BIG DATA AND HADOOP TECHNOLOGY-A STUDY

S. Vanitha [1] and Dr .P. Balamurugan [2]

[1] PhD Research Scholar in the Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, India,

[2] Assistant Professor in the Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, India,

ABSTRACT- *This paper aimed at study and analyzing big data concepts and Hadoop technology. In this study designed about big data concepts, its challenges and categories, characteristics and its applications. This paper also presented how big data works and how to manage big data from the collected resources, and its tools used for implementation. This paper also consists of hadoop concepts and HDFS file system and its architecture for the various applications.*

Key Words: *Big data, Hadoop, HDFS.*

I.INTRODUCTION

Big data is a term also refers to data sets that are too large or complex for traditional data-processing application software to adequately deal with [3]. The term "big data" tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data. This paper is organized as follows. Section II Describes the Challenges in Big data, Section III Describes the Categories of Big Data, Section IV covers about the Characteristics of Big Data, Section V describes the Applications of Big data, Section VI covers about the Big data analytical tools, Section VII describes about Hadoop Technology and its Architecture, Section VIII deals with the Conclusion and future work.

II.CHALLENGES IN BIG DATA

The challenges in big data are,

- ✓ It is difficult to find ways to effectively to store data.
- ✓ Data must be valuable and accurate to store the collected data.
- ✓ It requires a lot of work to clean that data, relevant to client with meaningful.
- ✓ Data scientists have to spend more time for accurating and preparing the data for the future use.

III. CATEGORIES OF BIG DATA

Big data' could be found in three forms:

1. Structured
2. Unstructured
3. Semi-structured

Structured Data

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.

Examples Of Structured Data. An 'Employee' table in a database is an example of Structured Data

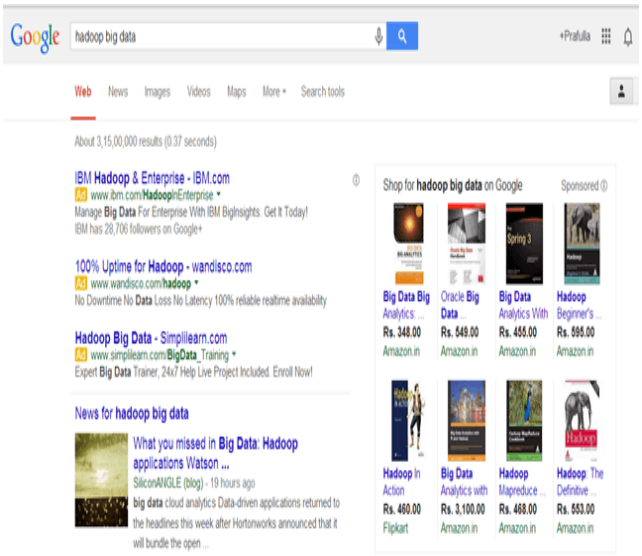
Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000

Unstructured Data

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. Typical example of unstructured data is, a heterogeneous data source containing a combination of simple text files, images, videos etc.

Example of Un-structured Data

Output returned by 'Google Search'



Semi-structured Data

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in XML file.

Examples of Semi-structured Data

Personal data stored in a XML file-

<rec><name>PrashantRao</name><sex>Male</sex><age>35</age></rec>

<rec><name>SeemaR.</name><sex>Female</sex><age>41</age></rec>

<rec><name>SatishMane</name><sex>Male</sex><age>29</age></rec>

<rec><name>SubratoRoy</name><sex>Male</sex><age>26</age></rec>

<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>

IV.CHARACTERISTICS OF BIG DATA

The characteristics of big data contain Volume, Variety, Velocity, Veracity, and Value.

Volume

Which needs to be considered while dealing with Big data. The volume of data is determined by its size [11, 12].Whether a particular data can be actually being considered as a Big data or not is dependent upon volume of data. This might be tens of terabytes of data [14].

Variety

Variety refers to heterogeneous sources and nature of data. The variety of data based on internal, external, behavioural, or/and social type. Data can be structured, semi structured, or unstructured type [11, 12].

Velocity

Velocity refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

Veracity

Veracity **refers** to the data quality of captured data can vary greatly, affecting the accurate analysis. [13]

Value

Value is the result of the data that are in the form of statistical, events, correlations, and hypothetical.

V.APPLICATIONS OF BIG DATA

Big data was applied in real time applications such as

- Big Data Applications in Healthcare.
- Big Data Applications in Manufacturing.
- Big Data Applications in Media & Entertainment.
- Big Data Applications in IoT.
- Big Data Applications in Government

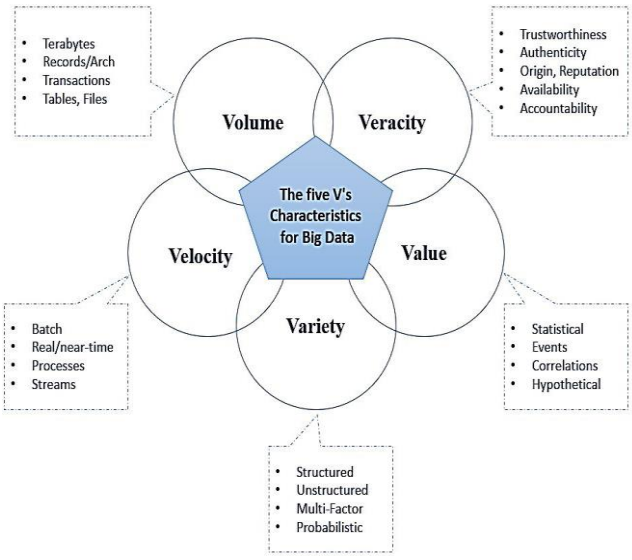


Fig 1:Five Vs Big Data Characteristics [15]

VI. BIG DATA ANALYTICAL TOOLS

1) Hadoop:

The Apache Hadoop software library is a big data framework. It allows distributed processing of large data sets across clusters of computers. It is designed to scale up from single servers to thousands of machines.

2) HPCC:

HPCC is a big data tool developed by LexisNexis Risk Solution. It delivers on a single platform, a single architecture and a single programming language for data processing.

3) Storm:

Storm is a free and open source big data computation system. It offers distributed real-time, fault-tolerant processing system. With real-time computation capabilities.

4) Qubole:

Qubole Data is Autonomous Big data management platform. It is self-managed, self-optimizing tool which allows the data team to focus on business outcomes.

5) Cassandra:

The Apache Cassandra database is widely used today to provide an effective management of large amounts of data.

6) Statwing:

Statwing is an easy-to-use statistical tool. It was built by and for big data analysts. Its modern interface chooses statistical tests automatically.

7) CouchDB:

CouchDB stores data in JSON documents that can be accessed web or query using JavaScript. It offers distributed scaling with fault-tolerant storage. It allows accessing data by defining the Couch Replication Protocol.

8) Pentaho:

Pentaho provides big data tools to extract prepare and blend data. It offers visualizations and analytics that change the way to run any business. This Big data tool allows turning big data into big insights.

9) Flink:

Apache **Flink** is an open-source stream processing Big data tool. It is distributed, high-performing, always-available, and accurate data streaming applications.

10) Cloudera:

Cloudera is the fastest, easiest and highly secure modern big data platform. It allows anyone to get any data across any environment within single, scalable platform.

11) Openrefine:

Open Refine is a powerful big data tool. It helps to work with messy data, cleaning it and transforming it from one format into another. It also allows extending it with web services and external data.

12) Rapidminer:

RapidMiner is an open source big data tool. It is used for data prep, machine learning, and model deployment. It offers a suite of products to build new data mining processes and setup predictive analysis.

13) DataCleaner:

DataCleaner is a data quality analysis application and a solution platform. It has strong data profiling engine. It is extensible and thereby adds data cleansing, transformations, matching, and merging.

14) Kaggle:

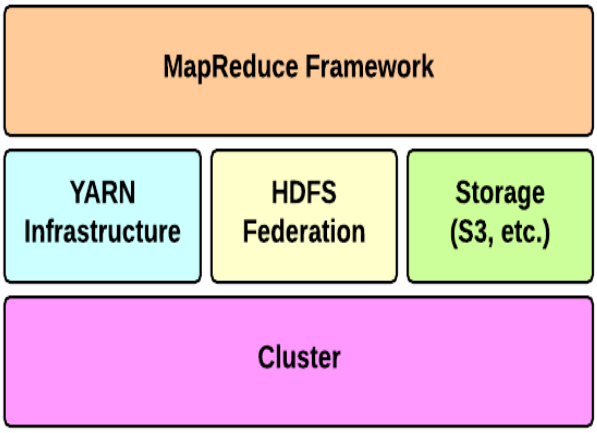
Kaggle is the world's largest big data community. It helps organizations and researchers to post their data & statistics. It is the best place to analyze data seamlessly.

15) Hive:

Hive is an open source-software big data too. It allows programmers analyze large data sets on Hadoop. It helps with querying and managing large datasets real fast.

VII. Hadoop Architecture Overview

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. Here are mainly five building blocks inside this runtime environment (from bottom to top):

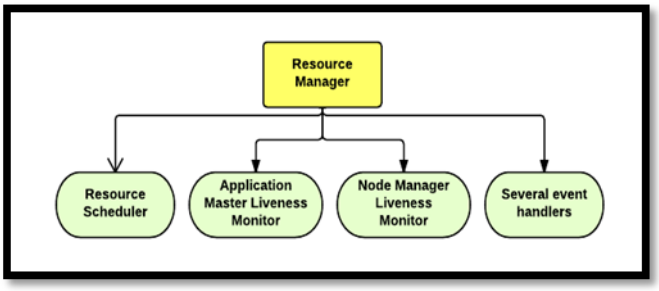


The **cluster** is the set of host machines (**nodes**). Nodes may be partitioned in **racks**. This is the hardware part of the infrastructure.

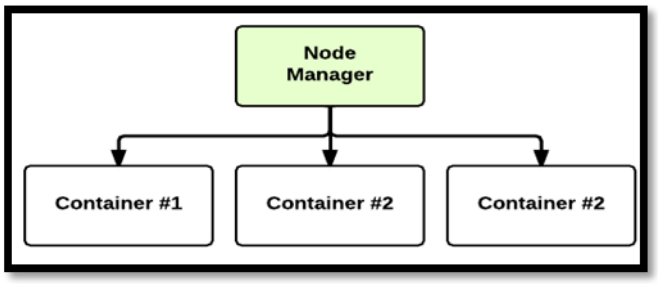
The **YARN Infrastructure** (Yet Another Resource Negotiator) is the framework responsible for providing the computational resources (e.g., CPUs, memory, etc.) needed for application executions.

Two important elements are:

The **Resource Manager** (one per cluster) is the master. It knows where the slaves are located (Rack Awareness) and how many resources they have. It runs several services; the most important is the **Resource Scheduler** which decides how to assign the resources.



- The **Node Manager** (many per cluster) is the slave of the infrastructure. When it starts, it announces himself to the Resource Manager. Periodically, it sends a heartbeat to the Resource Manager. Each Node Manager offers some resources to the cluster. Its resource capacity is the amount of memory and the number of cores. At run-time, the Resource Scheduler will decide how to use this capacity: a **Container** is a fraction of the NM capacity and it is used by the client for running a program.



The **HDFS Federation** is the framework responsible for providing permanent, reliable and distributed storage. This is typically used for storing inputs and output (but not intermediate ones).

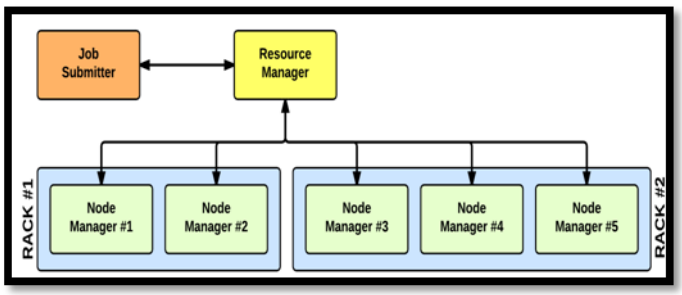
Other alternative storage solutions. For instance, Amazon uses the Simple Storage Service (S3).

The **Map Reduce Framework** is the software layer implementing the Map Reduce.

The YARN infrastructure and the HDFS federation are completely decoupled and independent: the first one provides resources for running an application while the second one provides storage.

The Map Reduce framework is only one of many possible frameworks which run on top of YARN (although currently is the only one implemented).

YARN: Application Startup



In YARN, there are at least three actors:

- The **Job Submitter** (the client)
- The **Resource Manager** (the master)
- The **Node Manager** (the slave)

The application startup process is the following:

1. A client submits an application to the Resource Manager
2. The Resource Manager allocates a container

The Resource Manager contacts the related Node Manager

1. The Node Manager launches the container
2. The Container executes the **Application Master**

VIII. CONCLUSION

Nowadays 80% of data captured today is unstructured which is being collected from various sources. All this data is also big data. In this paper, discussed the two important concepts that is basics ideas about big data and hadoop technology. The major challenge of this paper should be surveyed all the basic concepts of big data and hadoop technology. This work opens up the new interesting work that big data applied in real time applications with hadoop technology. Future contributions will be concentrating on developing an intelligent system using hadoop in medical health care application.

REFERENCES

[1]. https://www.guru99.com/what-is-big-data.html

[2]. https://data-flair.training/blogs/big-data-applications-various-domains/

[3] Breur, Tom (July 2016). "Statistical Power Analysis and the contemporary "crisis" in social sciences". Journal of Marketing Analytics. **4** (2–3): 61–65. doi:10.1057/s41270-016-0001-3. ISSN 2050-3318.

[4] Laney, Doug (2001). "3D data management: Controlling data volume, velocity and variety". META Group Research Note. **6** (70).

[5] Goes, Paulo B. (2014). "Design science research in top information systems journals". MIS Quarterly: Management Information Systems. **38** (1)

[6] Marr, Bernard (6 March 2014). "Big Data: The 5 Vs Everyone Must Know"

[7] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. **7**: 1–5.

[8] Dedić, N.; Stanier, C. (2017). "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery". **285**. Berlin ; Heidelberg: Springer International Publishing. ISSN 1865-1356. OCLC 909580101

[9] Fox, Charles (2018-03-25). Data Science for Transport. Springer.

[10] **https://www.oracle.com/bigdata/guide/wha is-big-data.html**

[11]. Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review". Martinhilbert.net. Retrieved 7 October 2015.

[12]. DT&SC 7-3: What is Big Data?. YouTube. 12 August 2015.

[13] Big Data's Fourth V

[14] Yiu, C., *The big data opportunity*, 2012, Policy Exchange

[15] International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-2, Issue-1, Jan.-2015 Big Data And Five V's Characteristics 1.Hiba Jasim Hadi, 2.Ammar Hameed Shnain, 3.Sarah Hadishaheed, 4.Azizahbt Haji Ahmad 1Ministry of Education, Islamic University College, Third Author AffiliationE-mail: nassirfarhan@yahoo.com, [s802371, s802370, s93456]@student.uum.edu.my

[16] https://www.edureka.co/blog/big-data-applications-revolutionizing-various-domains/

[17] POSTnote 468 July 2014 Big Data: An Overview

[18] *Judge, Peter (2012-10-22). "Doug Cutting: Big Data Is No Bubble". silicon.co.uk. Retrieved 2018-03-11.*

[19] https://readwrite.com/2013/05/23/hadoop-what-it-is-and-how-it-works/

[20] https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

[21] http://hadooptutorial.info/hadoop-evolution/

[22] https://www.quora.com/Why-is-Hadoop-important

[23] http://ercoppa.github.io/HadoopInternals/HadoopArchitectureOverview.html