Cauvery College for Women (Autonomous)

Nationally Accredited (III Cycle) with 'A' Grade by NAAC

Annamalai Nagar, Tiruchirappalli-18.

Name of the Faculty	: Dr Sinthu Janita
Designation	: Professor & Head
Department	: Computer Science
Contact Number	: 9894484436
Programme	: MSc Computer Science
Batch	: 2016-2017 Onwards
Semester	: IV
Course	: Big Data Analytics
Course Code	: P16CSE5A
Unit	: IV
Topics Covered	: Hadoop Architecture

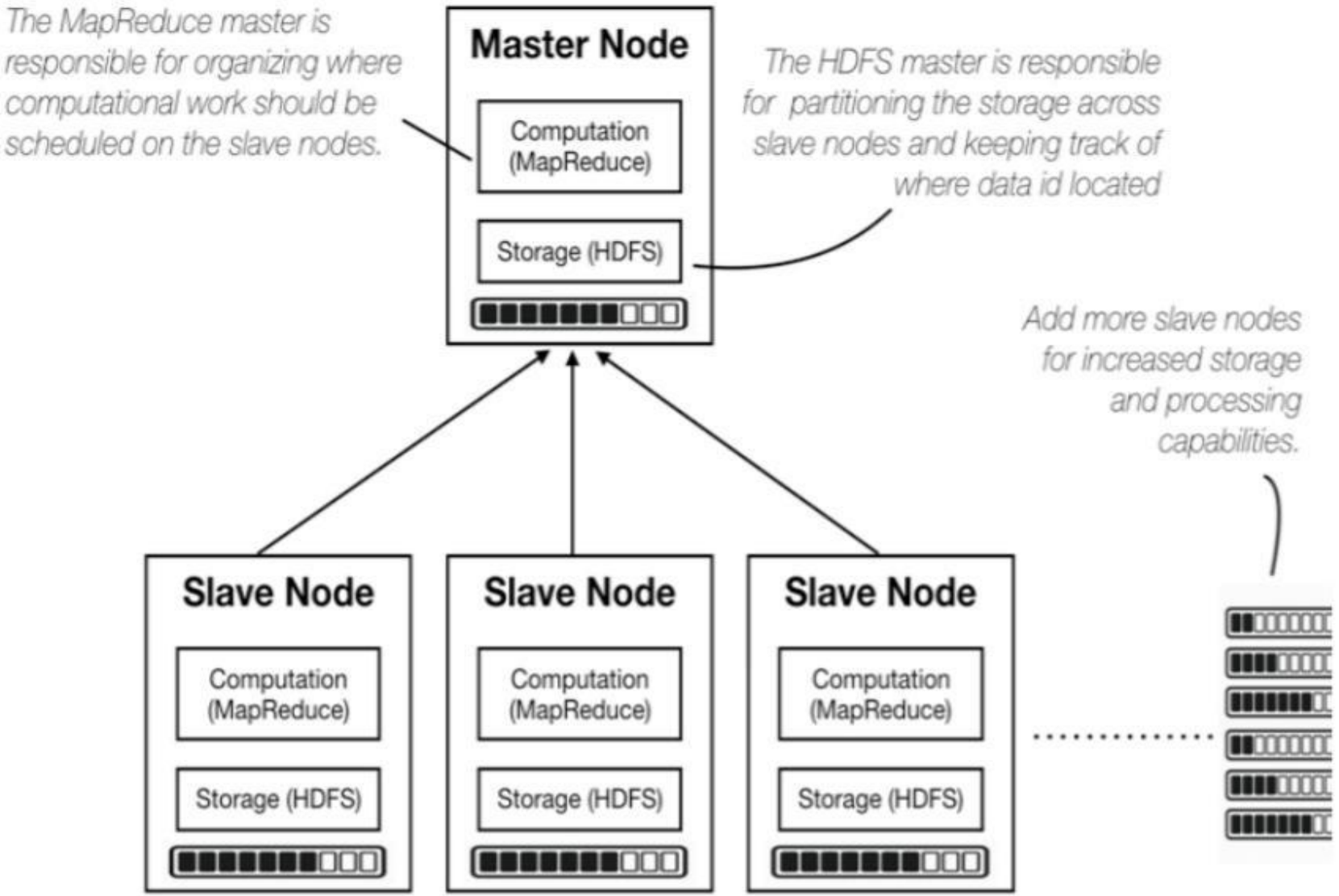
Hadoop Foundation for Analytics

Unit IV

Hadoop Foundation for Analytics:

History, Needs, Features, Key Advantage and Versions of Hadoop, Essential of Hadoop ecosystem, RDBMS versus Hadoop, Key Aspects and Components of Hadoop, Hadoop Architectures

Hadoop Architecture



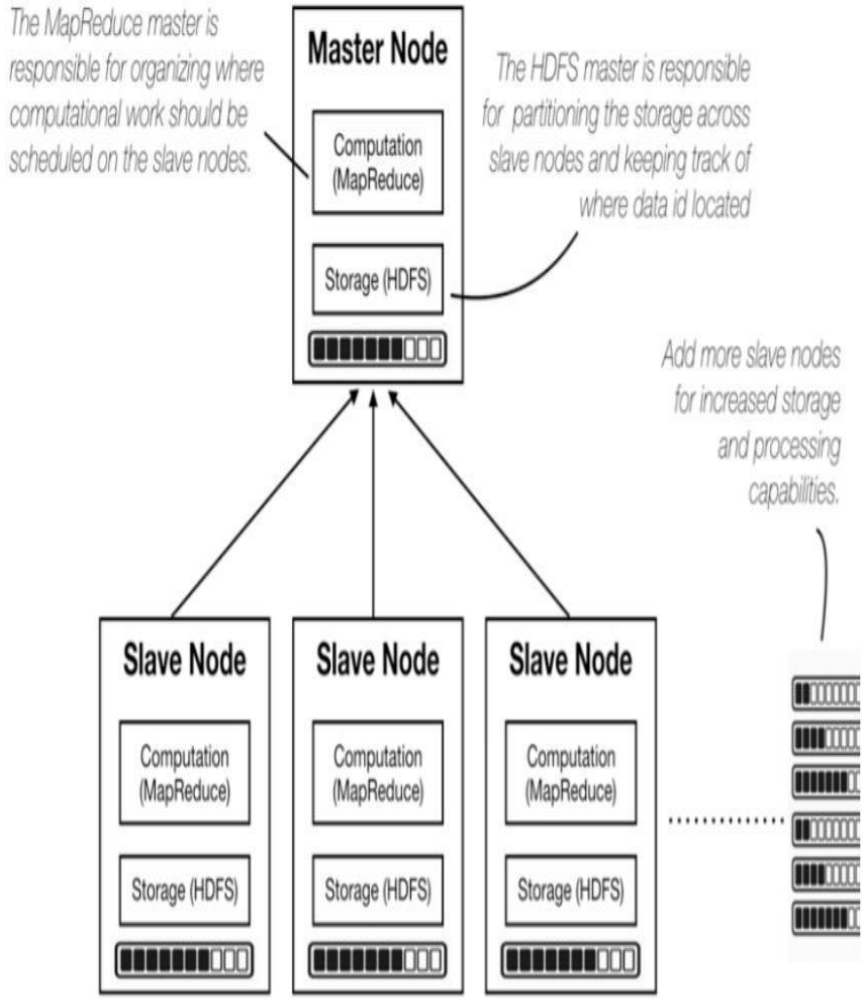
Hadoop Architecture

Hadoop is a distributed Master-Slave Architecture.

Master node is Name Node and Slave Node is Data Node

Master HDFS: Its main responsibility is partitioning the data storage across the slave nodes. It also keeps track of locations of data on Data Nodes

Master MapReduce: Decides and schedules computation task on slave nodes.

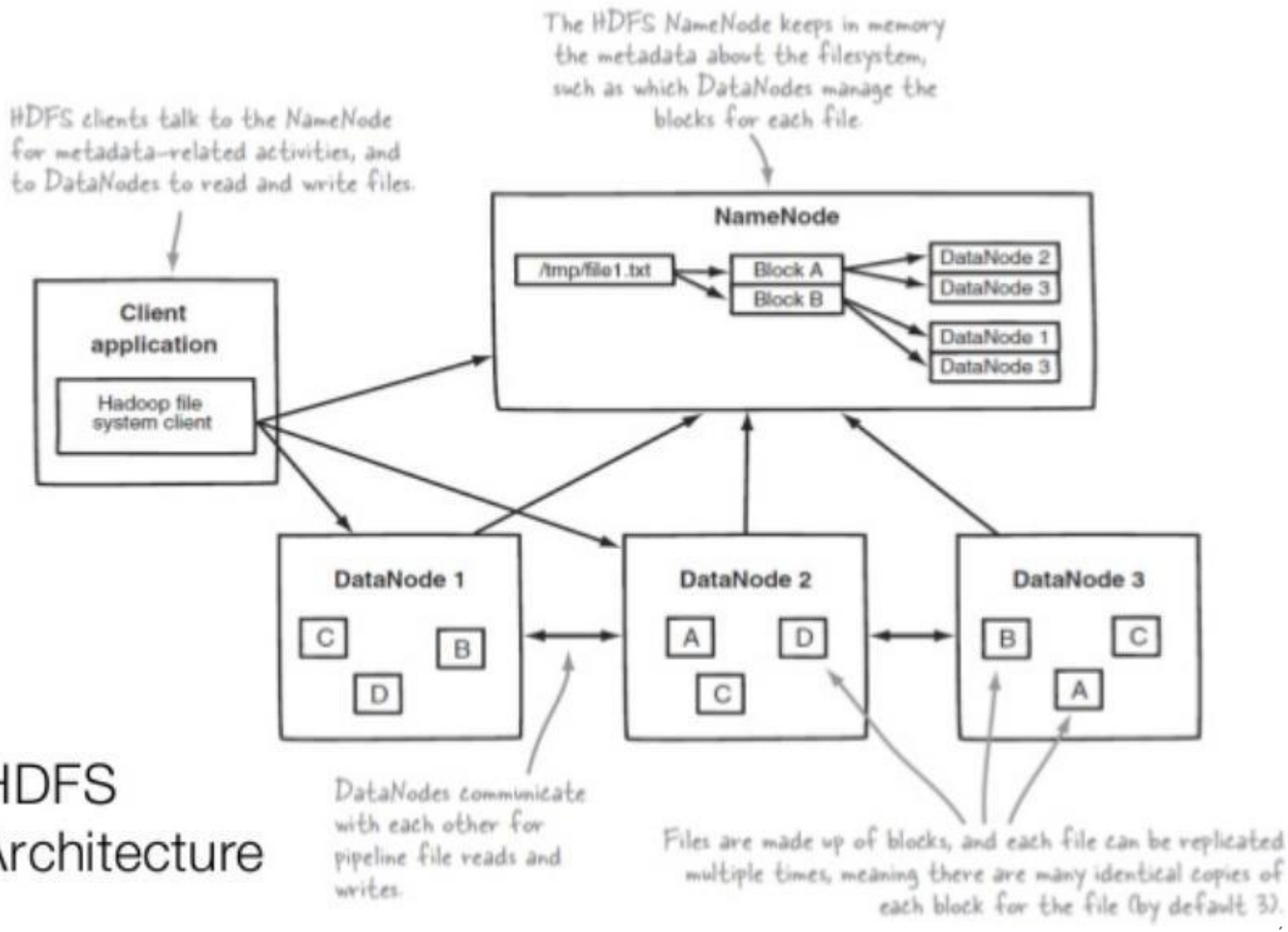


HDFS (Hadoop Distributed File System)

Key Points

- Storage Component of Hadoop
- Distributed File System
- Modeled after Google System
- Optimized for high throughput
- You can replicate a file for a configured number of times which is tolerant in terms of both software and hardware
- Replicates data blocks automatically on nodes that have failed
- You can realize the power of HDFS when you performread or write on large files (gigabytes and larger)
- Sits on top of native file system such as ext3 and ext4

HDFS Architecture



HDFS Architecture

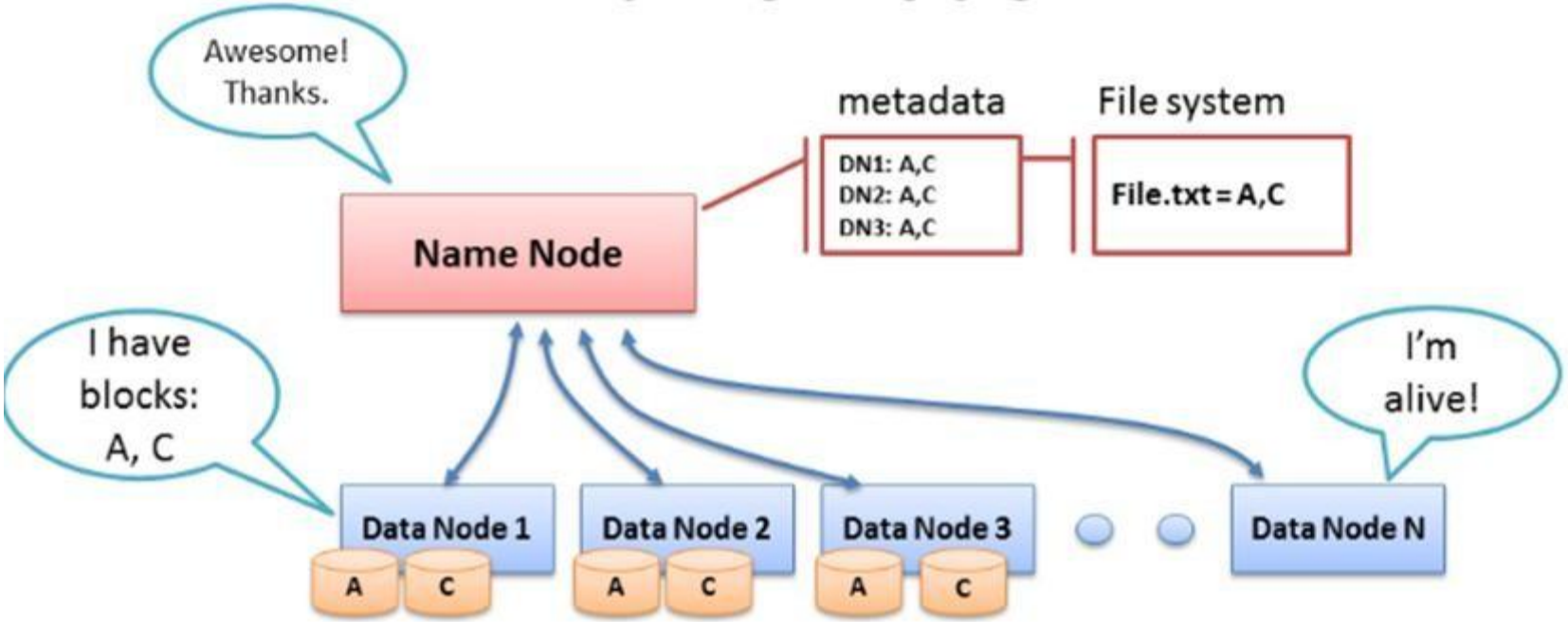
HDFS Daemons

NameNode

- HDFS breaks a large file into a smaller pieces called blocks
- NameNode uses a rackID to identify DataNodes in the rack
- A rack is a collection of DataNodes with in the cluster
- NameNode keeps tracks of block of a file as it is placed on various DataNodes
- NameNode manages file related options such as read, write, create and delete
- Its main job is managing the File System Namespace
- A file system Namespace is a collection of files in the cluster

•Name Node stores HDFS namespace

Name Node



- Data Node sends Heartbeats
- Every 10th heartbeat is a Block report
- Name Node builds metadata from Block reports
- TCP – every 3 seconds
- If Name Node is down, HDFS is down

HDFS Daemons ctd..

DataNode

- There are multiple DataNodes per cluster
- During Pipeline read and write DataNodes communicate with each other
- A DataNode also continuously send “heartbeat” message to NameNode to ensure the connectivity between the NameNode and DataNode
- If there is no heartbeat from a DataNode , the NameNode replicates the DataNode within the cluster and keeps on running as if nothing has happened

HDFS Daemons ctd..

Secondary NameNode

- The Secondary NameNode takes a snapshot of HDFS metadata at intervals specified in the Hadoop configuration

- Since the memory requirement of Secondary NameNode are the same as NameNode, it is better to run NameNode and Secondary NameNode on different machines

- In case of failure of NameNode , the secondary NameNode can be configured manually to bring up the cluster however the Secondary NameNode doesn't record any real time changes that happened to the HDFS metadata .