AN EXHAUSTIVE STUDY: BIG DATA

Manisha Valera, Ankit Virparia, Om Mehta

Abstract— **Big Data is a novel term used to identify the datasets that due to their outsize and complexity. Big Data is Heterogeneous, Large Volume, and Distributed Data. Big Data are now swiftly escalating in all science and engineering domains, including physical, biological and biomedical sciences. Data mining is the capability of extracting useful information from these large datasets, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. This study paper includes the information about what is big data, Data mining with big data, Challenging issues and its related work. Big Data is thus very significant to increase productivity growth in the entire world since it is affecting public sectors. Big data refers to capacious data which ranges in Exabyte's (1018) and beyond. The archetype of processing huge datasets has been shifted from centralized architecture to distributed architecture. In this paper, we provide a widespread survey of Big data research, while prominence the specific concerns in Big data world. In this paper we have also discussed the challenges of Big data with various advantages and a disadvantage of these technologies. We represent the key issues in this area, and discuss the various methods to tackle these issues.**

Index Terms—**Big Data, Hadoop, Map Reduce, Data Mining**

I. Introduction

In era of Internet. The thing which is unfamiliar to us, we Search it. And in fractions of seconds we get the number of links as a outcome. This would be the better example for the processing of Big Data. This Big Data is not any diverse thing than out regular term data. Just big is a keyword used with the data to identify the collected datasets due to their large size and complexity? We cannot manage them with our current methodologies or data mining software tools. The online discussion provide a new means to sense the public interests and generate feedback in real-time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. The data collection has grownup enormously and is beyond the ability of commonly used software tools to capture, manage, and process within an endurable time.

In Big data the information comes from multiple, heterogeneous, autonomous sources with complex relationship and continuously growing. Up to 2.5 quintillion bytes of data are created daily and 90 percent data in the world today were produced within past two years [1].for example Flicker, a public picture sharing site, where in an average 1.8 million photos per day are receive from February to march 2012 [2].this shows that it is very difficult for big data applications to manage, process and retrieve data from large volume of data using existing software tools. It's become challenge to extract knowledgeable information for future use [3]. There are different challenges of Data mining with Big Data.

II. What Is Big data?

We create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This much amount of data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This huge amount of the data is known as "Big data". Big data is a buzzword, or catch-phrase, utilizes to describe a massive volume of both structured and unstructured data that is so huge that it's complicated to process using traditional database and software techniques. In most enterprise scenarios the data is too large or it moves too fast or it exceeds current processing capacity.

Big data has the potential to help organizations to improve operations and make quicker, more intelligent decisions [4]. Big Data, this term becomes communal in IT industries. As there is a huge amount of data lies in the industry but there is nothing before big data comes into trend. Big data is actually an evolving term that describes any capacious amount of structured, semi structured and unstructured data that has the potential to be mined for information. Big data doesn't refer to any specific quantity, so this term is often used when discussing about petabytes and exabytes of data [5]. When dealing with bigger datasets, organizations face complications in being able to create, manipulate, and manage big data. An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations.

Manuscript received June, 2016.

Manisha Valera, *Computer Engineering, Indus University, Ahmedabad, India*

Ankit Virparia, *Computer Engineering, Indus University, Ahmedabad, India*

Om Mehta, *Computer Engineering, Indus University, Ahmedabad, India*

FIVE V'S OF BIG DATA

There are many properties associated with big data. The prominent aspects are Volume, Variety, Velocity, Variability and Value [6].

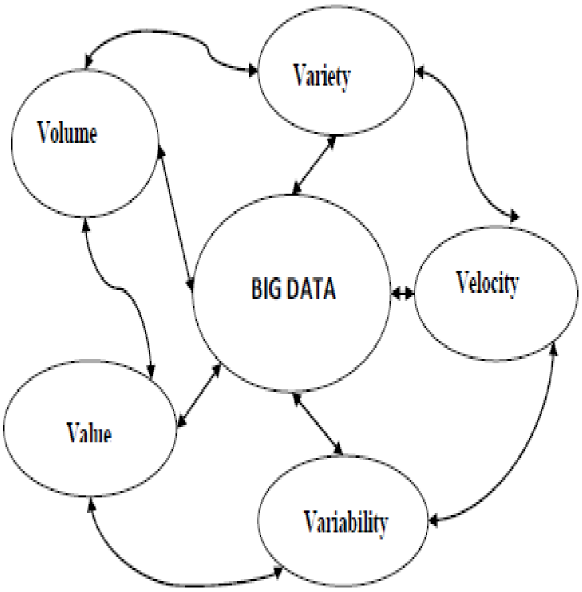

Figure-1. Five V's of Big Data.

A. Volume:

The volume of big data is exploding exponentially day to day. The data gathered through social websites and sensor networks going to cross from petabytes to Zeta bytes. Many factors contribute to the increase in data volume. Transaction-based data stored through the years. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

B. Variety:

Data produced are from different categories, consists of unstructured, standard, semi structured and raw data which are very difficult to be handled by traditional systems. Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, and financial transactions [4].

C. Velocity:

This concept indicates the speed at which the data generated and become historical. Big data is capable enough to handle the incoming and outgoing data rapidly. Data is streaming in at unprecedented speed and must be apportioned with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

D. Variability:

It describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

E. Value:

All enterprises and e-commerce systems are acute in improving the customer relationship by providing value added services. For that, study on customer attitudes and trends in the market are to be analyzed. Moreover, users can also query the data store to find business trends and accordingly they can change their strategies. By making big data open to all, it creates transparency on functional analysis. Supporting real time decisions and experimental analysis in different locations datasets can do wonderful things for enterprises.

TECHNICAL CHALLENGES IN BIG DATA

Big data faces many technical challenges which are on the roadway of the research.

A. Failure handling

Systems can be devised in such a way that the possibility of failure must fall within the permitted threshold. Fault tolerance is a technical challenge in big data. When a process started it may involve with numerous network nodes and the whole computation process becomes clumsy. Retaining check points and fixing the threshold level for process restart in case of Failure, are greater concerns.

B. Data heterogeneity

Big data deals with unstructured, semi-structured and structured data. Linking unstructured data with structured data, transforming data from one form into another requisite form needs a lot of research.

C. Data quality

For predictive analysis or for better decision making amount of relevant data helps a lot. But the quality of such data is based on the source through which they are resultant. Though big data stores large significant data, the accurateness of data is completely dependent on the source domains. Hence, there is a question of how far the data can be reliable and it definitely requires appropriate trust agent filters.

III. BIG DATA AND DATA MINING

The Big Data is nothing but a data, available at heterogeneous, autonomous sources, in extreme large amount, which get updated in fractions of seconds. For example, the data stored at the server of Facebook, as most of us, daily use the Facebook; we upload various types of information, upload photos. All the data get stored at the data warehouses at the server of Facebook. This data is nothing but the big data, which is so called due to its complexity. Also another example is storage of photos at Flicker. These are the good real-time examples of the Big Data. Another best example of Big data would be, the readings taken from an electronic microscope of the universe. Now the term Data Mining, Finding for the exact useful information or knowledge from the collected data, for future actions, is nothing but the data mining. So, collectively, the term Big Data Mining is a close up view, with lots of detail information of a Big Data with lots of information [8].

KEY FEATURES OF BIG DATA

The features of Big Data are:

- It is huge in size.
- The data keep on changing time to time.
- Its data sources are from different phases.
- It is free from the influence, guidance.
- It is too much complex in nature, thus hard to handle.

CHALLENGING ISSUES IN DATA MINING WITH BIG DATA.

There are three sectors at which the challenges for Big Data arrive. These three sectors are:

- Mining platform.
- Privacy.
- Design of mining algorithms.

Basically, the Big Data is stored at different places and also the data volumes may get increased as the data keeps on increasing continuously. So, to collect all the data stored at different places is that much expensive. Suppose, if we use these typical data mining methods (those methods which are used for mining the small scale data in our personal computer systems) for mining of Big Data, and then it would become an obstacle for it. Because the typical methods are required data to be loaded in main memory, though we have super large main memory. To maintain the privacy is one of the main aims of data mining algorithms. Presently, to mine information from big data, parallel computing based algorithms such as MapReduce are used. In such algorithms, large data sets are divided into number of subsets and then, mining algorithms are applied to those subsets. Finally, summation algorithms are applied to the results of mining algorithms, to meet the goal of Big Data mining. In this whole procedure, the privacy statements obviously break as we divide the single Big Data into number of smaller datasets. When we divide the Big Data in to number of subsets, and apply the mining algorithms on those subsets, the Results of those mining algorithms will not always point us to the actual result as we want when we collect the results together.

For those people, who intend to hire a third party such as auditors to process their data, it is very important to have efficient and effective access to the data. In such cases, the privacy restrictions of user may be faces like no local copies or downloading allowed, etc. So there is privacy-preserving public auditing mechanism proposed for large scale data storage [7]. This public key-based mechanism is used to enable third-party auditing, so users can safely allow a third party to analyze their data without breaching the security settings or compromising the data privacy. In case of design of data mining algorithms, Knowledge evolution is a common phenomenon in real world systems. But as the problem statement differs, accordingly the knowledge will differ.

IV. HADOOP: SOLUTION FOR BIG DATA PROCESSING

Hadoop is an Apache open source framework written in Java that allows distributed processing of large dataset across cluster of computers using simple programming model Hadoop creates cluster of machines and coordinates work among them. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop consists of two component Hadoop Distributed File System (HDFS) and MapReduce Framework [9].

A. *HDFS (Hadoop Distributed File System)*

HDFS is a file system designed for storing very large files with streaming data access pattern, running clusters on commodity hardware. HDFS manages storage on the cluster by breaking incoming files into pieces called ‘blocks’ and storing each blocks superfluously across the pool of the server. HDFS stores three copies of each file by copying each piece to three different servers. Size of each block 64MB. HDFS architecture is broadly divided into following three nodes which are Name Node, Data Node, and HDFS Clients/Edge Node.

1. Name Node

It is centrally placed node, which contains information about Hadoop file system. The main task of name node is that it records all the metadata & attributes and specific locations of files & data blocks in the data nodes. Name node acts as the master node as it stocks all the information about the system. And provides information which is newly added, modified and removed from data nodes.

2. Data Node

It works as slave node. Hadoop environment may contain more than one data nodes based on capacity and performance. A data node performs two main tasks storing a block in HDFS and acts as the platform for running jobs.

3. HDFS Clients/Edge node

HDFS Clients are sometimes also known as Edge node. It acts as linker between name node and data nodes. Hadoop cluster there is only one client but there are also many depending upon performance needs.

B. *MapReduce Framework*

MapReduce is defined as a programming model for processing and generating large sets of data. There are two phases in MapReduce, the "Map" phase and the "Reduce" phase. The system splits the input data into multiple chunks, each of which is assigned a map task that can process the data in parallel. Each map task reads the input as a set of (key, value) pairs and produces a transformed set of (key, value) pairs as the output. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (key, value) pairs to reduce task, which groups them into final results.

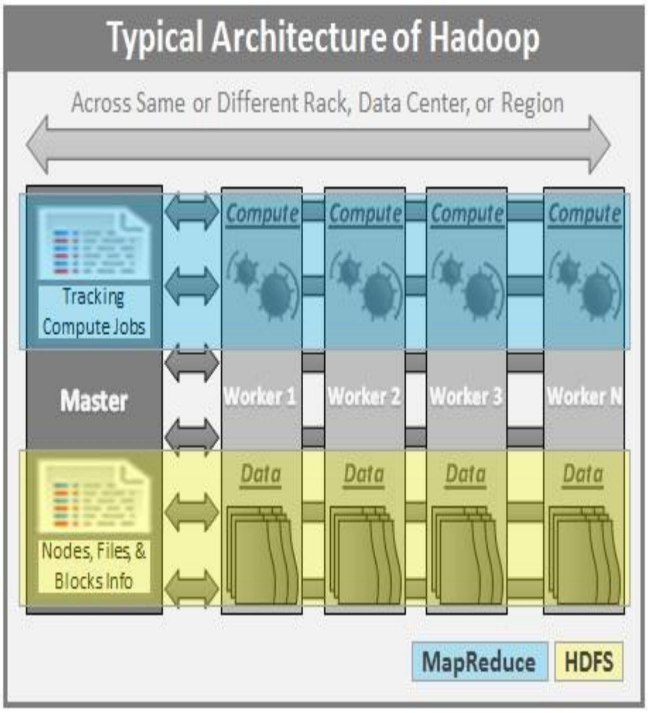

Fig -2: Architecture of Hadoop

Map-Reduce was introduced by Google in order to process and store large datasets on commodity hardware. Map Reduce is a model for processing large scale data records in clusters. Map function performs the task as the master node takes the input, divide into smaller sub modules and distribute into slave nodes. A slave node further divides the sub modules again that lead to the hierarchical tree structure. The slave node processes the base problem and passes the result back to the master Node. The Map Reduce system arrange together all intermediate pairs based on the intermediate keys and refer them to reduce() function for producing the final output. Reduce function works as the master node collects the results from all the sub problems and combines them together to form the output [10].

Map(in_key,in_value)--->list(out_key,intermediate_value)

Reduce(out_key,list(intermediate_value))--->list(out_value)

The parameters of map () and reduce () function is as follows:
map (k1,v1) ! list (k2,v2) and reduce (k2,list(v2)) ! list (v2)
A Map Reduce framework is based on a master slave architecture where one master node handles a number of slave nodes. Map Reduce works by first dividing the input data set into even-sized data blocks for equal load distribution. Each data block is then assigned to one slave node and is processed by a map task and result is generated. The slave node interrupts the master node when it is sluggish. The scheduler then assigns new tasks to the slave node. The scheduler takes data locality and resources into consideration when it disseminates data blocks.

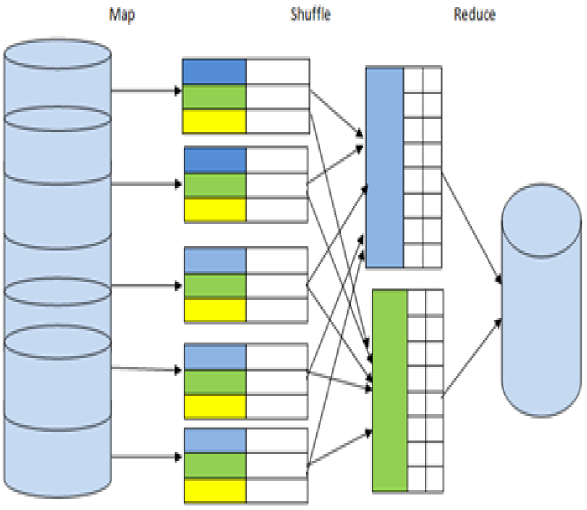

Fig. 3 Architecture of Map Reduce

Figure 3 shows the Map Reduce Architecture and Working. It always manages to allocate a local data block to a slave node. If the effort fails, the scheduler will assign a rack-local or random data block to the slave node instead of local data block. When map () function complete its task, the runtime system gather all intermediate pairs and launches a set of condense tasks to produce the final output. Large scale data processing is a difficult task, managing hundreds or thousands of processors and managing parallelization and distributed environments makes is more difficult. Map Reduce provides solution to the mentioned issues, as it supports distributed and parallel I/O scheduling, it is fault tolerant and supports scalability and it has inbuilt processes for status and monitoring of heterogeneous and large datasets as in Big Data . It is way of approaching and solving a given problem. Using Map Reduce framework the efficiency and the time to retrieve the data is quite manageable. To address the volume aspect, new techniques have been proposed to enable parallel processing using Map Reduce framework [12]. Data aware caching (Dache) framework that made slight change to the original map reduce programming model and framework to enhance processing for big data applications using the map reduce model [11]. The advantage of map reduce is a large variety of problems are easily expressible as Map reduce computations and cluster of machines handle thousands of nodes and fault-tolerance. The disadvantage of map reduce is Real-time processing, not always very easy to implement, shuffling of data, batch processing.

V. CONCLUSION

Big Data is rapidly mounting in this era, and we will have to manage much more amount of data in coming years. This data is going to be more miscellaneous, superior, and quicker. We discussed some perceptions about the topic, and what we consider are the main concerns and the main challenges for the future. Big Data is becoming the new Final Edge for data research and for business applications. Big Data mining will help us to discover knowledge that no one has discovered before. This paper discussed an architecture Using Hadoop HDFS distributed data storage, and MapReduce distributed data processing over a cluster of commodity servers. The main goal of our paper was to make a survey of various techniques which handle a massive amount of data from different sources and improves overall performance of systems.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu, and Wei Ding" Data Mining with Big Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014

[2] F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?" http://www.flickr.com/photos/franckmichel/6855169886/, 2012.

[3] Hadoop. http://hadoop.apache.org/.

[4] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule," Survey Paper On Big Data" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939

[5] Apache Pig. Available at http://pig.apache.org

[6] S. Justin Samuel, Koundinya RVP, Kotha Sashidhar and C.R. Bharathi," A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES", ARPN Journal of Engineering and Applied Sciences,VOL. 10, NO. 8, MAY 2015 ISSN 1819-6608

[7] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Secure Cloud Storage" IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[8] Rohit Pitre, Vijay Kolekar"A Survey Paper on Data Mining With Big Data"International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1, Issue 1, April 2014

[9] Varsha B.Bobade," Survey Paper on Big Data and Hadoop", International Research Journal of Engineering and Technology (IRJET) ,e-ISSN: 2395-0056 Volume: 03 Issue: 01 | Jan-2016 www.irjet.net, p-ISSN: 2395-0072

[10] Sabia and Love Arora," Technologies to Handle Big Data: A Survey", International Conference on Communication, Computing & Systems (ICCCS–2014)

[11] Wang, J.; Xiao, Q.; Yin, J.; Shang, P. Magnetics, "DRAW: A New Data-gRouping-AWare Data Placement Scheme for Data Intensive Applications With Interest Locality "IEEE Transactions (Vol: 49), 2013, 2514 – 2520s

[12] Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issue.1,2010, 72-77.

Manisha Valera has completed M.E in Computer Engineering from GTU in 2013. And right now she is working as an Assistant Professor in Indus University. Her areas of interest are Data mining, Big data. She has 5 research publications, in those 4 are of international journal and 1 is of international conference.

Ankit Virparia has completed M.Tech in Computer Science and Engineering University from Nirma University and right now he is working as an Assistant Professor in Indus University. His areas of interest is Big data. He has 2 research publications in international conference.

Om Mehta has completed Master of Technology (ICT) in 2013 from Nirma University and right now he is working as an Assistant Professor in Indus University. His areas of interest are Networking, Big data. He has 5 research publications, in those 2 are of international journal and 3 are of national conference.