Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

Associate Professor

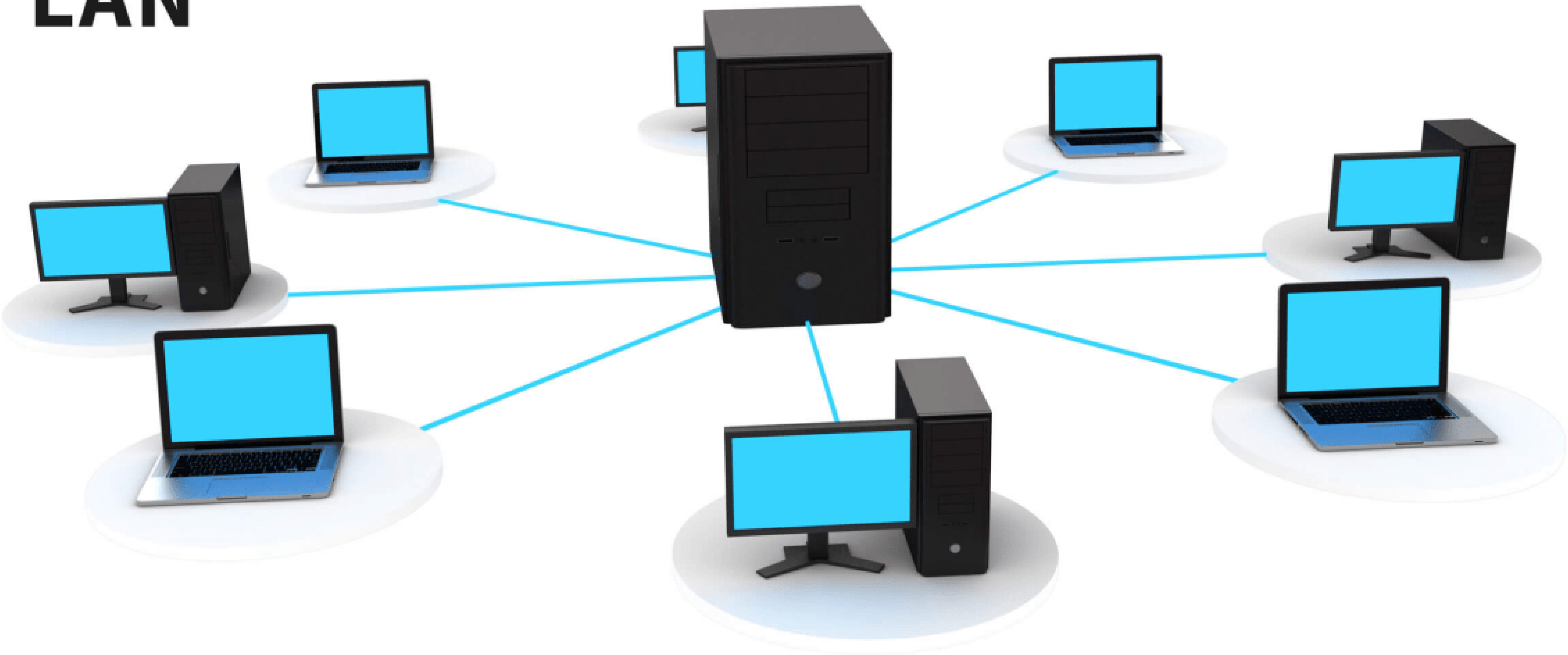
Learning Outcomes

After this lecture, you will be able to

- What is Hadoop Cluster
- Learn Architecture of Hadoop Cluster
- Learn data storage in hdfs
- HDFS Architecture
- HDFS Hadoop Features

What is Hadoop Cluster?

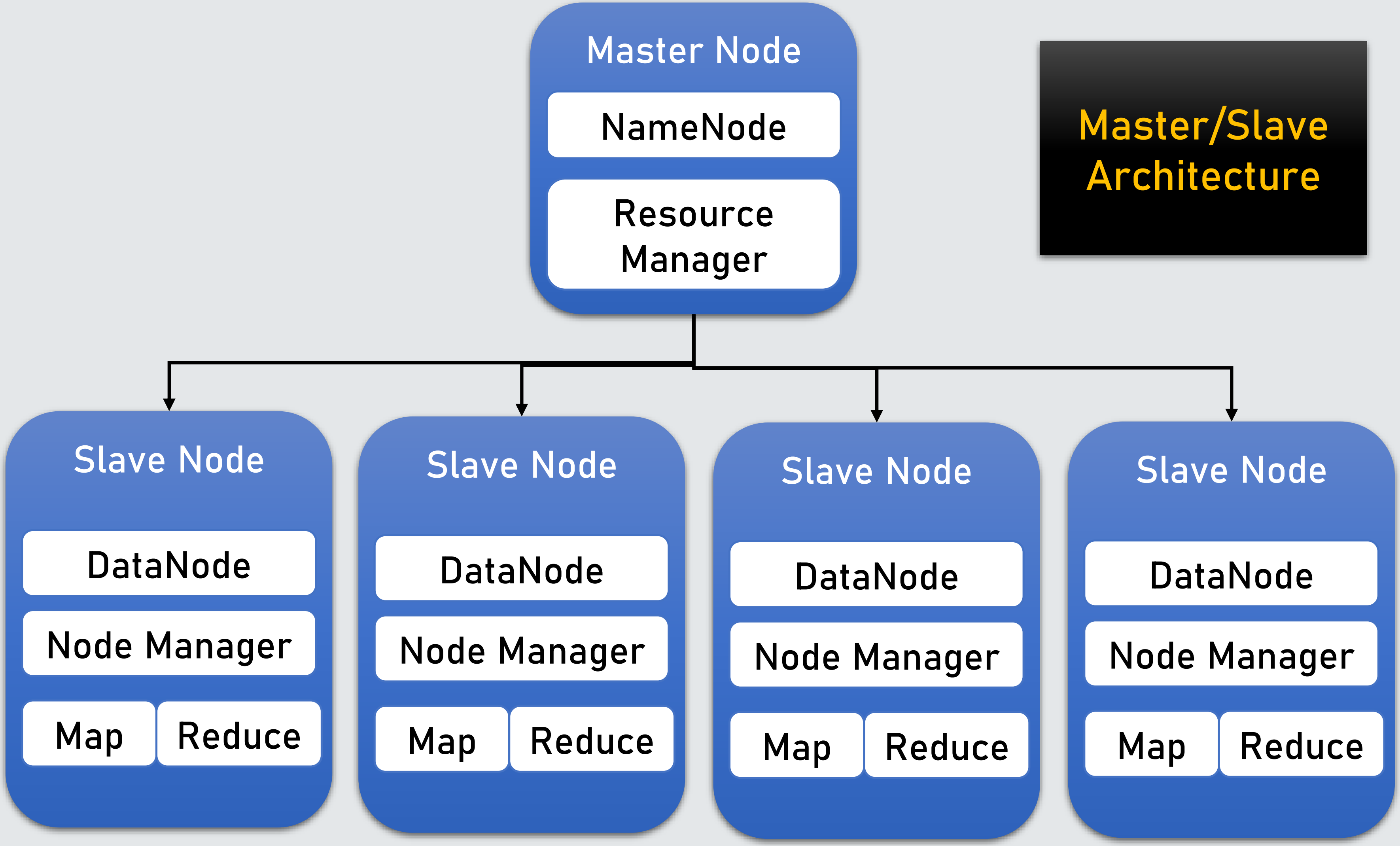
LAN



What is Hadoop Cluster?

- Storing and processing large data sets
- Commodity hardware connected together.
- Communicate with a high-end machine.
- Master and slaves implement distributed computing

Architecture of Hadoop



Master Node

NameNode

Resource Manager

Slave Node

DataNode

Node Manager

Map | Reduce

Slave Node

DataNode

Node Manager

Map | Reduce

Slave Node

DataNode

Node Manager

Map | Reduce

Slave Node

DataNode

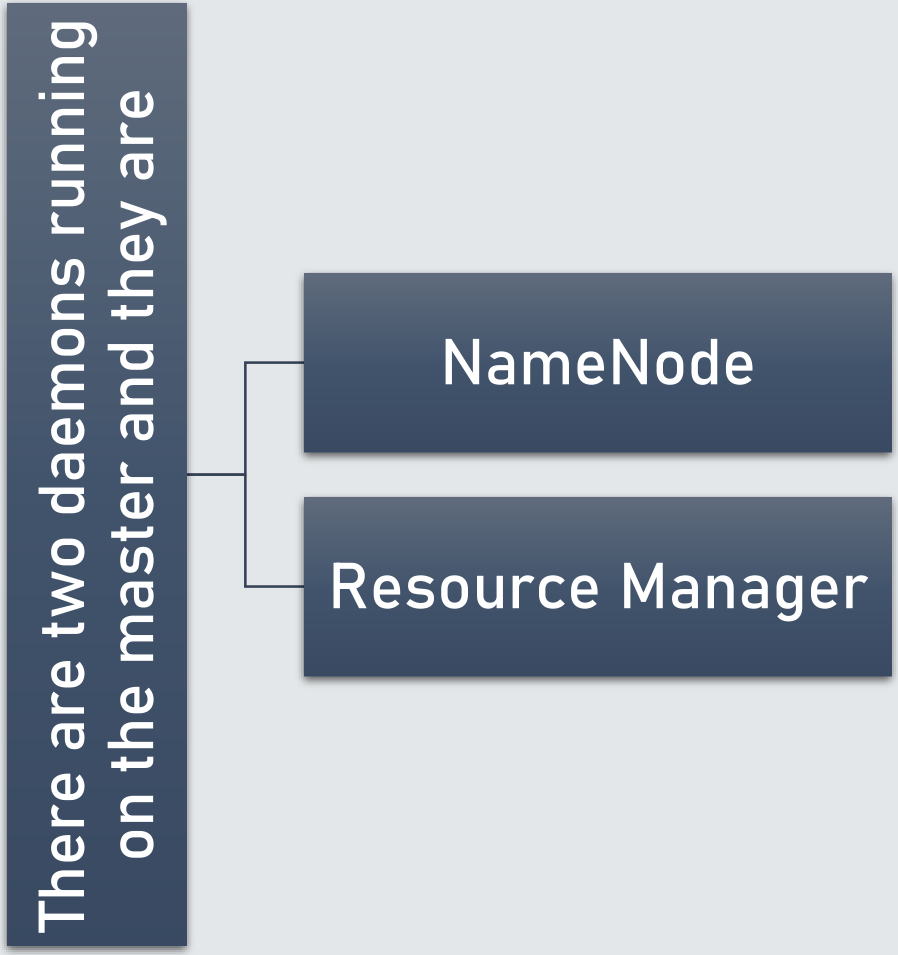
Node Manager

Map | Reduce

Master/Slave Architecture

Architecture of Hadoop

Master in Hadoop Cluster



There are two daemons running on the master and they are

NameNode

Resource Manager

Architecture of Hadoop

i. Functions of NameNode

- Manages file system namespace
- Regulates access to files by clients

Architecture of Hadoop

i. Functions of NameNode

- Stores metadata of actual data Foe example – file path, number of blocks, block id, the location of blocks etc.

Architecture of Hadoop

i. Functions of NameNode

- Executes file system namespace operations like opening, closing, renaming files and directories

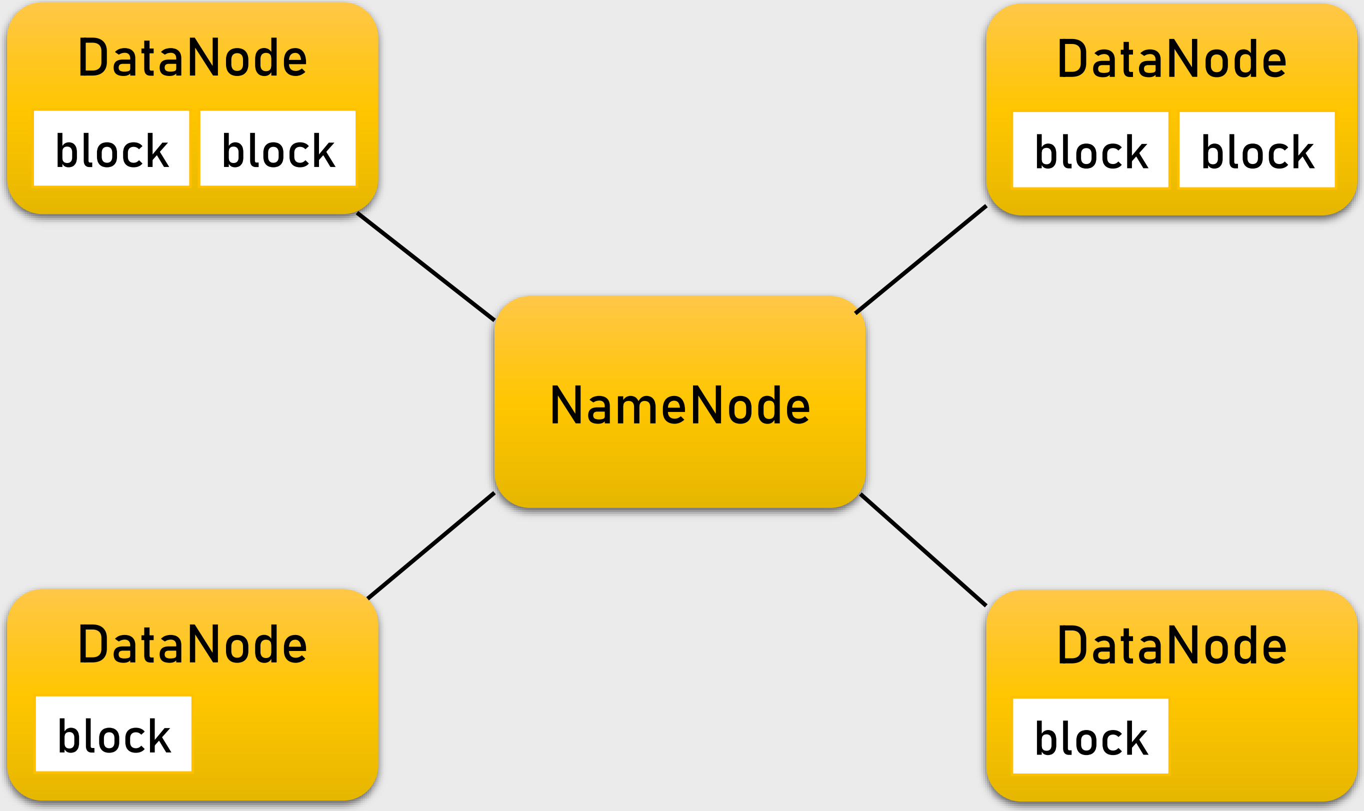
Architecture of Hadoop

i. Functions of NameNode

- The NameNode stores the metadata in the memory for fast retrieval. Hence we should configure it on a high-end machine.

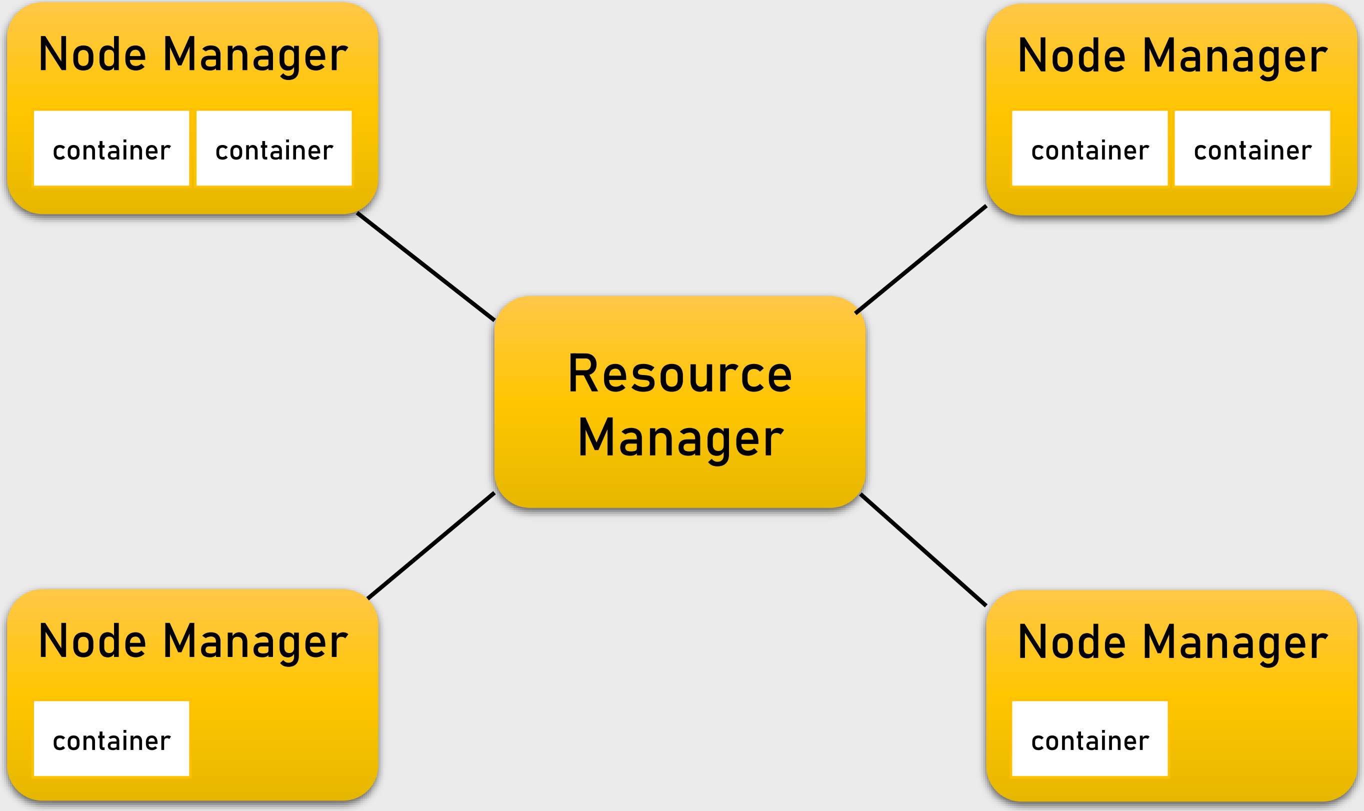
Architecture of Hadoop

i. Functions of NameNode



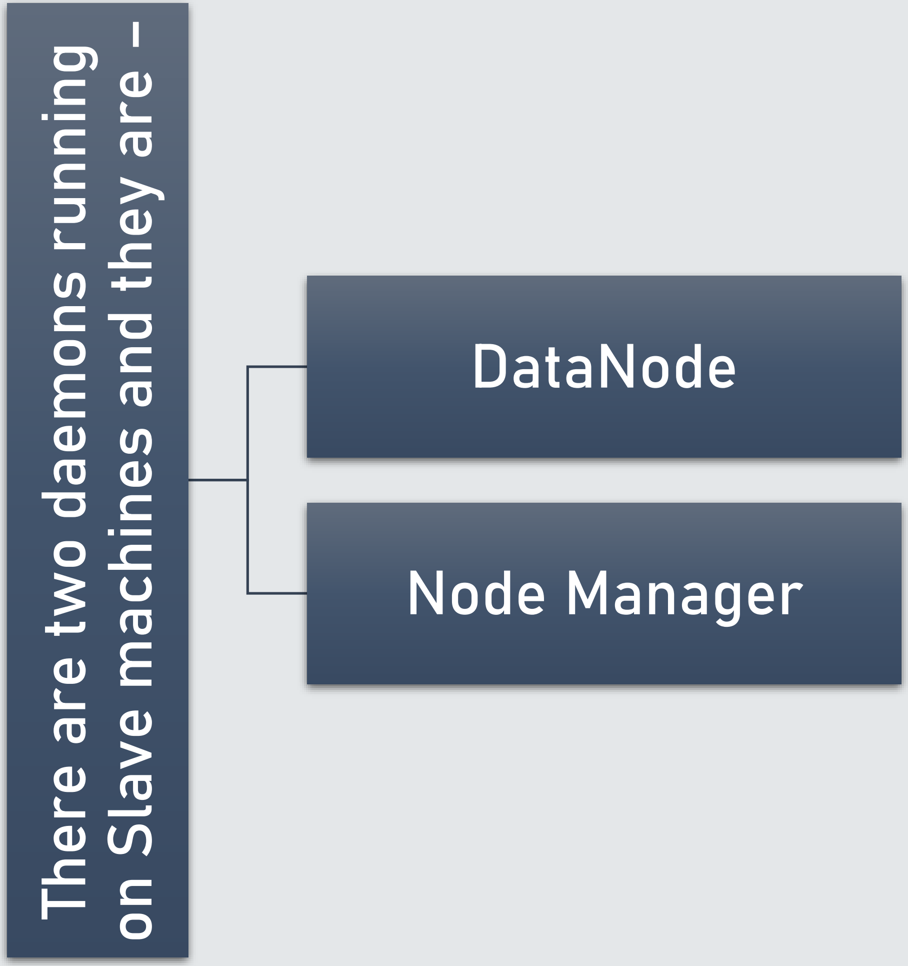
Architecture of Hadoop

ii. Functions of Resource Manager



Architecture of Hadoop

Slaves in the Hadoop Cluster


There are two daemons running on Slave machines and they are –
DataNode
Node Manager

Architecture of Hadoop

Functions of a Data Node

- It stores the business data
- It does read, write and data processing operations
- Upon instruction from a master, it does creation, deletion, and replication of data blocks.

Architecture of Hadoop

Functions of a Node Manager

- It runs services on the node to check its health.
- Scale Hadoop cluster

Architecture of Hadoop

Client nodes in Hadoop cluster – We install Hadoop and configure it on client nodes.

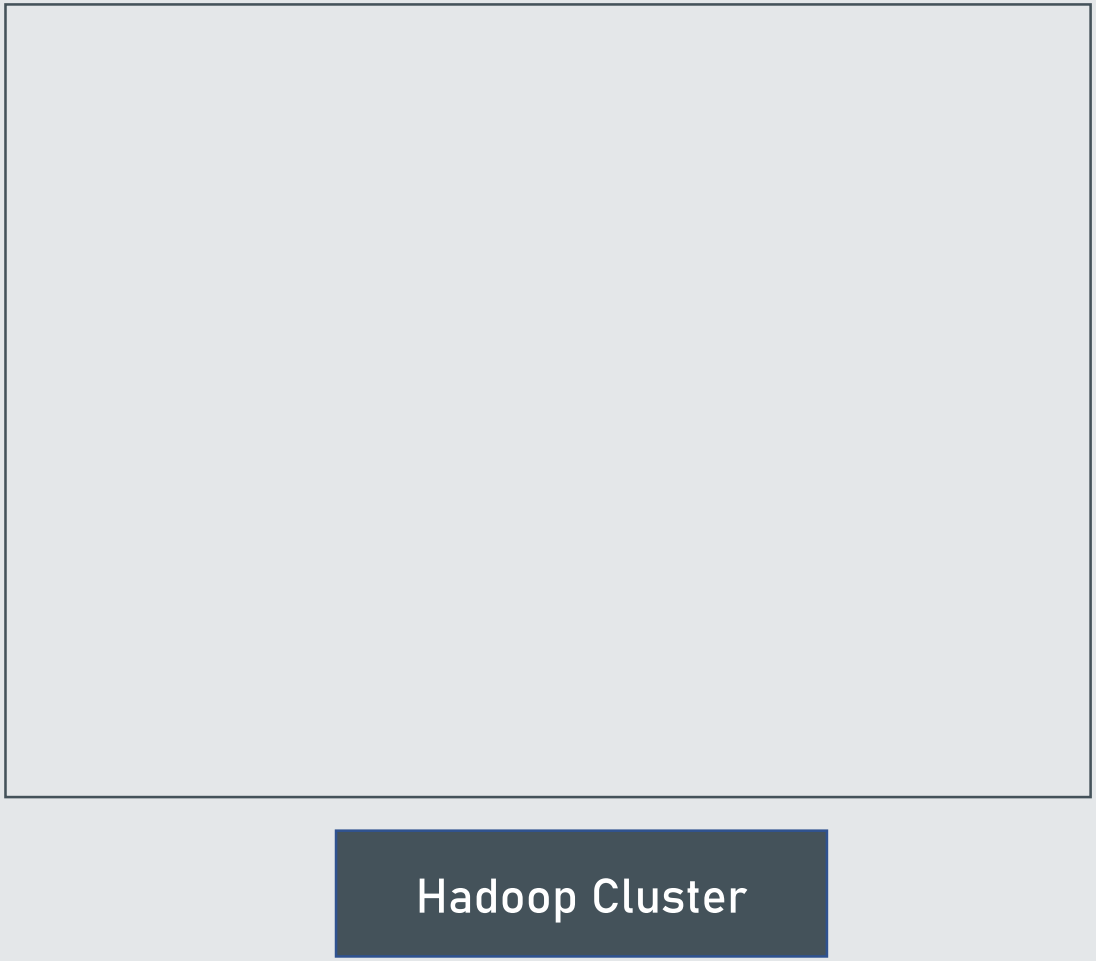
Architecture of Hadoop

Functions of the client node

- To load the data on the Hadoop cluster.
- Tells how to process the data by submitting MapReduce job.
- Collects the output from a specified location.

Data Storage in HDFS

Example



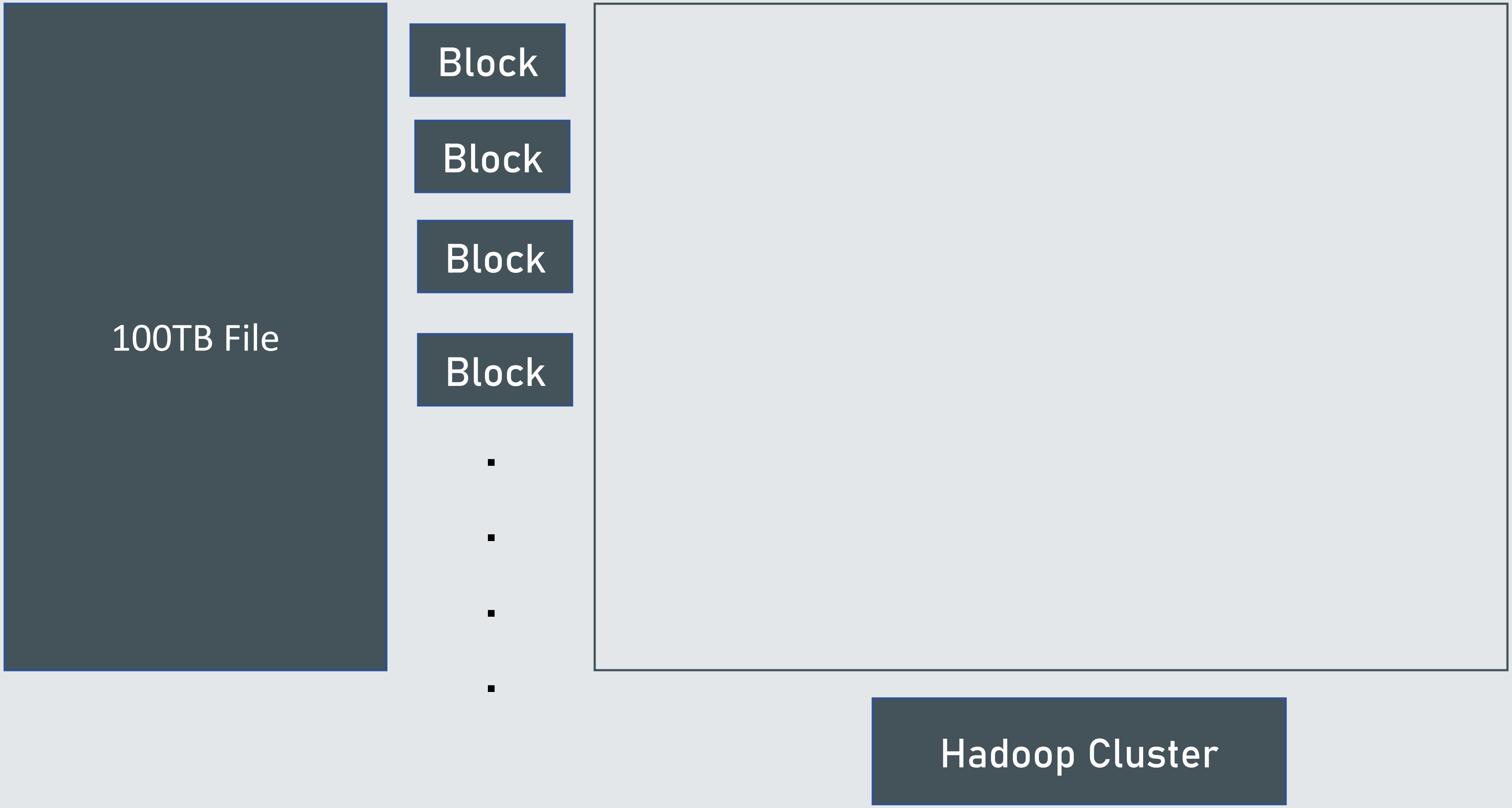
Hadoop Cluster

Example

100TB File

Hadoop Cluster

Example



100TB File

Block

Block

Block

Block

.

.

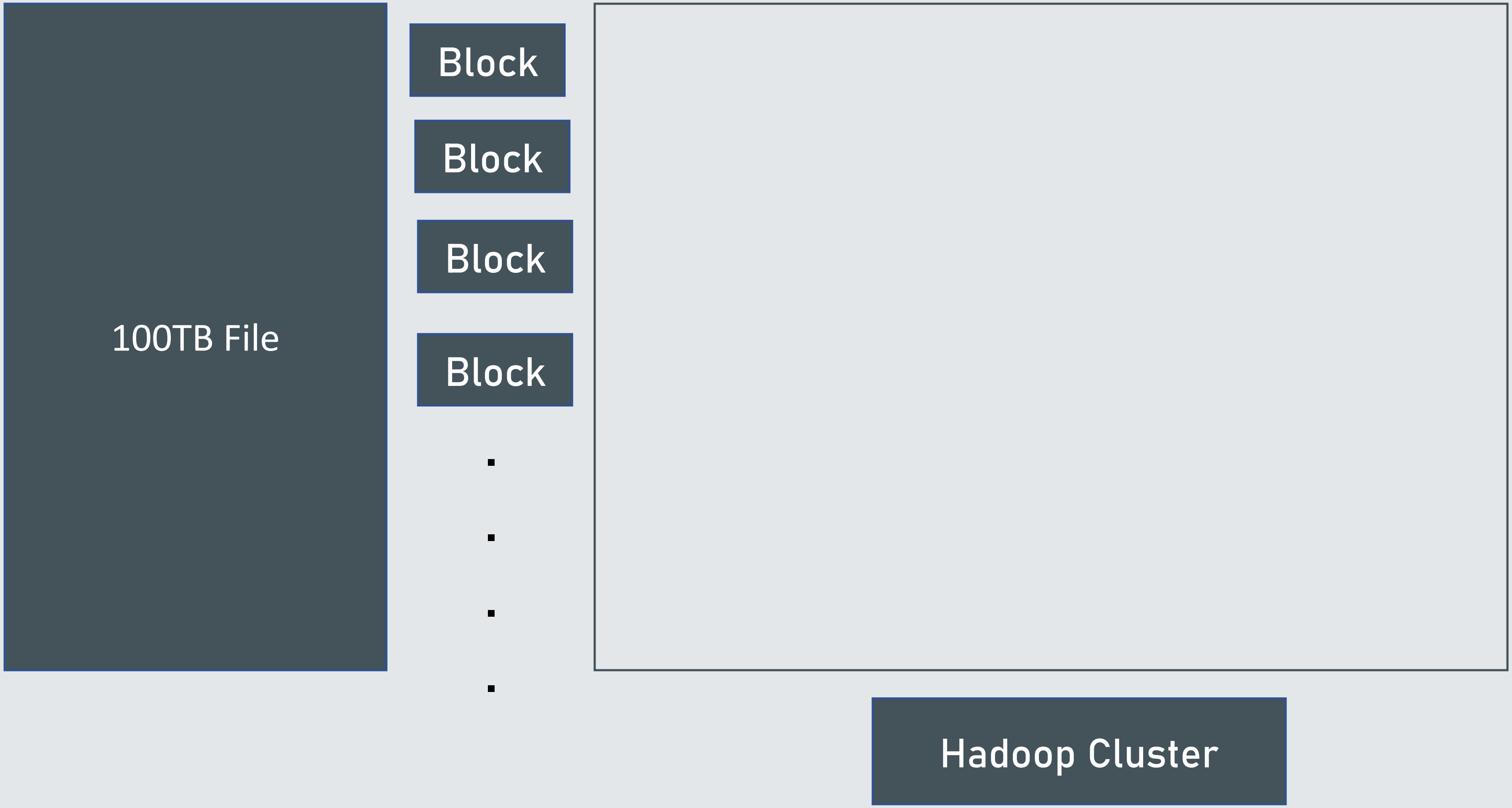
.

.

Hadoop Cluster

Example

When client copy this data in Hadoop



100TB File

Block

Block

Block

Block

.

.

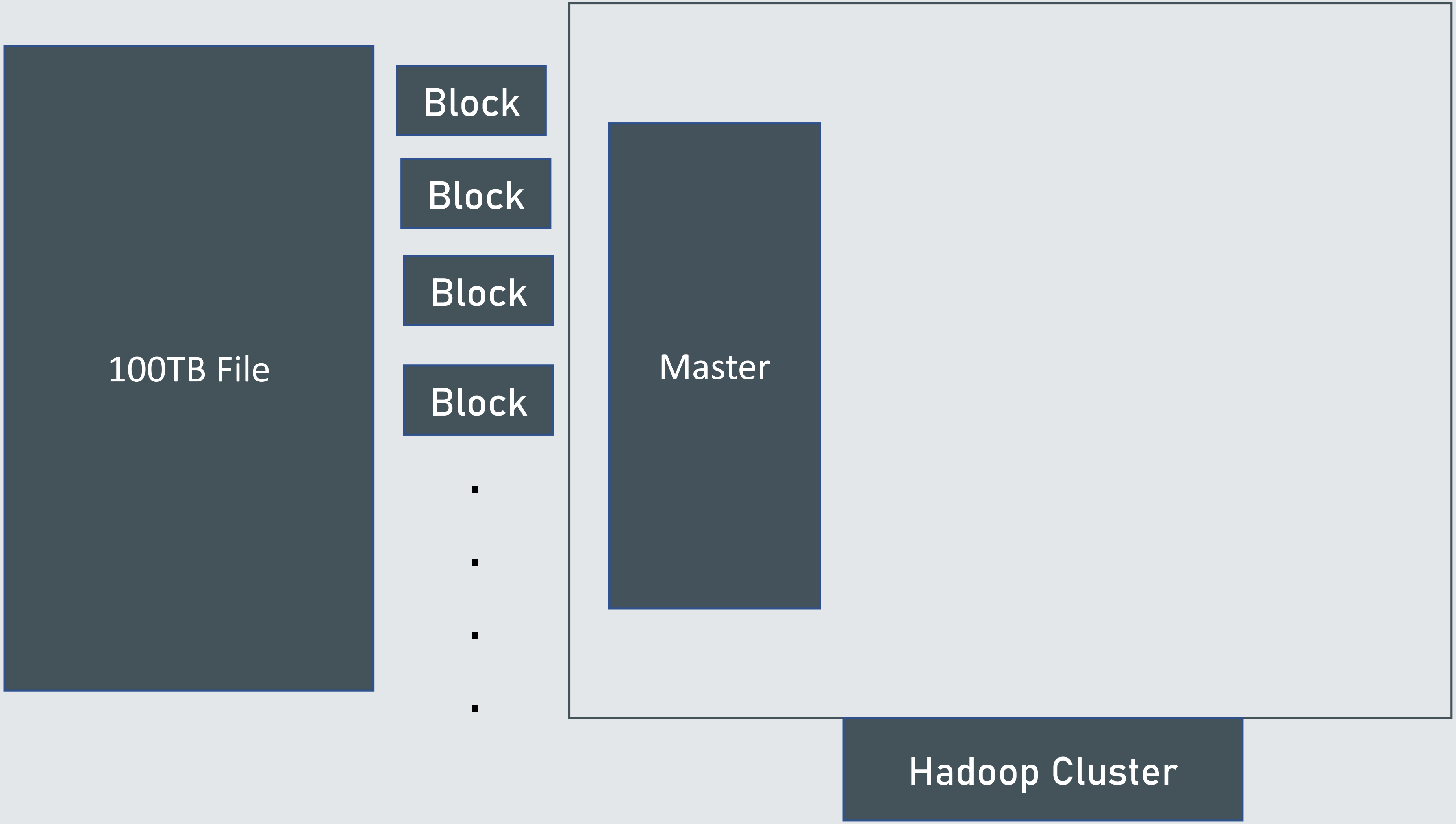
.

.

Hadoop Cluster

Example

When client copy this data in Hadoop



100TB File

Block

Block

Block

Block

.

.

.

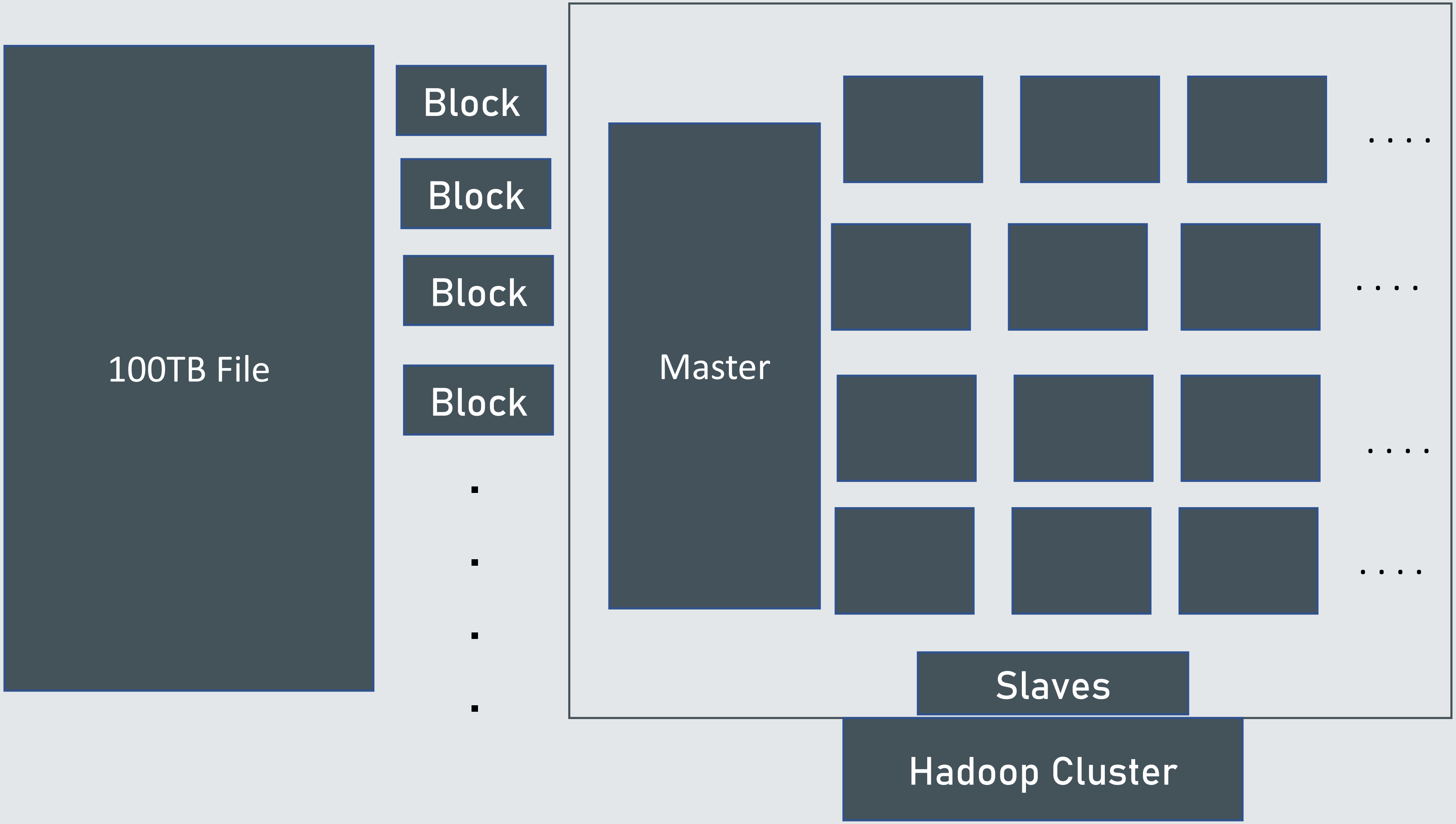
.

Master

Hadoop Cluster

Example

When client copy this data in Hadoop



100TB File

Block

Block

Block

Block

Master

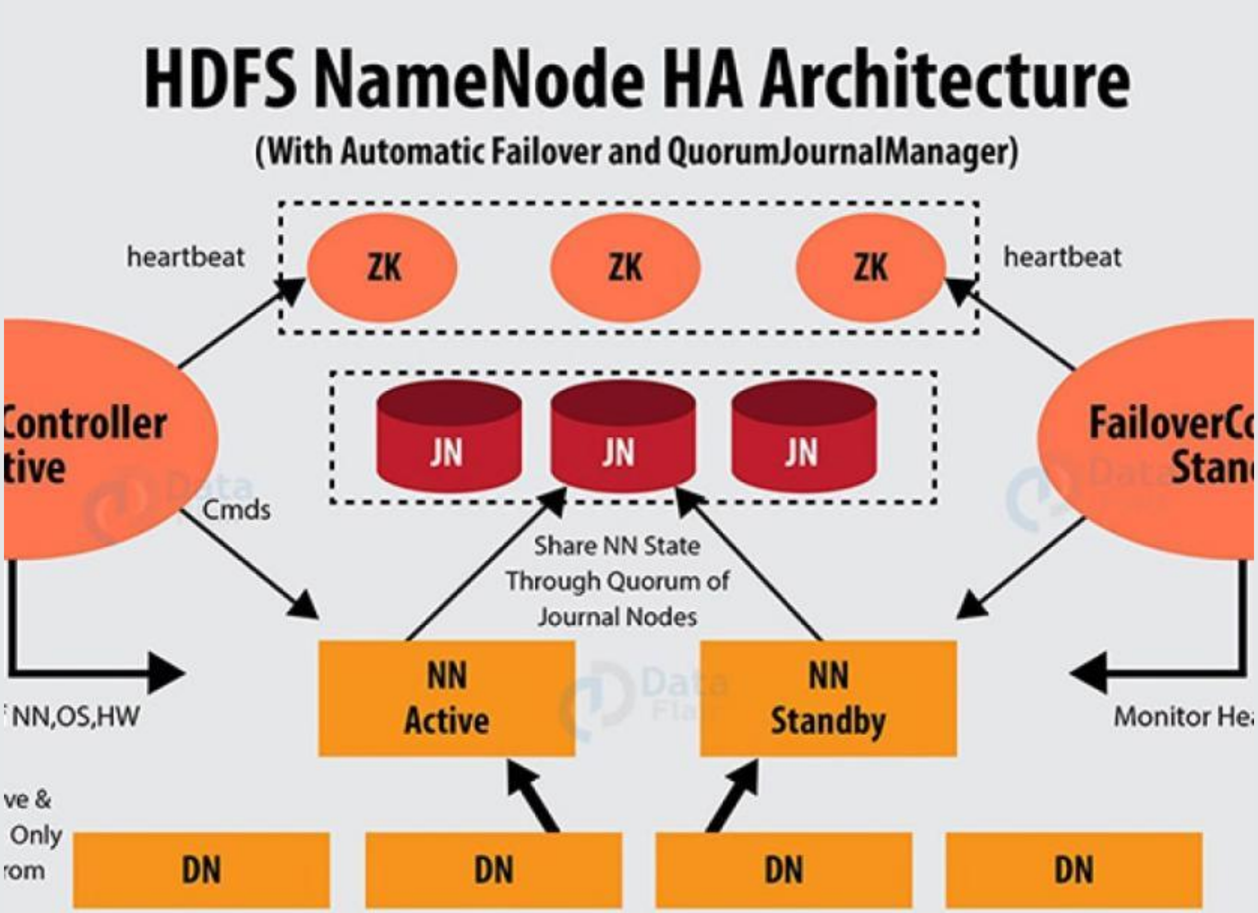
Slaves

Hadoop Cluster

What is Hadoop High Availability?

Single Point of Failure

What is Hadoop High Availability?

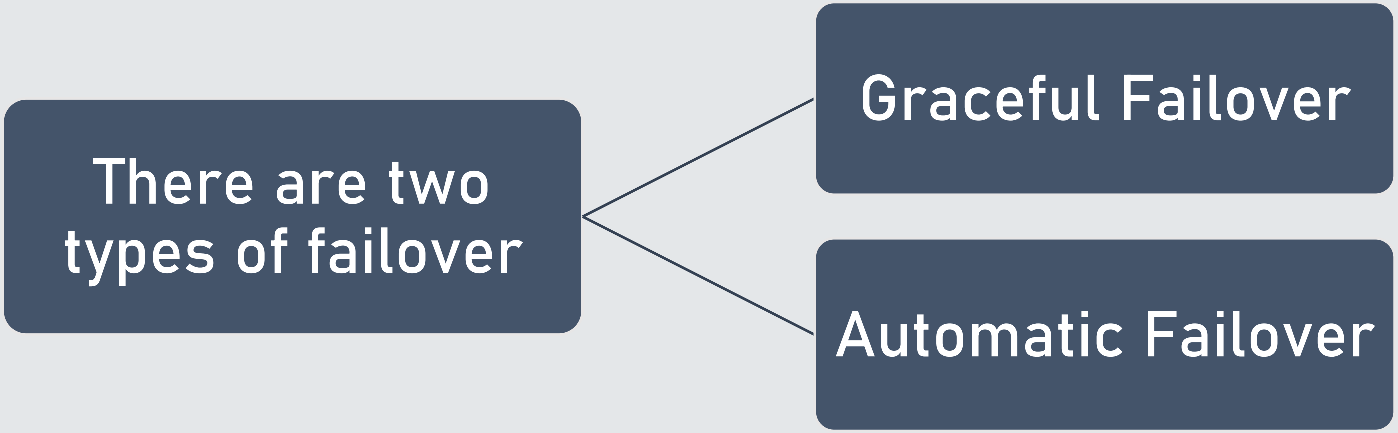


(Passive Standby Name Node) for automatic failover

What is Hadoop High Availability?

What is Failover?

Failover is a process in which the system transfers control to a secondary system in an event of failure.



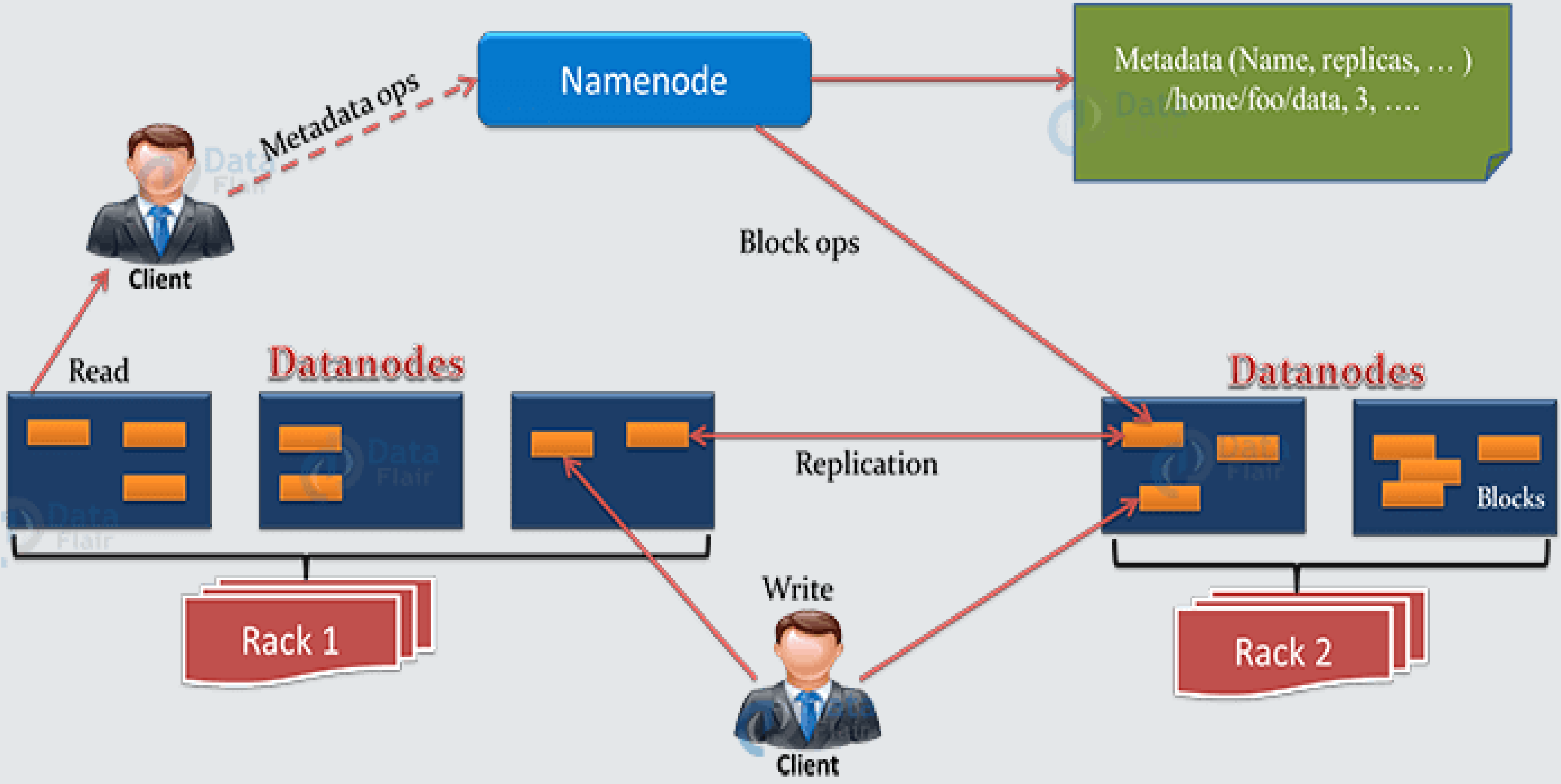
There are two types of failover

Graceful Failover

Automatic Failover

HDFS Architecture

HDFS Architecture



Hadoop HDFS Features

Distributed Storage

Blocks

Replication

High Availability

Data Reliability

Fault Tolerance

Scalability

- Vertical Scaling
- Horizontal Scaling

High throughput access to application data

Hadoop HDFS Features

Distributed Storage

Blocks

Replication

High Availability

Data Reliability

Fault Tolerance

Scalability

- Vertical Scaling
- Horizontal Scaling

High throughput access to application data

Hadoop HDFS Features

Distributed Storage

Blocks

Replication

High Availability

Data Reliability

Fault Tolerance

Scalability

- Vertical Scaling
- Horizontal Scaling

High throughput access to application data

Hadoop HDFS Features

Distributed Storage

Blocks

Replication

High Availability

Data Reliability

Fault Tolerance

Scalability

- Vertical Scaling
- Horizontal Scaling

High throughput access to application data

Hadoop HDFS Features

Distributed Storage

Blocks

Replication

High Availability

Data Reliability

Fault Tolerance

Scalability

- Vertical Scaling
- Horizontal Scaling

High throughput access to application data

Hadoop HDFS Features

Distributed Storage

Blocks

Replication

High Availability

Data Reliability

Fault Tolerance

Scalability

- Vertical Scaling
- Horizontal Scaling

High throughput access to application data

Hadoop HDFS Features

Distributed Storage

Blocks

Replication

High Availability

Data Reliability

Fault Tolerance

Scalability

- Vertical Scaling
- Horizontal Scaling

High throughput access to application data

Hadoop HDFS Features

Distributed Storage

Blocks

Replication

High Availability

Data Reliability

Fault Tolerance

Scalability

- Vertical Scaling
- Horizontal Scaling

High throughput access to application data

That’s all for now…