SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)

Re-accredited by NAAC with A+ grade, Accredited by NBA(CSE, IT, ECE, EEE & Mechanical)
Approvedy by AICTE, New Delhi, Recognized by UGC, Affiliated to Anna University, Chennai

Department of MCA

Topic: Hadoop Cluster

COURSE

19CAT702

Big Data Analytics

UNIT - II

Hadoop

CLASS

III Semester / II MCA

Session Objectives

❑ Understand the architecture of Hadoop cluster

❑ Prepare a cluster of nodes for Hadoop to run the user job

Hadoop Cluster

- ❑ Collection of interconnected computers with the capable of communicating with each other and work as a single unit on a given task
- ❑ Hadoop cluster deigned to store and management huge volume of data and to perform analysis
- ❑ Basically it has master and number of slaves
- ❑ **Advantages**: Scalability, cost effective, flexible, resilient to failure

Example

32,000 Nodes in a cluster

YAHOO!

4,000 Nodes in a cluster

facebook

5,000 Nodes in a cluster

kubernetes

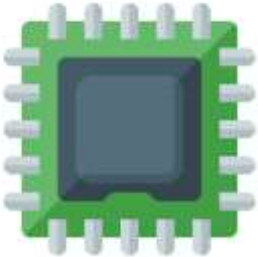
<100 to 1000 Nodes in a cluster

IBM

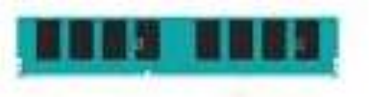
Hadoop Cluster: H/W Configuration

- Typical choice of computer node running a Hadoop datanode and tasktracker would have specifications (Yr 2010)

Processor2 quad-core 2-2.5GHz CPUs



Memory 16-24 GB ECC RAM*



Storage 4 × 1TB SATA disks

Network Ethernet

Hadoop Cluster Architecture

Master Nodes

- ❑ NameNode, Secondary NameNode, and JobTracker
- ❑ Utilize higher quality hardware
- ❑ Responsible for storing data in HDFS & overseeing MapReduce operations

Worker Nodes

- ❑ Virtual machines, running both DataNode and TaskTracker services
- ❑ Do the actual work of storing and processing the jobs as directed by the master nodes

Hadoop Cluster Architecture

Client Nodes

- ❑ loading the data into the cluster
- ❑ First submit MapReduce jobs describing how data needs to be processed,
- ❑ Then fetch the results once the processing is finished

Cluster Type

Single Node

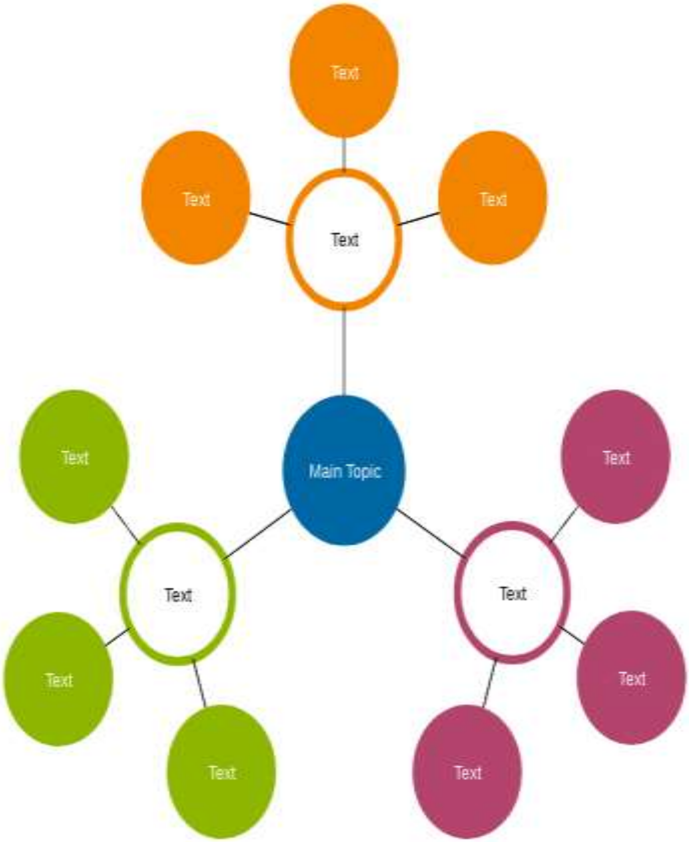
- ❑ All daemons like NameNode, DataNode run on the same machine
- ❑ All the processes run on one JVM instance
- ❑ Need not make any configuration setting

Multiple Nodes

- ❑ Daemons run on separate host
- ❑ Slave daemons run on low commodity h/w
- ❑ Slave machines can be present in any location

Cluster Management

- ❑ Tool should provide
  - work-load management
  - security
  - resource provisioning
  - performance optimization
  - health monitoring
  - policy management
  - job scheduling
  - back up and recovery
- ❑ High availability of NameNode
- ❑ Policy based controls



Network Topology

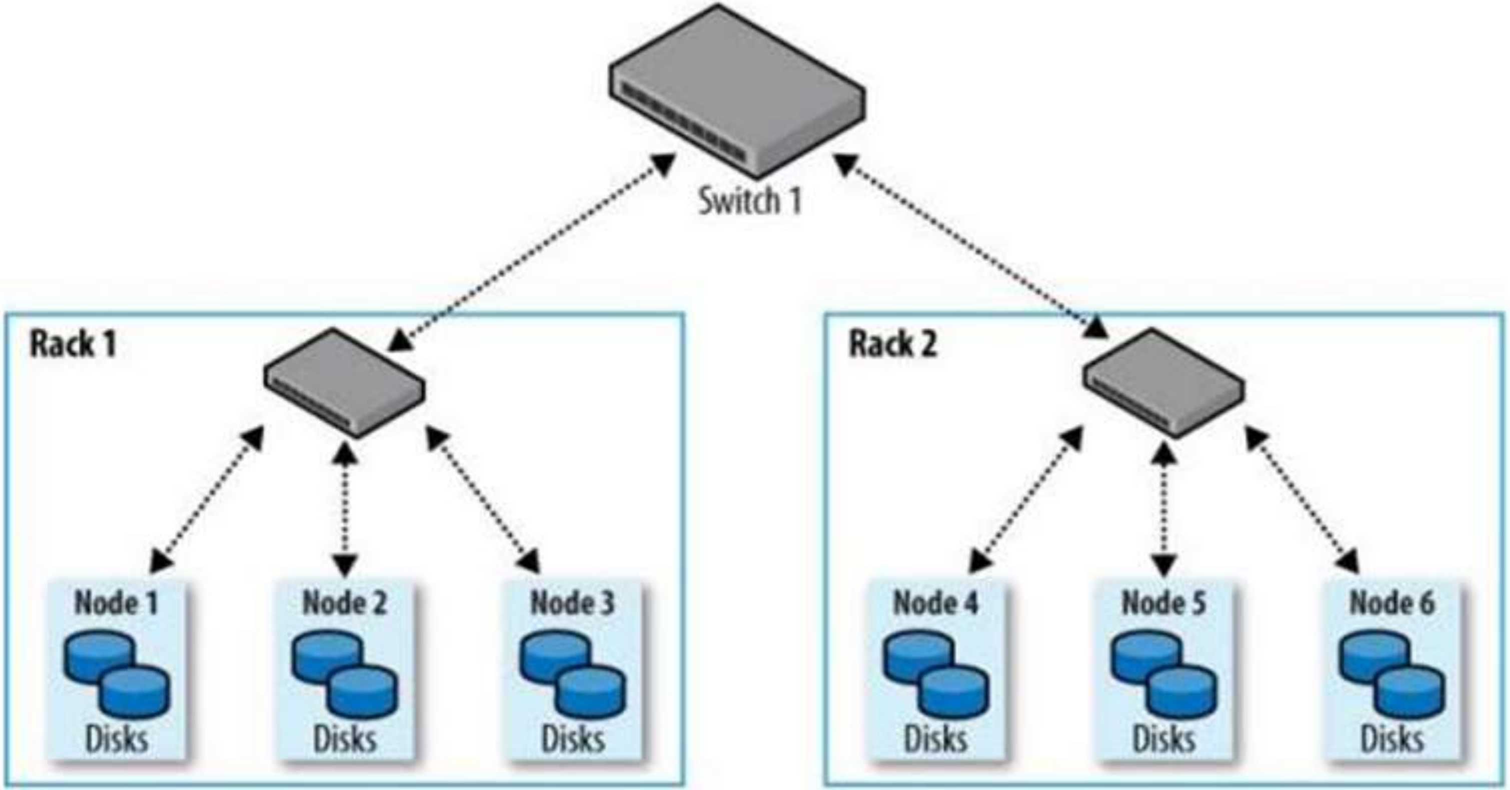
- ❑ Hadoop cluster architecture consists of a two-level network topology
- ❑ There are 30 to 40 servers per rack, with a 1 GB switch for the rack, and an uplink to a core switch or router
- ❑ If cluster runs on a single rack, then there is nothing more to do
- ❑ Namenode uses the network location to determine where to place block replicas

Network Topolgy

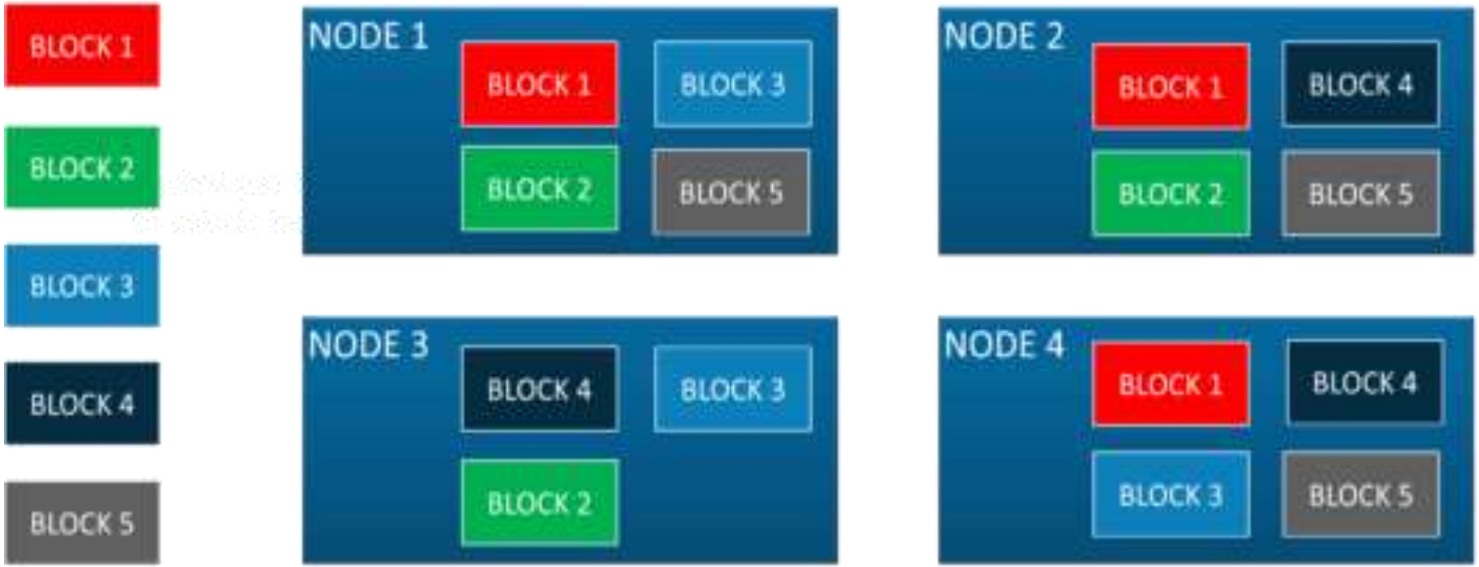
- Jobtracker uses network location to determine where the closest replica is as input for a map task that is scheduled to run on a tasktracker

- There are two network locations
  - /switch1/rack1 and
  - /switch1/rack2

Network Topology



Rack Awareness



BLOCK 1

BLOCK 2

BLOCK 3

BLOCK 4

BLOCK 5

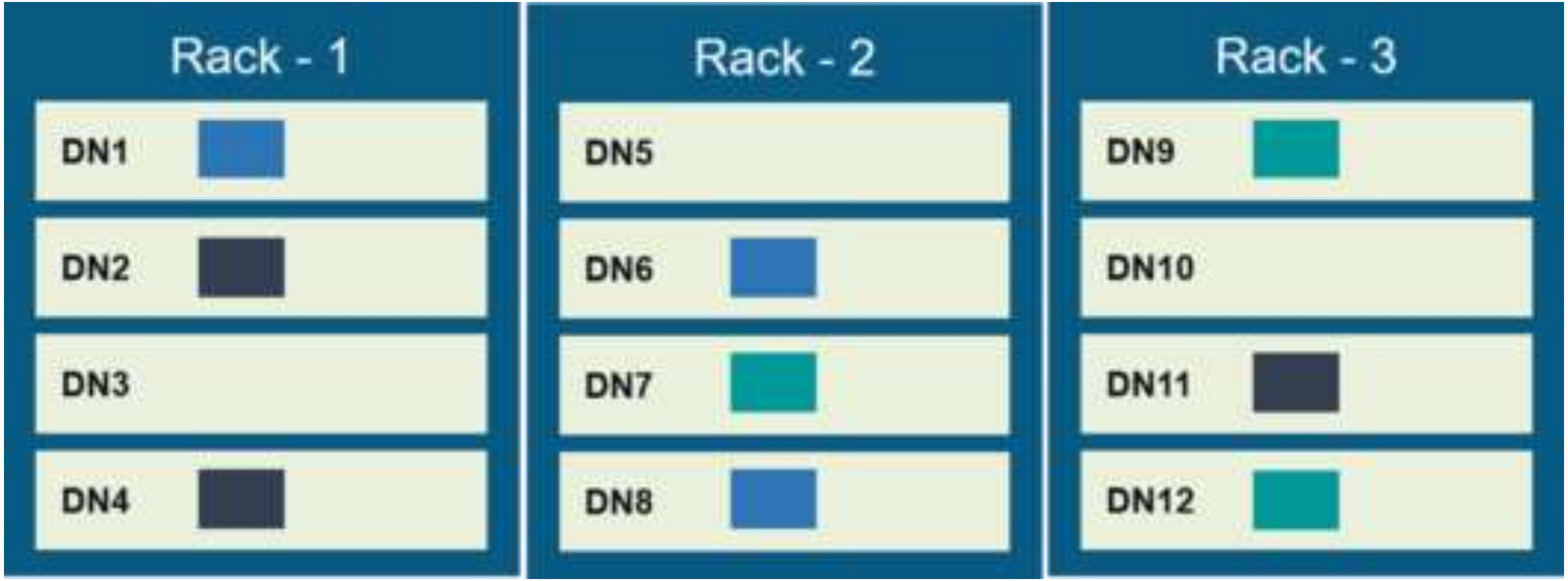
NODE 1
- BLOCK 1
- BLOCK 3
- BLOCK 2
- BLOCK 5

NODE 2
- BLOCK 1
- BLOCK 4
- BLOCK 2
- BLOCK 5

NODE 3
- BLOCK 4
- BLOCK 3
- BLOCK 2

NODE 4
- BLOCK 1
- BLOCK 4
- BLOCK 3
- BLOCK 5

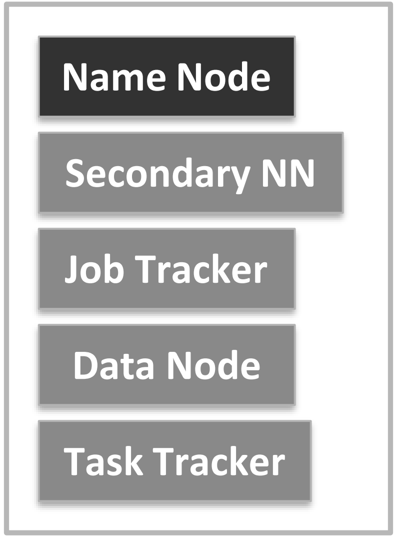
Rack Awareness



Rack - 1	Rack - 2	Rack - 3
DN1 🟦	DN5	DN9 🟩
DN2 ⬛	DN6 🟦	DN10
DN3	DN7 🟩	DN11 ⬛
DN4 ⬛	DN8 🟦	DN12 🟩

Cluster Management

Single Node Cluster



Name Node

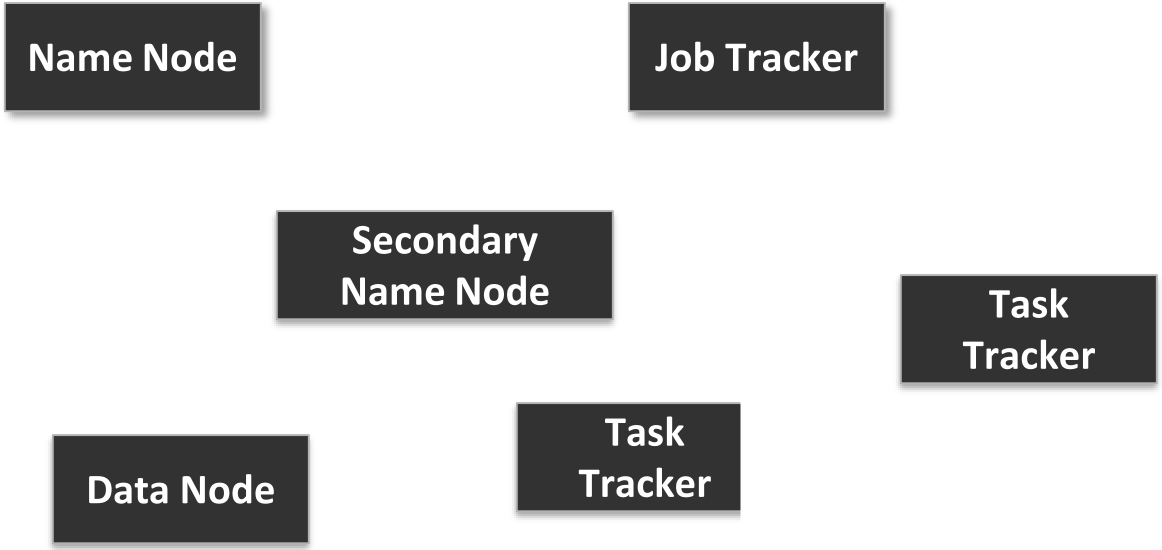
Secondary NN

Job Tracker

Data Node

Task Tracker

Multi Node Cluster



Name Node

Job Tracker

Secondary Name Node

Task Tracker

Task Tracker

Data Node

Facebook Data Center

https://www.youtube.com/watch?v=4A_A-CmrqpQ

Assessment

1. Client node in cluster performs

   A. Manages HDFS Ops B. Performs data Ops. C. Load data into cluster

2. Default replication factor for multinodes cluster is

   A. 1 B. 2 C. 3 D. 4

3. Configuration setting is not required for

   A. Single node cluster B. Multi nodes cluster C. Multidimensional Cluster

References

- ❑ Tom White, “ Hadoop: The Definitive Guide” Third Edition, O’reilly Media, 4th Edition, 2012

Web Resources

- ❑ https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html
- ❑ https://docs.cloudera.com/HDPDocuments/HDP2/HDP-2.1.2/bk_getting-started-guide/content/ch_typical-hadoop-cluster.html
- ❑ https://techvidvan.com/tutorials/hadoop-cluster/

Thank U