UNIT- II

Introduction to Hadoop

__/__/____

L A Lalitha (17)

Hadoop:

→ Hadoop is fundamentally infrastructure software for storing and processing large dataset

→ It is an open source project under apache.

To understand Hadoop, we should understand two fundament things about Hadoop
- How it stores the data
- How it processes the data

1. Storing Data

Hadoop is a cluster system, HDFS (Hadoop Destributed file System) as part of Hadoop.

Imagine we have data which can't be accomodated in our PC. Hadoop lets you store data that can't be stored in a single PC/node it also lets you store many files. (because it has multiple nodes of servers, eg: 2 nodes, 20 nodes, thousands of nodes ...)

2. Framework for Processing Data - MapReduce.

The MapReduce is clearly explained in the following figure.

MapReduce is done is several levels. like
Input
Splitting
Mapping
Shuffling
Reducing
Final Result

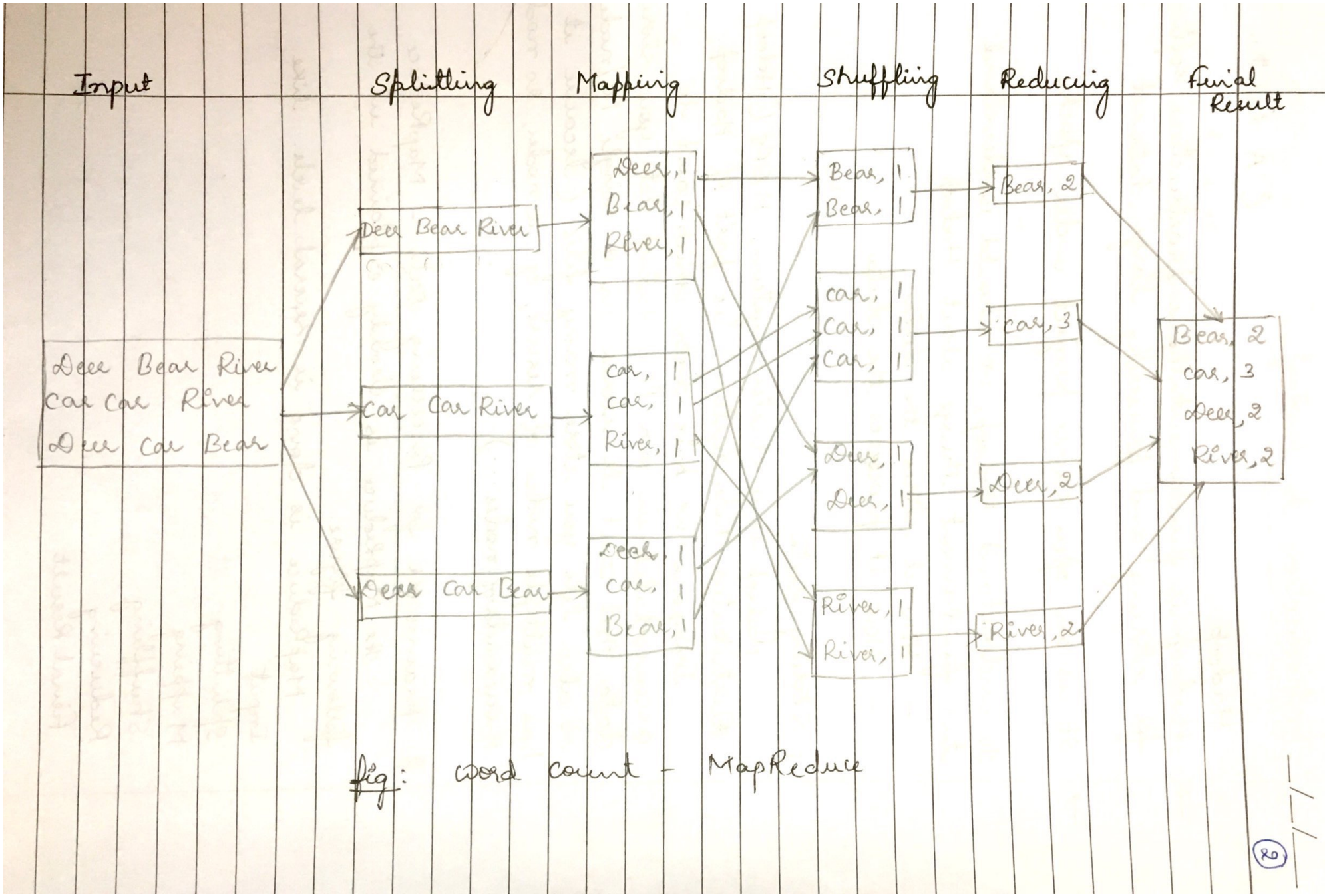

fig: Word count - MapReduce

Hadoop is fundamentally a batch processing environment. It is not good for adhoc queries.

Hadoop Cluster: It is a collection of different racks. Racks are collection of different nodes (30-40 nodes)

Eg:

Rack - 1	Rack - 2	....	Rack n
Node 1	Node 1		Node 1
Node 2	Node 2		Node 2
⋮	⋮		⋮
Node 3	Node 3		Node 3

A rack is a collection of 30-40 nodes that are physically placed close together and all connected to the same network switch.

Hadoop has two main components:

- Destributed File System - HDFS
- MapReduce Engine.

HDFS:

- Hadoop File System that runs on top of existing file system.
- Designed to handle very large files with streaming data access patterns
- Uses blocks to store a file or parts of a file

__/__/____

L A Lalitha (A)

Why Hadoop?

Its capability to handle massive amounts of data, different categories of data - fairly quickly.

The other considerations are,

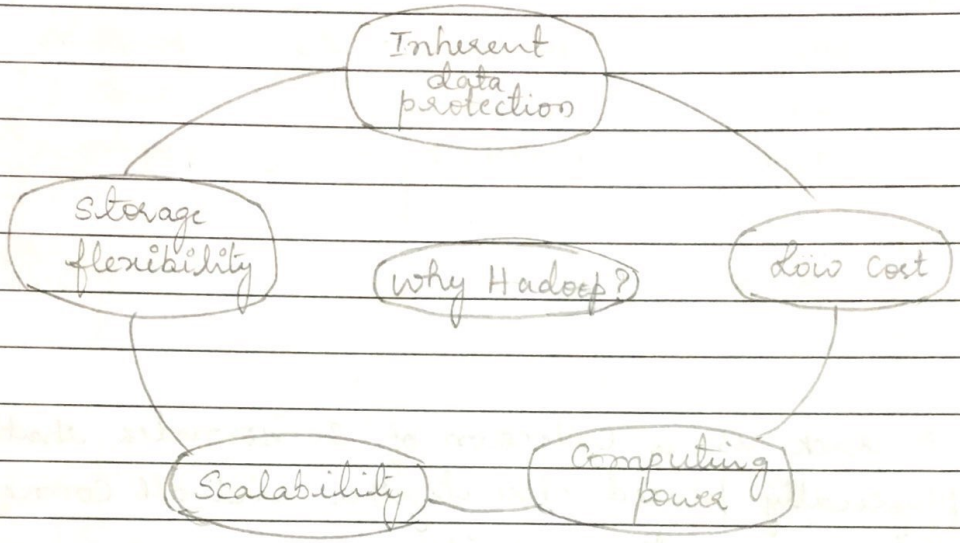

fig: Key Considerations of Hadoop.

1. Low-Cost:

Hadoop is an open-source framework and uses commodity hardware (commodity hardware is relatively inexpensive and easy to obtain hardware) to store enormous quantities of data.

2. Computing power:

Hadoop is based on distributed computing model which processes very large volumes of data fairly quickly. The more the number of computing nodes, the more the processing power at hand.

3. Scalability:

This boils down to simply adding nodes as the

system grows and requires much less administration.

4. Storage flexibility:

Unlike the traditional relational databases, in Hadoop data need not be pre-processed before storing it. Hadoop provides the convenience of storing as much data as one needs and also the added flexibility of deciding later as to how to use the stored data. In Hadoop, one can store unstructured data like images, videos, and free-form text.

5. Inherent data protection:

Hadoop protects data and executing applications against hardware failure. If a node fails, it automatically redirects the jobs that had been assigned to this node to the other functional and available nodes and ensures that distributed computing doesnot fail. It goes a step further to store multiple copies (replicas) of the data on various nodes across the cluster

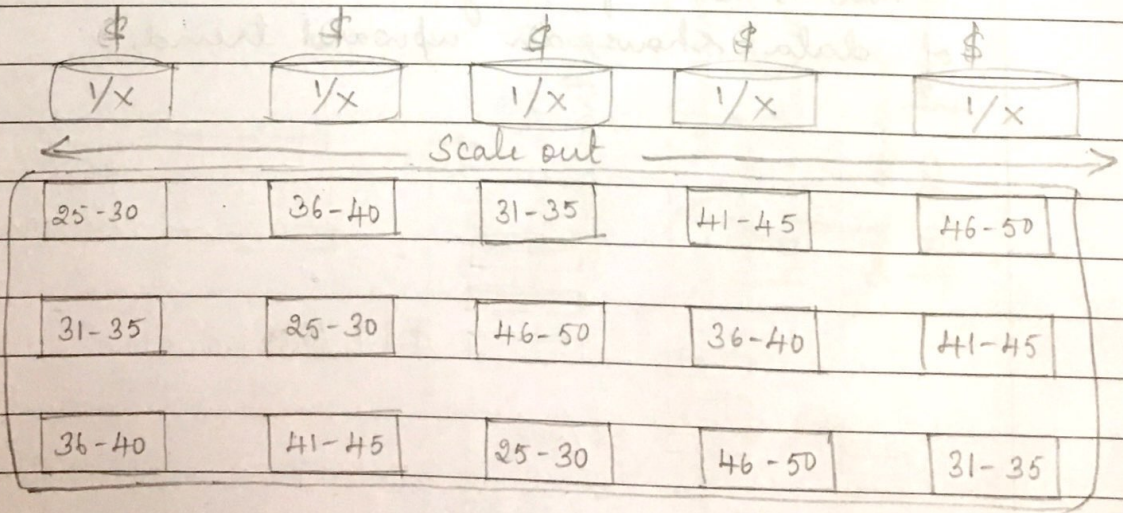

fig: Hadoop framework (DFS, Commodity hardware)

Hadoop makes use of commodity hardware, DFS, and distributed computing as shown in the above figure. In this new design, groups of machine are gathered together; it is known as a cluster

With this new paradigm, the data can be managed with Hadoop as follows:

1. Distributes the data and duplicates chunks of each data file across several nodes,
   Eg:- 25-30 is one chunk of data in the above fig.
2. Locally available compute resource is used to process each chunk of data in parallel.
3. Hadoop framework handles failover smartly and automatically.

Why not RDBMS?

RDBMS is not suitable for storing and processing large files, images and videos.

RDBMS is not a good choice when it comes to advanced analytics involving machine learning.

RDBMS call for huge investment as the volume of data shows an upward trend.

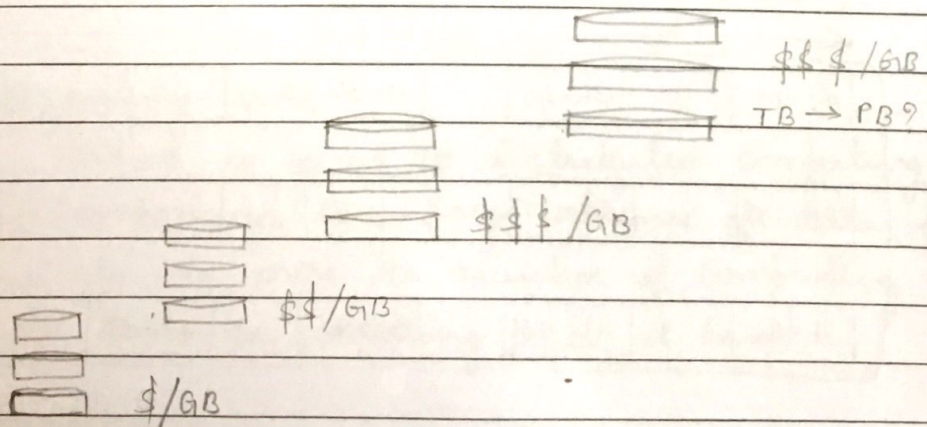

fig: RDBMS with respect to cost/GB of storage.

RDBMS Versus HADOOP

Parameters	RDBMS	HADOOP
System	Relational Database Management System	Node based flat structure
Data	Suitable for structured data	Suitable for unstructured data, structured data. Supports variety of data formats in real time such as XML, JSON, text based flat file formats etc.
Processing choice	OLTP when the data needs consistent relationship	Analytical, Big Data Processing Big Data processing, which does not require any consistent relationships between data
Processor	Needs expensive hardware or high-end processors to store huge volumes of data	In a Hadoop cluster, a node requires only a processor, a network card, a few hard drives
cost	Cost around $10,000 to $14,000 per TB of storage	cost around $4,000 per TB of storage

L A Lakshmi (7)

L A Lalitha (8)

Hadoop Overview

Open source software framework to store & process massive amounts of data in a distributed fashion on large clusters of commodity hardware

Basically Hadoop accomplishes two tasks:

1) Massive data storage
2) Faster data processing

1. Key Aspects of Hadoop

→ **Open Source Software:**

It is free to download, use & contribute to.

→ **Framework:**

Means everything that you will need to develop and execute and application is provided. - programs, tools, etc.

→ **Distributed:**

Divides and stores data across multiple computers. Computation / Processing is done in parallel across multiple connected nodes.

→ **Massive Storage:**

Store colossal amounts of data across nodes of low-cost commodity hardware.

→ **Faster Processing:**

Large amounts of data is processed in parallel, yielding quick response.

__/__/____

L A Lab Thr (9)

2. Hadoop Components

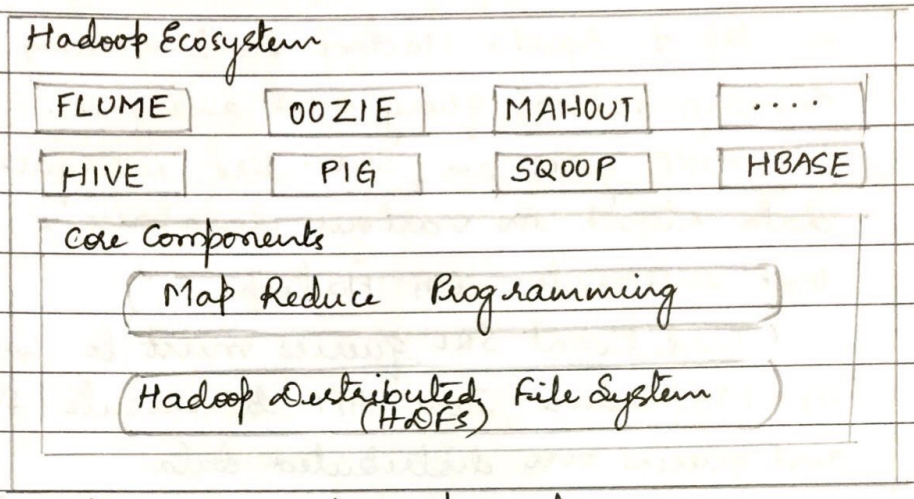

fig: Hadoop Components

Hadoop Core Components

1) HDFS:
   a) Storage Component
   b) Distributes data across several nodes
   c) Natively redundant
2) MapReduce:
   a) Computational framework
   b) Splits task across multiple nodes
   c) Processes data in Parallel.

Hadoop Ecosystem

Hadoop Ecosystem are support projects to enhance the functionality of Hadoop Core Components. The Eco projects are as follows:

1. HIVE
2. PIG
3. SQOOP
4. HBASE
5. FLUME
6. OOZIE
7. MAHOUT

3. Hadoop Conceptual Layer

It is conceptually divided into Data Storage Layer which stores huge volumes of data and Data Processing Layer which processes data in parallel to extract richer and meaningful insights from data

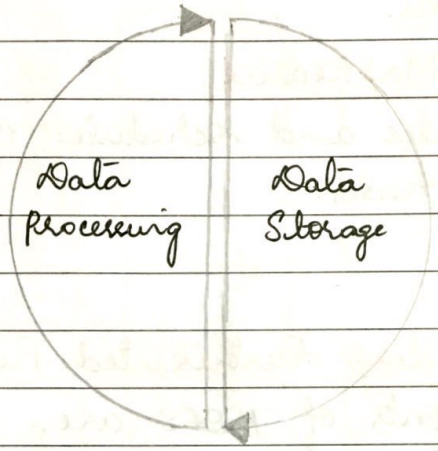

fig: Hadoop Conceptual layer

4. High-Level Architecture of Hadoop

Hadoop is a distributed Master-Slave Architecture. Master Node is known as NameNode and Slave Nodes are known as DataNodes

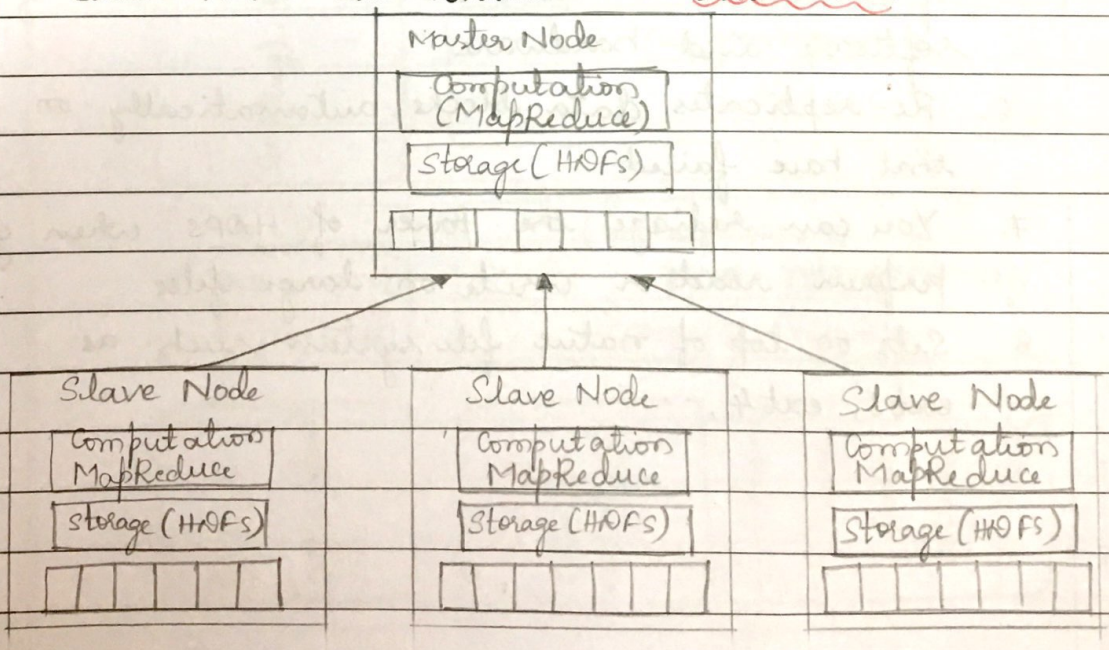

fig: Hadoop high-level architecture

Key components of the Master Node:

1. Master HDFS:

Its main responsibility is partitioning the data storage across the slave nodes. It also keeps track of locations of data on DataNodes.

2. Master MapReduce:

It decides and schedules computation task on slave nodes.

HDFS (Hadoop Distributed File System)

Some key points of HDFS are,

1. Storage Component of Hadoop
2. Distributed File System
3. Modeled after Google File System
4. Optimized for high throughput
5. You can replicate a file for a configured number of times, which is tolerant in terms of both software and hardware
6. Re-replicates data blocks automatically on nodes that have failed
7. You can realize the power of HDFS when you perform read or write on large files
8. Sits on top of native file system such as ext3, ext4, ..