UNDERSTANDING HDFS

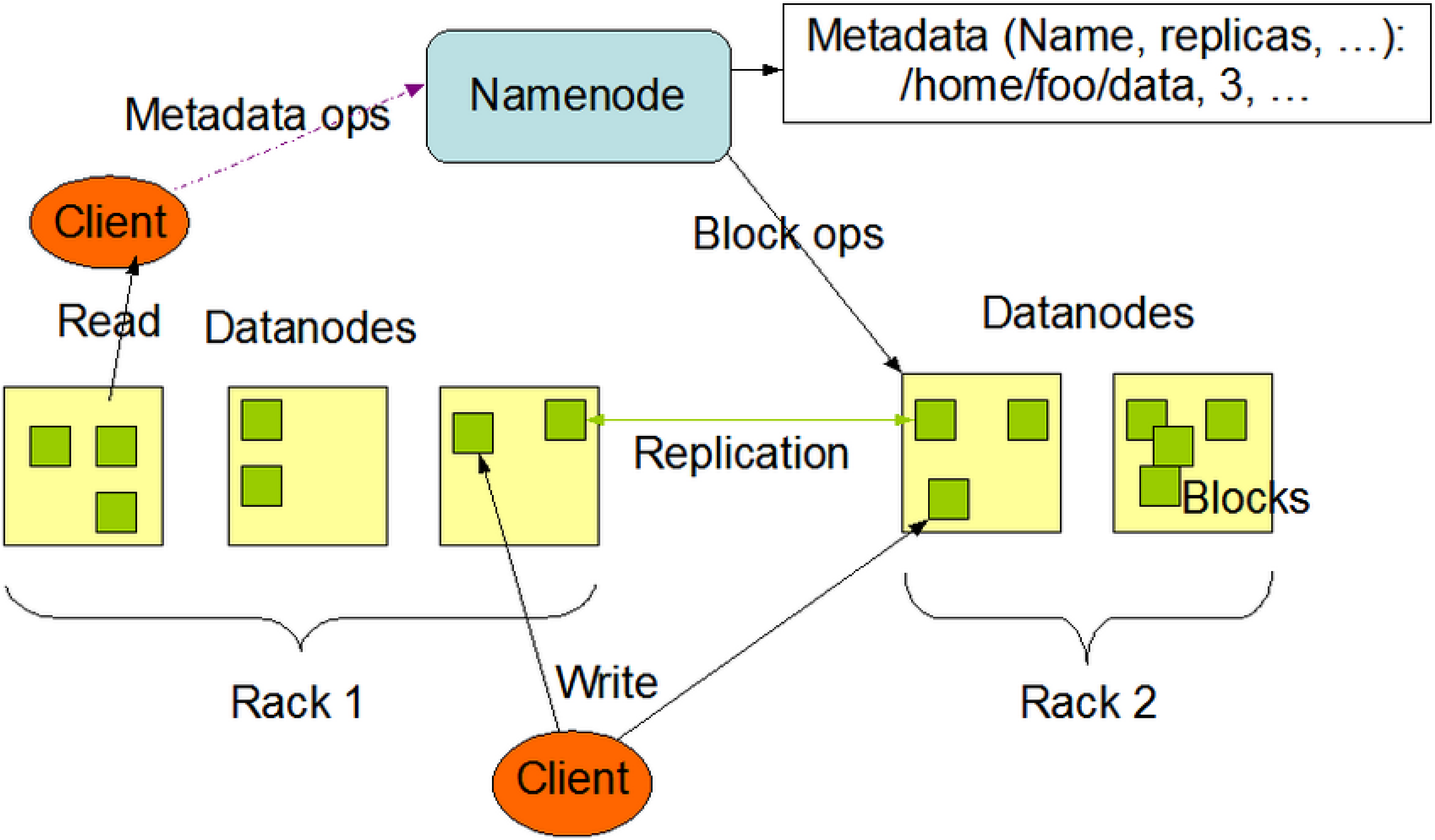
Introduction

HDFS (Hadoop Distributed File System) is the primary storage system used by Hadoop applications.

Introduction

HDFS is fault-tolerant and designed to be deployed on low-cost, commodity hardware.

HDFS provides high throughput data access to application data.


Metadata (Name, replicas, …):
/home/foo/data, 3, …
Namenode
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

HDFS Architecture

It focuses on NameNodes and DataNodes.

The NameNode is the hardware that contains the GNU/Linux operating system and software.

HDFS Architecture

NameNode works as a Master in a Hadoop cluster that guides the Datanode(Slaves).

Namenode is mainly used for storing the Metadata i.e. the data about the data.

HDFS Architecture

Namenode instructs the DataNodes with the operation like delete, create, Replicate, etc.

HDFS Architecture

A DataNode is hardware having the GNU/Linux operating system and DataNode software.

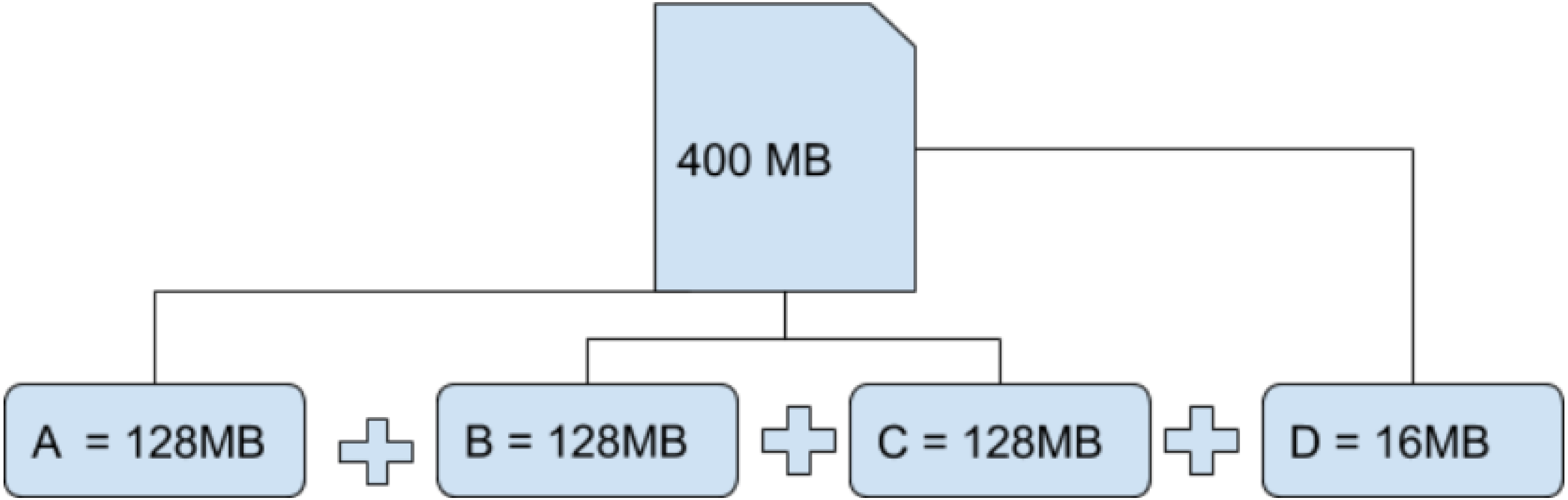
DataNodes are mainly utilized for storing the data in a Hadoop cluster.

File Block Size

Data in HDFS is always stored in terms of blocks.

So the single block of data is divided into multiple blocks of size 128MB

File Block Size



Replication Factor

Replication ensures the availability of the data.

Replication is making a copy of something and the number of times you make a copy of that particular thing can be expressed as it’s Replication Factor

Replication Factor

By default, the Replication Factor for Hadoop is set to 3 which can be configured means one can change it manually as per your requirement

Replication Factor

By default, the Replication Factor for Hadoop is set to 3 which can be configured means one can change it manually as per your requirement

File System Namespace

HDFS supports a traditional hierarchical file organization.

A user or an application can create directories and store files inside these directories.

File System Namespace

The file system namespace hierarchy is similar to most other existing file systems.

One can create and remove files, move a file from one directory to another, or rename a file.

Rack Awareness

The rack is nothing but just the physical collection of nodes in our Hadoop cluster (maybe 30 to 40).

A large Hadoop cluster is consists of so many Racks .

Rack Awareness

With the help of this racks information Namenode chooses the closest Datanode to achieve the maximum performance while performing the read/write information which reduces the Network Traffic.

Advantages of Hadoop Distributed File System

Fault tolerance

Speed

Compatibility and portability

Scalable

Advantages of Hadoop Distributed File System

Data locality

Cost effective

Stores large amounts of data

Flexible