Unit 05 Introduction to Big Data.

Big Data.

Big data refers to large and complex datasets that can not be efficiently managed, processed, and analyzed using traditional data processing techniques.

Types of big data.

1. Structured.
2. Unstructured.
3. Semi-structured.

5 V's of Big data.

Volume: volume refers to the vast amount of data generated and collected. It includes both structured and unstructured data. With the growth of the technology and the internet, Data volumes have significantly increased, leading to storage and processing challenges.

Velocity

Velocity refers to the speed at which data is generated, collected and processed. It refers to the real time or near real time nature of data streams, where the data needs to be analyzed and acted upon quickly. Example includes streaming of the data from social media, sensors financial transactions and more.

Variety

Variety refers to the diversity of the data types and sources. It encompasses structured, unstructured, and semi structured data. The variety of data presents challenges in terms of data integration, management, and analysis.

Veracity

Veracity related to the quality and reliability of data. it refers to the trustworthiness, accuracy and consistency of data.

Value

Value represents the ultimate goal of big data analytics. It refers to the ability to extract meaningful insights patterns, and value from big data.

Challenges

1. Lack of understanding of big data.
2. Dealing with data growth.
3. Confusion while big data tool selection
4. Generating insights in timely manner.
5. Recruiting and retaining big data talent.
6. Integrating disparate data sources.
7. Securing big data.
8. Organizational resistance.

Commonly used tools for Big data.

1. Hadoop
2. Map Reduce programming.
3. HDFS.
4. Spark.
5. Apache Hive.
6. Apache Pig
7. Apache Kafka.
8. HBase.
9. NoSql Database: MongoDB, HBase, Cassandra etc.

Map Reduce.

Map Reduce is a software framework and programming models used for processing huge amount of data. Map reduce program work in two phase, namely map and reduce.

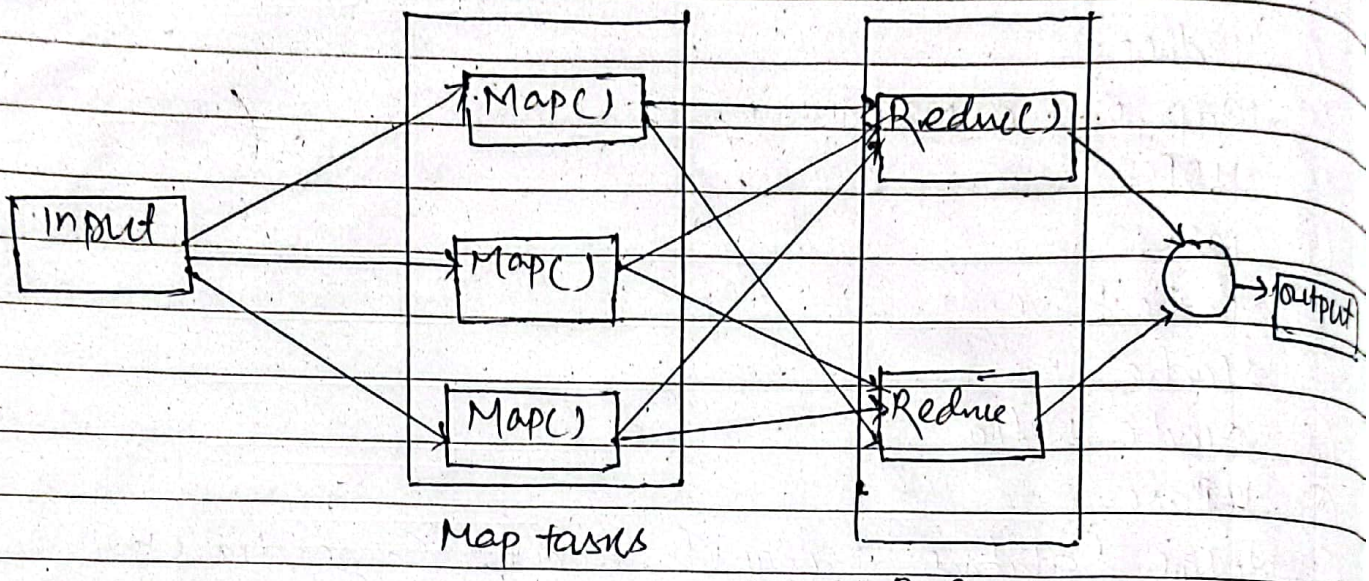
Map

It's tasks deal with splitting and mapping of data. It takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs)

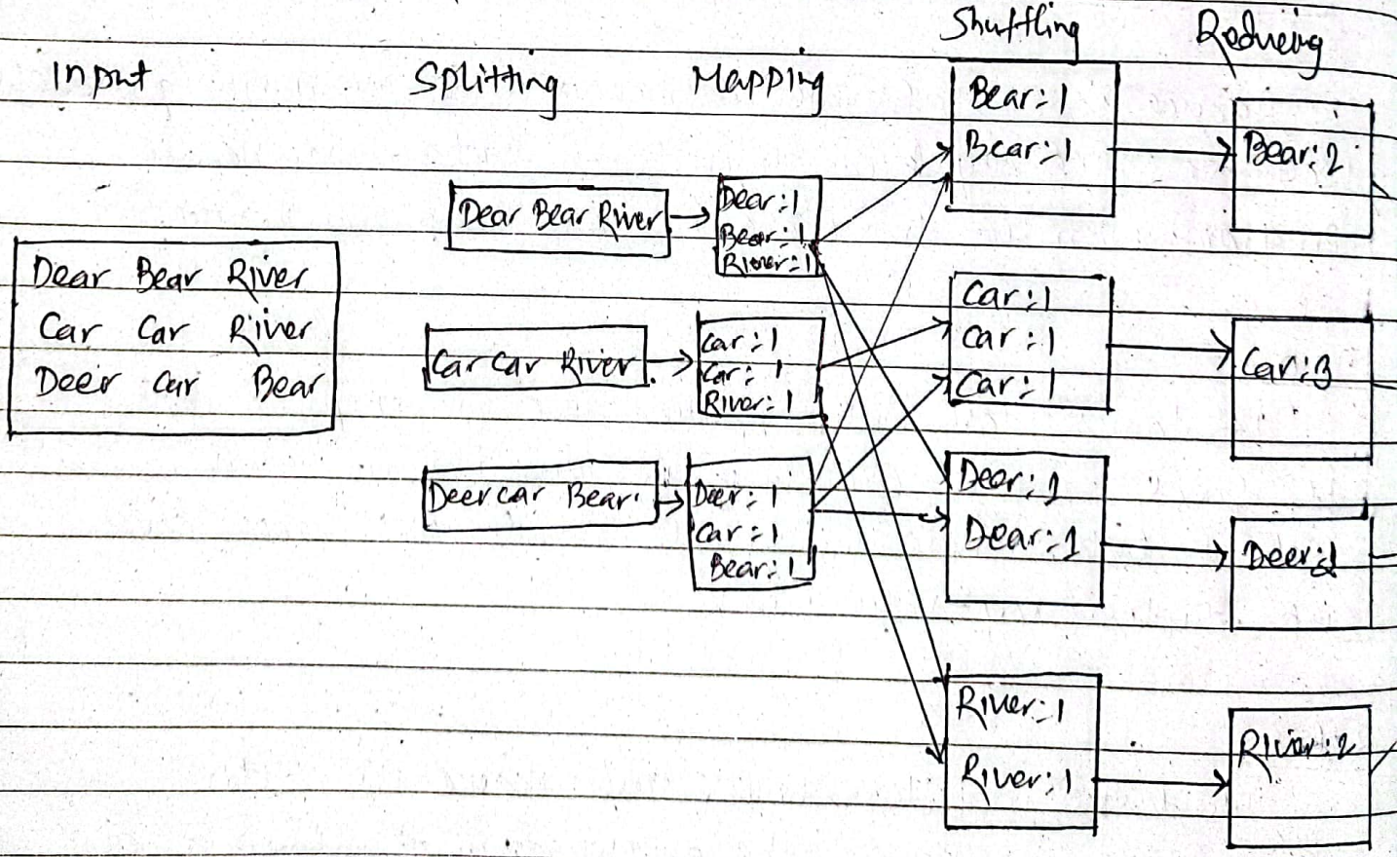
Reduce

It's tasks includes shuffle and reduce the data. The reduce tasks takes the output from the map as an input and combines those data tupples (key-value pairs) into a smaller set of tuples.

Input



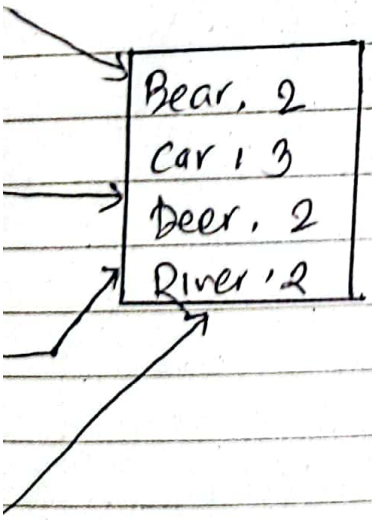
Example:: Map Reduce word count process.



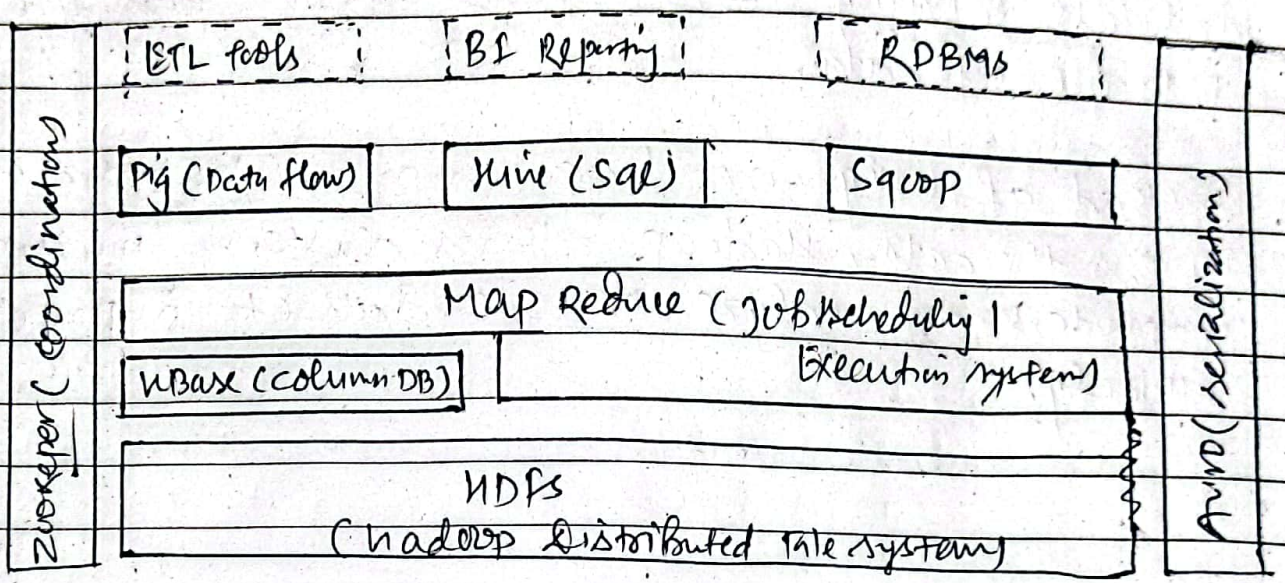
Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging from in size from gigabyte to petabyte of data.

Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel quickly.

Final Result:



Hadoop Ecosystem



Hadoop Ecosystem is shown in figure above.

The Hadoop Ecosystem consists of following components

1. Hadoop common.
2. Hadoop Distributed file system (HDFS)
3. YARN.
4. Hadoop MapReduce.

Hadoop common.

It consists of the common utilities that support other Hadoop modules.

→ It is also known as Hadoop core.

→ ~~It is also known.~~

→ It provides file system and operating system level abstractions

It contains Java archive files and scripts needed to start Hadoop.

YARN:

- It provides the solution to the scalability problems of classical MapReduce.
- It is framework for job scheduling and cluster resource management.
- It enables Hadoop to support more varied processing approach and broader array of applications.

MapReduce

It is method of distributing computations across multiple nodes.
Each node processes the data that is stored at that node.

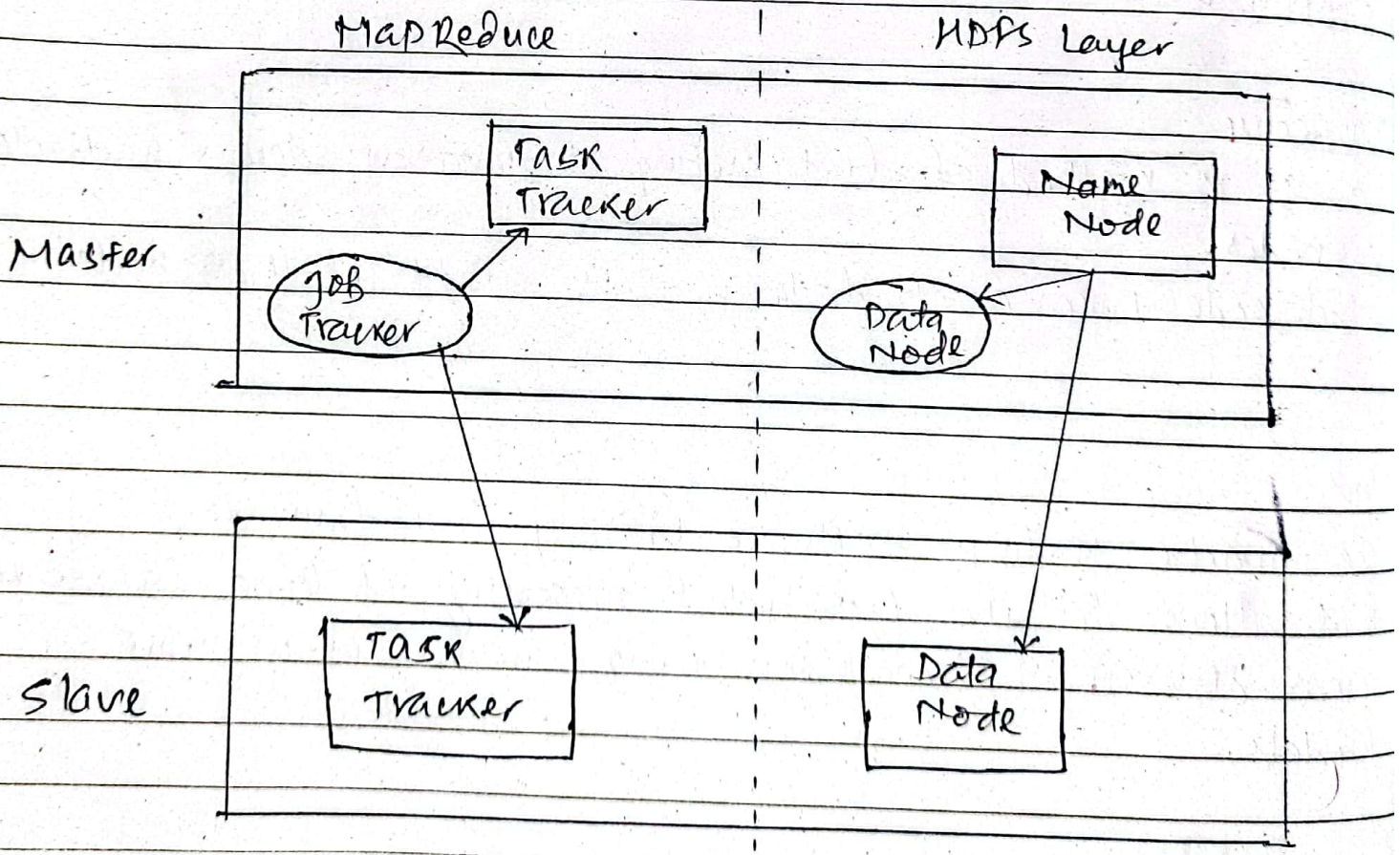
Note

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

Hadoop Architecture

The hadoop architecture is a package of the file system, MapReduce engine and HDFS (Hadoop Distributed file system). The Map reduce engine can be MapReduce/MR1 or YARN/MR2.

A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes job tracker, Task Tracker, NameNode and Datanode whereas the slave node includes Datanode and Task Tracker.

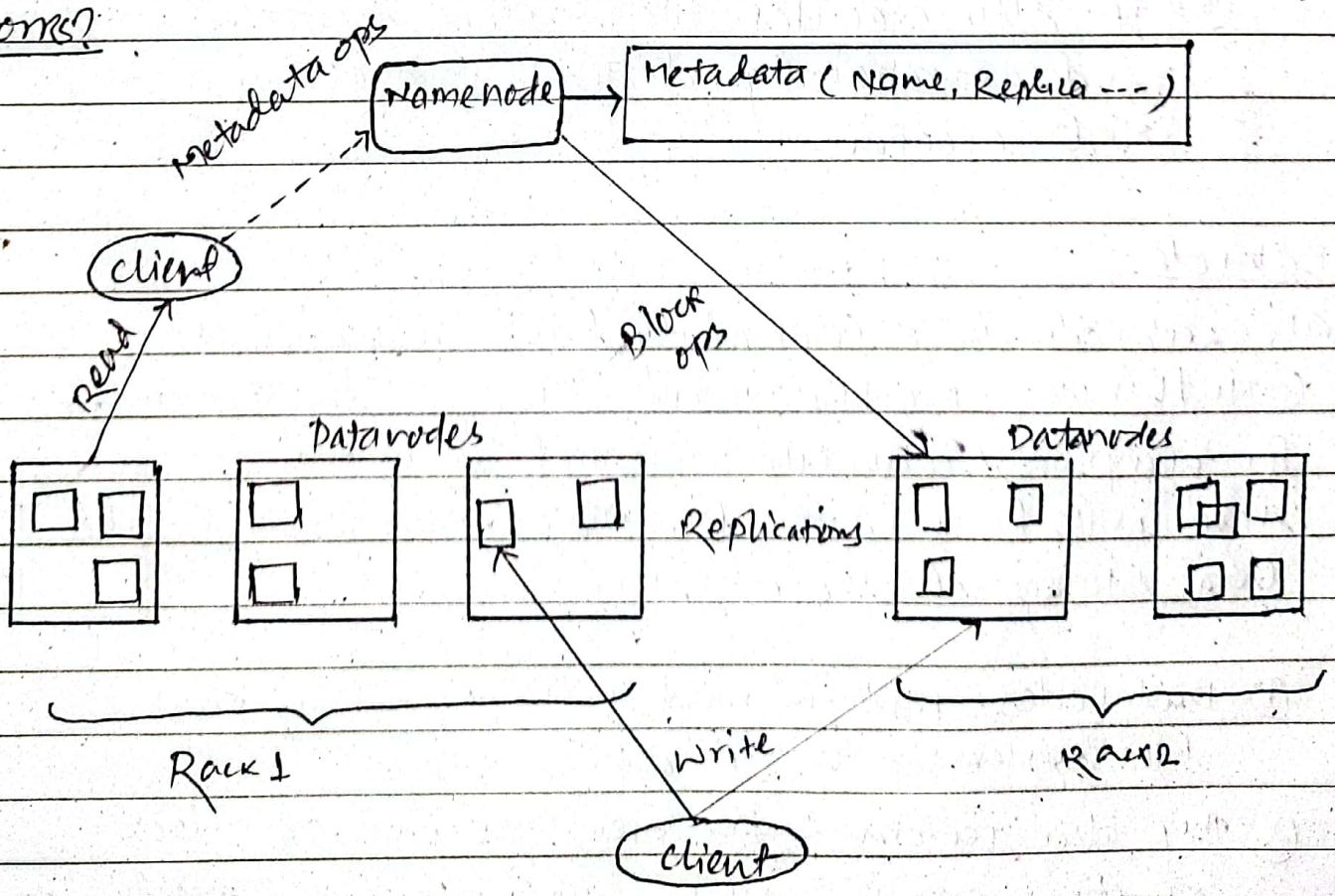


HDFS

The Hadoop distributed file system (HDFS) is the primary data storage system used by Hadoop applications. HDFS employs a NameNode and DataNode architecture to implement a distributed filesystem that provides high performance access to data across highly scalable Hadoop clusters.

Hadoop itself is an open source distributed processing framework that manages data processing and storage for big data applications. HDFS is a key part of the many Hadoop ecosystem technologies. It provides a reliable means for managing pools of big-data and supporting related big data analytics applications.

Works?



HDFS follows the master-slave architecture and it has the following elements.

NameNode

The Namenode is commodity hardware that contains the GNU/LINUX operating system and the namenode software. It is a software that can be run on commodity hardware.

The system having namenode acts as the master server and does the following tasks.

① Manages the file system namespace.
② Regulates client access to file
③ It also executes file system operations such as renaiming, closing, and opening files and directions

Datanode

The datanode is a commodity hardware having the GNU/Linux operating system and datanode software. For every node (commodity hardware/system) in a cluster. there will be a datanode. These nodes manages the data storage of their system

① Datanodes perform read-write operations on the file systems, as per client request
② They also perform ~~same~~ operations such as block ~~operations~~ creation, deletion, and replication according to the instructions of namenode.

Blocks.

Blocks are nothing but the smallest continuous location on your hard drive where data is stored.

In general, in any of the file system, you store the data as a collection of blocks. Similarly, HDFS stores each file as blocks which are scattered throughout the Apache Hadoop cluster.