Big Data Security

Dr. Prerna Mahajan*
Geetika Gaba**
Nagendra Singh Chauhan***

Abstract

In today's world, requirement of huge data is increasing rapidly with major security concern. Everyone wants their data to be secured with every aspect and dimension. If we talk about our personnel life everyone have data in huge quantity and of a huge variety (audio, video, text, chat, photos) and Professionally we have a large amount of data in field of media, healthcare, technology, private sector, science, sports which has to be stored so that security could be maintained with every aspect. Big data refers to management and analysis of huge amount of data that exceeds the capability and efficiency of traditional data with every dimension. Cloud computing can be viewed as a solution to store huge amount of data, but there are certain security concerns to deal with. Measures can be taken to provide incremental enhancements in securing the cloud that will ultimately provide us with a secure cloud.

To manage, identify and analyze complex data it is essential to store and share huge amount of data i.e. why big data is introduced to store large amount of data with great security. Big data was introduced to handle a large amount of data and also be processed massive quantity of data. Google has announced mapreduce framework to process large amount of data on hardware. Later on apache hadoop distributed file system is evolved as an efficient hardware component for cloud computing along with integrated part like map reduces. However hdfs and mapreduce were not quite efficient because they do not provide security to protect sensitive data. That is why hadoop was introduced to encourage security measures by using different technologies such as combining data mining technology. In this paper, we come up with some solution to provide security aspect in storing big data.

Keywords: FLUME, Hadoop, HBASE, HIVE, PIG, SQOOP, ZOOKEEPER

I. Introduction

Big data refers to the huge volume of data that cannot be stored and processed with in a time frame in traditional file system.

The next question comes in mind is how big this data needs to be in order to classify as a big data. There is a lot of misconception in referring a term big data. We usually refer a data to be big if its size is in gigabyte, terabyte, Petabyte or Exabyte or anything larger than this size. This does not define a big data completely. Even a small amount of file can be referred to as a big data depending upon the content is being used.

Let's just take an example to make it clear. If we attach a 100 MB file to an email, we cannot be able to do so. As a email does not support an attachment of this size. Therefore with respect to an email, this 100mb file can be referred to as a big data. Similarly if we want to process 1 TB of data in a given time frame, we cannot do this with a traditional system since the resource with it is not sufficient to accomplish this task.

As you are aware of various social sites such as Facebook, twitter, Google+, LinkedIn or YouTube contains data in huge amount. But as the users are growing on these social sites, the storing and processing the enormous data is becoming a challenging task. Storing this data is important for various firms to generate huge revenue which is not possible with a traditional file system. Here is what Hadoop comes in the existence.

II. Big Data

Big Data simply means that huge amount of structured, unstructured and semi-structured data that has the ability to be processed for

Dr. Prerna Mahajan*
Department of IT, Institute of Information Technology & Management, GGSIPU, New Delhi
Geetika Gaba**
Student - MCA, Institute of Information Technology & Management, GGSIPU, New Delhi
Nagendra Singh Chauhan***
Student - MCA, Institute of Information Technology & Management, GGSIPU, New Delhi

information. Now a days massive amount of data produced because of growth in technology, digitalization and by a variety of sources, including business application transactions, videos, picture , electronic mails, social media, and so on. So to process these data the big data concept is introduced.

Structured data: a data that does have a proper format associated to it known as structured data. For example the data stored in database files or data stored in excel sheets.

Semi-Structured Data: A data that does not have a proper format associated to it known as structured data. For example the data stored in mail files or in docx. files.

Unstructured data: a data that does not have any format associated to it known as structured data. For example an image files, audio files and video files.

Big data is categorized into 3 v's associated with it that are as follows:[1]

Volume: It is the amount of data to be generated i.e. in a huge quantity.

Velocity: It is the speed at which the data getting generated.

Variety: It refers to the different kind data which is generated.

A. Challenges Faced by Big Data

There are two main challenges faced by big data [2]

i. How to store and manage huge volume of data efficiently.

ii. How do we process and extract valuable information from huge volume data within a given time frame.

These main challenges lead to the development of hadoop framework.

Hadoop is an open source framework developed by duck cutting in 2006 and managed by the apache software foundation. Hadoop was named after yellow toy elephant.

Hadoop was designed to store and process data efficiently. Hadoop framework comprises of two main components that are:

i. HDFS: It stands for Hadoop distributed file system which takes care of storage of data within hadoop cluster.

ii. MAPREDUCE: it takes care of a processing of a data that is present in the HDFS.

Now let's just have a look on Hadoop cluster:

Here in this there are two nodes that are Master Node and slave node.

Master node is responsible for Name node and Job Tracker demon. Here node is technical term used to denote machine present in the cluster and demon is the technical term used to show the background processes running on a Linux machine.

The slave node on the other hand is responsible for running the data node and the task tracker demons.

The name node and data node are responsible for storing and managing the data and commonly referred to as storage node. Whereas the job tracker and task tracker is responsible for processing and computing a data and commonly known as Compute node.

Normally the name node and job tracker runs on a single machine whereas a data node and task tracker runs on different machines.

B. Features Of Hadoop:[3]

i. Cost effective system: It does not require any special hardware. It simply can be implemented in a common machine technically known as commodity hardware.

ii. Large cluster of nodes: A hadoop system can support a large number of nodes which provides a huge storage and processing system.

iii. Parallel processing: a hadoop cluster provide the accessibility to access and manage data parallel which saves a lot of time.

iv. Distributed data: it takes care of splinting and distributing of data across all nodes within a cluster .it also replicates the data over the entire cluster.

v. Automatic failover management: once and AFM is configured on a cluster, the admin needs not to worry about the failed machine. Hadoop replicates the configuration Here one copy of each data is

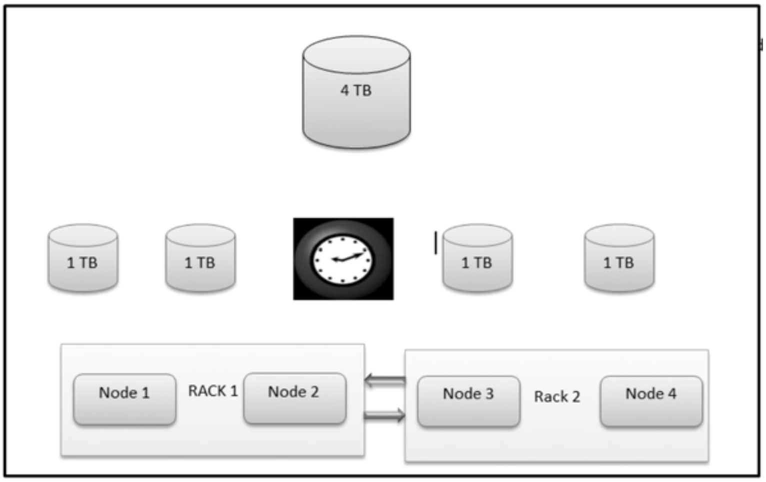

Fig. 1: Hadoop architecture

copied or replicated to the node in the same rack and the hadoop take care of the internetworking between two racks.

vi. **Data locality optimization:** This is the most powerful thing of hadoop which make it the most efficient feature. Here if a person requests for a huge data which relies in some other place, the machine will sends the code of that data and then other person compiles it and use it in particular as it saves a log to bandwidth

vii. **Heterogeneous cluster:** node or machine can be of different vendor and can be working on different flavor of operating systems.

viii. **Scalability:** in hadoop adding a machine or removing a machine does not effect on a cluster. Even the adding or removing the component of machine does not.

C. Hadoop Architecture

Hadoop comprises of two components

i. HDFS

ii. MAPREDUCE

Hadoop distributes big data in several chunks and store data in several nodes within a cluster which significantly reduces the time.

Hadoop replicates each part of data into each machine that are present within the cluster.

The no. of copies replicated depends on the replication factor. By default the replication factor is 3. Therefore in this case there are 3 copies to each data on 3 different machines.

III. Big Data Over Cloud [4]

The rise of cloud computing and data storage over cloud have facilitate in emergence of big data it is very easy to store big data and we can also say huge data over cloud and access more efficiently. Cloud computing have reduced computing time increased data storage capacity by using standardized technologies. Moreover we have many other advantages of big data storage over cloud it has major advantages over traditional file system as cloud platforms comes in different forms and sometimes have to be merged and integrated with traditional platforms

As cloud is an enabler for advanced analytics with big data with its cost effective delivery models it is providing analytics-as-a-service(AaaS) for cloud based big data analytics.

Bid data along with cloud computing is a compelling combination to handle huge amount of data with more efficiency and with more accuracy as data is becoming more valuable. Today conventions are shifting from what to store? To where to store with security and privacy

Big data refers to huge data sets that are orders of eminence (volume) more manifold structured, semi structured, and unstructured data that are and approaching faster (velocity) . This outpouring data is generated by different places. It is assorted and comes

Table 1: Hadoop Eco-system

1.	**PIG**	It is a scripting language used to write data analysis program for large dataset present within the hadoop cluster and known as PIG Latin
2.	**HBASE**	It is a column oriented database that allows reading and writing of data into the HDFS on a real time bases.
3.	**SQOOP**	Apache scoop is an application used to transfer data from Hadoop to any database management system.
4.	**HIVE**	Apache hive is sql like language which allows enquiring data from hdfs. The sqlvariant of hive is hiveql.
5.	**FLUME**	Apache Flume is an application that allows the streaming data into the hadoop cluster
6.	**ZOOKEEPER**	provide coordination between all the software to function properly
7.	**HDFS**	It is hadoop distributed file system, a major component of hadoop.
8.	**MAP REDUCE**	It is comprises of map() i.e. procedure that performs the task of sorting and filtering & reduce which performs a summary operation. It manages all data communication and transfer between different parts of system & controlling redundancy and fault among data.

in many formats, including text, document, image, video, and more. The actual significance of big data is in the insights. When big organizations store there sensitive data in data warehouse unaware of all these facts some data is already located over cloud. Depending on requirement of security and place where to store data IT industry is focusing on budget which leads to analytics-as-a-service which supports internal, external as well as hybrid cloud.

Analytics-as- a-service aims to solve big data security and privacy problem. IT are not just solving the problem of storing data and also providing an infrastructure to perform analytics-as-a-service. Many companies done have cloud to store big and sensitive data so it allows to store data on public cloud which helps to reduce cost of storage with security and maintenance also

DAaas(Data Analytics-as-a-service) represents an opportunity to extend platform that can provide cloud based analytical capability with huge variety and volume. It is a platform which provides end-to-end capabilities to handle data with a n innovative concept

IV. Big Data Security and Challenges [5]

*A. Insecure computation***:** It simply means that an untrusted computation program used by an attacker to submit to our big data solution to extract and turnout sensitive data from data sources.

Apart from information leak, it can also corrupt the data which result in unpredicted results from data.

It can also perform some denial of services on big data solution too.

B. Input validation and filtering: since data is collect data from variety of sources, so the biggest issue to validate the input which involves making a decision of what kind of data is trusted and what kind of data is untrusted. Data Filtering: It also needs to filter the malicious data from good data.

Since GBs or TBs of continuous data flow over the internet there is a big challenge to filter out the data Signature data can lead to reduce the problem here.

*C. Granular Access Controls***:** Designed for performance and scalability with no security in mind. Table, Rows and cell level access control went missing in big data.

By default access control is disabled. No sql is provided for that and a user has to depend on third party software to enable its accessibility.

D. Insecure Data Storage: data at various nodes, Authentication, Authorization & encryption is challenging. Encrypting of real time data can have

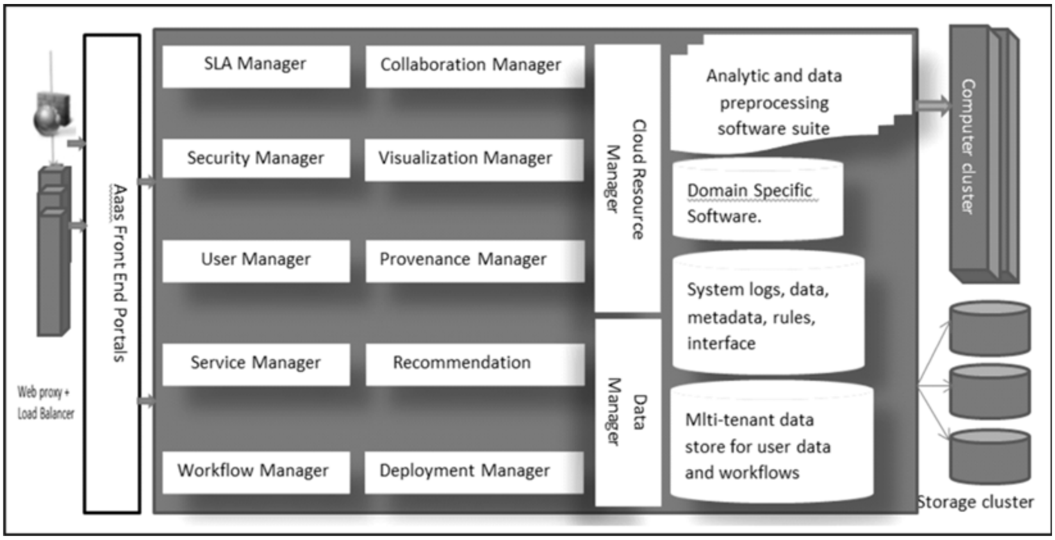

Fig. 2: Analytics as- a Service

performance impacts as well. You cannot encrypt all the data you share on the drive.

E. Privacy pertains in Data Mining and Analytics: Legitimatize of big data generally involves Data mining and Analytics.

Here sharing of data involves multiple challenges such as invasion of Privacy, Invasive Marketing and an unintentional disclosure of information.

Example: AOL releases of anonym zed search logs, User can easily be identified.

V. Solution to Achieve Big Data Security:[6]

At some extent we can achieve big data security with the following solutions.

1. Secure your computation code just by implement access control, code signing, dynamic analysis of computational code.
2. Strategy to prevent data in case of untrusted code.
3. Implement comprehensive input validation and filtering: consider all internal and external sources, Evaluate input validation and filtering of your big data solutions
4. Implement Granular access Control: Review Permission to execute ad-hoc queries and enable access control explicitly which is disabled by default.
5. Secure you data storage and computation: Sensitive data should be segregated, Enable data encryption for sensitive data which is again disabled by default, Provide API security and audit administrative access on data nodes.
6. Review and employ privacy maintaining Data Mining and analytics: here users make sure that the analytics data should not disclose sensitive information and get your Big data pen tested.
7. Security for big data can also be achieved through Hybrid cloud i.e. hybrid cloud means private and public cloud together large companies are providing space to store huge amount of data over hybrid cloud which reduces the cost of storing and maintenance with public cloud and a maintaining security with private cloud

VI. Conclusion

The goal of big data along with cloud computing and hadoop architecture is to achieve privacy and security for big data i.e. huge amount of data with variety, velocity, volume. Somehow privacy and security can be achieved with hybrid cloud computing, if big data is integrated with hybrid cloud architecture. We can acquire security if big data is encrypted and divided into small clusters then it can be managed and processed with privacy. Data on clusters can be secured by using digital signatures. Security of data can also be maintained through analytics-as-a-service (Aaas) .It also reduces the cost of storage and maintenance.

References

1. Roger Schell"Security –" A Big Question for Big Data"in 2013 IEEE International Conference on Big Data
2. A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.
3. Groenfeldt, Tom. "Big Data—Big Money Says It Is a Paradigm Buster." Forbes (January 6, 2012). forbes.com/ sites/ tomgroenfeldt/2012/01/06/big-data-big-money-says-it-is-aparadigm-buster
4. N, Gonzalez, Miers C, Redigolo F, Carvalho T, Simplicio M, de Sousa G.T, and Pourzandi M. "A Quantitative Analysis of Current Security Concerns and Solutions for Cloud Computing.". Athens: 2011., pp 231 – 238, Nov. 29 2011- Dec. 1 2011
5. "Cloud Security Alliance Top Ten Big Data Security And Privacy Challenges "by CSA Big Data Working Group
6. Katina Michael, Keith W. Miller "Big Data: New Opportunities and New Challenges," Published by the IEEE Computer Society 2013