International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 07
June 2017

Review on Big Data

TajinderKaur*
Sainik Institue,Ropar
tajindersaini1992@gmail.com

Abstract— Big data is a term that describes the large volume of data – both structured and unstructured. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data is popular term used to describe the exponential growth and availability of data, structured, semi-structured and unstructured data has the potential to be mined for information. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Big data is a broad term for data sets so large or complex that traditional data processing is inadequate.For the big companies like Facebook, Google, Yahoo etc. needs big data analytics to handle such unstructured data. Big data can be analyzed for insights that lead to better decisions and strategic business moves. This paper explains about big data analyzes and what kind different tools to handle it.
Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Hadoop is an open source software framework for storing data and running application. Hadoop makes it possible to run applications on system with thousands of nodes involving thousands of terabytes.

Keywords— Big data, Analyzing, Parameters.

I. INTRODUCTION

Defintion - Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tool, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. With the growth of technologies and services, the large amount of data is produced that can be structured and unstructured from the different sources. Such type of data is very difficult to process that contains the billions records of millions people information that includes the web sales, social media, audios, images and so on. The need of big data comes from the Big Companies like yahoo, Google, facebook etc for the purpose of analysis of big amount of data which is in unstructured form [1]

Big data often refers to simply handle of predictive analytics and user behavior analytics. Big data includes 3 V's process called 1. velocity, 2.volume, 3.variety . The big data can be used in education system, organizations, banking sector and companies . With the growth of technologies and services, the large amount of data is produced that can be structured and unstructured from the different sources. Data is very difficult to process which contains the billions records of millions people information that includes the web sales, social media, audios, images and so on.

Big data analytics analyze the large amount of information used to uncover the hidden patterns and the other information which is useful and important information for the use

Big Data Parameters

As the data is too big from various sources in different form, it has 3 parameters called as 3 Vs. The three Vs of Big Data are: Variety, Volume and Velocity

Fig 1- Big data parameters

Volume refers to the amount of data, Variety refers to the number of types of data and Velocity refers to the speed of data processing. According to the 3Vs model, the challenges of big data management result from the expansion of all three properties, rather than just the volume alone -- the complete amount of data to be managed.

II. Evolution of Big Data

In the last twenty years, the data is increasing day by day across the world .Some facts about the data are, there are 277,000 tweets every minute, 2 million searching queries on Google every minute, 72 hours of new videos are uploaded to YouTube, More than 100 million emails are sent, 350 GB of data is processing on facebook and more than 570 websites are created every minute. During 2012, 2.5 quintillion bytes of data were created every day. Big data and its analysis are the center of modern science and business areas. Large amount of data is generated from the various sources either in structure or unstructured form. Such type of data stored in databases and then it become difficult to Extract, transform and load. IBM indicates that 2.5 exabytes data is created everyday which is very difficult to analyze. The estimation about the generated data is that till 2003 it was represented about 5 exabytes, then until 2012 is 2.7 Zettabytes and till 2015 it is expected to increase 3 times [2]

III. BIG DATA ANALYZING PROCESS

Big data analytics is the process of examining large and varied data sets -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions.

After the big data storing process then it come anlayis process there are many critical requirements for big data. The first one is fast data loading and accesing data. The big data and network traffic interferes with the query exeute during data will be loading it is compulsory to deduct the loading time. And the next requirement is fast query process. In this process fulfillthe requirement of heavy workload and realtime request so many queries areresponse time are dangers. Thus data placement structure must capable for handling high qurery processing speed as the amount of queries fastly increase The third next requirements for big data processing is highly proper management Of storage space. They are fastly growth in user activity in demand for storage Capacity and power, data storage well management during process issue how To store data that space utilization is maximized be addressed. The final requirement is strong adaptivity to highly big workload patters. The big data sets are analyzed by different user and application for different purpose and in various way. The map reduce is a parallel process program model and develop by a map & reduce of functional languages which is accessible for big data processing. It is core of Hadoop and perform the data processing and analysis function. MapReduce process is work on adding more computer and handler also it increase the power efficiency and large space for in a specific computer. in another way to express it is divide the operation in part of task in stages and stage execute in parallel in process to reduce the time.[3]

IV. Summary

The process of the research into complex data basically concerned with the revealing of hidden patterns. Big data samples describe the review about the atmosphere, biological science and research. Life sciences etc.By this paper, we can conclude that any organization in any industry having big data can take the benefit from its careful analysis for the problem solving purpose. Using Knowledge Discovery from the Big data easy to get the information from the complicated data sets.[4]

The overall Evaluation describe that the data is increasing and becoming complex. The challenge is not only to collect and manage the data also how to extract the useful information from that collected data.

Grid Computing offered the advantage about the storage capabilities and the processing power and the Hadoop technology is used for the implementation purpose. Grid Computing

provides the concept of distributed computing. The benefit of Grid computing center is the high storage capability and the high processing power [4]. 1) According to Big Data, facebook has 1.11 billion people active accounts from which 751 million using facebook from a mobile [5].

2) In olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as "big data". In this they told that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems [6].

PROBLEM DEFINITION

The size of the data is growing day by day with the exponential growth of the enterprises. For the purpose of decision making in an organizations, the need of processing and analyses of large volume of data is increases. The various operations are used for the data processing that includes the tagging, highlighting, searching etc. Data is generated from the many sources in the form of structured as well as unstructured form. Big data sizes vary from a few dozen terabytes to many petabytes of data. The processing and analysis of large amount of data or producing the valuable information is the challenging task. As the Big data is the latest technology that can be beneficial for the business organizations, so it is necessary that various issues and challenges associated with this technology should bring out into light. The two main problems regarding big data are the storage capacity and the processing of the data [2].

TECHNIQUES AND TECHNOLOGY

A. Hadoop Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's Mapreduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, Mapreduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper [7].

B. HDFS HDFS is a block-structured distributed file system that holds the large amount of Big Data. In the HDFS the data is stored in blocks that are known as chunks. HDFS is client-server architecture comprises of Name Node and many Data Nodes. The name node stores the metadata for the NameNode. Name Nodes keeps track of the state of the Data Nodes. NameNode is also responsible for the file system operations etc [8].

C.HPCC HPCC is open source computing platform and provide the services for management of big data workflow. HPCC data model is defined by the user. HPCC system is designed to manage the most complex and data-intensive analytical problems. HPCC system is a single platform, a single architecture and a single programming language used for the data processing. HPCC system is based on Enterprise control language that is declarative, on-procedural programming

V. CONCLUSION

In the review paper on big data they are completely different topic. The paper decribes the cocept of big data along with V3 Volume, velocity and Variety of big data. This paper also proper study on big data processing problems. The Big data are challenges for better work and fast working process of big data. So they more data will store big data in that manner they will be easy to access. The big data analysis can be apply to converting business change and large Decision making by using advance analytic processes on big data, valuable knowledge. So the big

REFERENCES

[1] https://en.wikipedia.org/wiki/Information_security
[2] https://simple.wikipedia.org/wiki/Cryptography
[3] www.garykessler.net/library/crypto.html#types
[4] https://simple.wikipedia.org/wiki/Cryptographic_hash_function
[5] http://all.net/edu/curr/ip/Chap2-4.html