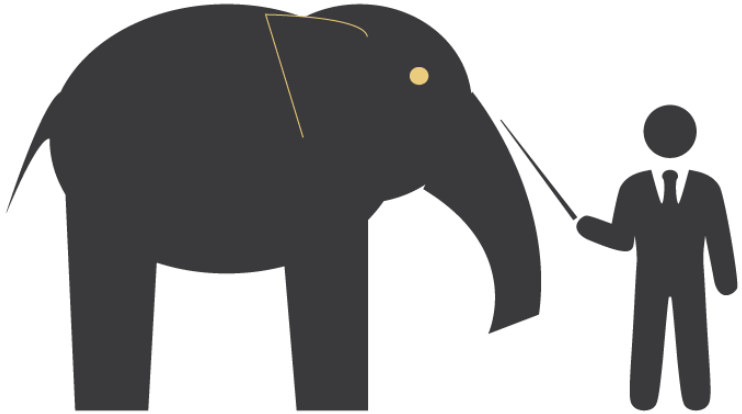
edureka!



TOP 5 HADOOP ADMIN TASKS

View Hadoop Administration Course at www.edureka.co/hadoop-admin

For Queries:
Post on Twitter @edurekaIN: #askEdureka
Post on Facebook /edurekaIN

For more details please contact us:
US : 1800 275 9730 (toll free)
INDIA : +91 88808 62004
Email Us : sales@edureka.co

Objectives

edureka!

At the end of this module, you will be able to

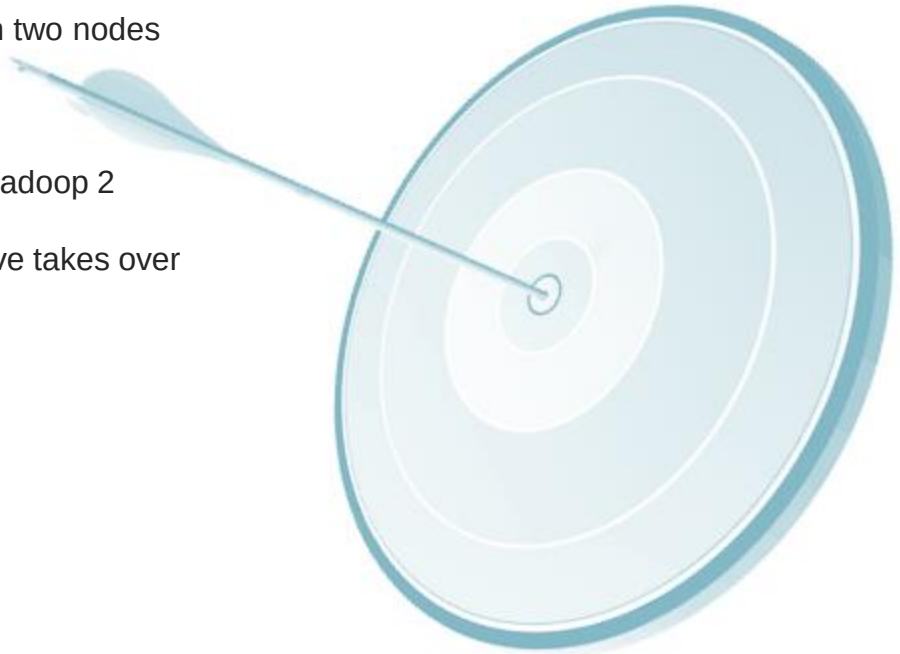
→ Understand Cluster Planning

→ Understand Hadoop fully distributed cluster setup with two nodes

→ Add further nodes to the running cluster

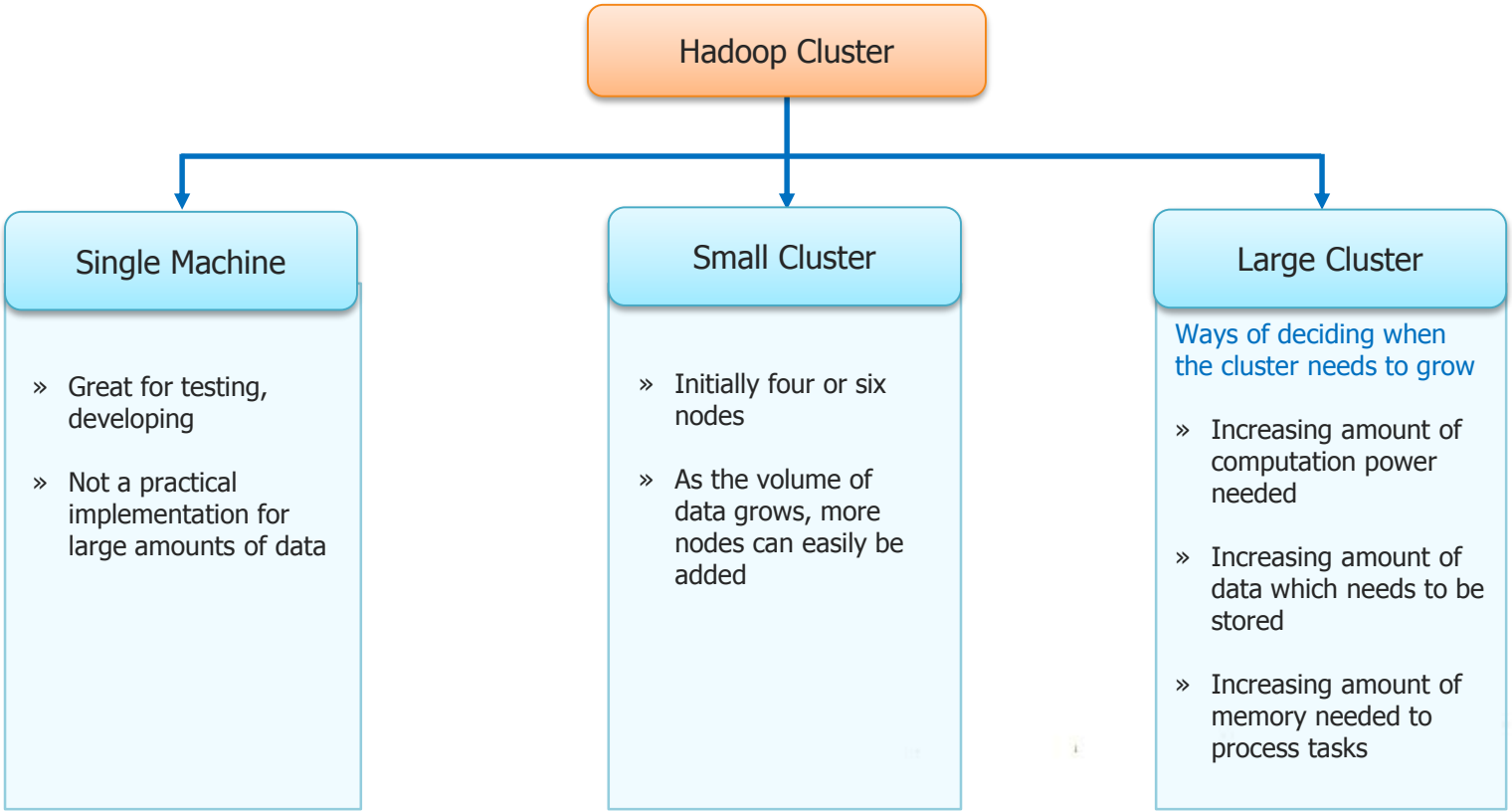
→ Upgrade existing Hadoop Cluster from Hadoop 1 to Hadoop 2

→ Understand Active NameNode Failure and how passive takes over



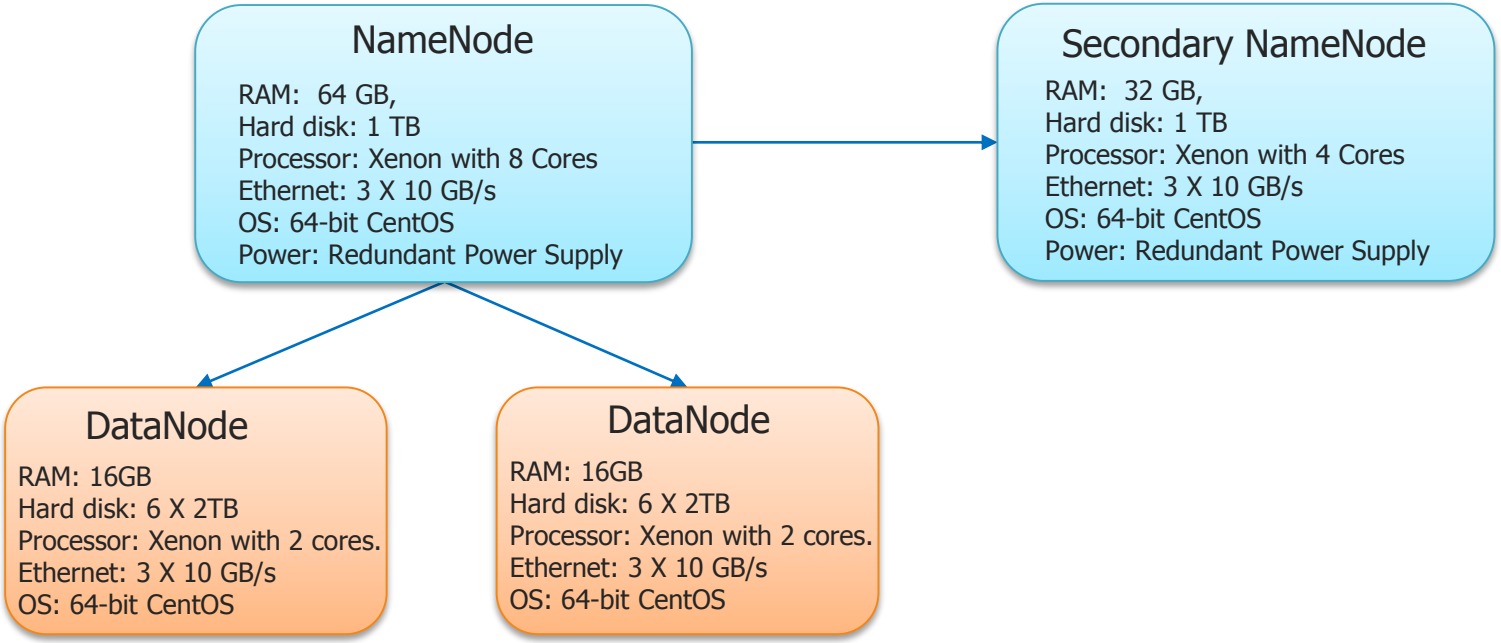
Hadoop Cluster: Thinking About The Problem

edureka!



Hadoop Cluster: A Typical Use Case

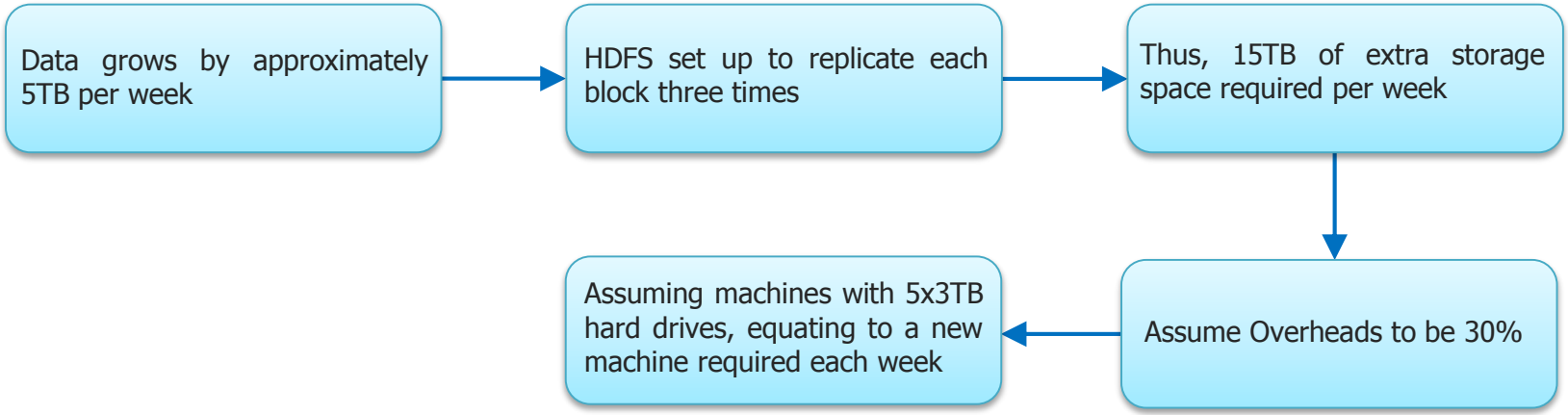
edureka!



Cluster Growth Based On Storage Capacity

edureka!

→ Seeking cluster growth on storage capacity is often a good method to use!

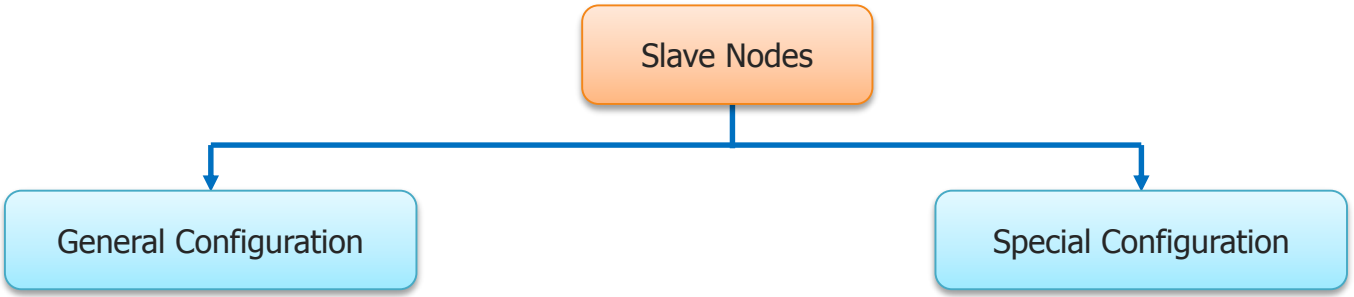


Slave Nodes: Recommended Configuration

edureka!

→ Higher-performance vs lower performance components

Save the Money, Buy more Nodes!



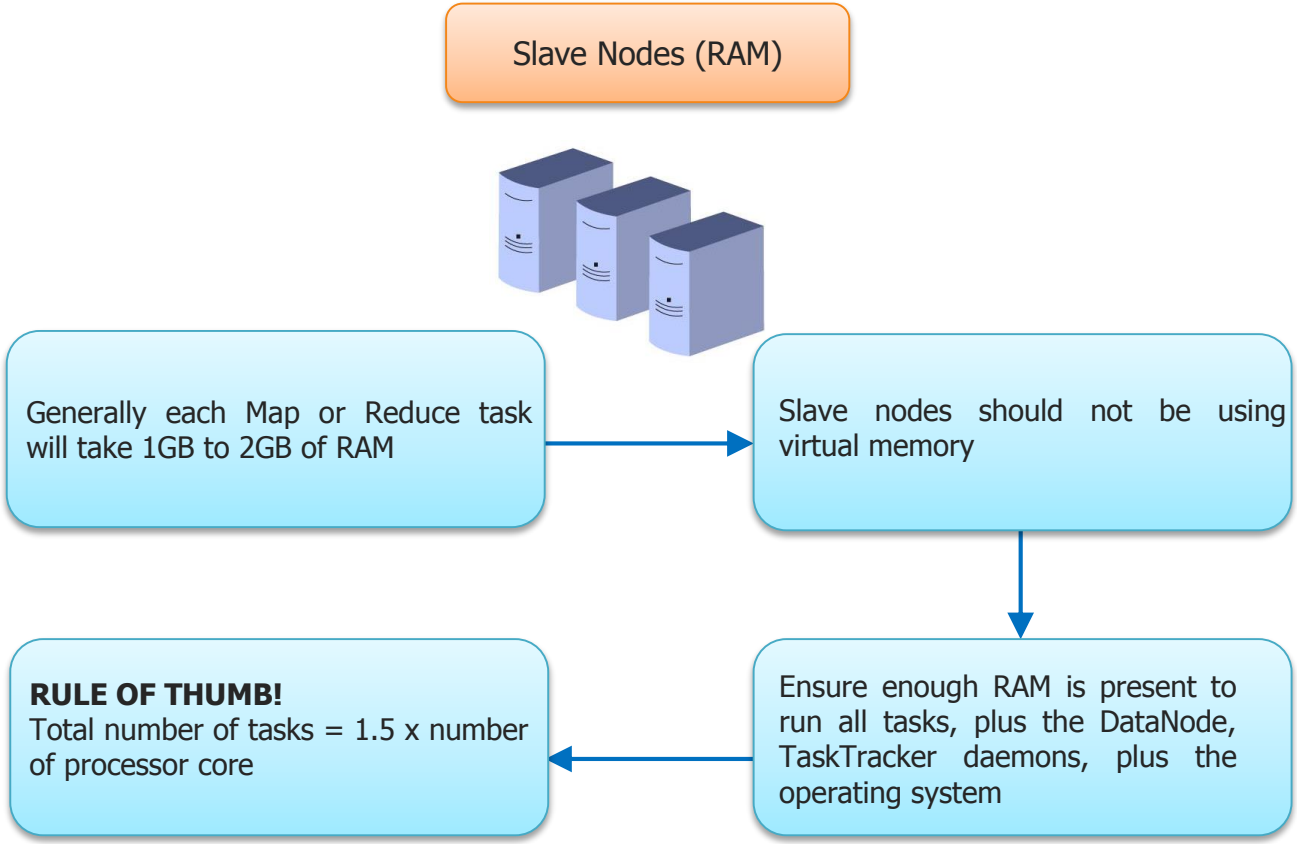
→ General (Depends on requirement 'base' configuration for a slave Node
- » 4 x 1 TB or 2 TB hard drives, in a JBOD* configuration
- » Do not use RAID!
- » 2 x Quad-core CPUs
- » 24 -32GB RAM
- » Gigabit Ethernet

→ Multiples of (1 hard drive + 2 cores + 6-8GB RAM) generally work well for many types of applications

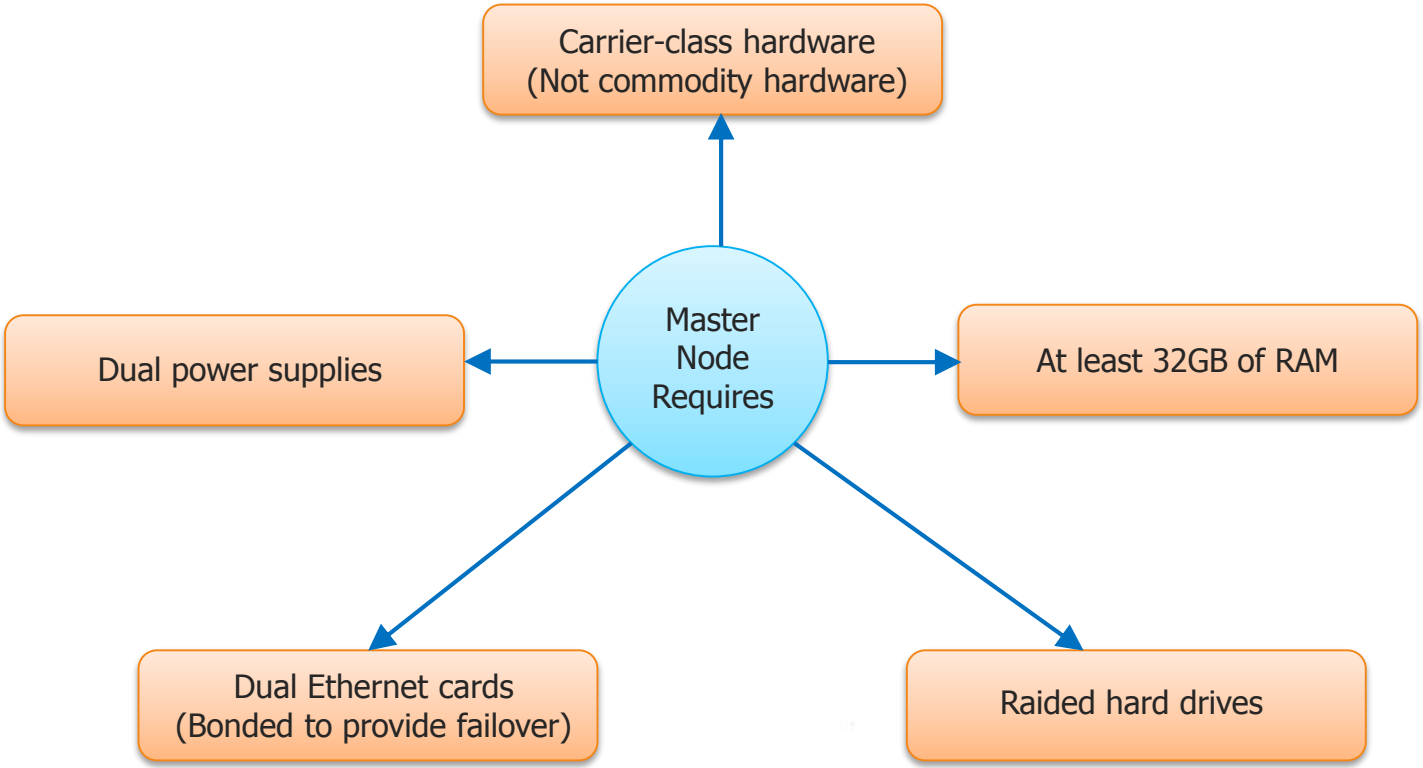
"A cluster with more nodes performs better than one with fewer, slightly faster nodes"

Slave Nodes: More Details (RAM)

edureka!



Master Node Hardware Recommendations



Fully Distributed Mode Cluster

edureka!

→ Hadoop requires certain ports on each nodes accessible via the network

→ However, the default firewall iptables prohibit these ports being accessed

→ To run hadoop applications, you must make sure that these ports are open

→ To check the status of iptables, you can use these commands under root privilege:

```
/etc/init.d/iptables status
```

→ You can simply turn iptables off, or at least open these ports:

```
9000, 9001, 50010, 50020, 50030, 50060, 50070, 50075, 50090
```

Hadoop Cluster Modes

edureka!

Hadoop can run in any of the following three modes:

Standalone (or Local) Mode

→ No daemons, everything runs in a single JVM
→ Suitable for running MapReduce programs during development
→ Has no DFS

Pseudo-Distributed Mode

→ Hadoop daemons run on the local machine

Fully-Distributed Mode

→ Hadoop daemons run on a cluster of machines

Hadoop Cluster

edureka!

→ Create Dedicated User and Group

» Hadoop requires all the nodes in the cluster have exactly the same structure of directory in which hadoop was installed
» It will be beneficial if we create a dedicated user (e.g."hadoop") and install hadoop in its home folder
» You must have root privilege on each nodes to carry on the following steps
» To change to "root", type in "su -" in the terminal and input the password for "root"

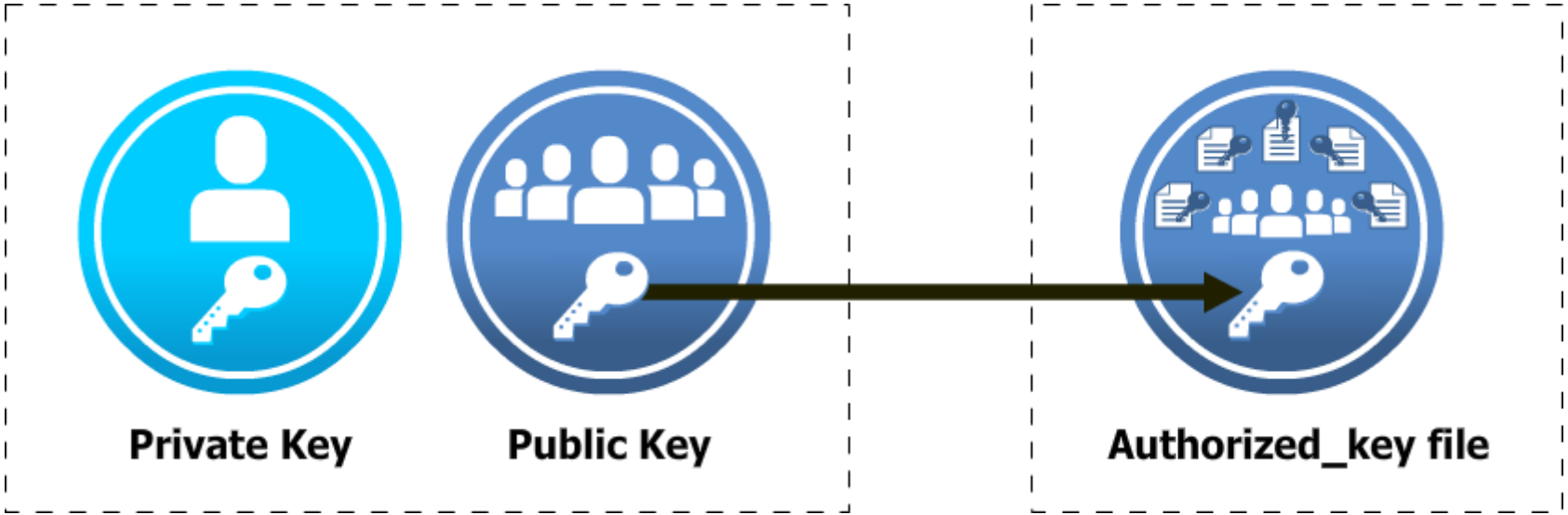
Create group "hadoop user":
groupadd hadoop use

Create user "hadoop":
useradd -g hadoop user -s /bin/bash -d /home/hadoop hadoop
in which -g specifies user "hadoop" belongs to group "hadoop user", -s specifies the shell to use, -d specifies the home folder for user "hadoop".

Set password for user "hadoop":
passwd hadoop

→ Then type in the password for user "hadoop" twice.
→ Then type in "su - hadoop" to change to user "hadoop".

Passwordless ssh



Configuration Files

edureka!

Configuration Filenames	Description of Log Files
hadoop-env.sh yarn-env.sh	Settings for Hadoop Daemon's process environment.
core-site.xml	Configuration settings for Hadoop Core such as I/O settings that common to both HDFS and YARN.
hdfs-site.xml	Configuration settings for HDFS Daemons, the Name Node and the Data Nodes.
yarn-site.xml	Configuration setting for Resource Manager and Node Manager.
mapred-site.xml	Configuration settings for MapReduce Applications.
slaves	A list of machines (one per line) that each run DataNode and Node Manager.

Configuration Files (Contd.)

edureka!

The core functionality and usage of these core configuration files are same in Hadoop 2.0 and 1.0 but many new properties have been added and many have been deprecated

For example:

→ 'fs.default.name' has been deprecated and replaced with 'fs.defaultFS' for YARN in core-site.xml

→ 'dfs.nameservices' has been added to enable NameNode High Availability in hdfs-site.xml

Deprecated Property Name	New Property Name
dfs.data.dir	dfs.datanode.data.dir
dfs.http.address	dfs.namenode.http-address
fs.default.name	fs.defaultFS

→ In Hadoop 2.2.0 release, you can use either the old or the new properties

→ The old property names are now deprecated, but still work!

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/DeprecatedProperties.html

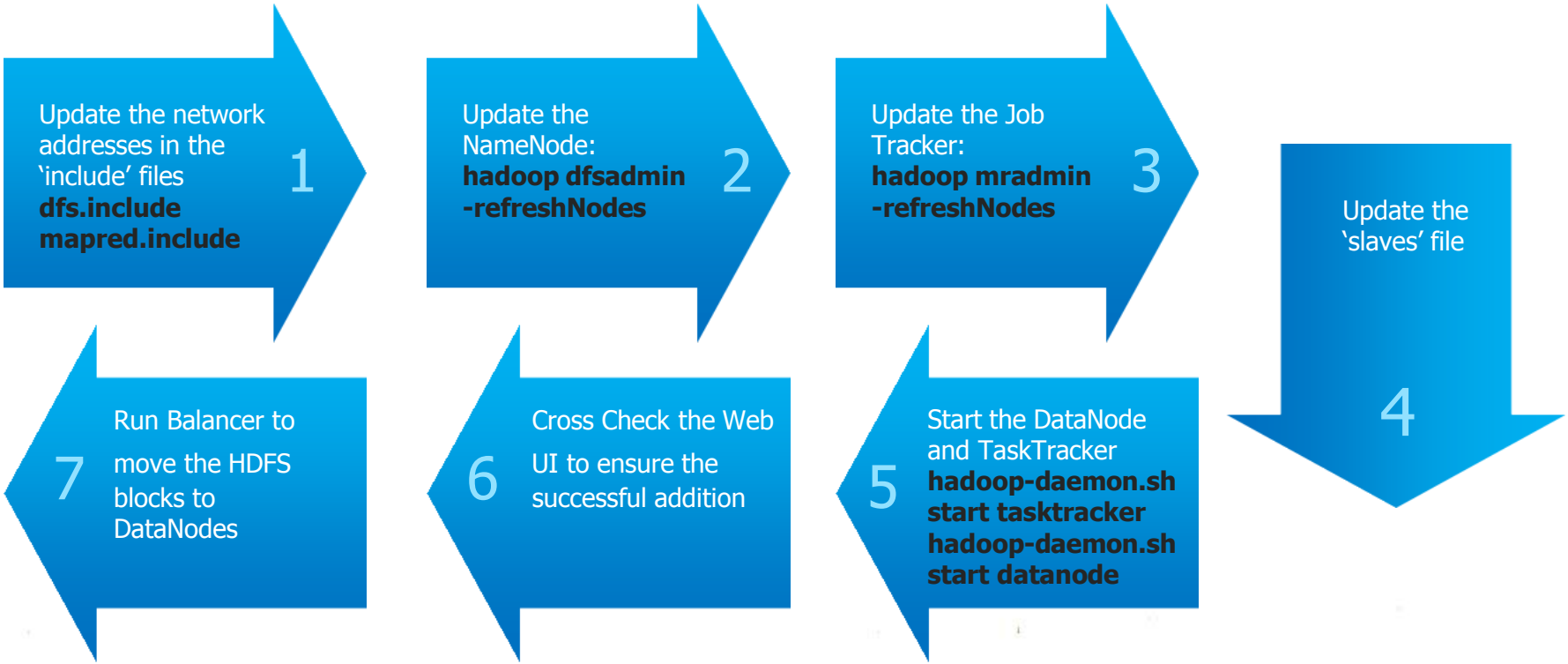
Commission Data Nodes

edureka!

→ **Add New Data Nodes**

Add (Commission) DataNodes

edureka!



Hadoop Upgrade from 1 to 2

edureka!

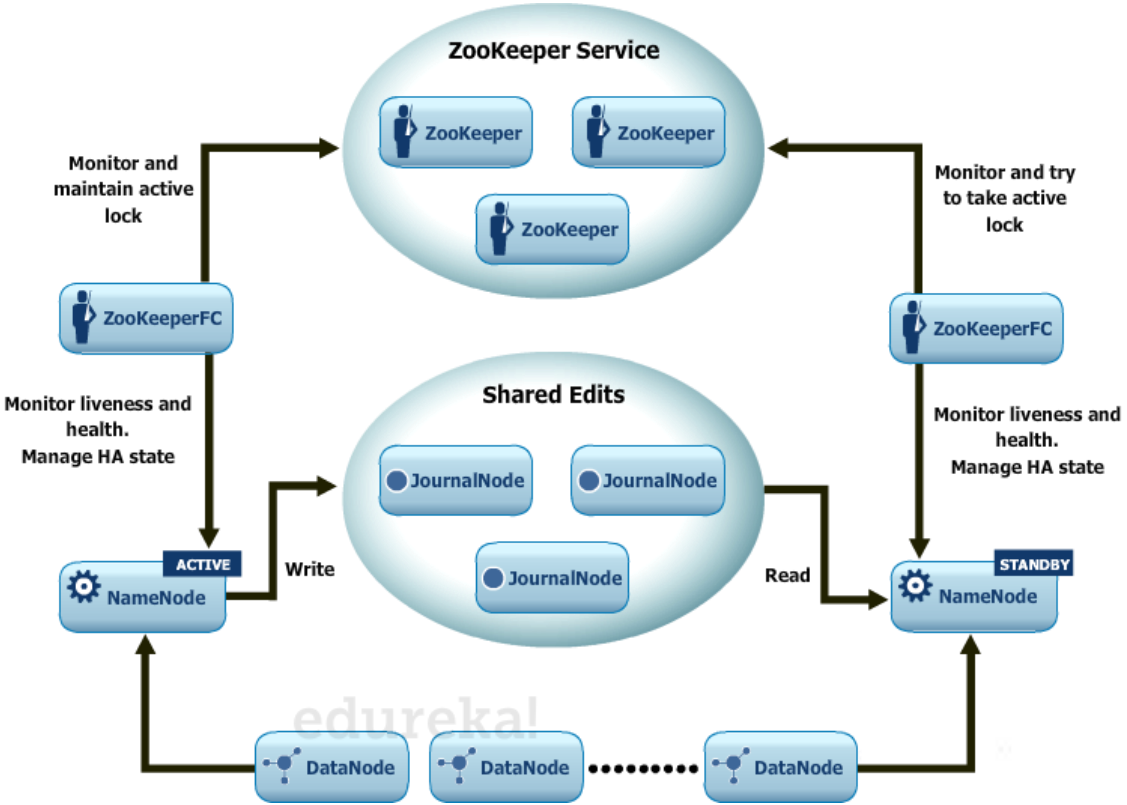
→ Run Reports
  - » FSCK
  - » LSR
  - » DFSADMIN

→ Take backup
  - » Configurations
  - » Applications
  - » Data and Meta-data

→ Install new version of Hadoop
→ Upgrade
→ Run New Reports
  - » FSCK
  - » LSR
  - » DFSADMIN

→ Compare old and new reports
→ Test new cluster
→ Finalize upgrade

NameNode HA


ZooKeeper Service
ZooKeeper
ZooKeeper
ZooKeeper
Monitor and maintain active lock
Monitor and try to take active lock
ZooKeeperFC
ZooKeeperFC
Monitor liveness and health. Manage HA state
Monitor liveness and health. Manage HA state
Shared Edits
JournalNode
JournalNode
JournalNode
ACTIVE
NameNode
Write
Read
STANDBY
NameNode
DataNode
DataNode
DataNode

DEMO

How it Works?

edureka!

LIVE Online Class

Class Recording in LMS

24/7 Post Class Support

Module Wise Quiz

Project Work

Verifiable Certificate

QUESTIONS

Twitter @edurekaIN, Facebook /edurekaIN, use #askEdureka for Questions

Course Topics

edureka!

→ Module 1
- » Hadoop Cluster Administration

→ Module 2
- » Hadoop Architecture and Cluster setup

→ Module 3
- » Hadoop Cluster: Planning and Managing

→ Module 4
- » Backup, Recovery and Maintenance

→ Module 5
- » Hadoop 2.0 and High Availability

→ Module 6
- » Advanced Topics: QJM, HDFS Federation and Security

→ Module 7
- » Oozie, Hcatalog/Hive and HBase Administration

→ Module 8
- » Project: Hadoop Implementation

edureka!