Big Data: Tools & Applications

Ankush Arunrao Parise[1], Rupali Ganesh Rajurkar[2]

[1,2]*Student, Dept. of CSE, Prof. Ram Meghe College of Engineering & management Badnera, Maharashtra, India.*

---***---

Abstract - *Every day lots of data are generated, this is very difficult task to stored and manage this data in a proper manner. This huge of data is called as a "Big Data". The size of data is increased day-by-day and all traditional technology are not able to handle this large amount of data. To deals this problem, new technology is introduced called as a "Big Data Tools" technology its able to handle large amount of data as well as stored in a proper manner. Also provide a mechanism for data analyzing and processing. This is an open source technology provided by Apache community. The Hadoop software library framework is providing facility for distributing processing of huge datasets a cross clusters of computers using simple programming. It is intended to scale up from single servers to a huge number of machines, each offering to local computation and storage. The library handle itself detect and handle failure at the applications layer. It's not compulsion, a data is required in a structured manner. Its support unstructured datasets also such as a NOSQL database. There are One Hundred and Fifty plus tools are available for data analyzing. Every tools have different methods for implementing as well as data analytic. The applications of big data are used anywhere. Because this technology helps to interpret large amount of data in very faster manner also improve the production efficiency and driven new data service for any business.*

Keywords - Big Data, Tools, Application.

1. INTRODUCTION

In digital world, everyone is connected to the internet for using a digital service. Data are generated from a different source. They are help to increase the growth of Big Data. In simple Big Data is a collection of hug and complex dataset which are very difficult as well as time consuming to process using traditional dataset management tools or data processing application. Data size is double in every two years. "Big data tool" is a helpful mechanism to arrange this big data in systematic manner for data analytic purpose. In "Big Data Tool" consist various tools are available for data storing as well as analytic. There are various filed in with data is much more important such as government sector, agriculture, call center, education, technical, insurance, healthcare etc. And they also application of big data.

2. Literature Review

In October 2003 Doug Cutting and Mike Cafarella published paper on Google File System. This paper is main genesis of Hadoop. After research on that paper with the help of Google new technology are invented that's called as a "MapReduce: Simplified Data Processing on Large Clusters". Hadoop framework is totally based on Java technology. Actual Development is start using Apache Nutch, project. But move to subproject of Hadoop start in January 2006.In that time Doug Cutting working at Yahoo! and rename that technology is "Hadoop" on his son's toy. All framework of Hadoop is developed in a Nutch. In that, 5000 plus lines of code for HDFS and 6000 lines of code for MapReduce.

Despite the fact that the idea of enormous information itself is generally new, the inceptions of big data collections return to the 1960s and '70s when the big data was simply beginning, with the principal server farms and the advancement of the relational database [3].

Around 2005 peoples are realize the large amount of data are generated though Facebook, YouTube, and others technology based on internet. A Hadoop is developed that the same year. A Hadoop is an open source framework used for storing and analyzing huge amount of data. NOSQL database are introduced as same year. That time this dataset is very popular. Because in this database no need to data are stored in a fix manner.

The development of open source framework is like Hadoop and in recently spark very essential for the growth of big data problem. Because they tool provide facility easier to work with cheaper to store. Todays is the volume of big data is a very high. But all users are still generating a huge of datasets.

The advent of IOT (Internet of Things), more objects and electronic gadgets are connected to internet, collect data on customers uses way and product performance. The advent of machine learning has produced more data. Because of cloud computing is possible to expand big data possibilities.

Over the past 25 years' data are generated in a huge scale. Every day introduced a new technology because of that scale of big data are increased. And a big data very difficult to handle. And According to International Data Corporation report (IDC) report in 2011in the world volume of data is 1. 8ZB.This data increased in nine times on every five years. As compared to the traditional database in big data consist the mass unstructured data that need to more time for analytic. Recently the industry interested in the high potential of big data and many government sectors are allowed to research on big data applications.

Nowadays, service of internet companies related to big data grow rapidly. For example, Google processed 20 Peta Bytes of data every day. Facebook also proceed a large amount of data.

In April 2006 version of Hadoop 0.1.0 was released.

Table 1: Latest Versions of Hadoop

Released Date	Versions
06 February 2019	3.1.2
16 January 2019	3.2.0
19 November 2018	2.9.2
15 September 2018	2.8.5
31 May 2018	2.7.7

In Table 2: shows latest versions of Hadoop. Currently, Apache Hadoop 3.1.2 are a new version. And development on Hadoop frame work is proceed now to make it very advance [1][2].

3. Types of Big Data

The collection of large amount of datasets is known as "Big data". Big data are categories in to three different types are as follow.

3.1 Structured

Any type of data can be stored, processing, accessing in a fix format it's called as a Structured data. The best example of structured database is a DMBS (Database Management System).

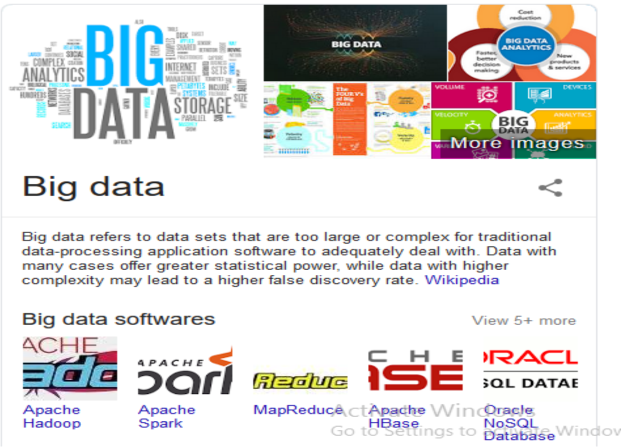
Table 2: Structured Data Format

Roll_No	Name	Sex	City
01	Ankush	M	Daryapur
02	Aniket	M	Wardha
04	Amit	M	Pune
05	Rupali	F	Akola
06	Harshal	F	Nagpur

In Table 1 shows the structure data. This type of data is stored in a table format. In this data are stored and processing in a fix format.

3.2 Unstructured:

Any type of data it doesn't have recognizable form it's called as an Unstructured data. This type of data is very complex for processing. The example of an unstructured data is heterogeneous data source containing combination of simple text, image file, video, document file etc. Now day association have wealth of data available but unfortunately they don't know how to use it?



Big data

Big data refers to data sets that are too large or complex for traditional data-processing application software to adequately deal with. Data with many cases offer greater statistical power, while data with higher complexity may lead to a higher false discovery rate. Wikipedia

Big data softwares

View 5+ more

Fig 1: Untrusted Data Format

In Fig.1 show the unstructured data in combination two different format such as plain text and images. That's type a data is a challenge to processing and analytic.

3.3 Semi-Structured

The combination of structured and un-structured data is called as Semi-Structured data. The semi-structured data are form of the structured data but not actually defined it. Like a table definition of a relational database. Like xml file.

```
<contact-info>
        <contact1>
                <name> Ankush Parise </name>
                <mobile> 99521510408 </mobile>
          </contact1>
          <contact2>
                 <name> Rupali Rajurkar</name>
                <mobile> 78758855885</mobile>
         </contact2>
</contact-info>
```

Above type of data is a best example of a Semi-Structured. The information of person and contacts are stored in that xml file in semi-structured format [3].

4. Characteristics of Big Data

- **Volume:** We already known as the Big Data is a large amount Volume of data that is generated a different data sources like government sector, agriculture, social media, technology, call center etc. In simple word the quantity of data.
- **Variety:** This term is belonging to the format of data. Its structured, unstructured and semi-structured data generated from different sources. Understand the variety of a data is a very important for data analytic purpose.
- **Velocity:** This term is belonging to the speed of data generation in a real time or how data generated fast and proceed. And also determine the real potential in the data. It's all about come in velocity.
- **Variability:** This term is belonging to inconsistency in a data. This can be disturbing the process of being able to manage and data in effective manner. It also refers to the inconsistent speed at which big data is loaded into your database.
- **Veracity:** Veracity is all about the accuracy of a data also check the noise and abnormality in the data. And perform some operations on that noisy data.
- **Visualization:** In today's world visualization is very important. Using graphs and chart to visualize large amount of complex data is more effective to spreadsheets
- **Value:** Value is a last diversion, but is most important of all. The all of other characteristics of Big Data are meaning less if we don't drive business value from data.
- **Complexity:** Data management is a very difficult task specially when data are coming in different sources [4].

5. Big Data Tools

We know that the data in a large unit is called as a "Big Data". And it very difficult task using traditional database to data storing in a proper manner and analyze it. This data size is a very Huge. Because different types of data that is encompasses, big data always come with more challenges about its volume and complexity. According to latest survey 80% of data are created in the world are unstructured. The biggest challenge is how this unstructured data can be in a structured data before we attempt to understand and capture most important data from different sources. And another challenge it how to stored it in a proper form. Here is top most tools are available for storing and analyzing a big data. There are two categories of big data tools such as storage and analysis. Using this tools, challenges about big data is become easiest. Because big data tools are providing a proper mechanism to deals with big data related challenges [5].

5.1 Apache Hadoop

Apache Hadoop is a free software framework based on a java technology it can very efficient to stored huge amount of data on data cluster. This tools provide facility to framework run parallel on a data cluster and has process to ability data across all nodes. The HDFS (Hadoop Distributed File System) is a storage system of Hadoop. In with stored big data and distributed across number of nodes and clusters. This tool also provides high availability of data.

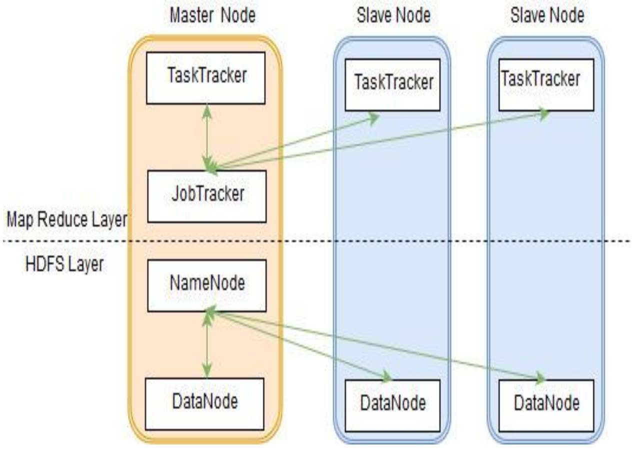


Fig 2: High Level Architecture of Hadoop

Fig. 2 shows Hadoop fallows a master slave architecture for data storage HDFS for data storing and MapReduce for distributed data processing. According to Architecture of Hadoop the master node for data storage is Hadoop HDFS is the NameNode and the master node for parallel processing of data using Hadoop MapReduce is the Job Tracker. The slave nodes in the Hadoop architecture are the other machines in the Hadoop cluster which store data and perform complex calculations. Every slave node has a Task Tracker and a DataNode that synchronizes the processes with the Job Tracker and NameNode respectively. In Hadoop architectural design implementation, the master or slave systems can be setup in the cloud [5] [6].

Features:

- Improve authentication when using HTTP proxy server.
- Data processing speed is high.
- Provide Data flexibility when data is proceeding.
- Robust ecosystem.
- Support POSIX-type file system for extended attributes.
- Suitable to meets analytic needs for developers.

5.2 Hive

Hive is an open-source software. This tool is used to distributed data management for Hadoop framework. Also support like a SQL language for accessing a big data. This tool mainly used to data mining. Its run on top of Hadoop.it also allow to programmer to analytic huge of datasets.

Features:

- Support query language like SQL for interaction with datasets.
- Compile language based on two important task such as map and reducer.
- Support java and python technology.
- Managed and querying only structured data.
- It allows to JDBC (Java Database Connectivity) for interaction.

5.3 Spark

Apache Spark is also open source tools of big data used for data analytic. It allows to over 80 high-level operators that's help to make apps parallel. In simple Apache spark is a unified analytic engine for huge data processing. Using this tool, we can archive high performance batch as well as streaming data.

Features:

- Provide fast processing.
- Help to run application on Hadoop cluster 100 times faster on memory and 10 times faster on a disk.
- It also provides lighting fast processing.
- Sophisticated analytics.
- Ability to integrate Hadoop data and existing Hadoop data on cluster.
- It can allow to built-in-API in Java, Python.

5.4 Sqoop

Sqoop tool is used to provide interface between relational database and Hadoop file system. It is also used to import as well as export data from a various relational database.

Features:

- Parallel data import as well as export.
- Used connector between all relational database.
- It also provides facility import result from SQL query.
- It allows the load full table using single Sqoop command.
- Support compression.

5.5 Microsoft HDInsight

A Microsoft HDInsight is a solution of big data Microsoft powered by Apache Hadoop. In with provide facility to Spark and Hadoop services in the cloud. Provide big data cloud in two categories such as stander and premium. It is also used to windows Azure Blog storage as the default file system. The biggest advantage of spark tool is high data availability in low cost [7].

Features:

- It can provide reliable analytic with industry-leading SLA.
- Also provide enterprise-grade security and monitoring.
- High productivity platform for developers.
- Provide cloud services in cheaper cost.

6.6 Apache Storm

Storm is also free and open source big data tool. It is a distributed real time framework for reliable processing the unbounded data stream. This frame works any programming language. Also capable real time processing. It can similar to Map Reduce job.

Features:

- It supports multiple language.
- Also provide massive scalability.
- It uses the parallel computing that runs across the cluster of machine.
- Easy tool for big data analytic.
- It works on "fail fast, auto restart".
- Run on Java Virtual Machine.

6.7 Apache Cassandra

Apache Cassandra database is used to provide a proper management of a big data. It is a main and best big data tool processing for unstructured datasets. Also provide a high available service with no signal point of failure. Supports contests and services are available from third parties.

Features:

- Simple operations.
- It can allow to across the data center easy distribution of data.
- High scalability.
- Provide continuous availability of a data sources.
- It can allow to data automatically to replicated to multiple nodes for fault- tolerance

6.8 MongoDB

MongoDB it is an open source NoSQL database. It can provide compatibility with a cross platform. It deals with a business real-time data for instant decision. It can run on MEAN software stack, NET applications and, Java platform [7].

Features:

- It can provide flexibility on cloud base services.
- It can allow to any type of data for storing or support multiple datatypes.
- It referred dynamic schema.
- Cost saving.

6.9 Neo4j

Neo4j is a very important tool related to business analysis. The business goal is that our products are simple and fit in use case, whatever it may be. When we want to analyze graphs for transaction, market analysis, oration optimization, or anything else it will provide analytic result in graphics form. Also Neo4j is a that is widely used to graph database in a big company's. It can support fundamental structure of graph database [7].

Features:

- It can allow ACID transaction.
- Provide high availability of data.
- Also scalable and reliable.
- Flexible, doesn't need of schema for data.
- Also integrated with another's databases.

6.10 R Programming Tool

This is also open source big data widely used in a big data industry for statistical analysis. R language it has own public library CRAN (Comprehensive R Archive Network) with consist 9000 plus modules and algorithm for statistical analysis.

R language can run on Windows server, Linux servers as well as SQL server. It also supports Hadoop and Spark. It allows portability. Hence developed and test R module on local data source can be easily implement. Also provide a wide variety of statistical tests.

Features:

- Data handling and storing facility.
- It can allow operators for computation on arrays.
- Also provides coherent, integrated collection of big data tools for data analysis
- Provide graphical facilities for data analysis.

7. Big Data Applications

Big Data Technology is a very popular technology for data handling as well as storing found in many applications in a various field.

Government
Technology
Agriculture
Big Data Application
Health
Marketing
Banking

Fig 3: Big Data Application

Fig. 3 shows the application of a big data. In every sector who's deals with large amount of data they are used "Big Data Tools" for Handling Big Data [8].

3.1 Government

A big data analysis is very useful in a Government sector. In government sector number of challenges are the integration and interoperability of big data across different government departments and various organization. The Indian Government used numerous techniques to ascertain how the Indian electorate is responding to government action, as well as new ideas for policy agreement. Big data is analyzed from various government agencies and is used to protect the country.

7.2 Education

From a technical point of view, a biggest challenge in the education industry is to incorporate big data from various sources and vendors and to used it on platforms that were not designed for the varying data. And also used to challenges to integrate data from different sources, on different platforms and from different vendors that were not designed to work with one another.

From a practical point of view, staff and institutions need to learn the new data management and analysis tools.

7.3 Banking

Big data tools are much more important in a banking sector. Almost people are used a bank and perform number of transaction. Such as credit and debit. Because of transition a lots of data are generated and it’s very important to maintain those data. That’s why big data tools are very important. According to research 62% of bankers are cautious in their use of big data due to privacy issues. The biggest challenges in a banking industry security as well as privacy issue. By uncovering hidden connections between seemingly unrelated pieces of data, big data analytics cloud potentially reveal sensitive personal data.

7.4 Agriculture

A biotechnology firm also uses sensor data to optimize crop efficiency. It plants test crops and runs simulations to measure how plants react to different changes in condition. Its data environment constantly adjusts to changes in the attributes of various information it collects, consisting temperature of environment, water levels, soil composition, growth, result, whether information and gene sequencing of each plant in the test bed. These simulations provide facility to discover the optimal environmental conditions for specific gene types.

7.5 Marketing

Marketers have start to use face recognition software to learn how well their advertising succeeds or fails at stimulating interest in their products. A recent survey in the Harvard Business Review looked at what types of advertisements compelled viewers to continue watching and what turned viewers off. Among their tools used are recognize or analyze facial expression to reveal what viewers are feeling. The research was designed to discover what kinds of promotions induced watchers to share the advertise with the help social media, helping marketers create advertises most likely to “go viral” and improve sales.

7.6 HealthCare

Now healthcare is yet another industry which is bound to generate a large amount of big data. Health care staff now have access to promising new threads of knowledge. This information is a form of “big data,” so called not only for its sheer volume but for its complexity and diversity as well as timelines. Pharmaceutical industry exports, payers, and providers are now starts to analyze big data to obtain insights. Recent technologic advances in the industry have improved their efficiency to work with such data, even though the files are enormous and often have various database structures and technical characteristics. So maintain all types of information related to different department big data tools is much more important in healthcare sector [9] [10].

8. Conclusion

In digital world every day lots of data are generated. It’s very complex job to handle this big data and stored in efficient manner. We have phase daily lots of challenges related to large amount of data. Such as data analytic for business analysis. To deals with problem related large amount of data, “Big data tool” is a very powerful mechanism.

Big data used in various sector Such as banking, agriculture, chemistry, data mining, cloud computing, finance, marketing, stocks, healthcare, education, technology, call center etc. Above all application of big data.

9. References

[1]https://en.wikipedia.org/wiki/Apache_Hadoop#History

[2] https://www.oracle.com/in/bigdata/guide/what-is-big-data.html

[3] https://www.guru99.com/what-is-big-data.html

[4]https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples/

[5]https://www.dezyre.com/article/hadoop-architecture-explained-what-it-is-and-why-it-matters/317 Hadoop

[6] https://www.whizlabs.com/blog/big-data-tools/

[7] https://neo4j.com/developer/get-started/

[8] https://www.simplilearn.com/big-data-applications-in-industries-article

[9] https://bigdata-madesimple.com/top-big-data-tools-used-to-store-and-analyse-data/

[10]https://www.digitalvidya.com/blog/big-data-applications