ISSN:0975-9646

Big Data: An Emerging Trend In Future

Sampada Lovalekar

Depart of IT, SIES Graduate School of Technology,
Nerul, Navi Mumbai, India

***Abstract*— The amount of data in world is growing day by day. Data is growing because of use of internet, smart phone and social network. Big data is a collection of data sets which is very large in size as well as complex. Generally size of the data is Petabyte and Exabyte.Traditional database systems is not able to capture,store and analyze this large amount of data. As the internet is growing, amount of big data continue to grow. Big data analytics provide new ways for businesses and government to analyze unstructured data. Now a days, Big data is one of the most talked topic in IT industry. It is going to play important role in future. Big data changes the way that data is managed and used. Some of the applications are in areas such as healthcare, traffic management,banking,retail, education and so on. Organizations are becoming more flexible and more open.New types of data will give new challenges as well. This paper highlights important concepts of big data.**

***Keywords*— Big data, Hadoop, HDFS, Map Reduce, NoSQL**

I. INTRODUCTION

In the past, type of information available was limited. There was a well-defined set of technology approaches for managing information. But in today's world, the amount of data in our world has been exploding. It has grown to terabytes and petabytes. Big data distinct from large existing data stored in various relational databases. It refers to a collection of large data sets which are very complex. Big data can be described using following terms:

1) *Volume*: Some small sized organizations may have gigabytes or terabytes of data storage. Data volume will continue to grow, regardless of the organization's size.

Many of these companies' datasets are within the terabytes range today but, soon they could reach petabytes or even exabytes. Machine generated data is larger in volume than the traditional data.

2) Variety: Different types of data are captured. It may be structured, semi structures or unstructured. Refers to the many different data and file types that are important to manage and analyze more thoroughly, but for which traditional relational databases are poorly suited. Some examples of this variety include sound and movie files, images, documents, geo-location data, web logs, text strings, web contents etc [1].

3. *Velocity*: The data is arriving continuously as streams of data. It is about the rate of change in the data and how quickly it must be used to create real value [2].

4. *Veracity*: If the data coming in large volume is not correct, it can create a problem and is of no use. So, it should be correct.

There are three types of big data. Structured, unstructured and semi structured. Similar entities are grouped together in structured data. Entities in the same group have the same descriptions. Examples are number, words, figures etc. Relational databases and spreadsheets are examples of structured data.Unstructured data is complicated information. Data can be of any type and does not follow any rule. It cannot be analyzed with normal statistical methods. For big data, different tools are required. Examples are social media, email, photos, multimedia etc. In semi structured data, similar entities are grouped together. Entities in same group may not have same attribute. Emails, EDI are example of this type of data.

Figure 1[3] shows organizations which are implementing or executing big data.

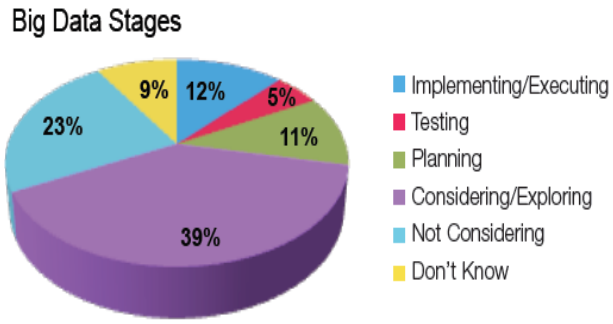

Fig 1 Big Data Stages

Big data is getting lot of attention now days**.** It will help to create new growth opportunities and entirely new categories of companies. Intelligence is combined with the production process. Constantly improving processing power and new techniques for data analysis mean that "Big Data" can be created from variety of sources. The creation of Big Data therefore permits organizations to create information about data that were never apparent or intended in the source information. If big data is used properly, enterprise can get a better view on their business. Some of the emerging applications are healthcare system, traffic management and many more.

The big data can come from various sources. It may be digitally generated and can be stored using a series of ones and zeros, and can be manipulated by computers. It may be mobile phone location data or call duration time or it may be a byproduct of our daily lives or interaction with digital Services. It may be generated using unconventional methods outside of data entry like, RFID, Sensor networks etc. It is the data generated by social networking sites.

II. DIFFERENCE BETWEEN TRADITIONAL AND BIG DATA ANALYTICS

Big data analytics can be differentiated from traditional data-processing architectures.In traditional data, sources are internal and structured. Data integration tools are used to extract, transform and load the data from transactional

databases. Then data quality and data normalization occur and the data is modelled into rows and columns. The modelled data is then loaded into an enterprise data warehouse. Big data is data that is too large to process using traditional methods. As the volume of data explodes, organizations will need analytic tools that are reliable, robust and capable of being automated .Traditional data warehouse is not able to handle processing of big data as data is coming from different sources like social media, video etc. This type of data grows at very high speed. The database requirements are very different in the case of big data. With big data analytics data can be anywhere and is in large volume. Big data analytics provides useful information. Hidden patterns are discovered. It focuses on unstructured data. Some technologies like Hadoop, NoSQL and Map Reduce are required for the analytics of big data.In big data analytics, the Hadoop system captures datasets from different sources and then performs functions such as storing, cleansing, distributing, indexing, transforming, searching, accessing, analyzing, and visualizing .So the unstructured data is converted into structured data. The working principle behind Hadoop and all big data is to move the query to the data to be processed, not the data to the query processor. Various languages used in the big data analytics are Java, Oracle JavaScript etc. Big data requires many different approaches to analysis, traditional or advanced, depending on the problem. It depends on the type of that particular problem. Some analytics includes traditional data warehouse concept.But some requires more advanced techniques. The IT techniques and tools to execute big data processing are new, very important and exciting. Big data technologies work faster than tradiditional data warehousing techniques.

Fig 2: Big data management

Figure 2. [4] Shows management of big data. Data is collected from various sources. It is not just a text data. It contains images, audio or video. It may be social data, machine generated data or documents. This data is unstructured. It is very critical to understand, categorize and analyze this large volume of data. A system is required to organize process and store this data into database so that it is analyzed efficiently. [4]. Table 1[5] explains the difference between traditional data analytics and big data analytics.

TABLE 1: DIFFERENCE BETWEEN TRADITIONAL DATA AND BIG DATA ANALYSIS

Traditional Data warehouse Analytics	**Big Data Analytics**
Traditional Analytics analyzes on the known data terrain that too the data that is well understood. Most of the data warehouses have a elaborate ETL processes and database constraints, which means the data that is loaded inside a data warehouse is well under stood, cleansed and in line with the business metadata.	The biggest advantages of the Big Data is it is targeted at unstructured data outside of traditional means of capturing the data. Which means there is no guarantee that the incoming data is well formed and clean and devoid of any errors. This makes it more challenging but at the same time it gives a scope for much more insight into the data.
Traditional Analytics is built on top of the relational data model, relationships between the subjects of interests have been created inside the system and the analysis is done based on them.	In typical world, it is very difficult to establish relationship between all the information in a formal way, and hence unstructured data in the form images, videos, Mobile generated information, RFID etc. have to be considered in big data analytics. Most of the big data analytics databases are based out Columnar databases.
Traditional analytics is batch oriented and we need to wait for nightly ETL and transformation jobs to complete before the required insight is obtained.	Big Data Analytics is aimed at near real time analysis of the data using the support of the software meant for it
Parallelism in a traditional analytics system is achieved through costly hardware like MPP (Massively Parallel Processing) systems and / or SMP systems.	While there are appliances in the market for the Big Data Analytics, this can also be achieved through commodity hardware and new generation of analytical software like Hadoop or other Analytical databases.

III TECHNOLOGIES USED FOR BIG DATA ANALYTICS

NoSQL database can handle unstructured and unpredictable data.The data stored in a NoSQL database is typically of a high variety.A NoSQL database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. Relational and NoSQL data models are very different. The relational model takes data and separates it into many interrelated tables that contain rows and columns. But document-oriented NoSQL database takes the data into documents using the JSON format. JSON is Javascript Object Notation.Another major difference is that relational technologies have rigid schemas while NoSQL models are schemaless. Many NoSQL databases have excellent integrated caching capabilities. So,the frequently used data is kept in system memory.NoSQL database types are[6] :

1) Document database: pair each key with complex data structure known as document. Document may contain nested document. This type of database store

unstructured (text) or semi-structured (XML) documents which are usually hierarchal in nature.

2)Graph stores: Graph databse is based on graph theory. It is used to store information about network.

3) Key value stores: Every single item is stored as an attribute name together with its value.

4) Wide column stores: They are optimized for queries over large datasets and store column of data together instead of rows.

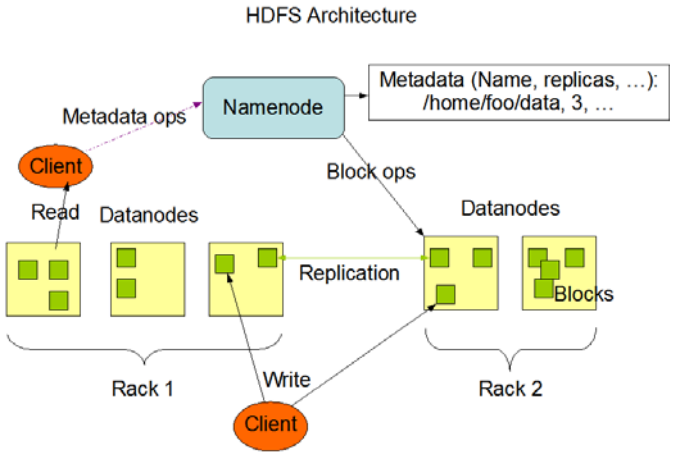

Fig. 1 HDFS Architecture

Apache Hadoop is a fast-growing big-data processing open source software platform. Hadoop can handle all type of data like structured, unstructured, pictures or audio. It runs on Linux, OS/X, Windows, and Solaris. Hadoop is scalable, flexible and fault tolerant.It contains HDFS. Hadoop HDFS is scalable, distributed file system written in Java. Figure 3 explains [7] HDFS architecture. HDFS has master/slave architecture. An HDFS cluster consists of a single NameNode. It manages the file system namespace. The name node is the equivalent of the address router for the big data implementation. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. New nodes can be added as needed and added without needing to change data formats The DataNodes are responsible for serving read and write requests from the file system's clients. DataNodes also performs function like block creation, deletion and replication as per the instruction of NameNode [7].

Hadoop creates *clusters* of machines and coordinates work among them. If any of the clusters fails, then Hadoop continues to operate the cluster without losing data. Map Reduce is a programming model and software framework first developed by Google. It works like a UNIX pipeline. A Map Reduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result. MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster node. The master is responsible for scheduling the job's component tasks on the slave, re-executing the failed task [8]. The slave executes the task as directed by the master.

Some of the Hadoop related projects [9] are described as:

1) Pig: It is a Scripting language and run time environment. It allows users to execute MapReduce on a Hadoop cluster. Pig's language layer currently consists of a textual language called Pig Latin.

2) Hive: It provides SQL access for data in HDFS. Hive's query language, HiveQL, compiles to MapReduce. It also allows user-defined functions.

3) HBase: A scalable, distributed database that supports structured data storage for large tables. It is column based rather than row based.

4) Mahout: Library of machine learning and data mining algorithm. It has four types of algorithm.

5) Oozie: Oozie is a Java Web-Application that runs in a Java servlet-container – Tomcat. It is job coordinator and workflow manager.

6) BigTop: It is used for for packaging and testing the Hadoop ecosystem.

IV CHALLENGES AND OPPORTUNITIES WITH BIG DATA

Big data analytics faces different challenges. These are described s follows:

1) Heterogeneity and Incompleteness [1]: Machine analysis algorithms expect homogeneous data, and cannot understand nuance. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain.

2) Timeliness [1]: There are many situations in which the result of the analysis is required immediately. Given a large data set, it is often necessary to find elements in it that meet a specified criterion The larger the data set to be processed, the longer it will take to analyze. It is difficult to design a structure when data is growing in very high speed.

3) Human Collaboration [1]: A Big Data analysis system must support input from multiple human experts, and shared exploration of results.

4) Privacy and security: This is another big challenge preserving individual privacy. For example in the healthcare industry, record of individual is very personal. But it can be available from multiple sources. So,it is difficult to maintain privacy and security.

5) Data Quality: A large volume of data is processed. Analyzing which data is important and to capture it is a big challenge.

6) Analysis: Big data is coming from various data sources. So analytics is a challenge.

7) Skill: Big data require people with new skill sets. Managing big data effectively requires the right people.

Opportunities with big data are summarized as,

1) The use of big data [10] will become a key basis of competition and growth for individual firms.All the companies will use big data. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and capture value from deep and up-to-real-time information.

2) Big data has opportunities in the field of education. More detailed information for school can be generated.This is beneficial for teacher and parents.
3) Almost all sectors like computer and electronic products, insurance, and government will increase their productivity from the use of big data[10].
4) Concept of big data have practical application in the area of healthcare research. In health care, data is coming from medical records, radiology images, human genetics etc. More information is analyzed regarding patient care and disease. Hence studies can be completed faster. Big data will help better future diagnoses and treatment of the patient.
5) Use of smart phone and tablet leads to high amount of mobile data traffic. Big Data is important for mobile networks.
   It is useful to improve network quality, traffic planning, prediction of hardware maintenance etc [11].
6) Various branches of science generates large amount of experimental data. Fulfilling the demands of science requires a new way of handling data [11].

V CONCLUSION

Big data is relatively new phenomenon. The field of big data is going from different perspective. Big data analytics provide new ways for businesses and government to analyze unstructured data. Research and development is required in this field. There are many technical challenges that must be addressed. Research is required to find new way of handling the data. For many IT decision makers, big data analytics tools and technologies are now a top priority. Big Data is going to play very important role in the future. There is need of analytical software which can handle huge storage as well as processing requirment of big data.To extract more and new value, there will be a focus on developing effective analytics. ew big data technologies and tools have been and continue to be developed.

REFERENCES

[1] Payal Malik, Lipika Bose," Study and Comparison of Big Data with Relational Approach", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013 pp 564-570
[2] Wei Fan, Albert Bifet," Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2
[3] Big Data Survey Research Brief", Tech. Rep.SAS, 2013
[4] Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole Velicanu," Perspectives on Big Data and Big Data Analytics", Database Systems Journal vol. III, no. 4/2012
[5] Srinivasan,"SOA and WOA Article, Traditional vs. Big Data Analytics, "Why big data analytics is important to enterprises", [Online].Available: http://soa.sys-con.com/node/1968472
[6] Available: http://www.mongodb.com/learn/nosql
[7] Dhruba Borthakur, The *Hadoop Distributed File System: Architecture and Design*
[8] [Online] Available: https://hadoop.apache.org
[9][Online]Available*:* http://www.revelytix.com/?q=content/hadoop-ecosystem
[10] Available: http://www.mckinsey.com
[11] NESSI-Big Data White Paper," Big Data –a new world of opportunities" December 2012