A REVIEW PAPER ON BIG DATA ANALYTICS

Kirti Bhatia[1], Lalit[2]

[1]HOD, Department of Computer Science, SKITM Bahadurgarh Haryana, India
bhatia.kirti.it@gmail.com
[2]M Tech 4th sem SKITM Bahadurgarh, Haryana, India
lalit.dang3@gmail.com

Abstract

The basis of this dissertation is the rise of "big data" and the use of analytics to mine that data. Big data is basically used to analyze the large amount of data. Companies have been storing and analyzing large volumes of data. In the recent times, data warehousing was mostly used for this purpose which was invented in the early 1990s. Big Data warehouses means terabytes that is data stored in the warehouses was terabytes in storage. But now it is petabytes, and the data is growing at a high speed. The growth of data is because the companies and organizations store and analyze the greater levels of transaction details, as well as Web data and machine-generated data, to achieve a better understanding of customer behavior and market behavior. Big data helps to achieve this goal. A large number of companies are using Big Data. It seems in coming years the growth of big data concept must be higher.

Keywords: *Data Analytics, Big Data, Large Volume Data, Web Data, Machine-Generated Data.*

----------***----------

1. INTRODUCTION

This dissertation is in the area of data mining in the large organizations to keep pace with the desire to store and analyze ever larger volumes of structured data. The vendors of RDBMS have given a number of special platforms that provide higher levels of price improvement and also a higher level of performance as compared to general purpose Relational DBMS. The platforms used for this purpose are available in a variety of shapes and sizes, from software only databases [1]. A large number of persons who make survey or we can say a large number of survey respondents said that they have an analytical platform that is applicable to this description.

As we know, new technologies are coming also have come to handle the complex data. These technologies can handle a large volume of this complex data which also includes the Web Data.

By web data, I mean the social media data like Facebook data. As we all use Facebook that is have an account on Facebook. A large number of pictures are uploaded to the Facebook server the content on Facebook server is very large and this type of content is social media content. New technologies also handle or overlook the machine generated data or the sensors and data like GPS that is Global Positioning System. These type of data is big data. By Big Data means very big data. [2]

I have discussed in above paragraph that big data or large volume of data is handled or we can also say analyzed by the new technologies. Now what are these new technologies? Let's discuss them in brief. A large number of companies, which includes Internet and also social media companies, are using new open source frameworks such as Hadoop and MapReduce to analyze, store and process large volumes of data which can be structured data also unstructured data, can be Web data or any typical data. Which also includes the data in batch jobs that run on a plenty of database servers.

The term BIG DATA is used to address the data sets that are too large and too complex than the traditional data that the traditional processing applications are not enough to handle or process this large and complex data. Traditional processing applications cannot process such large volume of data because of some issues. These issues include analysis, capture, search, sharing storage transfer, querying, and information privacy. These are some challenges which comes while processing such large volume data or complex data by traditional processing system. Big Data refers to the use of some analytics to extract value from the complex data or it refers to other methods like this to extract the value or to summarize the complex and large data. These analytical methods can also be used to summarize the web data. The main advantage of Big Data is its accuracy which leads to more confidence in decision making. [4] If decisions are better they can result in greater performance, greater efficiency, increase productivity and also the cost effective that is can help in the cost reduction also can help in the reduction of risks.

Big data includes very large volume data sets and very complex datasets with the sizes which cannot be processed by commonly used software tools to process and analyze the traditional data within a small limit of size and tolerable time period [1]. In concept of Big data size is a constantly increasing ranging from gigabytes to terabytes and now to petabytes. This requires a new set of techniques with the extended power to process the large volume data.

In an article published in IEEE Access Journal the writer of the article defined the big data definitions into three types: Attribute Definition, Comparative Definition and Architectural Definition. The author also showed a big-data map that shows the key features of big data. By analysis of data sets, we can find new techniques to find market trends also find ways to prevent serious diseases and so on [12].

2. LITERATURE SURVEY

Traditional applications require more resources to analyze, process and manage the big data. But sometimes these resources may not be available to the machines which may be due to the budget of inexpensive machines. In the area of information technology many organizations want a single cost effective solution to process the big data which may be available or not but if it is available than its cost must be much higher. A solution can be a single system having a number of processors and memories but it will be very expensive to build such a system. Another solution can be the clustering that is to build a system with high availability clusters. This type of clustering system looks like a single system and also it requires proper installation, management and administration services. This type of system can be expensive. Organizations were looking for a solution which can be cost effective. Another solution cloud computing comes into picture which is not so much expensive i.e. very economical. It also contains the required resources which are necessary to perform computations. This solution is based on the single instruction multiple data algorithm. In which a large volume of data is transformed. In this processing of each and every data items happen independently that is processing of each data item doesn't depend on another one and also processing of one data item doesn't have any impact on the other data item. [1] The Hadoop framework which is used widely now, for big data analytics provides such type of solution. Hadoop is an open source framework model which support cloud computing and also it supports distributed file system. Hadoop is based on the MapReduce model. MapReduce model was introduced by Google. Google uses this model to solve the large problems. This model includes the two step processing i.e. it performs the data analysis in two steps. First is Map step and second is Reduce. In the first step input is provided which is transformed into smaller element. The input to the first step are the data sets and its output is the transformation of these data sets into the smaller elements. The output of the first step i.e. Map task is the input to the second step i.e. Reduce task. In the Map task each and every set is processed independently in parallel. A number of map task can be run in parallel into a single cluster. In the Hadoop framework two classes and two methods are used. The first class reads the input and converts or transform them into a key/value pair for each record. And the second class transforms the key/value pair into the required results. Two methods used in the Hadoop frameworks are Map and Reduce method. The Map function takes the data sets as input and transforms each input record into a key/value pair. The output from a Map function is that key value pair. For each input record Map function only run for once. It may produce a key/value pair and also multiple key/value pairs. This output is sorted by the keys. Then the Reduce function is performed for each and every key/value pairs. For each key/value Reduce method is called. It is called one time for each and every key/value pair. Then this Reduce function produces the key/value pair in arbitrary number. Then these pairs are written to the output file. There is no need to sort the keys if they are unchanged to the keys produced by Map function, they must be in the same order if they are unchanged. This Hadoop framework consists of two main processes which are TaskTrackers and JobTrackers. TaskTracker tracks the map and reduce function or map and reduce task for each and every node in the cluster. As map function is run in parallel in many nodes in a cluster. It is the task of JobTracker to manage the parallel tasks processed by Task Tracker. The main task of the JobTracker is the job scheduling. It manages the task distribution which are submitted to TaskTracker. JobTracker may serve on priority basis or may serve on the First Come First Serve basis. It depends on its configuration. In a cluster there may be multiple TaskTrackers and a single JobTracker. This JobTracker controls and manages all the TaskTrackers in a cluster of nodes. If a TaskTracker gets down, then system continues to work but if a JobTracker gets down the system also downs. So, it is a single point of failure.

3. BIG DATA ANALYTICS

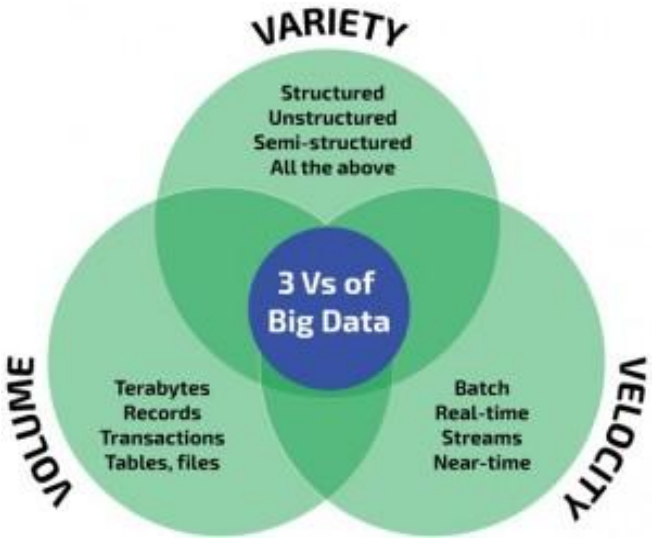
The "Big Data" which is most commonly used term nowadays refers to the datasets which grows in large volume and the term "Big Data Analytics" refers to the processing, analyzing, collecting, managing and processing the big data in order to extract information from the data. Big data analytics is also used for pattern recognition. It is used to extract the information from the data by analyzing the data. It helps organizations to better understand the information contained within the big data. Nowadays most of the organizations are using this approach in order to extract the information for their business perspective. Also these organizations know the benefit of using the big data. The main benefit of the big data is its accuracy. It is very accurate in computation which gives organization a better relief on the information they get. Also big data framework is an open source system which is easily available to the users. Another main benefit is it is inexpensive i.e. within budget of the organizations. Also to analyze the big data some software tools and proper skill sets are required. A variety of tools are now available which is described further. Also an important question about big data is how and where this big data is stored. As we all know big data includes structured data, unstructured data or it may be semi structured data or any other type of data. This all types of data is stored at a place which is called a Big Data Storage. Although big data storage is same as other storage devices. But it handles and stores a large amount of data and obviously Big Data Storage has more storing capacity than any other storage. These all concepts about big data like Storage and management, big data tools and big data analytics processing are described later. Now question arises why big data is becoming so much important and growing at a very high speed. Some key characteristics of big data are Volume, Variety and velocity which are also called 3 Vs of the big data. These characteristics are described below.

3.1 Characterstics of Big Data

Key characteristics of Big data which make big data so important and significant are Velocity, Volume and Variety which is also known as 3 Vs of big data.

By volume we mean the amount of data which is processed that is it used to process a large amount of data. Volume is the characteristic which conforms that we can process extremely large amount of data by using big data analytics. Organization like Facebook and Twitter are using this Big Data Analytics approach because the amount of data they have is very large in volume.

Another key characteristic of Big Data is Variety. By variety it means that by using Big Data Analytics approach we can process a wide variety of data which means that data can be in any form. From structured data to semi structured or unstructured it can be any form. Also data can be in text form also it can be in audio or video format. And this approach is capable of analyzing the data in any form. [5]
And the last V in 3 Vs of big data is Velocity which is its key characteristic. By velocity it means that data can be processed at a very high speed. In the hadoop framework data is processed around a large number of nodes in a cluster. And this data is processed at a very high speed.



These are some key characteristics of the Big Data. In short, Big data analytics process a verity of data in large volume at a very high speed. [7]

3.2 Big Data Analytics Tools and Methods

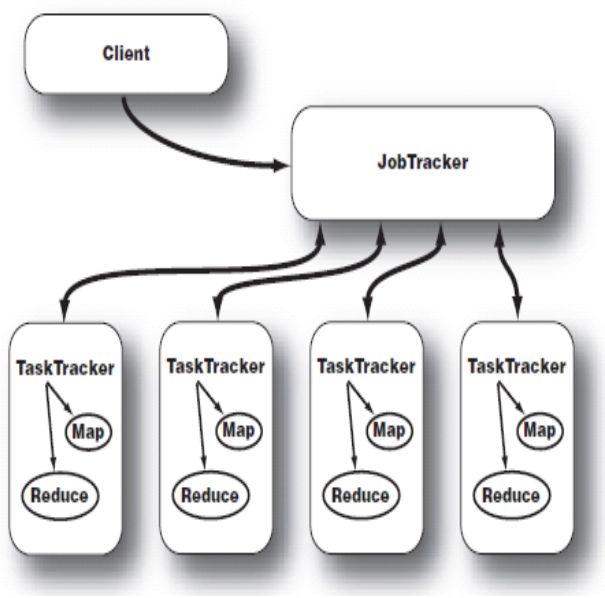
Organizations now agreed the fact that big data is very important from the business perspective and the have decided to use this approach as one of the key approach to success. With the amount of data available in an organization it is very difficult to take the decisions that is why companies are adopting this approach. As technology is increasing every day, the need to analyzing tools is becoming more and more. At present, a variety of tools are available in the market. Some of the tools are Hadoop, MongoDB and Cloudera. But the most important and most widely used tool for big data analytics is Hadoop. Hadoop is used synonymly with the big data. Hadoop is an open source framework for analyzing the large data sets on a distributed storage in a large cluster of nodes. By using Hadoop, we can store a large amount of data. It also processes this large amount of data at a very high speed. Hadoop process the computation or processing in parallel at each node in a cluster. Hadoop is based on the MapReduce model which consists of two phases of processing. First is Map phase and another is Reduce phase. In Map phase inputs in the form of datasets which can be structured and unstructured it performs sorting and filtering task and produces the output in the form of key/value pair. The output from Map phase is the input to Reduce phase. Then in the reduce phase summarization of data is performed. In this phase output produced is also in the form of key/value pair. Map function is performed at each and every data set from the input while Reduce function is performed only once for each key and its related data produced by the Map function. [10]

Another tool for big data is MangoDB. MangoDB is data storage and management system tool. This tool is a database management system. It is an alternative to Relational Database Management System. In this System only unstructured or semi-structured data is used. And it is used where data gets changed frequently.

There are a number of big data tools present in the market. Some of them are: Teradata, Qubole, BigML, Tableau, CartoDB etc. These all big data tool has their own significance and used in various organizations. [10]

3.3 Big Data Analytics Processing

Now I am going to discuss how big data analytics is performed on the big data by using Hadoop framework. As already discussed, Hadoop framework is an open source framework which is easily available to the users. It is based on Map Reduce programming model. Map Reduce model is a programming model which consists of two phase computation. MapReduce is the heart of the Hadoop which performs the big data analysis. Map reduce is based on two functions that are Map and Reduce functions. Its basic idea is to break down the task into smaller task and then process them in parallel. It is based on the strategy Divide and Conquer. As discussed it consist of two phase. In the first phase a Map function is performed. Map function takes data sets as input and perform some operations like filtering and sorting and produce the output in the form of key/value pairs. In the next phase which is known as Reduce phase, the summarization of data is performed on to the output key/value pairs produced in the first phase. For each and every key/value pair Reduce function is performed only once. The basic idea behind the MapReduce model is to add more nodes in a cluster rather than to increase the power of a single node and performing the task in parallel on all the nodes in a cluster. [6]



In a cluster of Mapreduce model there are two types of nodes available. The first node known as JobTracker and another known as TaskTracker. The main function of JobTracker is to manage all the jobs performed by TaskTrackers. The JobTracker nodes schedules the Map and Reduce functions on the TaskTracker nodes. They use some scheduling algorithm to perform this task. [11]

The main function of the TaskTracker is to actually perform the job. This TaskTracker performs the Map and Reduce function on the nodes scheduled by the JobTracker and then gives the result back to the JobTracker. [11]

To perform the Map Reduce function, at first client submits the job to the Hadoop. The Jobtracker of Hadoop then is notified that someone has fired the job. Then this JobTracker finds the free node available to run the job. When jobtracker finds a node or tasktracker which is ready to perform the job then job is assigned to that node. If no nodes are free then job is kept in a queue. When any one of the nodes is free then job is assigned. And at the node job is performed which is assigned by the JobTracker. After successful completion of the job the result is submitted to the JobTracker. Then at last client application is notified that job is done and results are ready to share.

4. BIG DATA CHALLENGES

As big data field is growing rapidly but it is not easy as it looks like. There are some challenges in analyzing the big data. Some of the challenges that are faced by the big data users are like: The first and main challenge is to understand the data available in an organization. It is not always possible to understand large volume of data for a use. Another challenge may be regarding security. Another problem faced by users is the quality of data. Good quality data means that it must be accurate and timely available. [12]

The main challenge in big data analytics is the availability of resources. These resources can be any software or any tool and it can be platform required and data availability and also can be human resources that is resources with proper skill sets are required in order to use this approach and get the results accurately and of good quality.

These types of challenge can impact the business badly. Their impact can be time delay of the product and quality of the product. A proper management and resources are required to face these challenges.

CONCLUSION

Big Data is very emerging field which is growing rapidly. It is just because of big data that organization can handle and process the data at higher speed more accurately than any other approach. It allows processing of unstructured data on different nodes in parallel which is big advantage of the Big Data Framework. It also has some challenges which are described in the above section. But it provides a reliable, cost effective and accurate solution to data analyzing problems. A large number of companies are using this approach to gain better quality analysis and more accurate results. In short, these companies would never exist without the advent of big data.

FUTURE SCOPE

Most of the organizations are dependent on the data available in their organizations. They process their data to extract the information and this data increase with the time. Here big data approach becomes so important. As we know big data is everywhere. But big data professionals are very few. For the big data professionals there are lot of opportunities. Demands for big data professionals are increasing day by day. Bigger companies are seeking for the big data professionals. There are a number of jobs in the field of big data. Also companies are providing big packages to the big data professionals. According to the research made, it is found that Big Data is everywhere and it is an important aspect for any organization. And this aspect is growing year by year. So, it is a very big opportunity who are seeking to make their carrier into the field of Big Data Analytics.

ACKNOWLEDGEMENT

I would like to thank my guide Ms. Kirti Bhatia for her indispensible ideas and continuous support, encouragement, advice and understanding me through my difficult times and keeping up my enthusiasm, encouraging me and for showing great interest in my thesis work, this work could not have finished without his valuable comments and inspiring guidance.

BIBLIOGRAPHY

[1]. Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012).s

[2]. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009).
[3]. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011).
[4]. Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013).
[5]. EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012).
[6]. He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208 (2011).
[7]. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A Self-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 261–272 (2011)
[8]. Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26–28 (2012)
[9]. Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011).
[10].http://www.import.io/post/all-the-best-big-data-tools/
[11].http://saphanatutorial.com/mapreduce
[12].http://www.sa.com/resources/asset/big-data-challenges-article