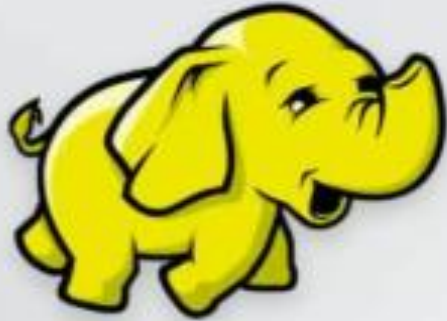# BIG-DATA HADOOP

BY BHOOMIKA R

4NI22CS038

What is Hadoop?

- Hadoop is an open source software programming framework for storing a large amount of data and performing the computation.

- Its framework is based on Java programming with some native code in C and shell scripts.

- Hadoop is used for some advanced level of analytics, which includes Machine Learning and data mining

Need for Hadoop

- Redundant, Fault-tolerant data storage
- Parallel computation framework
- Job coordination

Programmers

No longer need to worry about

Q: Where file is located?

Q: How to handle failures & data lost?

Q: How to divide computation?

Q: How to program for scaling?

History of Hadoop

- **Apache Software Foundation** is the developers of Hadoop, and it's co-founders are **Doug Cutting** and **Mike Cafarella**.

- It's co-founder Doug Cutting named it on his son's toy elephant. In October 2003 the first paper release was Google File System.

- In January 2006, MapReduce development started on the Apache Nutch which consisted of around 6000 lines coding for it and around 5000 lines coding for HDFS.
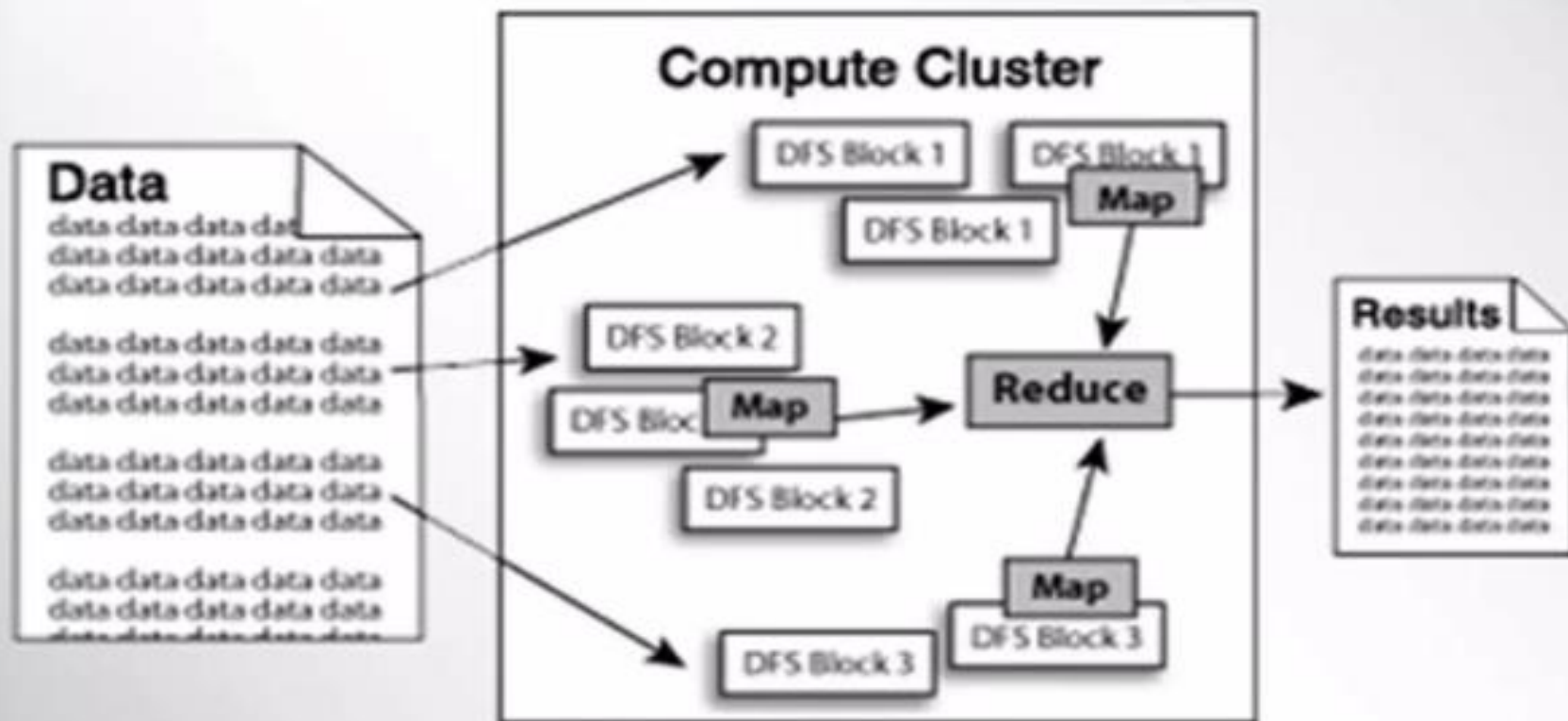
- In April 2006 Hadoop 0.1.0 was released.

Advantages and Disadvantages of Hadoop
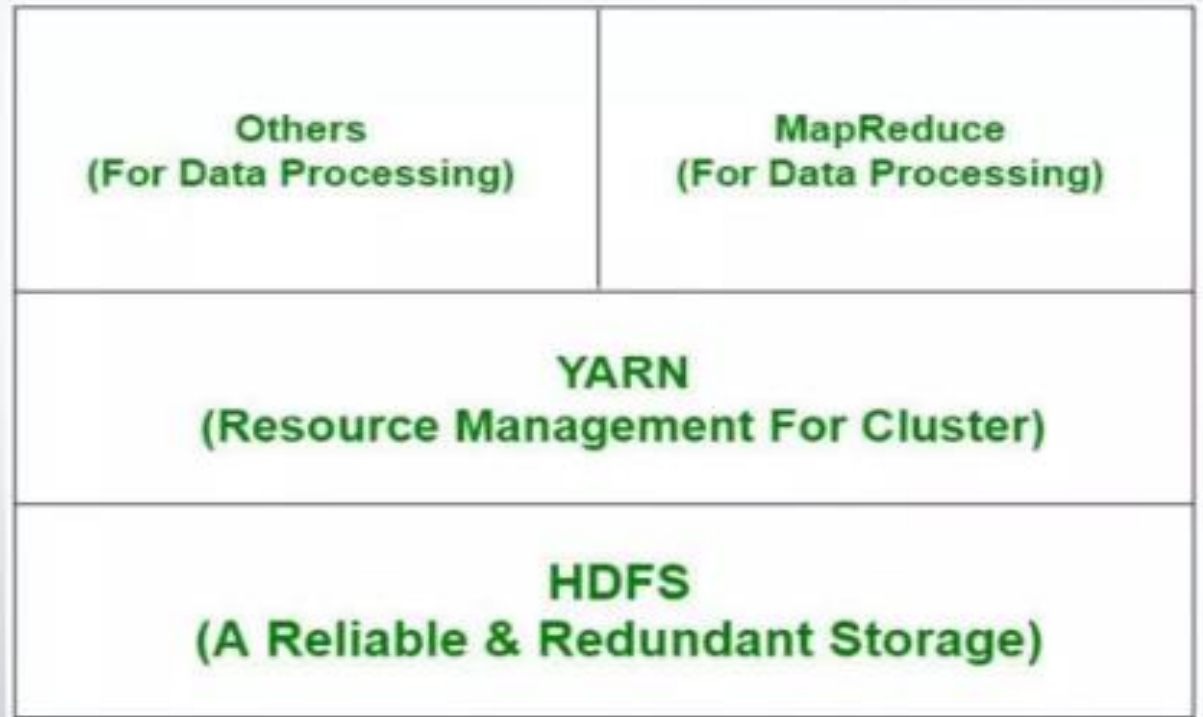
Advantages:

- Ability to store a large amount of data.

- High flexibility.

- Cost effective.

- High computational power.

- Tasks are independent.

- Linear scaling.

Disadvantages:

- Not very effective for small data.

- Hard cluster management.

- Has stability issues.

- Security concerns.

Hadoop Distributed File System

- It has distributed file system known as HDFS and this HDFS splits files into blocks and sends them across various nodes in form of large clusters.

- Also in case of a node failure, the system operates and data transfer takes place between the nodes which are facilitated by HDFS.

Others (For Data Processing)	MapReduce (For Data Processing)
YARN **(Resource Management For Cluster)**	
HDFS **(A Reliable & Redundant Storage)**	

Comparing: RDBMS vs. Hadoop

	Traditional RDBMS	Hadoop / MapReduce
Data Size	Gigabytes (Terabytes)	Petabytes (Hexabytes)
Access	Interactive and Batch	Batch – NOT Interactive
Updates	Read / Write many times	Write once, Read many times
Structure	Static Schema	Dynamic Schema
Integrity	High (ACID)	Low
Scaling	Nonlinear	Linear
Query Response Time	Can be near immediate	Has latency (due to batch processing)

Advantages of HDFS:

- It is inexpensive, immutable in nature, stores data reliably, ability to tolerate faults, scalable, block structured, can process a large amount of data simultaneously and many more.

Disadvantages of HDFS:

- It's the biggest disadvantage is that it is not fit for small quantities of data. Also, it has issues related to potential stability, restrictive and rough in nature.

Some common frameworks of Hadoop

- **Hive**- It uses HiveQl for data structuring and for writing complicated MapReduce in HDFS.

- **Drill**- It consists of user-defined functions and is used for data exploration.

- **Storm**- It allows real-time processing and streaming of data.

- **Spark**- It contains a Machine Learning Library(MLlib) for providing enhanced machine learning and is widely used for data processing. It also supports Java, Python, and Scala.

- **Pig**- It has Pig Latin, a SQL-Like language and performs data transformation of unstructured data.

- **Tez**- It reduces the complexities of Hive and Pig and helps in the running of their codes faster.

Modules of Hadoop frameworks

Hadoop framework is made up of the following modules:

1. **Hadoop MapReduce-** a MapReduce programming model for handling and processing large data.

2. **Hadoop Distributed File System-** distributed files in clusters among nodes.

3. **Hadoop YARN-** a platform which manages computing resources.

4. **Hadoop Common-** it contains packages and libraries which are used for other modules.
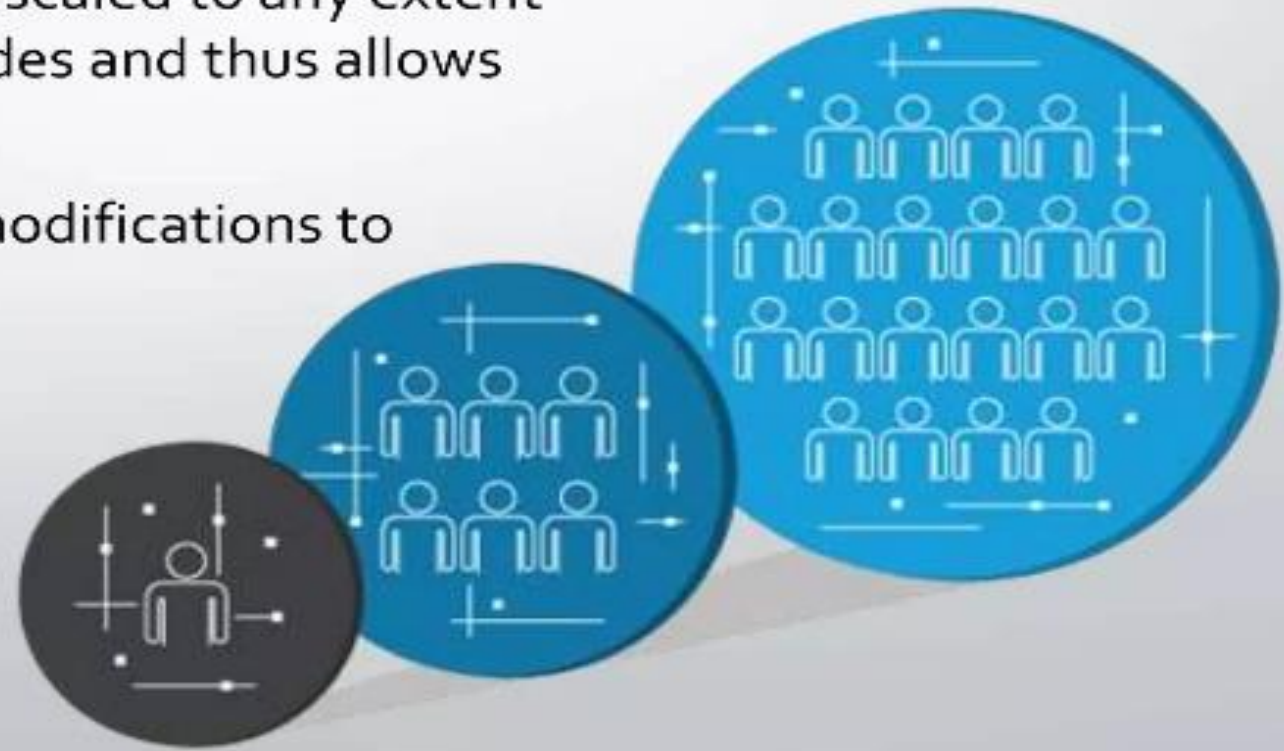
Features of 'Hadoop'

Suitable for Big Data Analysis

- As Big Data tends to be distributed and unstructured in nature, HADOOP clusters are best suited for analysis of Big Data.

- Since it is processing logic (not the actual data) that flows to the computing nodes, less network bandwidth is consumed.

- This concept is called as **data locality concept** which helps increase the efficiency of Hadoop based applications.

Scalability

- HADOOP clusters can easily be scaled to any extent by adding additional cluster nodes and thus allows for the growth of Big Data.

- Also, scaling does not require modifications to application logic.

FAULT TOLERANCE

- HADOOP ecosystem has a provision to replicate the input data on to other cluster nodes.

- That way, in the event of a cluster node failure, data processing can still proceed by using data stored on another cluster node.

Hadoop Analytics Tools

- There is a wide range of analytical tools available in the market that help Hadoop deal with the astronomical size data efficiently.

- Let us discuss some of the most famous and widely used tools one by one. Below are the top 10 Hadoop analytics tools for big data.

APACHE Spark™

- Apache spark in an open-source processing engine that is designed for ease of analytics operations.

- It is a cluster computing platform that is designed to be fast and made for general purpose uses.

- Spark is designed to cover various batch applications, Machine Learning, streaming data processing, and interactive queries.

Features of Spark:

- In memory processing

- Tight Integration Of component

- Easy and In-expensive

- The powerful processing engine makes it so fast

- Spark Streaming has high level library for streaming process

- MapReduce is just like an Algorithm or a data structure that is based on the YARN framework.
- The primary feature of MapReduce is to perform the distributed processing in parallel in a Hadoop cluster, which Makes Hadoop working so fast Because when we are dealing with Big Data, serial processing is no more of any use.

Features of Map-Reduce:

- Scalable
- Fault Tolerance
- Paraller Processing
- Tunable Replication
- Load Balancing

APACHE HIVE

- Apache Hive is a Data warehousing tool that is built on top of the Hadoop, and Data Warehousing is nothing but storing the data at a fixed location generated from various sources.

- Hive is one of the best tools used for data analysis on Hadoop.

- The one who is having knowledge of SQL can comfortably use Apache Hive.

- The query language of high is known as HQL or HIVEQL.

Features of Hive:

- Queries are similar to SQL queries.

- Hive has different storage type HBase, ORC, Plain text, etc.

- Hive has in-built function for data-mining and other works.

- Hive operates on compressed data that is present inside Hadoop Ecosystem.

APACHE mahout

- The name *Mahout* is taken from the Hindi word **Mahavat** which means the elephant rider.

- Apache Mahout runs the algorithm on the top of Hadoop, so it is named Mahout.

- Mahout is mainly used for implementing various Machine Learning algorithms on our Hadoop like classification, Collaborative filtering, Recommendation.

- Apache Mahout can implement the Machine algorithms without integration on Hadoop.

Features of Mahout:

- Used for Machine Learning Application

- Mahout has Vector and Matrix libraries

- Ability to analyze large datasets quickly

Apache Pig

- This Pig was Initially developed by Yahoo to get ease in programming.
- Apache Pig has the capability to process an extensive dataset as it works on top of the Hadoop.
- Apache pig is used for analyzing more massive datasets by representing them as dataflow.
- Apache Pig also raises the level of abstraction for processing enormous datasets.
- Pig Latin is the scripting language that the developer uses for working on the Pig framework that runs on Pig runtime.

Features of Pig:

- Easy To Programme
- Rich set of operators
- Ability to handle various kind of data
- Extensibility

APACHE HBASE

- HBase is nothing but a non-relational, NoSQL distributed, and column-oriented database. HBase consists of various tables where each table has multiple numbers of data rows.

- These rows will have multiple numbers of column family's, and this column family will have columns that contain key-value pairs.

- HBase works on the top of HDFS(Hadoop Distributed File System).

- We use HBase for searching small size data from the more massive datasets.

Features of HBase:

- HBase has Linear and Modular Scalability

- JAVA API can easily be used for client access

- Block cache for real time data queries

APACHE SQOOP

- Sqoop is a command-line tool that is developed by Apache.

- The primary purpose of Apache Sqoop is to import structured data i.e., RDBMS(Relational database management System) like MySQL, SQL Server, Oracle to our HDFS(Hadoop Distributed File System).

- Sqoop can also export the data from our HDFS to RDBMS.

Features of Sqoop:

- Sqoop can Import Data To Hive or HBase

- Connecting to database server

- Controlling parallelism

APACHE STORM™

- Apache Storm is a free open source distributed real-time computation system build using Programming languages like Clojure and java.

- It can be used with many programming languages.

- Apache Storm is used for the Streaming process, which is very faster.

- We use Daemons like Nimbus, Zookeeper, and Supervisor in Apache Storm.

- Apache Storm can be used for real-time processing, online Machine learning, and many more. Companies like Yahoo, Spotify, Twitter, and so many uses Apache Storm.

Features of Storm:

- Easily operatable

- each node can process millions of tuples in one second

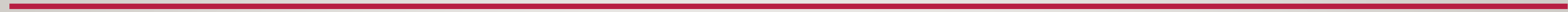- Scalable and Fault Tolerance

Companies Using Hadoop

Common Hadoop Distributions

- Open Source
  - Apache

- Commercial
  - Cloudera
  - Hortonworks
  - MapR
  - AWS MapReduce
  - Microsoft Azure HDInsight (Beta)

THANK YOU