IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138
INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG
An International Open Access, Peer-reviewed, Refereed Journal

BIG DATA ANALYTICS: USE OF HADOOP MAPREDUCE

Mrs. Jadhav Jayshree M.[1], Mrs. Kulkarni Chandraprabha V.[2]
Department Of Information Technology
Rajarshi Shahu Mahavidyalaya, Latur (Autonomous)
[Maharashtra]

ABSTRACT:

Big Data is a huge collection of data that comprises both structured data found in traditional databases and unstructured data like text documents, video and audio. Big Data is not merely data but also a collection of various tools, techniques, frameworks and platforms.. Different sources and the system at various rates are used to generate the data's approach. HADOOP is the popular tool for implementing BIG DATA. HADOOP is an open source technology that enables the distributing process of large data sets of fault tolerance with a very high degree. This paper deals with the technology aspects of BIG DATA for its implementation in organizations by using HADOOP MapReduce technique.

KEYWORDS: BigData , Hadoop, HDFS, Map Reduce

I. INTRODUCTION

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

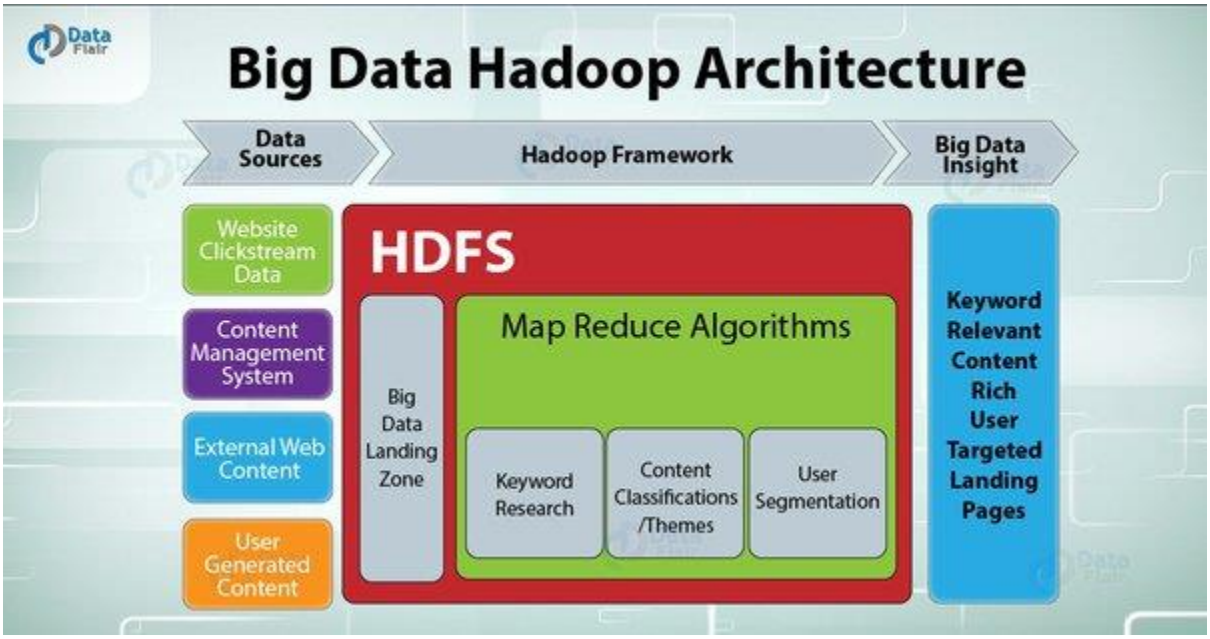
Hadoop is an open source, Java based framework used for storing and processing big data. The data is stored on inexpensive commodity servers that run as clusters. Its distributed file system enables concurrent processing and fault tolerance.

Characteristics of big data	Details
Volume	Organizations have to constantly scale their storage solutions since big data clearly requires large amount of space to be stored.
Velocity	Since big data is being generated every second, organizations need to respond in real time to deal with it.
Variety	Big data comes in variety of forms. It could be structured or unstructured, or even in different formats such as text format, videos, images, and more
Veracity	Big data, as large as it is, can contain wrong data too. Uncertainty of data is something organizations have to consider while dealing with big data.
Value	Just collecting big data and storing it is of no consequence unless the data is analyzed and a useful output is produced.

Big Data is not about the volume of the data, but more about what people use it for. Many organisations like business corporations and educational institutions are using this data to analyse and predict the consequences of certain actions. After collecting the data, it can be used for several functions like:

- Cost reduction
- The development of new products
- Making faster and smarter decisions
- Detecting faults

Architecture of HADOOP:



Today, Big Data is used by almost all sectors including banking, government, manufacturing, airlines and hospitality.

There are many open source software frameworks for storing and managing data, and Hadoop is one of them. It has a huge capacity to store data, has efficient data processing power and the capability to do countless jobs. It is a Java based programming framework, developed by Apache. There are many organisations using Hadoop — Amazon Web Services, Intel, Cloudera, Microsoft, MapR Technologies, Teradata, etc.

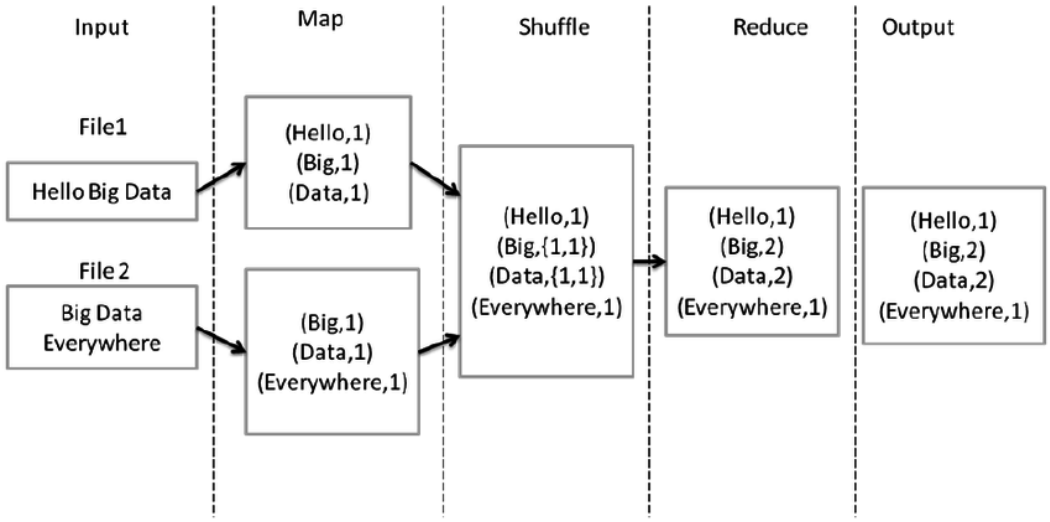
II. TOOLS & TECHNIQUES

A. Map Reduce:

MapReduce is a Hadoop framework used for writing applications that can process vast amounts of data on large databases.Map Reduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data on large clusters of commodity hardware in a reliable, fault-tolerant manner.It is a processing technique and a program model for distributed computing based on java. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples. Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. The name Map Reduce implies, the reduce task is always performed after the map job.

The major advantage of Map Reduce is that it is easy to scale data processing over multiple computing nodes. Under the Map Reduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the Map Reduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the Map Reduce model.

The stages of Map Reduce Program:



Generally Map Reduce paradigm is based on sending the computer to where the data resides! Map Reduce program executes in two stages, namely map stage and reduce stage.

i) Map stage: The map or mapper‟s job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

ii) Shuffle: The process of transferring data from the mappers to reducers is shuffling. It is also the process by which the system performs the sort. Then it transfers the map output to the reducer as input. This is the reason shuffle phase is necessary for the reducers.

iii) Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer‟s job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS

During a Map Reduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster. The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes. Most of the computing takes place on nodes with data on local disks that reduces the network traffic. After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

B. ***HADOOP DISTRIBUTED FILE SYSTEM (HDFS)***

This file system is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems.

HDFS sends data to the server once and uses it as many times as it wants. When a query is raised, NameNode manages all the DataNode slave nodes that serve the given query. Hadoop MapReduce performs all the jobs assigned sequentially.

Working Of HDFS:

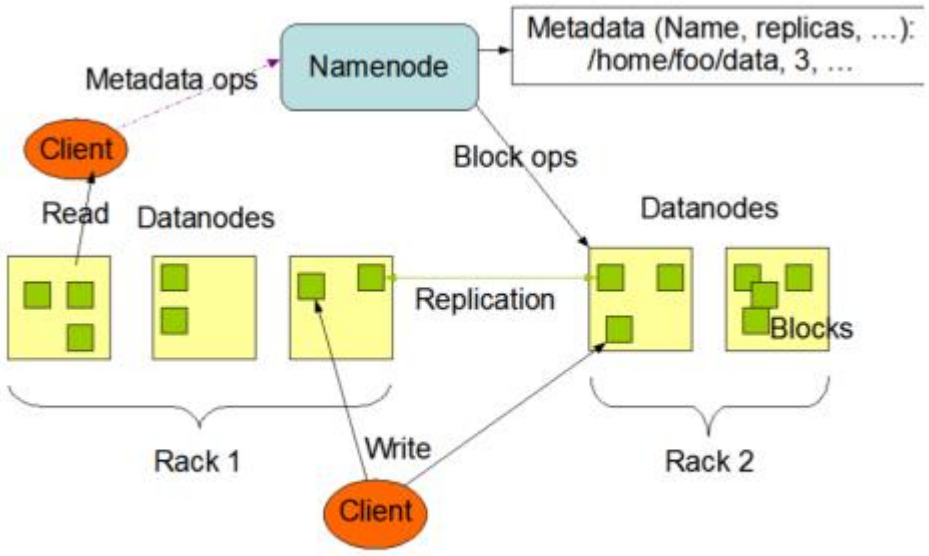


Figure 3. Procedure for storing Data in HDFS

Suppose a client is willing to put 150MB of data in a cluster and sends a request to the NameNode cluster as metadata. Metadata stores the data about the data given by the Client.

- 150MB of data is stored in a file with the file name as file.txt as shown in Figure2.
- The file is divided into 3 input splits a.txt, b.txt, c.txt of each 64MB block size (150MB / 64MB). a.txt – 64MB b.txt – 64MB c.txt – 22MB Figure 2. File.txt input splits
- NameNode responds to the client and requests to store 150MB data in the nodes which has space.
- Client store all the txt files in different DataNodes. However, all the files need not be in sequence order.
- DataNodes are commodity hardware which means if the system goes down the data doesn't lose since HDFS has been given 3 replications by default. Hence it has 2 more backup files for each text files stored in different DataNodes. Hence, the a.txt file occupies 450 MB (150 MB * 3) of files in the whole cluster because of the replication. The same way other text files are also allocated to DataNodes with their corresponding replications. All the DataNodes which are SlaveNodes for that NameNode give proper block report and heartbeat to the NameNode. This acknowledgment gives the information of the condition of the DataNodes. Block report shows the DataNodes are still allocated with some size of block and heartbeat gives the status of the nodes. This is how the data is stored in HDFS.

The importance of Hadoop

Hadoop is capable of storing and processing large amounts of data of various kinds. There is no need to preprocess the data before storing it. Hadoop is highly scalable as it can store and distribute large data sets over several machines running in parallel. This framework is free and uses cost-efficient methods. Hadoop is used for Machine learning,Processing of text documents,Image processing,Processing of XML messages,Web crawling,Data analysis,Analysis in the marketing field,Study of statistical data.

Hadoop has many useful functions like data warehousing, fraud detection and marketing campaign analysis. These are helpful to get useful information from the collected data. Hadoop has the ability to duplicate data automatically. So multiple copies of data are used as a backup to prevent loss of data.

III. CONCLUSION

Big Data is never complete without a mention of Hadoop . Hadoop is the preferred platform for Big Data analytics because of its scalability, low cost and flexibility. It offers an array of tools that data scientists need. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. In this review we also discussed some hadoop components which are used to support the processing of large data sets.

REFERENCES:

1. Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013. 1994 2/13/04
2. Kosha Kothari, Ompriya Kale "Survey of various Clustering Techniques for Big Data in Data Mining" Volume 1, Issue 7, 2014 IJIRT-2349-6002
3. https://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm
4. Etikala, P., Sultana, A., Schmidt, M., Beche, G. D., & Guster, D. (2015). Using Hadoop to Support Big Data Analysis: Security Concerns and Ramifications.
5. https://www-01.ibm.com/software/data/infosphere/hadoop/HIVE/
6. Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. "In: OSDI '04: 6TH Symposium on Operating Systems Design and Implementation (USENIX and ACM SIGOPS), pp.137-150.