

Big Data-A Review Study with Comparitive Analysis of Hadoop

Himani Tyagi¹

¹Department of CSE, BSAITM, Faridabad, Hariyana

Abstract - Data generated in daily life by human being is very large, this complete data is called big data. The data in this form is made useful for people by preprocessing it. To process this terabytes of data specialized hardwares and softwares are needed. And this data is going to get increased day by day. Therefore, big data analysis is a challenging and current area of research and development. The basic objective of this paper is to explore the potential impact, challenges, architectures, and various tools associated with it. As a result, this article provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues. A comparative study of hadoop, spark is also shown.

Keywords- spark, hadoop, big data analysis, challenge.

1. INTRODUCTION

Big data is a wide term that covers the non traditional strategies and technologies used to process, organize and gather insights from large datasets . It is not shocking or new that to work with the data that is not in the capability of a single computer system was a tedious task. With the introduction of big data with hadoop, gave a lot of ease and flexibility to store this 'big data'.

The data which follows following criteria is considered as big data,

1. **Velocity**:- it is defined as the speed of generation of data for example 500TB data per day can be considered as big data.
2. **Volume**:- this is defined as the size of the data for example 500TB data .
3. **Variety** :- it re presents the variety of data. there are three categories of data that is structured, semi-structured and unstructured. mostly the data is unstructured.
4. **Veracity** :-it can be defined as the truthfulness of data.

The most important feature of hadoop which makes it different from spark is hadoop works on batch processing and spark works on stream processing.

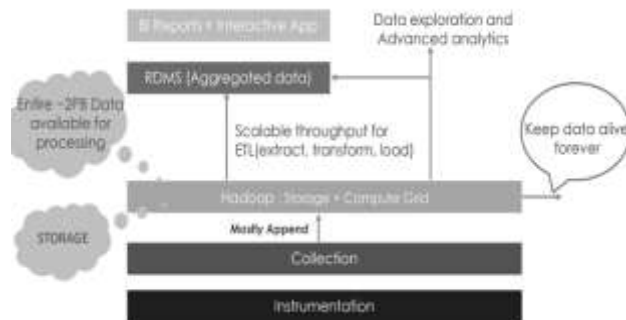


Fig1:big data concept

There are two types of data in big data[2]:

1. **Unstructured Data**:. The data generated from the social medias like Facebook, twitter, LinkedIn, instagram, Google+ like audios, videos etc.
2. **Machine data**: this data is generated from RFID chip readings and global positioning results(GPS).
3. **Structured data**: The data collected by the companies on their operations or the sales or the data stored in the form of tables is called structured data. These data are structured because data has a defined length and format for big data. Examples of structured data include numbers, dates, and groups of words and numbers called *strings*.

1. HADOOP COMPONENTS

Hadoop has two components HDFS and map reduce

A view of hadoop distributed file system(HDFS):-

The difference between regular file systems and HDFS is that in regular file systems each block of data is small, approximately 51 bytes and the problem for multiple seek because of large access for Input/Output unlike HDFS where each block of data is large, 64 MB and a single seek is required for reading huge data sequentially.

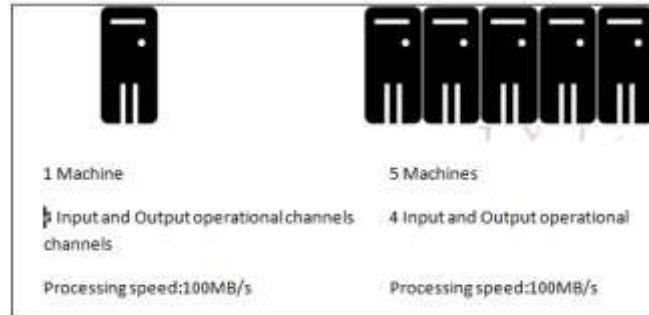


Fig2:distributed file system

The architecture of HDFS is considered as master/slave. It consists of a Single NameNode, a master server that manages the file system namespace and regulates access to files by clients. The data node is usually responsible for management of the storage attached to the nodes. Due to the replication of data on different data nodes Hadoop distributed file system is called as highly fault tolerant. The important feature of HDFS is that it is the storage system for the Map reduce jobs, both for input and output.

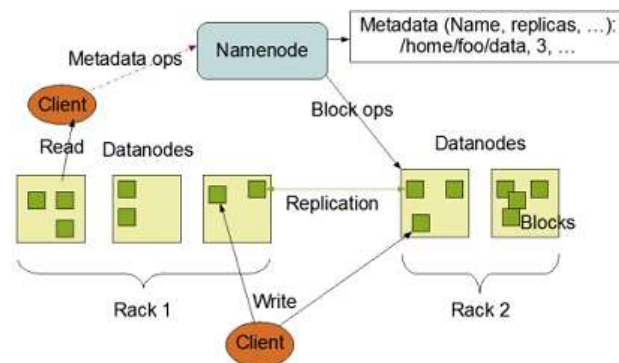


Fig3:representation of data nodes

There are three modes of Hadoop configuration

1. **standalone mode**:-in this mode, all Hadoop services runs in a single JVM on a single machine.
2. **pseudo-distributed mode**:- in pseudo-distributed mode, each Hadoop runs on its own JVM, but on a single machine.
3. **Fully Distributed mode**:-in Hadoop distributed mode, Hadoop services runs on individual JVM, but these reside in separate commodity machine in single cluster.

Hadoop services- the main services of Hadoop involves

Data node:-data nodes in Hadoop distributed file system(HDFS) are the slaves[10] that is responsible for storing blocks of data.

Name node:-It is the master node that is responsible for the management of data blocks that resides in data node. It is centrally placed node, which contains information about Hadoop file system[10].

Secondary name node:-the secondary name node is a specially dedicated node in HDFS to take a checkpoint of the file system metadata that is present in name node. It keeps track of the data that it is alive. It cannot be considered as the replacement for name node but can be considered as the helper node. If the name node fails, the data can be recovered from secondary name node's logs.

In Hadoop 1.0 map reducer was responsible for performing all the tasks like “job tracker”, allocation of resources so, what if map reducer crashes so here comes YARN with Hadoop 2.0 that brings many flexibilities for map reducer and the allocation and management of resources is been allocated to YARN and map reduces is supposed to do only processing of data.

YARN(yet another resource navigator) it is a resource management framework for scheduling and handling resource requests from distributed applications and can be called as resource manager. YARN split the two major functionalities of job tracker into two separate daemons :

1. Application manager
2. Application master

The application master runs the job and give results to application manager.one application master will be allocated for one job submission.it is responsible for containers and managing the application.it also manages the resources required for the job and give the report to the resource manager.

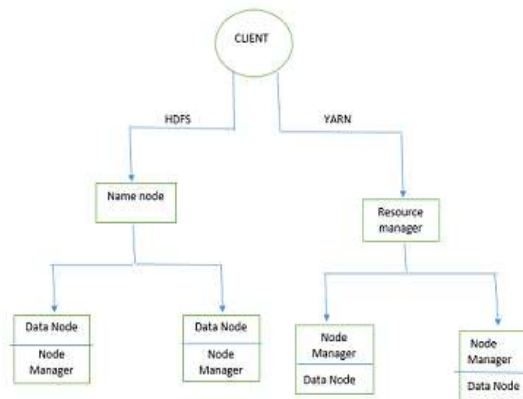


Fig 4:Hadoop cluster architecture

B. MAP REDUCE

The other component of hadoop apart from HDFS is map reduce. It is the programming model and an implementation for processing and generating large data sets with parallel and distributed algorithms on a cluster. Map reduce has become a ubiquitous framework for large-scale data processing [3].It is an initial ingestion and transformation step, where individual input records can be processed in parallel. Reduce process is the aggregation or summation step, in which all associated records must be processed together in a group. Task tracker keeps track of individual map tasks, and can run parallel. A map reduce job runs on a particular task tracker slave node. Jobs of a mapreducer

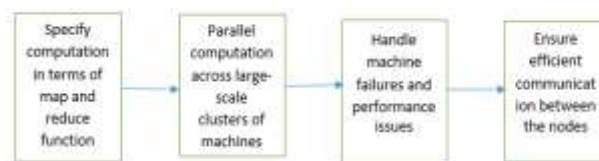


Fig5:map reduce

The steps involved in map reducer job

- 1.input
- 2.split
- 3.map
- 4.shuffle
- 5.Reduction

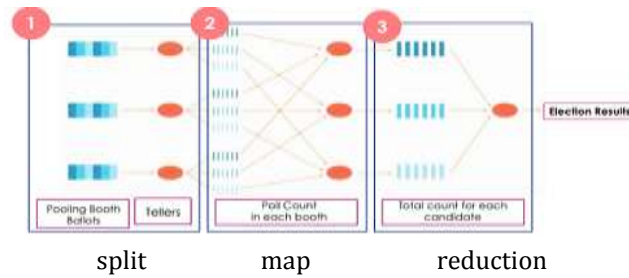


fig6:steps involved in map reduce

2. THE HADOOP ECOSYSTEM

Hadoop is a free and open source software framework. Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google’s MapReduce that is a software framework where an application break down into various parts. The Current Appache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper, HDFS and Map Reduce.[1].The Hadoop architecture can now be concluded and the place of HDFS, Map Reduce and YARN can be seen in the fig there are two more tools in Hadoop ecosystem that are important are FLUME and SQOOP known as the data ingestion tools.

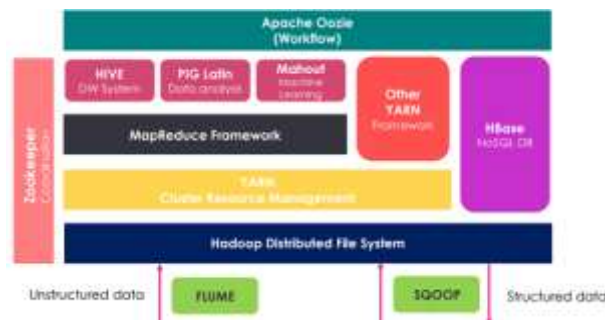


Fig7:hadoop architecture

Table1:Difference between traditional database and Hadoop

	Traditional DBMS	Hadoop
Processing	Traditional RDBMS cannot be used to process and store large amount of data or big data	but Hadoop has two main components HDFS that is responsible for storage of big data and Map Reduce that is responsible for processing large data by splitting it into several blocks of data and then distributing these blocks across the nodes on different machines.
Throughput	throughput can be defined as the total amount of data processed in a particular time period so that the output is maximum.RDBMS could not achieve higher throughput in any case when the data is very large(big).	This is the main reason behind the success of Hadoop over RDBMS.
Type of data	RDBMS can only be used for either structured(data in tabular form) or semi structured data(e.g, JSON data)	but because of the variety feature explained above in hadoop, it can be used for either structured, semi-structured or unstructured data.

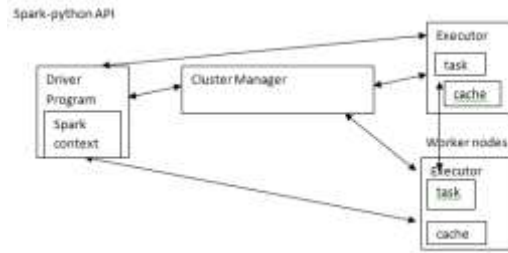


Fig8:spark context

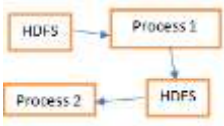
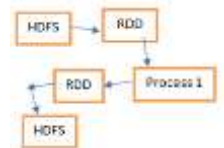
3. COMPONENTS DESCRIPTION:

Spark context-the spark context is the driver program that is responsible for the creation of RDD(resilient Distributed Datasets).The RDD represents a collection of items distributed across the cluster that can be manipulated in parallel. The datasets gets converted into blocks of data but into same RDD. The blocks are called atoms of the dataset. Spark context assigns an executor to each worker node for instance py4J whose responsibility is to transform by default spark context session into java-spark session for further processing of data. The transformation used most commonly are filter(),map(),join(), flatmap().

Cluster manager: the link between spark context and worker node is cluster manager. This manager works in three modes-standalone,YARN,MESOS.

Worker nodes: These are the nodes which runs the application code in the cluster. Here two worker nodes represents two machines. It is not recommended to run more than one worker node at a time.

Table2:The comparative analysis of Hadoop with spark

PARAMETERS	MAP REDUCE	SPARK
SPEED	<p>Slow in speed as compared to spark because of the storage and fetching of data in the following manner.</p>  <p>The diagram shows a sequence: HDFS -> Process 1 -> HDFS -> Process 2 -> HDFS.</p>	<p>There are fast as the processing of data is done in the following way. There is in disk and in_memory transfer which is in general fast.</p>  <p>The diagram shows a sequence: HDFS -> RDD -> Process 1 -> RDD -> HDFS.</p>
IMPLEMENTATION LANGUAGES	JAVA,C++,PEARL,RUBY	PYTHON,R,PROGRAMMING LANGUAGE,SCALA,JAVA,SQL
DIFFICULTY LEVEL	It is difficult to code in map reduce for data processing	Because of the presence of high-level operators it becomes easy to code in spark.
Fault Tolerance	Not fault tolerant	Spark is fault tolerant because the computations on RDD are represented as a lineage graph, a directed acyclic graph in spark.

4. CONCLUSION

We have entered in an era of billion or trillions of data that is also called Big Data. The paper describes the concept of Big Data based on 3 Vs, stands for volume, velocity and variety of Big Data. The technology associated to deal with big data -Hadoop, HDFS, Map Reduce. The challenges with Hadoop in comparison with spark. The paper also describes Hadoop ecosystem which is an open source software used for processing of Big Data. Hadoop is the need of today for processing and dealing with big data.

5. REFERENCES

- [1] International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153 www.ijsrp.org A Review Paper on Big Data and Hadoop Harshawardhan S. Bhosale¹, Prof. Devendra P. Gadekar.
- [2] International Journal of Engineering Technology, Management and Applied Sciences www.ijetmas.com March 2017, Volume 5 Issue 3, ISSN 2349-4476 19 AnjanaRaviprolu The Big Data and Market Research Anjana Raviprolu, Dr.Lankapalli Bullayya.
- [3] Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013).
- [4] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013
- [4] Kiran kumara Reddi & Dnysl Indira "Different Technique to Transfer Big Data : survey" IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}
- [5] Kenn Slagter · Ching-Hsien Hsu "An improved partitioning mechanism for optimizing massive data analysis using MapReduce" Published online: 11 April 2013
- [6] Umasri.M.L, Shyamalagowri.D, Suresh Kumar.S "Mining Big Data:- Current status and forecast to the future" Volume 4, Issue 1, January 2014 ISSN: 2277 128X
- [7] S.Vikram Phaneendra, E.Madhusudhan Reddy, "Big Datasolutions for RDBMS problems- A survey", In 12thIEEE/ IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).
- [8] Kaisler, S., Armour, F., Espinosa, J. A., Money, W. (2013). Big Data: Issues and Challenges Moving Forward. International Conference on System Sciences (pp. 995-1004). Hawaii: IEEE Computer Society.
- [9] Marr, B. (2013, November 13). The Awesome Ways Big Data is used Today to Change Our World. Retrieved November 14, 2013, from LinkedIn: <https://www.linkedin.com/today/post/article/20131113065157-64875646-the-awesome-ways-bigdata-is-used-today-tochange-our-world>.
- [10] Varsha B.Bobade] Survey Paper on Big Data and Hadoop, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 03 Issue: 01 | Jan-2016 www.irjet.net p-ISSN: 2395-0072
- [11] Shilpa Manjit Kaur, "BIG Data and Methodology- A review", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.
- [12] D. P. Acharjya, S. Dehuri and S. Sanyal Computational Intelligence for Big Data Analysis, Springer International Publishing AG, Switzerland, USA, ISBN 978-3-319-16597-4, 2015.
- [13] Jyoti Kumari, Mr. Surender, Statically Analysis on Big Data Using Hadoop, IJCSMC, Vol. 6, Issue. 6, June 2017, pg.259 – 265
- [14] C.L. Philip Chen, Chun-Yang Zhang, "Data intensive applications, challenges, techniques and technologies: A survey on Big Data" Information Science 0020-0255 (2014), PP 341-347, elsevier.
- [15] Katarina Grolinger At. Al. "Challenges for Map Reduce in Big Data", IEEE (10th World Congress on Services) 978-14799-5069-0/14, PP 182-189.
- [16] Gayathri Ravichandran ,Big Data Processing with Hadoop : A Review , International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 02, Feb -2017 .
- [17] Abdelladim Hadioui!!", Nour-eddine El Faddouli, Yassine Benjelloun Touimi, and Samir Bennani Machine Learning Based On Big Data Extraction of Massive Educational Knowledge, iJET – Vol. 12, No. 11, 2017.
- [18] Kache, F., Kache, F., Seuring, S., Seuring, S., Challenges and opportunities of digital information at the intersection of Big Data Analytics and supply chain management. International Journal of Operations & Production Management 37, 10–36, (2017) <https://doi.org/10.1108/IJOPM-02-2015-0078>
- [19] Zhou, L., Pan, S., Wang, J., Vasilakos, A.V., Machine Learning on Big Data: Opportunities and Challenges. Neurocomputing, (2017).