HADOOP

Hadoop Overview

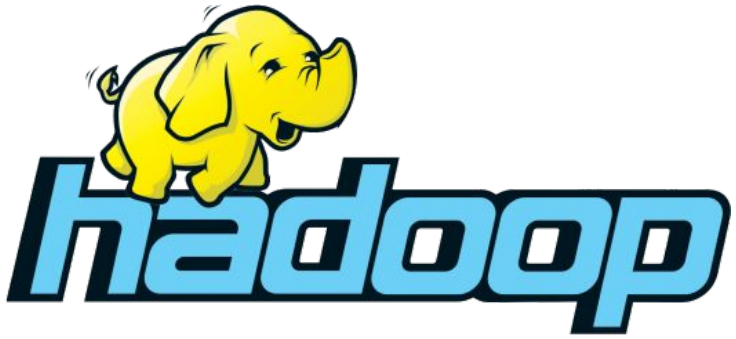
- Introduction to Hadoop
- RDBMS vs Hadoop
- Key aspects of Hadoop
- Hadoop Components
- High-level Architecture of Hadoop

What is Hadoop?

The Technology that empowers Yahoo, Facebook, Twitter, Walmart and others

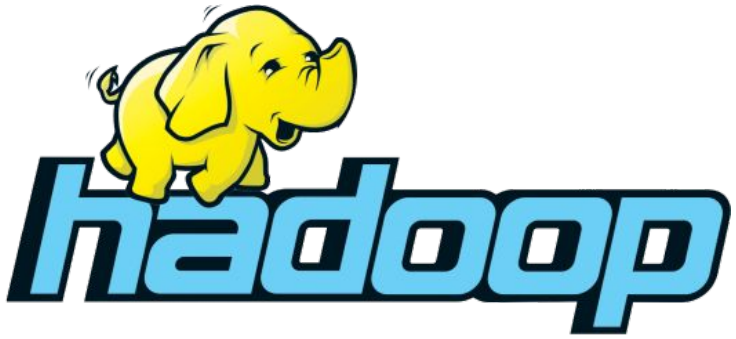
What is Hadoop?

An Open Source framework that allows distributed processing of large data-sets across the cluster of commodity hardware



What is Hadoop?

An open source framework that allows Distributed Processing of large data-sets across the cluster of commodity hardware

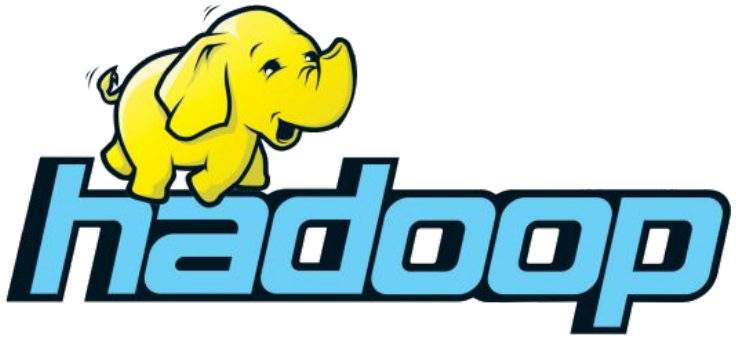


Distributed Processing

- Data is processed distributedly on multiple nodes / servers
- Multiple machines processes the data independently

What is Hadoop?

An open source framework that allows distributed processing of large data-sets across the Cluster of commodity hardware

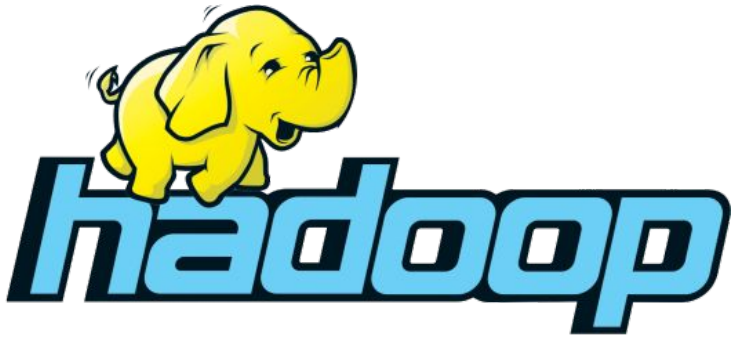


Cluster

- Multiple machines connected together
- Nodes are connected via LAN

What is Hadoop?

An open source framework that allows distributed processing of large data-sets across the cluster of Commodity Hardware

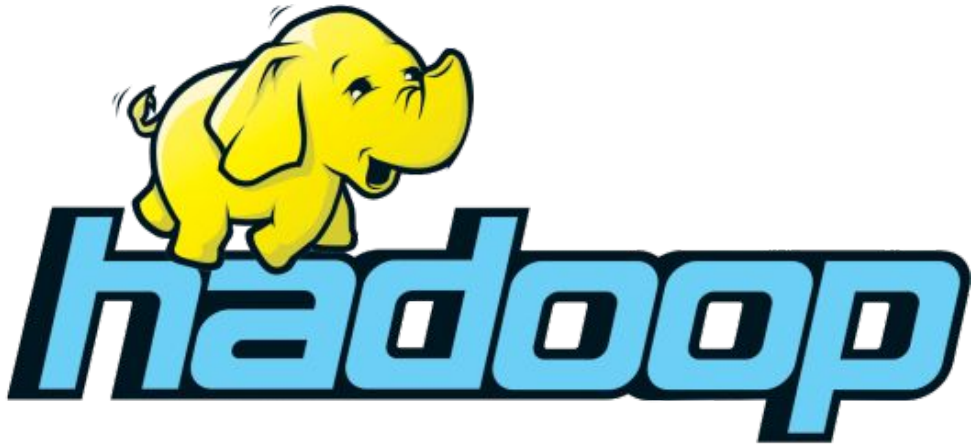


Commodity Hardware

- Economic / affordable machines
- Typically low performance hardware

What is Hadoop?

- Open source framework written in Java
- Inspired by Google's Map-Reduce programming model as well as its file system (GFS)



Introduction to Hadoop

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 an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environment.
- It is designed to handle big data and is based on the MapReduce programming model, which allows for the parallel processing of large datasets.

Introduction to Hadoop

- Hadoop also includes several additional modules that provide additional functionality, such as Hive (a SQL-like query language), Pig (a high-level platform for creating MapReduce programs), and HBase (a non-relational, distributed database).
- Hadoop is commonly used in big data scenarios such as data warehousing, business intelligence, and machine learning.
- It’s also used for data processing, data analysis, and data mining.

Introduction to Hadoop

- The Hadoop framework allows for the distributed processing of large data sets across clusters of computers using simple programming models.
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.
- It is used by many organizations, including Yahoo, Facebook, and IBM, for a variety of purposes such as data warehousing, log processing, and research.
- Hadoop has been widely adopted in the industry and has become a key technology for big data processing.

History of Hadoop

- The Hadoop was started by Doug Cutting and Mike Cafarella in 2002. Its origin was the Google File System paper, published by Google.
- It's co-founder Doug Cutting named it on his son's toy elephant.

Let's focus on the history of Hadoop in the following steps: -

- In 2002, Doug Cutting and Mike Cafarella started to work on a project, Apache Nutch. It is an open source web crawler software project.
- While working on Apache Nutch, they were dealing with big data. To store that data they have to spend a lot of costs which becomes the consequence of that project. This problem becomes one of the important reason for the emergence of Hadoop.
- In 2003, Google introduced a file system known as GFS (Google file system). It is a proprietary distributed file system developed to provide efficient access to data.

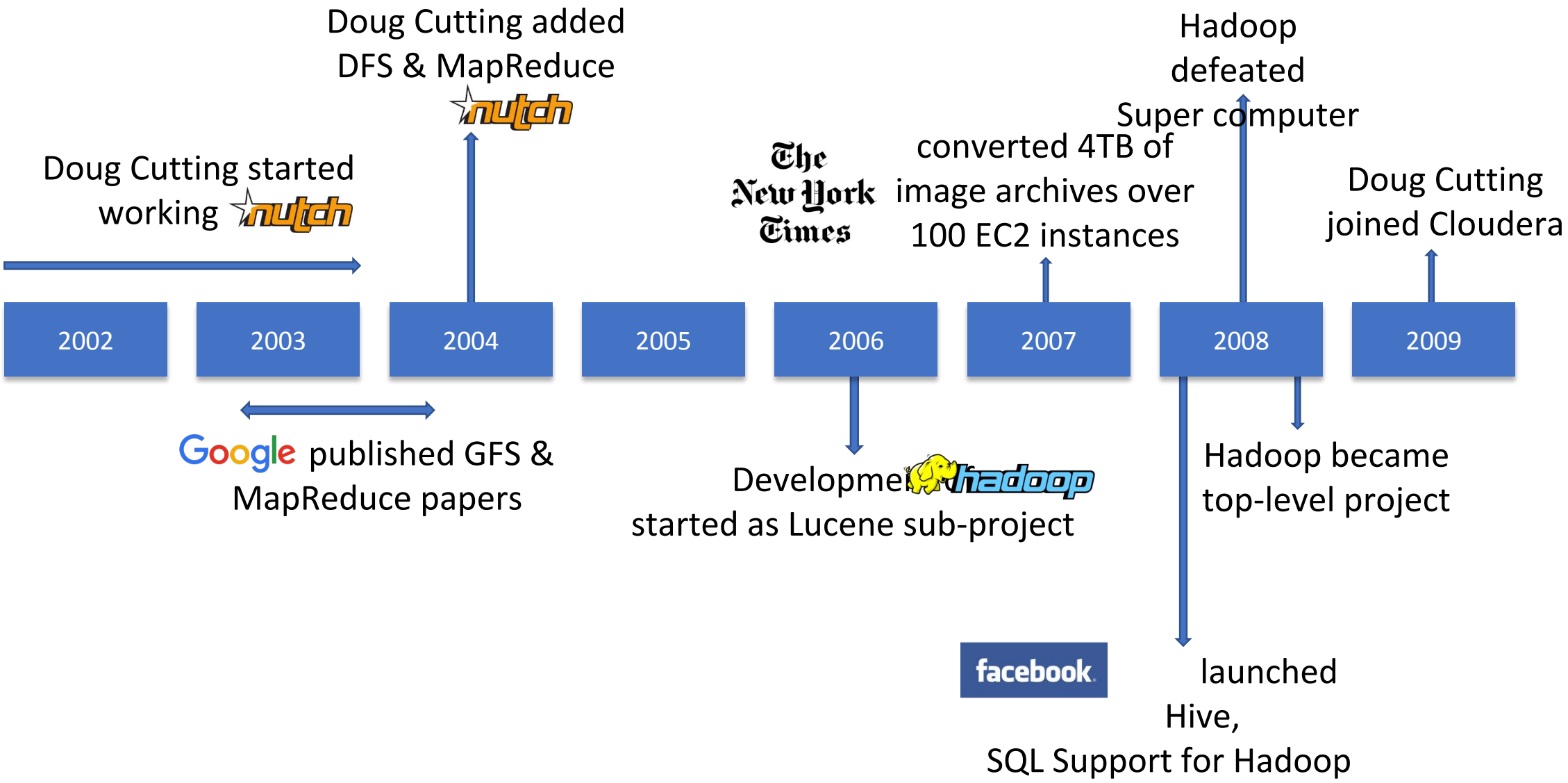
History of Hadoop

- In 2004, Google released a white paper on Map Reduce.
- This technique simplifies the data processing on large clusters.
- In 2005, Doug Cutting and Mike Cafarella introduced a new file system known as NDFS (Nutch Distributed File System).
- This file system also includes Map reduce.
- In 2006, Doug Cutting quit Google and joined Yahoo.
- On the basis of the Nutch project, Dough Cutting introduces a new project Hadoop with a file system known as HDFS (Hadoop Distributed File System).
- Hadoop first version 0.1.0 released in this year.
- Doug Cutting gave named his project Hadoop after his son's toy elephant.

History of Hadoop

- In 2007, Yahoo runs two clusters of 1000 machines.
- In 2008, Hadoop became the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 seconds.
- In 2013, Hadoop 2.2 was released.
- In 2017, Hadoop 3.0 was released.

Hadoop History



The Difference Between RDBMS and Hadoop

RDBMS is a system software for creating and managing databases that based on the relational model.	Hadoop is a collection of open source software that connects many computers to solve problems involving a large amount of data and computation.
Data Variety	
RDBMS stores structured data.	Hadoop stores structured, semi-structured and unstructured data.
Data Storage	
RDBMS stores average amount of data.	Hadoop stores a large amount of data than RDBMS.
Speed	
In RDBMS, reads are fast.	In Hadoop, reads and writes are fast.
Scalability	
RDBMS has vertical scalability.	Hadoop has horizontal scalability.
Hardware	
RDBMS use high-end servers.	Hadoop uses commodity hardware.
Throughput	
RDBMS throughput is higher.	Hadoop throughput is lower.

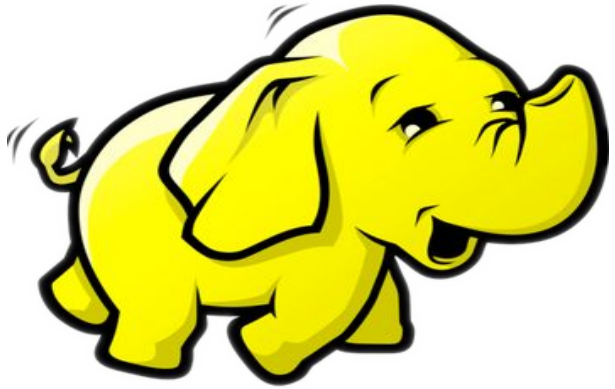
Key Aspects of Hadoop

Open Source

Distributed Processing

Fault Tolerance

Easy to use



Reliability

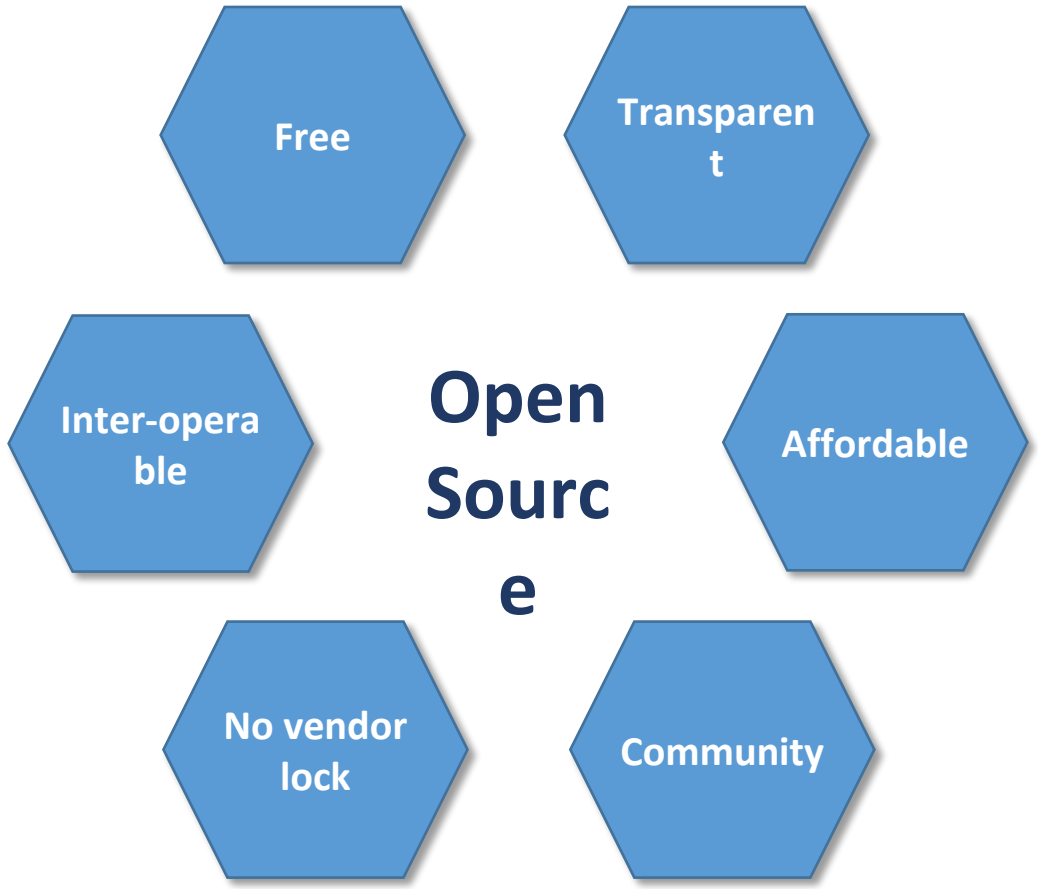
Economic

Scalability

High Availability

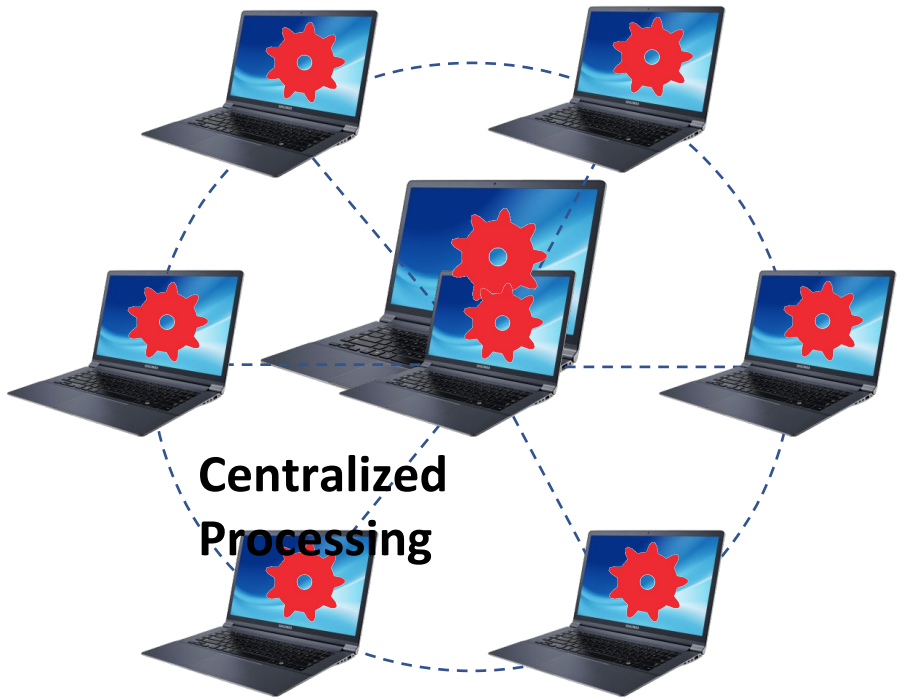
Open Source

- Source code is freely available
- Can be redistributed
- Can be modified



Distributed Processing

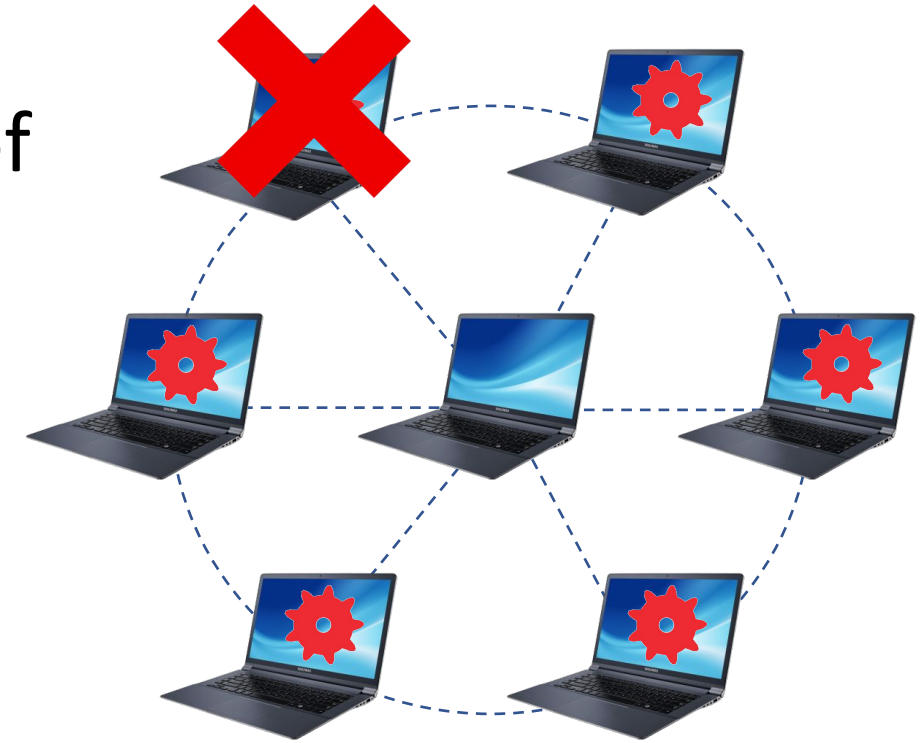
- Data is processed distributedly on cluster
- Multiple nodes in the cluster process data independently



Distributed Processing

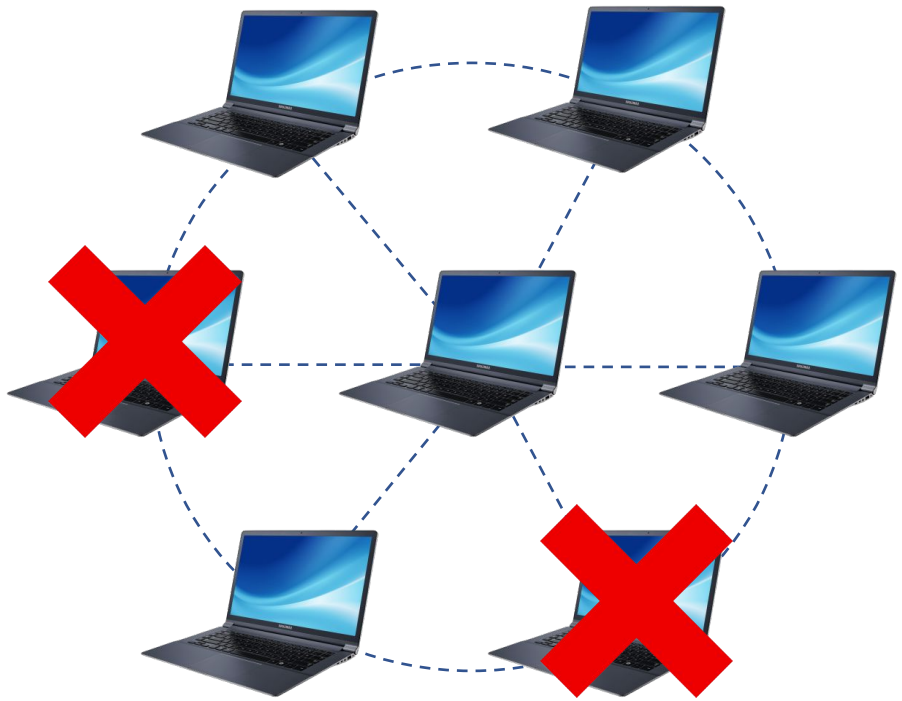
Fault Tolerance

- Failure of nodes are recovered automatically
- Framework takes care of failure of hardware as well tasks



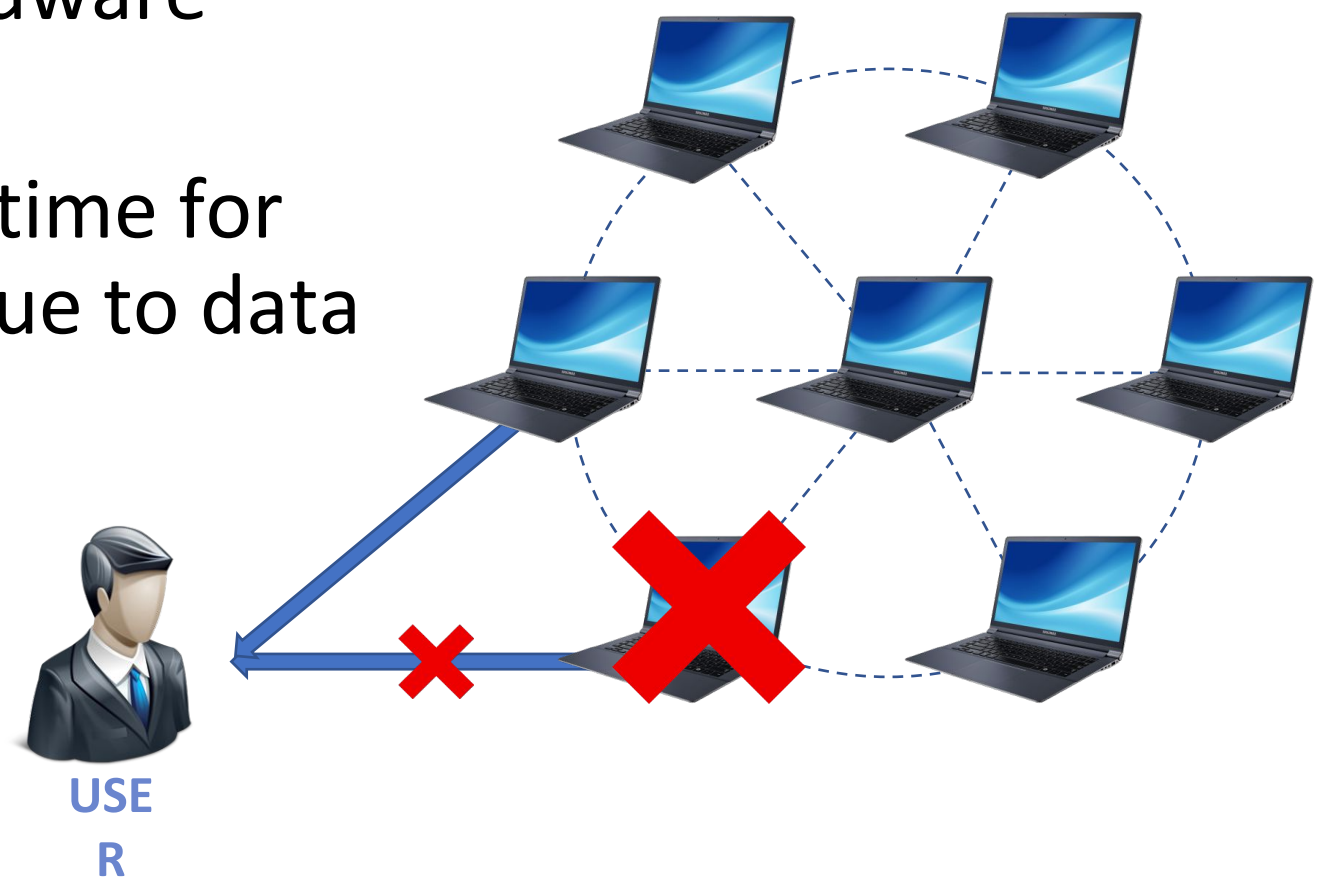
Reliability

- Data is reliably stored on the cluster of machines despite machine failures
- Failure of nodes doesn’t cause data loss



High Availability

- Data is highly available and accessible despite hardware failure
- There will be no downtime for end user application due to data



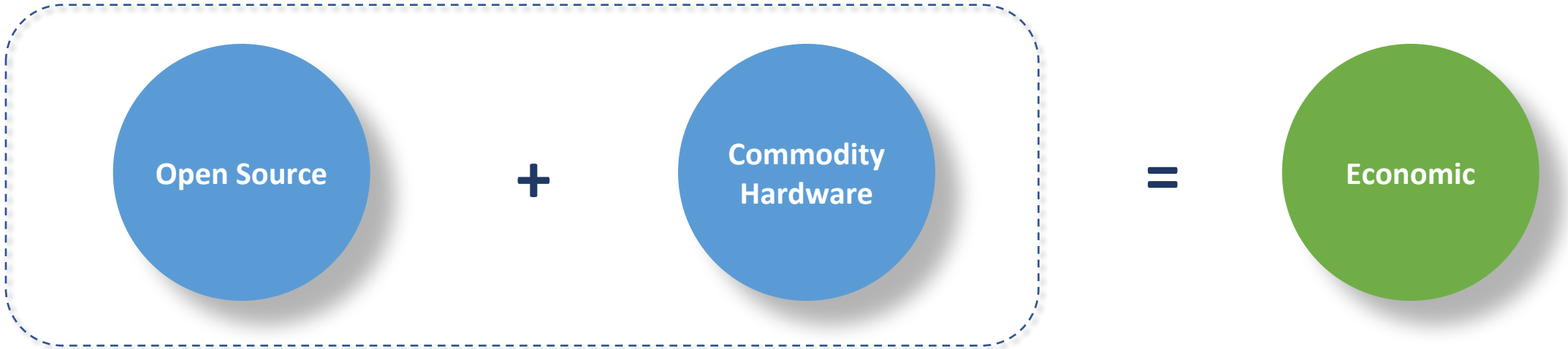
Scalability

- Vertical Scalability – New hardware can be added to the nodes

- Horizontal Scalability – New nodes can be added on the fly

Economic

- No need to purchase costly license
- No need to purchase costly hardware

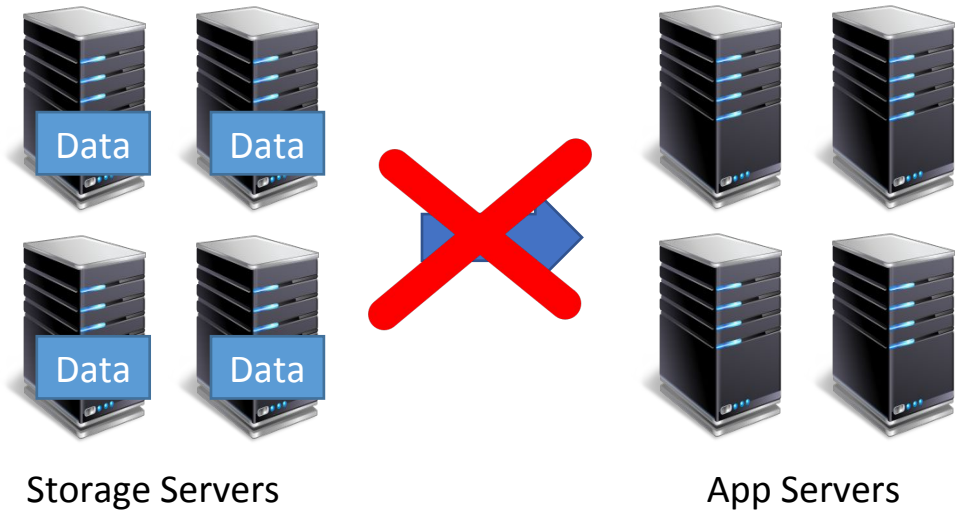


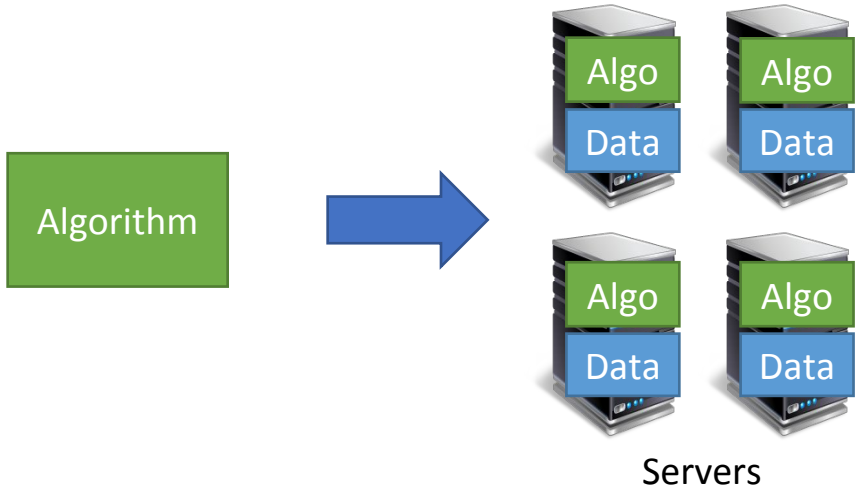
Easy to Use

- Distributed computing challenges are handled by framework
- Client just need to concentrate on business logic

Data Locality

- Move computation to data instead of data to computation
- Data is processed on the nodes where it is stored

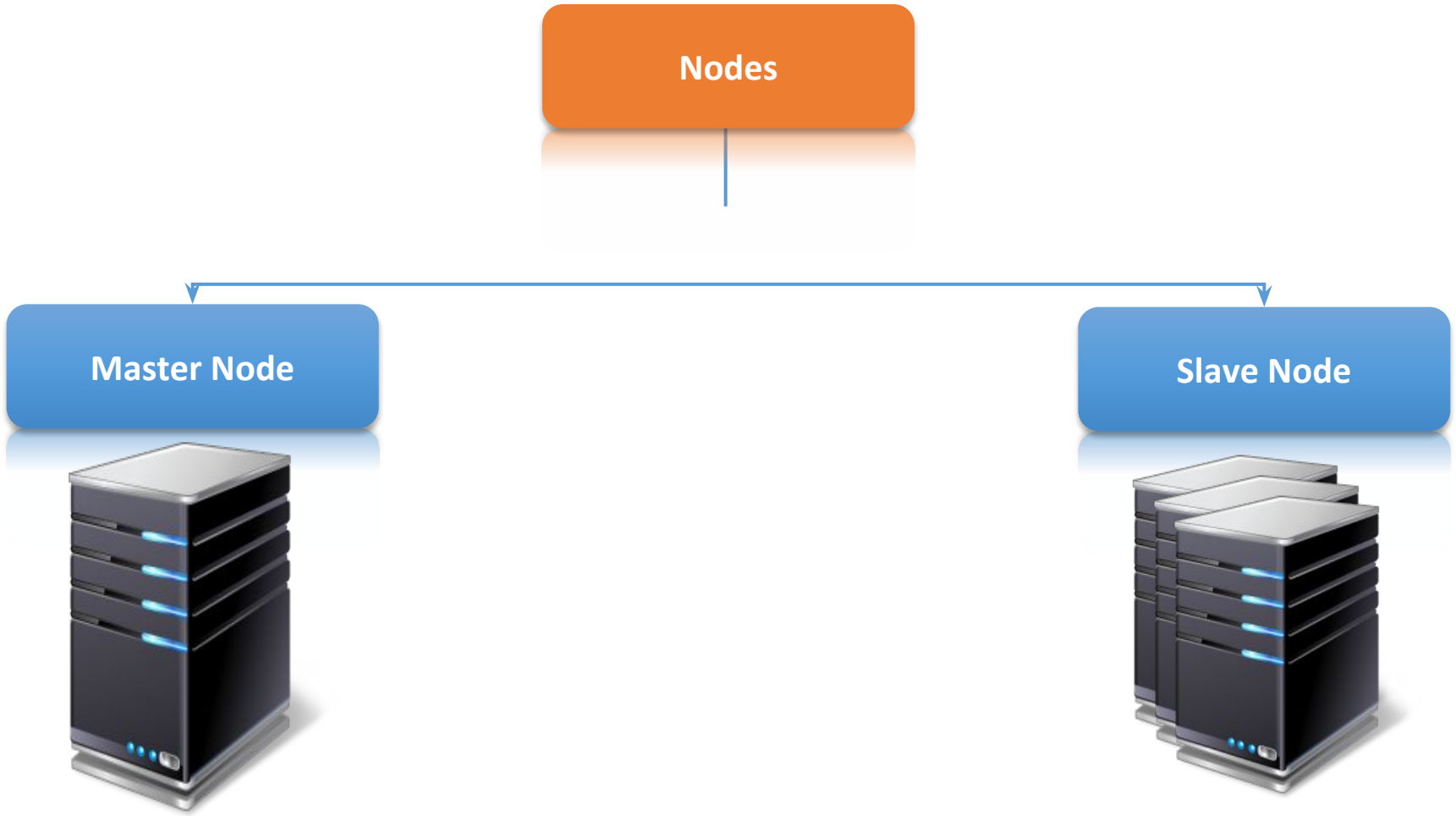




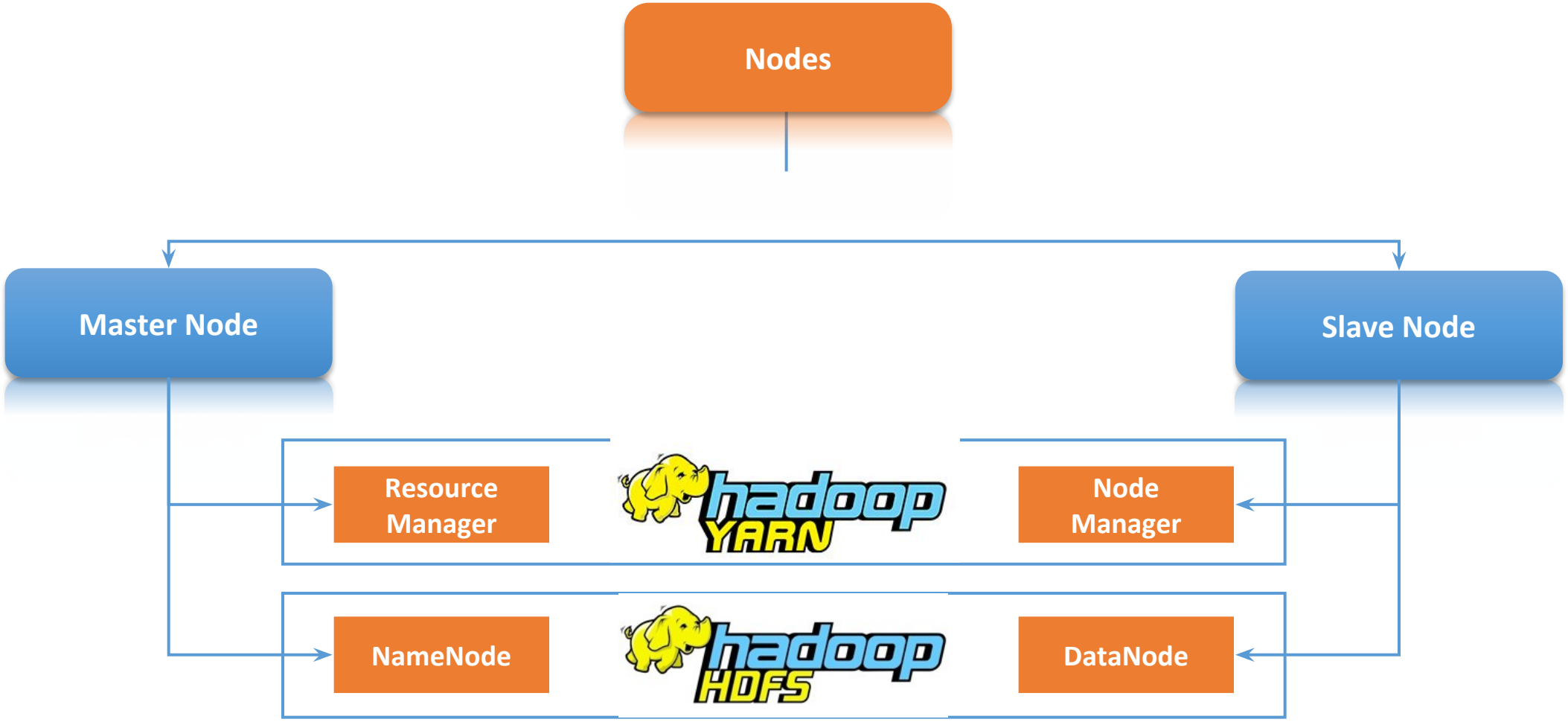
Hadoop Components

- HDFS
- MapReduce

Hadoop Nodes



Hadoop Daemons

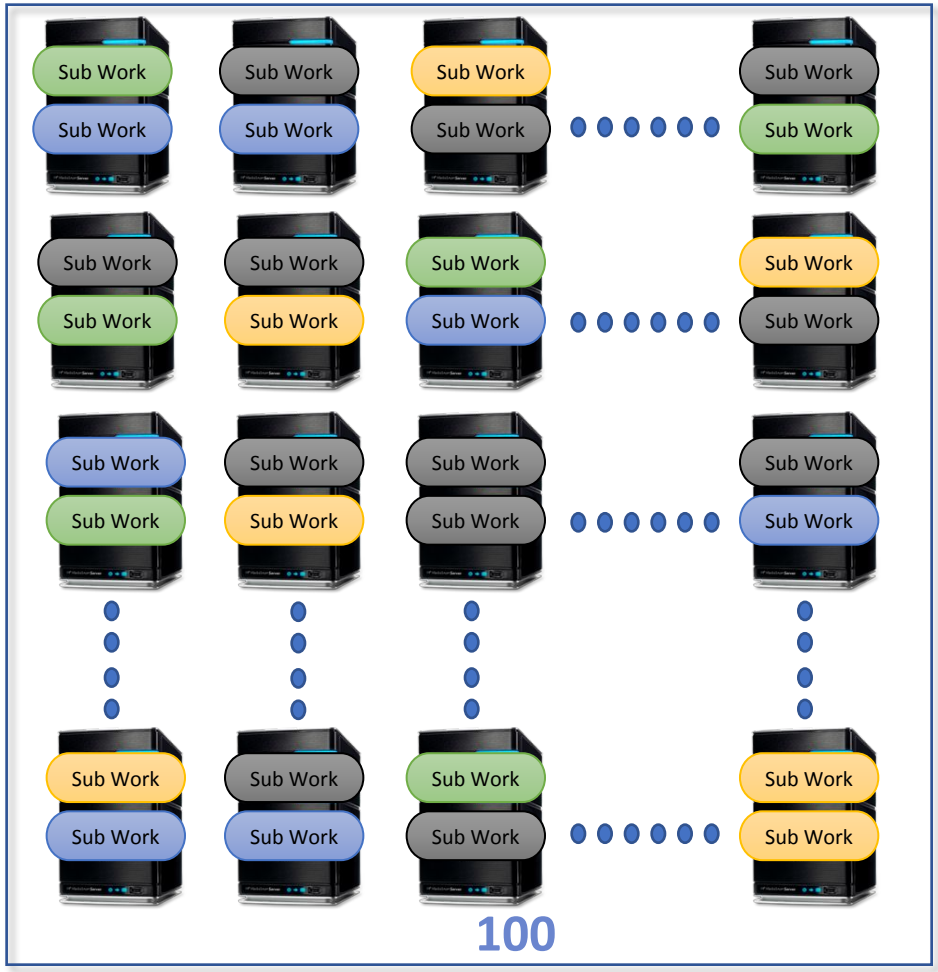


Basic Hadoop Architecture

USE
R

Work

MASTER(S
)



Hadoop Architecture

- The Hadoop architecture is a package of the file system, MapReduce engine and the HDFS (Hadoop Distributed File System).
- The MapReduce 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 TaskTracker.
- Both NameNode and DataNode are capable enough to run on commodity machines. The Java language is used to develop HDFS. So any machine that supports Java language can easily run the NameNode and DataNode software.

Hadoop Architecture

NameNode

- It is a single master server exist in the HDFS cluster.
- As it is a single node, it may become the reason of single point failure.
- It manages the file system namespace by executing an operation like the opening, renaming and closing the files.
- It simplifies the architecture of the system.

Hadoop Architecture

DataNode

- The HDFS cluster contains multiple DataNodes.
- Each DataNode contains multiple data blocks.
- These data blocks are used to store data.
- It is the responsibility of DataNode to read and write requests from the file system's clients.
- It performs block creation, deletion, and replication upon instruction from the NameNode.

Hadoop Architecture

Job Tracker

- The role of Job Tracker is to accept the MapReduce jobs from client and process the data by using NameNode.
- In response, NameNode provides metadata to Job Tracker.

Task Tracker

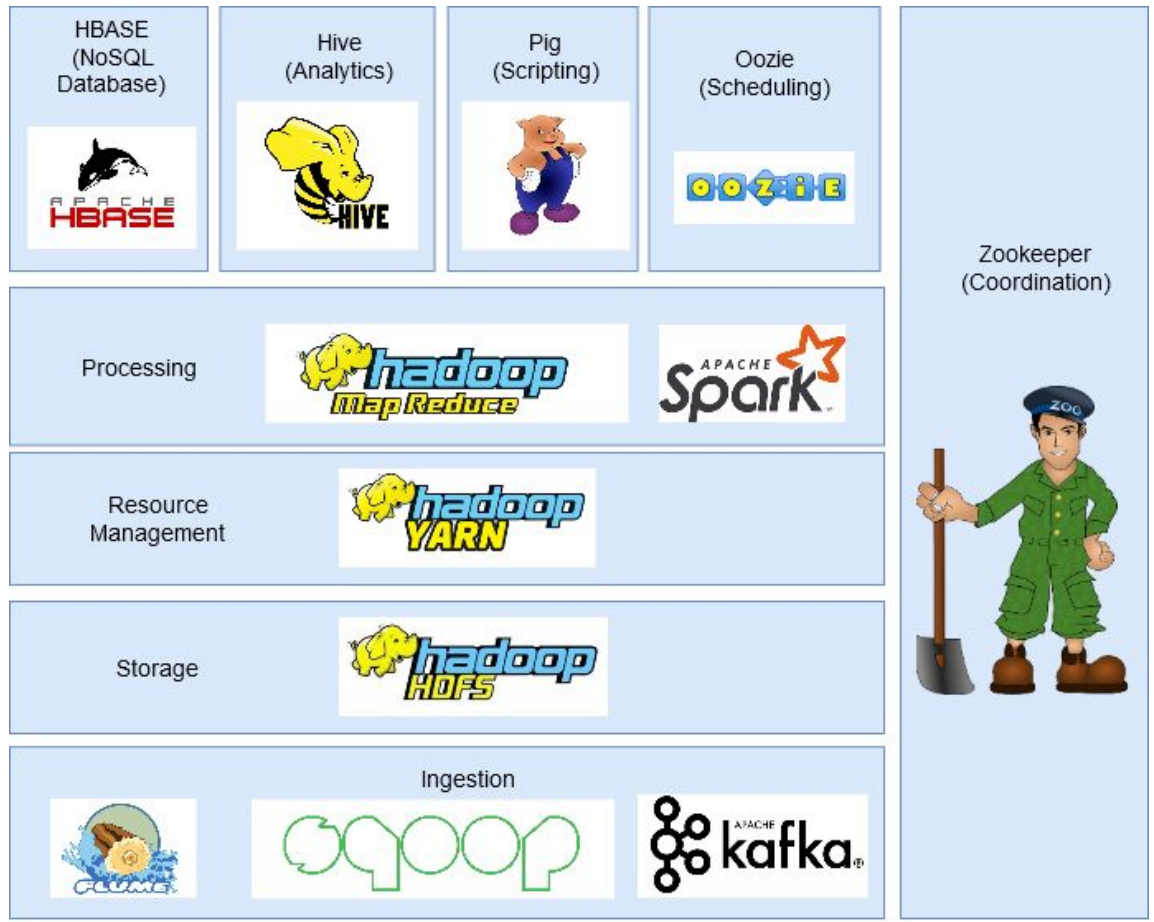
- It works as a slave node for Job Tracker.
- It receives task and code from Job Tracker and applies that code on the file. This process can also be called as a Mapper.

Hadoop Architecture

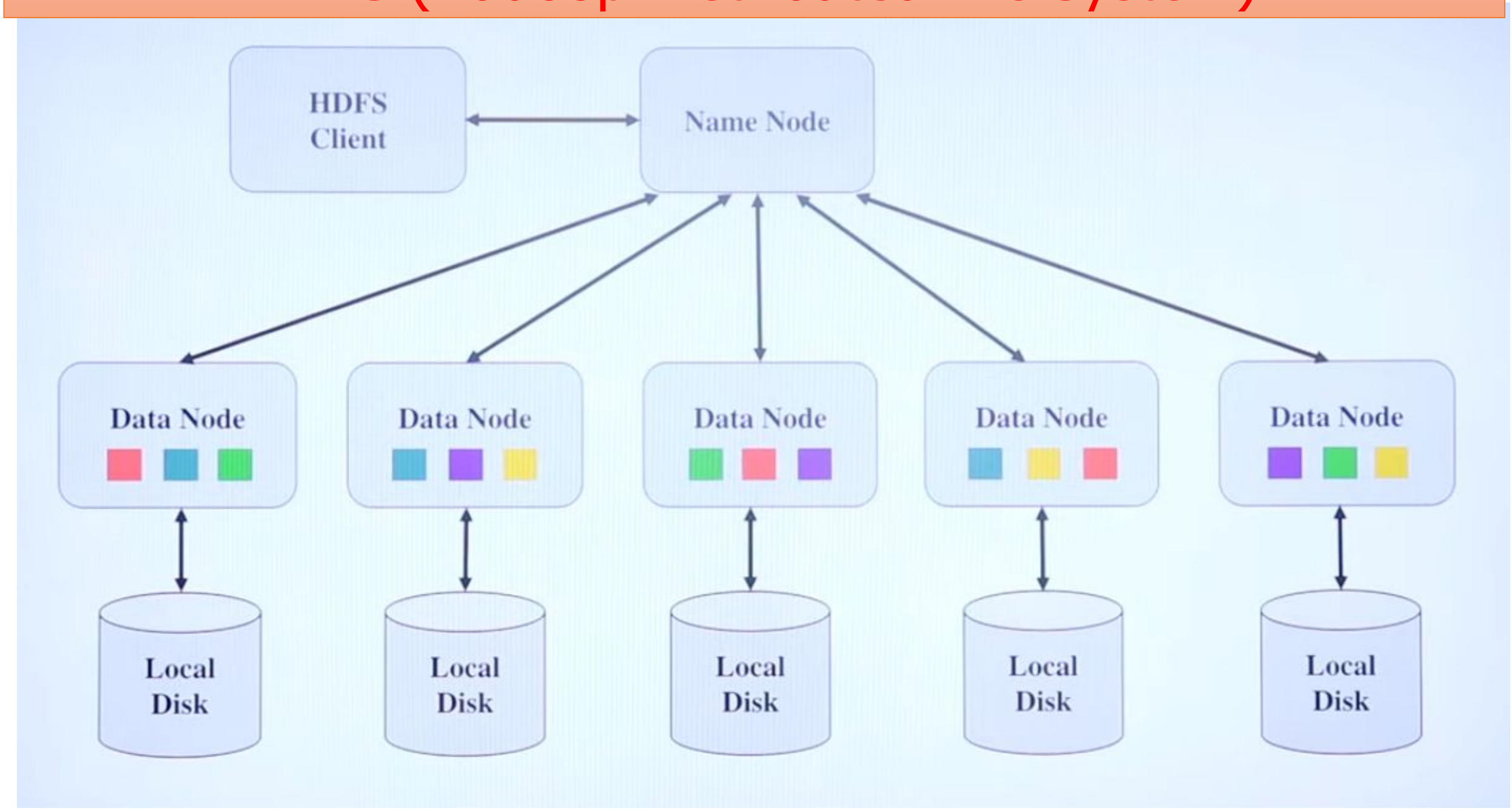
MapReduce Layer

- The MapReduce comes into existence when the client application submits the MapReduce job to Job Tracker.
- In response, the Job Tracker sends the request to the appropriate Task Trackers.
- Sometimes, the TaskTracker fails or time out.
- In such a case, that part of the job is rescheduled.

Hadoop Components

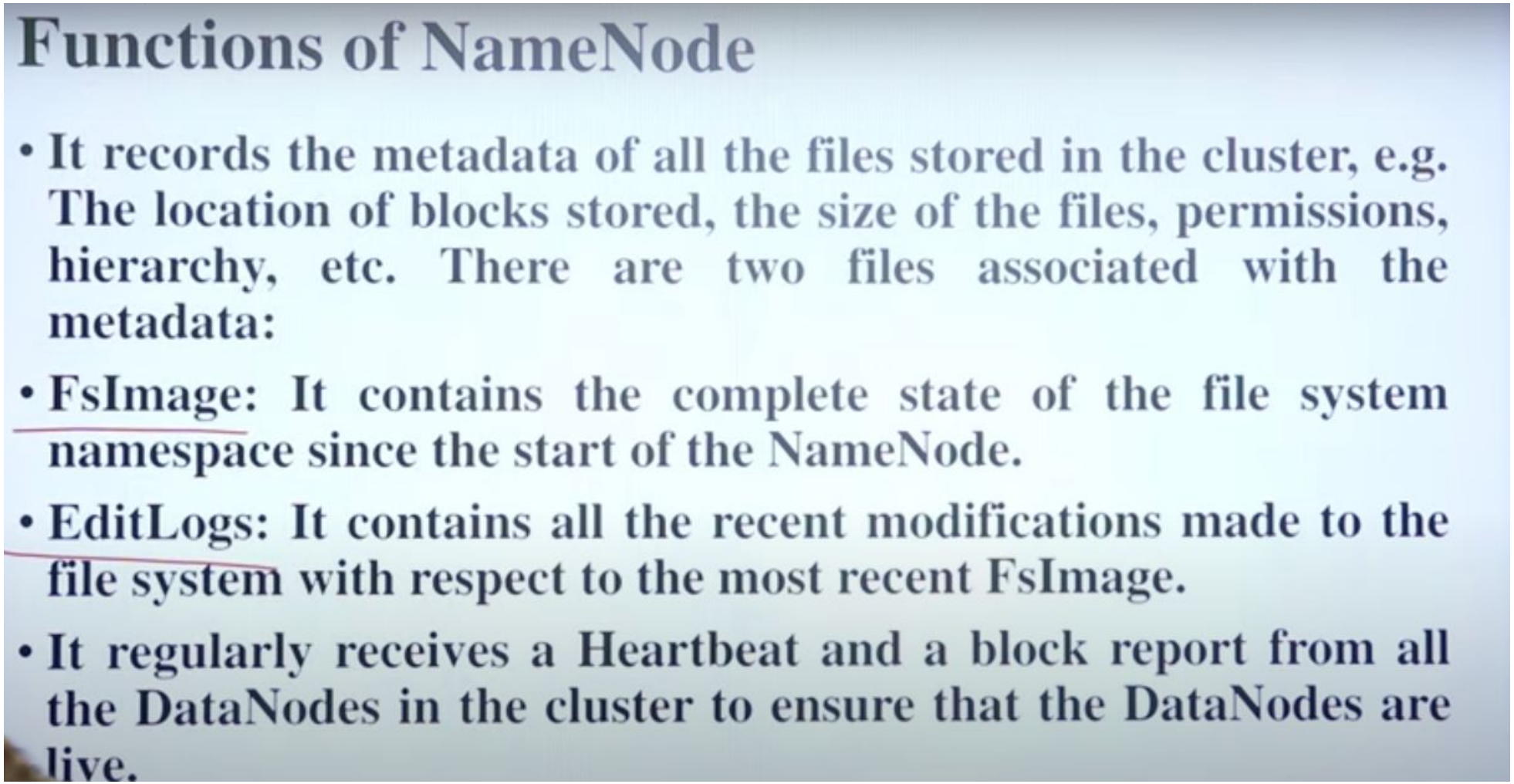


HDFS (Hadoop Distributed File System)



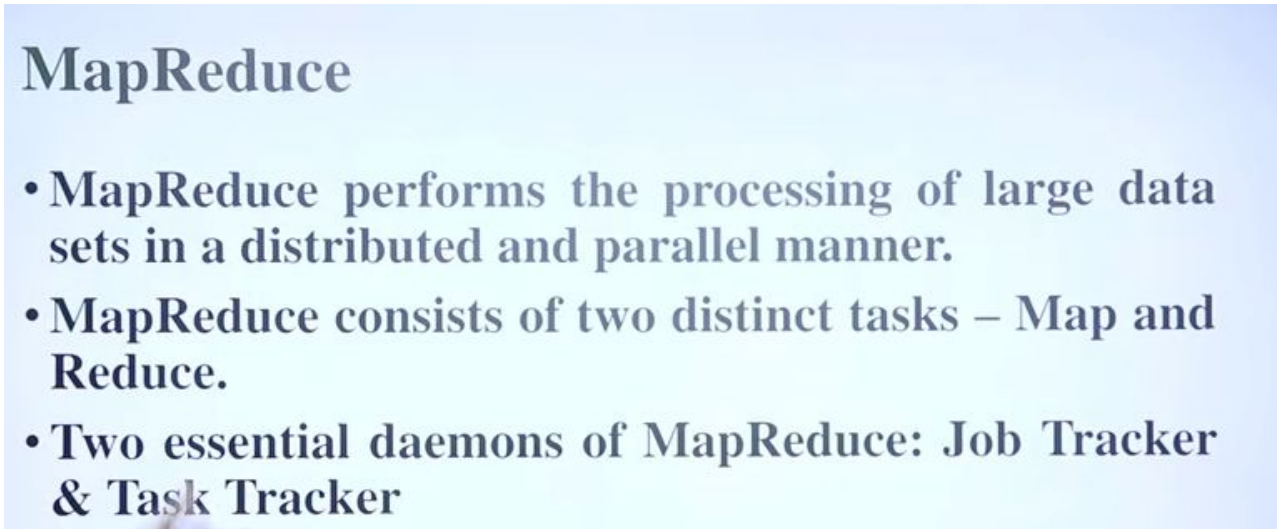
HDFS (Hadoop Distributed File System)

- **HDFS (Hadoop Distributed File System):** This is the storage component of Hadoop, which allows for the storage of large amounts of data across multiple machines.
- It is designed to work with commodity hardware, which makes it cost-effective.

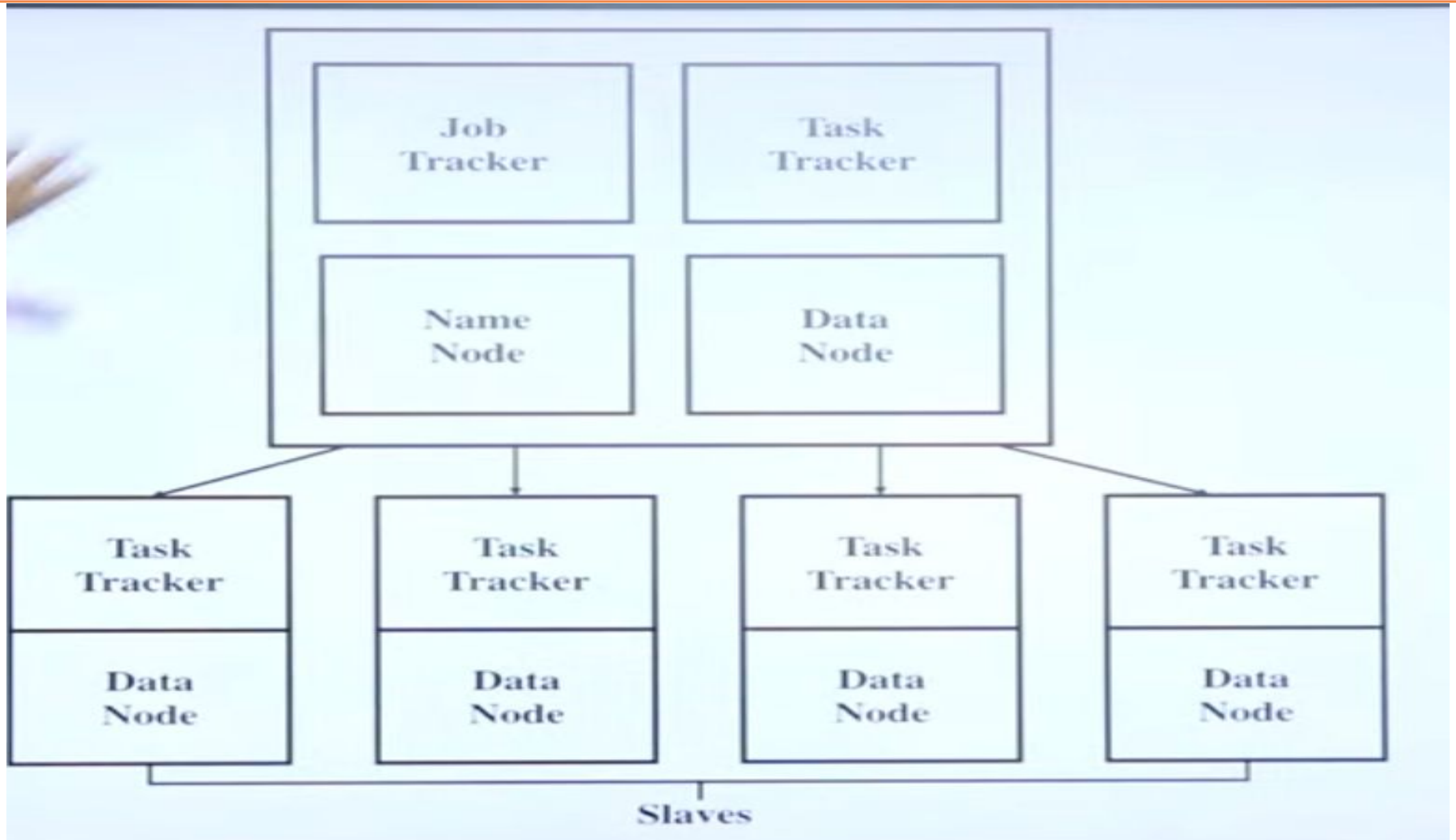


Map Reduce

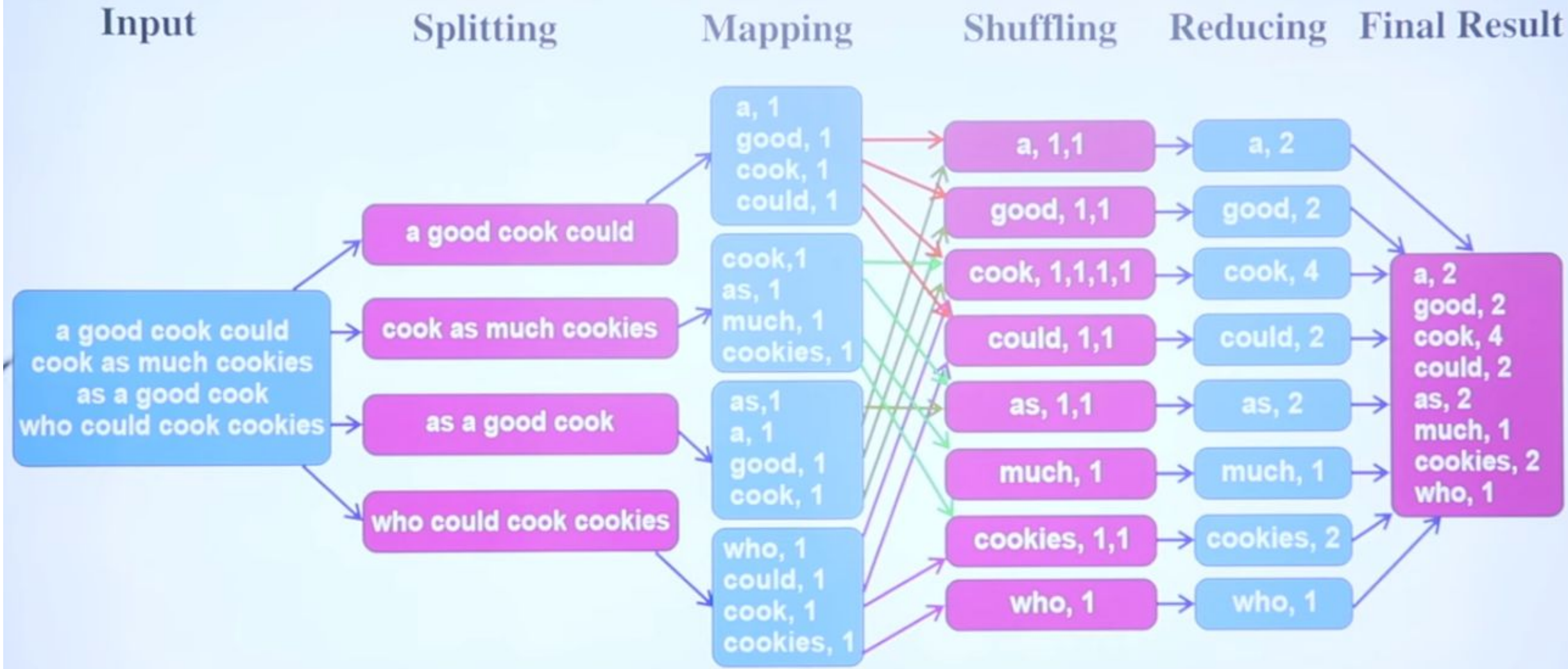
- **Map Reduce:** This is a framework which helps Java programs to do the parallel computation on data using key value pair.
- The Map task takes input data and converts it into a data set which can be computed in Key value pair.
- The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.



Map Reduce


Job Tracker
Task Tracker
Name Node
Data Node
Task Tracker
Data Node
Task Tracker
Data Node
Task Tracker
Data Node
Task Tracker
Data Node
Slaves

Map Reduce


Input
Splitting
Mapping
Shuffling
Reducing
Final Result
a good cook could cook as much cookies as a good cook who could cook cookies
a good cook could
cook as much cookies
as a good cook
who could cook cookies
a, 1
good, 1
cook, 1
could, 1
cook,1
as, 1
much, 1
cookies, 1
as,1
a, 1
good, 1
cook, 1
who, 1
could, 1
cook, 1
cookies, 1
a, 1,1
good, 1,1
cook, 1,1,1,1
could, 1,1
as, 1,1
much, 1
cookies, 1,1
who, 1
a, 2
good, 2
cook, 4
could, 2
as, 2
much, 1
cookies, 2
who, 1
a, 2
good, 2
cook, 4
could, 2
as, 2
much, 1
cookies, 2
who, 1

cloudera

Cloudera

Hive
(SQL Query)

Pig
(Scripting)

Mahout
(Machine Learning)

Oozie
(Workflow)

MapReduce
(Data Processing)

YARN
(Cluster Resource Management)

HDFS
(Hadoop Distributed File System)

Flume
(Data Collection)

Sqoop
Collection)

Zookeeper
(Co-ordination)

HBASE
(Columnar Stone)