Volume 6, Issue 1, January 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey on MapReduce, Hadoop and YARN in Handing Big Data

Sudha P. R

Assistant Professor, Department of ISE,

JSS Academy of Technical Education, Bangalore, India

Abstract- ***Today, the data is not only generated by people, but massive data is generated by machines also and it surpasses human generated data. This data is spread across different places, in different formats, in large volumes ranging from Gigabytes to Terabytes, Petabytes, and exabytes. In different areas of technology, data is being generated at different speeds. A few examples include stock exchange data, tweets on Twitter, status updates/likes/shares on Facebook,data from sensors, images from medical devices, surveillance videos, satellites data and many others. "Big Data" refers to a collection of massive volume of heterogeneous data that is being generated, often at high speed, from different sources.Traditional data management and analysis systems fall short of tools to analyze these data thus there is a need of innovative set of tools and frameworks to capture, process and manage these data within a tolerable elapsed time.Thus the concept of Big data is catching popularity faster than anything else in this technological era. Big Data demand cost-effective, fault tolerant, scalable and flexible and innovative forms of information processing for decision making. This paper emphasis on the features, architectures, and functionalities of Big data, Hadoop, Map Reduce, HDFS and YARN.***

Keywords- Big data, Hadoop, YARN, HDFS, MapReduce.

I. INTRODUCTION

Ancient days, man used primitive way of storing data on wood, carving on stones and then he started storing it on paper, cloths. New inventions made him to store the data on magnetic drums, floppy disks, magnetic tapes, Compact Discs, Hard Discs, pen drive etc. From this trend, the capacity of data storage has been increasing exponentially, and today data is stored on the cloud infrastructure. In Big data, the information comes from multiple, heterogeneous, autonomous sources and continuously growing. Upto 2.5 quintillion bytes of data is created daily and 90 percent data in the world today were produced within past two years [2] [7]. Hundred Terabytes of data is uploaded to Facebook every day, Twitter generates twelve Terabytes of data every day, YouTube users upload 48 hours of new video content every minute of the day, five hundred plus new websites are created every minute of the day [9]. Traditional data management and analysis systems are not enough to analyze these data thus there is a need of innovative set of tools and frameworks to capture, process and manage these data within tolerable elapsed time[1]. Thus Big Data demand reliable, fault tolerant, scalable, less expensive, flexible and innovative forms for processing the data.

II. BIGDATA

A. Characteristics of Big Data

Important Three V's of Big Data [10] are:

*1) Volume:*With the invention of social media and with the advancement of technology, the amount of data generated and collected is growing very rapidly. This data is getting collected in different formats, from different sources, in massive volume ranging from Terabytes to Petabytes, and even exabytes. Today human aloneis not involved in generating the data, but a huge amount of data is being generated by machines also and it overtakes human generated data. This size feature of data is referred to as Volumein the Big Data world.

*2) Velocity:*Velocity refers to the how fast the data is generated. In today's competitive world, decision makers need the necessary information in the least amount of time. Data is prolifering at different speeds and from different areas. A few examples include stock exchange data, tweets on the Twitter, updates and likes on Facebook etc. The speed at which data is generated is referred to asVelocityin the Big Data.

*3) Variety:*Variety refers to the data being generated and stored in different formats. Different applications will generate the data in different formats. Industry didn't have any powerful and reliable tools/technologies which can work with such voluminous unstructured data till the advancements in Big Data technologies. Unstructured data are stored in the form of audio files, sensor data, images, video files, web logs, etc. This aspect of varied formats of data is referred to as*Variety*in the Big Data.

B. Sources of Big Data

Sources of Big Data are broadly classified into six different categories as shown below.

*1) Public Data:*Public data includes data that is publicly available like data generated by government sectors, weather data, Wikipedia, research data, open source data and other data which is freely available to the public. This type of publicly accessible data is referred to as*Public Data*.

*2) Transactional Data:*Every enterprise will have some kind of applications which will perform different kinds of transactions like Mobile Applications, Web Applications and many more. In order to support the transactions of these applications, there are one or more relational databases which works at backend. These data are structured data and is referred to as*Transactional Data*.

*3) Social Media:*Huge amount of data is being generated on social networks like Twitter, LinkedIn, Facebook, etc. Thus social media has to capture and manage unstructured data formats which includes images, text, videos, audio etc. This category of data source is referred to as*Social Media*.

4) Enterprise Data: Huge amount of data comes from enterprises in different formats. Formats may be in the form of flat files, Word documents, emails, spreadsheets, PowerPoint presentations, HTML pages, pdf files, XMLs, legacy formats, etc. This data which is spread across the organization in different formats is referred to as*Enterprise Data*.

*5) Activity Generated data:*Data that has been generated by machines that surpasses the data volume generated by humans. These include data from various machines like images from medical devices, data from sensors, surveillance videos, satellites data and data from mobile towers. These types of data are referred to as***Activity** Generated*data.

*6) Archives:*Archives are the data which is very rarely required or which is not required anymore for any organization. Now a day's cost of the hardware is so cheap that none of the organization would like to discard any data, they would like to capture and store as much data as possible. Archived data includes records of ex-employees, old bank transactions, scanned documents, agreements copies, completed projects, this type of data which is less frequently accessed is referred to as *Archive Data*.

- ***Structured Data*** : Structured data is the datawhich has a well-defined data schema, and can be easily managed and processed using the traditional tools and techniques.Relational databases, XML files, data from Customer Relationship Management systems like call centers are some examples.
- ***Unstructured Data***. Unstructured data is the data, which do not have a well-defined schema or does not fit well into the relational world. Unstructured data includes spreadsheets, PDF files, Word documents, e-mails, audio files, video files, images, scanned documents, etc

C. Big Data Statistics [9]

100 Terabytes of data is uploaded to Face book every day,
Twitter generates 12 Terabytes of data every day,
YouTube users upload 48 hours of new video content every minute of the day,
500+ new websites are created every minute of the day.

III. HADOOP

Hadoop is a scalable, distributed and fault tolerant open source software framework written in java, for distributed storageand distributed processing any kind of huge data [22].

Some of the Characteristics of Hadoop are[6][8]:

- Hadoop provides distributed and reliable storage (HDFS) and processing system (MapReduce).
- Hadoop can scale massively, it can scale from hundreds to thousands of servers that does not require expensive high-end hardware.
- Hadoop is highly flexible as it can process both structured and unstructured data.
- Hadoop is highly fault tolerant. Same Data is replicated across many nodes and if a node goes down, that data can be read from another node.
- Hadoop works on the principle of read multiple times and write once .
- Hadoop is optimized for massive data sets. It takes more time to process less data than traditional systems.

A. Apache Hadoop Ecosystem[16]

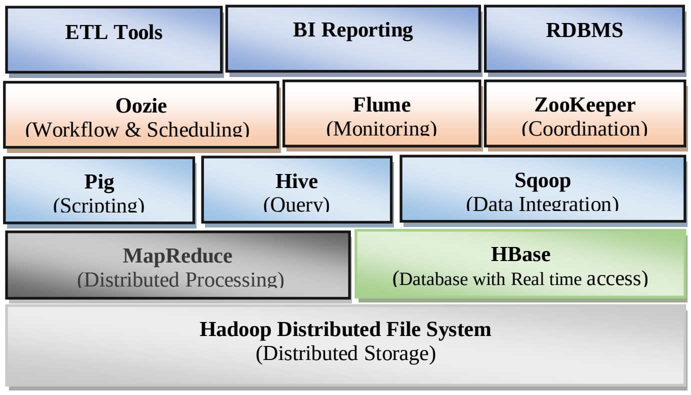

ETL Tools
BI Reporting
RDBMS
Oozie (Workflow & Scheduling)
Flume (Monitoring)
ZooKeeper (Coordination)
Pig (Scripting)
Hive (Query)
Sqoop (Data Integration)
MapReduce (Distributed Processing)
HBase (Database with Real time access)
Hadoop Distributed File System (Distributed Storage)

Fig. 1 Apache Hadoop Ecosystem

Hadoop Components and their functionalities[19]

HDFS	:	Distributed Storage.
MapReduce	:	Distributed Processing
Hbase	:	Column-oriented database with real-time access
Flume	:	Monitoring
Pig	:	Scripting
Hive	:	SQL
Oozie	:	workflow and Scheduling
Zookeeper	:	Configuration Management and Coordination
Sqoop	:	Imports data from relational databases

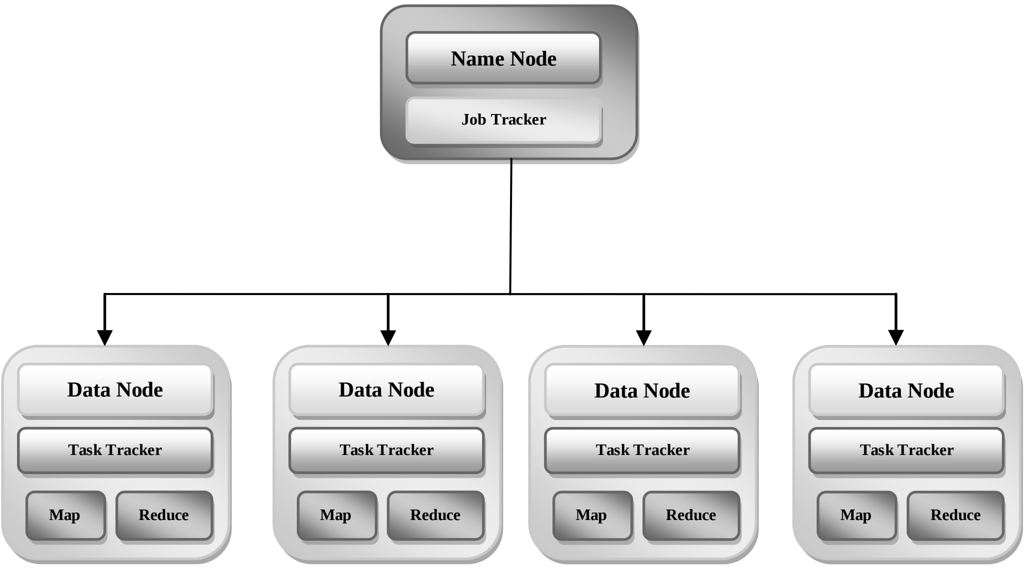


Fig. 2 Apache Hadoop master slave architecture.

1) Hadoop Distributed File System (HDFS): Hadoop distributed File System (HDFS) is a default distributed big data storage for Apache Hadoop, which runs on a clusters of low cost commodity hardware to store the data. Replicas of the data blocks are stored across multiple machines in redundant fashion to avoid data losses in case of failure and for faster access. HDFS is fault tolerant, highly available[3] and it also supports parallel processing..

Some of the key concepts [10].

- *Data Node and Name Node [5]:* data node is a Slave Node which actually exectues the tasks, NameNode is a Master Node which maintains the information of data Node , blocks and free space etc.
- *Job :* From the Hadoop perspective, Job is the unit of work to be done as requested by the client / user.
- *Tasks:* Jobs are divided into multiple sub-jobs known as Tasks. These tasks can run independently on various data nodes across the cluster.
- *JobTracker:* JobTracker node is responsible for scheduling, allocating and executing jobs on slave nodes, It also re-executes the failed Tasks and monitors the overall progress of the Job. There is only one JobTracker node for one Hadoop Cluster.
- *TaskTracker:*Any number of TaskTracker nodes can exist in a Hadoop Cluster. TaskTracker receives the necessary information from JobTracker for execution of a Task and then it executes the Task and Sends the Results to the JobTracker.
- *Map():* Map() is a function of MapReduce, which is responsible for processing data and producing the intermediate results.
- *Reduce():* Reduce() is a function of MapReduce which consolidates all the intermediate results produced by Map().
- *Data Block [7] :* Data Block can be considered as the standard unit of data/files stored on HDFS. By default each file is broken into 64MB, any file whose size exceeds 64 MB will be broken down into blocks of 64 MB, except the last block, which will be less than 64 MB that depends on the size of the file.

Data node maintains a block report for the blocks in its node and periodically sends this block report to the Name node so that Name node will have updated information about the location of replicas of data block in the clusters.Data node will also send heartbeat messages to Name Node every ten minutes which indicates that Data node is active and available, even after ten minutes if Name Node does not receive heartbeat messages from Data node, then Name node assumes that data node is lost and replicates data of the lost datanode from the other data nodes.

2) MapReduce: Mapreduce is a programming Model used in Hadoop for distributed processing of large datasets across different nodes in clusters [4]. MapReduce consists of two functions map and reduce, where map functions reads the data from the HDFS , processes that data and generates multiple intermediate results. Whereas Reduce function will combine all these intermediate results and generate the final results and written backs to HDFS.

There are three kinds of failures in MapReduce [13]:

- *Task Failures:* Causes for Task failure may be an interruption on a running mapper or reducer, which needs re-execution of the interrupted task. There exist several other reasons for a task failure: *Bad records,Media corruption*, Bugs in the jobs or in third-party software and *Contention i.e*, Sometimes Nodes may have to wait for shared resources during the task execution which will slow down the task and thus that task will be considered as failed.
- *Slave Failures[18]:* This type of failure may occur because of overheating of CPU, hard disk failures or any hardware failure. This failure stops the data node from accepting the new tasks from the name node.
- *Master failures[18]:* It is nothing but failure of Name node, this can be easily solved by having a Secondary Name node, which maintains the status ofthe Name node. Thus, the Secondary Name node can take over the role of Name node in case of failure.

Working of Map Reduce [10]: The following diagram shows the logical flow of a MapReduce programming model.

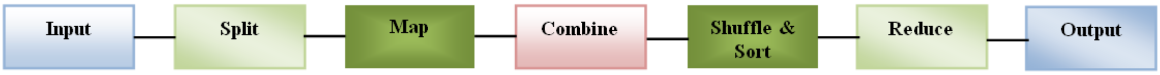
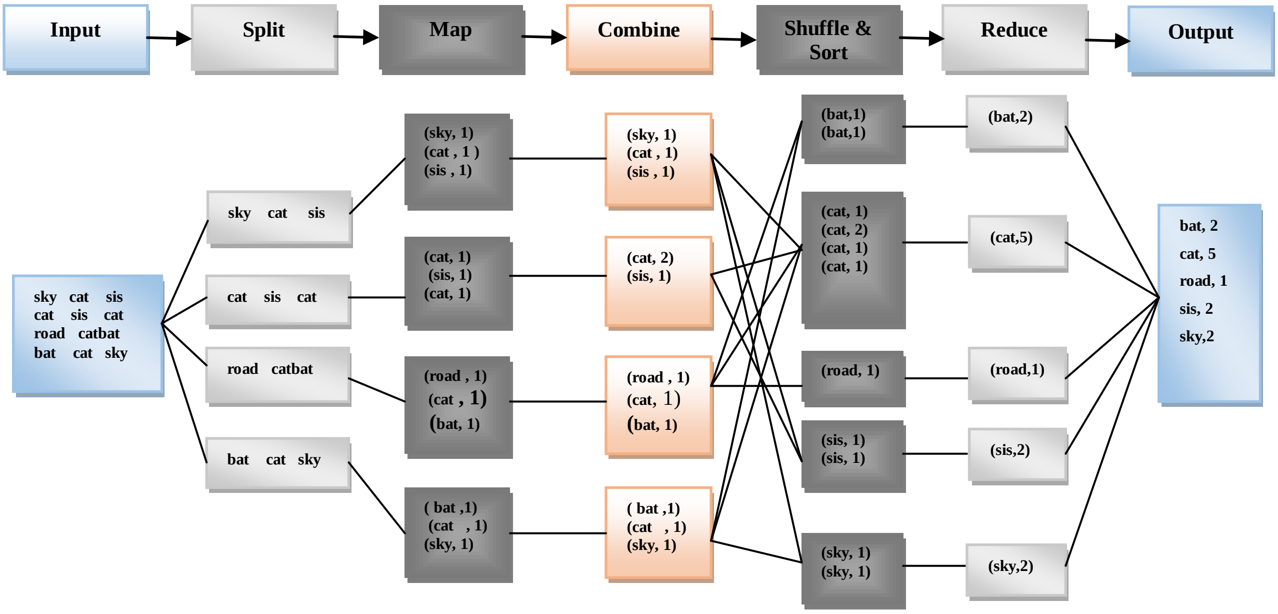


Fig.3 Mapreduce- Logical flow of data

MapReduce Word Count Example

Here is an example to understand the working of Map Reduce. File contains the following four lines of text, now how counting operation is performed when Map reduce takes this file as an input and how it counts the number of occurrences of each word is shown below.

sky cat sis
cat sis cat
road cat bat
bat cat sky



3) Yet Another Resource Negotiator(YARN):There were many Limitations for Hadoop 1.0

- JobTracker which was the part of MapReduce Framework, and which runs on a single machine was responsible for Resource management, Job Sheduling and monitoring across the cluster which limited the scalability. Failure of Jobtracker caused restarting of all the jobs.
- Map and Reduce slots are predefined and fixed. Say when map slots are full and Reduce slots are empty, resources allotted for reduce slots would sit idle but which were very essential for Map slots. Thus created problems in resource utilization.
- Only supporting Applications which obeys MapReduce framework can run on Hadoop1.0.

Hadoop 2.0 overcomes all these problems with YARN (Yet Another Resource Negotiator). Yarn performs the cluster Resource Management and Node Management Tasks and MapReduce does only data Processing tasks. There is no fixed slots for map and Reduce. No Job Tracker and TaskTracker are required for Hadoop 2.0. Any application thatfallows or that do not follow MapReduce Model can run on YARN*[15].*

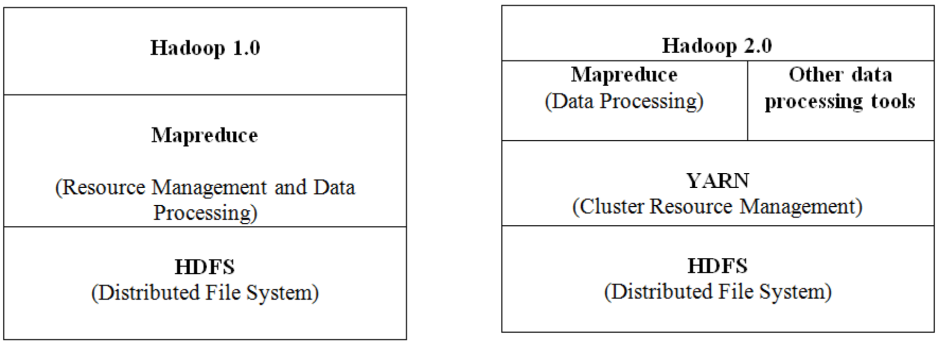

Hadoop 1.0
Mapreduce
(Resource Management and Data Processing)
HDFS
(Distributed File System)
Hadoop 2.0
Mapreduce
(Data Processing)
Other data processing tools
YARN
(Cluster Resource Management)
HDFS
(Distributed File System)

Fig. 4 Architecture of Hadoop 1.0 and Hadoop 2.0

IV. CONCLUSION

The data is generated and proliferating worldwide both from machines and human beings at different speeds and in different formats due to tweeters, facebook, stock trading sites, news sources and so on. Big Data is becoming the new area of research. Big data analysis helps business people to make better decisions and researchers to identify new opportunities. This paper presents fundamental concepts of Bigdata like characteristics, sources, statistics, frameworks and technologies to handle big data, differences between Hadoop 1.0 and Hadoop 2.0.

ACKNOWLEDGMENT

I would like to thank JSS Academy of Technical Education, Bangalore for providing me the facilities for carrying out this work.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu, and Wei Ding,*”Data Mining with Big Data”* ,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014

[2] Shital Suryawanshi, Prof. V.S.Wadne, “*Big Data Mining using Map Reduce: A Survey Paper*”, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. VII (Nov – Dec. 2014), PP 37-40 www.iosrjournals.org

[3] Singh and Reddy, ”A Survey on platforms for big data Analytics” Journal of Big Data 2014, 1:8

[4] J Dean, S Ghemawat, “*MapReduce: simplified data processing on large clusters*”, Communications of the ACM, 2008 – dl.acm.org.

[5] Suman Arora, Dr.Madhu Goel, “*Survey Paper on Scheduling in Hadoop*”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014

[6] http://hadoop.apache.org/

[7] Ms.Vibhavari Chavan, Prof. Rajesh. N. Phursule. “*Survey paper on Big Data*”, International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939

[8] https://www.dezyre.com

[9] http://wikibon.org/blog/taming-big-data

[10] https://www.mssqltips.com

[11] Jorge-Arnulfo Quiané-Ruiz, Christoph Pinkel, Jörg Schad, and Jens Dittrich “RAFTing MapReduce: Fast Recovery on the Raft” , Information Systems Group, Saarland University

[12] Sabia, Love Arora,” Technologies to Handle Big Data: A Survey”, International Conference on Communication, Computing & Systems (ICCCS–2014)

[13] http://hortonworks.com/hadoop/yarn/

[14] Harshawardhan S. Bhosale , Prof. Devendra P. Gadekar, “ A Review Paper on Big Data and Hadoop“, International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014, ISSN 2250-3153.

[15] Jeffrey Dean and Sanjay Ghemawat,”MapReduce: Simplified Data Processing on Large Clusters”

[16] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani , “Big Data: Survey, Technologies, Opportunities, and Challenges”, Hindawi Publishing Corporation, The Scientific World Journal Volume 2014, Article ID 712826.

[17] Sagiroglu, S.; Sinanc, D.,”Big Data: A Review”,2013,20-24

[18] Puneeth Singh Duggal, Sanchita paul, “Bigdata Analysis: challenges and Solutions”,Internal conference on cloud , big data and Trust 2013, Nov 13-15, RGPV