Privacy and Security in Big Data Management

A Kiran
Research Scholar, JNTUH, Hyderabad

Abstract—Finding new ways to take advantage of big data, small or big organizations need high security techniques and efficient policies to guarantee their systems. In big data, it is very difficult to store data and analyse these data with conventional techniques. It is though that the traditional mechanisms are not efficient on big data security and privacy challenges. Different challenges include Hadoop security, cloud security and monitoring and auditing, etc. Big data deals with different data formats, data collection done from different sources, streaming of data may cause security problems the existing systems led to the violation of the individual privacy control. Data should be shared in a secure environment without violating people’s privacy to overcome such complex issues, it is necessary to implement some set of rules and regulations in performing data sharing, authentication process, in generating personal information, and misusing of the system. With respect to these problems, proposed system introduced some mechanism challenging security and privacy issues for both user and application-oriented.

Keywords—Big data, Data mining, Cloud Security, Hadoop Security, Key Management, Anonymization.

I. INTRODUCTION

An idea of big data has been introduced in the computer science, due to enormous use of information. Big data originally meant for collecting huge volume of data that could not be processed using the traditional computing, database tools with high efficiency in the aggregation of data, data managing capabilities social media analytics and real time data. Data analytics is the activity of examining large amounts of data. In big data the information came from multiple heterogeneous autonomous sources as with complex and continuously evolving like digital data, business data scientific data audio video, etc. . . . . The need for a big data generated from the large companies for a specific purpose of analysis of large amounts of data as they are dealt with a large set of information like Facebook YouTube yahoo and google also maintain the large amounts of information. Big data analysis targets at an extraction of useful information or extraction of selective patterns from large amounts of datasets. Big data is the most popular buzz word among the researchers whose size specifies in the volume rate of increasing speed in data processing and also specifies heterogeneous data-structured and unstructured, such data are very difficult to capture, manage processing and integrating of datasets with the conventional tools such as relational databases. Currently big data processing depends on parallel programming models like map reduce which will provide a platform for big data analytics. There are various challenges which are faced by the researchers and developers to maintain scalability heterogeneity and incompleteness availability and privacy of data currently we are using different tools for processing big data which is helping us to prove the difference between the imagination and reality world. Different challenges include Hadoop security, cloud security and monitoring and auditing key management and Anonymization etc. big data deals with different data formats data collection done from different sources streaming of data may cause security problems. The existing system failed to provide high security to the system by violation of the individual privacy control. Data should be shared in a secure environment without people privacy. Data to be generated, analysed, stored in a secure database. Data should be retrieved in less time and with high availability. It is necessary to implement some set of rules and regulations in performing data sharing, authentication process, in generating personal information, and in misusing of system resources.

II. STANDARD BIG DATA PLATFORMS

1) HDFS

It is a file system that delivers extraordinary throughput access to distributed data which helps to organize large amounts of data in an efficient manner. It consists of two nodes, namely master node and slave node. The master node is called primary name node and it is used to manage metadata. Slave node is used to store data, it is also called as data node, which controlled by the master node.

2) Hadoop Map reduce and Big table

Map reduce performs job scheduling and resource management in cluster environment. Google proposed a model for programming, map reduce and it’s distributed file system and Google File system
(GFS) [13]. It is made for programming in the Hadoop to store large volume of data in a distributed environment [14]. For storage in the distributed environment and process structured data for a large size of data sets Google introduced a new technology called a big table in the year 2006 [15].

Fig 2. Different platforms of big data

3) Hbases

It is a Column oriented data base, it is very scalable. Distributed database to store huge structured data. It is also called Hadoop database, which runs on the top of the Hadoop file system. It is a non-relational distributed system it initiates input and output operations to the map reduce. H base is a management system that is versioned, distributed, open source, and works based on the big table and google.

4) Pig

It is a frame work for parallel computation by using a high level data flow language. The pig is also a platform for big data analysis and processing. One of the main feature that pig frame work has is parallel processing, which is to handle very large data groups. Pig can process terabytes of data.

5) Hive

On top of the Hadoop, hive is a data warehouse frame work. By using hive we can write sql like queries to process and analyses the data stored in HDFS. The advantage of hive is scalability and performance it works like normal sql language, and runs fast over big data sets.

6) Sqoop

It is a software tool which is used to transfer bulk data between relational data bases into Hadoop environment. Sqoop performs important or export operations and data from or to external databases (HDFS or, Hive or HBase).

7) Zookeeper

It is designed to provide synchronization purpose, coordination services and group services. It maintains some configuration details, named registry etc. in Hadoop the zookeeper is very much important to find any node failures. If particular node is down it takes necessary communication protocol for a particular node failure.

8) Avro

It is a data serialization system. It relies on schemas for read and write data operations. Avro has popular features like rich data types, fast and compact serialization. It works in many languages like c, c++, java, ruby and python.

9) Mahout

It is used to maintain information about machine learning and data mining techniques. Mahout is a scalable machine learning algorithm primarily focuses on four areas- categorization, classification of data sets, data clustering and collective filtering.

10) Spark

Apache spark is for fast data analytics used to process data on a large data sets. Spark is incorporated with Hadoop technology having enhanced analytical engines which is very much faster than Hadoop map reduce.

11) Flume

It is used to maintain log information for each and every session. It is a reliable service to process and manage log information. It provides capability to the users to collect log data from dissimilar sources and process high volume of log data in real time.

12) SciDB

SciDB is an open-source to manage the data it primarily works on multidimensional data model. SciDB deals with petabytes it has many applications areas like geospatial, scientific research area, financial departments, and high scale business.
13) Nimble

It’s a tool which is used to implement parallel machine learning and data mining algorithms on map reduce for large volume of data sets. It allows the users to implement any a machine learning or data mining algorithms that are supported by any parallel programming models.

III. CHALLENGING ISSUES FOR SECURITY IN BIG DATA

Conventional solutions are insufficient when dealing with big data to obtain privacy and security. Access permissions, firewalls, Encryption schemes, transport layer security can be broken; source of data can be unknown; even unidentified data can be re-identified [2]. To overcome these reasons, some advanced techniques and technologies are developed to monitor, protect and audit big data processes. Considering the related survey, this paper has classified privacy and security issues for big data under 5 titles as Lack of High secure Design, Hadoop security, cloud security, monitoring and auditing, key management and Anonymization[4].

Fig 1. Different Challenging issues of big data

1. Lack of High Security Design

One of the main challenging issues that big data has facing in recent development is, it could not have a well-structured system design for security that supported by all big data platforms. The Main drawback of big data while dealing with high volume of data is it provides many high level capabilities for business, but they don’t have access in before. Many of the platforms don’t have specific standard encryption, decryption algorithms, policy ability, obedience, and other features of security. Another drawback is skills gap that many high level business organizations are facing. Only two or three people have expert knowledge as a staff, which is a tough task to challenging security issues.

2. Security for Hadoop

Hadoop is a distributed process model for processing and analysing large data sets. The main functionality of Hadoop is to provide a platform to work with trusted environments. Some safety measures have been started to provide Security. While constructing a Hadoop model it will guarantee that security and privacy of information should be taken care from unauthorized access. Hadoop contains three nodes- name node, data nodes, an edge node. Name node does the main functionality of HDFS (Data Processing is done here) [5], Data node is used to store data, and edge node maintains relationship between name node and data node. A security mechanism has been employed between the user and name node, using this mechanism user must validate them self to access name node. some hashing techniques used for authentication like SHA -256 which is a popular hashing technique, some Random encryption techniques are like RSA Algorithm, Advanced encryption Methods, and RC6 also used for data in order to protect from an attacker, they do not gain access on whole data.

3. Security for cloud

Cloud computing is mainly used to provide on demand services, resource pooling and broad network access. Now a day’s Data storage on cloud is one of the main problem, it generates new kind of attacks, therefore some precautions must be taken by the service provider because of this, a secure way to handle shared data on cloud platform as have been presented [6].There are several security measurement techniques like authentication, encryption, decryption and compression methods etc., authentication is done through email-id and password for every authorized user. To prevent data from security issues data encryption and decryption should be done; there are two or three backup servers should be maintained for back up from natural disasters. In every server data has been stored in an encrypted format if any server is damaged, remaining servers will provide the data by decrypting with a secret key. The classical encrypted technique is not enough for data security on cloud a new scheme to be proposed to secure a big data storage on cloud [3, 7].

4. Key management and Anonymization
Key management, which is a process of generating keys or distribution of keys to the users or to the servers. These keys are used to authenticate the users and servers. In big data, a fast and dynamic key management protocol is to be developed to generate the keys and for distribute the keys between servers and users for authentication because it is a big challenging issue in big data now. To increase the efficiency in using the keys, following two operations must be implemented 1) reducing the key search 2) restricting the passive attacks. Big data provides the services to the single or group of users, so every group need to share the keys in secure manner for increase confidentiality. To adopt this Diffi-helmen key exchange algorithm is widely used. Anonymization is the process of de-identifying an individual information or a particular person details from particular data base by using some methods. By applying such methods it is impossible or moost difficult to find individual information. There are several methods to overcome an Anonymization problem, but in big data it is not possible due to the scalability issue.

5. Monitoring and auditing

Monitoring can be defined as controlling a transform of information and observes the flow of data in a system. Auditing can be defined as regular inspections on security procedure for the entire system to catch the intrusions. Two techniques play a crucial role in providing security. Firstly, intrusion detection, which is used for to detect the intruders and observes the abnormal behavior of the system. Secondly, prevention, which is for preventing the system from intruders. It is more difficult to practice these two techniques on the whole network traffic. To overcome this problem a monitoring architecture for security is developed by analyzes network traffic, IP packet flow and http traffic [8]. The proposed technique includes processing and storing data in indifferent sources through Correlation techniques. In a big data environment, finding abnormal sequences and identifying the alerts should be found by Network security systems. An event monitoring system has been proposed in the big data security which consists of four modules: data gathering, data integration, data analysis, and data interpretation [9]. Data gathering contains security related information, network related device logs and action information. The data integration process is done by data filtering and classification techniques. Correlations and association rules are determined to catch events using data analysis. Finally, data interpretation provides graphical and statistical results which are useful to the database to find network behavior and respond to the events.

IV. ANALYSISOF DATA FOR BIG DATA

The big data analysis used to examine the information and obtain useful information from all heterogeneous sources[2]. Analyzing is carried out to make the data secure examining the system not attack patterns network traffic and security related information and alerts are useful to discover anomaly detection intrusion detection and identifying frauds by using big data has been accomplished[10]. By analysing abnormal behaviour and suspicious parents from security that are used to construct automatic production system real time application prediction of attack [9, 11, 12]. Big data store data in terms of terabytes to process and analysed Tera bytes of data would take a long time on a single system it is difficult task to store such a huge amount of data on a personal computer because the memory of RAM is used for normal calculations by such activity system will struck due to heavy load considering DMart as examples which is a popular shopping place. It is a place to purchase all needs in daily life. DMart data will be stored in huge Central service the customer will go to D mart store at least once in a week is Chand every customer shopping details will be maintained by storing all this information Like customer information shopping information and transaction details of every customer of D Mart that I may go round 1 for a week this information will help the developers to promote their skills which will increase the profit to do all these need of big data comes into picture by doing big data analysis in D Mart system which will benefit its existing customers as well as welcomes new customers to buy more gold from it stores.

V. APPLICATIONS

As data continuous to grow the big data will become progressively important too large and small commercial Enterprises the most important big data applications are

1. Fraud detection:

A big data platform that can observe privileges and transactions in real time recognizing comprehensive patterns across various transactions to discover electricity is from user as it shows how fraud detection can take place in maximum circumstances is discovered after the damage is occurring we can't predict the home at an earlier stage but some precautions are taken to minimize the damage and adjust few guidelines to prevent it from repeat once again. It is very critical job to the big data team to
detect frauds from a group of legitimate users fraud detection is one of the well-known challenging tasks in Big Data because big data dealt with huge data.

2. Social media data analysis

Social media is the best example of the big data which will give an exact insight into how the data is generated and how it is analysed. Now a days social media are trending topic it is important to analyse social media data such as Facebook YouTube Twitter and LinkedIn in the present world everyone has an account in each of these sites to share personal information with their friends or relatives by lighting pages on Facebook or tweeting on Twitter more that is generated now it is very important to know how this data is analysed big data playing if we roll in browsing analysing of the high volume of data that is produced by social media

3. Real time data analysis

Analysis of real time data is an important challenge for big data because data generation is enormously increasing day by day with the development of information technology. In real time use that are produced with the online discussion and debates on controversial topic it will provide millions of data by means of tweet and public opinion and online public feedback. In real time for example flicker it is a site for sharing public pictures which is receiving millions of photos everyday it is a very difficult task to analyse search data big data provides some analytical processing techniques to process analytical data which is stored offline also process by big data such as store the data transaction and decisions.

4. IT Environment

Big data playing a key role in IT sector every day. IT Department is producing high volumes of the logs to trace data it is a challenging task for it professionals to store process and manage companies data without big data it required for manpower and resources to educate all the information by hand when you have this data must go and examine in most of the big data applications it lost data using post publicly applicable many businesses with the large information will be benefited from speedily recognizing large scale patterns to help in getting and protecting from regular problems you have any organization welcome the opportunity to improve the performance of the organization.

5. Educational system

Now a days data mining is playing a vital role in the field of an educational system the process of examining knowledge for beginners occasion system which contains large data volume this field data is generated from students stands for management department important feedback information from professionals[17]. Some kind of information that is generated by clean observations. This data may be utilized to improve the performance of the system by address in the students mistakes and to improve the quality of the study by this method week and summaries students details from different aspects are we can measure the student’s details by the performance results which will help students in improving their higher education [18].

6. Healthcare

Big data playing an important role in the field of health care [19]. It stores patient details doctor information pharmacy details medical laboratory information list of different diseases and diagnosis information the process of storing and extracting of such information is a critical task for database using traditional system which deals with small datasets to become such issues big data archived some goal in predicting diseases and in diagnosing them[20]. Big data and hands its services towards Medical Services minimizing medical cost identifying The Rest by using big data it is possible to generate medical reports of the patient and status report for every patient and disease related information big data is playing a crucial role in analysing classifying storing each and every patient details.

VI. CONCLUSION

Big data required attention in data gathering, processing, storing, analysing and transforming for security and privacy concern. In this proposed system I explain basic challenging is use of big data and simple analysis of big data analysis with DMart as an example. According to the literature communication should be done through encryption and encryption every user should be authorized to access systems monitoring and auditing is carried out in a secure Manor to prevent from threats data privacy and security is the biggest challenging problems in big data so advanced security tools and techniques for security to be implemented in future solutions in preserving the privacy need to be developed on existing technology is should be enhanced for accurate results this study will help in understanding the fundamentals of big data and its basic
characteristics, platforms of big data security challenges in big data not only for no but also for the future purpose.

REFERENCES


[15] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes,Robert E. Gruber, Bigtable:


