A Comparative Analysis of Big Data Analytical Tools

1st Dr. Jyotindra N. Dharwa
Associate Professor,
A. M.Patel Institute of Computer studies,
Ganpat University, Kherva, Gujarat, India.
jyotindra.dharwa@ganpatuniversity.ac.in

2nd Mr.Sanjaykumar B. Patel
Assistant Professor & Ph.D. Research Scholar,
A. M.Patel Institute of Computer studies,
Ganpat University, Kherva, Gujarat, India.
sbp05@ganpatuniversity.ac.in

***Abstract*— In recent years, the internet applications or communication medium are frequently producing very large volume of data, different types, variety and various digital data called big data. We are now in the period of a large amount of data generation. There are different ways of data generating. For Example, E-commerce, Social Media, Health care, etc. This research paper discuss and compare the data storing tool like Hbase, MongoDB and Sqoop. It also discuss and compare data analyzing tool like Hive and PIG. The objective of this research paper is to give insight about the various tools available in the big data analytic for the new researcher before performing data analysis.**

Keywords – Big Data Analytics tools

I. INTRODUCTION

Now a day, Different type of application system and people use different social network platforms like Facebook, Instagram and Twitter. Web portal and digital devices are generating large amounts of data insistent basis. Now a day anybody cannot deny the internet facility because it has improved the mode of a lifestyle of person, business advertises, the functioning of the government, online education learning, Health care etc. around the world. Making meaning full information analysis of data can help the organization and provide good economic benefits. Understanding this dataset is very useful as this is a critical and very important entity of an organization. The top-level management of big data makes data correctness or availability for new business intelligence. Its analysis also further helps in excellent decision-making capabilities and good economic benefits.

With the day by day increasing worldwide volume of data, the big data analysis process is commonly used in large datasets. Compared with a further conventional dataset and its process, big data types include structured, unstructured and semi-structured data required new real-time analysis using Hadoop system**.** This data set is using details about new things prospect to set new value. It will support us to expand a thorough understating of the unseen values and issues arise. It is challenging issues demanding to immediate resolutions. It is used to big data visualization process to generate a report for data analytics. [1]

II. BIG DATA AN OVERVIEW

Big Data is a term for collection of data sets so complex and large. Big Data is a collection of one or more datasets using processing in an open-source Hadoop system. It is impossible to process using traditional data processing application because it supports a small dataset or particular web application. The Traditional data set is to store the single application dataset who has generated the web application, blogs, etc. It is big challenge for data processing as compared to the traditional way. Big data challenges include analysis data, capture data, curate, search data, sharing data, storage dataset, transfer data, visualization of data, and information privacy. [2] Hadoop is the platform of choice for working with an extremely huge dataset. It stored data in a distributed architecture manner and not required specific schema. Big data is more useful for an organization to understand the demand to the person on a specific domain or product. It is a benefit for reducing the communication gap between both sides.

III. BIG DATA CHARACTERISTICS

This section explores the basic requirements for working with dataset is the same as the requirement for working with datasets of any size. It is different characteristics that can help to categorize as big data from other data. The figure below shows the characteristics of big data.

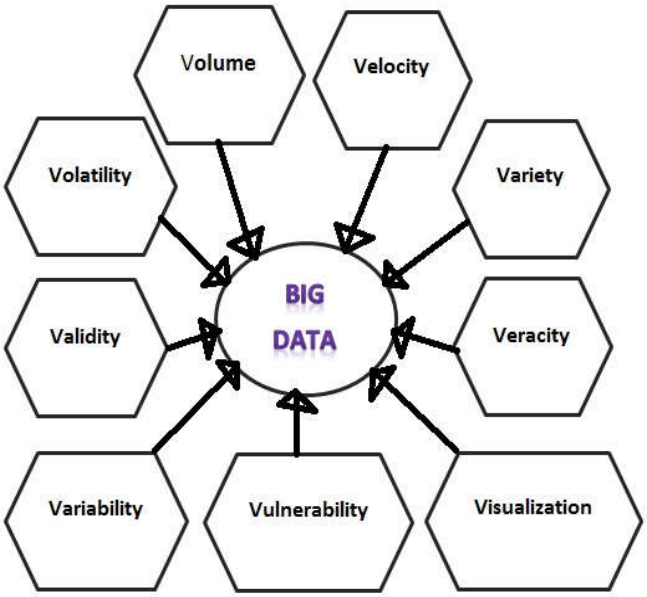

Fig. 1: Types of Big data characteristics

V**olume:** The amount of valuable data is massive with big data. The data can be generated in quantity and stored data can be in text, image, audio and video. Now a day data is generated each day in size of 2.3 trillion gigabytes. It is being created by such as point of sale in the online transaction and banking sector, GPS Sensor, Social Media, etc.

Velocity: The data can be arriving at fast speed and create a dataset in valuable period of time. This data can be processing in big data. [3]

Variety: The variety means different types of format of data. Variety in digital data types such as structured data in tabular format, unstructured data not in fixed size and stored image and video and semi-structured data in tabular, XML and JSON. It is data types generally require distinct processing skills and set of rules. It is challenging for terms of enterprises in the ETL System. ETL system is used for data integration, transformation, processing and storage in a Hadoop system. [3]

Variability: This produces data on a multitude of data dimensions. It produces a subsequent from various data types and sources. But now it is the viewpoint of producing a number of differences in the information. This data can be found by irregularity in order to any significant analytics to occur in big data. It is suggested to the unpredictable speed when generated data is stored in your database. [4]

Veracity: This characteristic is used to derivation or reliability of the dataset. For example – here we use the statistics data source to the analysis based on which item is more sold compared to other items. It will compare to items price in the last one year. You might ask the question: who created data source? Which approach did they collect the data source? Its data create summarize the information? Its information has been changed by anyone? Now given to the question are required for the evidence. Information on the data's veracity is around the help to improved understand the risk factor related to a new market trend exploration and business strategy based on data set.[4]

Validity: Validity discusses in how perfect and accurate the data it uses to do analysis. It might be used to a data set is useless information to cleansing their data set before used to analysis. Dataset is used for accurate analytics. This practices to ensure good data quality and reliable sources.

Vulnerability: This is used for new security concerns. Three types of vulnerability related to big data management and process were identified (1) privacy (2) security (3) absence of standards.[5] whenever the uses social media or web application to provide personal data for authenticating and after they have felt their behavior. It means your data may be selling them things for different commercial business to earn more money. [6]

Volatility: It means you have used dataset which is measured unrelated, ancient data or not useful longer period. Its volatility needs to establish rules for data availability and currently recovery of data when it is necessary. Make sure these are obviously tied to your business needs and processes with big data. The costs and complexity of storage and retrieval process is magnified. [4]

Visualization: It is a challenging task to view the data in big data. There are available different tools in big data visualizations to appearance some challenges because of store data of in-memory technology and poor scalability. It gives additional time for a response due to a huge dataset. [4]

Value: Value is useful information. In big data, it is a significant value can be found and including understanding view customer betters, improving processes and business performance.

IV. BIG DATA HADOOP TOOLS:

Now using Big Data Hadoop tool is Data Extracting, Storing, Cleaning, Mining, Visualizing, Analyzing and Integrating process in the dataset.

EXTRACTING TOOLS:

(1) Talend: This is a tool for data integrations that helps to effectively manage ETL Systems. This system refer the process in database usage in data storing on the warehouse. Data Extraction means where data is extracted from data sources. Data transformation means the record is transferred for storing in the prescribed format. Data Load means data is loaded into the destination database. This software is a free and open-source license. Its components and connector are Hadoop and NoSQL. [8]

(2) Pentaho: This software is used for data integration also called Kettle. It is a component. This tool is liable for ETL Processes. It is created with a graphical tool which works as a without writing code. PDI is an open-source tools which supports a massive array of input and output designs format including data sheets, commercial and text files. This is a complete visual big data integration tool.[8]

(3) Improvado: This tool is used in a data pipeline. This is used in marketing platforms and piping it into any destination data warehouse or visual tool.

(4) MarkLogic: This tool is a data warehousing solutions which makes data integration using an array.

(5) CloverETL: It is clean data using the ETL system. CloverETL software is generating quick development and professional skill available in the light footprint package. [8] It is also a leading data integration mechanization platform contains a designer, a visual development tool. Server side is an enterprise-grade data integration runtime platform, and cluster which enables parallel data processing on several data nodes. [9]

DATA STORING TOOL:

(1) Hbase: Hbase is a part of the Hadoop ecosystem. Apache Hbase is purely written in a java programming language. Hbase schema is a column-oriented database and stored on Hadoop Distributed files system. Hbase is supporting on a non-relational and distributed database. Hbase is good for storing data in the database and also data manipulations in an available dataset. Hbase is looking like a Big-table. Hbase is also supporting to analyze on HDFS. Hbase has java API for the client. It is a large table it offers fast lookups. Hbase does not support real-time transactions. It is structure and semi-structured data types supported [12] .The following figure shows the Hbase structure.

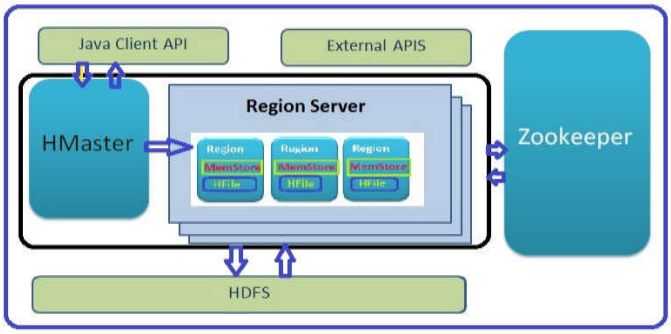

Fig. 2 Hbase Structure

(2) MongoDB: It is an open-source database. MongoDB is a document-oriented file system in data model. This is a store's data using a collection of flexible column's value in each database. It is a flexible document data model that is the same as JSON. MongoDB is the main feature of a file system with load balancing and data repetition over one or more machines for storing file data. This document contains one or many fields. It stores data in an array, binary data and sub-document format.

(3) Sqoop: Apache Sqoop is a tool in the Hadoop ecosystem. This tool is used to import and export data between Hadoop system to the relational database server. Relational database server likes Oracle RDB, SQLite, Teradata, MySQL, etc. Sqoop is SQL to Hadoop and

Hadoop to SQL data loading and stores. This tool benefits the import data that newly added records in a database table whenever it update the database. Sqoop is supported by automating the data process of using import and export facilities. [11]

TABLE I

Comparison of Big Data Storage Tools

Name of Tools	Mode of Software / OS	Description of Tool	Type of Data Model support	It is Concurrency	Secondary Indexes	API Support
Hbase	Apache Open Source / It support all Operating system with a Java VM.	It is Column– oriented database, It is also written a java language.	It is support Non-Relational DBMS, It creates a scheme.	Yes	No	Operation based and query-based APIs
MongoDB	Open Source / It is support Linux, Windows, Solaris Operating System	It is a document-oriented data stores. It is a C++ Language Implementation	It is Document oriented data store, It is schema-free.	Yes	Yes	CURD API
Sqoop	Apache Open Source/ support all Operating system with a Java VM.	It is support for bulk Import and export data between Hadoop and Relation database.	Relational DBMS, Structured Data	No	No	REST API and Client API

DATA ANALYZING TOOL:

(1) Hive: Apache Hive is open-source software. It is a data warehouse structure made on top of the Hadoop ecosystem system. It manage query, analysis and data summarizations. It looks like SQL to data stored in a different database. Hive is used as stored file system that integrates with the Hadoop system. HiveQL is automatically translating SQL-like queries into MapReduce. Mapreduce jobs executed on the Hadoop system. Hive is a different type of storage such as RCFile, ORC, HBase database, Plain text, and another file format. [10]

(2) PIG: Pig is a part of the Hadoop ecosystem. Apache pig is a high-level platform for developing a program that runs on top of the Hadoop system. Pig is supported data store, data processing and analysis. Pig is providing the facility to do all required data manipulations. It enables data workers to write complex data conversion without knowing Java language. It is familiar with the scripting language and SQL. User can develop his own function and also write a pig script, execute and run. [13]

TABLE II

Comparison of Big Data Analyzing Tools

Name of Tools	Hive	Pig
Mode of Software	Apache Open Source	Apache Open Source
Developed By	Facebook	Yahoo
It is support Schemas	Yes, but data can have many Schemas	Yes , It is optional
Language	It is SQL – Like Query Language	PIG – Latin, It is a Procedural data flow language (Scripting Language)
Partitions	Yes, It is store in small block.	No , It is Not possible
Create a User Defined Function	Yes, It is created on Java Language	Yes, It is created on Java Language
Streaming	Yes	Yes
Level of Abstraction	High Level	High Level
Line of Code by Programmer	Less line of code the MapReduce and Pig	Less line of code then the Mapreduce
Used by	It is used in server side	It is used in Researcher and Programmer
It is Support web Interface	Yes	No

DATA INTEGRATING TOOL:

Data Integration is the process of combing data from many different sources. They enable the application to access data related with other application and transfer data from one platform to another, transforming it is necessary.

(1) **Zookeeper:** It is a coordination service between Zookeeper and the data model. Zookeeper is also provided that server failure notification. Apache Zookeeper is a services used by cluster to coordinate between themselves and maintain shared data with robust synchronization techniques. This is a service provided by Zookeeper who is naming services, Configuration management and cluster management, locking and synchronization services, Reliable data registry. It provides client-

server architecture. The client is one node in cluster and access information from the server-side. The client sends a request to the server that the client is active. If the client is not connecting the server then redirect to another server. Server-side is a one-node that have collective and gives information to the client to inform that server is active. [14]

VISUALIZATION TOOLS:

Data Visualization is the exchange of information into visual contexts such as graphs or maps, to make data easier for the human brain to understand form. The main object is to make identify the patterns and trends in large data sets. There are many open-source visualization tools available as below. [15]

(1) R Tool: R Tool is open source software in windows and Linux operating system supports. R has a CLI Interface and Rstudio is GUI base tools. It is used for Machine Learning, data analytics, statistical computing, scientific research and graphics supported by R Foundations. It is used to analysis and visualization for a large amount of dataset. [16] This R Tool is used in few companies make statistical or calculated decisions. The names of companies are Twitter, Ford, Microsoft, Google and New york times. [17] This R Tool provides facilities like data storage, data manipulation, analysis and data visualization. [18]

(2) Tableau: This Tool is used to data visualization and Reporting. It supports query translation for online analytics processing cube, spreadsheets, relational database and cloud database to produce a graph. It is the entire data stored in-memory data engine [19]. This tool is data visual as dashboard, public, online, reader and server. It is also supporting real-time analysis processing of the data set. [20]

(3) Infogram: Infogram is visual representations of information. There are different ways to view information such as the charts, symbols, maps, text, and icons. Inforgram has provided the inbuilt visual templates. It is also support for shared information to the publisher for research, business work and education domain.**[** 21**]**

(4) Chart Blocks: This tool is free online available. We have no additional knowledge of complex coding. It builds report or visual context from database and spreadsheet.

V. CONCLUSION

In this research paper, we discussed about big data and its characteristics. Also we have done comparative analysis of extracting tools, data storing tools, data analyzing tools, data integration tools and visualization tools. So it is very helpful for the researcher who has started the research in the domain of big data and supposed to work in different big data analytical tools. It will give him the clarity about the different tools and help to select the best tool for his concern area.

VI. REFERENCES

[1] M. D. A. Praveena and B. Bharathi, "A survey paper on big data analytics," 2017 International Conference on Information Communication and Embedded Systems (ICICES),IEEE,Chennai, 2017, pp. 1-9.

[2] https://gennet.com/big-data/big-data-important/

[3] Paul Buhler, Wajid Khattak, Thomas Erl. "Big Data Fundamentals: Concepts, Drivers & Techniques", Prentice Hall,January 2016

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

[5] Fernando Almeida, "Big data: Concept, Potentialities and Vulnerabilities "Emerging Science Journal Vol.2 No.1 February, 2018. www.IJournalSE.org

[6] https://www.datasciencecentral.com/profiles/blogs/vulnerability-introducing-10th-v-of-big-data

[7] https://www.edupristine.com/blog/top-big-data-hadoop-tools

[8] https://www.guru99.com/etl-extract-load-process.html

[9] https://www.datamation.com/big-data/top-open-source-data-integration-tools.html

[10] https://data-flair.training/blogs/Hive/

[11] https://www.tutorialspoint.com/sqoop/sqoop_codegen.htm

[12] https://data-flair.training/blogs/Hbase/

[13] https://data-flair.training/blogs/hadoop-pig-tutorial/

[14] https://www.tutorialspoint.com/zookeeper/zookeeper_overview.htm

[15] https://searchbusinessanalytics.techtarget.com/definition/data-visualization

[16] https://en.wikipedia.org/wiki/R_(programming_language)

[17] https://www.datamentor.io/r-programming/

[18] **https://www.tutorialspoint.com/r/r_overview.htm**

[19] **https://en.wikipedia.org/wiki/Tableau_Software**

[20] https://www.guru99.com/what-is-tableau.html

[21] **https://infogram.com/blog/infogram-in-2017**