BIG DATA: A LITERATURE SURVEY

[1]Jasdeep Kaur, [2]Vijay Dhir
[1]Research Scholar, SBBS University, Punjab
[2]Director R&D, SBBS University, Punjab

Abstract **:** Big Data is a fast growing research area in today's world and is used in almost every field. Big Data means huge amount of data generated every second through different sources on the Internet. This paper, is a study on Big Data and gives the snapshot on the basic fundamentals of Big Data, its tools and technologies. Understanding the term Big Data into its 5Vs and discuss the Hadoop ecosystem and usage areas of Big Data.

*IndexTerms***: Big Data,Hadoop ,MapReduce, Hadoop Ecosystem.**

I. INTRODUCTION

With our growing dependency on World Wide Web, last few years has shown massive data explosion. Every single click on the Internet generates data. More and more data is generated with the in creeping of technologies and Internet in our daily lives. We can't imagine a world without data storage. Every transaction performed, event occurred in any organisation or in any field requires data to be stored somewhere for extracting the useful information from the available data. The data generated is so huge that it cannot be handled by the traditional datasets. So the epidemic growth of data in terms of Gigabytes, Terabytes, Petabytes and Zettabytes is only handled by Big Data for data processing and extracting useful information through big data Analytics.

II. MEANING OF BIG DATA

As the name implies huge amount of data is called Big Data. Big Data can handle structured data sets; semi-structured data sets unstructured data sets which cannot be done by the traditional database management systems. Data comes from different sources like social media, mobile phones, online transactions, sensors etc. Big data has brought revolution in every sector such as transportation, E-commerce, medicines, smart cities traffic management, weather forecasting and many more. For example a retailer can follow user clicks on the web sites; like Amazon tracts the user clicks on the web site and suggest products according to the user requirement on his next visit to the web site; and even Google retains the user search history pattern just because of Big Data.

III. 5VS OF BIG DATA

Volume

Volume means size of data. As the name implies when data becomes bigger in size termed as Big Data. Data generated from social media, online business, hospitals, mobiles or any organisation or individual is so huge and reaching up to zettabytes, which cannot be stored on a single device.So the huge amount of generated data is stored on distributed systems, that is data is stored on different places over the network and brought together by a software when required. There are 3.5 billion searches on Google per day. Even a single Facebook or twitter "like" button click generates data. With the advancement of Internet of Things 2.5 quintillion bytes of data is generated every day.[2]

Velocity

Velocity refers to speed. Speed is a rate at which data is generated, stored and processed. When we are transferring a movie its size is in Gigabytes and it will take few minutes. But what happens when data is Big Data that is data is bigger in size like in petabytes; at that time it becomes difficult to capture and analyze the data in real time. The speed at which debit card transactions are made and accounts are updated instantly is just few seconds. We can analyze the data while it's being generated with the help of Big Data. Velocity also involves data input/output streams, creation of structured records and analyzes the data and data delivery. Amazon processed over an incredible 17 million transactions during prime day 36 hours of sale on July 2016**.** Google processes on average more than 3.5 billion searches per day.

Variety

Variety means different types of data formats. Data comes from different variety of sources like social networking sites, mobile phones, sensors, GPS etc. in different file formats images, videos, text ,audio, logs etc. Today data is no longer in the form of structured data that can be represented nicely into the table rather almost 80% of data is unstructured all over the world.

Veracity

Veracity means accuracy and correctness of data. Data should be certain and consistent on which the analysis is to be conducted. Only reliable and accurate data can be used for further processing from the huge amount of data generated. It takes lots of efforts to process the data. In Big Data semi structured and unstructured data forms are messy in nature and it takes lot of time and expertise to process that data for further analysis.

Value

Value refers to worth of data being extracted from the big datasets .This is the most important V of Big Data and the process of extracting valuable information from the big data sets is called Big Data Analytics. It enables the organisations to exploit the data according to their requirements. Big Data can provide value to any business .For example recommendations made by Amazon for you or Netflix notifying you about your most liked movies by analysing your past history on Netflix, Uber can predict demand, price of journey and provide closest cab to you by analysing GPS(Global Positioning System).

IV. HADOOP

Hadoop is one of the technologies designed for the processing of Big Data. Hadoop is an open source platform that has computational power and analytical technologies to work with huge amount of data. A Hadoop cluster consist of single master node and multiple slave nodes. Hadoop works on distributes file system and developed HDFS (Hadoop Distributed File System) that can process the huge amount of data without losing any data while processing. In large cluster, if any failure occurs HDFS continues its working by taking snapshot of the name node on secondary name node of a name node server to the host file system and secondary name node replaces the primary name node thus it can prevent data loss. HDFS stores the data on cluster by breaking down the incoming files into pieces called blocks and store a copy of each block across the pool of cluster. HDFS creates three copies of each file on three different servers.Hdoop requires JRE1.6 (Java Runtime Environment) or higher version of JRE.Fig-1 gives the architecture of Hadoop cluster.

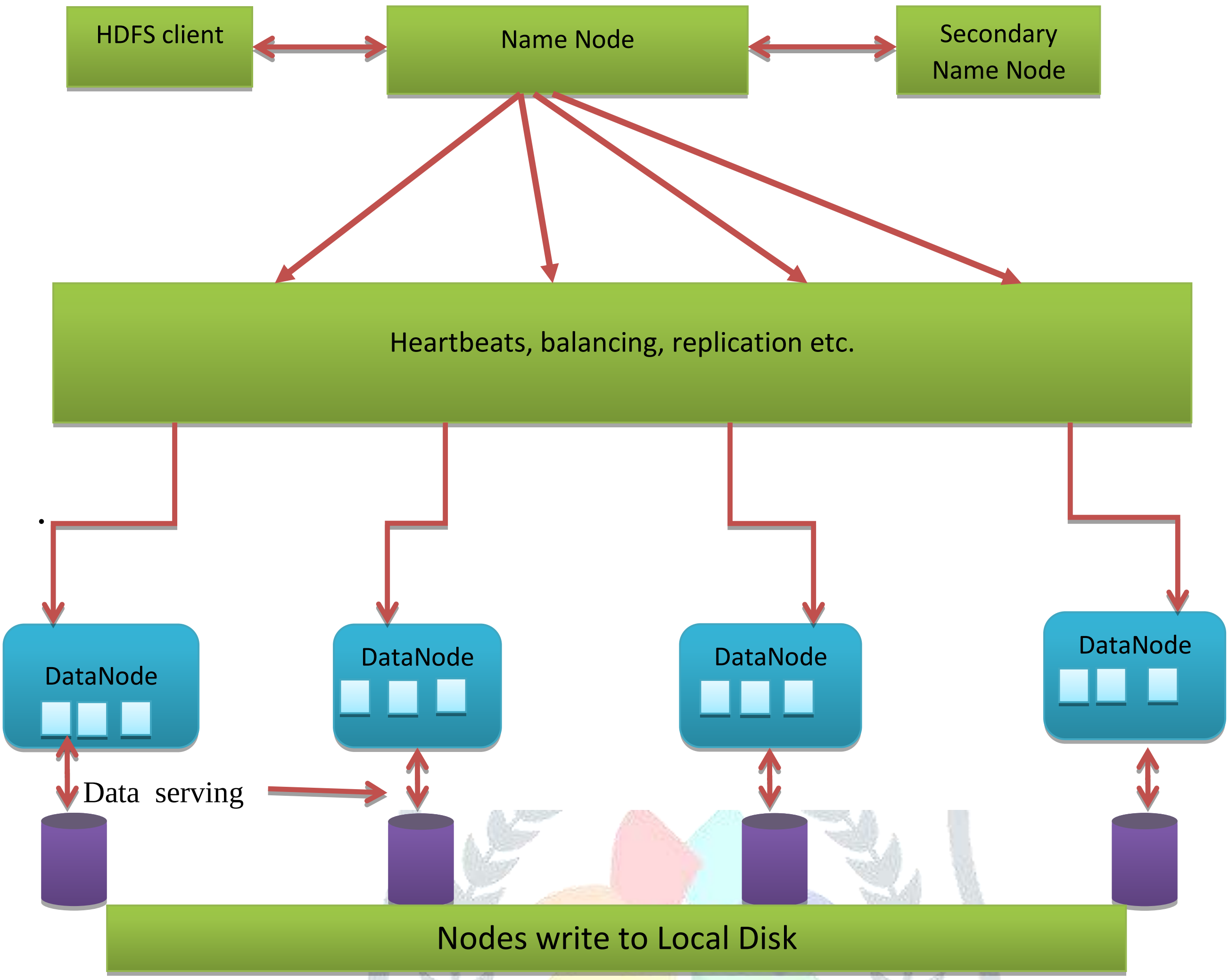

Fig. 1: Hadoop cluster architecture

V. MapReduce

MapReduce is software for the processing of large datasets discovered by the Google for the cluster of machines. MapReduce helps the developer to write the programs to process the huge volume of unstructured data parallel over a distributed and standalone architecture and provide aggregated useful results. Hadoop MapReduce is computational model framework used for mathematical calculation in Hadoop on a parallel and distributed implementation of MapReduce algorithm for high performance. MapReduce architecture have two functions Map () and reduce (). Map () function is responsible for the mapping of subtask to different nodes and reduce () function is responsible for reducing the responses from different nodes to a single unit. MapReduce has following components:

JobTracker

Job tracker is a master node that manages all the jobs and resources, allocate and monitor the job in the cluster of machines.

TaskTracker-

Task tracker runs the map and reduce task in the cluster of machines.

JobHistoryServer-

It keeps the track of all completed jobs

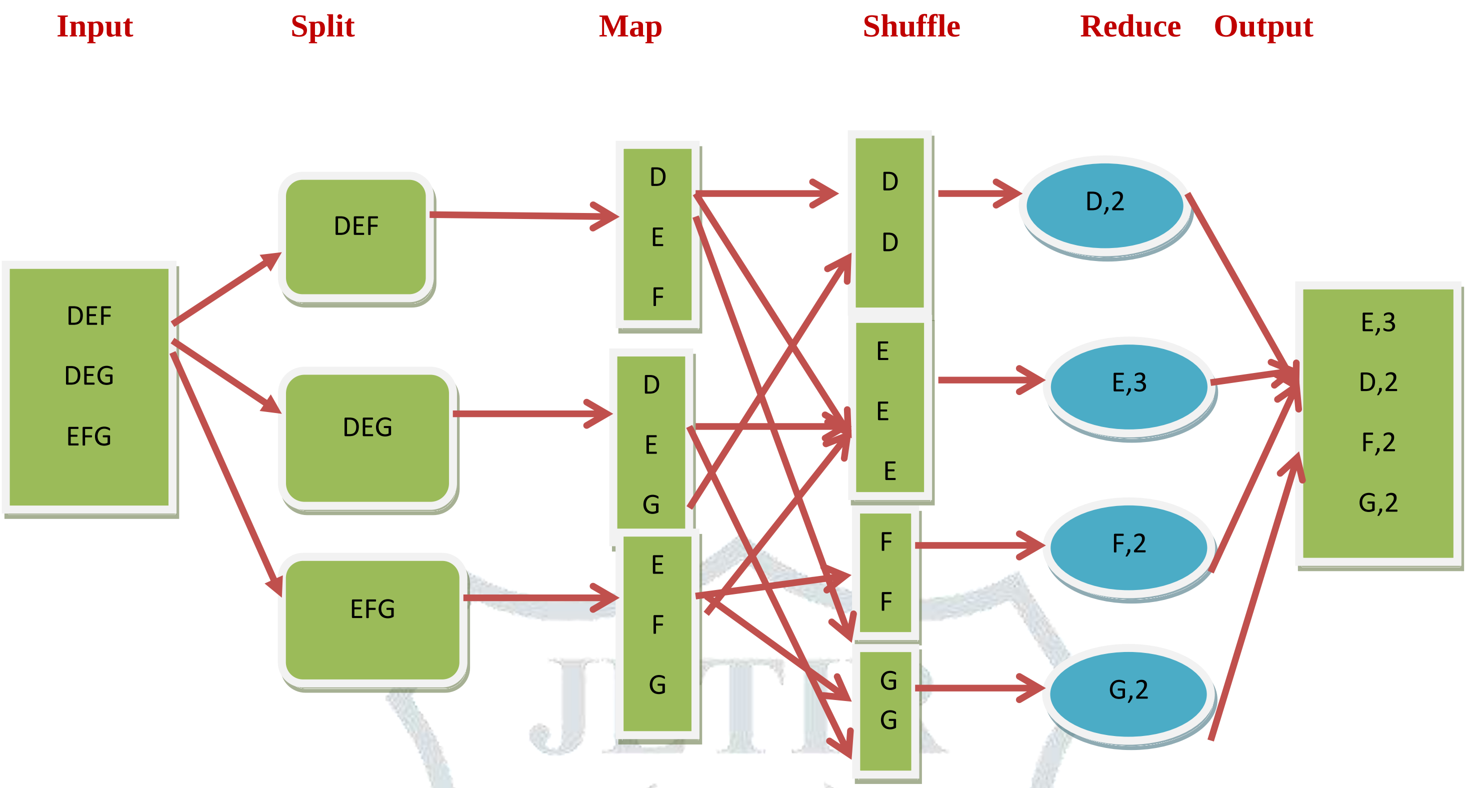

Fig. 2 : MapReduce Architecture

VI. Hadoop ECOSYSTEM

Hadoop ecosystem is a framework of various types of complex tool and technologies that are different in architecture but implemented and deployed to process Big Data in cost effective manner. Core components of Hadoop ecosystem are HDFS and MapReduce, along with these Haddop ecosysytem has various other elements to support the Big Data like Sqoop, Flume, Zoopkeeper, Oozie, Pig, Hive, Yarn, HBASE.

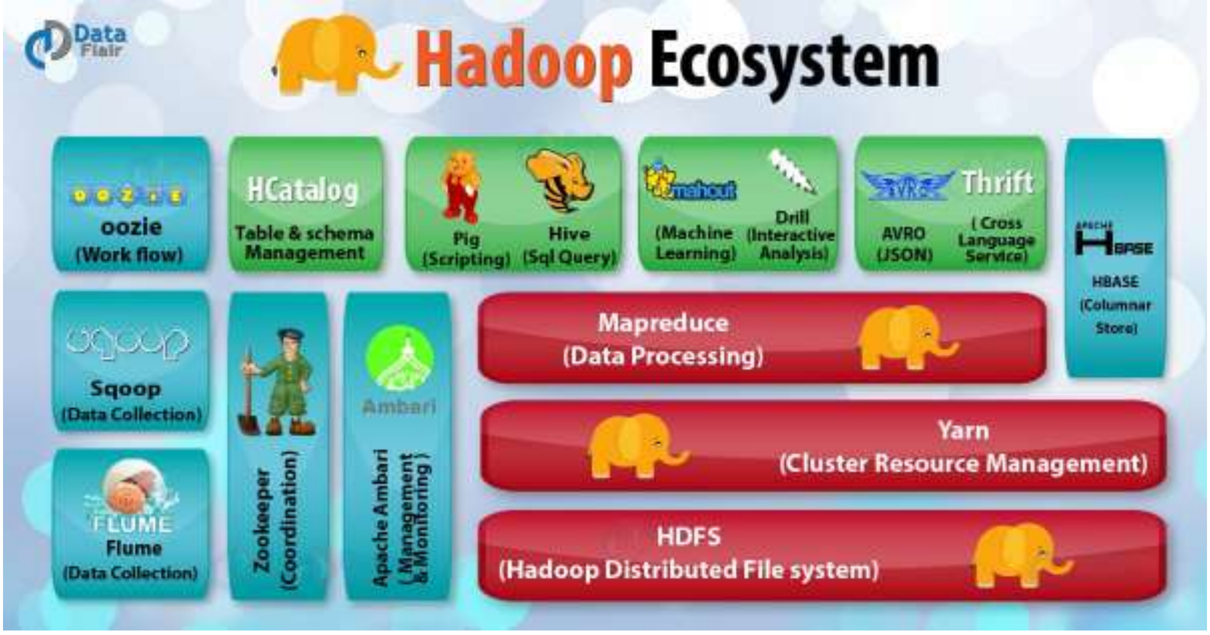

Fig 3: Hadoop Ecosystem (Image source [16])

VII. APPLICATIONS OF BIG DATA

HealthCare

Heath care industries and pharmaceutical companies uses big data to improve their research practices like invention of new drugs, clinical trail and data analysis. It helps the doctors to analyse medical history of every patient and provide services to them according to their requirements. With a added mHealth ,eHealth and other technologies, data is increasing tremendously .

Telecom

In telecommunication industry a lot of data is generated through mobile and internet. Big Data Analytics allow telecom industry to extract the useful information for improved customer services, call detail record analysis, mobile user location analysis through GPS, launching of new schemes by keeping in control on network and maximization of profit by lowering the cost.

Retail sector: In retail sector there is large volume of data generated from customer loyalty cards, inventory, employees etc. Big data extract the useful information for campaigning management, customer loyalty programs, current market trends, increase /decrease of demand of a product, so the timely decision can be taken for better results.

Financial institutions: Financial institutions like banks need third party evaluation for the customer credit score but with the evolution of big data banks can calculate its customer credit score by analysing their savings, incomes, credit card history etc. Big Data helps in detecting frauds like misuse of debit and credit cards of the customers.

Travel: Travel and tourism industry can generate more revenues with the help of big data analytics. They can offer more discounted packages to their consumers by analysing previous consumer demand patterns. For example an airline company can track the record of their consumer's travel, to analyse their frequent visited destinations, Window seat preferences, pricing in the competitive market etc.

Education

In education sector universities, colleges and other educational institution uses big data .Big Data developed new approaches like E –Learning, E-Books for teacher to impart knowledge and for students for their better performance. A lot of online data is available with the universities for faculty about the students like interest of their subject , evaluation results, analysing the behaviour signal of the students by which teacher can give more attention to the weaker students to improve their performance and can reach out the students especially distance learning over vast geographical area like bijous.

Transportation

Big Data helps in real time traffic management through the sensors installed on the roads .It can sense the traffic jam in particular area and route the traffic from some other routes to avoid traffic jams. For example GPS(Global positioning system) uses big data to analyse the traffic information and shows red line on traffic jam area ,shows green line on clear roads and can suggest you alternative routes.

Aviation Industry

Big Data plays an important role in the commercial aviation industry. Airline industry maintains a detailed record of their customer and analyse data like customer personal details, flying preferences etc. Every aircraft generates lot of data during operation, this data can be analysed for identifying the parts that require repairs, any technical fault in air craft, whether conditions and air traffic information etc.

Government

Data generated through online transactions, social media, adhaar cards etc. can be analysed for various public sector areas. By introducing GST (Goods and Services Tax) in India Big data analyses the data to bring transparency in the business and to know how the entire trade works over the Goods and Services Network. Government uses Big Data analytics in agriculture for better productivity of land and crops. As a result geotagging is used for the agriculture infrastructure. Big Data analytics is used in making public policies on health, education, preventing fraudulent activities etc.

VIII. CONCLUSION

We have entered the world of Big Data. This paper describes the fundamental concept of Big Data and its 5Vs.The basic understanding of Big Data tools and technologies. Hadoop , MapReduce Architecture and Hadoop ecosystem. Now we can have better understanding of HDFS. This paper also focuses on various application areas of Big Data that changes our lives for betterment.

REFERENCES

[1]. Harshawardhan S. Bhosale1 , Prof. Devendra P. Gadekar2. October 2014.International Journal of Scientific and Research Publications, Volume 4, Issue 10.pp-

[2]. Available online. http://www.internetlivestats.com/google-search-statistics/

[3].Mrs. Mereena Thomas. Dec-2015. A Review paper on BIG Data. International Research Journal of Engineering and Technology (IRJET) . Volume: 02 Issue: 09.pp-2

[4].Rahul Beakta. 2015. Big Data And Hadoop: A Review Paper.ResearchGate. Volume 2, Spl. Issue 2.PP-1-2

[5].Umar Ahsan, Abdul Bais. 2016. A Review on Big Data Analysis and Internet of Things. IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems.PP-1.

[6].Available online. Published March 5th 2018https://www.brandwatch.com/blog/47-facebook-statistics/

[7].Jeffrey Dean and Sanjay Ghemawat .2004 MapReduce: Simplified Data Processing on Large Clusters. *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA (2004), pp. 137-150

[8].By Xiufeng Liu, Christian Thomsen, Torben Bach Pedersen.2016. RESEARCH PAPER: MAPREDUCE-BASED DIMENSIONAL ETL MADE EASY.PP3-4.

[9].Soma Hota.2018. BIG DATA ANALYSIS ON YOUTUBE USING HADOOP and MAPREDUCE . International Journal of Computer Engineering in Research TrendsVolume-5, Issue-4 ,2018 Regular EditionPP-.

[10].Anurag Sarkar**,** Abir Ghosh, Dr. Asoke Nath .2015.MapReduce: A Comprehensive Study on Applications, Scope

and Challenges . International Journal of Advance Research in Computer Science and Management Studies. Volume 3, Issue 7.pp-

[11].Available online.2018. https://www.janbasktraining.com/blog/introduction-architecture-components-hadoop-ecosystem/

[12].G. Kalpana.2017. TECHNIQUES OF BIG DATA ANALYTICS: A LITERATURE SURVEY.International Journal of Advance Research. pp—

[13].Kuchipudi Sravanthi, Tatireddy Subba Reddy.2015. Applications of Big data in Various Fields.IJCSIT

[14].Ms. Swati S.Tawade.2018. Applications of Big Data: Review Paper. International Research Journal of Engineering and Technology (IRJET). Volume: 05 Issue: 02 .pp-

[15].S. Vanitha 1 , Dr .P. Balamurugan 2 .2018. BIG DATA AND HADOOP TECHNOLOGY-A STUDY . International Journal of Management, Technology And Engineering.Volume 8, Issue XII.pp—

[16].Availableonline.https://www.google.com/search?q=hadoop+ecosystem&source=lnms&tbm=isch&sa =X&ved=0ahUKEwjW5dCTh7jfAhVCaCsKHexUBg0Q_AUIDigB&biw=1366&bih=634#imgrc=gE90Yj uTiyZFmM:

[17].Available online. https://www.digitalvidya.com/blog/big-data-applications/

[18].Kuchipudi Sravanthi, Tatireddy Subba Reddy.2015. Applications of Big data in Various Fields. Kuchipudi Sravanthi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies. Vol. 6 (5) .pp-

[19].Available online. https://www.smartdatacollective.com/envisioning-tourist-demand-big-data/

[20].Available online.https://www.newgenapps.com/blog/how-big-data-is-taking-the-travel-industry-to-places

[21].Available online. https://www.analyticsinsight.net/how-indian-government-is-using-big-data-analytics-to-improve-economy-and-public-policy