution of normal metric data and exponential moving average value.

In Figure 3.10, the blue line depicts normal metric data and red one gives EMA in short EMA values. From the figure we could see for spikes that it is trying to smooth the data which seems to work like normal average. This function works better if the spike stays a bit, which could be due to some load, thus distinguishing a problem with normal loaded situation.

Now we will see the difference between the normal standard deviation and EMSD.

From Figure 3.11 it is evident that EMSD is even a far more refined version of standard deviation, where it is trying to smooth the data and track any real problem. We could see only the extreme variations the standard deviation value also spiked up; however, the case is not the same for EMSD which contains historical knowledge and tries to smooth the data and detect any anomalies.
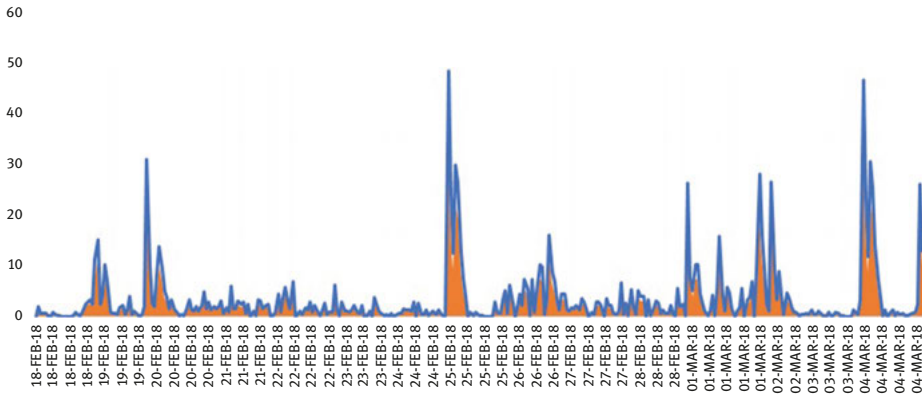
Figure 3.11: Standard deviation (SD) versus exponential moving standard deviation (EMSD).

Figure 3.12 depicts the alert mechanism being set with the help of EMA and EMSD. The orange dots on top are the cases when the alerts got triggered. Here we could see in the highlighted portion that the spikes seem to be regular and this mechanism did not trigger alert for this case.
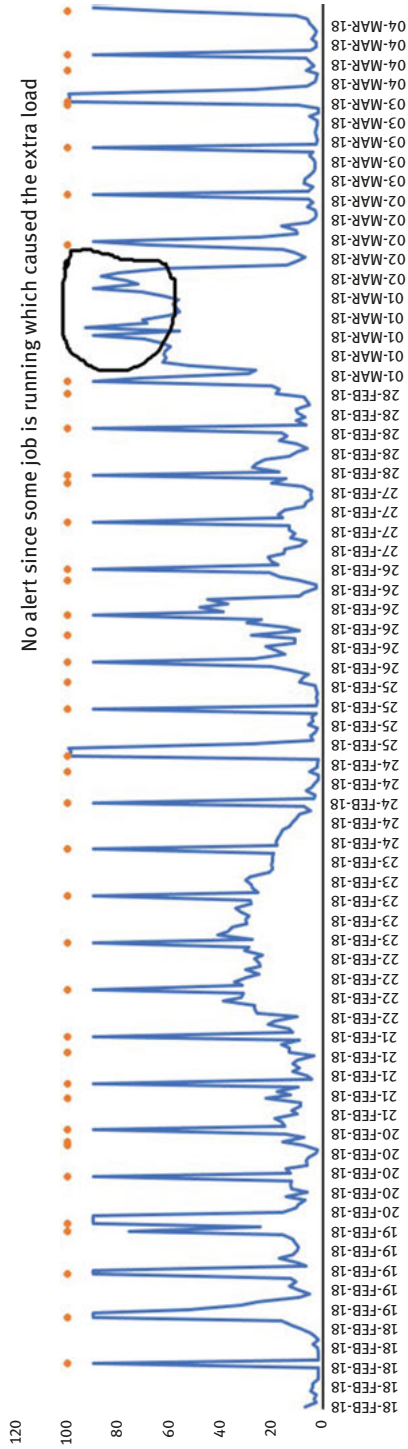
Figure 3.12: Alert mechanism setup using exponential moving average and exponential moving standard deviation.

3.5 Methodology section

We can use the below sample query to get the EMA value and this is specific to Oracle Big database SQL.

```
with t1 as (
select <column names>, row_number() over (partition by <partition column> order by
<ordered column>) r_n,
--2 / (1 + row_number() over (partition by <partition column> order by <ordered col-
umn>)) k_i
0.5 k_i
from <metric table>
), t2 as (
select <column names>, (case when r_n = 1 then 1 else k_i end * c_i) a_i, case when
r_n = 1 then 1 else (1 - k_i) end b_i from t1
), t3 as (
SELECT <column names>,
ai,
xmlquery(REPLACE(wm_concat(bi) over(PARTITION BY <partition column> ORDER BY <ordered
column> rows BETWEEN unbounded preceding AND CURRENT ROW), ',', '*') RETURNING con-
tent).getnumberval() mi
FROM t2
), t4 as (
select <column names>, m_i, (a_i / m_i) x_i from t3
)
SELECT <column names>,
round(m_i * SUM(x_i) over(PARTITION BY <partition column> ORDER BY <ordered column>
rows BETWEEN unbounded preceding AND CURRENT ROW), 3) ema
FROM t4;
```

We can use the following sample to query EMA using simplified method. Here we have two options to select the weight factor *w* mentioned in Section 3.3.3.1.

```
Special case w = 0.5. Here the weight factor has been fixed to 0.5, that is, giving
equal weightage to current metric data and past metric data, thus leaving a trail of
historical metric data in calculation:

with t1 as (
select <column names>, row_number() over (partition by <partition column> order by
<ordered column>) r_n, amount * power(2, nvl(nullif(row_number() over (partition by
<partition column> order by <ordered column>) - 1, 0), 1)) c_i
from <metric table>
)
select <column names>, round(sum(c_i) over (partition by  <partition column> order by
<ordered column> rows between unbounded preceding and current row)    / power(2, r_n),
3) ema
```

```
from t1;
Special case w = 2/(1 + i). Here it is dynamic; hence, for various levels it would
have different behavior:

with t1 as (
select <column names>, row_number() over (partition by  <partition column> order by
<ordered column>) r_n, amount * row_number() over (partition by  <partition column>
order by <ordered column>) c_i
from <metric table>
)
select <column names>, round(sum(c_i) over (partition by  <partition column> order by
<ordered column> rows between unbounded preceding and current row) * 2   / (r_n * (r_n
+ 1)), 3) ema
from t1;
```

Now we will discuss about various jobs and procedure setup to achieve this task. The challenge was to achieve good performance because the real-time calculation of EMA and EMSD is a difficult one and with enormous number of targets and its metric for a cloud big database, cloud datacenter is even getting worse. We concentrated on SQL tuning part and tried to make the query as fast as possible. We included parallelism in query. But still that was not good enough. We took another option to parallelize the job itself. Hence, the current model has been depicted as follows:

Figure 3.13 depicts the job flow. Here, a master job is responsible for all job allocations. It was being tested to use parallel 20 jobs. The main job ANOMALY_COUNT_REFRESH analyzes currently running jobs and forks a new session and attaches the job to the session (Figure 3.14).
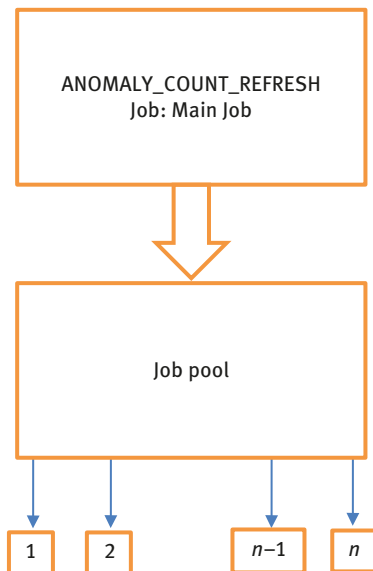
Figure 3.13: Job flow.
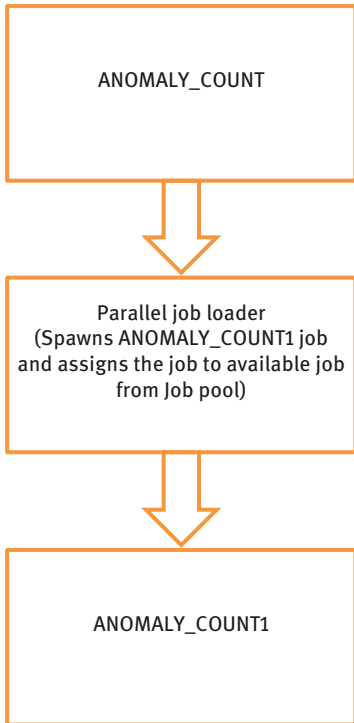
Figure 3.14: Job flow(1).

```
BEGIN
<Scheduler job create module> (
    job_name        => 'ANOMALY_COUNT_REFRESH',
    job_type        => <Object Type>,
    job_action      => 'ANOMALY_COUNT',
    start_date      => <Job start date>,
    repeat_interval => 'FREQ=<minutely,hourly etc>;INTERVAL=<interval duration 5
mins,1 hour etc>;',
    end_date        => NULL,
    enabled         => TRUE,
    comments        => <Job Description>);
END;
/
```

Once this job is assigned to a session, the main job comes in which it collects all necessary data and stores the output (Figure 3.15).

ANOMALY_COUNT1

Metric data collection
from data source

Metric data collection
From data source

Calculate EMA and EMSD based on
previous historical data
Update last load timestamp

Invokes ALERT_TRIGGER
Mechanism to trigger alert for
anomalies detected
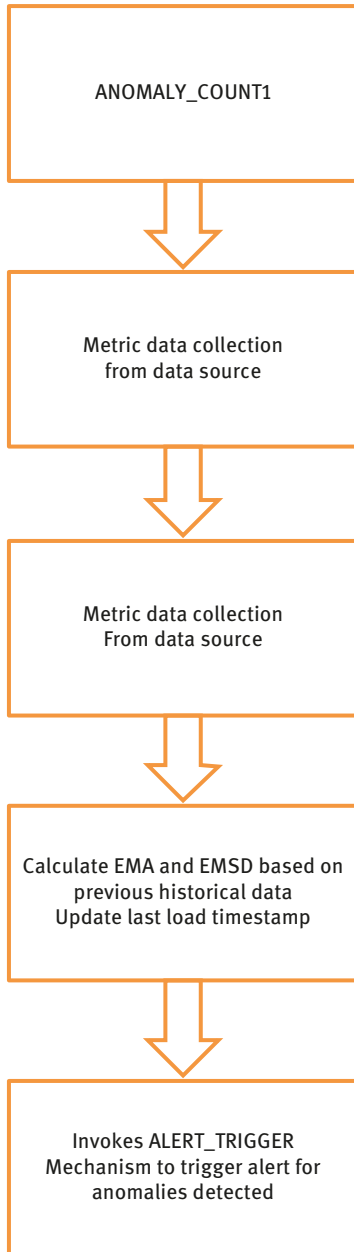
Figure 3.15: Job flow(2).

We have used the following sample commands to create pool of jobs. At present, only two jobs have been shown.

```
BEGIN
  <Scheduler job create module> (
    job_name          => 'ANOMALY_COUNT_JOB1',
    job_type          => => <Object Type>,
    job_action        => 'ANOMALY_COUNT1',
    end_date          => NULL,
    enabled           => FALSE,
    number_of_arguments => 2);
END;
/
BEGIN
    <Scheduler job create module> (
    job_name          => 'ANOMALY_COUNT_JOB2',
    job_type          => <Object Type>,
    job_action        => 'ANOMALY_COUNT1',
    end_date          => NULL,
    enabled           => FALSE,
    number_of_arguments => 2);
END;
/
```

3.6 Future work

This chapter tested on CPU utilization and memory utilization of big database servers. This approach can be extended to other types of hosts, servers, VM, and so on. This method of new monitoring mechanism should be tested for network devices as well. For network device monitoring, there lies special challenge because network devices have lot of interlinked metrics to detect a problem, for example, for a network device the reduction in traffic does not necessarily mean reduction in connected users. It could also be due to error, discards, packet loss, and so on. The method can be extended to datacenter monitoring for complete stack, say from security appliances to bare metal server. This kind of anomaly detection-based mechanism can be helpful for capacity planning as well. Capacity planning for network devices of a complete cloud datacenter footprint is a tedious job. Because it consists of lot of network devices, and a cisco switch or juniper device can have 1,000 or 2,000 network interfaces attached to it. Hence, doing capacity planning or pinpointing problem for a network interface is a difficult task.

References

[1]   Zhu, Qingsheng., & Liu, Renyu. "A Network Anomaly Detection Algorithm based on Natural Neighborhood Graph", "International Joint Conference on Neural Networks 2018".

[2]   Wang, Lin., Xue, Bai., Wang, Yulei., Li, Hsiao-Chi., Lee, Li-Chien., Song, Meiping., Yu, Chunyan., & Li, Sen. et al. "Iterative anomaly detection", "IEEE International Geoscience and Remote Sensing Symposium 2017".

[3]   Edisanter Lo. "Hyperspectral anomaly detection based on a generalization of the maximized subspace model," "2013 5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing".

[4]   Yao, Danfeng., Shu, Xiaokui., Cheng, Long., Stolfo, Salvatore J., & Bertino, Elisa.., Ravi et al. "Anomaly Detection as a Service: Challenges, Advances, and Opportunities, Synthesis Lectures on Information Security, Privacy, and Trust, Electronic".

[5]   Zhao, Rui., Du, Bo., & Zhang, Liangpei. "GSEAD: Graphical score estimation for hyperspectral anomaly detection, 2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote".

[6]   Chen, Mingqi., Jiang, Ting., & Zou, Weixia. "Differential physical layer secret key generation based on weighted exponential moving average, 2015 9th International Conference on Signal Processing and Communication".

[7]   Alexander, Belyaev., Ivan, Tutov., & Denis, Butuzov. "Analysis of noisy signal restoration quality with exponential moving average filter, International Siberian Conference on Control and Communications,2016"

[8]   Md. Emdadul, Haque., Md. Nasmus, Sakib Khan., & Md. Rafiqul, Islam Sheikh. "Smoothing control of wind farm output fluctuations by proposed Low Pass Filter, and Moving Averages, International Conference on Electrical & Electronic Engineering 2015".

[9]   Lutfi Al-Sharif, Ahmad Hammoudeh. "Estimating the elevator traffic system arrival rate using exponentially weighted moving average(EWMA), IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies 2017".

[10] Tajiri, Hiroki.. "Input filtering of MPPT control by exponential moving average in photovoltaic system, Teruhisa Kumano IEEE International Conference on Power and Energy 2012".

[11] Xie, Y. J., Xie, M., & Goh, T. N. "A MEWMA chart for a bivariate exponential distribution, IEEE International Conference on Industrial Engineering and Engineering Management 2009".

[12] Hansun, Seng., & Kristanda, Marcel Bonar. "Performance analysis of conventional moving average methods in forex forecasting, International Conference on Smart Cities, Automation & Intelligent Computing Systems 2017".

[13] Guo Feng Liu, Chen-Yu., Bin, Zhou., & Su-Qin, Zhang. "Spares Consumption Combination Forecasting Based on Genetic Algorithm and Exponential Smoothing Method, Fifth International Symposium on Computational Intelligence and Design 2012".

[14] Akpinar, Mustafa., & Yumusak, Nejat. IEEE International Conference on Environment and Electrical Engineering and IEEE Industrial and Commercial Power Systems Europe 2017".

[15] Wang, Shuo., & Li, Ning. "MYCAT Shard Key Selection Strategy Based on Exponential Smoothing, 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference 2018".

[16] Morimoto, J., Kasamatsu, H., Higuchi, A., Yoshida, T., & Tabuchi, T. "Setting method of smoothing constant in exponential smoothing, SICE 2004".

[17] Yanwei, Du.. "Research on Requirement Forecasting of Raw Materials for Boiler Manufacturing Enterprise Based on Exponential Smoothing Method, Second International Conference on Computer Modeling and Simulation, Electronic 2010".

[18] Setiawan, Wawan., Juniati, Enjun., & Farida, Ida.. "The use of Triple Exponential Smoothing Method (Winter) in forecasting passenger of PT Kereta Api Indonesia with optimization alpha, beta, and gamma parameters, 2nd International Conference on Science in Information Technology 2016".

[19] Li, Jing., Xu, Danning., Zhang, Jinfeng., Xiao, Jianhua., & Wang, Hongbin.. "The comparison of ARMA, exponential smoothing and seasonal index model for predicting incidence of Newcastle disease, World Automation Congress, Electronic 2010".

[20] Wi, Young-Min., Kim, Ji-Hui., Joo, Sung-Kwan., Park, Jong-Bae., & Oh, Jae-Chul. "Customer baseline load (CBL) calculation using exponential smoothing model with weather adjustment, Transmission & Distribution Conference & Exposition 2009".

[21] Chu, Chao-Ting., Chiang, Huann-Keng., Chang, Chao-Hsi., Hong-Wei, Li., & Chang, Tsung-Jui.. "A exponential smoothing gray prediction fall detection signal analysis in wearable device, 6th International Symposium on Next Generation Electronics 2017".

Shibakali Gupta, Ayan Mukherjee

4 Use of big data in hacking and social engineering

Abstract: Nowadays, in the fast-paced world of Google and Facebook, every detail of human being could be considered as a set of data or array of data that can be stored, verified, and processed in several ways for the benefits of users. Big data would be perfectly described with humongous large and complex data entities, where classic approach application software is incompetent for them. Big data epitomizes the evidence chattels classified by a high volume, velocity, and variability to require precise technology and analytical approaches for its transformation into value. Big data include netting data, search, data stowing, transmission, updating, data scrutiny, visualization, sharing, querying, data source, and information confidentiality. Big data can castoff in innumerable sectors like defense, health care, and Internet of things. The most famous example probably being Palantir, which was primarily sponsored by the CIA (Central Intelligence Agency). Its primary function was to deliver analytics sway in the war against terrorism of any kind but with accumulative dependency on big data, the menace of exploitation of this data also arises. The prominence of big data does not gyrate around data magnitude or dimensions rather it revolves around how you process it. You can consider stats from whichever cradle and analyze it to discover answers that facilitate cost diminutions, interval time declines, fresh product development and elevated offerings, and smart management. When you conglomerate big data with efficient and dynamic analytics, you can achieve business-correlated tasks such as detecting fraudulent behavior, recalculating entire risk portfolios in shorter span of time, determining root causes of failures, disputes, and blemishes in near real time. Few instances such as Cambridge Analytica lighten the insight of the exploitation of the big data. There are several instances where large amount of data has been stolen like in 2014, Yahoo Inc., where 3 billion accounts were effectively compromised according to official sources or in 2016, Adult Friend Finder where 412.2 million accounts were effected with credit card details compromised as well.

Deprived of the encompassing span of big data, it is taut to perceive a consequence where dappled endeavors and marginal verboten deeds would be a newsflash. It is merely with the inclusion of big data, does the sheer extent of this statistics turn heads. If one individual cheats during a test, it is just earnest of a quip from the instructor. If the entire class collaborates and cultivates a structure of cheating, it becomes newsworthy. The Panama Papers are an exceptional specimen

Shibakali Gupta, Department of Computer Science, University Institute of Technology, The University of Burdwan, Burdwan, West Bengal, India.
Ayan Mukherjee, Cognizant, Kolkata, India.

of an event that is not a requisite illegal, but sketchy to say the least. The statistic that several sets of international figures were acknowledged in this bulk data set is what marks the news. With the evolution of big data, it makes treasured visions for hackers invariably tempting, but it also provides a big structure of data that converts it to payload utmost necessary to protect.

In such a scenario, the security of big data is very important. This chapter shares sheer insight of how big data can be used in hacking and social engineering. This chapter will try to list down the ways big data is mined from various sources such as Google Services of Android and Facebook. It will list the various ways the big data is used in day-to-day life by the given companies and other advertising companies. This chapter will try to enlist all the major ill ways this data can be used against us and the ways the important and private data can be protected from the data-collecting companies.

Keywords: big data, ethical hacking, social engineering, Cambridge Analytica, big data security, data privacy, risk and threat

4.1 Introduction to big data

Big data is humongous information collections, multifarious that customary processing application software for data set that is derisory to pact. Big data encounters comprise several functionalities such as netting, exploration, and examination of information along with features like sharing and visualization. It also involves querying and updating information along with confidentiality. There are numerous theories concomitant with big data that are veracity, volume, velocity, variety, and value.

The nomenclature big data implies to the practice of prognostic analytics, user behavior analytics, or few other unconventional data analytic techniques that excerpt value from statistics, and seldom to a certain magnitude of dataset. Analysis of datasets can reveal new correlations to highlight business trends, practitioners of medicine, prevent diseases, combat crime, and so on. Often issues related to several portfolios like government, researchers, marketing, and business executives alike are met with the help of enormous datasets in ranges, including Internet search, urban informatics, financial technology, and business informatics. Even scientists encounter several limitations in various topics, including genomics, meteorology, and environmental research comprising complex physics simulations.

The motivation of big data can be best demarcated as quoted by Carly Fiorina, "The Goalmouth is to convert Facts into information which in-turn can be converting to insight." There is a massive growth in datasets because Internet of things devices such as mobile devices, aerial (remote sensing), radiofrequency identification readers, and wireless sensor networks, they are gradually gathered by cheap and abundant information sensing. Big data exemplifies the asset of information with features

such as volume, variety, and velocity to oblige explicit technology and analytical techniques for its transformation into value. Additionally, a new *V*, veracity, is added by some officialdom to describe it, revisionism challenged by some industry authorities. The three *V*'s have been primarily expanded to other harmonizing characteristics of big data [1].

The systematic study of big data can lead to:
- **Tuning according to target audience** – Big data is used by business today for scrutinizing gushes of the target audience and entertain them with optimized services to upsurge the business.
- **Cost cutting in various sectors** – Scrutiny of such mammoth bulk of data has also aided business in cutting down their overhead expenses in various sectors. Several bucks are being saved by enhancements in operational efficiency and more.
- **Intensification in operating boundaries in different sectors** – Big data also aids businesses in increasing operational brims in different sectors. With the help of big data, lot of blue-collar labor can be converted into machine task and this helps in growing operating precincts.

Big data can be described by the following characteristics:
- **Volume:** Volume can be defined as the magnitude of generated data and stowed data. The volume of the data regulates the value and potential insight and whether it can be deliberated as big data or not.
- **Variety:** Variety can be defined as the category and attitude of the data. This is usually for the people who scrutinize the data to increase the efficiency of the resultant insight. Big data concludes the missing or the omitted pieces from data fusion; it derives its information from sources like text to video anything.
- **Velocity:** This can be termed as the promptness at which the facts and figures are bred and treated to fulfill the requirements and dares that lies in the route of progress and expansion. Big data is frequently usable, reachable, and affordable in real time.
- **Veracity:** This can be defined as data eminence of netted data that can vary prominently, manipulating the precise analysis.

Cyber-physical and workshop systems may have a 6C system:
- Connection (sensor and networks)
- Cloud (computing and data on demand)
- Customization (personalization and value)
- Content/context (meaning and correlation)
- Cyber (model and memory)
- Community (sharing and collaboration)

Data as a requisite should be treated with cutting-edge analytics and algorithms to reveal evocative statistics.

For example, to achieve success in a factory one must contemplate both visible and invisible concerns with various components. Information generation algorithms must distinguish and address obscure issues such as machine degradation and component wear.

4.1.1 Application of big data

Big data helps in transmuting cream commercial progressions by appropriate and precise analysis of accessible statistics. These processes generally embrace:

i.   **Procurement with big data**: Ultimatum of requirements or necessities can be appropriately conjectured as per various conditions and features offered with big data.

ii. **Big data in product improvement:** It can approximately predict the type of invention compulsory to intensify sales.

iii. **Data warehousing in manufacturing industry**: Data warehousing is a major analytical methodology for categorizing apparatus or measures the practice deviance from the quality benchmark.

iv. **Data warehousing system for product dissemination:** Grounded depending on statistics presented; records scrutiny is considered useful to confirm symmetric circulation in arcade.

v.   **Data warehousing system in product advertisement:** Data warehousing system aids in significant advertisement stratagem that could upsurge sale by several folds.

vi. **Price administration using data warehousing system:** Data warehousing system helps business in studying market chart. This is an important part to sustain position in arcade and price management.

vii. **Merchandising:** Retail arcade relies majorly on data warehousing system and analytics to identify the recent trends of the goods.

viii. **Data warehousing system in sales:** Data analytics assists in optimizing product mix. It helps in aggregating sale for the commerce. It is also consignment of sales resources and accounts, and other operations.

ix. **Store maneuvers using data warehousing system:** Stored procedures can be observed by various analytical tools that lead to shrink in manual work. It regulates several factors like training of demographics or inventory echelons based on predicted procurement patterns.

x.   **Data warehousing system in HRs:** Data warehousing system has an altered way of recruitment and other human resource maneuvers. You can also discover the physiognomies and behaviors of efficacious employees, as well as other employee insights to accomplish talent better.

xi. **Data warehousing system in banking:** Data warehousing system has provided major prospect to corporations to visualize the larger scenario due to harmonizing the delicate trend of the records for prioritizing the privacy and shielding of information along with conveying value adds for customers. It has been fully embraced by several companies to drive business and advance the services they offer to customers.

xii. **Data warehousing system in finance:** Financial amenities have extensively espoused data warehousing system analytics to advise enhanced investment assessments with constant returns.

xiii. **Data warehousing system in telecom:** According to reports in "Global Data Warehousing System Analytics Market in Telecom Industry 2014–2018," it was found that the usage of data analytic tools in telecom segment is predicted to propagate at a compound annual growth rate of nearly 28% over the next four years.

xiv. **Data warehousing system in retail:** Retailers hitch data warehousing system to suggest that consumer has personalized shopping experiences. Evaluating customer is one-way data warehousing system technology in making a spot in retail. Two-thirds of retailers have made financial gains in customer management and CRM through data warehousing system.

xv. **Data warehousing system in healthcare:** Data warehousing system is used for scrutinizing data in the electronic medical record system with the objective of sinking costs and refining patient care. This data includes the amorphous data from physician notes, pathology reports, and so on. Data warehousing system and healthcare analytics have the technological advancement to predict, prevent, and cure diseases.

xvi. **Data warehousing system in media and entertainment:** Data warehousing system is altering the broadcasting and entertainment industry, providing users and viewers a much more tailored and enriched experience. Data warehousing system is utilized for growing revenues, analyzing real-time patron sentiment, increasing promotion effectiveness, ratings, and viewership.

xvii. **Data warehousing system in tourism:** Data warehousing system is renovating the global tourism. Information about the world is easily available than ever before. People have detailed itineraries with the help of data warehousing system.

xviii. **Data warehousing system in airlines:** Data warehousing system analytics provides with necessary tactics to the aviation industry. An airline now knows where each and every plane is heading, where any passenger is sitting in any of the flight, and what a passenger is watching on the IFE (In-flight Entertainment) or connectivity system.

xix. **Data warehousing system in social media:** Data warehousing system is a motivating influence behind every marketing resolution made by social media houses and it is driving personalization to the highest extent possible (Figure 4.1).

Big data application examples in different industries:

Retail/Consumer
- Merchandizing and market basket analysis
- Campaign management and customer loyalty programs
- Supply-chain management and analytics
- Event- and behavior-based targeting
- Market and consumer segmentations

Finances & Frauds Services
- Compliance and regulatory reporting
- Risk analysis and management
- Fraud detection and security analytics
- Credit risk, scoring and analysis
- High speed arbitrage trading
- Trade surveillance
- Abnormal trading pattern analysis

Web and Digital media
- Large-scale clickstream analytics
- Ad targeting, analysis, forecasting and optimization
- Abuse and click-fraud prevention
- Social graph analysis and profile segmentation
- Campaign management and loyalty programs

Health & Life Sciences
- Clinical trials data analysis
- Disease pattern analysis
- Campaign and sales program optimization
- Patient care quality and program analysis
- Medical device and pharmacy supply-chain management
- Drug discovery and development analysis

Telecommunications
- Revenue assurance and price optimization
- Customer churn prevention
- Campaign management and customer loyalty
- Call detail record (CDR) analysis
- Network performance and optimization
- Mobile user location analysis

Ecommerce & customer service
- Cross-channel analytics
- Event analytics
- Recommendation engines using predictive analytics
- Right offer at the right time
- Next best offer or next best action

Figure 4.1: Data warehousing system application in various industries.

4.1.1.1 Why big data is a lucrative target

As we race into the future, a swelling amount of modules concomitant to the infrastructure of our realm and enterprises are reliant on an Internet assembly. The probability of devastating cyberattacks from aggressive states, cyberterrorists, and hacktivists becomes much more real: This can be visualized pretty well in movie named Die Hard 4.0, where several unmanned cars crashing or rerouting of energy and electricity on a large scale thereby leading to blackout or tampering in traffic signal leading to accidents.

Few technological loopholes would never lead to an efficacious kinematic assault in a large scale. As an alternative to get access to the system, the invaders use several diverse but fundamental methodologies over the time. Data sabotage, that is, altering of data records can be considered one such cyberattack that seems to be minute but could be used by invaders for major advantages. Small manipulation in data could affect a lot in major sectors like stock market or defense agencies. A small manipulation of rating of a particular fake product in retail market could lead to its perception as a original product and major sale boost in retail sector or a simple tickle in financial figure of a company's remuneration could provide a major boost in stock market.

US agencies such as CIA and FBI are perceived as major fronts in 2016 for cybercrimes.

Several open confab concerning cyberterrorizations have been dedicated to the concealment and accessibility of information. In near future, we might also visualize several online maneuvers of manipulating major governmental decision, investors of stock market, or corporate decisions due to alterations and manipulation in veracity of the electronic figures provided to them.

4.1.1.1.1 New concerns for cybersecurity connoisseurs

Numerous sectors in recent years have seen ascending trends of data integrity outbreaks. A false news of President Obama's injury by Syrian hackers through Twitter account of Associated Press, leading to a sharp 150-point dip in stock market, can be seen as a simple but direct example of the same. The similar example can also be seen as minute altercation in a cooling system by Stuxnet worm, which lead to rescind Iranian nuclear program [2].

"Data veracity outbreaks have a number of dimensions to them," said Eddie Schwartz, universal vice president at ISACA, an international cybersecurity association. "If you get hold of a meticulous system like the power grid or water system that encompasses machinery operated by workstations and make minute alteration in the operational directives for that equipment, it can lead to some cataclysmic consequences – power outages or deviations in chemical balance."

4.1.1.2 Previous data warehousing system breaches in recent times

i. Yahoo
Date: 2013–14
Impact: 3 billion user accounts
Details: "In September 2016, the past prevailing Internet colossal, while in parleys to peddle itself to Verizon, indicated that it had been the prey of the humongous data breach in recent antiquity, probable by 'a state-sponsored artiste.' The outbreak compromised the original appellations, dates of birth, email addresses, and handset no. of Five hundred million patrons. The corporation published that the preponderance of the passwords had been hashed via the robust crypt algorithm.

Few months later, it buried that previous record with the revelation that a breach in 2013, by a different set of black hat hackers had compromised 1 billion records with names, dates of birth, security questions and answers, email addresses and passwords that were not well secured as those involved in 2014. In October 2017, Yahoo reviled that, all 3 billion-user accounts were being compromised.

The breaches bashed a probable $350 million off from Yahoo's sale amount. Verizon eventually remunerated $4.48 billion for Yahoo's core Internet industry. The pact stated that the two corporations to share regulatory and legal obligations from the breaches."

ii. Adult Friend Finder

Date: October 2016

Impact: More than 412.2 million accounts

Details: "The Friend Finder website, which comprised spontaneous hookup and adult content network like Adult Friend Finder, Cams.com, iCams.com etc., were penetrated in mid-October 2016. Hackers unruffled two decades of data from six databases that include complete details like names, email addresses and passwords.

The feeble SHA-1 hashing algorithm fortified most of the passwords, which was predestined that almost cent percent of them were been decoded by the time LeakedSource.com circulated its scrutiny of the entire data next month.

CSO Online stated at the stage that, 'a canvasser who has online Twitter identity as 1x0123 and as Revolver in other networks displayed images of Adult Friend Finder displaying an issue called LFI (Local File Inclusion vulnerability) which was being targeted. The ID stated said that the vulnerability was exposed in a service method on the production servers deployed by Adult Friend Finder. AFF Vice President Diana Ballou acknowledged the same and confirmed, that the issue was fixed which was prevailing due to injection vulnerability.'"

iii. eBay

Date: May 2014

Impact: Nearly 150 million user data

Details: "The Internet Giant known for its online auction testified a cyber-attack in May 2014 that is said to have exposed all the details and hashed passwords of all of its users. The company said, it was a scenario of Social Engineering where hackers accessed the company intranet using the credentials of three internal employees, and had a completed backdoor access for almost a year. In this tenure they paved their way to the consumer database.

As mitigation, it requested its users to update their passwords, but alleged that the financial statistics, such as credit card info, was stowed disjointedly and was not compromised. The venture was condemned for a dearth of communication notifying its consumers and pitiable implementation of the password-renewal procedure."

iv. Equifax

Date: July 29 2017

Impact: 143 million consumers along with Credit Card info for 209,000 consumers

Details: "Equifax that is one of the dominant credit agencies in America revealed on Sept. 7, 2017 that an application susceptibility on one of their networks led to a records breach that exposed roughly 147.9 million users. The breach was exposed on July 29, but the enterprise stated that it possibly happened in mid-May."

v. Uber

Date: Late 2016

Impact: 57 million along with 600,000 drivers exposed.

Details: "The Corporation came to know about the breach in late 2016 wherein couple of hackers was able to retrieve personal details of 57 customers of the Uber. They were also able to retrieve the driver license details of 600,000 Uber drivers. Credit card or Social Security numbers were secured as per the company. The hackers got access Uber's GitHub code repository account, where they retrieved user credentials to Uber's AWS account. Those authorizations should certainly not be on GitHub.

The Breach was made public a year later by Uber. They compensated the hackers with $100,000 to rescind the records with no clause or way to authenticate that same. The paid them stating it was a 'bug bounty.' Uber also sacked its CSO and placed the responsibility on him.

The breach is said to have affected Uber in both reputation and money. At the time that the break was announced, the business was in discussions to trade a stake to Softbank. Uber's valuation declined from $68 billion to $48 billion by the time the deal was signed."

4.1.1.3 Vulnerabilities in data warehousing system for big data

Data warehousing system for big data or big data analytics was defined by the connoisseurs with the help of terminology like value of the data, volume, and variety of the same, along with velocity and veracity of the data. This is also defined by 5V's. Recently, an additional V is gaining the focus of the market, let alone the experts of big data analysis. Vulnerability, as it gains focus in the market, distresses entire enterprise sector and is urging for major attention since, if this is not handled, rest all will be at stake. Due to numerous proficiency, it has now received the consideration of entire domain.

Due to its capabilities of further optimizing the business by better understanding of the habitué and enhanced productivity suggestion, it has made the life of decision makers a lot easier. Then a clause of confidentiality also comes into the picture, which mandates the enterprise to secure patron's data from any unauthorized scrutiny and due to this the vulnerability dispute needs to be addressed as an important contemplation.

4.1.1.3.1 Reason for all the 6V's
The data confidentiality is the major dimension several syndicates are dealing with, still there are numerous unquestionable cradle for procurement of user's personal data. As stated by Marr, "Vulnerability addresses the information that a mounting number of people are becoming comfortable on to the element that their delicate data, the sensitive data of many commercial initiatives, is being gulped up by the gigabyte, used to pry into their comportment and, eventually, peddle things."

Like several research organizations, an organization used for credit referencing named as Experian mentions similar views in their white papers and other research documents. To mien at the data susceptibility trait, one could mien at a sociological facet at the issue. Several experts like head strategist of Experian named as John Roughley states, "We think about things emotionally, and the emotion that's associated with data is sometimes one of nervousness, anticipation or liability. That's partly because it's new but it's also because everyone's seen stories in the various sources of Media about data breaches, and record number of individuals have experienced their records being tainted in some shape or form – the phone ringing off the hook with people asking about payment fortification indemnification."

The principal fears around data warehousing system susceptibility could be addressed through rudimentary questions like, how did my data reach these advertising companies contacting me? What extent of access they have to my data? What around the financial info? How much easily hackers can access my data? Will all of that be whipped by hackers to siphon off money from my account? Nevertheless, to address these disquiets, there needs to be some key steps such as reassuring customers, whose whatsoever information they offer to the company will be securely stored, will not be misplaced, or used for malevolent purposes.

As John Roughley explained, "It's about doing what you promise you will do, and as officialdoms we have a prerequisite to mark certain that we always perform with integrity and with regards to the custodianship of someone's data. It's about keeping it secure, keeping it safe, and not breaking any promises with regards to what we will do with it."

Data garage such as NoSQL have several security susceptibilities, which cause confidentiality issues. A conspicuous security blemish is that it is incompetent to encrypt records during the cataloging or logging of data or while dispensing it into diverse groups, when it is streamed or unruffled [2].

Out of all the majority of data warehousing system vulnerabilities being faced by business, common six can be mapped as follows:

a) Pitiable authentication for records
With the cradle of informative data records being flowing in and out of a company's data warehousing unit specially if discussed about Retails Department, and the ease of getting into the database with poor authentication system, it can act as an entry point for any malicious person. Through Rouge scanners, this can open the doors for fabricated transaction, improper rating systems, and so on into the functional system.

This can be mitigated through a granular level of control on the data with 5W's questions to maintain a trail for all the inbound and outbound data flow in the system endpoints along with a context-driven dogma podium for proper role settings.

As articulated in several security seminars, if you let the flow of improperly articulated and unstructured data with garbage-type data security, the same will be

haunting you for the rest of your life span in the same company. Hence, it needs to be ensured that the incoming records are from reliable sources, and that it is not tampered.

b) Apprehensive web consoles

Front-ends act as another security exposure for data warehousing systems. Considerable amount of interaction with data warehousing uses Internet-based web interface, which act as doors for cybercriminals due to their mostly unaddressed security loopholes. Using techniques like eves dropping, data can be captured easily, which flows in and out using Internet and can be altered to complete their causes. These types of scenarios do not depend on the size of data.

Authentication-related issues could also be instigated via techniques like SQL injection, where web front ends with less or no authentication are at high stakes. Stored procedures can be considered as a mandate to safeguard the data along with parameterized queries for data statements as engagement in modus operandi.

c) Rudimentary security controls unavailability

The lack or unavailability of robust security incorporation in major firms acts as another liability in data warehousing systems. Most of the mid-level or small-level firms do not consider security as a part of fundamental design of any solution, which leads to data leakage since several times these firms act as third-party source for any functionality.

Security deployment should be both preemptive and responsive to safeguard their data from getting into malicious hands. Firms should properly scrutinize for the same. Security threats or various other anomalies can be identified using threat scanning, which can be deployed along with perimeter defenses that can scrutinize in real time.

d) Pitiful encryption system clubbed with derisory masking protocols

Masking can also be represented as a small manipulation of data. This is mostly needed to cover the loopholes in the poorly defined encryption algorithm that can occur in the data flow during the integration points of systems.

Another issue with customer data is anonymity. The use of automated technologies like machine learning can lead to uncovering the user's identity via simple derivation or direct accidental visibility. These data need to be anonymized for the security of customer. Conceded concealment is a noteworthy area of data security while still industries contemplate of safekeeping with the value of missing data.

Ensuring that a given syndicate has comprehensive set of algorithms and guidelines in place for masking, and encryption is a major scenario in the current market. In ingestion points, to avert any further security concerns, apt dogmas need to be placed as well.

e) Record's improper lineage and respective audits

Records that are misplaced or scattered can be considered as another epitome of data warehousing system vulnerability. Improper or lost trail data creates an exposure to further details of the customer along with the configurations of security control shielding the data. Mostly, these scenarios can also arise due to records storage in multiple locations like On-premise and Cloud without any proper documentation.

As stated by Security Expert Morrell, security of data warehouse is an integral part for being innocuous. However, one should also take precautionary measure of developmental cybersecurity disputes. This can be done by data tracing with proper documentation. This is also critical for rules of compliance of confidentiality regulatory.

As a continuation of Morrell's statement, every single strand of data and its lineage requires a continuous audit trail to latent complications. It is very important to maintain detailed log of records. Starting from ingestion to usage, and from usage to authentication details and processing details, all should be maintained in a log to avoid missing any unknown threat. Pinpointed authentication information about the people acts as a precautionary measure to backtrack in case of security breach as well as real-time security to avoid any major security breach.

f) Huge dataset

Data warehousing system is already a set with huge data processing in a data warehouse. In such a scenario, data redundancy or data replication are not stringently required for future system, and failure can act as an overhead security concern. Rather securing networking with remote data and securing the same could provide better security.

As stated by Morrell, keeping several replicas of data floating for the system without proper trail cannot be considered as a way to assure data security. Hence, secured system and definitive measures are applied on cloistered data used for analysis to minimize Online surfacing and allow minimum public functions access to it. This eradicating security concerns due to insecure replicas.

4.1.1.4 Precautions needed to be taken

Will the above-mentioned issues and procedures thwart all data warehousing security susceptibilities? Perhaps not. They can only act as a base for a good start but it is rather important that all the firms should consider and focus seriously on data warehousing system security. Ease of data management along with several cost-saving processing methodology is the major reason for expanding popularity from large organizations to smaller and medium sized as well. Data analytics can be further broken down into service of data digging and data collection, which is nowadays being assisted by cloud-based stowing services. Nevertheless, the issue of

confidentiality and other security-related threats is still surrounding the integrated system of data warehousing and cloud-based storage system.

Old or traditional algorithm-based classic security application, which are designed to perform only on a particular volume of data, cannot handle this large volumes and leads to data leakage and other security and confidentiality threat. Dynamic data like that in stock markets that can be considered as of major importance in a nation's economy, also cannot be handled by these traditional security applications. Therefore, just a consistent security check will not be able to detect security blotches for constant streaming data. For this scenario, you need full-time seclusion while data streaming and data warehousing system analysis [3].

i. Security of transactional logs and shielding
Due to unavailability of data trails of their storage location, which is generally a fault of auto-tiering technique, more and more such encounters happen. Even though sensitive records like logs and other transactional records have erratic level of security but without shielding it is all inadequate. Instances like necessitating of auto-tiering system to manage the amplified accessible and scalable data due to huge transmission of data being stimulated by IT executive create an issue that is tough to handle at later stage.

ii. Percolation of input from end-points and validation
For data warehousing systems, front-end or end-point devices are the primary factors for preservation. Input data received from end-point are stowed and treated, and other obligatory tasks are accomplished. This is the major reason for acquiring legitimate end-point that is reliable to the maximum extend.

iii. Framework for fortifying dispersed calculation and procedures
Security arrangements and other fortifications in framework for dispersed calculation like MapReduce (which is a utility in Hadoop) is mostly lacking for computational safekeeping and other digital resources. Major preclusions are shielding the data and safeguarding the mappers in the manifestation of an unauthorized attacker.

iv. On premise of real-time data safeguarding and shielding
Large volumes of data generation were the major reason for most groups that were incompetent to preserve consistent validations. Nevertheless, favorable condition would be to execute security checks and scrutiny in real time or virtually in real time.

v. Encryption and defending access control method communication
To defend data from theft, a fortified data storage device can be considered as a smart step. Yet, encryption is still one important parameter in data stowing devices as vulnerability can arise anytime in any device.

vi. Data provenance
To categorize data facts, it is obligatory to be cognizant of its basis in mandate to regulate the data source precisely; authentication, endorsement, and access regulator could be achieved.

vii. Microlevel assessing and access controls
Low-level authentication control of data warehousing system by databases like NoSQL D.B or the Hadoop Distributed F.S entails a proper and updated validation methodology and obligatory authentication. Scrutinizing diverse records could be expedient and these statistics could be obliging in distinguishing any kind of security breach or other malevolent activity.

4.1.2 The result

Roughley stated that, "We can initiate to aid individuals with the methodology to extract the utmost value from their data, the way to treasure an economical supply of electricity or other energy source, benefits while acquire a bank mortgage, even advance and enhanced medic services using the data shared by their fitness tracker. However, we need to acknowledge the actual point that we have all become entities in the Data Warehousing system with gradually being habituated to data analytics and data sharing."

To put it simply, there are numerous ways that data exposure can be secured, let it be from securing and updating the servers up to date with security servers. Essentially following a virtuous line of disclosure for customer records, and using it for real worth ought to find a solution to the modern present-day problem.

4.2 Hacking

Hacking is an endeavor to abuse a computer system or a remote grid inside a larger network. In simple words, it is the illicit access to or control over computer grid-safety systems for some forbidden drive. The party engaged in hacking deeds are branded as a hacker. These hackers may alter structure or security topographies to achieve an objective that diverges from the original drive of the system. Hacking can also denote to nonmalicious actions, generally concerning scarce or improvised variations to equipment or processes.

Ethical hacking signifies the act of tracing flaws and vulnerabilities of workstation and information engines by duplicating the resolution and activities of malevolent hackers. Ethical hacking is also acknowledged as penetration testing or intrusion testing. An ethical hacker is a security expert who smears his/her hacking

abilities for defensive tenacities on behalf of the possessors of information systems. By piloting penetration tests, an ethical hacker gazes for answer to the following four basic questions like information/locations/systems: can an attacker gain access, can an attacker see on the target, the value of available information to attacker, and is the attempted hack recorded in the targeted system?

Hackers deploy the range of modus operandi for hacking, including
– Spoofing attack: It comprises of sites that fabricate information by imitating legitimate websites, and they are consequently treated as trustworthy sites by users or further programs.
– Vulnerability scanner: These types of programs scan remote computers on grids for known vulnerabilities.
– Password cracking: It can be termed as the method of retrieving passwords from information stockpiled or communicated by computer systems.
– Viruses: These are self-replicating set of codes that spread by injecting replicas of themselves into other executable programs or documents.
– Packet sniffer: These can be defined as those applications that seize data packets with intentions of viewing information and passwords in transit over the networks.
– Root kit: They epitomize a set of code packages that grind to sabotage functionality of an operating system from legitimate operators.
– Trojan horse: It functions as a back-door in a computer system to permit an intruder to achieve access to the system in future.
– Key loggers: These are the tools deliberated to record every keystroke on the infested machine for later retrieval.

Certain organizations hire hackers as part of their upkeep staff. These authentic hackers also recognized as ethical hackers or white hat hackers that use their capabilities to discover faults in the syndicate's security system, thus averting distinctiveness in individuality larceny and further computer-linked delinquencies. White hat hackers are typically perceived as hackers who use their expertise to assist people. They may be rehabilitated black-hat hackers or may merely be well proficient in the procedures and practices used by hackers. An organization can employ these professionals to test and implement best methodologies that make them less susceptible to malicious hacking efforts in the future.

4.2.1 Big data versus ethical hacking

While the syndicates currently are converging on exploring data warehousing system and analytics because of economical stowage, reachability, usability, and conception of distributed computing, they unknowingly also create a prospect for hackers in social engineering as well. A technique wherein the hacker can know the

inclinations and interests of employee in the enterprise that can assist in construct-ing an efficacious social engineering attack. For example: With the predisposed data warehousing system of the employee, the records can be excavated easily, whose sites are frequently logged by the employees and the frequency of stopover to the given site (Facebook, YouTube, etc.). With this information, a naive hyperlink in a spam e-mail can be twisted to disclose not only his individual minutiae but can also be enticed into providing corporate authorizations and thus providing numer-ous accesses to the hacker.

Currently, data warehousing system and networks deliver "Just-in-time" back-ing for governments, syndicates, and officialdoms during crises. It will also protect forthcoming scenario of national and international network security, new proce-dures of sovereignty. It also enriches the thoughtfulness of use, abuse, and net-working of broad topical. These statistics if in mischievous hand can be a base point for taking down an entire region or government off-guard.

Without the encircling span of data warehousing system, it's tough to conceive a scenario where dappled ventures and borderline illegal acts would make news. It's only with the application of data warehousing system does the utter scale of this evi-dence turn heads. If one individual gazes at another individual's sheet during a test, it's commendable of a red mark from the professor. If the whole class cooperates in an organized way and develops a coordination of cheating, it becomes newsworthy. The Panama Papers for illustration are an admirable example of something that is not obligatory illegally, but sketchy to say the least. The element that hundreds of high-profile global figures were acknowledged in this mass dataset is what makes the news. With the evolution of data warehousing system, it makes opportunities for hackers even more appealing, but it also creates a pool of data that becomes even more necessary to protect [4] (Table 4.1).

Table 4.1: Instances of harms versus benefits of data warehousing system.

Scenarios of issue	Pros	Cons
Incursion of cloistered communications	Shared and political engrossment on very enormous scale	Very truncated hurdles to intervention
Unrestricted revelation of anecdotal cloistered specifics	Analytics illuminates to enhanced and well-timed treatments in the healthiness domain/ commercials that might be concerned to you, and so on	Analytics can conjecture peculiar actualities from innocuous feedback data
Tracing, stalking	Location sharing can be used for triangulation, judging shorter routes, proximate allies, even evading natural calamities, and others	A criminal can utilize the info to raid house when empty

With data warehousing system set, hackers may possibly destruct or yank data warehousing system sets with reasonably trivial alterations in instruction to achieve benefit. Certain techniques might be anodyne to the community but hackers might even exploit annual economic corporate reports for individual advantage. Such vicissitudes in monetary reporting models might also alter the policymaking of management, investors, dealers, and further people who build their verdicts on these monetary reports.

Industries like Equifax, which is one of the distinct consumer credit agencies, functions on multibillion-dollar statistics advisor industry, which acts as a perfect example. They decorate an exhaustive depiction of an individual's life and that sketch is utilized to style resolutions with direct impressions. As a corporation swells to its stockpile of data, the worth matures exponentially; so, the imperative of dataset traders is to uninterruptedly hoard as much data as conceivable.

In nearby impending, hackers might have the capability to intrude into workstations that pedal vital technological paraphernalia that regulates water distribution, rail networks, gas distribution, and so on. By gaining admin access of this workstation, hackers can alter the operational configurations or manually construct anarchy. This would have a disparaging consequence. Grounded on the research steered by specialists, a point was established that this was undeniably conceivable. As per reports, events of such potential have not achieved the public news, yet it's a possibility that it could have occurred already.

Thankfully, the good people are keeping up and developing strategies to thwart modern cyberattacks. Let us compare how cyberattacks have traditionally been detected and how data-centric menace revealing system is updating the cybersecurity sphere, leading safekeeping enterprises to design a contextualized and analytical slant to threat recognition system.

4.2.1.1 Scalability and data amalgamation: to detect infringements, you have to validate each piece of data

Customary security incident and evidence management software was not capable enough to accumulate ample and adequate information to perceive up-to-date, erudite infiltrations. Furthermore, although they utilize chronological data, most of them do not have the storage or handling competences to scrutinize data later than 30 days, which leads to overlook significant idiosyncrasies. Additionally, these tools scrutinize diverse cradles of data discretely rather than in conjunction with one another [5–7].

Updated tools that have occurred take into account the speed, size, variety, and complexity of data in a mandate to distinguish the new era of cyberattacks. The fresh paradigm appeals for layering predictive analytics and machine learning systems on cream layer of all cradles of data in an organization's cyberinfrastructure (Figure 4.2).

Figure 4.2: Monitoring UI of HADOOP (data warehousing system scalability and analysis) tool.

4.2.1.2 Well-designed conception is crucial

Pictorial illustrations of infrastructure statistics can assist in making security exposures visible. Conversely, present-day safekeeping mavens are not well proficient in data conception. Stereotypically, their prescribed training includes just statistics, computer science, and security. In such circumstances, if records are detained

across much longer time horizons and from several disparate sources, well-designed visualization becomes indispensable to threat scrutiny.

Companies that use data conceptualization tools have customarily utilized them for post-destruction design and not for real-time threats monitoring. If platforms are integrated and paired with streamlined visualization, users can swiftly and accurately pinpoint system susceptibilities [5, 6].

4.2.1.3 Smaller companies more exposed due to unaffordability of cybersecurity

Traditionally hackers used to hit substantial establishments with comprehensive cyberattacks envisioned to disorder huge number of systems and make headline news. The modern cyberattack, however, has more low-profile outbreak on confidential records with the intent to go undetected. Small-scale corporations are most exposed, as they can't afford to implement and manage tech that traces the footprint of the data warehousing system over the endpoints of their organizations.

The artificial intelligence and human expertise for monitoring are not necessarily boomingly costly, but the hardware for treating of such gigantic volumes of data might be exceedingly exorbitant. Thus, the security tactic should majorly rely on the worth of the chattels that need to be protected [10, 11, 12].

4.2.1.4 Fundamental challenges to combating cybersecurity coercions

Day by day, malware outbreaks intensify in volume and intricacy; they are grim for traditional diagnostic tools and arrangement to tackle them because of majorly two factors: scalability and data density.

For example, each day at Sophos Labs, more than 300,000 new potentially mischievous files require scrutiny, and SQL-dependent infrastructure will not scale well and has high maintenance cost [7, 8].

4.2.1.5 Data warehousing system analytics as a path forward to cybersecurity

Detection of hacking attempts and countermeasures to instantly respond is a major focus area. Prevent, detect, and respond are collectively called as PDR paradigm. This can be considered as the doors where data warehousing system analytics comes in.

Corporations and analytical firms are now confirming that these encounters might probably be overwhelmed with data warehousing system analytics. Investigative corporations have been scripting reports and counseling their patrons about the impressions of data warehousing system analytics on cybersecurity across diligences: For example:

- IDC pinpoints that cloud and data warehousing system will avert cyberthreats to the health organizations.
- According to Gartner, one-fourth of universal corporations has already adopted methodology of data warehousing system processing [13].

4.3 Social engineering

Human user interface being used in various manipulative ways to accomplish mischievous activities with vivid range can be explained as a definition for social engineering. Social engineering, in the milieu of information security, discusses disclosure of confidential info with psychosomatic influence of people. Category comprising assurance tricks for the system access, information congregation tenacity, or deceit diverges from a customary classic "con" in that it is frequently one of several steps in a more intricate fraud structure.

Social engineering can also be explained as concomitant with the social sciences, which are deed of psychosomatic influence of a human, but this recent security vulnerability has surrounded the information security experts since few decades. The cognitive biases are the basis of entire social engineering procedures that are pinpointed on explicit characteristics of human judgment, sporadically referred to as human hardware bugs [14].

Numerous combinations of mix and match are used to generate attack techniques. The assaults cast off for this attack are utilized by the hackers to snip secluded data of the users. There are several examples of this type of attack like the one in which the user is called in their cell phones by people posing as bank employees to fetch their card details for malicious transaction or those in which a mail is sent to the user with a link to click which will in turn take the user to a malicious infected page to load virus into the system, and so on.

Generally, these types of attacks are transcripted with one or several bookmarks. Out of those steps, for a well-defined carefully planned attack, the first step is generally the homework on the victim to collect circumstantial records that range from user's likes/dislikes to chalking out entry points via carefully validating the security guidelines and protocols.

In the next step mostly, the preposterous person tries to gain victim's trust to provide him the stimuli which will trigger the actual attacking actions like breach of security protocols and capturing subtle information.

4.3.1 Lifecycle of an well-organized attack

This can be considered as one of such attacks with no real-time protections except maintaining logs, which will lead to finding out later. It basically acts in manipulation

or other humanoid faults, which let them as insider's access without directly acting upon the intrusion system for the hackers. Since inside attacks cannot be fully predicted based on human manipulation of authentic users, it is more tauter to recognize such attacks, let alone the real-time protection. These types of attacks can only be tackled when users are trained thoroughly with all the modes of attack and how to be precautious for them (Figure 4.3).

Preparing the ground for the attack:
· Identifying the victim(s).
· Gathering background information.
· Selecting attack method(s).

INVESTIGATION

Social engineering life cycle

Deceiving the victim(s) to gain a foothold:
· Engaging the target.
· Spinning a story.
· Taking control of the interaction.

HOOK

EXIT

Closing the interaction, ideally without arousing suspicion:
· Removing all traces of malware.
· Covering tracks.
· Bringing the charade to a natural end.

PLAY

Obtaining the information over a period of time:
· Expanding foothold.
· Executing the attack.
· Disrupting business or/and siphoning data.

Figure 4.3: Lifecycle of a well-organized attack.

4.3.2 Types of social engineering

There are several and diverse methodologies, which along with human collaboration give shape to the attacks. Generally, these can be classified into five broad types of assaults [15].

i. Baiting

As a fish is caught with the help of a bait or a rat catcher uses a bait to trap it, same is this type of attack where the greed or curiosity of the victim is used as a bait to provide a false assurance. This greed or curiosity either lands them right into the deceptive trap compromising their personal information or wide opens their workstation for viruses. The baits are generally having an authentic look, which provides the victim with false assurance. Physical media is the mostly used form to disperse these types of malwares.

Such a scenario like a malware infected flash drive can be considered as an example, which contains the bait suitable for the target user. Due to inquisitiveness,

victim uses the flash drive in any workstation, thereby providing straight route for the malware to infest the system.

These types of attacks are not confined to physical world only; advertisements and other lucrative links to download any software act as a form of online bait. The baits are mostly generalized form and not targeted to any particular user.

ii. Scareware

All the online users have generally seen or faced scenarios where multiple alarms suddenly pop up in the browser or system. Series of fictitious threats are bombarded in the system. This type of attack is termed as scareware, where the victim is threatened in a cyberway to make them believe that their system is compromised and/or is infested with malware. This leads the user to actually download a software suggested by the attacker, which is the real payload for the attacker to compromise the system. So in short, a rogue scanner software or deceptive software that threatens the user to act according to the attacker can be termed as scareware.

Figure 4.4 can be considered as one of the most common scenarios being encountered by almost every Internet user, where popup banners with utmost legitimate looking banners are bombarded in the browser. These popups generally have threatening messages or texts like the one in Figure 4.4. The users are forced to install malicious software or click a link that redirects them to a payload containing site to compromise the system [9].

Figure 4.4: Example of a scareware.

Spam emails with threat and warnings are a mode of operandi for this type of attack, which lures the user to spend on worthless products.

iii. Pretexting

In this type of attack, series of well-planned manipulations are crafted by an invader to acquire information of the victim. The perpetrator often instigates the attack by pretending as someone else to the victim to requisite classified data to accomplish the assignment. All varieties of apposite information and records are congregated utilizing this swindle like as SSNs (Social Security Number) can be considered as input or output for this type of attacks.

In a classic mode, invader kicks off the attack according to the following steps:

Imitates as colleague, law enforcement agency, bank and tax officials, or other entities that under specific circumstances having authority-level access.

Enquires about classified, important but partial information of the victim to avoid major doubts to the victim.

Uses the data received to data mine the rest of the classified and more important data that can harm the victim in a major way.

iv. Phishing

One of the online' s most prominent and prevalent type of manipulation attack dependent on directly reaching the user via mailbox or messaging services can be defined as attack style of phishing. It depends majorly on the human tendency of receiving free services or earnestness or sense of distress. It focuses on a better form of a lie in which subtle info is spit out to the victim to generate the sense of curiosity or urgency, thereby leading them to clicking the malevolent link in the mails or chats, which redirect them to payload pages or attachments.

As shown in Figure 4.5, using an electronic mail false sense of affection or caring is injected in the user along with curiosity of knowing the identification of the source showing the affection. The link that shows a greetings being shared by an unknown user actually leads to a payload-containing website that is to be triggered as soon as the user navigates into the webpage. Once the payload is installed, the user falls on the mercy of predator only.

These types of attacks are generally send in a mass to huge set of receivers, with almost similarity to the original links, and regularly updating the mail servers with information from security platforms can actually help the admins to obstruct these types of attacks.

v. Spear phishing

Since the phishing attack is more generalized and can easily be obstructed, it does not have any specific target. The modified version is also available in manipulation attacks where phishing is specifically directed according to a chosen victim that can be an individual or a member of any large syndicate. They follow the below steps:
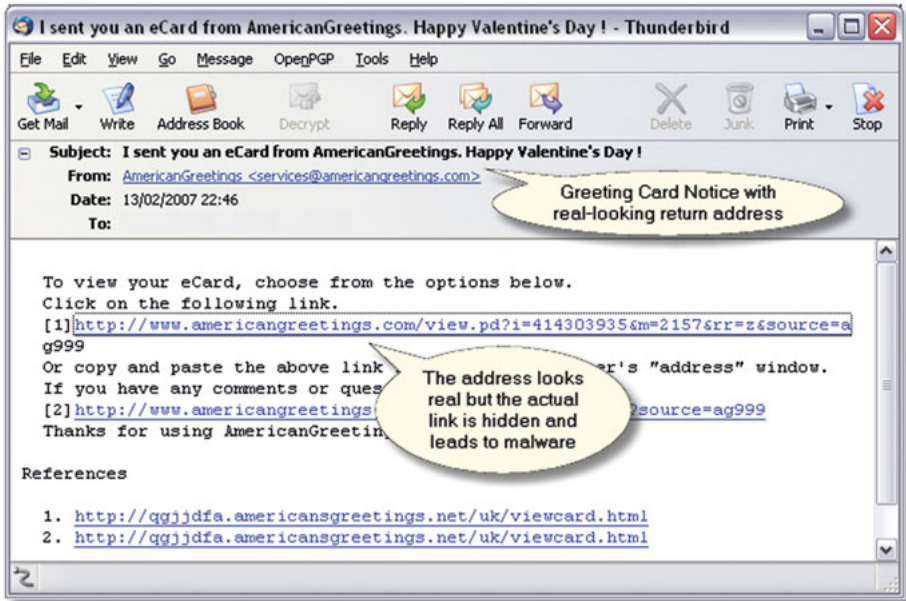
Figure 4.5: Example of phishing email.

Selection of a victim

Datamining more information about the victim like hobbies or interests, and job-related information to make attack less suspiciously.

Closely monitoring the victim to initiate attack in a proper time to attack with maximum success rate.

These types of attacks are generally long duration attacks but are ample tough to sense out and have enhanced triumph rates.

These types of attacks can be visualized as any assailant impersonating as an employee of the same organization as the victim but with higher authority or access to emergency services. After proper background studies and proper timing, a message is delivered by the assailant that are mostly urgent or emergency routine services which needs their authentications or other important details. The information shared by the assailant like victim's supervisor name and all are retrieved by the assailant during the prerequisite data mining, thereby forcing the victim to believe the authenticity of the call and disclosing all classified details or dispatching them via any web link.

4.3.3 Big data versus social engineering

Social schmoozing platforms are groundbreaking platforms because of their role in transitional behavior among users and third parties with their business orientation.

Entities analyze the users' data to operate commercial campaigns and, in lieu, foster the financial development of the platform itself, thus subsidizing to comprehend the visions of Internet pioneers, that is, to cultivate a digital grid where information is free and can be utilized for the well-being and the financial development of the entire humanity.

Not surprisingly, however, data analysis might be easily misused, for instance, by exploiting the detailed information about users toward morally questionable objectives (e.g., tailored persuasion techniques, for which we refer to another post of this blog). In addition, once disclosed to the acquiring party, data are not anymore in possession of the social network and, as such, might be illicitly forwarded to other parties. Given this scenario, we try to briefly explain what are the current capabilities and consequences of such capillary data production and analysis, that is, how much can be done starting from our digital shadow?

Nowadays, the combination of psychology and data analysis is so powerful that 70 likes on Facebook are enough to infer more about a users' personality than what their friends know about him; 300 likes are enough to know that user more than his partner. Hence, online social networks are such privacy-invasive that there is almost a coincidence between the daily life of a person and their digital shadow. Artificial intelligence techniques are the today's state of the art in many data analysis tasks and, while already performing excellently, their growth is not expected to stop [16].

Considering that the Internet is widespread at any level of our lives, with the online social networks acting as a giant magnifying lens on the society, and being particularly suitable to foster the political discussions, the inferences performed on our data should raise serious concerns. Data might be used to profile users, to encounter them in a much-tailored fashion, and consequently, leveraged to induce them doing something they would not do in their own to perform social engineering to the extreme, precisely. The more is known about users, the easier is also to employ persuasion techniques to propose them exactly what they like, or are scared of, thus opening the doors for a plague of our time: the widespread diffusion of fake news, which, in turn, have detrimental effects on the democracy of a country. In fact, a group of attackers with sufficient available resources can spread misconceptions and fake news on a global scale to influence the results of huge events by hacking the voters (which ironically has the same effect of vote rigging!) [9].

Very recently, the case of an alleged misuse of data carried out by a company operating in the marketing sector, named Cambridge Analytica, came under the spotlight of the media. It is a case worth discussing because it embodies much of the issues described throughout this post. First, some details about the fact: Cambridge Analytica is accused to have been involved in an illicit sharing of data with Aleksandr Kogan, a researcher who developed a Facebook-based application to gather information about users' personalities [17, 18]. Before 2014, Facebook's rules about data sharing were not as stricter as they are now. Specifically, a user allowing to disclose some of his/her data had also the capability to reveal pieces of

his friends' information. In this way, from the 270K users who deliberately shared their data with the application, it had been possible to profile up to 50 million American electors. With such information in hands, Cambridge Analytica is accused to have performed microtargeting campaigns to favor the election of Donald Trump, by employing unscrupulous means, such as the spread of fake news to create a significant shift in public opinion (Figure 4.6).
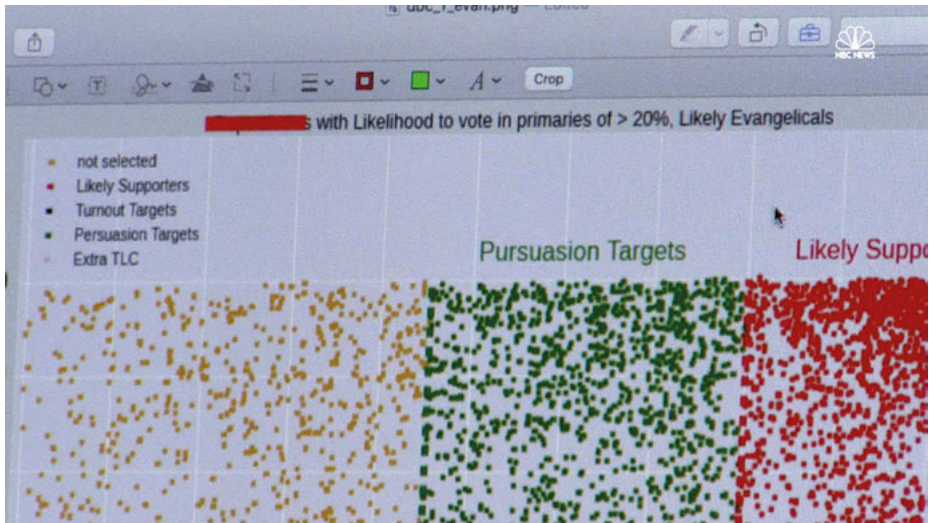
Figure 4.6: Sample of Cambridge Analytica analysis report.

In our view, four main lessons should be learnt from this story:

Today's data-driven business models come at the cost of sacrificing privacy and require a high level of trust on the entities managing our data. Once data have been disclosed, in fact, there is no guarantee that the party that is entitled to use them (e.g., the legitimate application) illegally forward them to other entities or not.

Although rules are mostly imposed to limit the control that users have on their friends' information (as Facebook did in 2014), the issue is inherently present in online social networks, since they are based on the friends/followers paradigm. Due to this model, in fact, the boundaries among users' information spaces have become blurred. Just think of a picture where a user is inadvertently tagged. Moreover, it has been shown that a target user's information (e.g., location) could be accurately inferred from the analysis of the profiles of his friends.

Social engineering benefits from the heterogeneity and volume of the available data, and widely employs persuasion techniques. The data-centric and all-interconnected world we live in represents the favorable scenario for the application of an extreme social engineering, that is, people can be easily profiled, contacted, and

deceived to induce effects that go far beyond the traditional industrial espionage. As a matter of fact, social engineering has the potential to spread ideologies and influence the result of huge political events by exploiting the structure of the democracy itself.

The Duolingo case, as explained in our project also, is an excellent example of how tracking of people's behavior on a large scale and inferring their behavioral habits is one of the solutions to improve the efficiency not only of the attack patterns, but also of the training systems.

4.4 Conclusion

Data warehousing system analytics is a major boom in current cyber industry. Data warehousing system analytics if used correctly helps in identifying, understanding customers, optimizing according to their needs, science and research, military, and other defense applications. Data warehousing system analytics can help identifying illegal or hacking attempts even from minute data availability. However, on the contrary, data warehousing system can also be used in corporate espionage, spying on people and even alter their decisions (e.g., U.S Elections) and with the rise in social networking applications every details on every individual can be considered to be achieved online in some way or the other.

Due to the above factor, data warehousing system security is one of the major concerns in cyberindustry. As described by Einstein on the context of nuclear energy, tool that can provide major and sustainable development can also be the cradle of foremost devastations. Data warehousing system security can be considered as important in current cyber market.

The data warehousing system, which primarily meant 3V's now, has been updated to 6V's, that is, volume, value, variability, velocity, variety, and veracity. Data warehousing system analytics and the related security measures are growing every day and in this chapter an insight has been given for the same. With continuous growth in data volumes and improvement and inclusion of new tools in the market for analyzing the same, in future, data warehousing system security needs to be revamped every single moment along with other methods to identify the hacking attempts as well.

References

[1] Bertino, Elisa "Big Data – Security and Privacy", 2015 IEEE International Congress on Big Data, (2015), doi: 10.1109/BigDataCongress.2015.126.
[2] Moreno, Julio, Serrano, Manuel A., & Fernández-Medina, Eduardo "Main Issues in Big Data Security", Alarcos Research Group, University of Castilla-La Mancha, 2016.
[3] Bertino, E., Jonker, W., & Pektovic, M. "Data Security – Challenges and Research Opportunities", SDM, 2013.

[4]   Toshniwal, Raghav, Dastidar, Kanishka Ghosh, & Nath, Asoke "Big Data Security Issues and Challenges". International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2016, ISSN: 2349–2163.

[5]   Chen, M. et al. "Big Data: A Survey". Mobile Networks and Applications, 19(2), 171–209, Jan. 2014.

[6]   Mayer-Schönberger, Viktor, & Cukier, Kenneth Big Data: A Revolution that Will Transform How We Live. Work, and Think. Houghton Mifflin Harcourt., 2013. ISBN 9781299903029. OCLC 828620988.

[7]   Paulet, R., Kaosar, Md. G., Yi, X., & Bertino, E. Privacy-Preserving and Content-Protecting Location Based Queries IEEE Transactions on Knowledge and Data Engineering, 2014, 26(5), 1200–1210.

[8]   Ongsulee, Pariwat, Chotchaung, Veena, Bamrungsi, Eak, & Rodcheewit, Thanaporn "Big Data, Predictive Analytics and Machine Learning", 2018 16th International Conference on ICT and Knowledge Engineering(ICT&KE), 2018.

[9]   Kappler, Karolin, Schrape, Jan-Felix, Ulbricht, Lena, & Weyer, Johannes "Societal Implications of Big Data", (2018). KI – Künstliche Intelligenz. 32(1), Springer. doi:10.1007/s13218-017-0520-x.

[10] Peter Kinnaird, Inbal Talgam-Cohen, eds. "Big Data". XRDS: Crossroads. The ACM Magazine for Students. 2012, 19(1), Association for Computing Machinery. ISSN 1528–4980. OCLC 779657714.

[11] Jagadish, H.V. et al. "Challenges and Opportunities with Big Data", 2011, [online] Available: http://docs.lib.purdue.edu/cctech/1/.

[12] Leskovec, Jure, Rajaraman, Anand, & Ullman, Jeffrey D. Mining of massive datasets. Cambridge University Press, (2014). ISBN 9781107077232. OCLC 888463433.

[13] Press, Gil. (9 May 2013). "A Very Short History of Big Data". forbes.com. Jersey City, NJ: Forbes Magazine. Retrieved 17 September 2016.

[14] Carminati, B., Ferrari, E., & Viviani, M. "Security and Trust in Online Social Networks", Morgan & Claypoo, 2013.

[15] Bag, Monark, & Singh, Vrijendra. (2012) "A Comprehensive Study of Social Engineering Based Attacks in India to Develop a Conceptual Model", DOI: 10.11591/ijins.v1i2.426

[16] Andrew McAfee & Erik Brynjolfsson "Big Data: The Management Revolution". hbr.org. Harvard Business Review.

[17] O'Neil, Cathy (2017). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Broadway Books. ISBN 978-0553418835.

[18] Batini, C., & Scannapieco, M. "Data Quality: Concepts Methodologies and Techniques", 2006.

[19] Breur, Tom "Statistical Power Analysis and the contemporary "crisis" in social sciences". Journal of Marketing Analytics, July 2016, 4(2–3), 61–65. doi:10.1057/s41270-016-0001-3. ISSN 2050-3318.

[20] Sh. Hajirahimova, Makrufa Sciences, Institute of Information Technology of Azerbaijan National Academy of; str., B. Vahabzade; Baku; AZ1141; Azerbaijan; Aliyeva, Aybeniz S. "About Big Data Measurement Methodologies and Indicators". International Journal of Modern Education and Computer Science. 9 (10), 1–9. doi:10.5815/ijmecs.2017.10.01.

[21] [online] Available: https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

[22] [online] Available: https://www.dataspace.com/big-data-applications/big-data-helps-detect-hacking/

[23] [online] Available: https://www.cnbc.com/2016/03/09/the-next-big-threat-in-hacking–data-sabotage.html

[24] [online] Available: https://www.csoonline.com/article/2130877/data-breach/the-biggest-data-breaches-of-the-21st-century.html

Srilekha Mukherjee, Goutam Sanyal

5 Steganography, the widely used name for data hiding

Abstract: Nowadays, global communication has no bounds. More information is being exchanged over some public channels that serve to be an important mode of communication. Without them, the field of technology seems to collapse. But awfully, these communications often turn out to be fatal in terms of preserving the sensitivity of vulnerable data. Unwanted sources hinder the privacy of the communication and may even temper with such data. The importance of security is thus gradually increasing in terms of all aspects of protecting the privacy of sensitive data. Various concepts of data hiding are hence into much progress. Cryptography is one such concept, and the others being watermarking and so on. But to protect the complete data content with some seamlessness, we incorporate concepts of steganography. It provides complete invisibility to any sensitive data that is being communicated. This prevents attracting unwanted attention from third-party sources, which helps to some extent with information safety. The field of big data is quite into fame these days as they deal with complex and large datasets. Steganographic methodologies may be used for the purpose of enhancing the security of big data since they also find ways of doing so.

Keywords: Steganography, Cryptography, Mean Squared Error, Peak Signal-to-Noise Ratio, Entropy

5.1 Introduction

The worldwide booming technological [1] trends bring along with it several disincentives, and overcoming that has become a challenging task. In these days modern research [2] has given a new dimension to almost all fields of technology. The new technologies have stressed on mediums being digital [3]. Communication [4] is the sole mediator between any sorts of technology that offers a service to mankind. Modern technologies are indeed a boon for the human race in many ways. Global network of computers serves as an error-free mediator in the source to destination delivery of any data/document [5]. Research is carried out at a rigorous level, and hundreds and thousands of newer techniques are coming up so as to sort complex day-to-day problems [6]. Technology has made life simpler and easier. But as it is said there is no rose without a thorn, so goes the way with these

Srilekha Mukherjee, Goutam Sanyal, National Institute of Technology, Durgapur, India.

new technologies. It is often necessary to transmit essential and confidential information globally. They face many problems. Their original goal and progress is often hampered by few things, and security [7] is one such thing. Internet also acts as a major source or base to numerous fields that want to endow a critical grip while bracing communication. Information is being communicated or exchanged over the same.

We know that for technologies to reach and serve mankind, they are always braced with communication. Without communication no technology can reach any life. In these days, security is an extremely important issue [8] that has taken over the attention and concentration of many. Of course there are enough valid reasons of being so. The threat of data piracy has been a high risk factor for quite a long time. The mass reproduction of information is also a huge problem. This leads to security concerns [9], which became a major issue in communication. Modern communication requires certain measures to repulse the attention of any third party, which affirms immense secrecy along with full confidentiality of information. The field of data hiding [10] hence came to their rescue.

A communication is a two-way [11] process. The first way is where it is being packed and sends from the sender's side. The second is where the receiver receives and unpacks it. Now the receiver [12] is expected to receive the same exact package what the sender has actually packed. No problem arises if the aforesaid is the case and the receiver receives the same thing that was send by the sender. But the problem arises if the receiver does not receive the same packed thing that was send from the sender's side. Here lies the main disadvantage and flaw. Some third-party attacks [13] are being made in the way of these communications. These third parties tend to temper with the technological benefits from being achieved. The full-fledged benefits of technologies are made to be jammed. Hence, the purpose is not achieved or fulfilled. Therefore, it is extremely important to protect [14] the need of originality for the information being communicated.

5.2 Security of information

In today's highly dynamic and competitive world, information security serves to be the actual fuel for the engine of global communication. Information can be defined as processed data or facts. In order to safeguard its originality, several processing [15] has to be made before communication. The main motive in these cases is to communicate privately [16], although it is actually done via a public communicator medium.

These days, the concept of covert communication is increasing at an alarming rate. The safe end-to-end [17] delivery of data is the prime issue of concern in information security. Also, this has to be ensured without any interference. For the rest of the world, a seemingly unimportant and transparent communication takes place.

Cryptography [18] is a well-known technique of data protection. In this technique, any data or information to be transmitted is first encoded, that is, it is converted to an encrypted [19] form. The main aim of cryptography lies in securing these contents in a possible way. This can be done with the help of some secret keys. Several policies of encryption, which completely facilitates data security, are actually effectuated. This objective is achieved by transforming the original data into an illegible form that cannot be understood by any eavesdropper [20] sitting in middle. By encrypted form, it is meant that it can be portrayed as any noisy form of the information. Any other form apart from the original one can be pointed as an encoded one. The output is actually an illegible one. It cannot be read or understood by any outsider. The original information is known as plain text [21] and the encoded information is known as cipher text [22]. Cryptography may be further classified as symmetric key cryptography and asymmetric key cryptography. The denoting difference is pointed out with the help of the keys used for encoding and decoding the information.

Shown next in Figure 5.1 are the types of cryptography that are used to encrypt information into some illegible form.

Plaintext ⟶ Ciphertext ⟶ Plaintext

A) Secret key (symmetric) cryptography. SKC uses a single key for both encryption and decryption.

Plaintext ⟶ Ciphertext ⟶ Plaintext

B) Public key (asymmetric) cryptography. PKC uses two keys, one for encryption and the other for decryption.

Hash function

Plaintext ⟶ Ciphertext

C) Hash function (one-way cryptography). Hash functions have no key since the plaintext is not recoverable from the ciphertext.

Figure 5.1: Types of cryptography.

– *Secret key cryptography*
  These methods use a single key [23] for both encryption and decryption. The sender first uses this key for encrypting the necessary plaintext and generates the cipher text. The generated cipher text is then send to the concerned

receiver. On receiving it, the receiver again uses the same key for decrypting the secret message. Thus, the receiver finally recovers the corresponding plaintext. A single key is applied for both encryption and decryption functions; therefore, secret key cryptography is again called as symmetric encryption. Basically, these methods may be applied and used for provision of confidentiality and privacy [24].

– *Public key cryptography*

These methods apply a single key for encryption purpose and one another for the purpose of decryption. Therefore, these methods are also known as asymmetric encryption/decryption. They can primarily be used for the purposes of authentication [25], key exchange, as well as nonrepudiation.

– *Hash functions*

Hash functions [26] are also known as message digests. They have one-way permitted encryption. Mainly these are algorithms, which generally use no key. Just a hash value of fixed length is estimated. This estimation is based on the plaintext. Also, this makes it nearly impossible for either the length of the plaintext or the contents to be recovered. The hash functions are most commonly used and applied by many operating systems for encrypting the entered passwords. Further, the hash functions provide a different mechanism to assure the integrity of a specific file.

Mainly, different methods or techniques of cryptography have been created for securing the secrecy of messages by encryption as well as decryption of data. The encrypted form of any data may often attract the attention of the third-party external sources. This makes it inquisitive for any unwanted external source and might intercept that this form may contain some hidden precious information from a source. This feature of cryptography may actually provoke any unintended third party about the covert communication that is being taking place. Hence, there arises an obvious requirement to hide any form of encryption made to protect and secure sensitive data.

Steganography [27] solves this issue by completely masking the information without making it visible to the outer sources. In order to avoid drawing any kind of suspicion, it has its methods to make some changes in the structures of the host so that it is not identifiable by any human eye. Therefore, the transmission of the hidden data is made in a completely undetectable manner. The communication in this case is completely kept hidden. It is a skill of hiding any confidential information within another media/entity such that nothing unusual appears in front of external sources. It hides the contents of data/any information within a carrier [28] medium, thus facilitating a seemingly invisible communication. Third parties are actually not able to see the information that is being communicated. Only the sender and receiver sides know and are aware about the secret communication being taking place. This particular advantage of steganography has increased its usage to a much higher level. It has given a new dimension to the concept of information

security. The safety and integrity of sensitive data is guaranteed. All the fields and sectors have started using techniques that safeguards individual safety and security. Due to its immense potential of secured connectivity, it has become widespread. Therefore, the concepts of steganography are having huge demands in today's world. It facilitates privacy for several legitimate purposes during communication. Third parties are actually not able to see the information that is being communicated. Only the sender and receiver sides know and are aware about the secret communication being taking place. More communications takes place electronically in these days [29]. Likewise, for steganographic communications to take place, multimedia signals [30] are mostly chosen as renowned message carriers necessary for secured communication. There are many techniques that are figured out after high-quality researches.

Another technique of watermarking [31] also has high usage in the field of information security. The main advantage in this case is to confirm the authenticity of any original data. Also, they may or may not be hidden in the host data. The watermark [32] is hidden within the host data in such a way that it possibly can never be removed. Even if its removal is made possible that can only be done at the cost of demeaning the concerned host data medium. Several watermarking applications, for example, copyright protection or source authentication may have an active adversary [33]. These stated groups may participate in making several attempts that removes, forges, or invalidates the embedded watermarks. Special inks have been used for hiding messages in the currencies as well. Steganography has its main goal of secure communication intact. The controlling factor is that the people are not by any chance aware of the presence of any hidden messages. This is what distinguishes steganography from any other forms of data hiding or information security.

5.3 Steganography

Steganography might be defined as the science and art of hiding data or information within another information, which appears to be harmless. The specific word "steganography" is a mere combination of two different Greek words, that is, "steganos" and "graphein," which means "covered" and "writing," respectively. The sensitive message can be hidden within a selected carrier known as the cover medium. This cover with the hidden data within is known as stego. The cover object serves to be any kind of medium within which any private message might be successfully embedded. This also aids to hide the presence of the very secret message, which is being sent. Referring to an image as a medium, we may say that the cover image is the seemingly unimportant image, within which the actual confidential image is to be embedded. On the other hand, the stego-image serves to be a carrier for communicating the private image across.

5.3.1 History of steganography

Right from the ancient days [34], the concept of steganography had been used. The ancient kings and rulers used many techniques for data hiding. One was shaving the head of a trusted slave and then writing the message on his scalp. Once the hair grew back, he was sent to the corresponding recipient with that message. The recipient king shaved his head to read the message or information. A Greek historian Herodotus mentions a remarkable history related to this. Histiaeus, the chief of Miletus (an ancient Greek city), had sent a secret message to his concerning vassal, Aristagoras, the leader of Miletus, by shaving the head of one of his trusted servants. He then marked the secret message on the shaved scalp [35] and had sent him on his advised way when the hair on his scalp had regrown. This was one of the many techniques of how communication was made during those days.

Demaratus, the king of Sparta, from 510 until 491 BC had used this strategy to sent an anticipated warning for a forthcoming attack to Greece, by inscribing it directly on the underlying wooden support of some wax tablet. The final covering step was applying and smoothening its beeswax covered surface. Also, these wax-covered tablets were commonly used at that time as popular reusable writing surfaces. Even quite for some time, they were used for shorthand purposes.

Mary, Queen of Scotland, used to hide several letters with the combination of some techniques of cryptography and steganography. She had her secret letters hidden in a bunghole of some beer barrel that could freely pass in as well as out of her concerned prison cell. During World War II, another steganographic method that was practiced by the United States Marines was mainly the use of Navajo "code talkers." They applied a kind of simple cryptographic technique and the messages used to be sent in all very clear text.

The vast uses of steganography were simply not limited to mere writing materials. The ancient use of large geoglyphs of the known Nazca lines in Peru can also be considered as a said form of the steganography. The figures vary in actual complexity. These geoglyphs are open to view, though most of them were not identified/detected until they were viewed directly from the above air. The designs are mainly shallow lines, which were made in the ground. It was done by removing the naturally existing reddish pebbles as well as uncovering the whitish or grayish ground underneath. Scholars have different opinion in interpreting their purposes. Moreover, in general, they accredit some sort of religious significance to those.

Another description of a human vector example does include writing secret messages on textures of silk. Later, this would be compressed and converted into one ball. A final covering with wax was the last step. The messenger then had to swallow this wax ball. In this case, the method for retrieving the secret message was not described in the sources.

Another example of steganography is the one that involves some specific use of the Cardano grille. Named after its very creator, Girolamo Cardano, this device

can be considered to be as simple as a sample of paper with some holes made in it. The intended message can only be retrieved, when this grille is placed over some printed text. Such techniques might be related to the stated Cardano grille, which employs classical steganography techniques including methods of pin punctures in any text materials (e.g., newspapers) or overwriting some printed texts with any pencil.

Some evidences support that prior to the stated Civil War, there were certain methods of providing private messages to captured slaves that aided in their own escape. The quilts, which were mostly left to dry by hanging them from window-sills, were used as the target source. Secret messages were passed to the captivated slaves by some sort of patterns made in the quilts. This guided them in their venture for freedom. We may consider the Bear Paw symbol as an example of one such said quilt pattern. This represented an advice given to follow the found bear tracks over some of the region of mountains.

Some of the other uses of the stenographic techniques involve one photograph of the few captured crew members of the U.S.S. Pueblo. There all the crewmembers had spelled the same word "snowjob" and this was done using various hand positions. During the Vietnam era, some instances were found where during photo ops, the captured members in the U.S. Armed Forces would also use several hand gestures. This was often just to make these gestures aired by the media. The techniques mostly employed were by using their eyelids to blink some hints in Morse code, for example, torture. Also, the prisoners of the ill-famed Hanoi Hilton use to have a "tap code" for communicating among each other. This code was mainly based on a 5×5 matrix, where each of the residing letters is assigned a tap sequence. The sequence is purely based on the stated matrix. The spaces or pauses between those characters were twice as much long as the gaps or spaces in those particular letter codes.

There are many other examples from history, which relates to the same purpose as well. Use of invisible inks that glows when heated was one such example. Communication through microdots was one technique used during the World War days. There were certain other techniques as well.

5.3.2 Modern steganography

Digital communication is the boon of the trending technology. With the progression in this field of digital communication, the need of steganography serves to be a backbone of global communication. The need of security for the information traversed being the prime concern increases the demand of steganography. The sole reason was to secure and hide sensitive data from everyone except its intended recipient. This was the built-in feature of the domain steganography. It can actually guarantee the covert communication. This heightens the application range of steganography to an enormous extent.

In the recent years, the global interest followed by research and development in this field has almost inflated to a high level. The presence of redundancy [36] in some of the representations of digital media (used as carrier or cover) is the targeted areas of data hiding in steganography. It attracted the attention of many researchers and developers, who decided to generate newer techniques of availing and sustaining covert communication [37]. During the communication stage, any unauthenticated people may only notice the transmission of a seemingly unimportant image.

A communication always takes place between two parties. One is the sender party and the other being the receiver party. Similarly in steganography, two processes take place: one in the sender side known as sender phase and the other at the receiver side, also known as receiver phase. Robustness as well as transmission security during communication are extremely essential for transmitting the vital message to its intended sources while declining access to unauthorized people. Hence, a secret point-to-point communication between the two trusted parties in the two sides should be ensured.

The communication channel is a public medium where any kind of untrusted source may be present. Their main aim might be mainly to uncover any secret data that passes by. Steganography evaluates and generates a number of ways until the attacker does not find some way to detect and trace the hidden information. With the communicating channel being selected, the communication proceeds with sending the stego. Many new techniques came up for the purpose of hiding information.

In these days communication has become digital. Therefore, the techniques used are digital steganographic techniques. A steganographic procedure has two phases: one taking place at the sender side and the other at receiver side. On the sender's side, the sender embeds the message within a chosen cover medium. On other hand, on the very receiver's side, the receiver extracts the hidden message from the received stego. The resemblance of the stego with its respective cover represents the efficiency of the procedure used. Also the efficiency of the algorithm lies in extracting the hidden information in a lossless manner. The lossy [38] extraction results in loss of data fields from the hidden information. This is definitely not what is expected out of a steganographic procedure, whose main aim is to communicate data secretly from sender to receiver. Therefore, if there is even a partial loss of hidden data, then the procedure is not fully efficient to what it promises.

5.3.3 Benefits of steganography

The primary demand of data integrity and authentication leads to absorption of certain effective measures in the respective systems. Government organizations have a wide range of use in this area. Various purposes of individual interests are some other important factors using the same. The self-conscience level of the modern crowd regarding the security attacks has increased a lot. People have become much more aware regarding protection of their personal and professional data. This self-

awareness has led to an increase in the use of enhanced security in the communication systems. Even several trade and business purposes make use of the potential of steganography to communicate new product launching information or any other trade secrets.

5.3.4 The major challenges of effective steganography

The major challenges of effective steganography should be successfully met so as to achieve a potently secured communication. These requirements are for enabling of a secured communication. Therefore, for a good steganographic system, the following parameters should be significant.

- *Robustness and security:*
  The term robustness [39] refers to resisting the attacks and securing the contents of the hidden data in a possible way. Thus, robustness is an important and challenging factor in any effective steganographic system. Also, for the facilitation of security, the data hidden must appear to be invisible to the external world. This concept of invisibility helps in achieving seamlessness in any carrier.
- *Size of payload:*
  By the size of payload [40], we can say that this term is the total possible amount of secret data/message that can be hidden in any carrier or cover. The steganographic approach focuses in transmitting huge amounts of information maintaining the factor of imperceptibility. This amount of information hidden in any carrier should be maintained without breaking any of the other requirements (such as robustness and invisibility). Since steganography promotes hidden communication and therefore the requirements for higher payload along with secured communication are sometimes contradictory.

Note: It is always not possible to maximize capacity simultaneously with the trait of imperceptibility and improve robustness in any data hiding scheme. Henceforth, an acceptable balance of the above stated parameters must be sorted based on the necessary goal/application. For example, some steganographic schemes may forgo a bit of robustness in the favor of capacity with low perceptibility. On the other hand, a watermarking scheme, for which large capacity and low perceptibility is not a requisite, would definitely promote high robustness. Since the prime aim of steganography is hiding data, so the methods must promote sufficient capacity.

5.3.5 Types of steganography

Then, in Figure 5.2 the types [41] of steganography are based on the medium in which data is hidden. For example, if data is hidden in any text file, then it is text

steganography [42]. If the cover medium is image, then it is image steganography and so on.
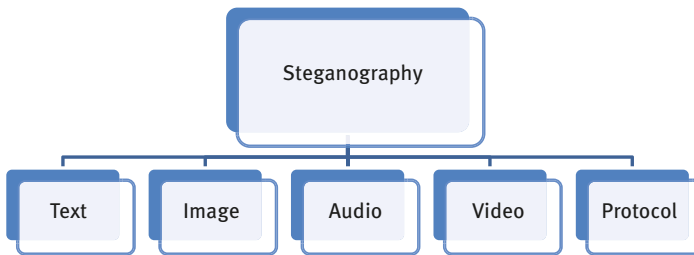

Figure 5.2: Types of steganography.

All these types are used in their relevant mediums, where hiding will be of utmost important in such medium. Various sectors of public importance emphasize in information hiding in the respective required medium, as per their necessity.

5.3.6 Steganography using image as a medium

Due to some subsistence of restricted potential of our human visual system [43], concealing information within digital images is asserted to be quite an efficient medium. Image steganography is also considered to be a potential for facilitating a secured communication globally.

5.3.6.1 Types of image files

There are two primary types of image files [44]: raster and vector (as specified in Figure 5.3). Any image may be catalogued in terms of either vector or any raster graphics. The image preserved in raster form is often said to be a bitmap. We may also define an image map as a file containing some information, which associates different locations on some given specified image with their hypertext links.
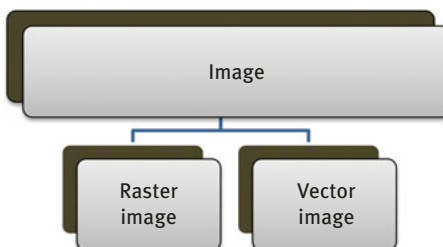
Figure 5.3: Two primary categories of image files.

– *Raster image*

Raster images (Figure 5.4) are made up of collection of dots called pixels. These are generally more common (like PNG, JPG, GIF, etc.) and are widely used over web. Each pixel is specified as a tiny colored square. Suppose we zoom in to any raster image, we may see a lot of these little tiny squares. Raster images are the ones that are created with pixel-based programs. They may also be captured with a camera or scanner. When an image is scanned, it is converted to a collection of pixels, which we call a raster image.
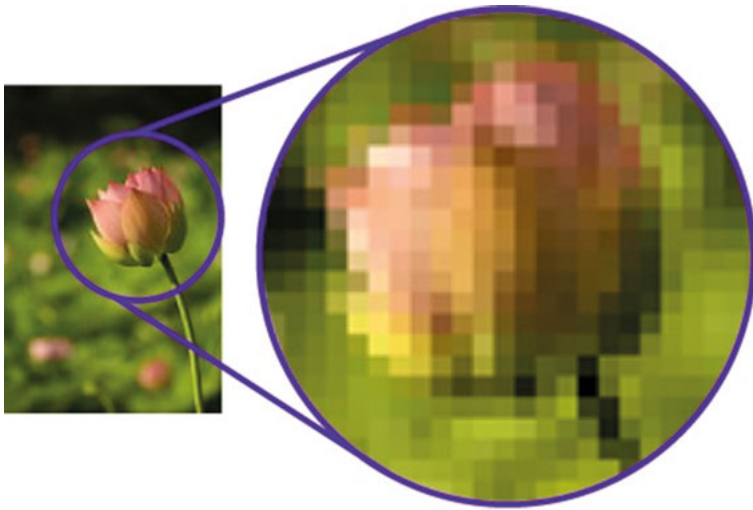
Figure 5.4: A raster image.

– *Vector image*

A vector image (Figure 5.5) is specified as one of the two major image file types. Vector graphics are those that are created with any vector software. These are more common for image files, which are applied onto any physical product. All vector images are object oriented while raster ones are pixel oriented. Since the vector graphics are not formed of pixels; therefore, they are resolution independent. Also, the vector shapes (called as objects) may be printed as large and at that highest resolution what the printer or output device allows. They always maintain all their details when zoomed in or out.

5.3.6.2 Pixel

A pixel [45] is denoted as a physical point present in a raster image. It is actually the smallest addressable element in any display device.
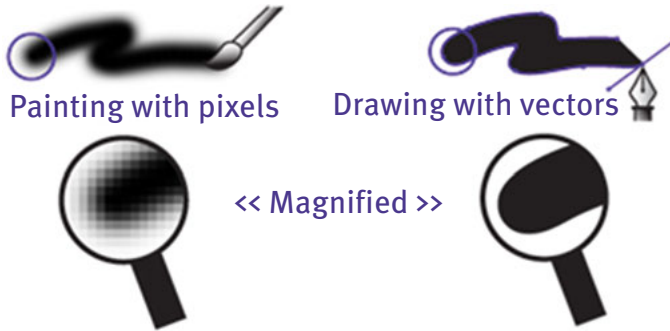
Figure 5.5: Rasters and vectors.

For example, in a 512 × 512 image that has 512 pixels considering side to side and 512 considering top to bottom has a total of 512 × 512 = 262,144 pixels.

5.3.6.3 Types of image steganography

There are several types of steganographic techniques that efficiently hide data. Broadly, it is categorized into two types of domain: spatial [46] and transform. Figure 5.6 shows few categories from both the spatial and transform [47] domain techniques.

Figure 5.6: Domains in steganography.

In the field of spatial domain, the secret data bits/messages are embedded directly into the cover bit planes. The least significant cover bits get directly replaced with the specific bits of secret message. There are a wide variety of procedures that use spatial domain techniques, like that of least significant bit [48] and pixel value differencing [49]. They are efficient in terms of several aspects like that of maximum data carrying c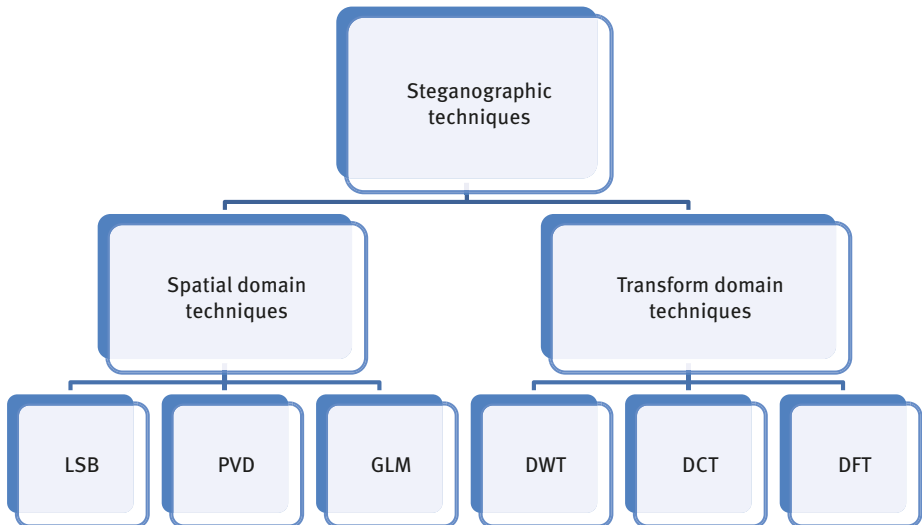apacity. In the transform domain, the secret message is embedded in their respectively transformed cover. There are a number of efficient transform domain techniques, like discrete wavelet transformation [50], discrete cosine transformation [51], and discrete Fourier transformation [52].

5.3.6.4 Some analytic metrics

There are several benchmark metrics based on which the efficiency of the constructed steganographic procedure may be found out. Accordingly, its strength can be determined. Given below are few such metrics, whose values may be computed and the efficiency of the steganographic output or rather stego may be found.

- *Payload* is stated as the total data-carrying capacity of that specific image, referred to as the carrier host or cover. The carried one is termed as the confidential or secret data. This carrier or confidential object might be any file, including some text, image, audio, and video. Henceforth, the embedding capacity of any cover or host image is the maximum capacity, which may denote that only on crossing this particular point distortion is recorded.

- *Mean squared error or MSE:* This is in correspondence to that expected value of the obtained squared error loss or the obtained quadratic loss. MSE [53] is actually a risk function. We may say that the difference occurs mainly due to the presence of randomness or may be the estimator do not account for any information that might generate a more accurate estimation. Thus, it supposedly incorporates both the variance of estimator along with its respective bias.

    If we consider a cover image "CI" that has M by N pixels such that $M==N$ and a stego-image "SI(i,j)," which is obtained after hiding data within "CI," the MSE is found out as

$$\text{MSE} = \frac{1}{(M^{\star}N)} \sum_{i=1}^{M} \sum_{j=1}^{N} [\text{CI}(ij) - \text{SI}(ij)]^2 \tag{5.1}$$

- *Peak signal-to-noise ratio (PSNR):* While the MSE represents the obtained cumulative squared error measured between the images, the PSNR [54] represents a specific measure of the existing peak error between the stego and original. Now, for color images (i.e., having three RGB component values per pixel), the PSNR definition is same. It's just that the MSE is now calculated as the sum of

all the squared value differences, which is divided by the respective image size and also by three. It is formulated as

$$\text{PSNR} = \frac{10\log_{10}\left(255^2\right)}{\text{MSE}}\,\text{db} \qquad (5.2)$$

– *Structural similarity index measure (SSIM):* In general, SSIM [55] is considered to be a full reference metric. This signifies that the measure of any image quality depends on an uncompressed initial or rather distortion-free image, which is used as a reference. Now, we may say that this structural information is the primary idea that the resident pixels have very strong interdependencies. Also this is especially the case when they are close spatially. These kinds of dependencies always carry some relevant and important information related to the very structure of objects in their visual scene. It is calculated as

$$\text{SSIM}(c, s) = \frac{(2\mu_c\mu_s + v1)(2\sigma_{cs} + v2)}{(\mu_c{}^2 + \mu_s{}^2 + v1)(\sigma_c{}^2 + \sigma_s{}^2 + v2)} \qquad (5.3)$$

where μ_c is the mean of cover, μ_s is the mean of stego, $\sigma_c{}^2$ is the variance of cover, $\sigma_s{}^2$ is the variance of stego, and σ_{cs} is covariance of cover as well as stego.

– *Mean and standard deviation:* In probability and statistics, the mean [56] and expected value synonymously refer to one particular measure of the central tendency. Also, this pertains to either of a probability distribution or random variable, which is characterized by the specific distribution. Also, we may say that the standard deviation [57] or "σ" is the measured square root of the variance of "X." Rather it is actually the square root of the estimated average value of "$(X - \mu)^2$."

– *Entropy:* Here, the term "message" may stand for any event, character, or sample, which is drawn from specific distribution of data stream. Thus "entropy" [58] is known to characterize our uncertainty regarding the source of information. Since entropy is greater for more random sources, so it is understood as the measure of uncertainty instead of certainty. Entropy may be defined in some context of any probabilistic model. We may say that independent fair coin flips always have some entropy of 1 bit each per flip. Considering a source generating one long string of character B's, it will always have entropy of 0. This is because the next character here will always be a "B." The next following points are to be considered:

  – The net amount of existing entropy might not always be some integer number of the bits.
  – Some data bits might even not convey any information. As an example, some data structures sometimes redundantly store information. Also regardless of the specific information residing in the very data structure, they might have identical sections. Given any data source, it gives the average of the bits that are needed for encoding it.

– *Skewness and kurtosis*: In case of any nonparametric skew, we can define it as $(\mu - v)/\sigma$. Here, μ is its mean, v is its median, and σ is its standard deviation. In cases where distribution remains symmetric, the mean becomes equal to median. Thus, such distribution will always have zero skewness [59]. Kurtosis [60] determines how sharp and tall the central peak could be when it is relative to some standard bell curve. Also, there are several interpretations for kurtosis, along with those how certain measures have to be interpreted. The primary measures are tail weight, peakedness (i.e., width of peak), and also lack of shoulders (i.e., when distribution primarily resides in peak and tails and not in between).

Consider some univariate data $D_1, D_2, \ldots, D_N$, then their skewness as well as kurtosis is found as follows:

$$\text{Skewness} = \frac{\sum_{i=1}^{N}(D_i - \mu)^3/N}{\sigma^3} \tag{5.4}$$

$$\text{Kurtosis} = \frac{\sum_{i=1}^{N}(D_i - \mu)^4/N}{\sigma^4} \tag{5.5}$$

where μ is mean, σ is standard deviation, and N is number of pixels.

5.4 Conclusion

The rapid sprout in the usage of sensitive information exchange through the Internet or any public platform causes a major security concern in these days. More essentially, digital data has given an easy access to communication of its content that can also be copied without any kind of degradation or loss. Therefore, the urgency of security during global communication is obviously quite palpable nowadays. Hence, the data hiding in the seemingly unimportant cover medium is perpetuated. The realm of steganography ratifies the stated fact to safeguard the privacy of data. Unlike cryptography, steganography brings forth various techniques that strive to hide the existence of any hidden information along with keeping it encrypted. On the other hand, any apparently visible encrypted information is definitely more likely to captivate the interest of some hackers and crackers. Therefore, precisely saying, cryptography is a practice of shielding the very contents of the cryptic messages alone. On the other hand, steganography is seriously bothered with camouflaging the fact that some confidential information is being sent, along with concealing the very contents of the message. Hence, using steganographic procedures in the field of big data enhances their security.

References

[1]   Mukherjee, S., & Sanyal, G. A chaos based image steganographic system", Multimed Tools Appl, Springer, 2018, 77 (21).

[2]   Gupta, B., Agrawal, D.P., & Yamaguchi, S. Handbook of Research on Modern Cryptographic Solutions for Computer and Cyber Security, 2016.

[3]   Saha, PK., Strand, R., & Borgefors, G. Digital Topology and Geometry in Medical Imaging: A Survey. IEEE Transactions on Medical Imaging, 2015, 34(9), 1940–1964.

[4]   Potdar, V., & Chang, E. Gray level modification steganography for secret communication, IEEE International Conference on Industrial Informatics, Berlin, Germany, 2004, 355–368

[5]   Dagadita, MA., Slusanschi, EI., & Dobre, R. Data Hiding Using Steganography. 12th International Symposium on Parallel and Distributed Computing, IEEE, 2013, 159–166

[6]   Katzenbeisser, S., & Petitcolas, F. A. Information Hiding. Artech House information security and privacy series, Artech House, 2015, ISBN 978-1-60807-928-5, pp. I-XVI, 1–299

[7]   Mukherjee, S., & Sanyal, G. (2018): A Multi-level Image Steganography Methodology Based on Adaptive PMS and Block Based Pixel Swapping, Multimed Tools Appl, Springer, 2018

[8]   Mukherjee, S., & Sanyal, G. Extended Power Modulus Scrambling (PMS) Based Image Steganography with Bit Mapping Insertion. In: Fahrnberger G., Gopinathan S., Parida L. (eds) Distributed Computing and Internet Technology, 2019, ICDCIT 2019. Lecture Notes in Computer Science, vol 11319. 364–379. Springer, Cham

[9]   Mukherjee, S., Roy, S., & Sanyal, G. Image Steganography Using Mid Position Value Technique, International Conference on Computational Intelligence and Data Science (ICCIDS), Procedia Computer Science, 2018, 132,461–468, Elsevier

[10] Mukherjee, S., & Sanyal, G. A Novel Image Steganography Methodology Based on Adaptive PMS Technique. In: Sa P., Sahoo M., Murugappan M., Wu Y., Majhi B. (eds) Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. Advances in Intelligent Systems and Computing, 2018, vol 518. 157–164. Springer, Singapore.

[11] Das, Shantanu, Tixeuil, Sebastien (Eds.). Structural Information and Communication Complexity. 24th International Colloquium, SIROCCO 2017, Porquerolles, France, 2017

[12] 6. Mukherjee, S., Ash, S., & Sanyal, G. A Novel Differential Calculus Based Image Steganography with Crossover, International Journal of Information and Communication Engineering, World Academy of Science, Engineering and Technology (WASET), (2015), 9(4): 1056–1062

[13] Zhengan, H., Shengli, L., Xianping, M., Kefei, C., & Jin, L. Insight of the Protection for Data Security Under Selective Opening Attacks. Information Sciences, 2017, 412–413, 223–241

[14] Mukherjee, S., & Sanyal, G. Enhanced Position Power First Mapping (PPFM) based Image Steganography, International Journal of Computers and Applications (IJCA), Taylor and Francis, 2017, 39 (2): 59–68,

[15] Mukherjee, S., & Sanyal, G. Edge Based Image Steganography with Variable Threshold, Multimed Tools Appl, Springer, 2018.

[16] Khosla, S., & Kaur, P. Secure Data Hiding Technique using Video Steganography and Watermarking. International Journal of Computer Applications, 2014, 95(20), 7–12.

[17] Mukherjee, S., & Sanyal, G. A Physical Equation Based Image Steganography with Electro-magnetic Embedding, Multimed Tools Appl, Springer, 2019

[18] Kaminsky, Alan., Kurdziel, Michael., & Radziszowski, Stanislaw. An Overview of Cryptanalysis Research for the Advanced Encryption Standard. Proceedings – IEEE Military Communications Conference MILCOM, 2010, 10.1109/MILCOM.2010.5680130.

[19] Khalaf, Abdulrahman. Fast Image Encryption based on Random Image Key. International Journal of Computer Applications, 2016, 134.

[20] Dai, Hong-Ning., Wang, Qiu., Dong, Li., & Wong, Raymond. On Eavesdropping Attacks in Wireless Sensor Networks with Directional Antennas. International Journal of Distributed Sensor Networks, 2013, 2013.

[21] Panda, M., & Nag, A. Plain Text Encryption Using AES, DES and SALSA20 by Java Based Bouncy Castle API on Windows and Linux. 2015 Second International Conference on Advances in Computing and Communication Engineering, 2015, 541–548.

[22] Wei, S., Sun, Z., Yin, R., & Yuan, J. Trade-Off Between Security and Performance in Block Ciphered Systems With Erroneous Ciphertexts. IEEE Transactions on Information Forensics and Security, 2013, 8, 636–645.

[23] Khalaf, Abdulrahman. Fast Image Encryption based on Random Image Key. International Journal of Computer Applications, 2016, 134.

[24] Ping, L., Jin, L., Zhengan, H., Tong, L., Chong-Zhi, G., Siu-Ming, Y., & Kai, C. Multi-Key Privacy-Preserving Deep Learning in Cloud Computing. Future Generation Computer Systems, 2017, 74, 76–85.

[25] Muhammad, K., Ahmad, J., Rho, S., & Baik, S.W. Image steganography for authenticity of visual contents in social networks. Multimedia Tools and Applications, 2017, 76, 18985–19004.

[26] Sobti, Rajeev., & Ganesan, Geetha. Cryptographic Hash Functions: A Review. International Journal of Computer Science Issues, 2012, 9, 461–479.

[27] Sedighi, V., Cogranne, R., & Fridrich, J. Content-Adaptive Steganography by Minimizing Statistical Detectability.: IEEE Transactions on Information Forensics and Security, 2016, 11(2), 221–234

[28] Steendam, H. On the Selection of the Redundant Carrier Positions in UW-OFDM. IEEE Transactions on Signal Processing, 2013, 61(5), 1112–1120.

[29] Zhu, L., & Zhu, L. Electronic signature based on digital signature and digital watermarking. 5th International Congress on Image and Signal Processing, CISP, 2012, 1644–1647

[30] Zhang, Weiming., Zhang, Xinpeng., & Wang, Shuozhong., "Near-Optimal Codes for Information Embedding in Gray-Scale Signals," IEEE Transactions on Information Theory, 2010, 1262–1270.

[31] Abdallah, E.E., Ben Hamza, A., & Bhattacharya, P. MPEG Video Watermarking Using Tensor Singular Value Decomposition, International Conference Image Analysis and Recognition, ICIAR 2007: Image Analysis and Recognition, pp. 772–783

[32] Li, J., Yu, C., Gupta, BB. et al. Color image watermarking scheme based on quaternion Hadamard transform and Schur decomposition. Multimed Tools Appl, 2018, 77(4), 4545–4561.

[33] Do, Q., Martini, B., & Choo K-K, R. The Role of the Adversary Model in Applied Security Research. Computers & Security

[34] Kahn, D. The History of Steganography. Lect Notes Comput Sci, 1996, 1174. 1–5.

[35] Siper, A., Farley, R., & Lombardo, C. The Rise of Steganography. Proceedings of Student/ Faculty Research Day, CSIS, Pace University, D1_1-7, 2005.

[36] Hamid, Nagham., Yahya, Abid., Ahmad, R. Badlishah., Osamah, M. Al-Qershi., Alzubaidy, Dheiaa Aldeen Najim., & Kanaan, Lubna. Enhancing the Robustness of Digital Image Steganography Using ECC and Redundancy. Journal of Information Science and Engineering, 2012.

[37] Carson, Austin., & Yarhi-Milo, Keren., Covert Communication: The Intelligibility and Credibility of Signaling in Secret, Security Studies, 2016, 26, 124–156.

[38] Hussain, A., Al-Fayadh, A., & Radi, N. Image Compression Techniques: A Survey in Lossless and Lossy algorithms, Neurocomputing, 2018, 300, 44–69

[39] Borges, PVK., Mayer, J., & Izquierdo, E. Robust and Transparent Color Modulation for Text Data Hiding. IEEE Transactions on Multimedia, 2008, 10(8), 1479–1489.

[40] Cem Kasapbaşi, M., & Elmasry, W. New LSB-based colour image steganography method to enhance the efficiency in payload capacity, security and integrity check. Sādhanā (2018) 43: 68.

[41] Febryan, A., Purboyo, TW., & Saputra, RE. Steganography Methods on Text, Audio, Image and Video: A Survey. International. Journal of Applied Engineering Research, 2017, 12(21), 10485–10490

[42] Ahvanooey, MT., Li, Q., Hou, J., et al. AITSteg: An Innovative Text Steganography Technique for Hidden Transmission of Text Message via Social Media. IEEE Access, 2018, 6, 65981–65995.

[43] Khalil, M., Li, JP., & Kumar, K. (2015): Color constancy models inspired by human visual system: Survey paper. 12th International Computer Conference on Wavelet Active Media Technology and Information Processing. 432–435

[44] Zhang, Y-M., & Cen, J-J. (2010) Research on method of transformation from bitmap to vector graphics based on Adobe Illustrator CS4. International Conference on Advanced Computer Theory and Engineering, IEEE, V3_75–77

[45] Olugbara, OO., Adetiba, E., & Oyewole, SA. Pixel Intensity Clustering Algorithm for Multilevel Image Segmentation. Mathematical Problems in Engineering, 2015, 1–19

[46] Hashim, M., Mohd, R., & Alwan, A. A review and open issues of multifarious image steganography techniques in spatial domain. Journal of Theoretical and Applied Information Technology, 2018, 96(4). 956–977

[47] Elham, Ghasemi., Shanbezadeh, Jamshid., & Nima, Fassihi. High Capacity Image Steganography using Wavelet Transform and Genetic Algorithm. Lecture Notes in Engineering and Computer Science, 2011, 1. 10.1007/978-1-4614-1695-1_30.

[48] Yang, C., Weng, C., Wang, S., et al Adaptive Data Hiding in Edge Areas of Images With Spatial LSB Domain Systems. IEEE Transactions on Information Forensics and Security, 2008, 3, 488–497

[49] Shen, S., & Huang, L. A Data Hiding Scheme Using Pixel Value Differencing and Improving Exploiting Modification Directions. Computers and Security, 2014, 48, 131–141

[50] Dey, N., Roy, A. B., & Dey, S. A novel approach of color image hiding using RGB color planes and DWT. International Journal of Computer Applications, 2012, 36(5), 19–24.

[51] Zhou, X., Yunhao Bai, Y., & Wang, C. Image Compression Based on Discrete Cosine Transform and Multistage Vector Quantization, International Journal of Multimedia and Ubiquitous Engineering, 2015, 10(6), 347–356

[52] Bhattacharyya, D., & Kim, T. Image Data Hiding Technique Using Discrete Fourier Transformation. In: Kim T., Adeli H., Robles R.J., Balitanas M. (eds) Ubiquitous Computing and Multimedia Applications. Communications in Computer and Information Science, Springer, Berlin, Heidelberg, 2011, 151

[53] Hansen, B. The Integrated Mean Squared Error of Series Regression and a Rosenthal Hilbert-Space Inequality. Econometric Theory, 2015, 31, 337–361

[54] Tao, D., Di, S., Liang, X., Chen, Z., & Cappello, F. Fixed-PSNR Lossy Compression for Scientific Data. 2018 IEEE International Conference on Cluster Computing (CLUSTER), 2018, 314–318.

[55] Dosselmann, R., & Yang, X.D. A comprehensive assessment of the structural similarity index. Signal, Image and Video Processing. SIViP. Vol. 5, pp 81–91 (2011)

[56] Malik, F. Mean and Standard Deviation Features of Color Histogram Using Laplacian Filter for Content-Based Image Retrieval. Journal of Theoretical and Applied Information Technology, 2011, 34

[57] Sung, Jungmin., Kim, Dae-Chul., Choi, Bong-Yeol., & Ha, Yeong-Ho. Image Thresholding Using Standard Deviation. Proceedings of SPIE – The International Society for Optical Engineering, 2014, 9024. 10.1117/12.2040990.

[58] Duncan, K., & Sarkar, S. (2012) Relational Entropy-Based Saliency Detection in Images and Videos. 19th IEEE International Conference on Image Processing. 1093–1096

[59] Koo, H., & Cho, N. Skew Estimation of Natural Images Based on a Salient Line Detector. J. Electronic Imaging, 2013, 22.

[60] Ferzli, R., Girija, L., & Ali, W. (2010) Efficient Implementation of Kurtosis Based No Reference Image Sharpness Metric. in Jaakko Astola & Karen O. Egiazarian (Ed.): Proc. SPIE 7532, Image Processing: Algorithms and Systems VIII

Santanu Koley

6 Big data security issues with challenges and solutions

Abstract: Big data is a collection of huge sets of data with different categories where it could be distinguished as structured and unstructured data. As we are revolutionizing to zeta bytes from Giga/tera/peta/exabytes in this phase of computing, the threats have also increased in parallel. Besides big organizations, cost reduction is the criterion for the use of small- and medium-sized organizations too, thereby increasing the security threat. Checking of the streaming data once is not the solution as security breaches cannot be understood.

The data stack up within the clouds is not the only preference as big data technology is available for dispensation of both structured and unstructured data. Nowadays, an enormous quantity of data is provoked by mobile phones (smartphone) or equally the symphony form. Big data architecture is comprehended among the mobile cloud designed for supreme consumption. The best ever implementation is able to be conked out realistically to make use of a novel data-centric architecture of MapReduce technology, while Hadoop distributed file system also acts with immense liability in using data with divergent arrangement.

As time approaches, the level of information and data engendered from different sources enhanced and faster execution is the claim for the same. In this chapter our aim is to find out big data security that is vulnerable and also to find out the best possible solutions for them. We consider that this attempt will dislodge a stride forward along the way to an improved evolution in secure propinquity to opportunity.

Keywords: Big data, Hadoop, MapReduce, data security, big data analysis

6.1 Introduction

Massive knowledge is employed by industries at all levels that have access to big data and therefore the means to use it. Software package infrastructures like Hadoop change developers to distribute storage. The distribution process of terribly giant knowledge sets on computer clusters. It simply leverages a lot of computing nodes to perform data-parallel computing [1]. With the mixture of the ability to shop for computing power on demand [2] from public cloud suppliers, such developments greatly dramatize the adoption of huge data processing methodologies. Therefore, new security

 Santanu Koley, Department of Computer Science and Engineering, Budge Budge Institute of Technology, Kolkata, India

challenges have turned out from the coupling of big data with public cloud environments classified by heterogeneous compositions of hardware with operating systems (OS), and software package infrastructures for storing and computing on knowledge.

Data security is a much-needed criterion in today's Internet-based world that is dependent on mobile phone technology. Real-time use of social media such as Facebook, Twitter, LinkedIn, blogs, WhatsApp, and further Internet-based sites produces an enormous quantity of data all the way through the world. Eric Schmidt, the erstwhile CEO of Google, once said about the production in such a rate with the aim that it is fashioned by this Internet world up to 2003 is around 5 exabytes. These data are in an increasing manner in its day-to-day applications in a multiplicative manner [3]. The reason behind this growth of unstructured (e.g., MS-Word, PDF, any Text, Media Logs) or semistructured (XML, CSV, JavaScript Object Notation, etc.) data along with structured (data with relational databases) one is mainly the grounds of facet data restricted by diverse associations due to dissimilar reasons approximating enhancing of sales, detail study, analysis, escalation of social media, continuous survey, shared projects, IoT, multimedia, and so on.

Big data expertise works with other well-known technology, that is, cloud computing, provides the user much flexibility in terms of services and money, pays as per the use of singular services endowed with, eradication of costly computer hardware, has little speculation on setup, moves threats in excess of contradictory systems, and furthermore has diminutive point of time to market make cloud systems remarkably acknowledged. Cloud computing has set out novel applications in mobile technology as the entire services and amenities are provided at its finest.

The traditional applications like RDBMS (relational database management system) come into play for structured data of tabular approach stored in .csv or .xls category of formats. Today the demand for applications like Facebook and WhatsApp involves in storing unstructured data besides the structured one. The unstructured data similar to images of diverse setup together with healthcare substantiation is analogous toward X-rays, ECG, MRI images, travel and logistics as well as moveable text layouts, forms of miscellaneous kind, video as well as audio, documents comparable to text, doc, rtf, and additional setups, manuals, contacts, automotive data, data associated with safety, accessory of electronic mail, energy/industry retails, and others Prearranged one enlightens us concerning tables of early database management systems .CSV's and .XLS's, where row and column are applicable, as conventional database management systems represent superior.

Big data follows an essential part with cloud computing as data stores in cloudlets. Five V's of big data provides strength to analyze the data with dissimilar approaches and finally find the results. Using the elucidation endows with Google, Doug Cutting and his panel developed an open-source project called HADOOP. Hadoop scuttles function with the MapReduce algorithm, where the data is routing in comparison with others. In short, Hadoop is used to build up applications that could carry out absolute numerical study on gigantic quantity of data.

Real-time privacy with big data analysis is required. The security problem arises when distributed frameworks are used like MapReduce function of Hadoop, which dispense huge processing tasks to dissimilar systems to save processing time and breach of security crop up. These tasks are taken as input to endpoint devices. These are the main factors of security violation as data processing, storage, and other tasks are performed here.

Storage on endpoints can stock up this streaming data into several tiers. When the next data is stored, the tiers are also changed as the criteria of the most used and least used concept are there. The manual tiering system is replaced with auto-tiering, thus transaction logs cannot cop up with and increase security crisis. Nonrelational data stores like NoSQL cannot encrypt data during distributing with endpoints when it is flowing or composed, labeling or logging also makes the same problem. At the time of storing data into storage devices, a proper encryption technique is needed. Similarly the access control, encryption, and validation are also necessary when users are associated with enterprise IT as a whole. Data mining solution is another security breach as data collected from the provider and collector provides this data to the miner. This technique involves a specified mining algorithm, which ensures the privacy and security of data.

Granular auditing is a kind of security check on logs to ensure the result on different parts, where data is stored in case of external attacks. It may be an unsuccessful hit, but auditing finds its consequences. Granular access control of big data requirements by means of NoSQL databases or the Hadoop distributed file system (HDFS) intended for a vigorous verification practice and obligatory admission power. Data provenance is done with metadata, where users can check, verify, and authenticate the data with high speed for any security issues.

Securing data storage endpoints is much necessary, but at the same time, it is critical too in case of distributed architecture. When the data volume can rise up to exabytes, autotiering system is essential. This will set streaming data involuntarily allocate to indicate and put together, organizing enormous dimensions of data uncomplicated to manage. Unverified services or contradictory protocols like crisis can occur as a result. This system generates logs that can store the data about storage in tiers that can be isolated and preserved accurately. Secure untrusted data repository is a network file system that can protect the data and logs from modifying them from external unauthorized users by constantly checking and monitoring.

Organizations that employ huge unstructured data like photographs, video, audio, and text cannot use typical structured query language (SQL), rather they utilize NoSQL. It does not have default administrative user enabled, weak authentication connecting server and client as communicating via plaintext. Weak password storage, lack of encryption support, susceptible to SQL injection, and denial of service attacks are also responsible for severe security issues. A bit solution to these may include server–client encryption algorithms approximating Rivest–

Shamir–Adleman, Advanced Encryption Standard, and SHA-256 Cryptographic Hash Algorithm, as well as Secure Sockets Layer encryption.

The distributed framework separated into diverse endpoint devices need high security. The reason behind the formulation of a trusted certificate at the point of all endpoints will ensure safekeepings. Ensuring endpoint security measures include testing resources of data or information on a regular basis and usage of reliable, trusted networking devices only through the use of a platform like mobile device management (MDM).

At every endpoint, the trusted certificates will facilitate to make sure that the data stay protected. Supplementary actions to help the association ought to exploit comprise usual resource testing and allowing merely trusted devices to associate with the network all the way through the economic consumption of an MDM program. After this hardware part, a data checking should also be there to guarantee valid data received on endpoints. Now to get rid of the hackers and protect from malware application, input devices as well as applications are defenseless. Sometimes an external attacker copies with several identifications as users and fills those endpoint storages with forged data to make the system unavailable.

6.2 Cloud computing

Cloud computing is the aspect of the present computing globe. It is a centralized approach in stipulations of data storage, improved operations, and ubiquitous, well-located, on-demand web access with smallest disbursement. It moves behind a collective pool architecture, where essential networks, special servers, enormous storage, numerous applications, and dissimilar services are incorporated [4]. The cloud model is an amalgamation of five indispensable distinctiveness, three service models, and four deployment models as looked for [5].

Cloud computing follows a service-oriented architecture (SOA), where it is divided into three main categories of services like IaaS (infrastructure as a service), PaaS (platform as a service), and SaaS (software as a service). This on-premises solution hardware or software is just like purchasing a car where driving, alteration, and changing routes as per requirement is possible as per the buyer/user. While IaaS can be an example of leasing a car from someone and drive it as we like, but in case of upgradation several others can be taken. On the other hand, PaaS can be described as taking a taxi for some time and pay the fare only, but cannot drive it either by modifying anything we dislike. Conversely, SaaS is similar to using a bus with predefined routes. Fundamentally, IaaS is employed by IT administrators; software developers are bringing into play; PaaS and SaaS make use of the simple end users.

IaaS cloud service providers are Amazon Web Services Elastic Compute Cloud (EC2), Microsoft Azure, and Google Compute Engine. The big players of PaaS are

Heroku, AWS Elastic Beanstalk, and Google App Engine. Finally, as an end user, we are aware of the names of SaaS such as Gmail, Trello, Salesforce CRM, EventPro, Office 365, and Google Docs.

The cloud does not stand for the traditional one as the name recommended is a flexible service by means of service-oriented structure. It makes use of the Internet intended for supplying certain services. The data, computer hardware along with software everything, is shared in this construction. "Share and use of applications and resources of a network environment to get work done without concern about ownership and management of the resources and applications" (M-S. E Scale, 2009) [6–8]. At this moment, the SOA put in the picture concerning the services provided through this technology is not something except contradictory transformation of several extraordinary technologies. The exceptional form of cloud service provides service and deployment models. Service model can be divided into three different varieties of perception, resembling PaaS, SaaS, and IaaS. Service models are described as the NIST model [9] for the above structural design. The discussion division for us is the IAAS cloud of the recently developed system.

The fantastically well-preferred cloud figuring is referred to as its notable highlights as pay-per-utilize model, which makes it a minimal effort, resource pooling, simple to introduce and usage capacity, arrangement of services as required by the client and ranch out capacity, QoS, wide network access, suppleness at a brief beat, self-stipulation, decided upgrade of administration, adaptable, relentlessness, easy upkeep with updegree, squat fence vitality, and so on [10].

"MCC at its simplest refers to an infrastructure where both the data storage and the data processing happen outside of the mobile device. Mobile cloud applications move the computing power and data storage away from mobile phones and into the cloud, bringing applications and mobile computing to not just Smartphone users, but a much broader range of mobile subscribers" [11, 12].

As the processing task isn't finished by the cell phone, the power and memory utilization is likewise less here and in the long run, the cell phone turned out to be quick [13].

IaaS cloud innovation is the suggested technique for putting away information on cell phones, apps, and so forth as second code portion or CDroid [14] server plot. To achieve the minimal effort phrasing, we should find the cloud server in a neighborhood land, where Fujitsu server is situated must have most reduced power cost for every unit in this world [15, 16].

6.3 Big data

Big data is an expression that is used to allude different sets of data that are exorbitantly tremendous or compound for customary information handling application

programming to viably manage. Information with numerous cases (lines) offers more prominent measurable power, while data with higher multifaceted nature (more qualities or segments) may prompt a higher false disclosure rate.

Big data challenges incorporate catching of data, data stockpiling, data analysis, seeking, sharing, exchange, perception, questioning, refreshing, information security, and data source problems. Big data are related to distributed (cloud) computing as data put away onto clouds can be depicted in Figure 6.1.
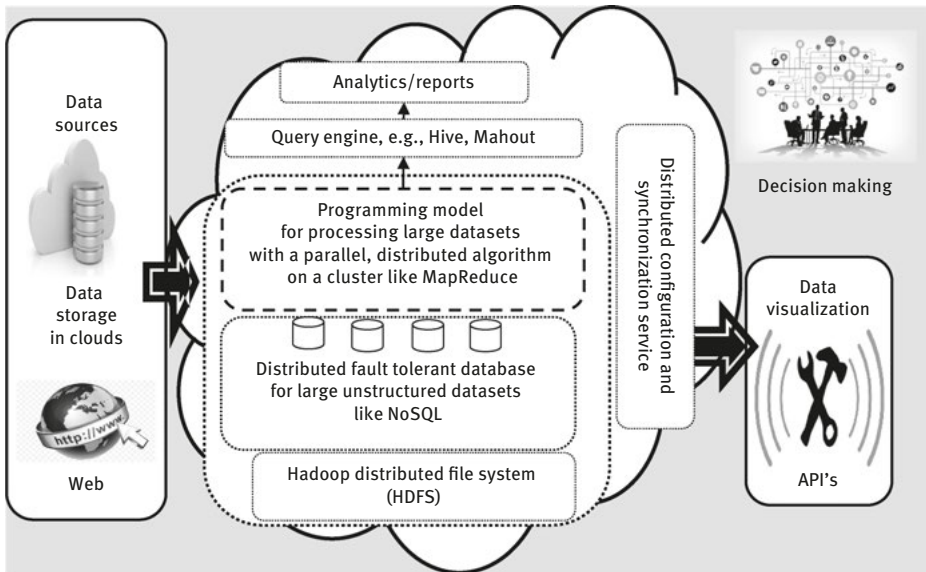
Figure 6.1: Cloud computing with big data.

The present routine with regard to the expression "big data" is slanted to submit to the utilization of prescient analytics, client conduct investigation, or certain other propelled data analytic techniques that separate an incentive from data, and sometimes to a specific size of the dataset. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem" [17]. Analysis of data collections can discover new relationships to "spot business patterns, anticipate ailments, and battle wrongdoing thus on" [18]. Scientists, business officials, experts of drug, publicizing, and governments alike consistently meet troubles with extensive datasets in territories including Internet look, fintech, urban informatics, and business informatics. Researchers experience restrictions in e-science work, including meteorology, genomics [19], connectomics, complex material science reproductions, science, and ecological research.

Big data is an idea to store, process, and break down gigantic sum (e.g., exabytes) of data that is almost incomprehensible with conventional RDBMS

framework. This is on the grounds that other than organized data it manages semi- or unstructured data. Big data is grouped into a few classifications like data sources, content arrangement, data stores, data organizing, and data preparing. The detailed order can be comprehended with the assistance of Figure 6.3.

Data sources are named as web and social media, a content format known as structured, semistructured, or unstructured configuration.

The data format is document oriented, column oriented, graph based, and key qualities while data staging can be depicted as cleaning, normalization, and change. At long last, data processing should be possible with the assistance of two unique strategies like batch processing and real-time processing.

Big data is able to illustrate by means of three ideologies of "V" such as volume, variety, velocity, and it can be extended to veracity, validity, volatility, and value (Figure 6.2).
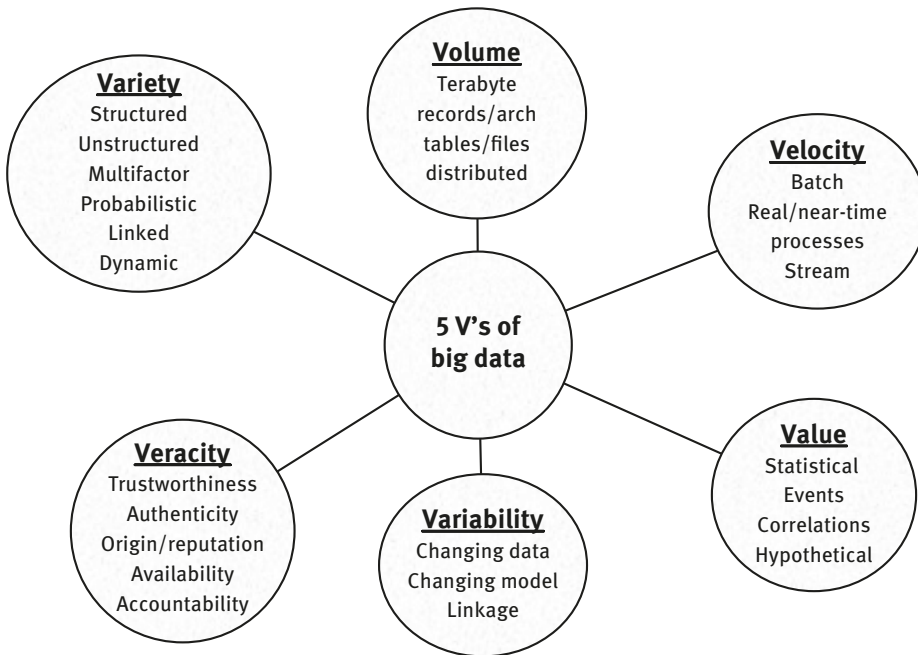
Figure 6.2: The V's of big data.

–   Volume: Volume talks on the extent of the data created from various sources and is prepared to extend as far as records, tables, or other forms. The size of the data may vary up to terabyte, petabyte, and zetabytes. Information made with definite data analysis and smartphone are the greatest maker of such sort of longitudinal data [21].

- Variety: Variety means the class of data, for example, in various structures. Furthermore, various sources will deliver big data, for example, sensors, gadgets, social networks, the web, and cell phones. For instance, data could be web logs, radiofrequency identification sensor readings, unstructured social networking data, gushed video, and sound. Other than unstructured data, both structured and semistructured assortment of data stores into big data.
- Velocity: This implies how regularly the data is produced, for example, data analysis. For instance, each nanosecond, millisecond, second, minute, hour, day, week, month, and year. Handling recurrence may likewise contrast from the client necessities. A few data should be prepared real time, batch, or stream too and some may not.
- Veracity: Veracity is the data in doubt, that is, uncertain, untrusted, and unclean. The uncertainty is due to data inconsistency and incompleteness, ambiguities, latency, deception, and model approximations. It is the management of the reliability and predictability of inherently imprecise data types.
- Validity: Validity of input data based on accurate processing is done on the data and provides a particular output as the product. The validity of data is very near to the veracity of the data. Through big data, one should be spared attentively regarding validity. For example, in the banking sector, data accumulated from various banks must be valid to show the growth of the nation. The finance ministry may plan their strategy for the next financial year regarding the valid data provided.
- Volatility: Volatility is very common when data is needed to transform into other types on a regular basis. If it is not accounted for analytical results, perhaps invalid at the instant they are produced. Such types of circumstances are very common, where businesses like the stock market or a telecom company (call data records related to one day). Volatility is directly linked by means of the challenges of invalidity and veracity.
- Value: In big data, value means exact data that are in reality and has some meaningful aspects that take out from some usable system. Apparently, this must be the output of big data processing. Value is an essential feature in the big data. Extracting the exact value is only possible when proper roads are there; here road refers to IT infrastructure arrangement to collect big data. These can be possible in dealing with a business that is meant for a return on investment.

There is another factor called inconstancy, which can be an issue for the individuals who break down and analyze the data. This alludes to the irregularity, which can be appeared by the data on occasion, along these lines hampering the way toward having the capacity to deal with the data adequately.

Big data are portrayed by various perspectives: (a) data sources, (b) data are various, (c) data can't be sorted into standard social databases, (d) content

arrangement, (e) data stores, (f) data organizing, and (g) data are created, caught, and processed quickly. The real classification of big data can be illustrated in the hierarchical structure as shown in Figure 6.3.
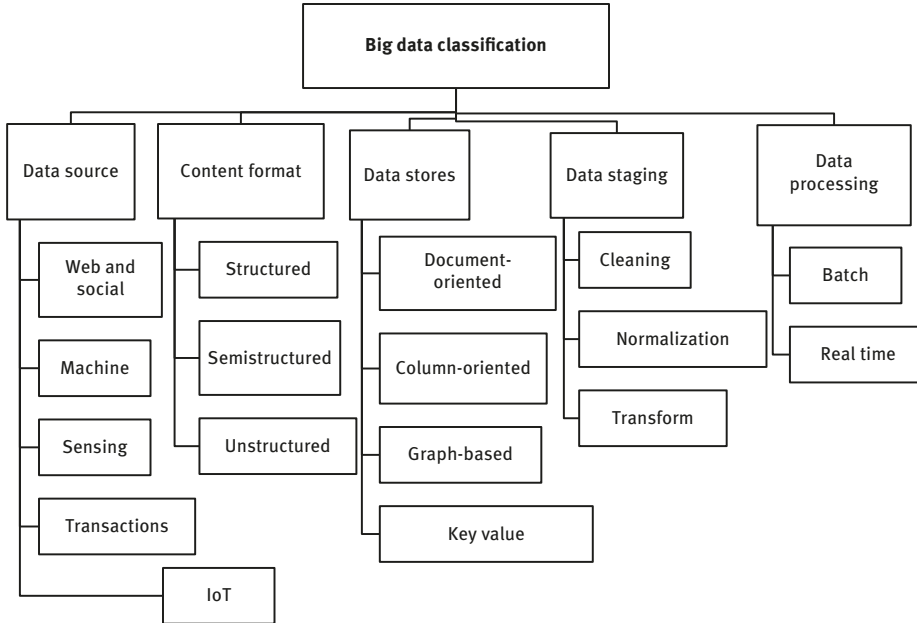
Figure 6.3: Big data classification [20].

The distributed file system structure is utilized to store in big data utilizing two unique strategies like HDFS and MapReduce programming system. The two are utilized to store and keep up the entire data structure (Figure 6.4).

Figure 6.4: Life cycle of big data.

6.3.1 Hadoop distribution file system

HDFS is responsible to collect the data in lumps. Here data is parted into squares of 64 MB each. It is made perfect for executing its file framework on a hardware

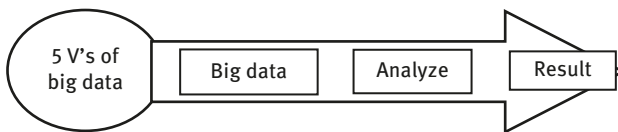platform where the cost is lower as well. It can endure the exceedingly imperfection framework as it keeps up block replication. The Internet searcher for the web application as named Apache-Nutch venture was the explanation behind the production of this engineering. Apache-Hadoop is the great introduction as a subproject of the resulting web crawler. An occasion of HDFS is made of a major number of server PCs; some of the time the amount can leave a couple of thousand as well. These servers reserve the data of the said record framework. They are not set up for a solitary collaboration by the clients, yet for cluster preparing. The data documents are a lot greater past evaluation, as they achieved the size close to TBs. Here the low latency of data access is ignored; however, high throughput is upgraded and that is much fundamental for a framework like HDFS. Data coherency and this throughput are the establishment of the idea of "compose once and read various occasions" idea of the files.

The Hadoop architecture contains data sources, Hadoop system, and big data understanding. The data sources contain site clickstream data, content administration framework, outside web substance, and user created content. The Hadoop system incorporates HDFS that is controlled with big data landing zone and MapReduce algorithms. These algorithms are constrained by keywords investigate, content characterizations, or subjects and client division. The big data knowledge holds keyword-applicable substance-rich client focused on presentation pages.

HDFS is set up to utilize Java technology that takes care of business. The basic plan of this document framework is given a name of NameNode – DataNode followed on master–slave development. Serving read–write operation/application from the file frameworks, customer is performed by the pair. The guidance from NameNode is completed by performing block creation, cancelation, and replication. They are expected to execute on item frameworks that might be OS as GNU/Linux as NameNode is somewhat software same as DataNode. NameNode runs file system namespace tasks like opening, shutting, and renaming files and directories. It is a master server and administrates the file framework namespace and controls access to documents by customers.

NameNode finishes up the mapping of different blocks to DataNodes. To guarantee the DataNodes working precisely, that is ordinarily single per bunch, the NameNode engages time-to-time refreshing of heartbeat and a block report from DataNodes in an ordinary interim. DataNode might be called as a vendor for the hub with its arrangement, as a gathering of DataNode is upright for amassing of a few blocks. These blocks contain part documents dispersed in changed blocks.

HDFS depicts a file system namespace and, put aside, client data to be put away in these documents are in blocks.

NameNode does not flood with user data; it is an arbitrator for all of HDFS metadata. HDFS is planned such a way that it has a solitary NameNode in a bunch to the most noteworthy degree that disentangles the design of the framework. Hadoop

system has the layer, in particular HDFS layer and MapReduce layer. The second one is the known execution engine in a multinode cluster. A job tracker colleague tasks trackers in both master and slave parts; then again, name node in the HDFS layer partner's data nodes in those parts as appeared in the outline in Figure 6.5.
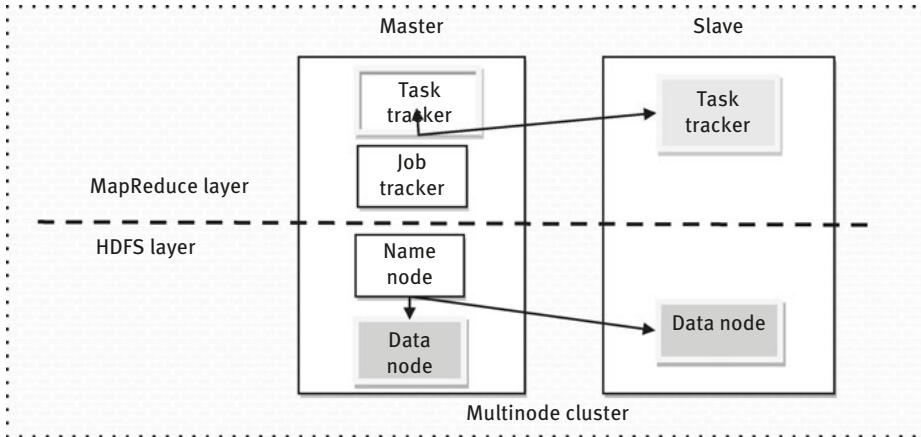

Figure 6.5: Master–Slave architecture of HDFS.

6.3.2 MapReduce

MapReduce is a data handling worldview, where a major measure of data is transformed into little. This programming model is combined with the task for doling out and causing cumbersome data collections with a parallel, distributed algorithm on a bunch [23, 24]. Hadoop is a framework that can store and control a lot of data all around effectively, in light of basic master–slave engineering. It is the center of Hadoop. Thoughtfully comparable methodologies have been very outstanding since 1995 with the Message Passing Interface [25] standard having decreased [26] and dissipate activities [22].

MapReduce is separated into a few applications, patterns, examples of overcoming adversity, utilities, capacities, highlights, and executions. Utilizations like questions and examination, works as map and reduce, includes in the vein of a programming model, huge-scale dispersed information handling, straightforward yet limited, parallel programming, extensible, motivated in useful programming however not comparable and once in a while think in recursive arrangements.

Executions looking like with Google, apache-Hadoop, and a wide range of innovations are utilized effectively in MapR. Different maps that lessen systems like Signalcollect and storm are utilized here. Security and privacy challenges in big data biological system are depicted in Figure 6.6 in the above section.
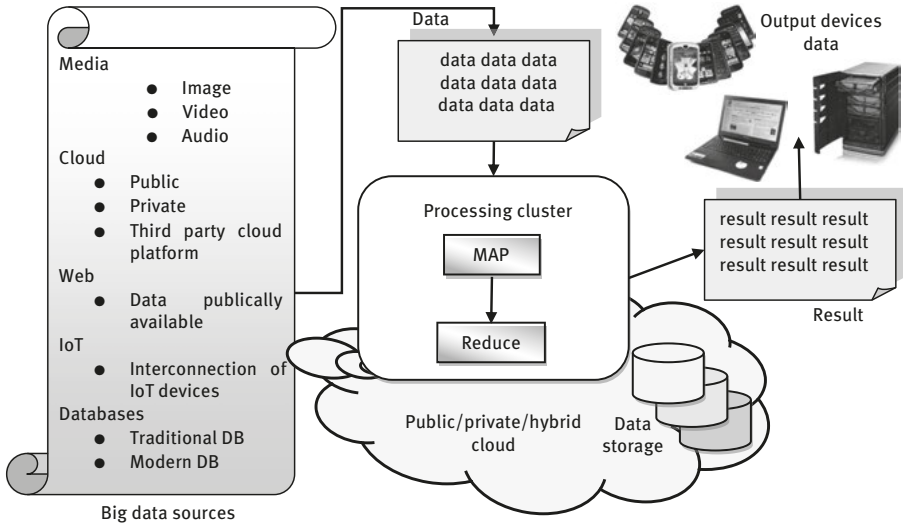
**Figure 6.6:** Security and privacy challenges in big data ecosystem [22].

6.3.2.1 Big data security structure

Big data security structure can be characterized into various classifications. Some of them can be named as infrastructure security, data protection, data management, and integrity and responsive security [27]. The outline depicts the design in detail. Foundation of security might be portrayed in cases like dispersed systems and nonsocial information stores, although information protection shows information mining and examination, cryptographically authorized information security and granular access control. Data executives depict data tiering, exchange logs, granular examining, and data provenance. Respectability and responsive security delineate constant protection and endpoint gadgets. The arrangement can be portrayed in Figure 6.7.

Big data is the most recent innovation utilized by associations that get vulnerability as we are uninformed of the vast majority of the things. The vast majority of the devices bringing into play are open source, and accordingly discover assaults of the hubs where information stores. Data stores here conveyed in nature; subsequently, ill-advised client verification happens. There is a significant prospect for pernicious data input and inadequate data approval.

The examination of the advancement of big data uncovers the high adequacy of big data as far as data handling. Be that as it may, the data preparing and information stockpiling of big data raises the issue of the data security ruptures and infringement of clients' protection. At the equivalent time, the fundamental exercises
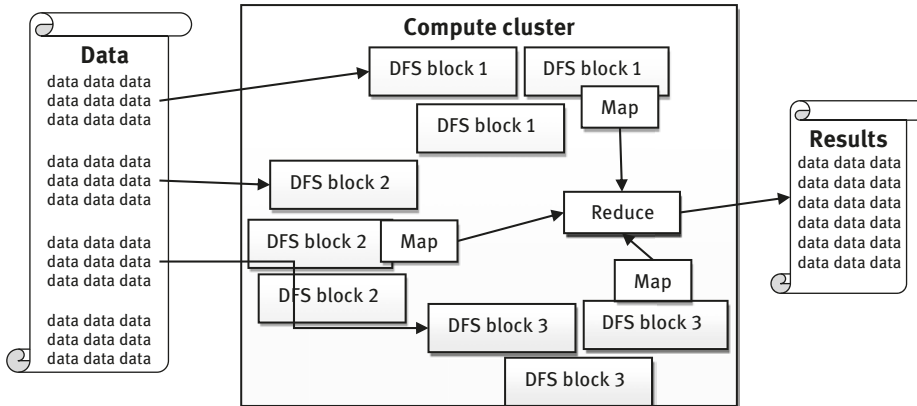
Figure 6.7: MapReduce architecture.

learned are the questionable idea of big data in light of the fact that, from one perspective, the advancement of big data raises data security dangers; in any case, then again, big data can possibly improve the data security in the event that they are legitimately utilized. The capability of big data is gigantic and the consideration regarding the data security is basic for the powerful improvement of big data and the avoidance of various dangers.

6.3.2.2 Real-time privacy

Constant supervision of real-time security is much demanding when dissimilar protection procedure engenders a gigantic amount of observant. The perceptive may go ahead numerous tribulations. They are frequently unobserved as the users cannot deal with the cut-off quantity. This may have a huge setback and also enhance further amid big data as the volume and velocity of data pour out. On the other hand, big data expertise endows with a prospect to facilitate this competence to carry out and consent to prompt processing and analytics of a remarkable type of data and this, in turn, can be worn out to formulate an existing real-time variance discovery associated with scalable safekeeping analytics as well. In real-time privacy, analytics and the use cases are diverse in dissimilar business applications where that particular industry will get benefitted (Figure 6.8).

For example, e-commerce and consumer marketing industry obtains huge support and profits in terms of monetary benefits. The same thing has to happen with healthcare industry when doctors need accurate data in terms of report generation for medical tests where doctors to prescribe. This situation is much similar when tax paying for a country as well. The tax calculations, returns, payback, advance
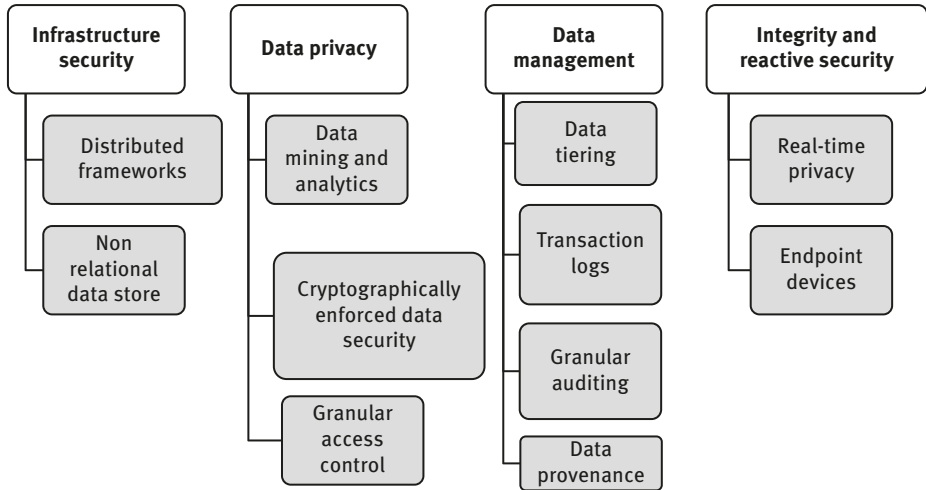
Figure 6.8: Different big data security and privacy challenges [22].

tax, and claims sometimes get frauds done on tax payments. Here the problem arises when accessing the data between different parties, resources of those data, and accessing the data in office or none office hours.

Today we move to computation on real-time data where big data faces most challenges. Here, real-time updating or keeping an eye on the websites and web pages is completed. The gigantic quantity of data (sometimes tera/petabytes) is composed on or after a variety of resources, sorted out, scrutinized by means of numerous data mining, data classification, and prediction algorithms, and consequently, reports are kept up of all these analyses. These prepared reports are exceptionally helpful when decision-making standards are satisfied. The carrying out of an organization depends to a great extent on those accounts. Language processing is a real-time data processing language used to process data streams coming from multiple sources (Figure 6.9).

IBM's Stream Processing Language has three singular varieties of operators: utility, relational, and arithmetic, which take data through input source operator and give output through output source operators. These multiple operators present in between the source filter, aggregate, join multiple data streams according to the need of the user. As per the necessities, the formulations of the operators can be executed manually by the users. Processing of streaming data is put in a more competent technique in big data, whereas it also props up ad hoc queries.

Here the end users can write their own query in SQL as in custom database application and straightforwardly submit them to relevant web applications too. Thus, it can get further flexibility and power. But there is a larger security concern of having ad hoc queries. The entire practice can be expressed in Figure 6.10.
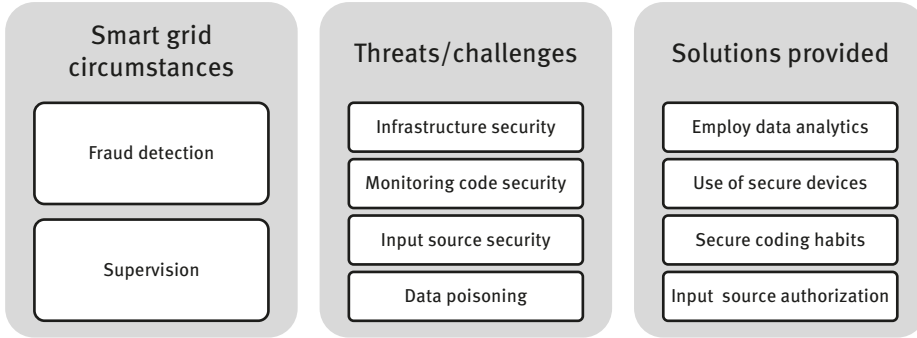
Figure 6.9: Time security monitoring (real-time privacy).
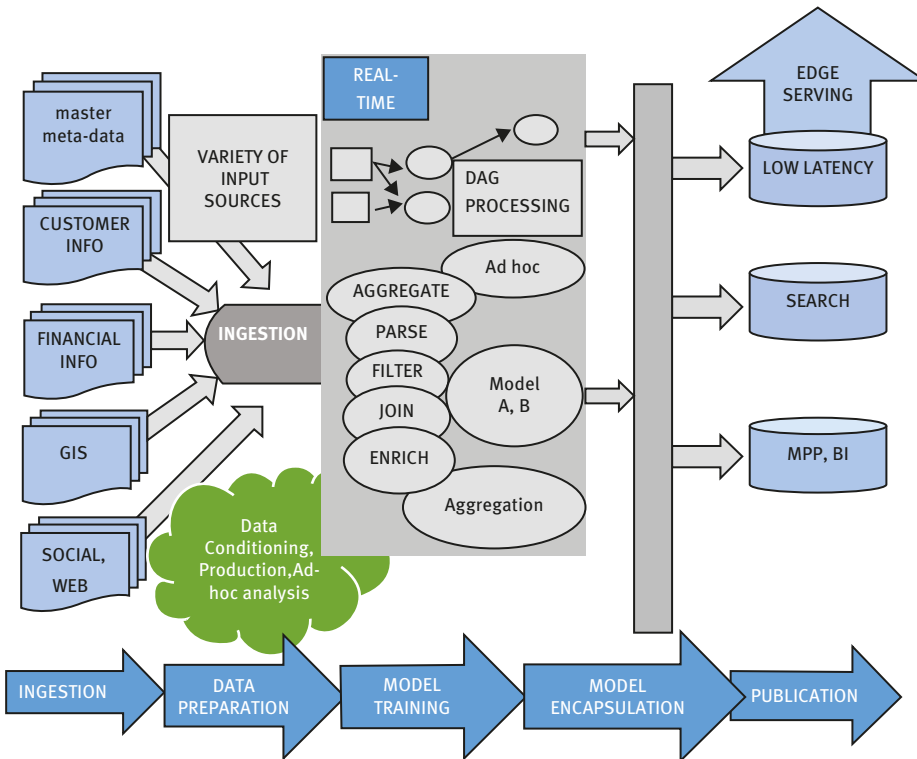

Figure 6.10: Real-time streaming and continuous computation.

Solution

At that place should be a control on the admittance to the data; moreover, it ought to be monitored. To put off illicit right of entry to the data, threat intelligence should be employed. Use big data analytics to set up confidence acquaintances to come together to make sure that merely authorized links to take place on a cluster. Scrutinizing tools like security information and event management (SIEM) way out can be brought into play to stumble on uncharacteristic associations. They may include:

6.3.2.2.1 Secure authentication gateways

Powerless confirmation system is a standout among the most widely recognized components that contribute toward information ruptures. Uncovering the vulnerabilities present in client confirmation work, a programmer can possibly access delicate information. Imperfect execution of client confirmation process must be counteracted at the plan organize. Guarantee that there are no broken confirmation tokens, which can be misused by any unapproved clients.

6.3.2.2.2 Utilize principle of least privilege

We ought to be in a perfect world to keep up a layered access control and actualize standard of least benefit. It advocates constraining client access to the insignificant dimension that will permit ordinary working. As it were, we should give a client just those benefits that are fundamental for that client to deal with his/her duties. It would keep unscrupulous IT experts from enjoying unlawful information mining exercises.

6.3.2.2.3 Utilize retrospective attack simulation

Not all associations can work in-house framework to help big data activities because of monetary requirements. Big data venture depends on an outsider cloud-based (public or private) arrangement; at that point review assault recreation can be utilized to discover vulnerabilities with the outsider application facilitated on the cloud. On the off-chance that the assault succeeds, at that point you ought to examine the issue further to locate changeless goals. Review reproduction would assist with identifying plausible shortcomings in the framework before a veritable programmer endeavors to abuse the helplessness.

6.3.2.2.4 Utilize latest antivirus protection

Numerous antivirus merchants have concocted security arrangements that are explicitly focused toward big data activities. So dependably ensures big data condition with the most recent Antivirus suite. Ensure that the updates and fixes are introduced when they are made accessible by the producer.

6.3.2.2.5 Utilize principle of least privilege

Big data is a developing business sector and the advancements are continually developing, making it hard for the current security answers to stay aware of the expanding request. Intermittent reviews will assist identifying new vulnerabilities as they make their essence felt. Subsequently, it can realign the security consistence with the present security guidelines.

6.3.2.2.6 Secure coding practices

While evaluating the code, one should ask himself the accompanying essential inquiries like, Am I ready to comprehend the code effectively? Is the code composed after the coding norms/rules? Is a similar code copied more than twice? Will I unit test/troubleshoot the code effectively to discover the underlying driver? Is this capacity or class too enormous? On the off-chance that truly, is the capacity or class having such a large number of duties? On the off-chance that one may feel that the appropriate response isn't tasteful to any of the above inquiries; at that point you can propose/prescribe code changes.

6.3.2.2.7 Input source authorization

The utilization to constrain which wellsprings of information are substantial for employment accommodation, including workstations, gadget perusers, hubs, and interior perusers. For instance, it should need to keep certain clients from entering employments from a specific workstation.

To approve the accommodation of work from explicit information sources, request that the security administrator can enact the class and characterize a profile for each information source. Furnish a security manager with info source and gadget names.

6.3.2.2.8 Employ data analytics

Big data analytics is the regularly mind-boggling procedure of looking at extensive and shifted data collections or enormous information to reveal data, including concealed examples, obscure connections, advertise patterns, and client inclinations that can enable associations to settle on educated business choices.

As a conclusion to real-time privacy, solutions endow with the use of security devices, secure coding habits, input source authorization, and employ data analytics. Figure 6.8 briefs the story in short.
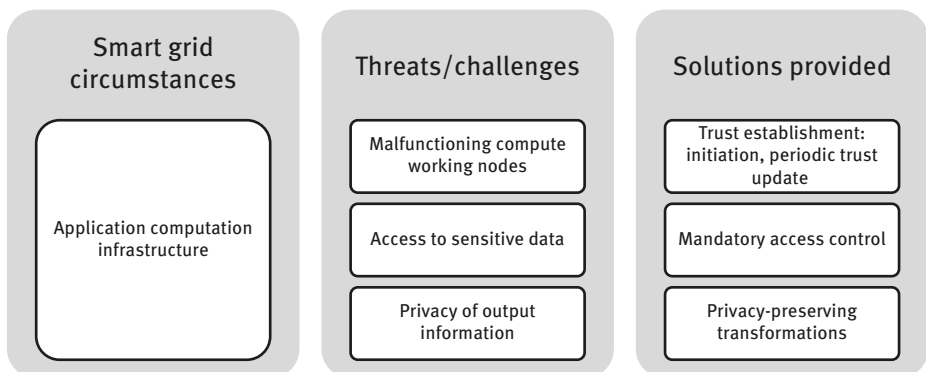
6.3.2.3 Distributed framework

The distributed computing framework is the modern approach today, where not only hardware, the software is distributed to. Here a software component shares

multiple computers in a network in order to get the faster processing capability. Distributed programming frameworks (DPF) tie together the primary, secondary storage, processing of data in an enormous quantity by size. Here in this framework (MapReduce), they split the parallel computation; storage functions process mammoth volumes of big data. It could identify wicked mappers, and look over the data entrusted mappers is the prevalent setback of big data and can potentially get in the way big data seclusions endeavor.

For example, MapReduce in Figure 6.7 is a programming model and a related usage for handling and creating substantial data collections with a parallel, distributed algorithm on a cluster. Normally, it segregates the input datasets into self-regulating pieces. These components of a set are processed by map jobs in a comprehensive equivalent approach. MapReduce, in its initial stage, is naught but a Mapper for each lump understands the data by a process called reading; there is a little calculation is executed with the data already read and amount produced a list of key/value couple. In the following stage, a reducer consolidates the qualities that are in the correct circumstance to each unique key and yields the outcome.

The attack anticipation assesses in two special ways: mapper security and the security of the data in the existence of an untrusted mapper. The untrusted mappers might revisit unidentified consequences, which will, consecutively, bring about an inaccurate cumulative outcome. When using large-scale datasets, it is incredible to recognize the outcome of momentous smash up in scientific and financial calculations. Different marketing and advertisement agencies collect and analyze buyer–retailer data to reach consumers they have already marked. These errands include a high measure of parallel estimations over expansive informational collections and are especially appropriate for MapReduce systems, for example, Hadoop. The information mappers may contain spillages that might be deliberate or unexpected in nature. For instance, a mapper may release an extremely particular incentive by breaking down individual data, undermining clients' security.

Figure 6.11: Distributed frameworks' secure computations.

To carry out a gigantic quantity of data, DPF makes use of parallel computations. DPF makes use of parallelism in computations with storage space to practice an enormous amount of data (Figure 6.11).

There are two different methods presented to make certain the trustworthiness of mappers: trust organization and mandatory access control (MAC). Throughout the first part, that is, trust establishment, "workers" must be genuine along with prearranged belongings by "masters," and only when they're experiencing can they be doled out mapper responsibilities. Subsequent to this requirement, periodic updates must be constructed to ensure mappers, again and again, and congregate the recognized procedure.

Alternately, MAC, the predefined security approach, will help to follow out. On the other hand, while MAC makes certain that the input of mappers is safe and sound, it does not put off data loss from mapper output. To keep away from this, it is important to influence data de-identification techniques that will set off the erroneous info from being circulated among nodes.

Solution

The seclusion and security problem take into account of a number of questions like auditing, access control, authentication, authorization, and privacy once bring into play the mapper and reducer process. The way out pays trusted third-party monitoring and security analytics (Apache Shiro, Apache Ranger, and Sentry) just as protection arrangement implementation with security to put off information spillage. In that regard is inalienable lack of clarity in administering compound application reconciliations on the creation-scale distributed framework. The system is to work by methods for an undertaking class programming model that has the office to grasp remaining task at hand strategies, tuning, and general observing and organization. At that point, when we make an application for a solitary office or different capacities, we influence an IPAF (to entomb protocol acceptability framework). The solution provided with the following:

6.3.2.3.1 Trust establishment

This blended gift additionally commands the connection between big data and trust. On the one hand, a lot of trust-related data can be used to build up creative data-driven methodologies for notoriety-based trust of the board. On the other hand, this is naturally attached to the trust we can put in the sources and nature of the basic data. There may be situation when trusted nodes in distributed structure malfunction with compute working nodes; the solution for this is to make regular updates on each node periodically.

6.3.2.3.2 Mandatory access control

MAC is a lot of security approaches obliged by framework arrangement, setup, and confirmation where access to sensitive data can take place. Macintosh arrangement, the board, and settings are built up in one secure system and restricted to framework

chairmen. Macintosh characterizes and guarantees a unified implementation of private security strategy parameters.

6.3.2.3.3 Privacy preserving transformations

Because of data gathering from various sources, odds of protection break have expanded. It is hard to apply existing protection models (security safeguarding procedures) in big data investigation due to 3Vs: volume (substantial measure of information), variety (organized, semistructured, or unstructured information), and velocity (quick age and preparing of information) – qualities of big data.

Thus, distributed framework solutions afford trust establishments like commencement, occasional trust update, obligatory access control, and protection safeguarding changes. This course of action is explained at this point in Figure 6.10.

6.3.2.4 Endpoint devices

The above framework of big data accumulates data from an assortment of resources. They are generally called as endpoint devices. The technique of collecting data split into two classes of perils such as data collection with validation and filtering of data. The first part (data collection with validation) is to collect data from several endpoints connected in a distributed network, where millions of hardware and software are associated with it in an enterprise network.

Here another problem arises when an input validation is performed on the piled-up data. Substantiation of the input data is important as infected data are too compiled with some malware application with computer viruses that may harm the data sources. Now the data should be filtered and modified as per the given format of the data requirements. The second part (filtering of data) provides the exact outcome improbable to validate and filter data. The data mapping can be done with the help of knowledge processing and business goals or assess the scope of customer data. Previous signature-oriented data filtering may perhaps be unsuccessful to solve the input validation and data filtering to slow down completely.

Solution
 Various solutions include the following:

6.3.2.4.1 Corruption free software

Tamper-proof software is required to apply the accumulated data from diverse endpoints. Finding out the proper software is awfully essential with the existing setup. The depravity of free software is much asked for swift operation, virus detection and removal, and use of original software obligatory.

6.3.2.4.2 Trust certificate and trusted devices

The trust certificates are necessary with the use of trusted devices. These are really difficult when collecting data from various sources, but during the filtration process, one can ensure the said certificate and devices. Filtration process can include firewall protection.

6.3.2.4.3 Analytics to detect outliers

The purpose of data analytics may help to find outliers. When special interpretations on data that move away from normal path and produce extreme values are called outliers, they will point out changeability during a measure, tentative inaccuracy, or a novelty. In alternative terms, an outlier is an associate inspection that deviates from a taken whole outline on a section.

Outliers can be divided into two categories, namely univariate and multivariate. Univariate outliers are originating on one occasion staring at an allocation of principles during a particular attribute area. Multivariate outliers are set up during an n-dimensional area (of n-characteristics). Staring at allotments in n-dimensional areas is often terribly tough for the human intelligence, that's why we need to prepare a replica to try and do it for us.

Outliers can even be available in completely diverse flavors, looking on the surroundings: point outliers, contextual outliers, or collective outliers. Point outliers are distinct data tip that lay off from the remainder of the distribution. Contextual outliers are noise in data, like punctuation symbols previously become conscious text analysis or background noise signal once doing speech identification. Collective outliers are subsets of novelties in data like a symbol that will point out the invention of recent trend.

6.3.2.4.4 Cryptographic protocol

Cryptographic protocols such as triple DES (put back the main data encryption standard utilizes three individual keys with all 56 bits. The entire key length implies 168 bits, yet masters would fight that 112 bits if key quality is progressively like it), RSA (an open key encryption estimation considered as amiss computation in light of its usage of two or three keys), Blowfish (symmetric consider parts messages along with squares of 64 bits and encodes them freely), Twofish (keys used in this figuring may be up to 256 bits in length and as a symmetric framework, only a solitary key is required), and AES (advanced encryption standard, incredibly successful in 128-piece structure; similarly, uses keys of 192 and 256 bits for significant encryption purposes) are there to compile the encrypted data collected from other origins.

They may be exceedingly useful when using semistructured or unstructured data. This procedure of endpoint input validations is exposed in Figure 6.12.
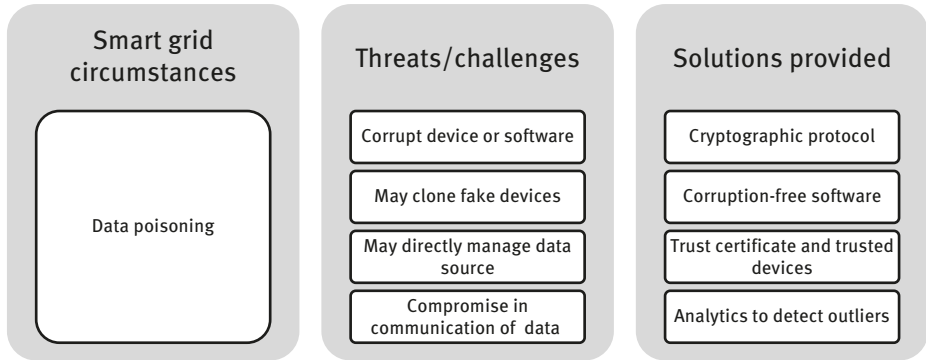
Smart grid circumstances	Threats/challenges	Solutions provided
Data poisoning	Corrupt device or software	Cryptographic protocol
	May clone fake devices	Corruption-free software
	May directly manage data source	Trust certificate and trusted devices
	Compromise in communication of data	Analytics to detect outliers

**Figure 6.12:** Endpoint input validation/filtering.

6.3.2.5 Data into several tiers

Tiered storage of data allocates to two or more dissimilar categories of storage media. They depend upon the business value that data collected so far. This business value enables the data tiering at an exceptionally low cost, but collecting the maximum data possible within. Sometimes the significance of data alters over time. There are some events that establish the altering worth of data, such as how often we access the data and for what time, what capacity you need to fend for its functions and exceptional circumstances that call for its quick retrievals, such as an account audit or financial account. Tiered storage includes the physical aspects, where hard disk, tape, or disk storage, as well as logical aspects like cloud storage, is associated.

Tiering approach of data storage in said devices depends upon the type of user data with its applications. As time progresses, this data tiering complexity increases (in a few cases, it decreases too) and it is classified into unusual types; thus considered necessary to store data in different types of storage medium. Generally, this is done manually, but the system may upgrade to automatic data storage tiering. The foremost companies help autotiering system, namely IBM, NetApp, and EMC. They can submit application, that is, policy-based rules to shift data involving dissimilar tiers automatically.

The advantages of the tiered storage system are cost efficient, operationally effective, and elastic. The tiered data storage system compels massive diminutions in storage costs compared with untiered storage. Today, the organizations realize instead of a single sort of depot for such data are a misuse of money for the majority of jobs (Figure 6.13).

Tiering storage is operationally very much efficient as per the different need of time and type. Such storage consists of the data that give out essential business functions that end up in high-performance storage media, such as solid-state drives

| Hot | Warm | Cold | Frozen |
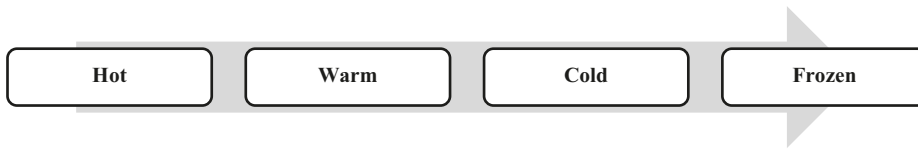
Figure 6.13: Multi-temperature splits big data.

(SSD) through archival data among little significance that ends up in tape storage or low-cost cloud storage services.

The normal manual train system is flexible enough, but automated tiering system has the largest flexibility ever, to shift data among diverse storage media as its value changes.

Big data describes distributed processing frameworks like Hadoop that are worn for data processing and tiering storage for big data applications. Hadoop's substantial storage potential appears on or after its clustering architecture, wherein data are distributed and lay-up in a network of compound computing possessions. If subscribed with a specific cloud setup, the data tiering part may be stored in the cloud data center; otherwise, Hadoop setup may be installed on-premise for enterprise-specific area.

At the very beginning, the datasets are created and afterward, they are used on a regular basis by several users from the diverse user from dissimilar endpoints. Seeing that large datasets come in Hadoop clusters, a component of the data is stock up on individual computing machines or nodes in a cluster.

Yet, dissimilar variety of data along with the regularity of data access is very frequent but with the progression of time, the value of data diminishes. Data tiering is supported by Hadoop, and by splitting the cluster into different tiers based on the frequency of use, it can considerably shrink the costs to pile up this data. This technique of reduction of cost in data storage is continuous in nature. It is publicized as (Figure 6.14).

Hot data: The real-time data analysis needs high-frequency entrance of data. The regular reporting or ad hoc queries are also associated with this type of data. The utmost computing power is utilized with this kind of data.

Warm data: This type of data accessed seldom. But it is functional at times so that it is necessary to be stored on hard disks or sometimes in SSD storage too. The computing power needed is obviously much less than hot data.

Cold data: Cold data is essentially a library type of data called archival data. This character of data is used in enterprises at times when the periodic report is generated or an enterprise wants to retain for compliance or once-off queries. This data is located in storage by means of negligible processing power obtainable of it.

Frozen data: The frozen data means it will almost never be used. This kind of data does not store in basic computer storage areas like HDD. Such kind of data is
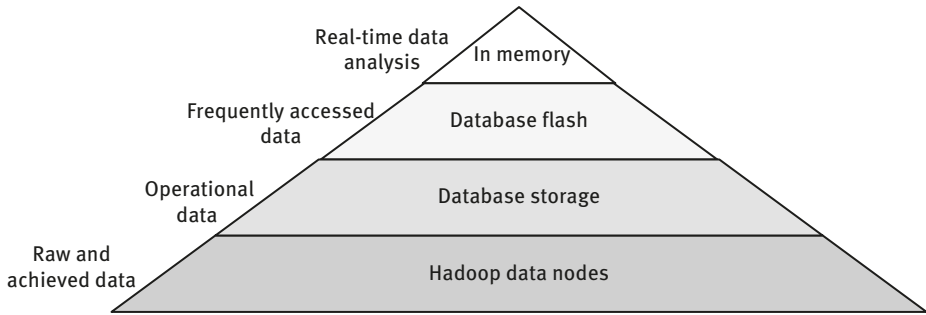
Figure 6.14: Data tiering hierarchy.

almost unusable, so the power consumption is near to zero. They can be stored in a node that uses minimal computing power, and less processing task is performed. Here data is stored in an achievable manner (Figure 6.15).
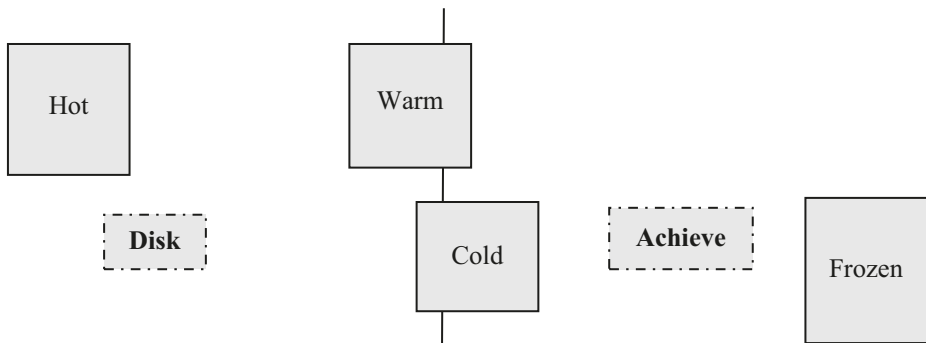
Figure 6.15: Mapping data to a storage tier based on temperature.

Solution

The cost reduction in data storage tiering is done for easy handling of data. Data is separated according to the frequency of use of data with a specific time frame stored in different nodes. Reduction of cost for data storage depends on the storage; when data stores on nodes with minimum computing power or in case of achieving data, the data with minimum computing power saves a huge amount of money. The data can move among said tiers via Hadoop tools, for instance, Mover. If the system turns out to be relatively dynamic, it sets aside for superior competence in the big data storage technique.

Now big data are classified into singular storage tiers per the occurrence of its usage. This is an excellent opening tip for the enterprise that wants to store information in a less pricey manner. After sometime, when this process is rationalized, user organization need not think about its storage and reallocation of the same data with its cost minimization as they move between tiers.

The development of every job is directly proportionate with the addition of data as big as it constructs the peak of big data that gathers by themselves. A small part within big data is useful in

this instance as the cold and especially archive data exist. Thus there is a perspective to categorize and tier enterprise level big data as soon as it gets into the clusters, leading to even greater efficiencies.

Storage tiering has immense prospective in a business world, where industries are under pressure to accomplish useful imminent commencing the bulky bind of data they gather on a habitual basis. Now the data will simply maintain to breed in volume and velocity in terms of big data. Tiered storage brings cost optimizations to the table that can guarantee organizations achieve the correct equilibrium between performance, capability, and cost on their big data clusters.

Storage tiering has extraordinary potential in this present reality where organizations are feeling the squeeze to increase valuable bits of knowledge from the extensive swathes of data they gather all the time; data that will just keep on developing in volume and speed. Layered capacity conveys cost enhancements to the table that can guarantee associations that accomplish the correct harmony between execution, limit, and cost on their big data groups.

6.3.2.6 Cryptographically enforced access control and secure communication

The big data analytics system can be made automated in terms of collecting data, but as a result, it enhances data loss during this period. Different encryption techniques and proper training of users can minimize the risk associated with it. Adequate protection method must be introduced as big data stored into clouds should be tested accurately. This can be done for extra protection of data on the user side while cloud service provider confers customary precautions review within the time frame.

Cloud service providers can be imposed penalties if the security standard does not meet up to the expected standards. The right to use of power strategy must be set up for authorized user access to both internal and external user sites. User authentication for the data coming from different sites must be controlled up to some level. To protect data from unauthorized access, a second-level security mechanism is very much useful, namely encryption (Figure 6.16).

The reason behind doing the same is the use of raw data as well as analyzed data. Confidentiality and integrity should be imposed on data as a measure of data protection.

Another aspect of the data security mechanism is the use of antivirus and firewall protection. Today the trend is the creation of special attacks like ransomware – a type of malware that is very common. They are breaching the defense of computers throughout the world on a regular basis. Security mechanism includes some other small techniques like disconnection of user devices with servers restraining important information when they are unused. The security mechanism must pay attention to the fortification of the function, rather than just safekeeping the device. Proactive and reactive security skill should be supplied to big data.

Data visibility manages to different entities differently as there are two methods for organizing, like systems, individuals, and organizations. The primary technique systematizes the visibility of data in the form of preventive admittance to the main scheme, for example, the OS itself or hypervisor. The secondary technique encapsulates the data itself in a self-protective shell by means of cryptography.

Symmetric encryption
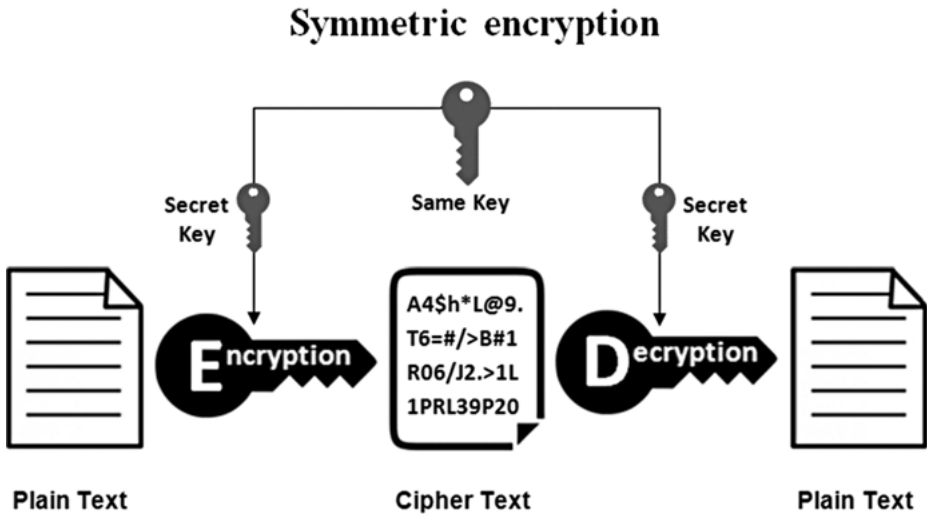
Figure 6.16: Cryptography as encryption and decryption.

Both the approaches put in their remuneration and detriments. In the past, the foremost procedure was simpler to carry out and, as integrated among cryptographically protected statement. These may be customary on behalf of the prevalence of computing and communication infrastructure.

The conventional security methods legitimate for an assortment of protective intimidation. They include replay attacks, password-guessing attacks, spoofing logins, intersession that has chosen plaintext attacks (Kerberos-specific attack), and session key's exposure. These attempts are really common in big data security techniques. They are pertinent to any of the customary verification schemes. The peculiar and more protected system, namely Kerberos fails sometimes. Today there is a need for implementing a scheme that can forestall the attempts.

Solution

Instead of using cryptography in security control, steganography can be used where data is covered up by some other media. In the case of cryptography, data are actually changed in its format but steganography creates an illusion for the intruders who want unauthorized access to user data.

To address the security problem like access control in big data is setting a revision in the secure remote password (SRP) protocol to have room for the access control of the clients in authentication level. This is the aim of assigning the access labels to the big data users to limit their access rights in the big data environment. Simple remote password protocol is a secure password-based authentication and key management protocol (Figure 6.17).

This protocol authenticates the clients to the server using a password-like mystery. This mystery must be known to the client only. No other secret information is needed to remember by the client. The server stores the verifiers for every user to

Smart grid circumstances	Threats/challenges	Solutions provided
Data integrity and privacy	Imposing access control	Group signatures with trusted third parties
	Search and filter	Identity and attribute based encryptions
	Outsourcing of computation	Encrypting techniques supporting search and filter
	Integrity of data and preservation of anonymity	Fully homomorphic encryption

Figure 6.17: Cryptographically enforced access control and secure communication.

authenticate the client, but if this verifier is compromised by an attacker, it cannot be used to by the attacker to impersonate as a client. The foremost benefit of the SRP protocol above further verification methods is that there is no need to store any password equivalent data and the systems are immune to the password attacks. When the client is verified by the server, a cryptographically strong secret is exchanged by the SRP protocol between the communication parties to pass securely.

Ultimately, the cryptographically enforced access control and secure communication solution present as a conclusion are identity and attribute-based encryptions, encrypting techniques supporting search and filter, fully homomorphic encryption, and group signatures with trusted third parties.
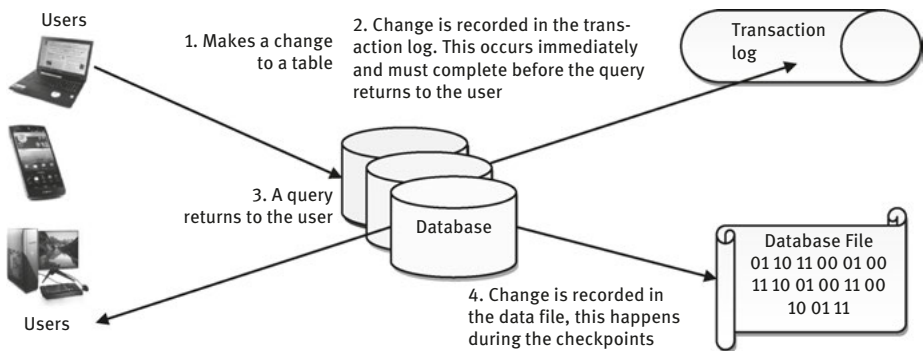
6.3.2.7 Transaction logs

The protected data storage and transaction logs are really much linked up in the midst of data warehousing, management, processing, and hence make the security tribulations. Sometimes external security threats may distort the memory due to unauthorized access within big data. This deformed data is then transferred from one source to another while the data transmission occurs between different nodes and users. The tiered approach has already described how data is stored in different tiers. This case is likewise for the transaction logs, and both data and transaction logs are stacked out in multitiered storage media. The data moves frequently among tiers manually.

This procedure assures the developer direct authority over vitally what data is moved and when. On the other hand, as the data set volume produces exponentially, scalability and accessibility require autotiering for big data storage management. The autotiering solution does not have any track of where the data is stocked up. This leaves a new problem of protected data storage. For instance, on that point is an establishment that wants to ingest data from different departments of their own. There are dissimilar types of data (every bit per data tiering approach) available within the

departments. The data that are virtually not used and mostly used both exist in parallel. An autotiering storage system will help a great deal and can put away the money via transferring the less used data to a lower grade and thus along. It may find that the data stored in a lower tier is sensitive information. Usually the lower level data have less security as companies do not desire to spend much for the unused ones, only in this event, they must convert the policy to send data toward lower level or increase the security of data.

Exchange logs can develop wild when not appropriately kept up. Each time data is changed in the database, records get added to the exchange log. On the off-chance that an exchange is kept running against an extremely huge table, the exchange log must record and store those data changes simultaneously until the exchange is finished. Since the log composes its data to disk, this can gobble up a ton of disk space all around rapidly (Figure 6.18).

Figure 6.18: Transaction log (basic working procedure).

Solution
Data tiering in big data analysis plays a major role, especially when an autotiering scheme is adopted. This autotiering must be fully log based so that the movements of data between tiers are noted and side by side the sensitive data should be marked properly and increase the security of the tier itself. Now, this may be the problem for huge data to provide high security, for this case data should be isolated and kept in a safe place. The encryptions, especially policy-based encryption, and a signature on data safe secure data storage and transaction logs. The proof of data possession and periodic audits of data are also recommended (Figure 6.19).

6.3.2.7.1 Recuperation models
To begin with, pick the correct database recuperation model. There are three that can be utilized with an SQL server database, yet the most normally utilized models are SIMPLE and FULL. The third alternative, BULK_LOGGED, is normally utilized briefly for expansive activities where execution is basic. The penance to get this execution help is potential information misfortune if something turns out badly.

Figure 6.19: Secure data storage and transaction logs.

6.3.2.7.2 Development limits

One method for evaluating a proper most extreme exchange log measure is to set it at any rate the extent of the biggest list in the table. Since revamping the list requires in any event a similar measure of room in the exchange log as the file itself, the exchange log will dependably have enough space to finish the task.

6.3.2.7.3 Backup

At last, take steady and robotized reinforcements. There are numerous methods for setting this up (and is past the extent of this post). On the off-chance that the recuperation model is set to SIMPLE, the main alternative is to perform reinforcements of the information; backing up the exchange log is beyond the realm of imagination. Every reinforcement will be a solitary depiction of the information in the database at the season of the reinforcement.

In the event that the recuperation model is set to FULL, at that point of exchange log reinforcements enable the database to be completely reestablished to a specific point in time and aren't obliged to just a single explicit minute in time. Log reinforcements likewise play out the significant errand of stamping existing log records as dormant. Idle log record space would then be able to be reused by the exchange log when it returns to that area of the log document since it keeps in touch with the sign in a consecutive request. It's likewise essential to take note of that playing out a "full reinforcement" from the reinforcement alternatives does not back up the exchange log, notwithstanding when the database is in the FULL recuperation model.

6.3.2.8 NoSQL

The contemporary big data analysis is the problem of synchronization between database systems, which host the data and make available SQL querying, by means of data analytic [28] packages that carry out numerous forms of non-SQL processing, like data mining and statistical analysis.

Nonrelational data stores recognized as NoSQL databases are even under pressure concerning security infrastructure.

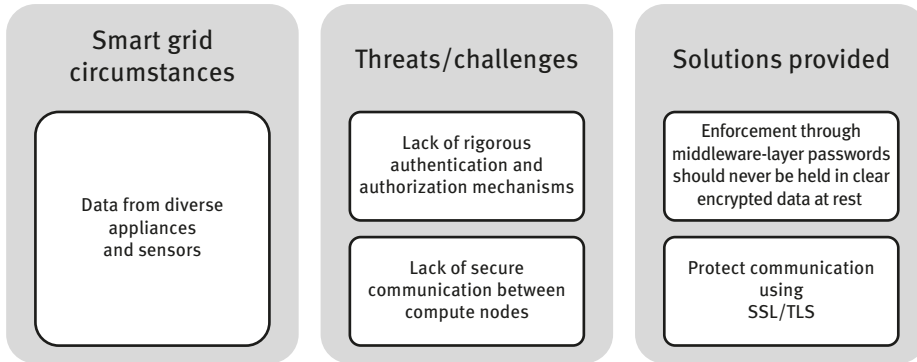The basic difference between NoSQL and SQL maybe thought of the difference in nonrelational and relational structure of data. Data stored here in has NoSQL more like document layout, whereas SQL has more table setup. This provides NoSQL to be more flexible and trouble free to deal with the new data models than in SQL. NoSQL is a open-source software; it signifies the minimum cost associated with it and can act with low configuration hardware too. As a consequence, small software companies make use of it. Quick processing on big data is obtainable using NoSQL tools. Elastic scalability is also useable with this database. Here no database models are employed with NoSQL and thus a huge time is preserved. All the above points are just opposite to SQL (shown in Figure 6.20). Examples of NoSQL database tools such as MongoDB, CouchDB, CloudDB, and Bigtable are set up to use a decent variety of difficulties presented by the examination world and subsequently security was never part of the model at any time of its origination stage (Figure 6.21). Most engineers utilizing NoSQL databases by and large incorporate security in the middleware. No financing is offered by NoSQL databases for implementing it expressly in the database. Such security rehearses represent extra difficulties. As far as hotel and preparing the colossal majority of data, associations managing enormous unstructured informational collections may pick up the preferred standpoint by moving from a customary social database to a NoSQL database. Subsequently, the security of NoSQL databases depends on outside upholding systems.
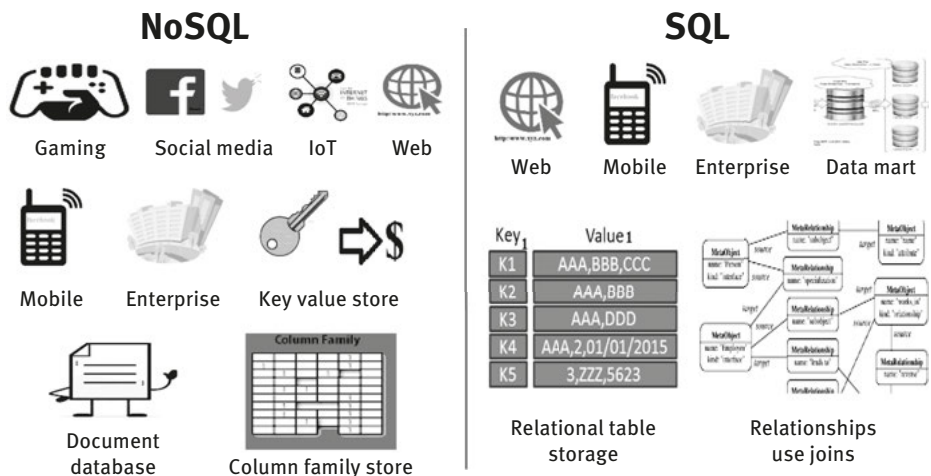
Figure 6.20: Basic difference between NoSQL and SQL.

For instance, well-molded answers for NoSQL infusion are as yet not built up. Each NoSQL databases was in parliamentary law to drop down the security episodes. The

Smart grid circumstances	Threats/challenges	Solutions provided
Data from diverse appliances and sensors	Lack of rigorous authentication and authorization mechanisms	Enforcement through middleware-layer passwords should never be held in clear encrypted data at rest
	Lack of secure communication between compute nodes	Protect communication using SSL/TLS

Figure 6.21: Security best practices for nonrelational data stores.

general public must continue through security arrangements for the middleware adding things to its motor and toughen NoSQL database itself to coordinate social databases without settling on its utilitarian qualities.

NoSQL database encryption solutions

Nonrelational or NoSQL databases are adaptable database arrangements that are composed for a lot of changed data types. Since they include more than customary database tables – utilizing objects and lists rather – they require an alternate way to deal with enormous data security.

Clients would now be able to ensure data in any NoSQL database, including driving database merchants, for example, MongoDB, Cassandra, Couchbase, and HBase.

Leverage file framework-level encryption answers for guaranteeing the documents, organizers, and offers that contain the files and articles listed in the NoSQL outline.

Coupled with strategy-based access controls, clients hold a fine purpose of control regardless of the gigantic information volumes.

Application-level big data encryption or tokenization arrangements append security legitimately to the data before it ever is spared into the NoSQL diagram.

Operations stay straightforward to the end-client while the database holds its capacity to direct inquiries, and convey data without reductions in execution.

6.3.2.9 Proper encryption technique

To ensure that the most touchy private data is completely secure and just open to the approved elements, data must be encoded depending on access control approaches. To guarantee confirmation, course of action, and decency among the dispersed substances, a correspondence system that is cryptographically verified must be completed. Delicate data is for the most part put away decoded in the cloud.

Lack of designing security measures is also creating security problems such as encryption, policy enablement, compliance, and risk management. If these

things are needed, they should be built on their own. Data masking policies and aggregating datasets may be used as security measures. Here re-identifying individuals are the proper tools that may perhaps locate datasets back simultaneously. Confining solitude might show the way to augmented safety measures menace, particularly if the data caught up be full of responsive and commercial information. Assortment of data is another important security measure, where data provider provides structured or unstructured data, but both are used by high-, middle-, or low-level users. This is the newest among all technologies using in today's computing world. It is very clear when any technologies are not well understood, certainly they become vulnerable.

The principle issue to scramble data is the win or bust recovery strategy of encoded data, which limits clients from effectively performing finely grained activities, for example, sharing records or ventures. Quality-based encryption (ABE) reduces this issue by using an open key cryptosystem where ascribes identified with the data scrambled serve to unscramble the keys. Then again, the decoded less touchy data helpful for investigation must be conveyed in a safe and settled upon way utilizing a cryptographically secure correspondence system.

Hadoop encryption solutions

The SafeNet data protection portfolio can secure data at different focuses in the Hadoop engineering – from Hive and HBase to singular nodes in the data spill with Gemalto, clients have a decision. Incorporate straightforward application-level security by means of APIs to ensure data without changing their database structure.

Select a column-level answer for Hive that licenses typical questioning. Obtain a file system-level arrangement with policy-based access controls. Every Hadoop big data encryption and tokenization arrangement is totally straightforward to the enduser and configuration saving encryption usefulness implies that customers will keep on benefitting from the examination instruments that suck up additional incentive from developing information stores.

Come up a center ground among obscurity and validation through a group signature, a cryptographic configuration where individuals can sign their data yet can just be named as individuals from a gathering. Just believed outsiders can perceive the substance (Figure 6.22).

6.3.2.10 Data mining

The use of big data can make out intrusion of security, obtrusive advertising, diminished common opportunities, what's more, increment state and corporate control. The special user data that appears really simple can be made usable when they make in use to predict something after analyzing it.
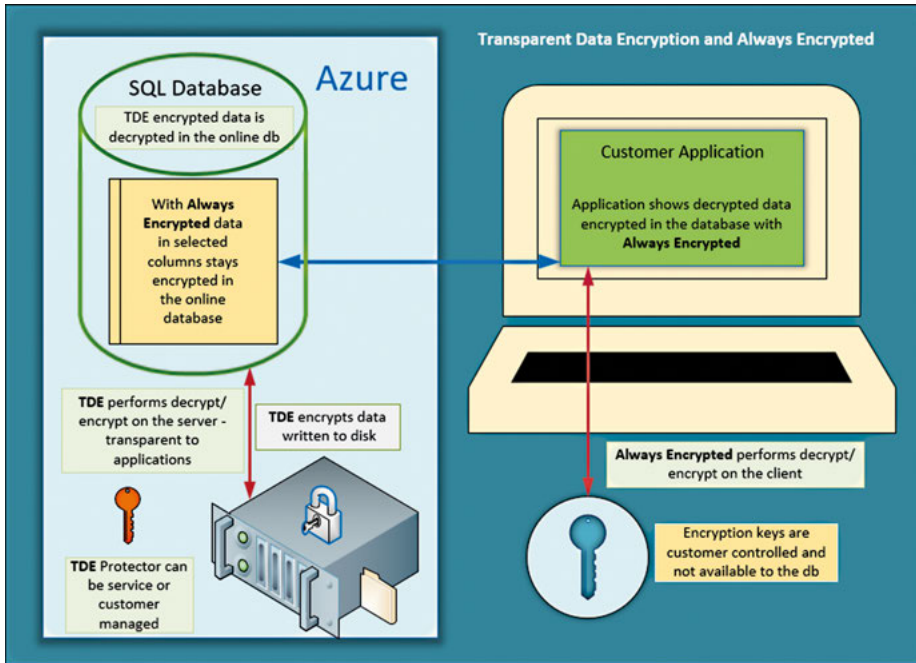
Figure 6.22: Transparent data encryption [29].

Anonym zing data for investigation isn't adequate to continue client privacy. For instance, Netflix (an American worldwide entertainment organization that blesses with spilling media and video on interest on the web) confronted an issue when clients of their anonym zed dataset collection were perceived by associating their Netflix motion picture scores with IMDB scores. Consequently, it is critical to decide rules and proposals for averting incidental security divulgences. User data marshalled by organizations and government experts are steadily mined and dissected by inside investigators and furthermore by outside contractual workers. A malevolent insider or unapproved accomplice can manhandle these datasets and get private information from clients. Also, knowledge organizations require the amassing of immense amounts of information. Hearty and adaptable, protection-saving calculations will expand the likelihood of gathering pertinent data to complement client security.

Solution

Data mining is a useful means to dispense with the vast amounts of big data, as important information can be pulled out, and then utilized to project future developments. The analysis of this data can help to solve problems and to shape strategies for predicting future trends (Figure 6.23).

A class of data mining is named as basket analysis. This analysis process reveals with reference to consumer buying options over a combination of different commodities at
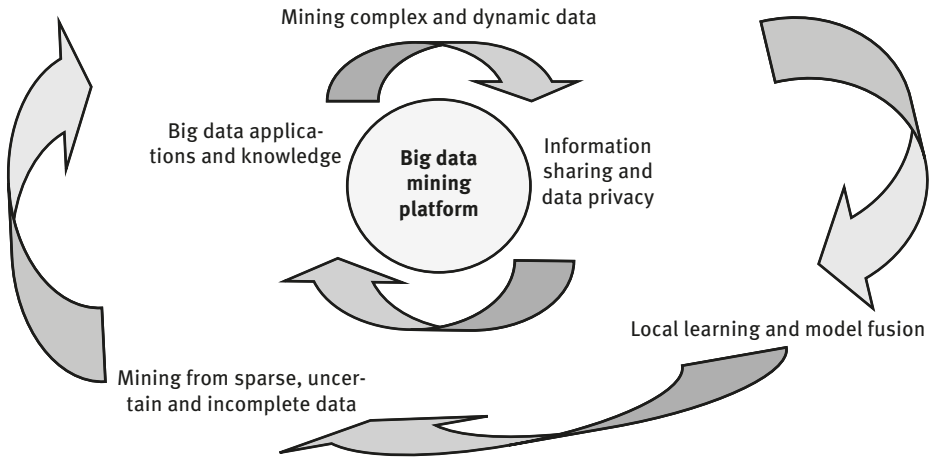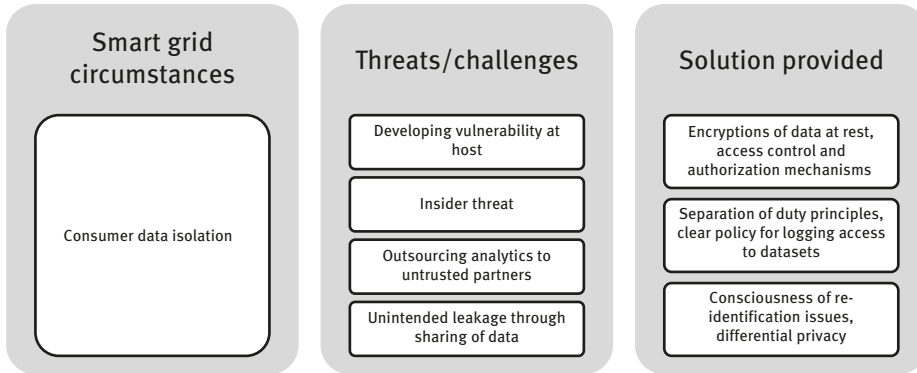
Figure 6.23: Data mining – system architecture.

a time for specific occupations. The pattern matching ability of data mining provides retailers spending capability and increase out the products bought collectively. Mining gives the analyzed data to the marketing section of the business, based on those different items they advertise to the customer.

Another brand of data mining is sales forecasting; the operative procedure for this is very simple. It memorizes which product a consumer bought at what timing. This forecasting technique now determines when the purchaser failed to buy the same product again and once more. For instance, coffee retailing is a lot higher in winter than in summer, so retailers are acquainted with increased stock for the winter months. It can be useful for the retailer to keep funds in monthly basis when sales (maintaining inventories) budgeting plays a significant part in modern commercial enterprise.

Finally, data mining helps to predict the customer needs at times and build proper databases in a particular area with the quantity and quality of different products (Figure 6.24).

This enables retailers, dealers, distributors, as well as manufacturers to know close to consumer needs and also launching new products at times. The encryptions are necessary for data at rest, so that no external attack can harm data. Access control and authorization mechanisms are likewise required to go through. The severance of duty, ideology, and comprehensible policy for logging access to datasets are the big factors in data mining. For the awareness of re-identification issues, differential privacy should also be preserved.

| Smart grid circumstances	Threats/challenges	Solution provided
Consumer data isolation	Developing vulnerability at host	Encryptions of data at rest, access control and authorization mechanisms
	Insider threat	Separation of duty principles, clear policy for logging access to datasets
	Outsourcing analytics to untrusted partners	Consciousness of re-identification issues, differential privacy
	Unintended leakage through sharing of data	

Figure 6.24: Scalable and compostable privacy-preserving data mining and analytics.

6.3.2.11 Granular auditing

In real-time security settings, the system keeps an eye on each and every attack with proper notification to the administrator of the system. Theoretically, the system is 100% protected, but in reality, it may not happen every time. There are many things to settle on like missed hit auditing of information is very much essential. Periodic system audit information is required to analyze the system to enhance security. What information is to be checked, what to ignore, what and when went wrong at times with other processes involved in it are also obligatory. The acquiescence, regulations, and forensic exploration are significant. System auditing is mandatory in case of distributed data processing applications in big data applications as data volume is so big that a standalone system cannot process them properly. Distributed systems enable the computer hardware like gateway, router, server; client computer system of numerous types, thus auditing system is put into practice over these systems as well. The auditing system must involve with the software systems too, which includes application software enabled and also the system software (OS).

The data-level safekeeping assistance could be comprehended by means of the pattern given. To analyze database activities, a standard statement is produced by the sample organization. The database administrators (DBA) analyze and review activity reports of their databases. The data accumulation and filtration are different jobs; DBAs have the farm duties to sort out data that is not important to them.

Figure 6.25 describes two managers, namely DBA manager and audit director. The DBA manager for Oracle databases can only see the data that belongs to the same database. This procedure is the same for the DB2 and MySQL databases as well. The DBA manager has permission to access all the data stored in the different databases. There is an audit manager associated with the auditing section for in-house audits. The SOX and HIPAA are two different auditors work under the same
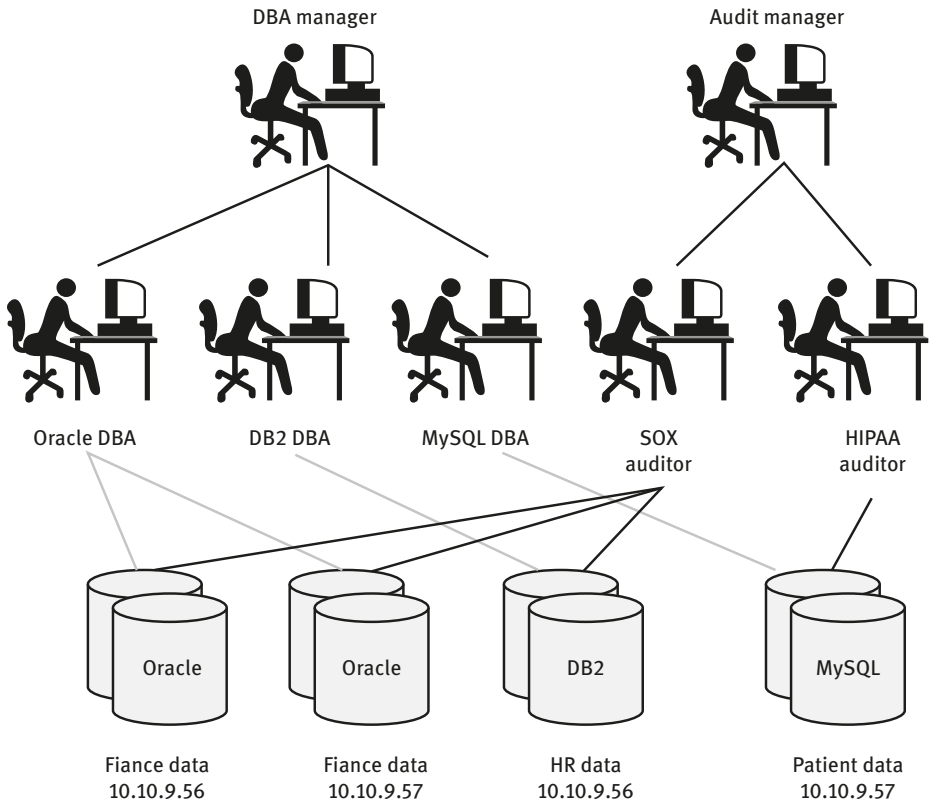
Figure 6.25: Granular audit process.

audit manager. The SOX auditor is accountable for the monetary data, i.e. sales, salaries, and orders, no matter where that data are stocked up. The HIPAA auditor is used in healthcare systems. It audits, data associated with patient information, no issue where that data is accumulated (Figure 6.26).
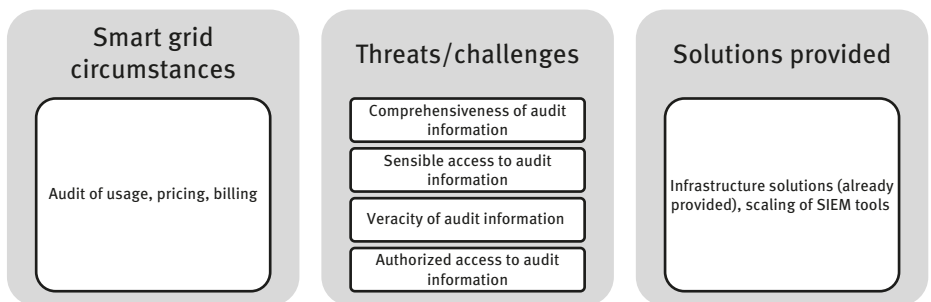
Figure 6.26: Granular audits.

Solution

Granular access audits in the majority of the cases can facilitate to come across the attack. Audit trials disclose the reason behind the nondetection of data at the beginning phase. The most significant mechanism of auditing may be talked about as the completeness of the required audit, well-timed approach to audit information, the reliability of the information, and endorsed entrance to the audit information.

Successful audit trial guarantees to incorporate the appropriate procedure and technologies in the big data infrastructure, including function logging, SIEM, forensic tools, and enabling SYSLOG on routers.

Split big data and audit data to differentiate among responsibilities. The audit data looks upon information concerning what has turned out in the big data infrastructure; however, it must be reserved taking, apart from the "regular" big data. A dissimilar network subdivision or cloud should be set up to host the audit system infrastructure.

6.3.2.12 Granular access control

The big data scheme acts as a significant task in data over the network of networks and storages. They carry out an excellent job regarding execution and versatility. But alas! It finds nearly no protection in mind. Existing RDBMS applications are secure enough with several protection attributes regarding access control, clients, tables, and rows and even at the cell level. However, a change of essential challenges put off big data solution to endow with comprehensive access control. The most important and foremost involvedness with course-grained get to components is that data that could somehow or another be shared is frequently cleared into an increasingly prohibitive gathering to ensure security.

Here granular access control is essential for diagnostic frameworks to adjust to this continuously progressive complex security environment. Keeping a log of jobs and experts of clients is one of the dangers alongside keeping up access marks crosswise over diagnostic changes. Big data examination and distributed (cloud) computing are progressively cantered around taking care of assorted dataset collections, both as far as assortment of diagrams and necessities. Legitimate and strategy confinements on data originate from different sources. Security arrangements, sharing understandings, and corporate strategy additionally forces prerequisites on data that dealt with. Dealing with this overabundance of the confinements has so far brought about expanded expenses for creating applications and a walled pationursery approach in which few individuals can take an interest in the analysis.
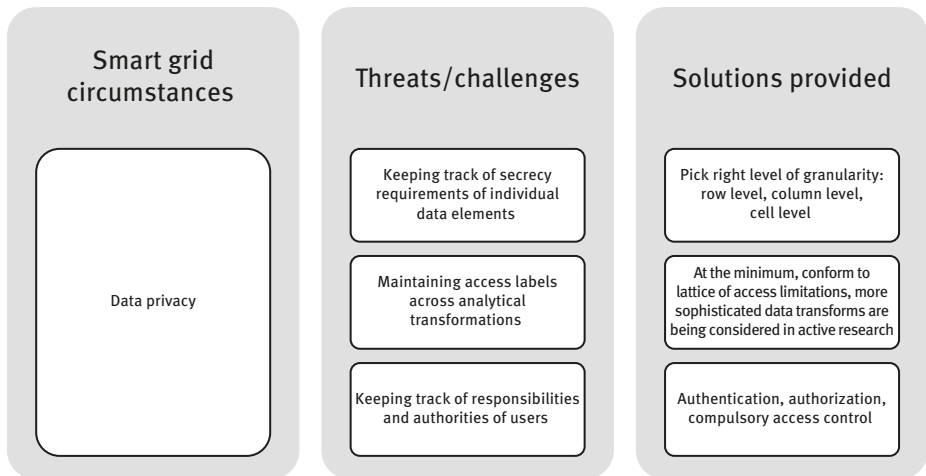
Solution

Granular access control as with any background is significant in a cloud environment to make certain that only the right people can have the right to use the right information and that the wrong people can't. The granular access control is vital to achieving this, and the cloud security alliance has acknowledged three specific tribulations in the realm of cloud data admission authority such as tracking confidentiality necessities that meant for individual essentials, supervising roles and authorities of users, and properly implementing secrecy requirements with MAC (Figure 6.27).

The resolution includes addressing these tribulations: associations may want to put into practice get to controls in the foundation layer, rearrange multifaceted nature in the application space, and receive confirmation and required access control. Nonetheless, scaling these technologies to the levels necessary to accommodate big data can present their own set of unique challenges. Use the typical single sign-on (SSO) method to condense the organizational work caught up in supporting a huge user base. SSOs transfer the trouble of user verification from administrators to publicly obtainable scheme.

As a conclusion, the data must pick the right dimension of granularity at the row level, column level, and cell levels. At the base, comply with cross section to get to impediments; progressively complex data changes are being considered in dynamic research. The authentication, authorization, and compulsory access control are mandatory.

Smart grid circumstances	Threats/challenges	Solutions provided
Data privacy	Keeping track of secrecy requirements of individual data elements	Pick right level of granularity: row level, column level, cell level
	Maintaining access labels across analytical transformations	At the minimum, conform to lattice of access limitations, more sophisticated data transforms are being considered in active research
	Keeping track of responsibilities and authorities of users	Authentication, authorization, compulsory access control

Figure 6.27: Granular access control.

6.3.2.13 Data provenance

The information and process integrity are needed in big data. Thus, the term data provenance comes into play, which integrates them by reporting the entities, frameworks, and procedures working on and adding to data of intrigue. The lifetime of data and its starting place must confer as unchangeable chronological evidence. The large provenance graphs and diagrams are analyzed to perceive metadata dependence for protection, and discretion application is computationally exhaustive. The beginning of an application or the place (process) of creation must need to know to the assortments of key security applications. This source of data is important for companies based on financial trading. They need to know the origin, process, precision of data for further research on market trends and future forecasting. Data provenances endow with the safety measures and they are completed within the time frame and involve speedy algorithms to switch the provenance metadata restraining this information.

Big data provenance architecture may be partitioned into few parts as big data admittance, distributed big data platform, provenance, and application employing provenance. Every partition will get in touch with particular live data, then database engines calculate the provenance data from there. The above diagram describes a reference design; system developers force choice as well as make mind up on component in each sub-framework that is dependent on the focused provenance practice circumstances along with competence.

The subsystem called big data access describes the most proficient method to get the various sorts of big data for provenance and distributed systems. System developers are required to find out the finest method, or tool to access the detailed datasets. Further work is needed for experimentation requisite on the datasets. The synchronization involving data and computer system is extremely desirable. Based on this synchronization, the provenance following the execution is an important aspect.

Distributed big data sub-framework manages development and execution support for big data applications and capacity support for big data provenance over appropriated big data platform (Figure 6.28). Before executing with the definite big data engine, the application designers will choose one big data work process framework to construct their submission. Some systems similar to Spark are capable to take action as both workflow production tool and distributed data parallel (DDP) execution engine. By means of DDP programming models, a big data submission can be constructed by covering inheritance tools or straight programming. The provenance storage too requires selecting the appropriate (distributed) databases or file handling systems (Table 6.1).

The provenance sub-framework settles on which provenance information will be reported and how to record it. The provenance can be separated into three extents: data, lineage, and the environment. Revamping the definite condition of the investigation over these three measurements is basic to repeat any information-driven logical analysis. Data provenance catches the condition of the input halfway
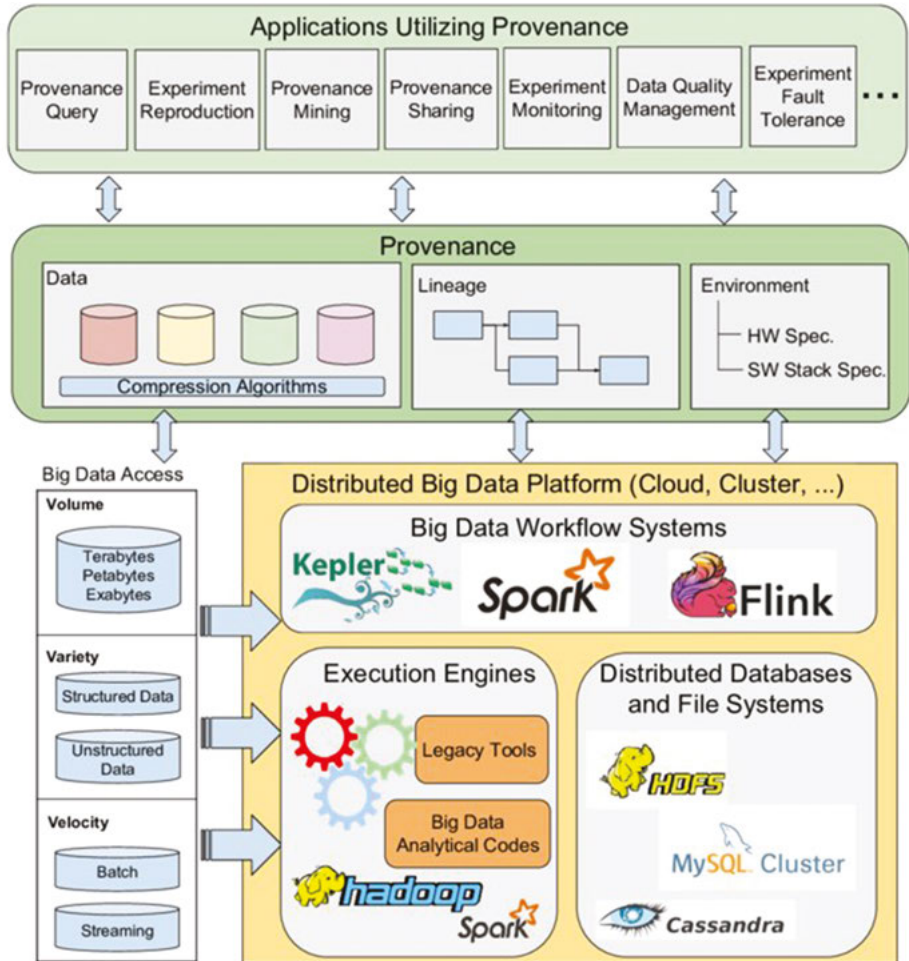
Figure 6.28: Data provenance architecture [30].

and yield data at the season of the investigation. The fundamental pressure calculation is favored by this provenance relying on the data informational index. Lineage provenance arranges the computational bustle of the trial, which is caught by putting away the guidelines that actuated on these information datasets. System provenance unites data concerning the careful condition of the framework design, which assesses both equipment determinations and framework programming details (OS, libraries, outsider apparatuses, and so on).

Large prospects of big data provenance carry on how we could make use of it. Big data provenance can be worn for provenance question, test multiplication, provenance mining, try checking, information quality administration, analyze adaptation to noncritical failure, and numerous others as an uncommon sort of big

Table 6.1: Big data provenance work comparison [30].

	Provenance recording	Applicable big data engines	Provenance usage	Environment provenance recording
Kepler	Parallel recording in a MySQL database	Unmodified Hadoop	Parallel query through MySQL Cluster	N/A
RAMP	Parallel recording in files	Extended Hadoop	Backward provenance tracing	N/A
HadoopProv	Parallel recording in files	Modified Hadoop	Parallel query through index files	N/A
Pig Lipstick	Parallel recording in files	Unmodified Pig/ Hadoop	Graph operation-based query	N/A

data. Each capability can be an individual application or an internal module in a superior scheme.

Big data workflows need to discover modeling and capturing provenance information by a small number of key learning. Here we make available a little outline of the efforts that attempt the challenges. Several new studies on big data provenance are not comprehensive in view of the fact that we spotlight on those that are in the circumstance of DDP also, and that have test analysis.

Kepler distributed provenance framework is done on big data provenance. It recommends a data model that can catch provenance contained by MapReduce occupations just as the provenance of non-MapReduce work process errands. It builds up the Kepler DDP engineering to record and inquire provenance in a disseminated manner on a MySQL cluster. It additionally offers an API to request the formed provenance. The WordCount application and a bioinformatics application called BLAST carry into play with the versatility of gathering and questioning provenance.

RAMP (reduce and map provenance) is an augmentation to Hadoop that underpins provenance catch and following for MapReduce work processes (). RAMP catches fine-grained provenance by wrapping Hadoop APIs. This constant wrapper-based methodology is perfectly clear to Hadoop and clients. RAMP implements various realities working expense all through provenance catch and empowers skillful in reverse following.

HadoopProv changes over Hadoop to incorporate provenance catch and investigation in MapReduce employments. The goal is to diminish provenance-limit expense. It follows to treat provenance in the map and reduces fragment freely. It likewise defers the structure of the provenance chart to the question stage by mixing transitional keys of the map and reduces provenance files.

Pig lipstick suggests a provenance system that combines database-style (fine-grained conditions) and work process style provenance (coarse-grained conditions)

Figure 6.29: Data provenance.

over Pig Latin. It advances a far-reaching and squashed diagram-based portrayal of fine-grained provenance for work processes that yield a more extravagant chart model than the OPM-Open Provenance Model standard utilized in work processes. It characterizes three graph revolution operations to ease the study of fiction work-flow analysis queries.

Solution

Cloud computing involves big data applications within it; thus, provenance of metadata matures in volume. They tend to be huge and thus intricacy grows very high. There are three most important intimidation to protect provenance metadata in big data relevance such as faulty infrastructure mechanism, infrastructure external hit, and infrastructure in the interior hit.

Secure data provenance is necessary to solve these threats. The system needs to improve trust; the external attack should be condemned. The usability of secure provenance can be accomplished by protecting origin finding technique, and data admittance is tuned up to some degree. The system is based on clouds, and the infrastructure of the cloud setup solves the problem where collection and preventing outside attacks, pairing a fast, lightweight authentication technique with current provenance tools are present. The insider attacks must be prohibited, for that a dynamic, scalable access control product is essential. To maintain the employability, connection ability of provenance graphs fine-grained access control technology offers data access attribution in big data application.

Design goals must be identified in big data provenance. The layer-based architecture is followed, where the access control mechanism is needed to address security. It can hold equally the structured and unstructured types of data, but this can handle simple queries only. For handling complex queries we will add additional components.

Finally, as a conclusion, it can be assured that the authentication techniques, message digests, access control through systems, and cryptography are the supporting points.

6.3.2.14 Other problems

Cloud computing technology enables dynamic collection of colossal data from different ends or nodes. These types of data generated by different organizations include unstructured as well as semistructured data. These data are commonly referred to as big data. These data are needed for security from both ends (internal

as well as external), privacy for specific data. Cloud computing involves equally with affirmative and unenthusiastic belongings.

Fake data generation is another problem in the big data analysis, where security breaches made by cybercriminals fabricate data and store it in the same place as it was there. The big data technology used so far is implemented through open-source code that can be modified by an individual to have proper knowledge about it. The attack is done using any one of the user sites that cannot be recognized, checked, or imposed penalties in big data; at the same time, the server side is also unaware about what measures to be maintained.

The most astonishing thing about this methodology is that the phony data doesn't have to coordinate the qualities of genuine data precisely. That data-to-data coordinating is exceptionally difficult to incorporate with manufactured data, so it is uplifting news that it is commonly not required. Rather, you simply need to coordinate key performance indicators (KPIs) among phony and genuine data when utilized as contribution for the procedure of intrigue. Note that these KPIs are in respect to the particular issue and models you are working with. Android malware [31] sometimes cause problems in big data, and proper use of antivirus helps to remove them.

The volume of big data market increases exponentially where the growth of enterprise-level data is the fastest. It is relied upon to twofold at regular intervals, from 2,500 exabytes in 2012 to 40,000 exabytes in 2020. Figure 6.30 suggests that about big data market today (2018) is about $24.3 billion and will be $49.1 billion in 2020 [32]. These will not only tend to increase complexity but also the problem of accessing and storing security measures that should be improved up to some extent.

Till now the discussion is all about hardware, and software associated with it, but another part of big data security is left off. The data, software, and hardware everything maintained by human beings are sometimes taking the system unanimous. The devices logged in but not in use yield access to further workforce as well as the threat of admittance data passing through unrestricted Wi-Fi. Automatic user log table with actual work seconds should be taken for an internal purpose for individual systems, and proper training on usage of data is also be useful.

Another way out of the big data security problem is to use software tools that can analyze and monitor in real time so that the system can get the happenings instantly. Besides the actual hazard, the alarming system can find some artificial data works as threats called false alarm. It must have the capability to differentiate between false alarms and real threats. The tools produce a huge amount of network data; again, it is problematic to maintain this data. Now each organization may not afford the cost of tools as well as it needs to update the hardware system to execute 24 × 7 checking capability. Big data analytics itself can resolve for the said problem by enhancing the services of network protection. Log tables are maintained to take data about changes made on the network; they can also be used for anomalous network connections.

Numerous large organizations, which are exercised on big data, their leading apprehension are the security of the cloud-based systems. Malevolent assaults on
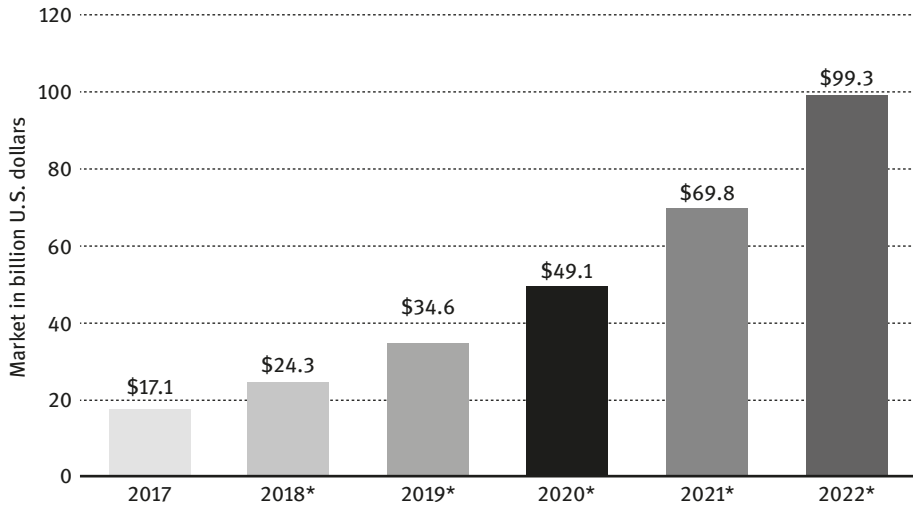
Figure 6.30: Year-wise big data market growth prediction.

IT systems are progressively thornier and an unmarked malware is being fashioned every now and then. Regrettably, endeavors that employed through big data come across such concern more or less on a daily basis. However, there is a clarification for each predicament and pronouncement one is certainly significant for the security apprehension.

The confrontation setup next to security, access control, compression, encryption, and compliance need to concentrate on in an organized manner as computing atmosphere turns out to be contemptible, and application atmosphere converted to the network along with system and analytics setting grow to be collectively more with the cloud. For the data section, further mechanism may possibly be added to grip composite queries.

Here, in the above section, we have marched from end to end and all the possible big data security challenges and have put down a number of suggestions intended for building big data processing and computation extra reliable and in turn making its infrastructure more secure. Some of the security issues discussed so far are very common in nature and are particular to big data crop up as of the numerous infrastructure tiers – (both computing and storage) employ in support of big data processing, the new-fangled computation infrastructures like NoSQL databases drawn on prompt throughput that are indispensable meant for huge dimensions of data are not comprehensively tenable as of most important security intimidation. The nonscalability of real-time scrutinizing practice, the diverse arrangement of devices that produce data, uncertainty by means of miscellaneous lawful limitations that come may lead to ad hoc approach for security and privacy.

For explicit big data tribulations, there must have a big ecosystem that exists. The subject matter here provides to illuminate unambiguous characteristic of the susceptible regions in the whole big data dispensation communications that require being investigated on behalf of definite intimidation.

The dispute with big data is that the unstructured character of the information creates it intricate to classify, model, and map the data as soon as it is confined to moreover accumulate. The dilemma is ended most horrible by the actuality that the data generally appear as of external sources, over and over again building it convoluted to substantiate its correctness.

Conflict resembling protection of data storage, data mining and analytics, transaction log and secure communication do subsist. The study on an assortment of safety measures faces up to roughly big data security and intact stack in large apparatus. Big data infrastructure security mechanism can be more stretched and modified to their directorial upbringing.

One of the modern approaches to secure big data is the use of blockchain. It could transform the technique we move toward big data. The quality of data would be improved and safekeeping on data has immediately two benefits to individuals and businesses as the blockchain is all around to grip information that can be digitized. The blockchains' prevalent improvement is its decentralization and thus no one owns the data entry or the integrity as it is established constantly by each computer on the network. Certainly, as much as necessary, blockchain and big data are a match made in heaven. The genuine issue at the present time is who will be the first to endow with the good number of appropriate and paramount skilled artificial intelligence/machine learning model working on top of distributed, transparent, and immutable blockchain generated data layers. The business to do this will roll in investments and engender enormous proceeds.

6.4 Conclusion

Big data analysis is flattering essential means for automatic determination of astuteness that is concerned in the recurrently stirring outline and secreted convention. This can facilitate companies to obtain an improved resolution, to envisage and recognize revolutionize and to categorize new fangled prospects. Dissimilar procedure in support of big data analysis as well as numerical analysis, batch processing, machine learning, data mining, intelligent investigation, cloud computing, quantum computing, and data stream preparing become possibly the most important factors. There is a gigantic open door for the big data industry in addition to plenty of possibility for research and enhancement.

Big data security solutions we have discussed so far may solve the current problem up to some extent. But it is to be said these are not the only solutions. They can

vary on a case-to-case basis, depending on the hardware devices or software applications used so far. The secure coding in a different language may fluctuate; encryption protocols are singular as per requirement. The layer of data may afford an unusual solution at times as authentication or authorization of data may be different. The proper analysis of static and streaming datasets in big data technology can assure to develop applications on medical and other scientific relevance and thus there can be a huge scope of business opportunities. The need for developing data-driven algorithms for creating applications on several other fields can open a new field of study.

Wish the problems and solutions given will be useful for the researchers and the big data fraternity at its best. The future research plan is to extend the other solutions possible as a new problem arises every corner of the world connected or to be connected.

Nomenclature

IoT	Internet of things
SaaS	Software as a service
Paas	Platform as a service
IaaS	Infrastructure as a service
NIST	National Institute of Standards and Technology
HDFS	Hadoop distributed file system
RDBMS	Relational database management system
MCC	Mobile cloud computing
QoS	Quality of service
SQL	Structured query language
IBM	International business machines
SIEM	security information and event management
DPF	Distributed programming frameworks
MAC	Mandatory access control
SSD	Solid-state drive
HDD	Hard disk drive
ABE	Attribute-based encryption
DBA	Database administrators
SSO	Single sign-on
DDP	Distributed data parallel
RAMP	Reduce and map provenance

References

[1]   A Cloud Security Alliance Collaborative research – Big Data working group. "Expanded Top Ten Big Data Security and Privacy challenges", 2013.
[2]   Rountree, Derrick., & Castrillo, Ileana. (2014), "The Basics of Cloud Computing", https://doi.org/10.1016/B978-0-12-405932-0.00001-3, ScienceDirect Elsevier Inc., pp 1-17

[3]   http://www2.sims.berkeley.edu/research/projects/how-much-info- 2003, Last Access: March (2019).

[4]   Koley, Santanu., Rashi, Agarwal., & Renu, Jain. "Cloud Computing: A New Era of Saving Precious Lives in the developing World". Skit Research Journal, 2015, 5(1), 2015(ISSN 2278-2508), 1–7.

[5]   Irion, K. Government Cloud Computing and National Data Sovereignty. Policy and Internet, 2012, 4(3–4), 40–71. DOI: 10.1002/poi3.10.

[6]   Zhou, Wenhao., Xiao, Yongbing., & Shen, Yulan. (2017) "Application of Cloud Computing in Telecom Operators" Creative Commons Attribution-NonCommercial 4.0 International License Computer System Networking and Telecommunications, WHIOCE Publishing Pte. ltd. (http://creativecommons.org/licenses/by-nc/4.0/).

[7]   https://www.economist.com/special-report/2010/02/25/data-data-everywhere., Last Access: May (2019).

[8]   Fung, PoTso., SimonJouet, Dimitrios P., & Pezaros, (2016) "Network and server resource management strategies for data centre infrastructures: A survey" Vol. 106, pp. 209–225.

[9]   Bohn, Robert. "Cloud Computing – A NIST Perspective and Beyond" Advanced Network Technologies Division, National Institute of Standard and Technology, U.S. Department of Commerce, 2016. (https://www.nitrd.gov/nitrdgroups/images/8/82/NISTCloudComputing.pdf).

[10] Koley, Santanu., & Ghosh, Shivnath. (2015) "A Life Saving Cloud Computing Architecture for Mobility of Sufferings", 9th INDIACom; INDIACom-2015; IEEE Conference ID: 35071, 2015 2nd International Conference on "Computing for Sustainable Global Development", 11th–13th March, 2015, ISSN 0973-7529; ISBN 978-93-80544-14-4, pp. 5.239–5.244.

[11] Akherfi, Khadija., Gerndt, Micheal., & Harroud, Hamid. "Mobile cloud computing for computation offloading: Issues and challenges." Applied Computing and Informatics, 2016, 14(1), 1–16.

[12] Wang, Yating., & Ing-Ray Chen, Ding-Chau. "A Survey of Mobile Cloud Computing Applications: Perspectives and Challenges". Wireless Personal Communications: An International Journal archive, 2015, 80(4), 1607–1623.

[13] Benslimane, Younes., Plaisent, Michel., Bernard, Prosper., & Bahli, Bouchaib. "Key Challenges and Opportunities in Cloud Computing and Implications on Service Requirements: Evidence from a Systematic Literature Review", CLOUDCOM '14 Proceedings of the 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, IEEE Computer Society, Washington, DC, USA, 2014. ©2014, ISBN: 978-1-4799-4093-6 doi>10.1109/CloudCom.2014.115, pp. 114–121.

[14] Koley, Santanu., & Ghosh, Shivnath. (2015) "Cloud Computing with CDroid OS based on fujitsu Server for Mobile Technology", Bilingual International Conference on Information Technology: Yesterday, Today, and Tomorrow, 19–21 February 2015, pp. 163–169.

[15] Koley, Santanu., & Singh, Navjot. "Cdroid: Used In Fujitsu Server For Mobile Cloud". Ge-International Journal of Engineering Research, 2014, 2(7), ISSN: (2321-1717), PP. 1–6.

[16] Koley, Santanu., & Ghosh, Shivnath. "Cloud Computing with CDroid OS based on Fujitsu Server for Mobile Technology". Skit Research Journal, 2014, 4(2), 2014(ISSN 2278-2508).

[17] https://moreign-technologies.co.za/2019/03/02/markets-trends-of-big-data-usage-in-health-care-during-the-next-5-years/, Markets Trends Of Big Data Usage In Health Care During The Next 5 Years, Last Access: May (2019).

[18] Reichman, O.J., Jones, M.B., & Schildhauer, M.P. "Challenges and Opportunities of Open Data in Ecology". Science, 2011, 331(6018), 703–5. Bibcode:2011Sci . . . 331.703R. doi:10.1126/science.1197962. PMID 21311007.

[19] Gaskell, Jill., Kersten, Phil., Larrondo, Luis F., Canessa, Paulo., Martinez, Diego., Hibbett, David., Schmoll, Monika., Kubicek, Christian P., Martinez, Angel T., Yadavi, Jagjit., Master,

Emma., Magnusonk, Jon Karl., Yaver, Debbie., Berka, Randy., Lail, Kathleen., Chen, Cindy., LaButti, Kurt., Nolan, Matt., Lipzen, Anna., Aerts, Andrea., Riley, Robert., Barry, Kerrie., Henrissat, Bernard., Blanchette, Robert., Grigoriev, Igor V., & Cullen, Dan. "Draft genome sequence of a monokaryotic model brown-rot fungus Postia (Rhodonia) placenta SB12". Genomics Data-ELSVIER, 2017, 14, 21–23.

[20] Demchenko, Y., Ngo, C., Laat, C. de., Membrey, P., & Gordijenko, D. (2014). "Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure" In W. Jonker & M. Petković (Eds.), Secure Data Management, Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-06811-4_13, pp. 76–94.

[21] Xueheng, Hu., Meng, Lei., & Aaron, D. Striegel.(2014) "Evaluating the Raw Potential for Device-to-Device Caching via Co-location" https://doi.org/10.1016/j.procs.2014.07.042 ELSEVIER Procedia Computer Science, Vol 34, pp. 376–383.

[22] Wang, Jianwu., Crawl, Daniel., Purawat, Shweta., Nguyen, Mai., & Altintas, Ilkay. (2015) "Big Data Provenance: Challenges, State of the Art and Opportunities" INSPEC Accession Number: 15679551, DOI: 10.1109/BigData.2015.7364047

[23] Terzi, Duygu Sinanc., Demirezen, Umut., & Sagiroglu, Seref. "Evaluations of big data processing." Services Transactions on Big Data (ISSN 2326-442X), 2016, 3, 1, 44–53.

[24] Geist, A. et al. (1996) "MPI-2: Extending the message-passing interface" In: L. Bougé, P. Fraigniaud, A. Mignotte, & Y. Robert (eds) Euro-Par'96 Parallel Processing. Euro-Par 1996. Lecture Notes in Computer Science, Vol. 1123, DOI https://doi.org/10.1007/3-540-61626-8_16, Online ISBN 978-3-540-70633-5. Springer, Berlin, Heidelberg.

[25] Kailai, Xu. (2017) http://stanford.edu/~kailaix/files/MPI.pdf, Last Access: May (2019).

[26] Alfredo Cuzzocrea, C. "Privacy and Security of Big Data: Current Challenges and Future Research Perspectives". ACM Digital Library, 2014, ISBN: 978-1-4503-1583-8 doi>10.1145/2663715.2669614, pp. 45–47.

[27] Boyd, Dana., & Crawford, Kate. (2011) "Six Provocations for Big Data " Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. doi:10.2139/ssrn.1926431.

[28] Cuzzocrea, A., Song, I., & Davis, K.C. (2011) "Analytics over Large-Scale Multidimensional Data: The Big Data Revolution!" In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104.

[29] Adhikari, M., Koley, S., & Arab, J. Sci Eng https://doi.org/10.1007/s13369-017-2739-0 (2017), "Cloud Computing: A Multi-workflow Scheduling Algorithm with Dynamic Reusability". Arabian Journal for Science and Engineering (Springer), Journal, 2017, 13369, Article: 2739.

[30] Koley, S., Nandy, S., Dutta, P., Dhar, S., & Sur, T. (2016), "Big Data architecture with mobile cloud in CDroid operating system for storing huge data," 2016 International Conference on Computing, Analytics and Security Trends (CAST), Pune, India, doi: 10.1109/CAST.2016.7914932, pp. 12–17.

[31] Bacci, Alessandro., Bartoli, Alberto., Martinelli, Fabio., Medvet, Eric., Mercaldo, Francesco., & Visaggio, Corrado Aaron. "Impact of Code Obfuscation on Android Malware Detection based on Static and Dynamic Analysis" 4rd International Conference on Information Systems Security and Privacy, Funchal, Portugal, 2018.

[32] Transparent data encryption or always encrypted?, Source: https://azure.microsoft.com/en-us/blog/transparent-data-encryption-or-always-encrypted/, Last Access: March (2019).

Shibakali Gupta, Indradip Banerjee, Siddhartha Bhattacharyya

7 Conclusions

Security of data is a major concern in this work-a-day world of information technology and communication systems. Unintended use and encroachment of data often lead to breach in the security of the underlying data. As a result, the integrity of data gets compromised, which leads to unsolicited system outcomes. With the advent of digital media technologies, a huge volume of data explosion has happened. Digital data content includes audio, video, and image media, which can be easily stored and manipulated. The superficial transmission and manipulation of digital content constitute an authentic threat to multimedia content engenderers and traders. Thus, the issue of security of data has become imminent.

Several approaches are in vogue for thwarting unwanted attacks on the integrity of data under consideration. The basic approach is based on ensuring a privacy policy of the intended users. The different authentication apparatuses help to launch the proof of identity of the end users. Access control regulations also add to the mechanism of data handling to a great extent.

Big data refers to datasets that are enormous in size as compared to normal databases. Big data generally consists of unstructured, semistructured, or structured datasets. Several algorithms as well as tools are in existence for processing these data within reasonable finite amount of time. The most prominent type of big data that has attached much attention is the unstructured form of data [1].

Big data is mainly characterized by the 4Vs (volume, velocity, variety, and veracity) [2–5]. Volume is a key characteristic of big data, which decides whether the information is a normal dataset or not. Velocity is the speed with a direction, which means the throughput or the speed of the data processing. It indicates as to how fast the information can be generated in real time to meet the requirements. Variety is important in this literature because it stands for quality and the type of data required in order to process it successfully.

This book is targeted to discuss the fundamental concepts of big data forms and the security concerns associated therein [7]. The first contributory chapter explains the common business models/platforms that use block chain as a platform for developing their processes based on digital identity. Each and every big data source or big database needs a security metric monitoring. The monitoring software collects various metrics with the help of custom codes, plugging, and so on. The next chapter describes the approach of modifying the normal metric thresholding to anomaly detection. The third contributory chapter deals with the social

Shibakali Gupta, Indradip Banerjee, Department of Computer Science, University Institute of Technology, The University of Burdwan, Burdwan, West Bengal, India
Siddhartha Bhattacharyya, RCC Institute of Information Technology, Kolkata, India

engineering aspect of big data hacking along with other hacking methodologies that can be used for big data and how to secure the systems from the same. The fifth chapter describes the information hiding and data consumption techniques in big data domain. The next chapter discusses some of the big data security issues with reference to some solution mechanisms.

Given the varied content of the book, the book would surely serve as a good treatise on the subject and would benefit the readers to grasp the inherent ideas of big data manifestation and security mechanisms involved therein.

References

[1]   Snijders, C., Matzat, U., & Reips, U.-D. "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science, 2012, 7(1).

[2]   Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development PolicyReview". martinhilbert.net. Retrieved 7 October 2015.

[3]   DT&SC 7-3: What is Big Data?. YouTube. 12 August 2015.

[4]   Cheddad, Abbas., Condell, Joan., Curran, Kevin., & Kevitt, Paul Mc. Digital image steganography: Survey and analysis of current methods Signal Processing, 2010, 90, 727–752.

[5]   Capurro, Rafael., & Hjørland, Birger. (2003). The concept of information. Annual review of information science and technology (s. 343–411). Medford, N.J.: Information Today. A version retrieved November 6, 2011.

[6]   Liu, Shuiyin., Hong, Yi., & Viterbo, E. "Unshared Secret Key Cryptography". IEEE Transactions on Wireless Communications, 2014, 13(12), 6670–6683.