ISSN:0975-9646

A Comparative Study of Hadoop Family Tools

Prachi Pandey
UIT RGPV, Bhopal

Dr. Sanjay Silakari
UIT RGPV, Bhopal

Uday Chourasia
UIT RGPV, Bhopal

***Abstract*— Digitization is spreading across the India and other countries. As a result of digitization data is increasing at a very fast rate. Data almost gets doubled every two years. Such fast growing data which cannot be handled by traditional systems is termed as Big Data. But this data is used in various sectors to take business advantage and to analyze the trend. To analyze such Big data it is very important to first store data and then to process the data. To provide solutions to this Big data problem a parallel processing framework Hadoop and other tools like Hive and Pig are introduced. In this paper a survey on Big data analysis tools and comparison is made.**

***Keywords*—Big data, Hadoop, Hive. Pig.**

I. INTRODUCTION

Big data is defined as large quantity of data which have need of new technologies and architecture to make possible to extort value from it by capturing and analysis process. New sources of big data include location specific data which has arrived from traffic management and from the tracking of personal devices such as Smartphone's. Big data has come into view because we are living in the world which makes mounting use of data intensive technologies. Due to such large size of data it becomes very difficult to achieve effective analysis using existing traditional techniques.

Since Big data is new upcoming technology in the market which can bring the huge benefits to the business organizations, it becomes necessary various challenges and issues associated in bringing and adopting to this technology are need to be understand. Big data concept means a dataset which continues grew so much that it becomes difficult to manage it using existing database models and tools. So at last Big data is data that exceeds the processing capacity of conventional database systems. The data is huge sized, moves too fast, or doesn't fit the structures of our database architectures. To gain value from this data, you must choose a substitute way to process it.

Big data define as the pool of information's that is unable to handled or analyzed using an existing or traditional data mining techniques or the tools. Constantly increase of computational power has brought tremendous flow of data in the past two decades. This remarkable flow of data is called as "big data" it cannot be deal with the aid of existing tools or any other procedure and this is more comprehensible to computers.

The size of BigData range from petabytes (PB) to Exabyte's (EB) or to zettabytes (ZB).The BigData created from the client server interaction which is known as customer call records or transaction histories etc.

Big Data system is getting lot of importance now a day from organizations to handle those data as well as using them in business growth. Some Big data examples such as data from Finance, Internet, Mobile device, Radio – Frequency Identification (RFID), Science, Sensor and Streaming are the top most seven data drivers.

A. What are the problems?

There are many problems to handle big data like storage, processing etc.

a. Data integration – The structure of merging data is not so easy task with a reasonable cost.
b. Data volume – The ability to process the volume at a suitable rate so that the information is available to result analyzers when they need it.
c. Skills availability –There are shortage of people. Who have the proficiency to bring all data mutually, analyze it and publish the results.
d. Solution cost –To ensure a positive ROI on a Big Data project; it is crucial to reduce the cost of the solutions.

B. What are the solutions?

Big data is very difficult to process and store. Mainly Hadoop is used to process the big data. Hadoop used HDFS to store the data efficiently and MapReduce framework for processing the data. MPI is also used to process the big data.

C. TYPES OF BIG DATA

Mainly Big data is divided in 3 types.

- **Structured data**: It includes all data which stored in the database in tabular form. Structured data represent only 5 to 10% of all informatics data. [8]
  Ex. Relational data.
- **Semi Structured data**: Data which does not exist relational database form but have some organizational properties that make it easier to analyze.
  Ex. CSV but XML and JSON documents are semi structured documents, NoSQL is also considered as semi structured.
- **Unstructured data**: All data which is not structured and is in free format is unstructured. In fact, most individuals and organizations achieve their lives around free data. Unstructured data represent around 80% of data.
  Ex. videos, photos, presentations, WebPages and many other kinds of business documents, audio

files, E-mail messages, Word, PDF, Text, Media Logs.

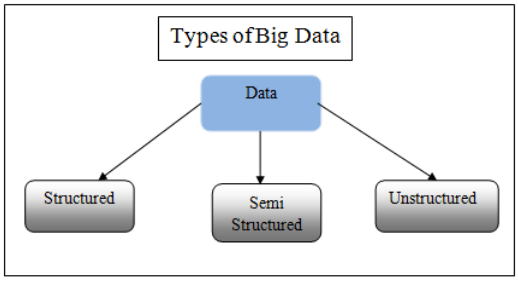

Fig 2: Types of Data

D. Challenges of Big Data

Big data may be the new trend in industry however, it can also mean to a considerable headache. The major challenges of big data are:

- Adoption of new technology
- Significant processing power
- Data integration
- Data volume
- Skill
- Solution cost
- Storage and Management
- High –Speed Networking
- Big Data Security

II. HADOOP

Hadoop is an open source parallel big data processing framework. It is included in Apache software foundation. Hadoop provides solution to Big data problem. It provides 3 layers:

- **Hadoop distributed file system(HDFS):** This layer provides distributed storage for big data across the cluster of nodes. For reliable data storage it also provide replication of each block.

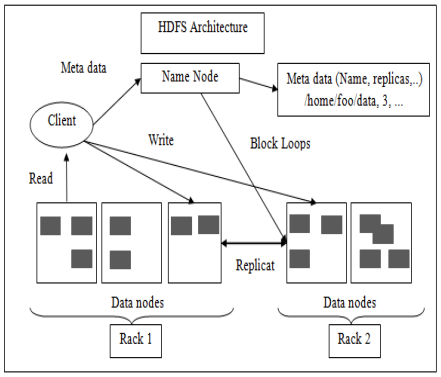

Fig 3: HDFS Architecture

- **Hadoop YARN:**It is a layer on top of HDFS which provides resource management and scheduling. On master node Resource manager is the daemon which is responsible for YARN and on worker nodes Node manager takes care of that.
- **Hadoop MapReduce:** It is a parallel processing framework for distributed processing.

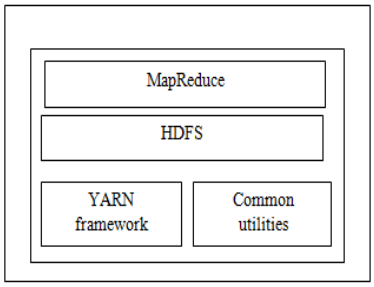

Fig 4: Hadoop Components

A. Hadoop MapReduce

Hadoop map/reduce is a parallel processing framework. Whenever any data is put on HDFS, data is divided into blocks with bock size of 128 MB. Namenode stores the metadata for every data. The resources are managed by namenode for data storage and resource manager manages processing on data nodes.

After successful distribution of data on HDFS whenever any job is submitted by the user to process the stored data, job is submitted to the resource manager. Resource manager asks namenode for the metadata of the data which is to be processed. And job is divided into tasks that is Mappers and Reducer. So the status of the whole job is monitored by Resource manager while status of the Mappers and Reducers is taken care by Node manager.

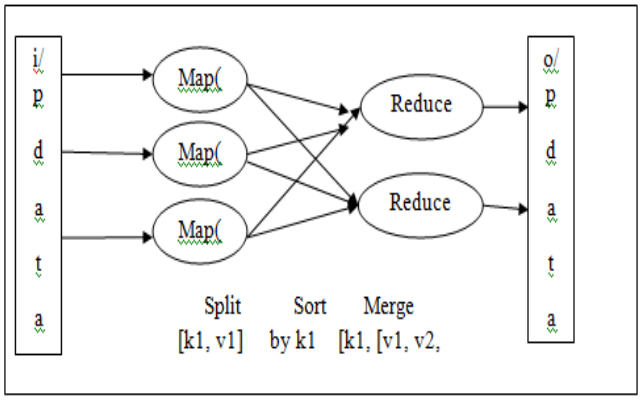

Fig 5: Map Reduce phases

B. Components of MapReduce

There are three basic components of MapReduce:

- Driver
- Mapper
- Reducer

There are various ways to execute MapReduce operations:

- The map/reduce is applicable for structured, semi structured and unstructured data.
- It can also be executed using scripting language like Pig.
- Or using SQL like language Hive.

III. LITERATURE REVIEW

In [2] Hardware acceleration of Mapreduce is proposed by utilizing multi core architecture of CPU. As MapReduce posses inherent parallelism and if multicore hardware can be utilized properly then acceleration can be achieved. Because sometimes number of mappers increases on a particular data node and CPU faces bottleneck. So to avoid this situation hardware acceleration is done in this paper. In [3] sequence alignment which is a basic method of processing information in Bioinformatics is done using Hadoop. It is used for finding sequences of proteins and nucleic acid. Most common local sequence alignment problem called BLAST is implemented in this paper. These algorithm posses some inherent parallelism so speedup is achieved when implemented on Hadoop. In [4] network failure detection system is built using Hadoop. As Hadoop have several daemons each for specific purpose. Job tracker for job submission task tracker for hadling running jobs over datanode. All these daemons create logs on respective nodes. These logs can be used to detect failures in network and used for network failure monitoring system in this paper. In [5] also a network failure monitoring system is proposed. In [6] energy aware scheduling of MapReduce jobs is done. As energy is a challenging issue in this era of computing. To save energy and to develop green algorithm is a big task. In this paper a new energy aware scheduling algorithm replaces traditional scheduling algorithm of Hadoop, which saves about 40% energy.

IV. PIG

Pig is a Hadoop extension that simplifies Hadoop programming by giving you a high-level data processing language while keeping Hadoop's simple scalability and reliability.

Pig has two major components:

- A high-level data processing language called Pig Latin .
- A compiler that compiles and runs your Pig Latin script in a choice of *evaluation mechanisms* .

The main evaluation mechanism is Hadoop. Pig also supports a local mode for development purposes.

Pig over MapReduce: Pig runs over the top of MapReduce thus all the Hadoop daemons must be running before starting Pig. Grunt is the name of the shell which runs over MapReduce.

V. HIVE

It is a platform used to develop SQL type scripts to do MapReduce operations. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic Map/Reduce.

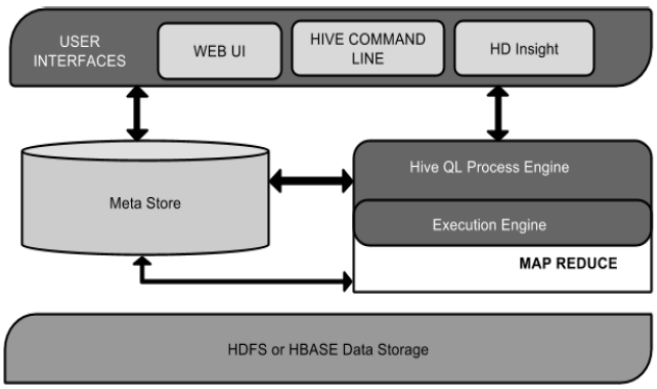

Fig 6: Hive Stack

VI. COMPARISON PIG & HIVE

	Pig	Hive
Type of Data	Semi Structured or unstructured both	Structured only.
Type of Tool	Scripting language	Query based (HiveQL)
MapReduce	Runs MapReduce in Background	Also Runs MapReduce in Background
Concept of data storage	Uses concept of Bags for data storage which are in turn stored over HDFS.	Uses concept of tables and databases which are in turn stored over HDFS.

VII. CONCLUSION

As Big data analytics is trending these days due to lots of applications areas. This survey puts light on Big Data analysis techniques. In this paper Hadoop, Pig, Hive have been studied and it is found that Pig and Hive are easy to use and high level utility tools. So we lots of customizations are not to be done in Mapper and Reducer these utility tools can be used. And if customization is needed then Map Reduce code has to be written. Hive and Pig can be used to write above mentioned queries and Map reduce can be used to write customized and efficient code for the same. In this survey we have analyzed Flight data in terms of the parameters mentioned and information is retrieved using Hadoop.

REFERENCES

[1] Oscar D. Lara, Weiqiang Zhuang, and Adarsh Pannu "Big R: Large-scale Analytics on Hadoop using R" in 2014 IEEE International Congress on Big Data.

[2] Toshimori Honjo, Kazuki Oikawa "Hardware acceleration of Hadoop MapReduce" in 2013 IEEE International Conference on Big Data.

[3] Ming Meng, Jing Gao, Jun-jie Chen "Blast-Parallel: The Parallelizing Implementation Of Sequence Alignment Algorithms Based On Hadoop Platform" in 2013 6th International Conference on Biomedical Engineering and Informatics (BMEI 2013).

[4] Madhury Mohandas, Dhanya P M "An Approach for Log Analysis Based Failure Monitoring in Hadoop Cluster" in 2013 IEEE.

[5] Ilja Kromonov, Pelle Jakovits, Satish Narayana Srirama "NEWT - A Resilient BSP Framework for Iterative Algorithms on Hadoop YARN" in 2014 IEEE.

[6] Lena Mashayekhy, Mahyar Movahed Nejad, Daniel Grosu, Dajun Lu, Weisong Shi "Energy-aware Scheduling of MapReduce Jobs" in 2014 IEEE International Congress on Big Data.

[7] Kiran M., Amresh Kumar "Verification and Validation of Parallel Support Vector Machine Algorithm based on MapReduce Program Model on Hadoop Cluster" in 2013 International Conference on Advanced Computing and Communication Systems (ICACCS - 2013), Dec. 19 – 21, 2013, Coimbatore, INDIA.

[8] Arshdeep Bahga, Vijay K. Madisetti, Analyzing massive machine maintenance data in a computing cloud, IEEE Trans Parallel Distrib. Syst. 23 (10) (2012) 1831–1843.

[9] Sergio Barbarossa, Gesualdo Scutari, Bio-inspired sensor network design, IEEE Signal Process. Mag. 24 (3) (2009) 95–98.

[10] Ron Bekkerman, Mikhail Bilenko, John Langford, Scaling Up Machine Learning: Parallel and Distributed Approaches, Cambridge University Press, 2012.

[11] Gordon Bell, Tony Hey, Alex Szalay, Beyond the data deluge, Science 323 (5919) (2009) 1297–1298.

[12] Yoshua Bengio, Learning deep architectures for ai, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.

[13] Yoshua Bengio, Aaron Courville, Pascal Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[14] Janine Bennett, Ray Grout, Philippe Pebay, Diana Roe, David Thompson, Numerically stable, single-pass, parallel statistics algorithms, in: IEEE International Conference on Cluster Computing and Workshops, 2009, CLUSTER '09, 2009, pp. 1–8.

[15] Paul Bertone, Mark Gerstein, Integrative data mining: the new direction in bioinformatics, IEEE Eng. Med. Biol. Mag. 20 (4) (2001) 33–40.

[16] Josh Bongard, Biologically inspired computing, Computer 42 (4) (2009) 95–98.