IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

A Study on Big Data Analytics Tools

Dr. R. Devi[1] and Dr. K. Sharmila[2]
Associate Professor, Department of Computer Science[1]
Associate Professor, Department of Computer Science[2]
Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India
devi.scs@velsuniv.ac.in[1] and sharmila.scs@velsuniv.ac.in[2]

Abstract: *In this digital era, the emerging trends and technologies such as Internet of things and cloud computing really deals with huge and huge volume of data [1]. [3]Big Data Analytics has been gaining much focus of attention lately as researchers from industry and academia are trying to effectively extract and employ all possible knowledge from the overwhelming amount of data generated and received. Study of these massive data requires many levels to extract knowledge for decision making. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques [2]. Such huge data has to be studied and handled to extract knowledge from these datasets. Hence Big Data analysis is a current research area that mainly explore the potential impact of big data challenges, various tools used for big data analysis. This article deals with big data at numerous levels and opens a new horizon for researchers.*

Keywords: Big Data, Data Mining, Analytics, Decision Making

I. INTRODUCTION

[9]As the information technology spreads fast, most of the data were born digital as well as exchanged on internet today. [1] **Big Data** is a collection of data that is huge in volume. It is a data with so large size and complexity. [2]In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. Collection of this large and complex datasets are difficult to process using traditional database management tools and techniques. Data warehouses have been used to manage the large dataset. [5]In this case extracting the precise knowledge from the available big data is a foremost issue. Most of the presented approaches in data mining are not usually able to handle the large datasets successfully. The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools such as data mining and statistical analysis. [1]These challenges generally arise when we wish to perform knowledge discovery and representation for its practical applications. A fundamental problem is how to quantitatively describe the essential characteristics of big data. There is a need for epistemological inferences in describing data revolution.

II. CHARACTERISTICS OF BIG DATA

The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent techniques.

Volume:[3] Big data consist of vast volume of data generated from many sources such as social media , network business processes and so on. Volume refers to the huge amount of data that are being generated every day. Volume of data can be categorized as megabyte, kilobyte, terabyte, petabyte etc.[4] Volume has not much problem when compare to other characteristics of V features. Every day each user creates enormous amount of data. The major problem is determined by decreasing storage rate whereas velocity [5].

Veracity: Veracity means how much the data is reliable. Veracity is the process of being able to handle and manage data efficiently for data analysis.[4] The quality of captured data, which vary so high. The Accurate analysis of data depends on the veracity of source data. It is very similar to validity. [3] It describes origin or consistency of the data sources, its circumstance and how significant it is to the analysis based on it.

Variety: Data is produced either by human beings or by machines. Received data is classified into various categorize. It can be structured data, or unstructured data. Structured data are such as image, text and videos. Unstructured data are

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

such as audio, hand writing text, ECG reading and emails. Various unstructured data causes definite problems for storage, mining and analysing data [4][5].

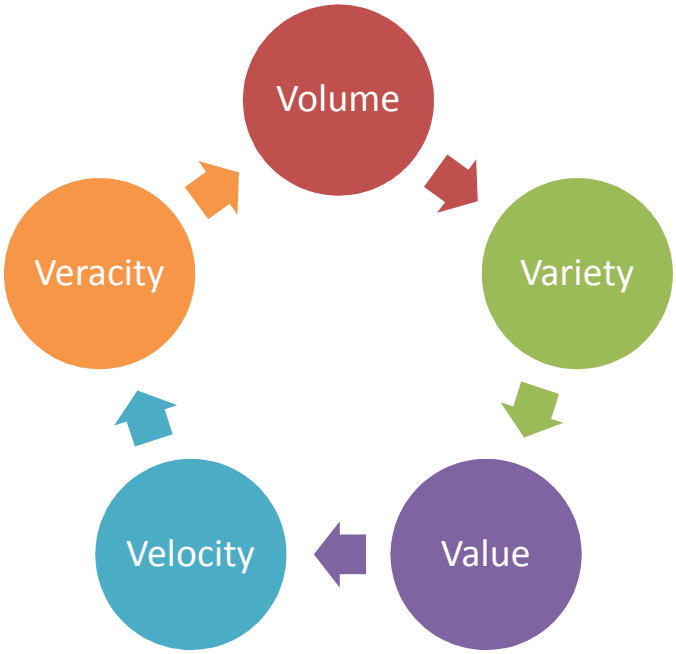

Figure 1: Characteristics of Big data Analytics

Velocity: It signifies the users how fast the data to be generated. Data velocity is fundamental task for some organizations. Many social medias are doing millions of photos uploaded and billions of searches on every day. It is a challenge for most organizations to reacting quickly enough to deal with data velocity. Velocity referred the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used. It streams into your enterprise in order to maximize its value.[4] Data is streaming in at extraordinary speed and must be dealt with in a timely manner. RFID sensors and smart metering are driving the need to deal with fast moving of data in near-real time.
Value: Value of the big data is used to understand the consumer better, aiming them consequently, enhancing processes and improving machine or well performances. It is changing a business to more competitiveness in world-wide stand. It suggests that big data bring big social value. There is pure connection between data and its visions.[5]

III. BIG DATA ANALYTICS TOOLS

Variety of tools are available for handling big data. Some of the tools discussed in this article are Apache Hadoop, HPCC, STORM, Cognos, MongoDB and Pentaho.

3.1 Hadoop

Hadoop is so popular because its ability to store and process huge amounts of any kind of data, quickly, its Computing power, its Fault tolerance, its flexibility in store as much data as you want and decide how to use it later[3]. That includes unstructured data like text, images and videos, its low-cost commodity hardware to store large quantities of data, its Scalability. Hadoop was started by Doug Cutting to support two of his other well-known projects, Lucene and Nutch.[5] Hadoop is Apache open-source software which runs on a cluster of commodity machines. Hadoop provides both distributed storage and distributed processing of very large data sets. [9][10] Hadoop is capable of processing big data of sizes ranging from Gigabytes to Peta bytes. Hadoop is a framework for performing big data analytics which provides reliability, scalability, and manageability by providing an implementation for the MapReduce program. Hadoop consists of two main components: the HDFS for the big data storage, and MapReduce for big data analytics. [10]The HDFS storage function provides a redundant and reliable distributed file system, which is optimized for large files, where a single file is split into blocks and distributed across.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

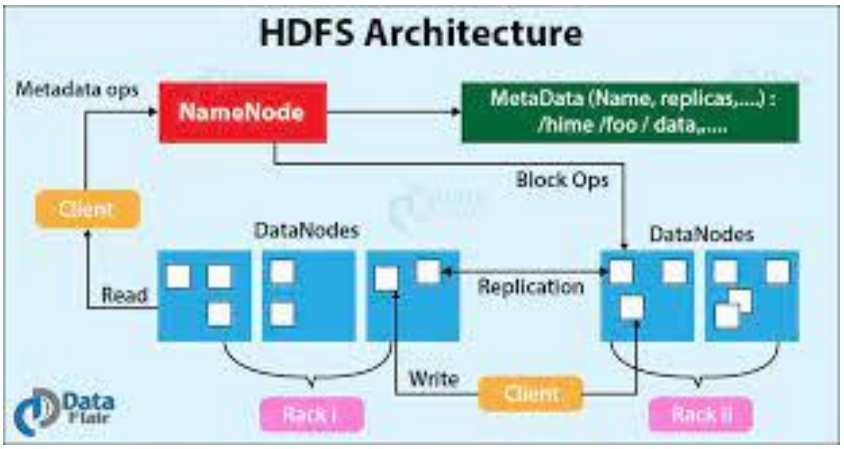

Figure 2: HDFS Architecture

Hadoop is a master/ slave architecture. The master being the name node and slaves are data nodes. The Name node controls the access to the data by clients. The data nodes manage the storage of data on the nodes that are running on. Hadoop splits the file into one or more blocks and these blocks are stored in the data nodes. Each data block is replicated to 3 different data nodes to provide high availability of the Hadoop system.[3][10].

3.2 Map-Reduce in Hadoop

Traditional System consists of centralized server to store and process data. The system is not suitable for huge volume of data. The centralized system creates too much of a bottleneck while processing multiple files simultaneously [5].The heart of Hadoop is MapReduce. It is this programming paradigm that allows for massive scalability across thousands of servers in a Hadoop cluster. Google solved this bottleneck issue using an algorithm called MapReduce. MapReduce divides a task into small parts and assigns them to many computers. Map Reduce is a parallel programming model, inspired by the "Map" and "Reduce" of functional languages, which is suitable for big data processing[10]. It is the core of Hadoop, and performs the data processing and analytics functions. The MapReduce function within Hadoop depends on two different nodes: the Job Tracker and the Task Tracker nodes. The Job Tracker nodes are the ones which are responsible for distributing the mapper and reducer functions to the available Task Trackers, as well as monitoring the results. The Map Reduce algorithm contains two important tasks, namely Map and Reduce[3]. The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs). The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples[9][10]. The reduce task is always performed after the map job.

3.3 HPCC

[10]The HPCC Systems platform consists of two integrated but distinct clusters: a back-end data refinery cluster for ingesting, refining, and transforming big data called Thor and a front-end data delivery cluster supporting high performance online querying of processed data called ROXIE. Both clusters run on commodity off-the-shelf hardware. A single, powerful programming language called Enterprise Control Language (ECL) creates the applications that run on the data refinery cluster as well as those that drive the data delivery cluster. In combination these components provide a comprehensive, massively scalable solution for big data processing and analytics.

3.4 Apache Mahout

Apache mahout aims to provide scalable and commercial machine learning techniques for large scale and intelligent data analysis applications[9][10]. Core algorithms of mahout including clustering, classification, pattern mining, regression, dimensionality reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce framework. The goal of mahout is to build a vibrant, responsive, diverse community to facilitate discussions on the project and potential use cases. The basic objective of Apache mahout is to provide a tool for elevating big challenges. The different companies those who have implemented scalable machine learning algorithms are Google, IBM, Amazon, Yahoo, Twitter, and Facebook[3].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

3.5 Apache Spark

Apache Spark Apache spark is an open-source big data processing framework built for speed processing, and sophisticate. It is easy to use and was originally developed in 2009 in UC Berkeleys, AMPLab. It was open sourced in 2010 as an Apache project. Spark lets you quickly write applications in java, scala, or python. In addition to map reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing. Spark runs on top of existing Hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster. The cluster manager allocates the resources and the worker nodes to do the data processing in the form of tasks. Each application will have a set of processes called executors that are responsible for executing the tasks. The major advantage is that it provides support for deploying spark applications in an existing hadoop clusters[3].

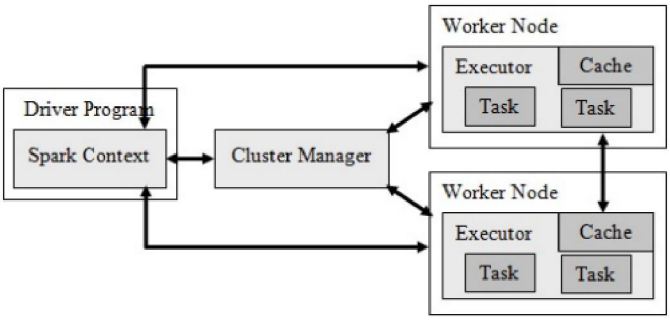

Figure 3: Apache Spark Architecture

3.6 STORM

Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with hadoop which is for batch processing. Additionally, it is also easy to set up and operate, scalable, fault-tolerant to provide competitive performances [9][10]. The storm cluster is apparently similar to hadoop cluster. On storm cluster users run different topologies for different storm tasks whereas hadoop platform implements map reduce jobs for corresponding applications.

3.7 IBM Cognos

It is the revolutionary new business intelligence release from IBM that breaks down the barriers to analytics. It is revolutionary because it expands traditional BI capabilities with planning, scenario modelling, real-time monitoring, and predictive analytics. These capabilities deliver power in an easy-to-use and unified experience that is collaboration and social networking enabled.IBM Cognos Real Time Monitoring (Cognos RTM), a component of IBM Cognos Enterprise, is software that provides visualization and analysis on real-time streaming analytics from Streams. Visualization is one of the major challenges that Big Data brings to business analysts; in fact, some universities today actually offer degrees in Big Data visualization [9][10].

3.8 MongoDB

The MongoDB database consists of a set of databases in which each database contains multiple collections shown in figure 4. Because MongoDB works with dynamic schemas, every collection can contain different types of objects. Every object – also called document – is represented as a JSON structure: a list of key-value pairs. The value can be of three types: a primitive value, an array of documents or again a list of key-value-pairs[10].

MongoDB supports two types of replications: master-slave and replica sets. In the master-slave replication, the master has full data access and writes every change to its slaves. The slaves can only be used to read data. MongoDB is in the forefront of NoSQL databases, providing agility and scalability to businesses. More than thousand companies and new start-up companies have acquired and are using MongoDB to develop new applications, refine client experience, fast track marketing time and minimize costs[10].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, February 2023
Impact Factor: 7.301

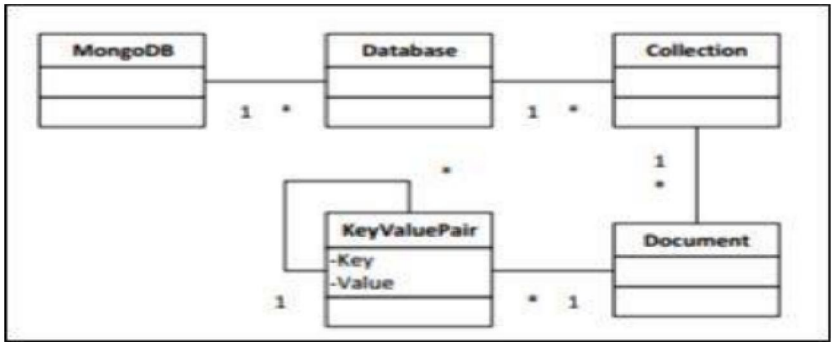

Figure 4 Mongo Db System

It's use of mostly bigger web applications like facebook, Amazon, Google etc. Due to the need, of huge amount of data storage for big data technologies we make use of one of the boosting invented technologies of NoSQL database i.e. MongoDB, Cassandra. MongoDB stores data in JSON structure, so get result in different-2 format it automatically.

3.9 Pentaho

[10] Pentaho provides a complete big data analytics solution that supports the entire big data analytics process. From big data aggregation, preparation, and integration, to interactive visualization, analysis, and prediction, Pentaho allows you to harvest the meaningful patterns. buried in big data stores.

IV. CONCLUSION

This paper gives a detailed survey on big data characteristics and tools used to analyse these big data. From this survey, it is understood that every big data platform has its individual focus. The paper also focuses on the importance of big data tools in real-world applications. In this research, the article has examined the innovative topic of big data, which has recently gained lots of interest due to its perceived unprecedented opportunities and benefits.

REFERENCES

[1]. What is Big Data? Introduction, Types, Characteristics, Examples (guru99.com)
[2]. Big Data Characteristics - JavaTpoint
[3]. D. P. Acharjy&Kauser Ahmed P,"A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016".
[4]. R.S. Karthiga, Senthil Kumar Janahan, U.V. Anbazhagu,"Research on Various Tools in Big Data"," International Journal of Innovative Technology and Exploring Engineering (IJITEE)",, Volume-8, Issue- 6S4, April 2019.
[5]. Mrs. Mereena Thomas," A Review paper on BIG Data",in "International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 09 | Dec-2015.
[6]. Nada Elgendy and Ahmed Elragal," Big Data Analytics: A Literature Review Paper", in in Lecture Notes in Computer Science ·21 September 2014.
[7]. K Sharmila, SA Vethamanickam," Survey on data mining algorithm and its application in healthcare sector using Hadoop platform" - International Journal of Emerging Technology and 2015
[8]. Lekha R. Nair&Sujala D. Shetty, Ph.D.," Research in Big Data and Analytics: An Overview",inInternational Journal of Computer Applications (0975 – 8887) Volume 108 – No 14, December 2014.
[9]. Tsai, CW., Lai, CF., Chao, HC. *et al.* Big data analytics: a survey. *Journal of Big Data* **2**, 21 (2015). https://doi.org/10.1186/s40537-015-0030-3
[10]. K.Yogeswara Rao 1 , S.Adinarayana," A Study on Tools of Big Data Analytics", in "International Journal of Innovative Research in Computer and Communication Engineering",Vol. 4, Issue 10, October 2016.