Big Data Analytics: A Review

Snehal N. Kathale
Asst. Professor
Department of Information Technology
Atharva College of Engineering,
Maharashtra,India

Supriya Mandhare
Asst. Professor
Department of Information Technology
Atharva College of Engineering,
Maharashtra,India

Chanda Chouhan
Asst. Professor
Department of Information Technology
Atharva College of Engineering
Maharashtra,India

Komal Mahajan
Asst. Professor
Department of Information
Technology
Atharva College of Engineering
Maharashta India

Jyoti Golakia
Asst.Professor
Department of Computer Engineering
Atharva College of Engineering
Maharashta, India

***Abstract*—In this paper we discussed an overview of big data analytics and what is the need of that. Big data is the data set which is big in volume, variety, velocity, value and varacity.Big data is mostly useful for the industries and in marketing for handling the larger datasets and management. We studied the different tools which is helpful for big data.It arises the many challenges and opportunities for the researchers and the developers.Big data is randomly changes our economic environment.**

Keywords—Big data; analytics,decision making

I. INTRODUCTION

In this era data increased day by day. Data is distinct pieces of information, usually formatted in a special way. All software is divided into two general categories: data and programs. Programs are collections of instructions for manipulating data. Data can exist in a variety of forms as numbers or text on pieces of paper, as bits and bytes stored in electronic memory, or as facts stored in a person's mind.Strictly speaking, data is the plural of datum, a single piece of information. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. [1]

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of fianding correlations or patterns among dozens of fields in large relational databases. After the data mining and data warehouse new field is come in computers i.e Big data. Nowadays through the advancements in technologies and the internet. With the in-crease in storage capabilities and methods of data collection, huge amounts of data have easily available. Every second, more and more data is being created and needs to be stored and analyzed in order to extract value. Furthermore, data has be-come cheaper to store, so organizations need to get as much value as possible from the huge amounts of stored data.[3].

The size, variety, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such sheer amounts of big data need to be properly analyzed, and pertaining information should be extracted.[1]Big data is rapidly changing our economic environment.We have studied the information about big data in next chapters.in this paper we discussing the Tools and methods,charecteristics,applications and need of big data.

II. BIG DATA

Big Data refers to enormous amounts of unstructured data produced by high-performance applications falling in a wide and heterogeneous family of application scenarios: from scientific computing applications to social networks, from e-government applications to medical information systems, and so forth.In big data, data is surprisingly changing from Gigabyets to TeraBytes and TeraBytes to Peta bytes.According to the survey of different social media sites; Every minute, over 300 hours of new video is uploaded Facebook, the most active of social networks, with over 1.4 billion active monthly users, generates the most amount of **s**ocial data – users like over 4 million posts every minutes – 4,166,667 to be exact, which adds up to 250 million posts per hour**.** Instagram, with 300 million monthly users in 2015, comes in second with 1,736,111 likes on photos each minute of the day, amounting to over 100 million likes per hour. Fig1. Shows that the data generation from different sources.

In a single minute social networking sites generates 50-60TB data..Big data is nothing but the data is in larger volume, velocity ,value, variety and varacity.Its depend on 5V's.

big data refers to two major concepts:

1.The breathtaking speed at which we are now generating new data

2.Our improving ability to store, process and analyze that data.

Database is the structured data but big data is the combination of structured as well as unstructured data. It contains images videos text file,etc. Big Data is a big thing. It will change our world completely and is not a passing fad that will go away.

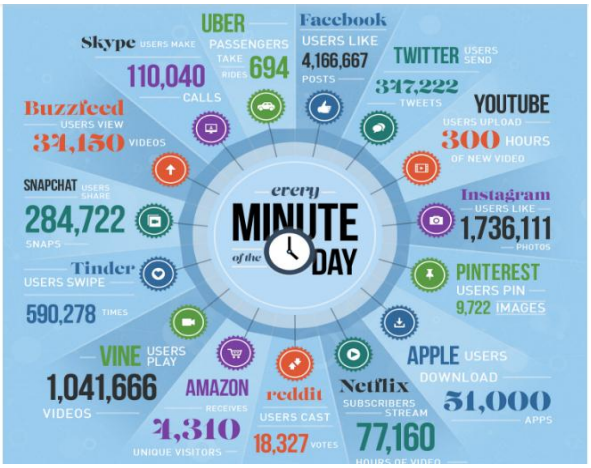

Fig.1 Data generation in single minute.

To understand the phenomenon that is big data, it is often described using five Vs: Volume, Velocity, Variety, Veracity and Value. This 5V's are the characteristics of big data

Volume refers to the vast amounts of data generated every second. Just think of all the emails, twitter messages, photos, video clips, sensor data etc. we produce and share every second. We are not talking Terabytes but Zettabytes or Brontobytes. On Facebook alone we send 10 billion messages per day, click the "like' button 4.5 billion times and upload 350 million new pictures each and every day. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute! This increasingly makes data sets too large to store and analyse using traditional database technology. With big data technology we can now store and use these data sets with the help of distributed systems, where parts of the data is stored in different locations and brought together by software[3].

Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds, the speed at which credit card transactions are checked for fraudulent activities, or the milliseconds it takes trading systems to analyze social media networks to pick up signals that trigger decisions to buy or sell shares. [3]

Variety refers to the different types of data we can now use. in the past we focused on structured data that neatly fits into tables or relational databases, such as financial data (e.g. sales by product or region). in fact, 80% of the world's data is now unstructured, and therefore can't easily be put into tables (think of photos, video sequences or social media updates). with big data technology we can now harness differed types of data (structured and unstructured) including messages, social media conversations, photos, sensor data, video or voice recordings and bring them together with more traditional, structured data.[3]

Veracity refers to the messiness or trustworthiness of the data. with many forms of big data, quality and accuracy are less controllable (just think of twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but big data and analytics technology now allows us to work with these type of data. the volumes often make up for the lack of quality or accuracy.[3]

Value: then there is another v to take into account when looking at big data: value! it is all well and good having access to big data but unless we can turn it into value it is useless. so you can safely argue that 'value' is the most important v of big data. it is important that businesses make a business case for any attempt to collect and leverage big data. it is so easy to fall into the buzz trap and embark on big data initiatives without a clear understanding of costs and benefits.[3]

III. BIG DATA OPPORTUNITIES

Recently, several US government agencies, such as the National Institutes of Health (NIH) and the National Science Foundation (NSF), ascertain that the utilities of Big Data to data-intensive decision-making have profound influences in Their future developments [4]. Consequently, they are trying to developing Big Data technologies and techniques to facilitate their missions after US government passed a large-scale Big Data initiative. This initiative is very helpful for building new capabilities for exploiting informative knowledge and facilitate decision-makers. From the Networking Information Technology Research and Development (NITRD) program which is recently recognized by President's Council of Advisors on Science and Technology (PCAST), we know that the bridges between Big Data and knowledge hidden in it are highly crucial in all areas of national priority. This initiative will also lay the groundwork for complementary Big Data activities, such as Big Data infrastructure projects, platforms development, and techniques in settling complex, data-driven problems in sciences and engineering. Finally, they will be put into practice and benefit society. According to the report from McKinsey institute [4], the effective use of Big Data has the underlying benefits to transform economies, and delivering a new wave of productive growth. Taking advantages of valuable knowledge beyond Big Data will become the basic competition for today's enterprises and will create new competitors who are able to attract employees that have the critical skills on Big Data. Researchers, policy and decision makers have to recognize the potential of harnessing Big Data to uncover the next wave of growth in their fields. Informing strategic direction, developing better customer service, identifying and developing new products and services, identifying new customers and markets, etc. The vertical axis denotes the percentages that how many enterprises think Big Data can help them with respect to specific purposes. [1]

IV. BIG DATA TOOLS

The fig 2 shows the architecture of big data. It shows the overall conceptual model of big data tools and methods and how the big data works.

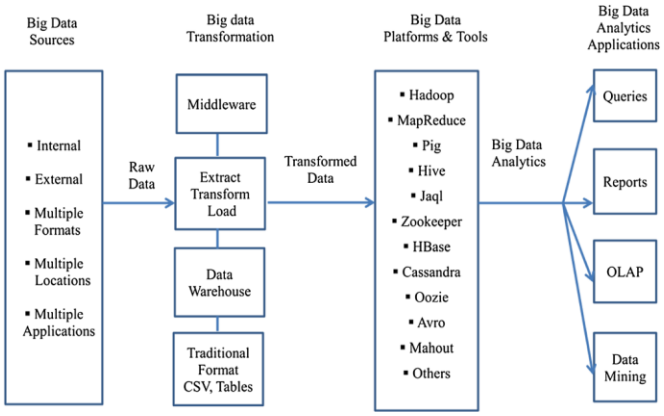

Fig 2.Architecture of big data [2]

There are the various sources for increasing data. Social networking sites play a very important role in larger data. There are the various applications which participate for database.Big data is the best for data store and management.Hadoop is the main software which is used for handling the database and manage well.Hadoop is nothing but the cominations of different tools.All this tools designed by Apache.

The name Hadoop has become synonymous with big data. It's an open-source software framework for distributed storage of very large datasets on computer clusters. All that means you can scale your data up and down without having to worry about hardware failures. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.[7]

Mapreduce is the basically combination of two words Map and Reduce.Map taken the data and converts into another set of data. Ad reduce performs summary of operations.It is the data processing unit for multiple computing nodes.Mapreduce frame work works on sinle master jobtracker and slave tasktracker on different nodes. The MapReduce framework operates on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types.[8]

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

Zookeeper behaves like his name. Keep watching on data base.zookeeper is the service which maintaining configuration information, naming, providing distributed synchronization, and providing group services.Its coordinate and managing all the services likes his child.

Hive is basically the query tool like sql.bt it performs queries on structured as well as unstructured dataset. Hive is used to query, summarize, explore and analyze that data. Hive supports all the common data formats such as BIGINT, BINARY, BOOLEAN, CHAR, DECIMAL, DOUBLE, FLOAT, INT, SMALLINT, STRING, TIMESTAMP, and TINYINT etc.[9]Hive is works on Hortonworks[8]

The Apache Mahout [9] aims to provide scalable and commercial machine learning techniques for large-scale and intelligentdata analysis applications. Many renowned big companies, such as Google, Amazon, Yahoo!, IBM, Twitter and Facebook, have implemented scalable machine learning algorithms in their projects. Many of their projects involve with Big Data problems and Apache Mahout provides a tool to alleviate the big challenges. Mahout's core algorithms, including clustering, classification, pattern mining, regression, dimension reduction, evolutionary algorithms and batch based collaborative filtering, run on top of Hadoop platform via the Map/reduce framework [46,47]. These algorithms in the libraries have been well-designed and optimized to have good performance and capabilities[1]

Oozie is the combination of workflow.it is tool which sort all the workflow between all the tools used in hadoop. Oozie combines multiple jobs sequentially into one logical unit of work. There are two basic types of Oozie jobs:Oozie Workflow jobs are Directed Acyclical Graphs (DAGs), specifying a sequence of actions to execute. The Workflow job has to wait Oozie Coordinator jobs are recurrent Oozie Workflow jobs that are triggered by time and data availability.

Habse is the nonsql database which is called as nosql.It is an open source nosql database that provide a real time access. HBase is A scalable, distributed database that supports structured data storage for large tables. Just as Bigtable leverages the distributed data storage provided by the Google File System, Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.

V. BIG DATA ANALYTICS APPLICATIONS

Big data is the latest software which is used in various applications. Big data is best for the industrial applications. Big data I basically used where the data size is too large..It has the various industrial applications. Here we see the different applications of big data.

1. Healthcare: Health data increasing day by day, big data is used in healthcare for clinical operations, for research and development, for patient analysisi, for monitoring the health index, Pre-adjudication fraud analysis, gene analysis etc.[2] Future applications of real-time data, such as detecting Infections as early as possible, identifying them swiftly And applying the right treatments (not just broad-spectrum Antibiotics) could reduce patient morbidity and mortality and even prevent hospital outbreaks.[4]

2. Data Mining and Clustering: to identify and address groups by customer type, text documents, products, patient records, click path, behaviour, purchasing patterns, etc. Decision trees illustrate the strengths of relationships and dependencies within data and are often used to determine what common attributes influence outcomes such as disease risk, fraud risk, purchases and online signups

3. Banking:In banks there is a huge data and day changes in a single minute,so its very difficult to manage all this data.So the bid data and hadoop is the very usegful to manage all this data of each and every user,transction details of all user, Further, outsourcing of data analysis activities or distribution of customer data across departments for the generation of richer insights also amplifies security risks

4. Agriculture: It plants test crops and runs simulations to measure how plants react to various changes in condition. Its data environment constantly adjusts to changes in the attributes of various data it collects, including temperature, water levels, soil. Composition, growth, output, and gene sequencing of each plant in the test bed. These simulations allow it to discover the optimal environmental conditions for specific gene types.

5. Finance: For performing its own credit score analysis for existing customers using a wide range of data, including checking, savings, credit cards, mortgages, and investment data.

It's useful also in various enterprises, for consumer goods, in credit cards, in stock exchange, in economy, in smart phones, in sap etc, This are the areas where hadoop is very useful for managing and clustering the data.

VI. CONCLUSION

As we have studies the big data technologies is the next door of innovation competition and productivity. It is a high performance application falling in wide and heterogeneous family of applications.As we seen in paper big data has the varios applications like in healthcare,in medicines in industries etc.Big data is very useful framework for the big industries to store and manage the data easily.It is the advances version of database and datamining.which changes the software era completely.The main purpose of this paper to give an idea about big data.and to explore the role of this in various fields.It is the discovery analytics for the new insights.

REFERENCES

[1] C.L. Philip Chen , Chun-Yang Zhang "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data" Information Sciences 275 (2014) 314–347
[2] Wullianallur Raghupathi1* and Viju Raghupathi2 "Big data analytics in healthcare: promise and potential" Health Information Science and Systems 2014, 2:3 http://www.hissjournal.com/content/2/1/3
[3] dilpreet singh and chandan k reddy "A survey on platforms for big data analytics" a Springer Journal of Big Data 2014, 1:8
[4] Kuchipudi Sravanthi, Tatireddy Subba Reddy "Applications of Big data in Various Fields" IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4629-4632
[5] http://enterprisearchitects.com/the-5v-s-of-big-data/
[6] Big data hub by IBM
[7] https://www.import.io/post/all-the-best-big-data-tools-and-how-to-use-them/
[8] https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.html
[9] Grant Ingersoll, Introducing apache mahout: scalable, commercial-friendly machine learning for building intelligent applications, IBM Corporation 2009
[10] Rui Maximo Esteves, Chunming Rong, Using mahout for clustering wikipedia's latest articles: a comparison between k-means and fuzzy c-means in the cloud, in: 2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom), 2011, pp. 565–56