Hadoop-Need for Big Data

Amanpreet Kaur

Department of Computer Science, Guru Nanak College for Girls, Sri Muktsar Sahib

ABSTRACT

In this paper I have discussed the upcoming need of today i.e. big data, area in which this is used. The way the data is being stored in three phases with types of data.. The solution was provided by the Google named hadoop an open source frame work that is used to store, process and analysis of data which is large, in a distributed environment across clusters of computers using simple programming models.

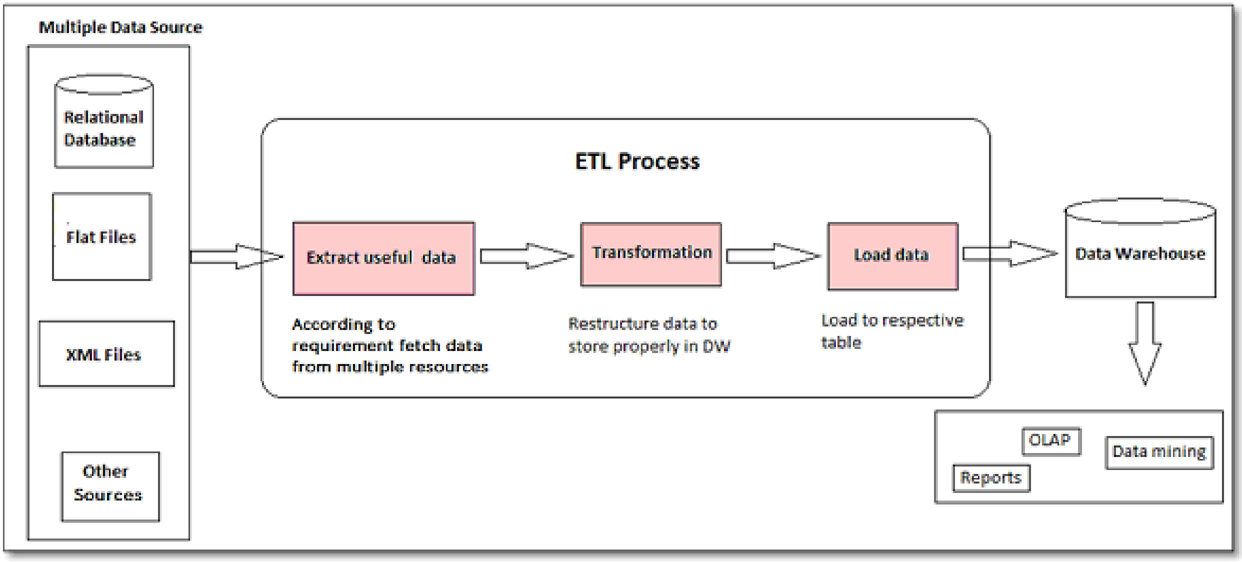
Keywords: Big data, Hadoop, HDFS, Map reduce, Yarn.

1. INTRODUCTION

Big data means collection of large datasets that cannot be processed using traditional computing techniques. There are basically three types of data needs to managed **Structured data** the data in the form of relations, **Semi Structured data** Extensible Markup Language data, Unstructured data in the form of Word, PDF, Text, Media Logs etc. Big data involves various tools, techniques and frameworks. Big data involves the data produced by different devices and applications**. Social Media Data**: Social media such as Face book and Twitter, LinkedIn hold information and the views posted by millions of people across the globe. Using the information kept in the social network, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums. **Production related Data:** This data holds information about the 'buy' and 'sell' decisions of product of different companies. Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production. **Medical Facilities:** Using the data regarding the previous medical history of patients, hospitals are providing better and quick service. Using the data lying online regarding the symptoms of diseases even a layman can diagnosis a problem and can have first aid by itself. **Search Engine Data**: Search engines retrieve lots of data from different databases.

2. STORAGE PHASES

First Phase: In this approach, an enterprise will have a computer to store and process big data. Here data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated software's can be written to interact with the database, process the required data and present it to the users for analysis purpose. This approach is suited to store the data up to certain limit in this huge amount of data becomes a tedious task to store and retrieve **Second Phase:** Before processing big data it must be recorded from various data generating sources. After recording, it must be filtered and compressed. Only the relevant data should be recorded by means of filters that discard useless information. In order to facilitate this work specialized tools are used such as ETL. ETL tools represent the means in which data actually gets loaded into the warehouse.ETL is a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse.



ETL process

Extracting data	The data is extracted from the source systems.
Transforming data	Data is transformed The phases are 1. Data analysis by applying various rules 2. Definition of transformation i.e. cleaning and mapping of values 3.Filtering of data by selecting certain columns to load and then split in into multiple columns 4. Verification and validation process (Simple or complex validation).
Loading data	Load into warehouse or repository for reporting in its format

RDBMS	HADOOP
RDBMS works on Structured data means data in predefined Scheme. It has primary key allocated by DBA.	Hadoop also works on semi structured data i.e. loosely followed and has no structure. In this Primary Key and Values are chosen by implementer at Processing time.
It is best for Real time and online system	It is best for offline batch processing System on large amount of data.
It can handle limited data.	It can handle large amount of data.
High throughput	Low Latency.
Costly hardware	Cheap hardware

COMPARISON BETWEEN RDBMS & HADOOP

Third Phase: But by using ETL process all **unstructured** data in such a vast ratio that is quite difficulty to store, extract, transform and load. But the problem is solved **by Google using an algorithm called Map Reduce** in the Hadoop Technique. This algorithm divides the task into small parts and assigns those parts to many computers connected over the network, and collects the results to form the final result dataset.

2.1. Hadoop: The solution was provided by the Google named hadoop an open source frame work that is used to store, process and analysis of data which is large , in a distributed environment across clusters of computers using simple programming models. In 2005 Dough Cutting and his team named hadoop , after toy elephant of his son..Now Apache Hadoop is a registered trademark of the Apache Software Foundation. Its written in java. It is used for batch/offline processing.It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. As the technology is hiking, the amount of data is growing rapidly from 5 hundred gigabytes in year 2003 to every 5 minutes in 2017 and this rate is still growing..It is being used by Face book, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster. Hadoop runs applications using the Map Reduce algorithm, where the data is processed in parallel on different CPU nodes. The framework of hadoop is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for huge amounts of data. As the storage capacity has grown but with the little change in the speed. With hadoop it is Possible to read bulk of data in few minutes.

Year	Storage	Transfer	Read
1999	1400 GB	4.5 MB/s	5 minutes
2000	1 TB	100 MB/s	3 Hrs
2017	**100 drives**	**100 MB/s**	**1 TB of data in 2 minutes**

TABLE -I

2.2. Principles of Hadoop:

Scale out	It lays the focus on scale out rather than scale up i.e. by adding more nodes/machine to an existing system. In scale up we add resources to the existing system like CPU,RAM etc. which increases the cost. Hadoop uses commodity hardware to store data so it really cost effective as compared to traditional RDMS. It’s a scalable system.
Code to Data Rather than Data to Code	As the size of data is large, When data moves from the processing node to the storage and vice versa, It may cause bottle neck in the network, in the hadoop system, code is moved as the size of code is smaller in size that is usually in KB.
Resilient to failure	Failures do occur in the Hadoop system but tackled as in this system as it has the property, with which data can be replicated over network, so if one node get prone to failure, then Hadoop takes the other copy of data. In Hadoop data is replicated thrice to have best recovery mechanism.

2.3 Modules of Hadoop

2.3.1. HDFS: Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture. Hadoop comes with a distributed file system called HDFS. In HDFS data is distributed over several machines and replicated to ensure their durability to failure and high availability to parallel application. It is cost effective as it uses commodity hardware. It involves the concept of blocks, data nodes and node name.

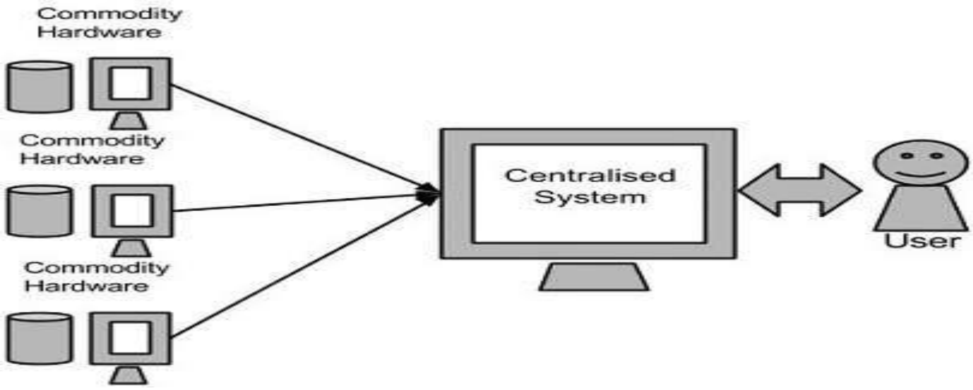


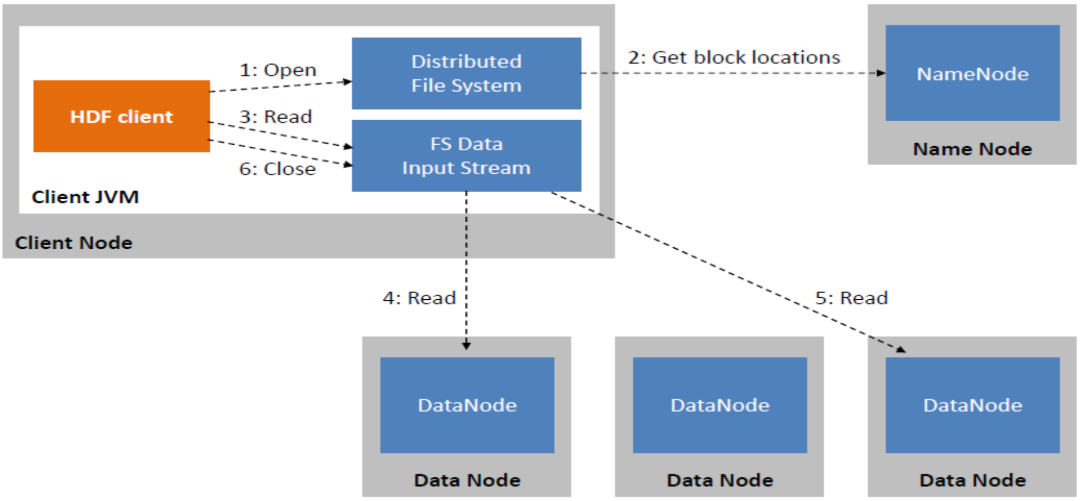
Diagram shows various commodity hardware's which could be single CPU machines or servers with higher capacity.

2.3.2. Concepts of HDFS

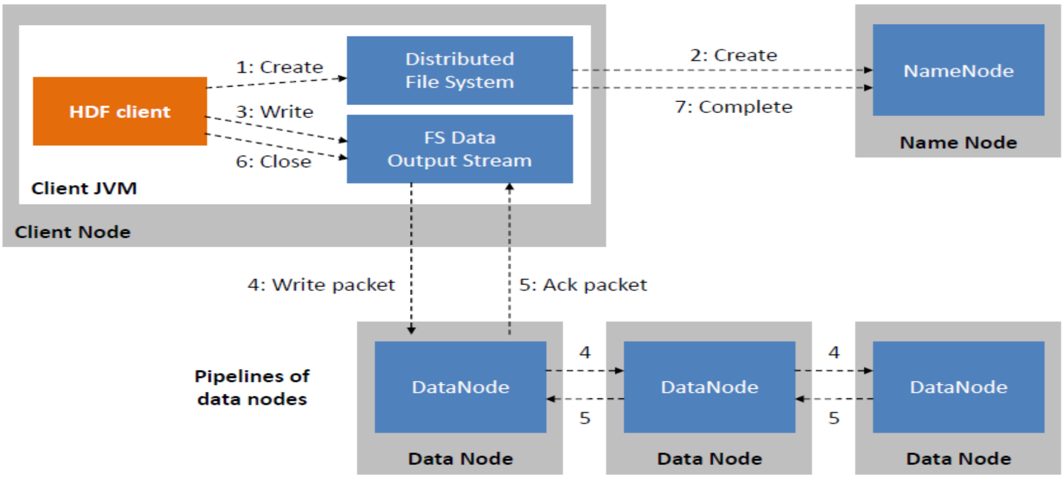
Blocks: A Block is the minimum amount of data that it can read or write. HDFS blocks are 128 MB by default and this is configurable. Files in HDFS are broken into block-sized chunks, which are stored as independent units. Unlike a file system, if the file is in HDFS is smaller than block size, then it does not occupy full blocks size, i.e. 5 MB of file stored in HDFS of block size 128 MB takes 5MB of space only. The HDFS block size is large just to minimize the cost of seek.

Name Node: It contains GNU/LINUX operating system and name node software. The system having name node act as a master server and manage the file system name space, regulate client access to files. As the Name Node is controller and manager of HDFS as it knows the status and the metadata of all the files in HDFS,the metadata information being file permission, names and location of each block. The metadata are small, so it is stored in the memory of name node, allowing faster access to data. The HDFS cluster is accessed by multiple clients concurrently, so all this information is handled by a single machine. The file system operations like opening, closing, renaming etc. are executed by it.

Data Node: It also contains GNU/LINUX operating system and data node software. These nodes manage data storage of their system. They store and retrieve blocks when they are told by client or name node the data node .Files are split into blocks that are managed by the name node and stored by data node. These blocks are replicated across machine at load time so that it can handle fault tolerant. Name node does not directly read or write data. Client interact with the name node to update and retrieve block location. They report back to name node periodically, with list of blocks that they are storing. The data node being commodity hardware also does the work of block creation, deletion and replication with the instruction from the name node.

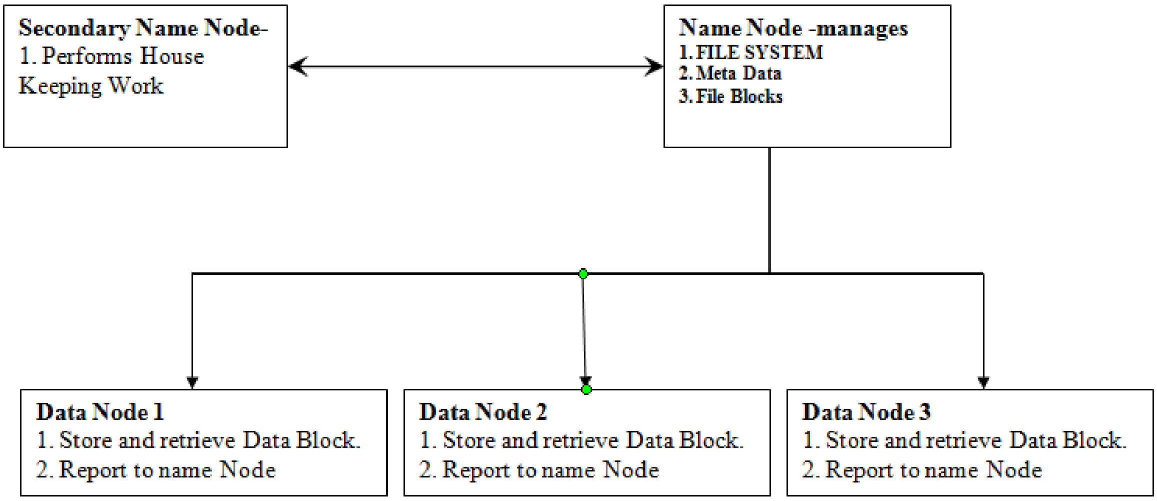


HDFS Read Image



HDFS Write Image

Since all the metadata is stored in name node, it is very important. If it fails the file system can not be used as there would be no way of knowing how to reconstruct the files from blocks present in data node. To overcome this, the concept of secondary name node arises.



HDFS Architecture

Secondary Name Node: It is a separate physical machine which acts as a helper of name node. It performs periodic check points. It communicates with the name node and take snapshot of meta data which helps minimize downtime and loss of data.

2.3.3. Yarn: Yet another Resource Negotiator is used for job scheduling and manages the cluster. Component of yarn that manages the job. **Client:** For submitting Map Reduce jobs. **Resource Manager:** To manage the use of resources across the cluster. **Node Manager:** For launching and monitoring the computer containers on machines in the cluster. **Map Reduce Application Master:** Checks tasks running the Map Reduce job. The application master and the Map Reduce tasks run in containers that are scheduled by the resource manager, and managed by the node managers.

YARN Benefits over Map Reduce: Scalability: Map Reduce 1 hits a scalability bottleneck at 4000 nodes and 40000 tasks, but Yarn is designed for 10,000 nodes and 1 lakh tasks. **Utilization:** Node Manager manages a pool of resources, rather than a fixed number of the designated slots thus increasing the utilization.**Multitenancy:** Different version of Map Reduce can run on YARN, which makes the process of upgrading Map Reduce more manageable

2.3.4. Map Reduce: This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the out of reducer gives the desired result. The Map Reduce is a paradigm which has two phases, the mapper phase and the reducer phase. In the Mapper the input is given in the form of key value pair. The output of the mapper is fed to the reducer as input. The reducer runs only after the mapper is over. The reducer too takes input in key value format and the output of reducer is final output.

How Map Reduce Works. Map takes a data in the form of pairs and returns a list of <key, value> pairs. The keys will not be unique in this case. Using the output of Map, sort and shuffle are applied by the Hadoop architecture. This sort and shuffle acts on these list of <key, value> pairs and sends out unique keys and a list of values associated with this unique key <key, list (values)>.Output of sort and shuffle will be sent to reducer phase. Reducer will perform a defined function on list of values for unique keys and Final output will<key, value> will be stored/displayed. The size of data to be processed decides the number of maps required. For example, we have 1000 MB data and block size is 64 MB then we need 16 mappers. The sort and shuffle occur on the output of mapper and before the reducer. When the mapper task is complete, the results are sorted by key, partitioned if there are multiple reducers, and then written to disk. Using the input from each mapper <k2, v2>, we collect all the values for each unique key k2. This output from the shuffle phase in the form of <k2, list(v2)> is sent as input to reducer phase**.**

3. CONCLUSION

Hadoop is in huge demand in the market. As over the network there is bulk of data but the problem arises is to manage and access the dataset. Hadoop handles such dataset. It have advantages that make it more useful in the market like it deploy on the low cast hardware machine and used by large set of audiences on huge amount of dataset. In this paper I have also discussed the difference between RDBMS and Hadoop. Machine learning algorithm for Big Data need to be more robust and easier to use. Therefore, still betterment is needed in Big Data solution.

4. REFERENCES

[1] V. S. Patil and P. D. Soni,*"HADOOP SKELETON & FAULT TOLERANCE IN HADOOP CLUSTERS "*,International Journal of Application or Innovation in Engineering & Management (IJAIEM) Vol . 2(2), Feb 2013,pp. 247-250

[2] https://www.javatpoint.com/hbase-data-model

[3] Intellipaat. "Hadoop Creator goes to Cloudera". *Intellipaat Blog*. Retrieved 2 February2016.

[4] https://www.tutorialspoint.com/hadoop/

[5] Apache Hadoop (2013). HDFS Architecture Guide [Online]. Available: https://hadoop. apache.org/docs/r1.2.1/hdfs_design.ht