HADOOP: The Foundation for Big Data Analytics and Data Science

Shaffy Girdher
Assistant Professor in Computer Science
Panjab University Constituent College, Sikhwala(Sri Muktsar Sahib)

Abstract

In this world of information the term BIGDATA has emerged with new opportunities and challenges to deal with the massive amount of data. BIG DATA has earneda place of great importance and is becoming the choice for new researches. **Apache Hadoop** is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming mode. To find the useful information from massive amoun t of data to organizations, we need to analyze the data.Mastery of data analysis is required to get the information from unstructured data on the web in the form of texts,images, videos or social media posts. This paper gives an introduction to Hadoop and its components and how hadoop has become founding technology for Big data processing, Analytics, and Data Science.

Keywords: Hadoop,Big Data,Data Science,Big Data Analytics

Introduction:*Big data* is a term that describes a *large* volume of structured, semi-structured and unstructured *data* that has the potential to be mined for information and used in machine learning projects and other advanced analytics applications. Big data analytics allows data scientists and various other users to evaluate large volumes of transaction data and other data sources that traditional business systems would be unable to tackle. Hadoop and mapreduce system ,Using the solution provided by Google, **Doug Cutting** and his team developed an Open Source Project called **HADOOP**.Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

Big data: Big data" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.3Vs (volume, variety and velocity) are three defining properties or dimensions of big data. Volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the speed of data processing. According to the 3Vs model, the challenges of big data management result from the expansion of all three properties, rather than just the volume alone -- the sheer amount of data to be managed.

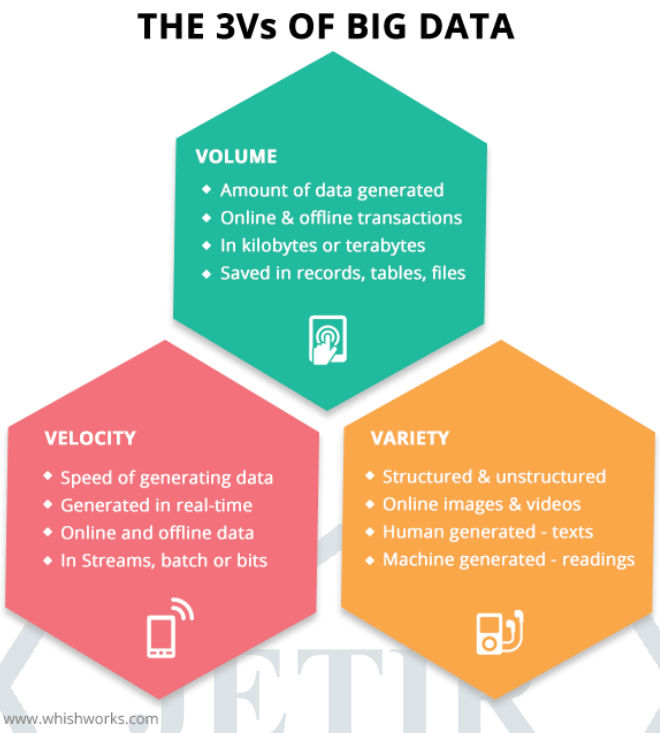

Fig:1 3vs of DATA

Big Data Analytics:

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes. Analysis of big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable. Businesses can use advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing to gain new insights from previously untapped data sources independently or together with existing enterprise data.

Data Science Using (multiple) data elements, in clever ways, to solve iterative data problems that when combined achieve business goals, that might otherwise be intractable.

Fig 2.Data Science

Traditional Approach

In this approach, an enterprise will have a computer to store and process big data. For storage purpose, the programmers will take the help of their choice of database vendors such as Oracle, IBM, etc. In this approach, the user interacts with the application, which in turn handles the part of data storage and analysis.

Limitation

This approach works fine with those applications that process less voluminous data that can be accommodated by standard database servers, or up to the limit of the processor that is processing the data. But when it comes to dealing with huge amounts of scalable data, it is a hectic task to process such data through a single database bottleneck.

Google's Solution

Google solved this problem using an algorithm called **MapReduce**. This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset.Google developed an open source project called Hadoop. Hadoop runs applications using MapReduce algorithm,where data is processed in parallel with others.in short hadoop is used to develop applications that could perform complete statistical analysis on huge amount of data. the hadoop framework application works in environment that provides distributed storage and computation across clusters of computers.Hadoop is designed to scale up from single server to thousand machines,each offering local computation and storage.Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed *storage* and *computation* across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop Architecture

At its core, Hadoop has major layers namely –

- Processing/Computation layer (MapReduce), and
- Storage layer (Hadoop Distributed File System).
- Yarn framework
- Common utilities

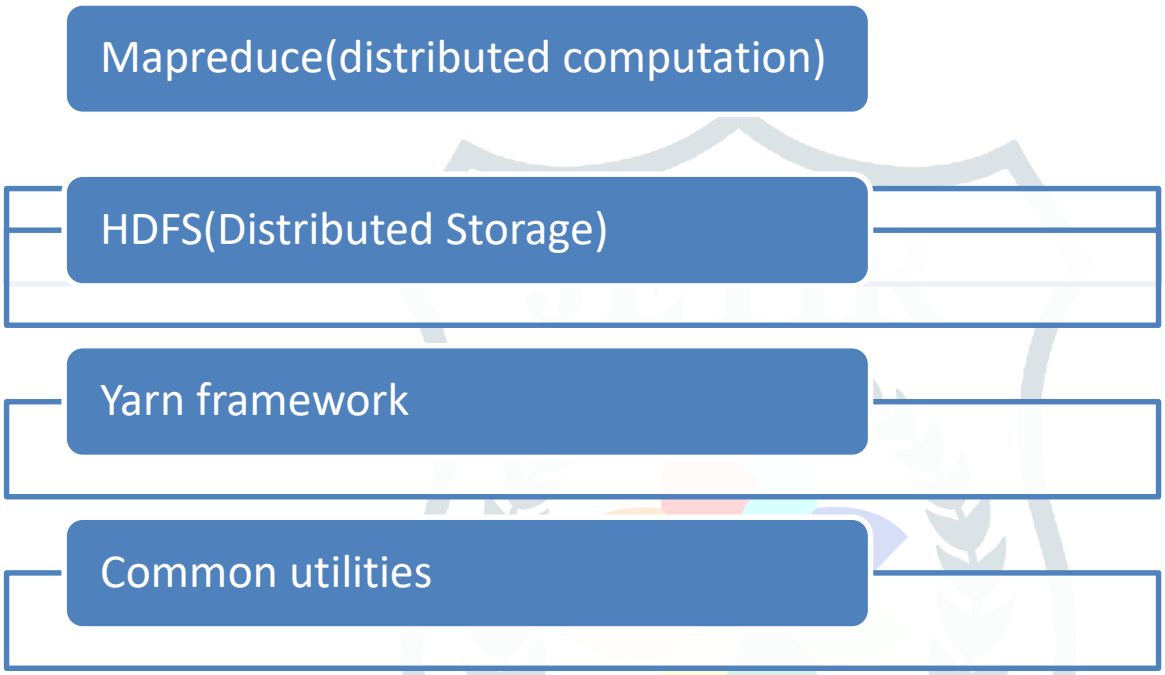

Fig:3 Hadoop Architecture

MapReduce: MapReduce is a core component of the Apache Hadoop software framework.Hadoop enables resilient, distributed processing of massive unstructured data sets across commodity computer clusters, in which each node of the cluster includes its own storage. MapReduce serves two essential functions: it filters and parcels out work to various nodes within the cluster or map, a function sometimes referred to as *the mapper*, and it organizes and reduces the results from each node into a cohesive answer to a query, referred to as *the reducer*.

How MapReduce works

The original version of MapReduce involved several component daemons, including:

- **JobTracker -- the master node that manages all the jobs and resources in a cluster;**
- **TaskTrackers -- agents deployed to each machine in the cluster to run the map and reduce tasks; and**
- **JobHistory Server -- a component that tracks completed jobs and is typically deployed as a separate function or with JobTracker.**

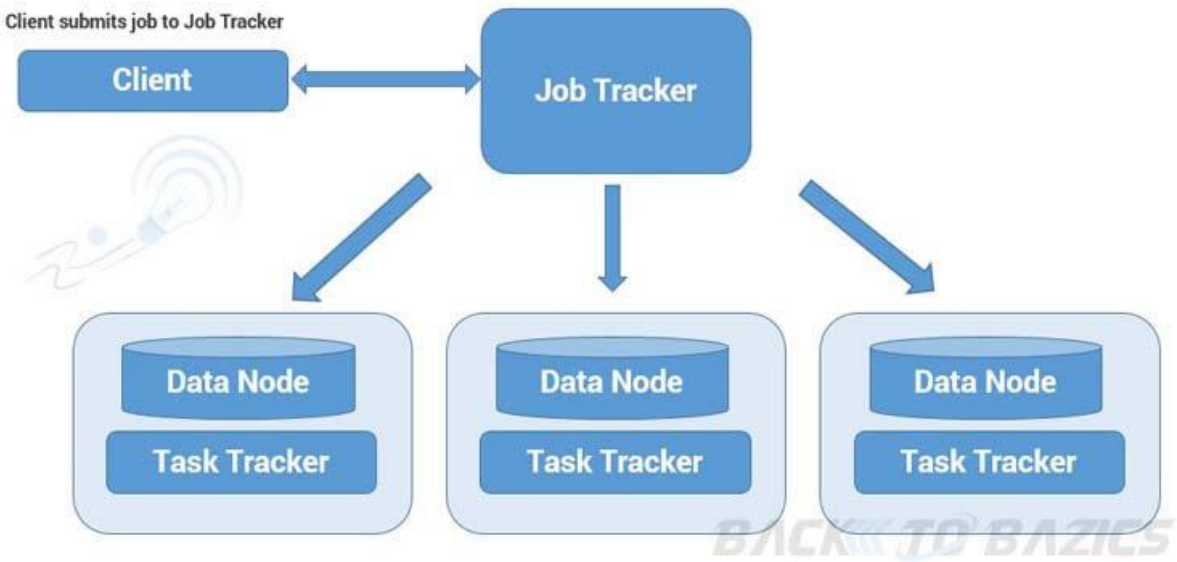

Fig.4 Working Of Map Reduce

HDFS: The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built infrastructure for the Apache Nutch web search engine project. HDFS is now an Apache Hadoop subproject.

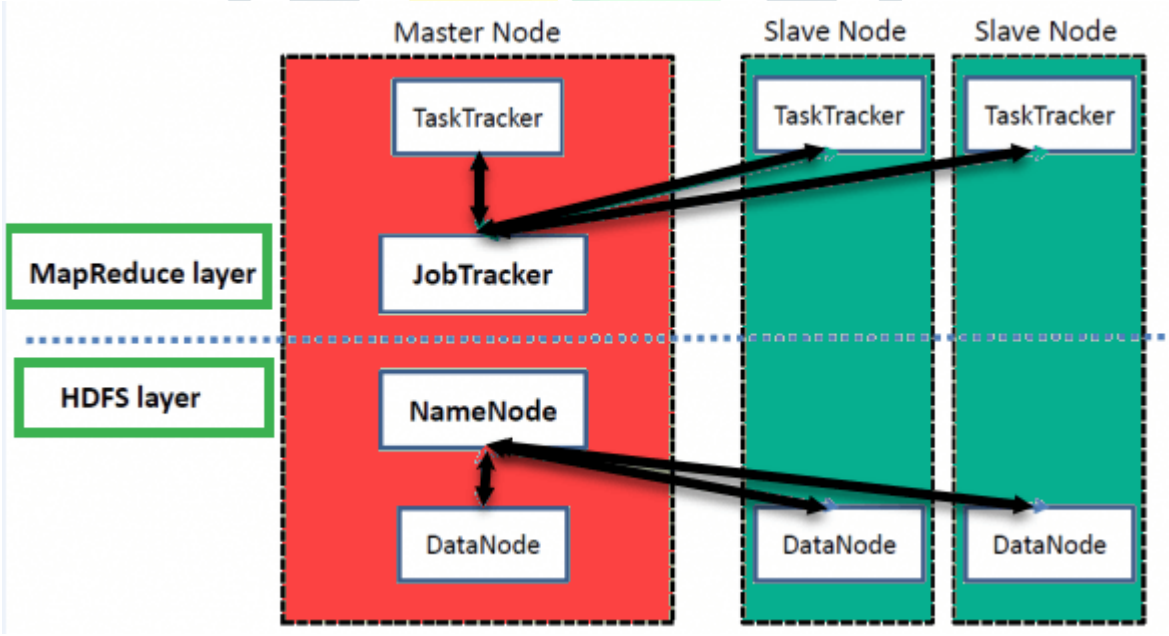

Fig.6 Map Reduce Architecture

Apache Hadoop HDFS Architecture follows a *Master/Slave Architecture*, where a cluster comprises of a single NameNode (Master node) and all the other nodes are DataNodes (Slave nodes). Though one can run several DataNodes on a single machine, but in the practical world these DataNodes are spread across various machines.

NameNode is the master node in the Apache Hadoop HDFS Architecture that maintains and manages the blocks present on the DataNodes (slave nodes). NameNode is a very highly available server that manages the File System Namespace and controls access to files by clients.

DataNode DataNodes are the slave nodes in HDFS. Unlike NameNode, DataNode is a commodity hardware, that is, a non-expensive system which is not of high quality or high-availability. The DataNode is a block server that stores the data in the local file ext3 or ext4.YARN provides open source resource management for Hadoop, so you can move beyond batch processing and open up your data to a diverse set of workloads, including interactive SQL, advanced modeling, and real-time streaming. The YARN-based architecture is not constrained to MapReduce.

YARN: The next generation of Hadoop's Compute platform

Let's slightly change the terminology now. The following name changes give a bit of insight into the design of YARN:

- ResourceManager instead of a cluster manager
- ApplicationMaster instead of a dedicated and short-lived JobTracker
- NodeManager instead of TaskTracker
- A distributed application instead of a MapReduce job
- In the YARN architecture, a global ResourceManager runs as a master daemon, usually on a dedicated machine, that arbitrates the available cluster resources among various competing applications. In YARN, MapReduce is simply degraded to a role of a distributed application (but still a very popular and useful one) and is now called MRv2. MRv2 is simply the re-implementation of the classical MapReduce engine, now called MRv1, that runs on top of YARN. YARN is a completely rewritten architecture of Hadoop cluster. It seems to be a game-changer for the way distributed applications are implemented and executed on a cluster of commodity machines.

Hadoop Common:The Hadoop Common package is considered as the base/core of the framework as it provides essential services and basic processes such as abstraction of the underlying operating system and its file system. Hadoop Common also contains the necessary Java Archive (JAR) files and scripts required to start Hadoop. The Hadoop Common package also provides source code and documentation, as well as a contribution section that includes different projects from the Hadoop Community.

Conclusion

We live in the information era where everything is connected and generates huge amount of data.such data,if well analysed,could aggregate value to society.Hadoop addresses the big data challenges,proving to be effiecient framework of tools.Hadoop is scalable,cost effective,flexible,fast and resilient to failures.now hadoop 2.0 supports the yarn resource manager.because of many such enterprise ready features.hadoop is making news and positive predictions. Hadoop 2.0, NameNode High Availability feature comes with support for a Passive Standby NameNode. Large amount of Data from multiple stores is stored in HDFS but you can only run MapReduce. Hadoop 2.0 provides YARN API's to write other frameworks to run on top of HDFS. YARN provides better resource management in Hadoop, resulting in improved cluster efficiency and application performance. Hadoop usage in other data processing applications.There are various challenges and issues aboutbig data. There must support and encourage fundamental research towards these technical issues if we want to achieve the benefits of big data. Big-data essentially convert operational, financial and commercial problems in aviation that were alreadyunsolvable within economic and human capital constraints using discrete data sets. By centralizing data acquisition and consolidation in the cloud, and by using cloud based virtualization infrastructure to mine data sets efficiently, big-data methods provide new insight into extantdata sets.

REFERENCES

[1]A Research Paper on Big Data and methodology by Shilpa and Manjit Kaur

[2]Review Paper on Use of Big Data in E-Governance of India by Shubham Kalbande, Sumant Deshpande and Prof. Mohit Popat

[3]Review paper on big data and Hadoop by Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar

[4]Big Data in Big Companies by Thomas H. Davenport Jill Dyché

[5]Big Data And Hadoop: A Review Paper by Rahul Beakta CSE Deptt., Baddi University of Emerging Sciences & Technology, Baddi, India

[6]A Survey on Big Data Analysis Techniques by Himanshu Rathod, Tarulata Chauhan

[7]A review on Hadoop —HDFS infrastructure extensions by Kala Karun A, Chitharanjan K

[8]http://www.tutorialspoint.com/Hadoop

[9]/hadoop_big_data_overview.htm

[10]Hadoop Map Reduce for Big Data by Judith Hurwitz, Alan Nugent, Fern Halper and Marcia Kaufman

[11]Suhas V. Ambade and Prof. Priya Deshpande. Hadoop Block Placement Policy for Different File Formats. International Journal of Computer Engineering and Technology, 5(12), 2014, pp. 249-256.

[12]Gandhali Upadhye and Astt. Prof. Trupti Dange. Nephele: Efficient Data Processing Using Hadoop. International Journal of Computer Engineering and