A Relative Scrutiny on Big Data and Hadoop Paraphernalia and Techniques

Ruchika[1] , **Gurtej Singh Ubhi**[2], **Maneet Kaur**[3]
[1, 2, 3] Mtech Scholar, Lovely Professional University
Email - gurtejubhi91@gmail.com, ruchikathakur065@gmail.com

Abstract: *Today world wants are changed day to day with respect to their data storage, amount and their types of data. Today each and every firm has lot of data and they essential a tool for processing and storing that data in distributed manner. Big data use Hadoop and MongoDB for storing purpose and Mapreduce algorithm for processing that data. Business intelligence tools also help to sort and manage the data. HDFS (Hadoop Distributed File System) manage Hadoop architecture and their system to work appropriately.*

Key Words: *Mapreduce, Hadoop, Business Intelligence tools, MongoDB, Recommendation System, Hive*

1. OVERVIEW:

In today world data is increased day –by-data in term of GB (Giga bytes) to TB (Tara Bytes) and TB (Tara Bytes) to PB (Peta Bytes) to ZB (Zeta Bytes) (In today world it term as big data). Due to this increase in data it turns out to be difficult to handle that huge amount of data and to process this data. To handle these volumes of big data various tools and techniques are used. Various filtering, clustering techniques are used for analysis and analytical purpose. Hadoop eco-environment system components are used for processing the data.

Nowadays , every filed produce huge amount of big data for handling that amount of data apache Hadoop is used.

2. INTRODUCTION:

There is no proper description for the big data. Big data is typically used in many business because business are primary Concerns to their unstructured data and 80 percent [1] of the enterprise and business data are unstructured data that is difficult to handle ,manage and process. For handling that data we want big data and their various tools and techniques.

Big data: Some of author describe big data as a collection of huge amount of data in term of Tera bytes whose scalability, diversity and verity prerequisite new data processing structure, new processing techniques that solve and mine the various hidden pattern that increase the business assessment [2]. There are mainly four main features of big data that are explained as below as:

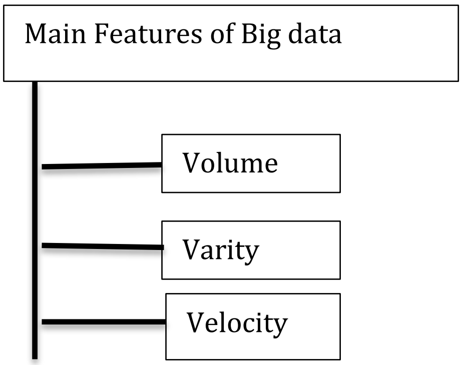

Figure 1: Features of Big data

- *Volume:* we can also state it as scale. Volume of data is increase day-to-day. It may increase from TB (Tera Bytes) to PB (Peta Bytes) and PB (Peta Bytes) to ZB (Zeta Bytes).
- *Varity:* Also refers as diversity or complexity of data. Big data handle dissimilar structure of data, different format of data from rational to raw text data.
- *Velocity:* it denotes to the speed for moving the streaming data and large volume data.
- *Veracity (optional)*: it handles uncertainty occur due to the data inconsistence, ambiguity and other problem that befall in the data.

Apache Hadoop: it is firstly developed and use for Google. It is open source software that handles verity of data with their virtual grid operating system architecture for storing huge amount of big data and help to process this homogenous and heterogeneous data [3]. It basically run the various algorithms for handling that distributed data processing, unstructured data and structured data processing by using Mapreduce.

Some of characteristics of Hadoop are:

- It process large amount of data in distributed way.
- It is open source software.
- It customs various analytical tools and algorithm for processing the large amount of data.
- Use HDFS
- Low cost and scalable
- Higher computing power and flexible with respect to storage.
- Automatic fault tolerance ability

There are mainly three component of Hadoop that are:

- HDFS (Hadoop Distributed File System)
- MAPREDUCE
- YARN()

HDFS: HDFS full form is (Hadoop Distributed File System) [4] that is use to handle huge amount of data and process that data in the form of clusters and apply several clustering algorithms for analysis purpose.it store the Meta data of the data.

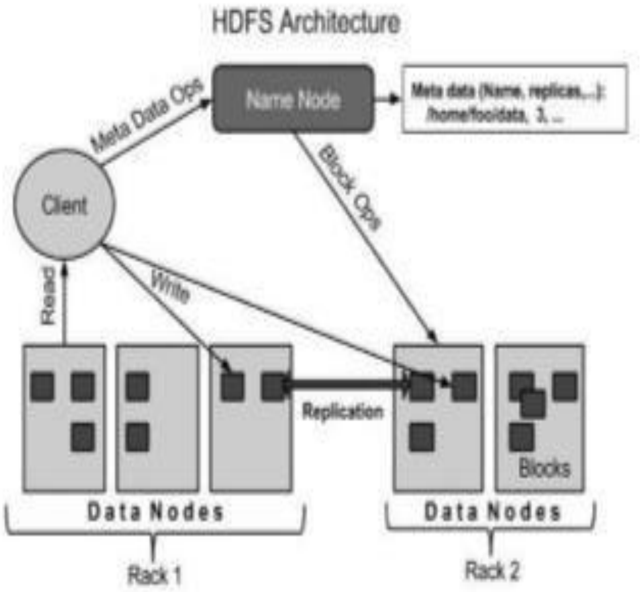

Figure 2. HDFS Architecture

3. SOME ISSUE RELATED TO BIG DATA:

There are various challenges and issues related to the handling and processing the big data because verity and volume of data is increase day to day. Due to large amount of data processing queries become more difficult to handle and rapidly increase in verity of data like images, text, link, numeric data, other type of data [5]. Processing speed get slow down due to this. Some important data may also damage due to this inflexibility. It is quite easy to handle homogeneous data but more difficult to handle heterogeneous data.

4. OPPORTUNITY RELATED TO BIG DATA:

There are lots of opportunity associated to big data that help any organization to handle their large amount of data, like in financial sector it store data related to finance, healthcare sector it store health related patient records, doctors detail and medicine ,medical equipment related details . In retail sector it is also used [5]. Web/social media/mobile companies also use it for storing their user detail and data like their likes, search pattern, calling and messaging records. Manufacturing and government sectors also use it.

5. MAPREDUCE PROCESS:

It is an algorithm that is used for handling the huge amount of data in distributed manner [6]. Many Mapreduce algorithms are designed by various companies like Google's Mapreduce, Apache's Hadoop Mapreduce [7]. It basically works in two phases. Those two phases are:

- *Map Phase:* In this phase it firstly divides the large amount of data into small- small parts. Then it finds the perfect match for those parts and generates intermediate result.
- *Reduce Phase:* in this phase it sorts the intermediate results and then coming it to form a result that is used by any firm.

Mapreduce process is explicated in following steps with the help of algorithm. That is explained as below as:

Step 1: Take bulky amount of data as Input.

Step 2: Breakdown the data in small function units.

Step 3: Map () use for mapping the related data to each other.
Map (Value, key)

Step 4: After mapping it generate intermediate key value.
Intermediate (Value, key)

Step 5: Reduce () sort the intermediate result and produce the output.
Reduce (Value, key) -------→ result ()

Step 6: End

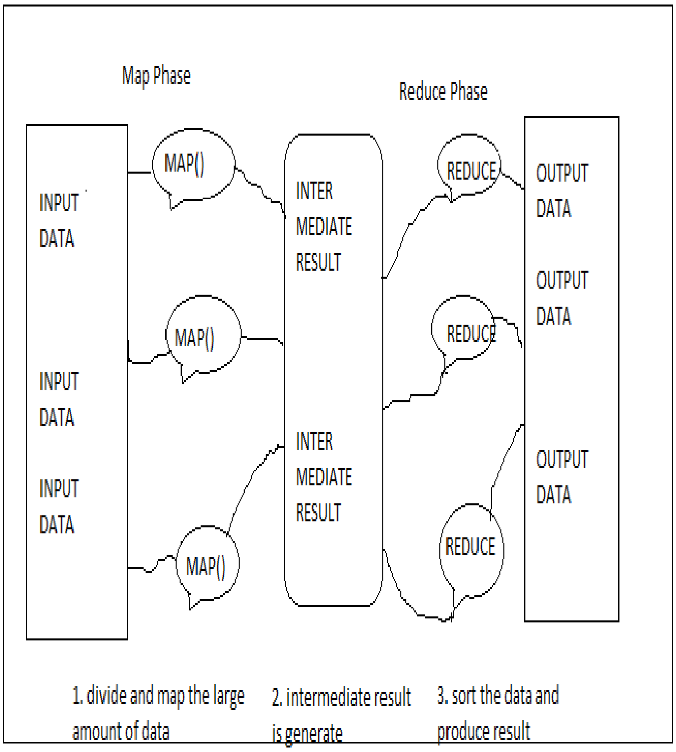

Figure 3: Mapreduce

6. LITERATURE REVIEW

6.1 *Big-Data Application: Study and archival of Mental Health Data, using MongoDB [8]*

Mental health basically measure by high rate of depression, disorder and type of disorder that the human is affected. In this paper genetic algorithm and big data tool MongoDB is used for analysis purpose. Genetic algorithm return optimized result by applying random operations on large amount of data.

MongoDB is used for both processing, searching and storing the data. It produces the result with in a lesser amount of time and cost which is necessary by health care. In paper various type of disorder are discussed.

In steps working of algorithm is explained here:-

Step 1: Selection of data is completed by creating subset of the data means categorized the data.

Step 2: Then check the result by reducing the random data and make relationship between that data.

Step 3: Then apply iterative process [9] for finding the better result from the set.

Step 4: Apply these command on MongoDB for insertion, updating and displaying the result.

6.2 *The use of Business Intelligence Tools for leadership and university administration [10]*

MS-SQL Server 2012 and intelligence tools are used for discovering, cleaning and processing the data and finding new fact from that data. These tools are used for finding the new facts from the university data and analysis can be perform on that data that how many students are studying which subjects, courses, areas and number of student choose which kind of course and way.

Step 1: Fetch and preprocessed the data

Step 2: After processing create tables and make analysis that how many students choose which course and areas, gender also checked.

Step 3: OLAP (Online Analysis processing) procedures are used for analysis that help in making decision.

Step 4: SQL Server used here for processing and accomplishment queries on that data.

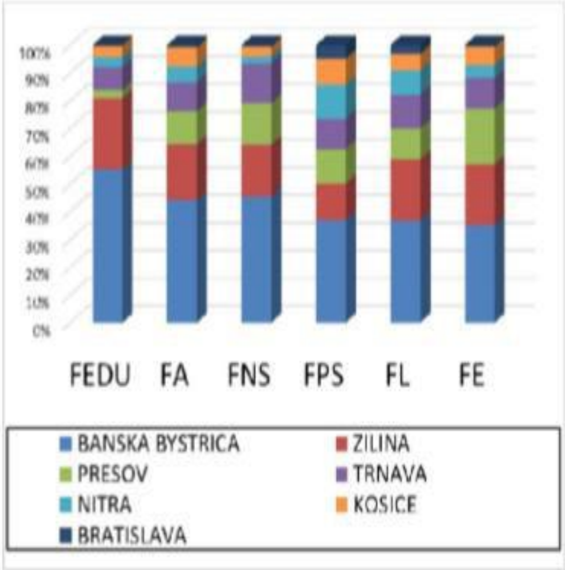

Figure 4: Ratio of student by region of their origin at MBU in their session 2013/2014

6.3 *Feedback Analysis using big data tools [11]*

Big data tools used for feedback analysis and sentimental analysis. Feedback data gather form various web sites. It is basically unstructured data that comes in various forms not in structure form. MongoDB and DynamoDB used for storing and processing that data.

It gather data from human to human, Human to machine, machine to machine etc. various algorithm and clustering algorithm are used.

Hadoop and Mapreduce algorithm used for processing the data. It process data and provide correct Feedback to their originator.

Figure 6: Social Media FeedBack Symboles

6.4 Big Data Analysis: Recommendation System with Hadoop framework [12]

Recommendation system also developed for helping the users for recommending the best option for choosing the best and most preferred sites, services and products. It gathers the data from various webs that data is available in various forms like in rating, reviews, feedbacks and likes, SMS etc.

Various filtering techniques are used for filtering the data like collaborative and content based etc. this analysis can be done on various social sites like Twitter, Amazon and other reputed sites.

7. CONCLUSION:

In this paper we show the use of big data and Hadoop and their tools. We also illustration the certain of current field are where the big data and business intelligence tools are used. These tools are used in Medical field for storing and processing patient data. Education Sector it is also used it. For feedback analysis and developing the recommend system it also used. Several other fields are there are not described here where we use big data and Hadoop.

REFERENCES:

1. Information system and management, ISM Book, 1st edition 2010, EMCZ, Willy Publishing
2. S. Case, S. Indonesia, and A. Uluwiyah, "Trusted Big Data for Official Statistics," 2016.
3. P. Deepika and A. R. G. R, "A Study of Hadoop-Related Tools and Techniques," 2015.
4. K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," 2010.
5. M. Dhavapriya, N. Yasodha, and I. Introduction, "Big Data Analytics : Challenges and Solutions Using Hadoop , Map Reduce and Big Table www.ijcstjournal.org," 2016.
6. S. Agarwal, "Map Reduce : A Survey Paper on Recent Expansion," 2015.
7. Z. Khanam and S. Agarwal, "M AP -R EDUCE I MPLEMENTATIONS : S URVEY A ND," 2015.
8. P. Dhaka and R. Johari, "Big Data Application : Study and Archival of Mental Health Data , using MongoDB," pp. 3228–3232, 2016.
9. L. Hao, S. Jiang, B. Si, and B. Bai, "Design of the Research Platform for Medical Information Analysis and Data Mining," pp. 1985–1989, 2016.
10. J. Gubalová, "The use of Business Intelligence Tools for leadership and university administration."
11. K. Yadav, "Feedback Analysis Using Big Data Tools," pp. 0–4, 2016.
12. J. P. Verma, "Big Data Analysis : Recommendation System with Hadoop Framework," 2015.