Evaluating Technology: Big Data

Sunny Kumar [#1],Vikash Kumar Garg[#2]
[#]*GZS PTU Campus, Bathinda, Punjab, India*
[1]Sunnykumar1018@gmail.com
[2]Vickyrocks.garg@gmail.com

Abstract— *The invention of new technology is changing the scenario of world day by day. Each and every sector is growing with a wide variety of changes at a very fast pace. This is a fact. Another fact that exists is that each sector wants perfection in their work with the evolution of new technology. This new technology leads to generating large amount of data to be stored and processed. This large amount of data so generated is called big data.This paper introduces the basic introduction of big data, the problems related to big data, some perspective of big data, big data platform, techniques to implement the concept of big data and inferences drawn.*

I. INTRODUCTION

In today's world we have seen from the past few decades that technology is spreading too vast and increases at a very high rate. The reason of this spreading up of technology is the invention and research that is growing so fast. This leads to the generation of large amount of data for keeping and processing. Now days data comes from everywhere like GPS, Smart phones, Sensor, Social networking sites. These all are generating large amount of data that is so difficult to store and process. The data produced from these systems is raw data, unstructured data, semi-structured data that is not stored in data warehouse. Therefore, it leads to a big problem of storing and processing of this large amount of raw data. This is called big data problem. From the analyser point of view to analyse the large amount of raw data it is also so difficult. So to analyse this information we need an efficient technique to implement on the big data.

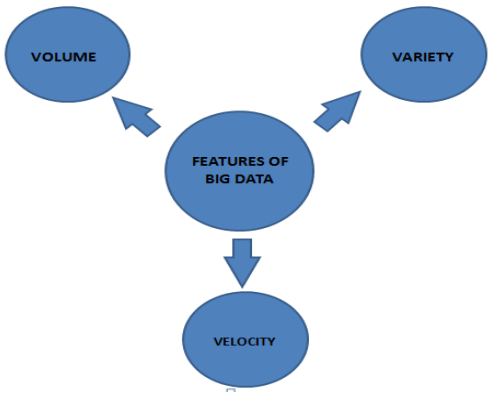
II. WHEN AND WHERE TO CONSIDER BIG DATA

Big data applies to that type of information that cannot be analysed or decoded using traditional system so invention to need of new systems. In current scenario the data that are available for any organization is very large. So the storage and processing ability to analyse that data is so difficult and to decode that massive data is also very limited. These organisations have access wealth of the information but do not know how to use that information, how to process that information and how to get the actual result from these information. This happening due to the random type of generating of data because the data that are generating are raw data unstructured or semi structured so that even they do not decide this data is worth keeping or not. Suppose we take a simple example of internet, the users are increases day by day and this leads to the generating of large amount of data. In 2010, 2.5 billions peoples that are use internet that are generating 15 Zeta bytes of data in our hands[1]. Now think up to 2020, how much this figure of generating data is raised. To analyses this large data and efficient storage and accessing mechanism is big challenge for everyone.

III. BIG DATA PERSPECTIVES

A. Not Ready For Volume:

Organisations are not ready to handle the bigness of data that they have in their hands. As technology and users produce massive amount of data. So divide the data and make a plan to handle all data. Divide across dimensions based on values and the usage frequency and size (Peta bytes, Gigabytes, Zeta bytes) and complexity. The internet of things performed various operations on data, the devices communicate with each other through wired and wireless and these connected devices produce large amount of data. The cost for collection, storage, and analysed data is a big matter in volume.



B. Velocity:

In today's scenario sheer volume and different varieties of data for storage and the collection mechanism is changed. So, if the term velocity is use to describe the concept of data it is the rate at which data is produced and needs to be handled. So it means how vastly or quickly data comes and how to maintain this type of data. Examples of high velocity data are click stream data that show the recorded users activities as the interact with web pages, GPS tracking mechanism.[2]

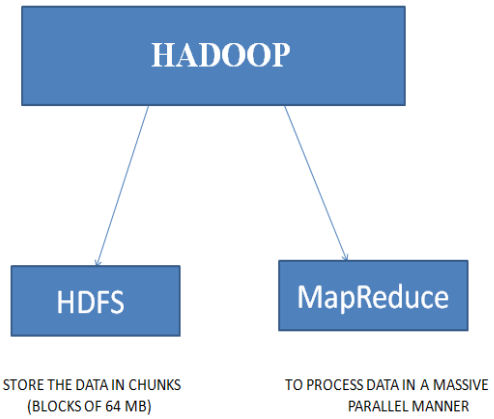
C. **Variety:**

As technology increases day by day so data produces by these technologies also comes into number of varieties like semi structured, unstructured (audio, video formats) and raw form. This is due to the different format of data that are generated from different technologies. So, this is the major challenging aspect that multitude of different formats comes and different structures and it needs to be analysed all of them.

IV. TOOLS TO IMPLEMENT BIG DATA SOLUTIONS

To deal with big data solutions first to consider about storing and processing of massive amount of data. Now the question arise that where to store this raw data? how to store and how to process in a timely manner. IBM provides some techniques that are used to implement big data. Now If consider the big data as from technology perspective then to think about Hadoop. So question arise in mind that what is hadoop? Is it a platform for big data or a solution to big data?

1. Hadoop: Hadoop is just a platform of big data. To handle large amout of data and perform large scale operations on data hadoop is used. The main features of hadoop are that it maintains the redundancy built in its environment so that if any failure occurs in any one machine into cluster it can handle easily. It is combination of Hdfs and Map Reduce systems.



2.HDFS*:* It is hybrid distributed database system. It is used to store the data efficiently but not concern about the processing and analysing of data. It is a file system used in hadoop that is based on google file system. In this the data is distributed among various nodes of the system and it maintains blocks of data size in nodes (64 MB) and analysed the data parallely. Simplicity, fault tolerance, cheap storage is features of HDFS.

3. Map Reduce: This is backbone of hadoop. It is for the processing of data and processing of large amount of data is done in a very small time and efficiently. Where map and reduce perform the different functions. In this firstly all data is mapped and convert it into another form where it can handle according to (key/value mechanism) after that the reduce is taken all data from map and reduces it and make a smaller set of data. This is main functioning perform by Map Reduce in hadoop. The goal of this programming model is that the scalability is for large data volume and 1000's of machine. Input data type is file to key value records and it used mapped and reduce functions.

4. Hbase: It uses column oriented database management system. The main focus of Hbase database is to put on column. Various attributes

are grouped together to make a column families. It makes the hadoop system to flexible so that at any time data come is easily added in the system. So that real time data generation problem easily removed in this database because the database schema is very flexible in Hbase. This is very powerful feature of Hbase database for real time systems where data are comes continuously. Hbase distributed database provide good functionality in terms of security, query language ,transactions, data layout max data size.

*5.*NO SQL: Not used only Sql language but there are more than Sql. Some new feature are added into it like map reduce framework, key value record, document block storage, graphs database. Columns are stored together and uses a key to value like in XML storage. As when data stored in XML in format of key to value access is fast and availability of data is high because it provide the replication of data. Like in acid it has base it means availably of data, consistency of data and flexibility soft state. As No Sql uses various data model like key value, column oriented, document oriented store, graph database in all these data model different factors like performance, scalability, complexity, flexibility does matter at a very big rate. These components give better results in all data model that is why No Sql uses it .

*6.*Stream computing: stream computing is one of the basic technique that is implement on big data but the problem is that is do not analyze and process data when the data is in rest state or when the data is already stored in a raw manner. It only apply to the data when all the data is in motion like that in stream computing it can easily identified that all people who are currently new in that area and you get continuous updated results because GPS system produces refreshed data in real time. This is used in stream computing

V. CONCLUSION

As data is produce at a large rate so handling is very necessary of that data. One way is that use existing data and analyzed it and produce new objects values from it. If new data comes store it efficiently likes indexing properly done. New technologies used for proper handling of that massive data.

As we all know now that big data is a big problem in all the sectors. So it is a evaluating technology which spreading its effect day by day and it has a far long effect if we not handle this in a proper way. So we all conclude that it's a very large problem that has to deal as soon as possible.

REFERENCES

BOOKS:

[1] Understanding Big Data
[2] Analytics for enterprise class Hadoop and streaming data by PAUL ZIKOPOULOS
[3] Resource library informatics PDF file study ONLINE
[4] http://bigdatauniversity.com/
[5] http://www.informatica.com
[6] http://hadoop.apache.org/
[7] http://WhiteHouse.gov/BigData.