Asian Journal of Convergence in Technology *Volume II Issue III*
Issn No.:2350-1146, I.F-2.71

Overview of BIG DATA and HADOOP

Prof. Ravindra V. Kerkar[1], Prof. Waman R. Parulekar[2], Prof. Harshada U. Salvi[3]

Department of Master of Computer Application[1,2,3]
Finolex Academy of Management and Technology Ratnagiri[1,2,3]
University of Mumbai
{ravindra.kerkar09[1], vamanparulekar [2]& salviharsh289[3]}@gmail.com

Abstract **- Big data is huge amount of data; it is a collection of large datasets which cannot be processed using traditional data computing tools and application. It's not a tool, application or framework, it is merely large datasets. Now a day's various websites, web application are generating large and complex data. Processing this generated data using traditional data processing applications and tools is inadequate. Aim of this paper is to take review of Big Data and what are the challenges with managing large volume of data. This paper also gives the overview about the hadoop framework and how we can handle this large volume of data sets using hadoop framework.**

Keywords - BIG DATA, Hadoop, Map Reduce, HDFS

I. INTRODUCTION

Data is beneficial for both businesses such as social networking sites like Facebook, twitter, Pinterest, online shopping sites like Amazon, Snapdeal, Flipkart etc. and for individuals to select appropriate policy. Prior the 80's e-Commerce did not exist and information section, stockpiling and handling were successive procedures. Also data was processed using monolithic computers like mainframes. Jobs were done using batch processing operating system computers. Data processing was used in non-critical areas like payroll and accounting systems. Data processing could be feasible for only large enterprises and institutions. Data processing could only support post-event analysis and long term planning.

In 1980's and before data creation was a controlled procedure. Additionally rate of data creation was known and reasonable. The procedure of information creation and preparing was co-found. Before the 1980's information was organized i.e. everything about the data must be known "apriori" to have the capacity to store it and procedure it. This was possible because data creation was under control.

In the 1990's, data creation was still a controlled step and data was structured. The volume of information produced was reasonable and information preparing was still brought together. Social Databases are utilized for information handling.

After 90's Internet era was started and everything changed. Diagram 1.1 shows the global internet traffic 1993-2013 in Petabytes per month [1]. Fig.1 depicts that after 2003 internet traffic increases rapidly. The measure of information delivered by us from the earliest starting point of time till 2003 was 5 billion gigabytes [2]. If you pile up the data in the form of disks it may fill an entire football field. The same amount was created in every two days in 2011 and in every ten minutes in 2013[2]. This rate is still growing enormously. Though all this information formed is significant and can be useful when processed.

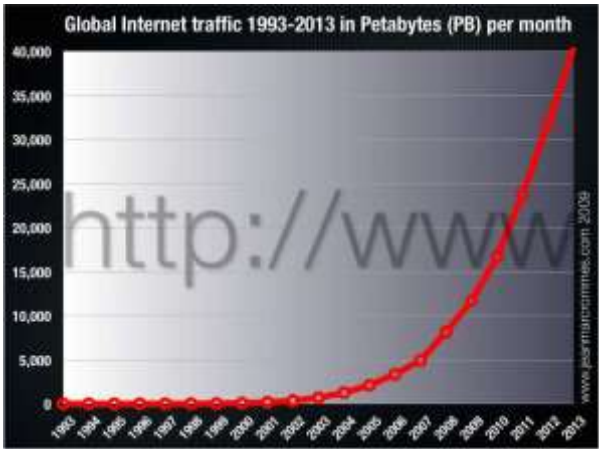

Fig.1 Internet Traffic Trends [1]

Early years of the Internet, internet used for e-commerce i.e. B2B Transactions and B2C Transactions. B2C Transactions includes the several sectors like Banking, Retail, Finance, Travel and hospitality and healthcare.

In the new millennium, the internet is adopted rapidly in e-commerce especially B2C. Customers easily find out best deal with the help of internet and businesses had to proactively attract customers to increase their product sales. Due to that data processing plays the business critical role instead of supportive role.

Then the new era of social networking and mobility came and which changes business because social networking has turned into a wellspring of important data to know the customer, their choices, preferences, and behavior. Due to the arrival of new technologies, high-end devices, and faster communication means like social networking sites such Facebook, twitter, the amount of data produced by mankind is growing rapidly with every year.

Now day's success of B2C business transaction relies on the ability to analyze customers' current and past behavior

real-time. Social Networking handles the unstructured data having extremely large data generation rates and it's highly distributed.

"Every two days now we create as much information as we did from the dawn of civilization up until 2003." - Erik Schmidt, former CEO-Google. Unstructured Data having very large data sets which are dynamic and increases rapidly, highly dispersed and distributed data generation, hard to move data to a single and / or central location and critical to process data and generate results real time.

II. WHAT IS BIG DATA

A. Introduction of Big Data

Big data is huge amount of data; it is a collection of large datasets which cannot be processed using traditional data computing tools and application. It's not a tool, application or framework, it is merely large datasets. Big data includes the data generated by different devices and applications. Given beneath are some of the fields that go under the umbrella of Big Data.

1. Social Media Data: Social media is the most popular media over the internet and people using mobile phones. In last several years' users of social media sites increased. Social media website like Facebook and twitter contain huge data. This social media data is big example of big data.
2. Black Box Data: Black box data is a device of airplanes. This device captures voices of the flight crew, voice recordings of microphones and the routine information of the aircraft.
3. Search Engine Data: Search engines like Google, Yahoo, AltaVista contains and process huge data. These engines deal with Big Data.
4. Stock Exchange Data: Shares data of different companies and users decisions like buy or sell.
5. Hospital Data: In modern era hospitals are using pervasive systems. These systems deals with patient monitoring, patient record keeping, medical history record keeping and many more tasks are there. This data is also good example of big data.
6. Web based Businesses Data: Includes large data of products which are available to sell online, transaction records, customer data.

In this way, Big Data incorporates tremendous volume, high speed, and an extensible assortment of information. The information in it will be of three sorts.

1. Structured data: Relational data which can be stored using DBMS and RDBMS. Here, first schema is created then data is stored.
2. Semi Structured data: XML data, It is not fully structured as compared to structured data.
3. Unstructured data: Word, PDF, Text, Media Logs etc. Here schema is not required usually not known, data is store before knowing schema e.g. social networking sites like Facebook, applications like whatsapp.

B. Benefits of Big Data

Big data are really essential to our life and it is an important part of modern world technologies. Following are some well known benefits of Big Data:

1. Using the information kept in social network sites like Facebook, Google-Plus, various marketing agencies are learning about the responses people for their campaigns, promotions and other advertising mediums.
2. Using the information in the social media like preferences and product awareness of their consumers, product companies and retail organizations are planning their production.
3. Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

C. Big Data Technologies

Enormous information advances are essential in giving more exact examination, which might prompt more solid choice making bringing about more noteworthy operational efficiencies, cost decreases, and diminished dangers of the business. To saddle the force of enormous information, you would require a framework that can oversee and prepare colossal volumes of organized and unstructured information progressively and can ensure information protection and security. There are different advancements in the business sector from various merchants including Amazon, IBM and Microsoft and so on. To handle enormous information. While investigating the advances that handle Big Data, we inspect the accompanying two classes of technology

1. Operational Big Data

   Operational Big Data includes systems like NoSQL Big Data system and MongoDB system. MongoDB provide operational capabilities for real-time, interactive operational Big Data workloads where data is mainly captured and stored. NoSQL Big Data systems are intended to take benefit of new cloud computing architectures. This makes operational big data workloads much easier to manage, cheaper, and faster to implement. Notwithstanding client cooperation's with information, most operational frameworks need to give some level of ongoing insight about the dynamic information in the framework. Some NoSQL frameworks can give

experiences into examples and patterns in view of continuous information with negligible coding and without the requirement for information researchers and extra foundation.

2. Analytical Big Data
   This includes systems MPP (Massively Parallel Processing) database systems and MapReduce that offer analytical capabilities for retrospective and composite analysis that may touch most or all of the data. MapReduce gives another strategy for dissecting information that is correlative to the abilities gave by SQL, and a framework in view of MapReduce that can be scaled up from single servers to a huge number of high and low-end machines. These two classes of innovation are corresponding and often conveyed together.

D. Challenges of BIG DATA

There are several challenges in managing big data. The key challenges associated with big data are as follows:

- Hard to capture the data due to its huge size
- Difficult to data curation
- Faster data generation rate
- Unstructured data
- Storing of data for its future use
- Searching specific information
- Sharing the data with other media
- Transferring the data from one location to another
- Difficult to analyze the data
- Present the data in meaningful format
- These are the key challenges associated with big data.

To fulfill the above challenges, organizations normally take the help of enterprise servers.

E. Solution to Big Data

In traditional approach, an enterprise will have a computer to store and process the big data. Here this data is stored using an RDBMS like MS SQL Server, Oracle, DB2 and sophisticated software's are used to interact with the database, process the essential data and provide it to the end users for analyze this.

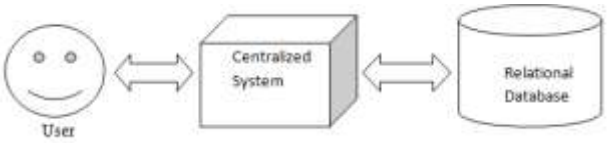

Fig. 2 Traditional data processing

This approach work well where we have small amount of data that can be stored by standard database servers or up to the limit of processor which is processing the data. But when it comes to dealing with large volume of data, it is tedious task to process such huge data through these traditional database servers.

Google resolve this problem by introducing an algorithm known as MapReduce. This algorithm actually divides the task into multiple small tasks and assigns those tasks to May computers connected over the network and finally collect the results to form the resultant dataset. Commodity Hardware means devices with general configuration such as 4 GB RAM, 500 GB hard disk etc.

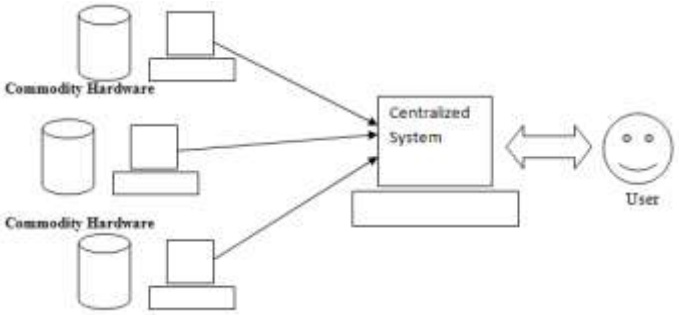

Fig. 3 Solution to Big Data

Fig 3 shows various commodity hardware's which could be single CPU machines or servers with higher capacity. The major task to store and process the unstructured data when it's structure is not known. First treat this data as "BLOB" (Binary large object). Attach some means of uniquely identifying each blob and then storing these blobs. Here size should not matter. For any kind of unstructured data, identify the "key" and "Value", here value is data associated with this key. Now each data item includes <key, value> pair. Key is used to identify the individual data items.

III HADOOP FRAMEWORK

A. Overview of Hadoop

Hadoop is open source framework written in java and is used to manage Big Data. This structure permits conveyed handling of extensive datasets crosswise over bunches of PCs utilizing basic programming models. A Hadoop outline worked application works in a domain that gives disseminated capacity and calculation crosswise over groups of PCs. Hadoop is intended to scale up from single server to a large number of machines, every offering nearby calculation and capacity.

Hadoop was created by Doug Cutting and Mike Cafarella [2] in 2005. Cutting, who was working at Yahoo! at the time

[3], named it after his son's toy elephant [3]. It was originally developed to support distribution for the Nutch search engine project [4].

Distributed computing is a wide and varied field, but the key distinctions of Hadoop are that it is

- Accessible—Hadoop runs on large clusters of service machines. It also runs on cloud computing services like Amazon's EC2.
- Robust—Since it is proposed to keep running on commodity hardware, Hadoop is architected with the presumption of successive equipment glitches. It can smoothly handle most such disappointments.
- Scalable—By adding more nodes hadoop scales linearly to handle big data in cluster
- Simple—Hadoop allows users to quickly write efficient parallel code.

Hadoop is an Open Source software framework which implements the Map Reduce technique; it is capable of parallel execution on 1000's of nodes. It uses Commodity hardware for its operation and provides fault tolerant job completion.

B. Hadoop Architecture

Hadoop frameworks contains several module which work together to manage large data set. Hadoop framework includes following four modules:

- Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide file system and OS level abstractions and contains the necessary. Java files are required to start Hadoop.
- Hadoop YARN: YARN is a framework for job scheduling and cluster resource management in Hadoop.
- Hadoop Distributed File System (HDFS): A distributed file system that provides high throughput access to application data.
- Hadoop MapReduce: YARN-based system for parallel processing of large data sets.

C. MapReduce

MapReduce is a problem solving technique which is based on the 'divide & conquer' pattern. This is amenable to parallel processing and distributed processing. It is also used to solve very large scale problems which involving large data sets. To process the big data using hadoop framework following have to be specified

- Input data location
- Output result location
- 'Map' function
- 'Reduce' function
- Other job parameters

All this is termed 'Job Configuration' in hadoop framework. Here two major processes at work:

1. Map: Map is used to process input data to generate a <key, value> pair. All mappers involved generate such pairs.
2. Reduce: Reduce is used to combines values with the same key and to generate an aggregated or consolidated output.

Following diagram shows the working of Hadoop - MapReduce

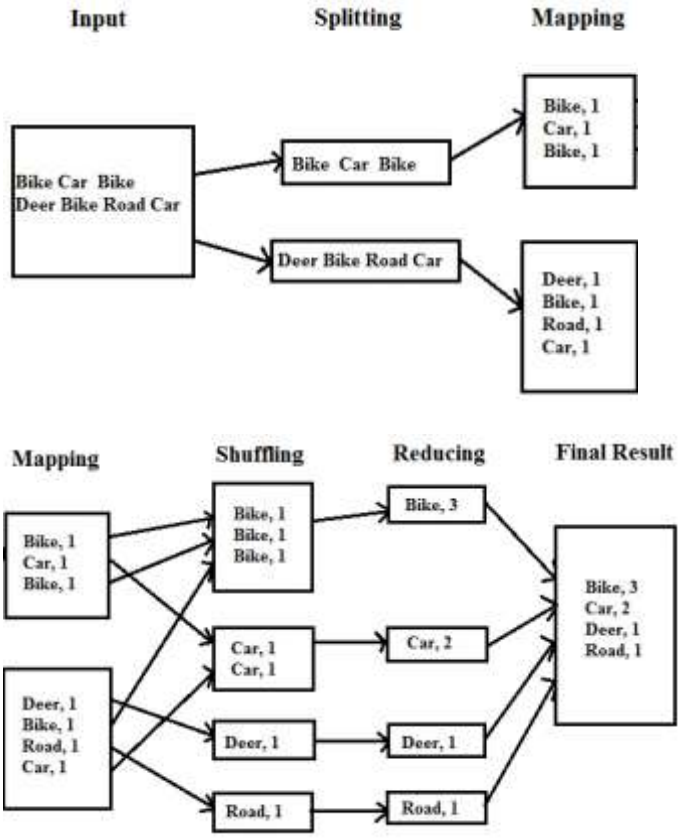

Fig. 4 Hadoop- MapReduce example

Consider the example which requires finding of number of occurrence of each word in specific file.

- In first step input data is split into two separate parts and assigns them to two separate mappers.
- Mapper then finds the occurrence of each word and generates output which includes all words names with count 1.
- Then shuffler takes this output and sorts this and generates separate part for each word.
- Then reducer reduces duplicate entries for each word and generates output which includes unique words with their occurrence count.
- Finally combines the output of all reducers single output is generated which includes the unique words with their occurrence count.

D. Hadoop Distributed File System (HDFS):

HDFS can store Petabytes of data which is highly scalable. It relies on commodity x86 servers and Open Source software. HDFS supports computation in each server. It treats failures as 'inevitable' and handles them like 'noise'.

Assumptions and Goals of HDFS

•Hardware failure is the norm rather than exception

•Designed more for batch processing rather than interactive use

•Tuned to support very large files

•Assumption: Apps need 'write once read many times' access model for files

•Moving computation to data is easier than moving data

•Easily portable from one platform to another

HDFS is configurable and comes with default configurations well suited for many applications. Following are important components of HDFS

- NameNode : which runs on master
- DataNode : which runs on master and slaves
- NameNode and DataNode are used for storage management.

Below diagram shows the components of Hadoop Distributed File System:

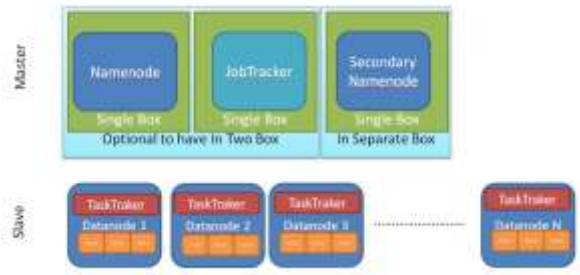

Fig. 5 Components of Hadoop Distributed File System
Source: blogs.msdn.com

- NameNode: The NameNode is the expert of HDFS that guides the slave DataNode daemons to perform the low-level I/O tasks. The NameNode is the clerk of HDFS; it monitors how your grinds are separated into record squares, which hubs store those pieces, and the general wellbeing of the circulated file system.
- DataNode: every slave machine in your cluster will host a DataNode daemon to carry out the grunt job of the distributed file structure—reading and writing HDFS blocks to real files on the local file system.
- JobTracker: It is runs on master and only one overall JobTracker schedules tasks on the slaves. It also monitors the Tasks and the Job. If any tasks failed then re-executes it.
- TaskTracker: One TaskTracker per slave node. It executes tasks as directed by the Master. Data storage and computation largely happen on the same node to maximize throughput.
- SecondaryNameNode: The Secondary NameNode (SNN) is a subordinate daemon for observing the condition of the group HDFS. Like the NameNode, every bunch has one SNN, and it normally dwells all alone machine also. No other DataNode or TaskTracker daemons keep running on the same server. The SNN varies from the NameNode in this procedure doesn't get or record any continuous changes to HDFS. Rather, it speaks with the NameNode to take previews of the HDFS metadata at interims characterized by the bunch design.

The term Hadoop has come to refer not just to the base modules, but also to the ecosystem [5], on the other hand gathering of extra programming bundles that can be introduced on top of or nearby Hadoop, for example, Apache Pig, Apache Hive, Apache HBase, Apache Phoenix, Apache Spark, Apache ZooKeeper, Cloudera Impala, Apache Flume, Apache Sqoop, Apache Oozie, and Apache Storm [6].

CONCLUSION

Due to the arrival of new technologies, high-end devices, and communication means for social networking sites like Facebook, twitter, the amount of data produced by mankind is growing rapidly every year. Traditional applications and tools are inefficient to handle such large volume of data sets. Hadoop allows us to store, process and retrieve Big Data efficiently. Hadoop framework is the best solution handle a large volume of the data set.

ACKNOWLEDGMENT

We are grateful to Finolex Academy of Management & Technology, Ratnagiri for providing technical support and guidance and giving permission to publish this work. We are pleased to our colleagues for their support and inspiration to complete this paper.

REFERENCES

[1] Jean Marc Rommes, "The Long View: How Big Is The Internet?". N.p., 2009. Web. 7 Feb. 2016.

[2] "Michael J. Cafarell**a**". Web.eecs.umich.edu. Retrieved 2013-04-05.

[3] Hadoop creator goes to Cloudera.

[4] Vance, Ashlee (2009-03-17). "Hadoop, a Free Software Program, Finds Uses Beyond Search". The New York Times. Archived from the original on 11 February 2010. Retrieved 2010-01-20.

[5] "Hadoop contains the distributed computing platform that was formerly a part of Nutch. This includes the Hadoop Distributed Filesystem (HDFS) and an implementation of MapReduce." About Hadoop.

[6] "Continuuity Raises $10 Million Series A Round to Ignite Big Data Application Development Within the Hadoop Ecosystem". finance.yahoo.com. Marketwired. 2012-11-14. Retrieved 2014-10-30.

[7] "Hadoop-related projects at". Hadoop.apache.org. Retrieved 2013-10-17.

[8] http://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm

[9] "Hadoop in Action" by Chuk Lam.