Big Data and Its Applications– A Review

Bhawana Gupta, Ashwini Kumar
School of Computing Science and Engineering
Galgotias University, Greater Noida
Uttar Pradesh, India
gbhawana9@gmail.com, ashwinipaul@gmail.com

Raghvendra Kumar Dwivedi
Department of Information Technology
KIET Group of Institutions, Ghaziabad
Uttar Pradesh, India
raghvendra.dwivedi@gmail.com

***Abstract*- Big Data is known as huge data set that is not processed in single system. To increase the processing speed and storing high volume data we prefer big data than traditional techniques. Big data helps in decision making for business prospective also. Big Data also refers a huge quantity of data sets with a large variety can be stored, managed and processed. It make secure correct information fast and provide benefit to the researchers and customers by taking the properties of volume, variety, veracity, velocity, value, visualization,** volatility, vulnerability, validity **and variability into consideration. With the help of traditional database system we cannot handle large data. To handle huge dataset, consider big data tools and techniques. In this review paper we first proposed the introduction and their characteristics and related issues faced by big data. It also provided the tools for big data which are help to managing data. This article is like a small seed of helping to know about the data storage and management of big data.**

Keywords- Big data, big data tools, Hadoop

I. INTRODUCTION

Over the last few times, the big data appears in society, application and organizations. Today's era is digital era. Over the last few years the data has increased around 1.8ZB. This amount will be increased double in few years. Big data, It is a large collection of data sets [4] [10]. Big data consider various kinds of datasets like unstructured, structured and semi structured. Unstructured data include audios, videos, images and text files. Structured data includes relational database like tables and semi structured data represents a log files and xml file [1]. To manage big Data by using traditional techniques are very complex. So here we discussed about tools and techniques to manage big data. There are some sources from which big data comes like facebook, twitter, blogs, sports, sensors, linked in etc. big data range start from terabytes to yotabytes. Big data technologies illustrate new technologies and framework planned to commercially mine data from huge volumes of data [2].

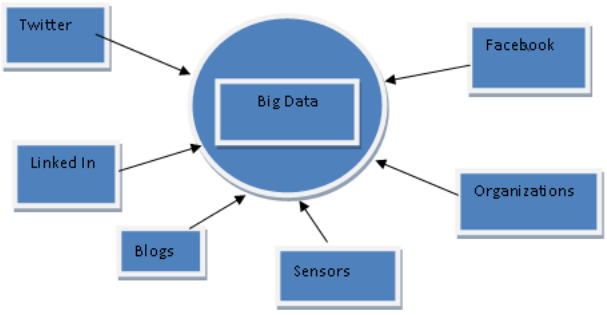

Fig 1- sources of big data

There are big data characteristics which are defined as volume, variety, veracity, velocity, value [12], visualization, variability, volatility, vulnerability and validity. Volume considers size of data in manner of terabytes, petabytes, exabytes, zetabytes and yotabytes. Approximately thousands of videos and audios uploaded in one minute to youtube. Every second on one click a new data is create. Variety considers unstructured data like text, images, and audio, video etc. semi structured data considers xml files and structured data such as tables [12]. Veracity refers confidant data or trusted data. It is similar to similar to, but not the same as, validity or volatility. Velocity in context of big data is the speed at which data is being created, generated, or refreshed. Normally more than 40, 000 search queries every second processed by Google alone. Value refers meaning. If big data does not take decision in context of business then it is meaningless. It considers the understanding and quality of big data. Visualization, look up technical problems owing to restrictions to adapt to increased demands, functionality, and reply time. It considers graph and chart etc. Variability refers inconsistencies in the dataset. Inconsistencies detected by anomaly and outlier detection methods. It refers inconsistent speed at which big data is loaded into database. Volatility refers accessibility. It ensures quick retrieval of information when needed. Vulnerability concerns about security. Validity is defined as how exact and right the data is [3].

ISBN: 978-1-5386-4119-4/18/$31.00 ©2018 IEEE

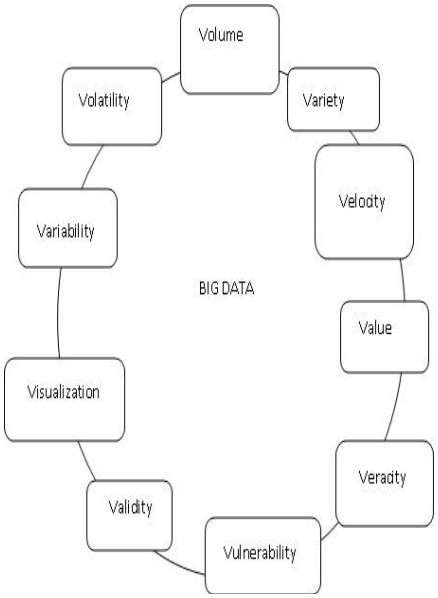

Fig-2 10 V's of big data

A. Sectors uses by Big Data

There are various applications of Big Data. Some of them are automotive engineering, Banking Sector,

Environment sector, Agriculture sector, Art and culture, Home affairs, Power and energy, Tourism sector[8], Commerce services, Information and communication services, Private and government organizations, Defense, Labor and employment Industry, urban and rural areas, statistics[9], Social development and food industry etc [7].

B. Big data's life cycle

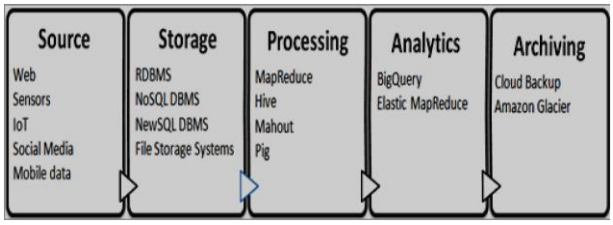

Fig-3 big data's life cycle

Big data gets from various sources like web, sensors, IOT, social media, and mobile data etc. when we find these large dataset we have to store it in different database systems as RDBMS, NoSql DBMS, NewSql DBMS, and File Storage Systems. Stored high volume data in storage system and then processed it by different processing techniques such as Map Reduce, hive, mahout, pig etc. big data analysis by Big Query and elastic Map Reduce and then archive with cloud backup and Amazon glacier.[5]

II. BIG DATA STORAGE AND MANAGEMENT

It is most important thing that institutions have to handle big data. We focused on where and how data will be stored. The conventional techniques of structured data storage and access include relational databases, data marts, and data warehouses. There is a tool Extract, Transform, Load (ETL). It is used for data stores. It extracts the data from external ways, convert the data, and upload data into the database. This type of data is error free and converted into desirable needs [12].

Difficult statistical techniques are used for present data analytics [5]. NotOnlySQL(NoSQL) was designed for storing and handling unstructured data. It provides the mechanism for scaling, flexibility, and easy application progress and usable. NoSQL databases split into data management and data storage [12].

In-memory databases handle the data in server memory. It is removing disk input/output (I/O) and allowing real-time reply from database. It improves the performance, and developed new applications.

Hadoop is a open source framework which provides reliability, scalability, and manageability. It performs big data analytics. It contains two components: First HDFS that is used for the big data storage, and second MapReduce that is used for big data analytics. It also provides the mechanism for data protection which ensures availability and reliability. HDFS nodes contain two types node i.e. the Data Nodes and the Name Nodes. Data is stored in duplicated file blocks and the Name Node acts as a control device between the client and Data Node, Data Node contains requested data [12].

III. BIG DATA TOOLS

Big data's characteristics create fresh problems of combination, classification, keeping data, and decrypted data. Big data tools are categorized into different categories as data gathering and keeping data, data retrieval and observing, and data managing, modeling, and analysis. There is essential tools named MapReduce, it is used for huge parallel data access and it is a programming model. Hadoop is free software based on MapReduce. It is designed for processing and storage large data. There are some useful profitable database sellers like Oracle, Microsoft and IBM. There is another free and open-source big data platforms consider MongoDB, Spark and Storm [11].

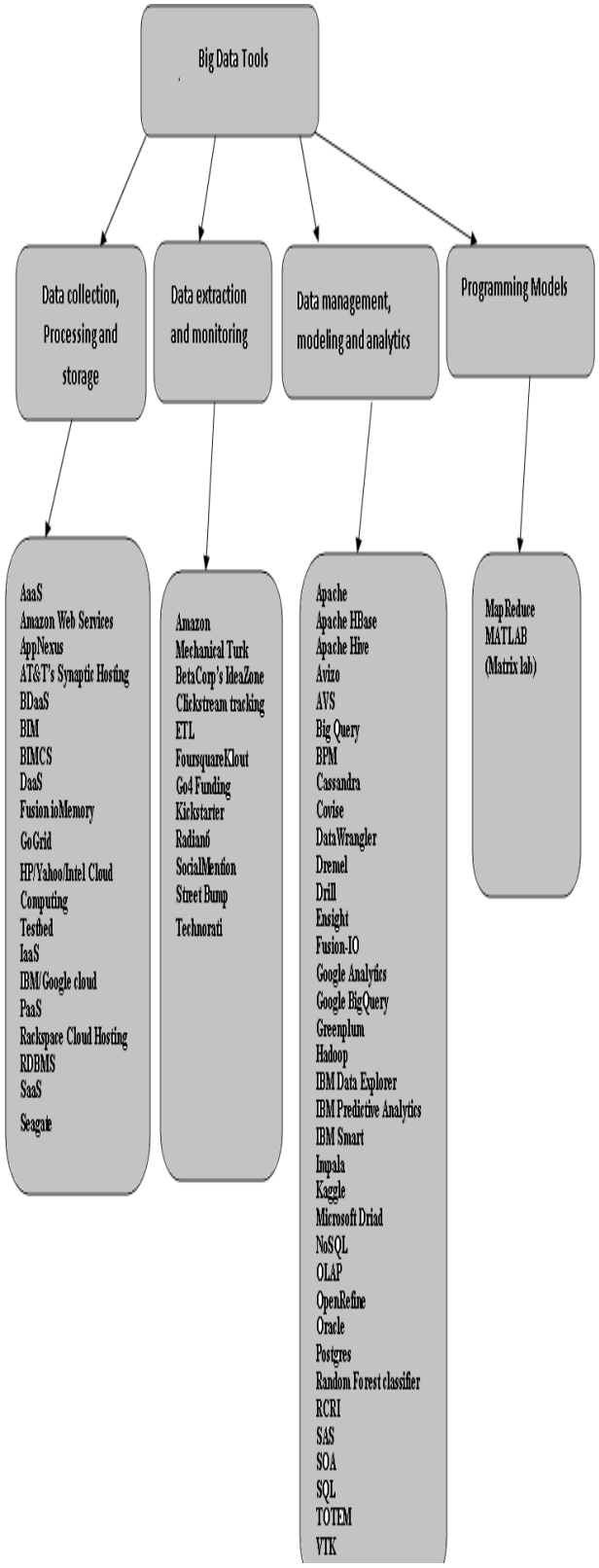

Fig -3 Big Data Tool's

IV. ISSUES OF BIG DATA

In digital era hurdles encounters by present big data. It is regarding to data gathering, processing and examination in manner of: fraud data collection, imperfect information and corrupt data, not typical, integrity and loyalty, and moral issues. Each issue may not be related to different kind of big data. Multiple issues may applicable in various manners [6].

A. Fake data collection

Conventional data gathering is generally collected by technical investigators, research institutions, or governmental organizations. These dataset is gathered by authorized scientific institutions. It has high data authenticity and reliability. These organizations have additional resources to perform those tasks. Still big data bear authenticity and reliability problems. For instance, social networking sites data are gathered from Twitter and other social media. These are profit-making sites which has profit purpose. These are not sure about data reliability or authenticity or integrity. These sites focused on profit only. There are three differences between a scientific research organization and profitable social networking sites.

First, social networking sites are collected data for profit but research institutions are collected data for research. Second, these social profitable big data suppliers can change the sampling methods and analysis techniques at any time. Scientists cannot identify these changes.

Third, profitable sites have no enthusiasm to assure the genuineness and credibility of the data sets they gathered [6].

B. Information incompleteness and noise of big data

There are stored bulky data but we cannot assure that the data has good quality. Several big data has good quality but it includes restricted information. This condition constrains the extra application of these data. For instance if we collect data from mobile phones or smart card or some other platform then it is not complete information. It is incomplete information and due to incompleteness and noise data we cannot solve the problems [6].

C. Representativeness problems of big data

Data suppliers usually do not accept a technical sampling method because the characteristics of those method and their profitable nature when data is collected. Almost the existing big data study that work on social networking platforms. That data has the .assumptions: collected data from these sites, such data have mainly bulky users. It means such type of data can represent the possible population who represent the data [6].

D. Consistency and reliability problems of big data

Without having consistent dataset we do not trust the information and due to these problems, big data analysis is possible. Big data, this is get from the various sources like social media, government organizations or other platforms may be inconsistent or fake(rumors) . So it is difficult to trust that information is consistent or reliable [6].

CONCLUSION

Big data has taken many positive changes in current era. With the use of big data tools and technologies we processed huge dataset which are not processed with traditional technologies in traditional database. There are some hurdles that may be effect big data in different ways. To avoid these issues we have to collect data in secure and authenticated manner. Big data is affected our business decisions and improve the performance of business. At present big data, is managed by Hadoop using MapReduce. With the help of different databases, Big data is managed. Big data tools processed data into many steps that increased the scalability, performance, speed of accessing data etc. In this review paper we discussed big data in brief, particularly characteristics, applications, issues and tools. Big data storage and management is also examined. Big data will increase day by day and it is managed by new tools and technologies. This review paper helps to researcher to find the issues of big data and in future researchers will resolve the big issues with the help of new ideas and techniques.

REFERENCES

[1] Chen, Shiwen Mao, Yunhao Liu "Big Data: A Survey Min", Mobile Netw Appl (2014) 19:171–209

[2] "Big Data and Data Mining A study of (Characteristics, Factory work, Security Threats and Solution for Big Data, Datamining...)", ConferencePaper, April2016, https://www.researchgate.net/publication/301513250

[3] https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

[4] "Trend on the Implementation of Analytical Techniques for Big Data in Construction Research" (2000–2014), ConferencePaper,May2016,https://www.researchgate.net/publication/303515244

[5] Manar Abourezq1, Abdellah Idrissi2, "Database-as-a-Service for Big Data: An Overview", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, 2016 157 | P a g e, www.ijacsa.thesai.org, Computer Science Laboratory (LRI) Computer Science Department, Faculty of Sciences Mohammed V University Rabat, Morocco

[6] Rethinking big data, "A review on the data quality and usage issues", Article in ISPRS Journal of Photogrammetric and Remote Sensing, December 2015, https://www.researchgate.net/publication/288918371

[7] Samiddha Mukherjee1, Ravi Shaw2, "Big Data – Concepts, Applications, Challenges and Future Scope " IJARCCE ISSN (Online) 2278-1021 ISSN (Print) 2319 5940 International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016 , Information Technology, Institute of Engineering & Management, Kolkata, India 1, 2

[8] Marko Grobelnik, "BIG DATA tutorial" marko.grobelnik@ijs.si Jozef Stefan Institute Ljubljana, Slovenia Stavanger, May 8th 2012

[9] https://data.gov.in/

[10] Anindita A Khad, "Performing Customer Behavior Analysis using Big Data Analytics", 7th International Conference on Communication, Computing and Virtualization 2016

[11] "An empirical study of the rise of big data in business scholarship",Article,June2016,https://www.researchgate.net/publication/295093727

[12] Big Data Analytics, "A Literature Review Paper", https://www.researchgate.net/publication/264555968