A Review on Big Data- Storage Techniques and Its Challenges

Vidyashree H D, Kavyashree E D, Sowmya Shree P

Assistant Professor, Department of Computer Science, College of Academy for technical and Management Excellence, Karnataka , Mysuru, India

ABSTRACT

Big data as the name indicates it is extremely large data sets collected from various sources like internet, camera, applications, and bank transactions and so on. RDBMS is not sufficient to store and process such large quantity of data. This paper introduces several storage and processing techniques to deal with the big data. Apache Hadoop, Microsoft HDInsight, NoSQL (Not Only SQL), Hive, Sqoop, PolyBase, Big data in EXCEL, Presto are some of the techniques to store and process big data. Storing and processing is the one issue of big data on the other hand privacy and security is the another issues of big data. In this paper we introduced challenges of big data security. Apart from these issues and challenges we also have some advantages of big data.

Keywords: Big data, Hadoop, NoSQL, Hive, Sqoop, Salami attacks, Trust relationship attacks, Session hijacking attacks.

I. INTRODUCTION

The term big data means large amount of digital information like video, audio, log files, transactional data, web data, and so on. These data are generated by the computer system, sensors, mobile phones, etc. So, big data is defined by three terms, volume, variety and velocity. Volume describes the amount of data measured in terms of terabytes. Variety describes the type of data. Velocity describes the frequency of the data generation. This means how much data generated in one millisecond or second or minute or hour or day or week or month or year. There are several challenges relating to volume, complexity and security of big data. Almost 80% of the data created in the world are unstructured. Structuring these unstructured data is one of the big challenges. One more challenge is to store big data and one of the major challenges big data is facing is security issue.

The availability big data is very much important because big data is necessary for many confidentiality and privacy tasks. For example, the availability of multiple datasets, that can be easily combined and analyzed, makes very easy to infer sensitive information. Collecting data from multiple data sources and devices, such smart phones, smart power meters, is also a problem of data privacy. The big data lifecycle is shown in the below fig-1.

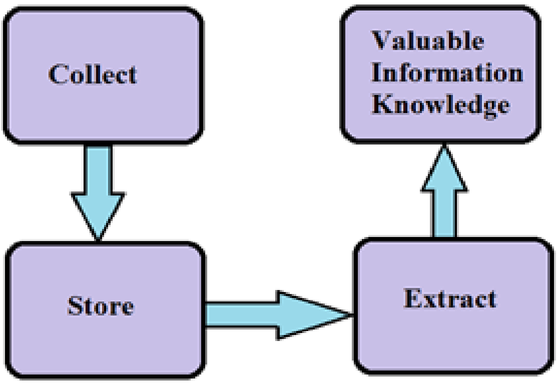

Figure 1. Life Cycle of Big data

II. HOW BIG DATA ARE STORED?

BIG DATA is a collection of large and complex data sets so that it is difficult to process using traditional applications/tools. Big data rise above Terabytes in size. Since big data created in the world are unstructured. It is very difficult to store and analyze these unstructured data. Some top tools are used to store and analyze big data are as follows. [10]

A. Apache Hadoop:

Apache Hadoop is an open source software platform for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. These clusters help us to process data across all nodes. Hadoop services provide for data storage, data processing, data access, data governance, security, and operations. This software runs in parallel on a cluster and HDFS- Hadoop Distributed File System. HDFS divides big data and assigns these data to nodes in a cluster. This reflects data in a cluster thus ensure high availability [11].

B. Microsoft HDInsight:

There are two types of HDInsight: Windows Azure HDInsight Service and Microsoft HDInsight Server for Windows. Both were developed in partnership with Hadoop software developer. Windows Azure HDInsight Service provides a software framework designed to manage, analyze and report on big data. It is simpler, scalable, and cost-efficient. Windows Azure HDInsight Service uses Azure Blob Storage as the default file system or we can store data in the Hadoop Distributed File System (HDFS). This also provides high availability with low cost [12].

C. NoSQL *(Not Only SQL):*

NoSQL Database, also known as "Not Only SQL" is an alternative to SQL database which does not require any kind of fixed table schemas unlike the SQL. Big Data NoSQL databases were developed by companies like Amazon, Google, LinkedIn and Facebook to overcome the drawbacks of relational DBMS. RDBMS cannot meet all requirements necessary for storing and processing unstructured data. These requirements grow exponentially, NoSQL is a dynamic and cloud friendly approach to dynamically process unstructured data with ease. This avoids join operations on the data as in RDBMS. Since relational database is a subset of NoSQL database, NoSQL can be referred as structured storage. NoSQL Database consists of various types of data storage model some of them are Graph, Key-Value pairs, Columnar and Document. There are many open-source NoSQL DBs available to analyse big Data.

D. Hive:

HDInsight supports many of the Hadoop query, transformation, and analysis tools, and we can install some additional tools and utilities on an HDInsight cluster if required. **Hive**, design a schema on the data to run query, and use a HiveQL for these queries. For example, we can use the **CREATE TABLE** statement to create a table and then execute **SELECT** command to extract the required data. **Hive** is developed by Facebook, Netflix and Financial Industry Regulatory Authority (FINRA). This is used for Data mining purpose and runs on top of Hadoop.

E. Sqoop:

Sqoop is a command-line interface application for transferring data between Hadoop and relational database servers which is shown in fig 2. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. This can be effectively used to transfer structured data to Hadoop or Hive [13].

F. PolyBase:

This works on top of SQL Server 2012 Parallel Data Warehouse (PDW) and is used to access data stored in PDW. PDW is a data warehousing appliance built for processing any volume of relational data and provides an integration with Hadoop allowing us to access non-relational data as well.

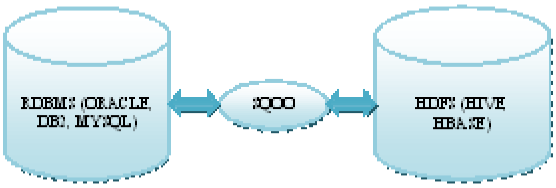

Figure 2. Sqoop for transferring data

G. Big data in EXCEL:

A tool from Microsoft is EXCEL which is comfortable for working. We can use EXCEL 2013 to connect data stored in Hadoop. Horton works is a big data software company. This developed Enterprise Apache Hadoop, provides an option to access big data stored in their Hadoop platform using EXCEL 2013. We can use Power View feature of EXCEL 2013 to summarize the data. Similarly, Microsoft's HDInsight allows us to connect to Big data stored in Azure cloud using a power query option.

H. Presto:

Presto is free distributed SQL query software for running queries against data sets of all sizes ranging from gigabytes to petabytes. This software is used by analysts who want response times ranging from second to minutes. Presto is used by Facebook for running queries over several internal data stores, including their 300PB data warehouse. Over 1,000 Facebook employees use Presto daily to run more than 30,000 queries that in total scan over a petabyte each per day.

III. CHALLENGES OF BIG DATA SECURITY:

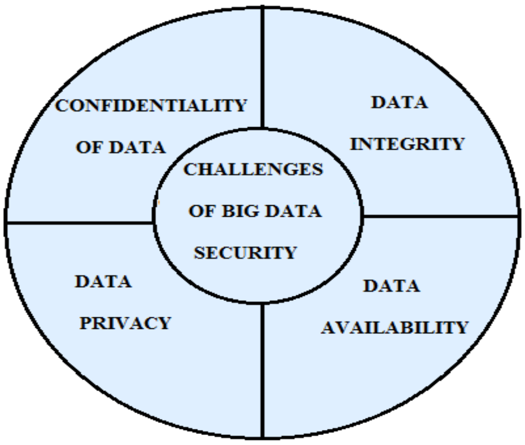

Figure 3. Challenges of big data

A. Confidentiality of Data:

Since large amount of data are collected every day, but service provider and owners of the big data has less memory to store big data, it is difficult to store, analyse and process such big datasets efficiently. Big data experts and tools are used to provide security to the big data. Tools mentioned in the previous topic are used to store, analyze and process the big data.

Let us take an example of data transaction, data transaction is nothing but data recorded while exchanging the information. Transactional data may include financial, logistical or work-related, involving everything from a purchase order to shipping status to employee hours worked to insurance costs and claims [1]. So these datasets consisting of some individual details and some sensitive data like debit card, credit card information. To ensure the privacy and confidentiality of the data there may be having several levels of analysis of data. Adding extra levels to the data analysis will take some preventive measures to protect any sensitive data from unauthorized access.

B. Integrity:

Protecting the data from unauthorized access this prevents data from altering is called Data integrity. Hardware errors, software errors, user errors, or intruder's error are the main reasons for data integrity problems [2]. There are several data integrity attacks some of them are as follows[3],

- ✓ Salami attacks- Series of minor attacks those together results in a larger attack.
- ✓ Trust relationship attacks- Exploits the trust relationship between a user and the web sites they visit.
- ✓ Man in the middle attack- Alters the communication between two parties who believe they are directly communicating with each other.
- ✓ Session hijacking attacks- Security attack on a user session over a protected network.

Maintenance of Integrity using following points:

a. Data provenance:

Documentation of all entities inputs and processes that affects the big data. Data provenance is a process of checking the data states from the beginning to the present state. Without this information we do not know from which source the data is coming from.[4]

b. Data trustworthiness:

Data trustworthiness means how much we can believe in data. If the data is not trustworthy then it is difficult to take decisions on that data and also difficult to analyze and process the data. The reason why data become untrustworthy is lack of protection from malicious users and error free data. Some of the techniques to ensure the data trustworthiness is data correlation and source correlation techniques [1].

c. Data loss and data de-duplication:

Two things mainly affect data integrity one is data leakage and the other data loss. To prevent data loss DLP (Data Loss Prevention) techniques are used. One technique to prevent data leakage in hadoop is Haddle framework which consisting of analytical and data collection methods. Using Haddle framework we can improve the performance of hadoop by collecting the data logs.[5-6]

Since big data is very huge it is very difficult to store such huge amount of data. To optimize the data storage we use data deduplication. Data deduplication is nothing but compressing the data, this will eliminate the redundant copies of data from big data storage. This helps in proper utilization of storage and the number bytes to be sent are reduced.

Classification of Data deduplication based on granularity, location and ownership are as follows:

- ✓ Granularity- File level and block level deduplication.
- ✓ Location- Client side and server side deduplication.
- ✓ Ownership- One user and cross user deduplication [7].

C. Availability:

Data must be available when authenticated user trying to access the data. To make data available all the time for the authorized user we use HA (High Availability) systems [2]. HA systems are designed using backup servers, communication links and replication. Now data availability issues are prevented by cloud computing. But we need to resolve some attacks that will violate the data availability. Some of the attacks are Although Denial of Service (DoS) attack, Distributed Denial of Service (DDoS) attack and SYN flood attack (Form of Denial of Service attack)[3].

D. Data privacy:

Data privacy is also called as information privacy. This ensures that the personal information of an individual never be shared with third party without informing the owner of that information. To ensure security of NoSQL data with privacy protection is elaborated in [8].

IV. ADVANTAGES OF BIG DATA

1. Big data identify threats and error quickly.
2. Increase in the loyalty of customer.
3. Help in making wise decisions in business.
4. Sentiment analysis can be performed by big data tools.
5. Helps in understanding market conditions.
6. Saves cost, tools like hadoop and cloud based analytics helps in identifying more efficient ways of doing business.
7. Hadoop identify new sources of data easily and make decisions based on learning's.
8. Helps in predicting specific situations and its insight.

V. CONCLUSION

Big data is very important in our daily life. It is necessary to properly store, manage and process these data. In this paper we try to give brief knowledge about big data storage and processing. Only storing and processing of big data is not sufficient. We need to maintain big data in such a way that it must be safe and secured. In this paper we mentioned what are the challenges that we are facing in big data security. Why we need to store, process and secure these big data? Why, because big data is very important to lead our day to day life. Almost every individual in this world is dependent on these data

VI. REFERENCES

[1]. Min Chen, Shiwen Mao, Yunhao Liu, Big Data: A Survey, Springer Science+Business media New York, 2014.

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

[3]. Types of Network Attacks against Confidentiality, Integrity and Availability, Omnisecu.com. Online].Available: http://www.omnisecu.com/ccna-security/types-of-network-attacks.php, 2017.

[4]. R. Alguliyev and Y. Imamverdiyev, "Big Data: Big Promises for Information Security," IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, 2014, pp. 1-4, 2014.

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

[6]. SA" NS Institute InfoSec Reading Room",Sans.org,2017. Online]. Available: https://www.sans.org/reading-room/whitepapers/dlp/data- loss-prevention-32883. Accessed: 24- Jan- 2017].

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

[8]. Heni and F. Gargouri, "A Methodological Approach for Big Data Security: Application for NoSQL Data Stores", Neural Information Processing, pp. 685-692, 2015.

[9]. https://www.newgenapps.com March 2018]

[10]. http://bigdata-madesimple.com/top-big-data-tools-used-to-store-and-analyse-data, March 2018]

[11]. https://hortonworks.com/apache/hadoopMarch 2018]

[12]. https://msdn.microsoft.com/en-us/library/dn749853.aspx March 2018]

[13]. https://sqoop.apache.org/docs/1.4.6/SqoopUserGuide.html March 2018]