ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

A Study on Big Data Computing

Vrind Gupta [1], Aman Chandok [2]

B.Tech Scholar, Department of Information Technology, Seth Jai Parkash Mukand Lal Institutre of Engineering & Technology, Radaur, Yamunanagar, Haryana, India [1]

B.Tech Scholar, Department of Information Technology, Seth Jai Parkash Mukand Lal Institutre of Engineering & Technology, Radaur, Yamunanagar, Haryana, India [2]

ABSTRACT: Big Data is not limited to the data in bits and bytes or its processing simultaneously but it is a full flesh combination of collecting, storing, processing and analyzing ample quantities of data currency that is heterogeneous in structure, size and values in order to produce valuable insights that are useful for businesses and other industries. In this paper many of the Big Data Dimensions and Analytics tools such as Apache Hadoop, Apache Spark, Hadoop Yarrn, Hadoop MapReduce and HDFS file system are defined with their utilities. Despite Apache Spark prove to be able to analyze the data with minor latency, it continues to be much developed next-gen software for Big Data processing been released.

KEYWORDS: Big Data; Apache Hadoop; Apache Spark; Hadoop Yarn; Hadoop MapReduce; HDFS.

I. INTRODUCTION

Smart phones with megapixel cameras, handheld computers, wireless sensor networks, omnipresent social media, earth orbiting/planetary orbiting satellites, space bound telescopes etc. Generate huge amount of data each and every while and this data need to be analysed timely. The nature of these types of data is termed as Big Data. We are submerged with the inrush of this type of data in our daily lives. Data, being accelerated at exceptional scale with the shoot-up of the ranges of different application areas is collected and need to be analysed every now and then. A huge amount of both structured and unstructured data is enable to be handled by traditional database system. This data is known as Big Data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

Big Data technologies are important to provide accurate analysis for better decision making in the businesses.

3 V's of Big Data

1. **Volume of Data:** It refers to the ever increasing amount of data.
2. **Variety of Data:** It refers different type of data and its sources like structured or unstructured data.
3. **Velocity of Data:** It refers to the pace of data processing.

Apart from these veracity(diversity among data sources) and value(different insights) can also be taken as the characteristic features of Big Data.

II. HADOOP

Doug Cutting charged to develop an open source version of MapReduce system to control the challenges to the exponential growth of data to cutting edge businesses such as Google, Yahoo etc. This system was called Hadoop which today is a core part of computing infrastructure for many companies. Hadoop is used for writing and running distributed applications with key distinctions:

- Accessible- Hadoop runs on huge clusters of commodity machines or on cloud infrastructure.
- Robust- Hadoop is made skilfully to handle hardware malfunctions and failure.
- Scalable- Hadoop can linearly ladders to handle larger data by adding more nodes to the cluster.
- Simple- Hadoop allows users to write efficient programs.

Hadoop Building Blocks

Hadoop employs a master slave architecture for both storage and computation.

- **Name Node-** It sits at the top of the for Hadoop File storage system that takes into account processing of each slave machine(DataNode) to perform various tasks. It acts as the accountant for HDFS that as it keeps the knowledge of the whereabout of file blocks as well as how the files are broken down into file blocks. Each cluster has only one NameNode.
- **Data Node-** Each slave computer machine in the cluster is a DataNode daemon that performs the desired work of distributed file system allotted to it by the NameNode that is reading and writing the HDFS blocks. A cluster can have many DataNodes.

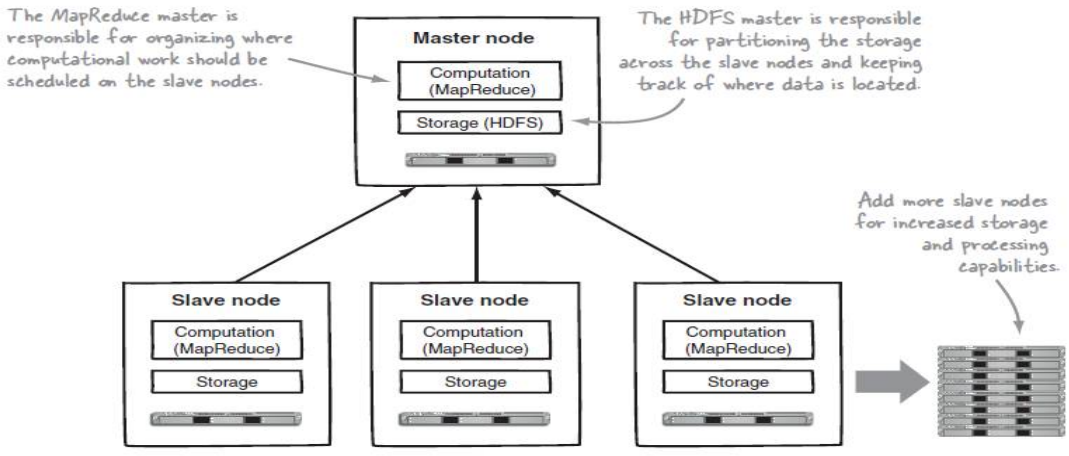


Figure 1.2 High-level Hadoop architecture

- **Secondary Name Node-** It is the assistant daemon for detecting the state of the cluster HDFS. As like the NameNode each cluster has only one secondary NameNode. It comes into play on the single point failure of Hadoop that is when the NameNode shut down. Thus, it helps to minimize the down time and loss of data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

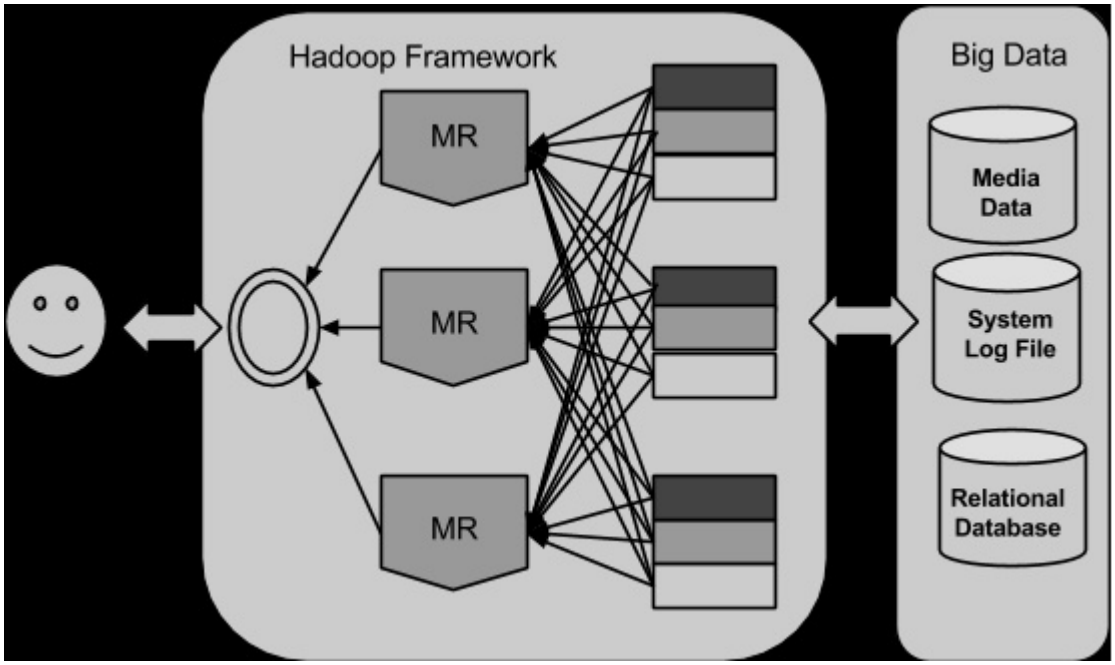
Vol. 5, Issue 10, October 2017

- **Job Tracker-** The Job Tracker Daemon is the co-operator between the computer application and Hadoop. Once the code is submit it follows the execution plan by determining the type of file to be processed; assigning the nodes to different tasks and monitoring the running tasks. There is only one Job Tracker per hadoop cluster.
- **Task Tracker-** task tracker follows the master/slave architecture that is it manages the execution of individual tasks on each slave node. The number of task trackers is equal to the number of slave nodes.

III. HADOOP FRAMEWORK

Hadoop Famework modules:

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



I. **Hadoop YARN:** It was not present in the initial version of Hadoop but due to the limitations of MapReduce developers included another open source community called YARN(Yet Another Resource Nagotiator). It was proposed in Hadoop 2.0 version. It eliminates the scalability limitations of 1st generation MapReduce paradigm. It enhanced the Hadoop by including scheduler and application manager to allocate the resources and accepting the job submissions. With YARN many applications can share single source at any point of time.

II. **Hadoop Distributed File System:** HDFS file system that is used at the backend of the Hadoop system. It performs according to the Master(NameNode)/Slave(DataNode) architecture. It divides the larger file input into 64mb blocks and stores it onto various machines or DataNodes.

III. **Hadoop MapReduce:** It is a parallel programming model which is used for writing distributed applications. It is the core of the Hadoop technology which was developed by Google in 1995. The essence of MapReduce is to divide large tasks into smaller chunks and then program them accordingly. It consists of two functions:

- **Map:** This function takes the input data and divides it into smaller sub problems and distributes them among the workers or slave nodes. There may be one or more Mapper functions. Each mapping operation is independent to the other.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

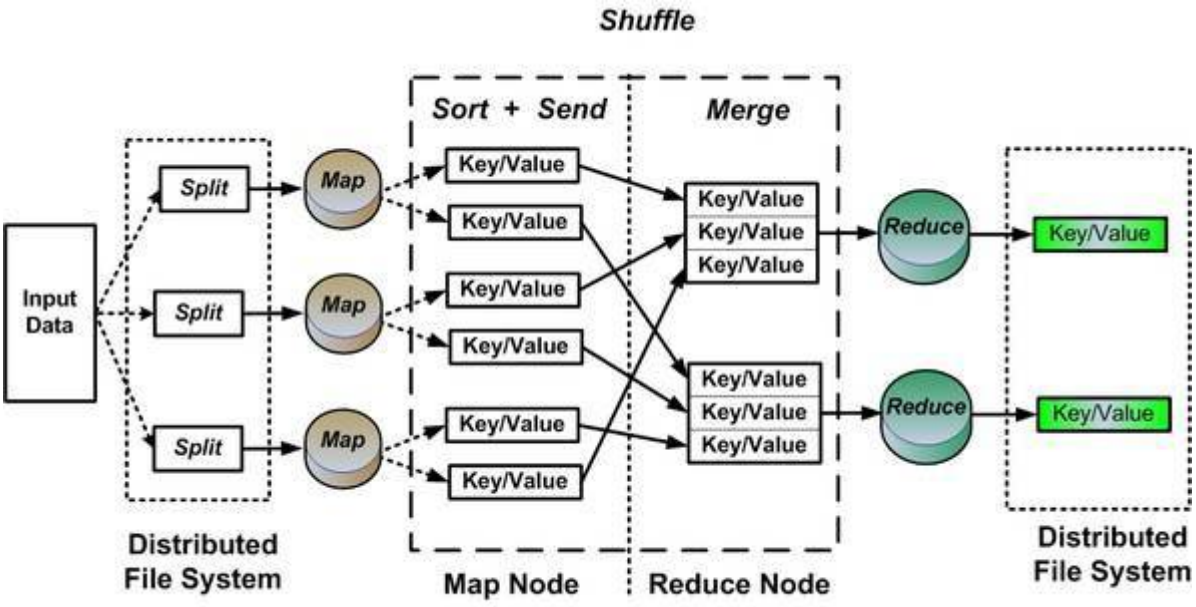
International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

- **Reduce:** It collects the processed data from mapper functions and sort them according to the keys. It then processes the data and aggregates the results to create the desired output. Reducers can also perform phase in parallel.



IV. APACHE SPARK

Apache Spark is an open source cluster computing framework which was developed at University of California Berkeley to provide a fast and general framework for large scale data processing much faster when compared with its predecessor Hadoop. Spark is almost 100 times faster than the Hadoop MapReduce in memory processing and ten folds on disk. Spark excels invincibly at streaming workloads interactive queries and machine based learning.

Spark can run both in distributed as well as standalone modes. Some scientists believe that Spark is about to diverge and substitute Hadoop for faster access to processed Data.

Apache Spark is a diffused and dispersed and highly scalable system with the upper hand to make programs using various programming languages like Java, Python, R, Scala etc. Spark does not have its own file system but it can use HDFS.

Apache Spark was donated to Apache Software foundation in 2013.

When not to use Spark

1. **Low tolerance to latency requirements:** If Big Data analysis is required on Big Data Streams and latency is the most crucial point than using Apache storm may results into better outcomes as expected.
2. **Shortage of memory resources**: Apache Spark maintains all its operations inside the memory. So it requires a huge amount of memory hence it must not be used when the memory requirements are more and available is less. In this case Apache Hadoop MapReduce can act as the best substitute.

V. CONCLUSION

This paper tells about the applications of the various Big Data analytic tools which introduced during the introduction of Hadoop and Spark..Both are specified to be as useful soft wares to provide the deep insights for the huge volumes of data in the real time along with their limitations on certain key aspects to provide the best of them at each ladders of time. A short comparison is also made to provide a clear outlook for using the software for various

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

purposes. As an outcome we see that spark is very well efficient in the streaming with a minor latency time. Thus analyzing data using spark with a couple of improvements should be persuaded as the important area of improvement in future.

REFERENCES

1. Meenakshi Jaiswal and Rubal Jeet,'A Narrative Study On Big Data' , International Journal of Emerging Trends & Technology in computer Science, Vol 5, Issue 4, pp. 6812-6815, 2017.
2. V.Sri Lakshmi,V.Lakshmi Chetna and T.P. Ann Thabitha,'A Study On Big Data Technologies', International Journal of Emerging Trends & Technology in computer Science, Vol 4, Issue 6, pp. 11350-11355, 2016.
3. Chucklam,'Hadoop In Action(Manning Publications)', pp. 21-25; 38-44, 2011.
4. V Srinivas Jonnalagadda at al, " A Review Study of Apache Spark in Big Data Processing" International Journal of Computer Science Trendsand Technology (I JCST) – Volume 4 Issue 3, May - Jun 2016.
5. Laney D, "3D Data Management: Controlling data Volume. Velocity & Variety", 2001.
6. Laney D, "Importance of Big Data: A Definition", 2012.
7. Lohr S, "The Origin of Big Data: An Etymological Detective Story", New York Times, 2012.
8. Big Data Analytics definition, Retrieved from http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics.
9. Big Data Analytics for Healthcare retrieved from https://www.siam.org/meetins/sdm13/sun.pdf.