Big Data Analytics

Sandeep[1], Sachin Kumar Chauhan[2], Reema[3], Shabnam Sangwan[4]
[1,2] Mtech scholar, Department of CSE, Sat Kabir Institute of Technology & Management, Bahadurgarh, Haryana, India
sk1034167@gmail.com
sonuchauhan@gmail.com
[3,4]A.P., Department of CSE, Sat Kabir Institute of Technology & Management, Bahadurgarh, Haryana, India
Shabnam022e@email.com
arorareema@live.com

Abstract: The basic of this dissertation is the rise of "big data" and the use of analytics to storehouse the data. Big data is mainly used to analyze the large amount of data. The Big data store large amount of data and provide helpful information in an efficient manner leads a system toward a serious computational challenges, like to analyze, mixture, and store, where information are remotely collected. In the recent times, data warehousing, data repository and data mining are mostly used for Bid data. Big data warehouse known as terabytes which mean data collect in warehouse was terabytes in storage but in recent time it is petabytes, and the data built at a high speed. The progress in companies and organization is mainly because it store and analyze the data at greater levels and in greater details, as well as meta data , Web data and code-generated data, to build better relationship between customer and market behavior. Big data provide helpful information to achieve a goal. A lot of companies in recent time use big data to improve their quality. Growth of big data increase at higher speed in recent time, even trend going in coming year.

***Keywords:** big data, data ware house, data repository.*

*****

I. Introduction

This paper is in the space of data warehouse in the large industry to keep pace with the desire to collect and analyze large data or volumes of informational data and structured data. Data warehouse is a collection of large amount of information as well as supporting system and this information is used to build project or software. The vendor of RDBMS has provided a different platform and each platform have different specialty which provide higher level of price replication and it also provide higher performance as compared to general purpose Relational DBMS. These platforms provide lot of information is available in a variety of shapes and size, to the database. there is large number of survey is done by people which gives real time description .therefore ,time to time updating is done by the company. We know that, new technologies are coming to handle complex data. These technologies provide large amount of complex data which include Meta data, web data and server data. The Meta data mean data about data which mainly provide large amount of data. Meta data basically provide descriptive information about data. The Web data gives data about social media data like what Sapp data, messenger data, Facebook data etc. A large amount of data is uploaded to the Facebook server, the data in Facebook server is very large and these types of data are social media content. New technologies provide machine generated data or code generated data and these technologies have also advance feature like sensors and data like GPS. The kind of data is known as big data. Big data mean very large data.

These kinds of data have in large volume and these kinds of data are used to build or analyze new technologies. Now question arise what is new technologies and how we use these technologies in real world. Let discuss in brief. A large number of company, which provide open framework it mean any person can use these open framework and work on them. And these Data is change structured into unstructured data. These data was including in batch jobs that run on server machine. The term BIG DATA is used to address the data set which mainly provides large and complex data. Traditional application is not enough to handle large and complex data. These mainly used for the purpose of analyze, queues, capture, for the purpose sharing data, solve complex queries, search data for their application and protect information from unauthorized user. This analytical method can be used to describe the web data. The basic advantage of Big Data is accuracy which leads your data more confidential in decision making. If decisions are better they can provide better performance, more efficient, effective and increase productivity and reduce risk with low cost so it can be cost effective.

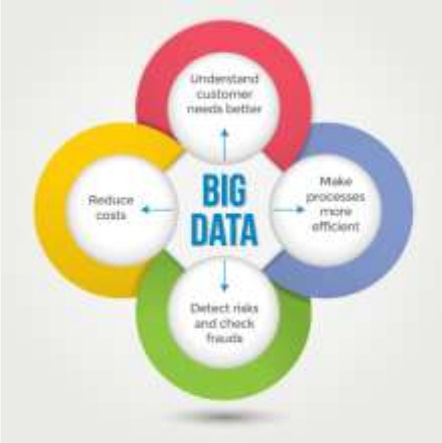
Big data analytics is basically a process which is used to check or examine bit of data to uncover hidden pattern. It well known about the customer preference and provide better result for their customer. Big data provide large data in small format. Data store in big data is in the range of gigabytes to terabytes and now to the petabytes. There is multiple data algorithm which provides information at each level in a specific manner and also in detail manner.

II. LITERATURE SURVEY

Traditional application requires more resources to check, process and handle big data. When application run or processes it require large amount of resources and take large time to run application. If it fail to gives output either break connection, break wire or some time traditional application can't support resources due less information available or due to technologies used in these application. Sometimes resources are available but machine can't support due to availability of machine which can't afford that kind of resources. These problem overcome by Big data, Big data store large amount of information, it support new technologies, take less resources and run application quickly. As name suggest, Big data is basically have large data which gives all information related to application, even same data run at different purpose. But it is available at higher cost, so built a new application with the help of big data then the manufacturing cost of application must be higher as compare to traditional application. A solution can be a single system having number of processors and memories. Another solution can be cloud computing, cloud computing used to store, manage, queue data on the server. Cloud computing have large amount of information available on the internet. It also contains large amount of resources which are necessary to perform computations. In these, multiple data algorithm is available which is mostly required in new technologies. Cloud computing load the data from host server rather than local server. It transforms large volume of data in a set which process data independently on different framework. There are different frameworks which are helpful in big data. The Hadoop technique is commonly used at that time. These techniques provide distributed storage in structure and unstructured manner which provide different level of distributed framework. It basically processing large amount of structure and unstructured data to gain business. Apache working on the Hadoop framework which is used to run server based program more efficiently. These is dynamic in nature it mean it run or processing large amount data parallel .different node are run at the same time. Therefore it required less time and run large amount of data in singe time. It support cloud computing framework. . It is basically graph based model that is used to build graph and then reduce nodes in the graph. It reduces complexity of Graph. The technique is Stream computing, which is used for data stream, including things on the web, aware about content on the web and it processing data in real-time and also analysis data in real-time. And Finally Navigation, this is providing real time information across the organization. It gives real time information that help organization to provide data in real time. It gives significantly better and faster decision. It runs parallel data into a single cluster. The first step is that it read the input and converts them into key pair for every record. And second it transforms value into required result. There are basically two method used Map and Reduce method. Map function is used to build a Map function which is used to provide different level of key node and second one is used to reduce node, that mean Reduce function is used to reduce complexity of the algorithm. It produces a single key pair as well as multiple key pair. The output is stored with the help of keys. Reduce function is performed for each and every key pairs. Reduce function is called for each key. It contain mainly two processes ResultSetMetaData and DatabaseMetaData, ResultSetMetaData provide method for obtaining information about the result contain in the result set. It gives data that in required to achieve goal. DatabaseMetaData provide method obtaining information about database. Big data provide DatabaseMetaData method which gives information about database; it gives method which is required to run application more efficiently. If DatabaseMetaData is down and can't work properly in that case ResultSetMetaData is useful to achieve goal.

III. BIG DATA ANALYTICS

The "BIG DATA" which is most commonly used term now days refers to set of data which grow in higher range. Big Data Analytics, as name suggest it analyzing, storing, collecting, processing, and managing the big data in order to extract information from the data to discover hidden patterns and other useful information. Big data analytics can help enterprises to better understand the information contained within the data and will also help search the data that are most important to the business and future business decisions. Analysts working with big data mainly want the knowledge that comes from analyzing the data. Big data means large volume. They constitute both structure and unstructured data that grow too quick that are not maintain and manage by traditional relational database system. Big data have large storage, big data tool. There are some key characteristics of big data like variety, volume and velocity which commonly known as V3. Of the big data.


Understand customer needs better
Reduce costs
BIG DATA
Make processes more efficient
Detect risks and check frauds

3.1 CHARACTERSTICS OF BIG DATA

Key characteristics of big data are volume, variety and velocity and commonly known as 3V of big data. Volume actually large volume, It basically storage area like in term of zeta bytes and terabytes. Velocity mean data growing very fast and quick in term of storage and processing, This is mainly Batch streaming and Real time streaming, Batch streaming is basically provide jobs in batch it mean these are run application in job e.g. data processing one after another. It means data break into small module and each module one after another. Another method data run each and every module at one time, it means each module run parallel throughout the application, it automatically stop run when result is come. Big data mean both structure and unstructured data. Variety handles structure, Semi structure and Unstructured. When these three combine it is difficult to handle by any system. Traditional Data have capacity of storing data in the range between gigabytes to terabytes but Big Data have capacity from petabytes to zettabytes. Traditional data store centralized where Big data store distributed data. Big data have structure and unstructured data; it means data comes from multiple formats, multiple resources, and multiple patterns. Last 2 year 90% of world's data created and 80% of world's data is unstructured. D-mart handles more than one million customer transaction every year. Whatsapp processes 1000TB per day. 80 hours of videos are uploaded to youtube every minute.

Variety
Velocity
BIG DATA
Volume

3.2 BIG DATA ANALYTICS TOOLS AND METHODS

Organizations now agreed that big data is very useful and very important from the business perspective and use this approach to find outcome. Traditional antique tool actually fail when it applied to large volume data which is used to analysis big data tool. Hadoop is basic way of big data platform in which we process large volume data cheaper and faster. Relatively fast on commodity platform. Different kind of tool is used in big data platform which are used to analytics big data. Hive is data housing tool, Hive used by yahoo and by Facebook. Hive provides SQL type of information we use sql for highly infrastructure. HBase is basically Hadoop database; it is large volume data store in database. ZooKeeper provide meta database, it mean it provide information data about data in detailed. Avro, large volume data structure ois civilized by Avro. Avro is basically used to civilized structure and unstructured data. RHadoop is another tool, R has basically mathematical, analytical and statically formula built in. R is basically programming language; R is basically popular tool in market. Sqoop is actually a tool that allows you to export data or further import data from traditional RDBMS tools. It transforms HDFS into RDMS. It is most important tool and used to aggregate tool from multiple algorithm at low cost. Flume is a type of flow machine to allow to process data to control data. These use for excel, flume casting is differ. Ooozie manage tab.

3.3 BIG DATA ANALYTICS PROCESSING

Now we discuss how big data analytics is processed on big data by using Hadoop framework. Map reduce is based on two function that are Map and Reduce function. Map reduces function based on Divide and Conquer. First it break down the task into small module and after that it run these entire module in a single task, It mean these run each and every task parallel. Once all modules run completely then it further compose each and every module into a single program. Divide and Conquer technique is based on this model. It first divides into small module and then combines all modules after each module run. It performs some operation which are filtering and sorting.

IV. BIG DATA CHALLENGES

As we know Big data growing rapidly but it is not so simple as it look like. There is lot of challenges in analyzing the big data. The first challenge is that it is not possible to understand that large volume of data. There is lot of tool which are used in big data and these is not possible to understand all these tool so understanding of big data is not simple and easy as it look like. The second challenge is security; Security is main perspective to every organization because every organization wants that it secure their data from unauthorized user. Third challenge is quality of data, every organization focus on quality because quality is main perspective from customer point of view. The main challenge is availability of resources. It is not possible to every resource at every time.

V. CONCLUSION

Big data is rapidly growing and it is more important for organization to build new application with the help of big data. Big data allow processing of unstructured data on different platform in parallel which is big advantage of big data. It is cost effective, reliable, accurate and secure as compare to traditional data. It provides better quality.

VI. FUTURE SCOPE

Requirement of big data increase day by day so demand for big data increase. For big data, there is lot of opportunities. There is large number of jobs in the field of big data. And these growing year by year. so there is lot of jobs opportunities in that field. According to data available, Big Data is everywhere and it important for every organization. Lot of people seeks to make carrier in the field of Big Data Analytics.

ACKNOWLEDGEMENT

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

BIBLIOGRAPHY

[1] Hilbert, Martin; López, Priscila (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information". Science. 332 (6025).
[2] Bakshi, K.: Consideration for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp.1-7(2012).s
[3] Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data". Release 2.0. Sebastopol CA: O'Reilly Media (11).
[4] "Big Data @ CSAIL". Bigdata.csail.mit.edu. 22 February 2013. Retrieved 2013-03-05.
[5] "The Government and big data: Use, problems and potential". Computerworld. 21 March 2012. Retrieved 12 September 2016.
[6] Survey on Big Data Using Data Mining" (PDF). International Journal of Engineering Development and Research. 2015. Retrieved 14 September 2016.
[7] "Degrees in Big Data: Fad or Fast Track to Career Success". Forbes. Retrieved 2016-02-21.
[8] Failure to Launch: From Big Data to Big Decisions, Forte Wares.
[9] Reips, Ulf-Dietrich; Matzat, Uwe (2014). "Mining "Big Data" using Big Data Services". International Journal of Internet Science. 1 (1): 1–8.
[10] "Big Data Definition". MIKE2.0. Retrieved 9 March 2013.