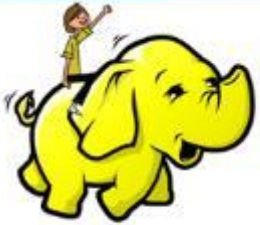
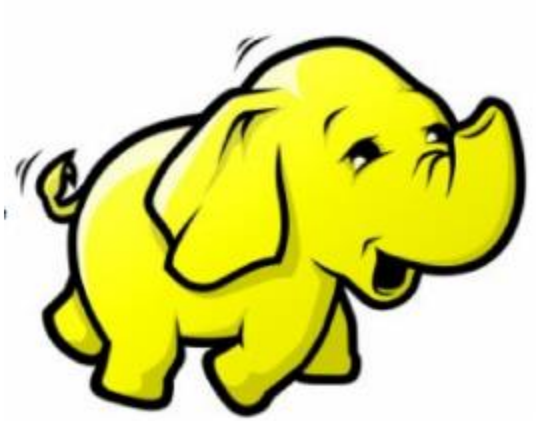
edureka!

Hadoop Administration



Hadoop Administration



Module 8:Project - Hadoop Implementation

Course Topics

edureka!

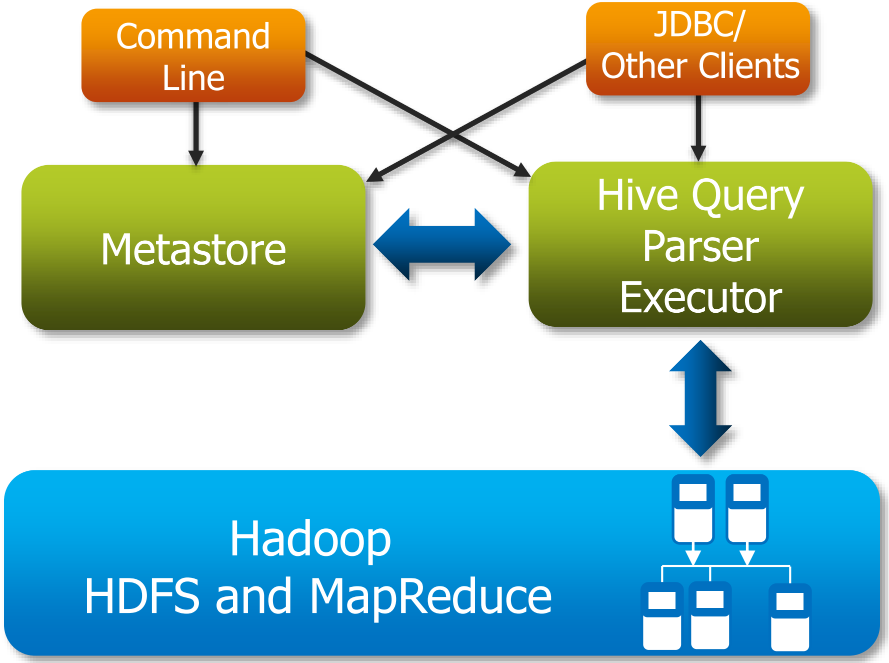
- ✓ **Module 1**
  - ✓ Understanding Big Data
  - ✓ Hadoop Components
- ✓ **Module 2**
  - ✓ Different Hadoop Server Roles
  - ✓ Hadoop Cluster Configuration
- ✓ **Module 3**
  - ✓ Hadoop Cluster Planning
  - ✓ Job Scheduling
- ✓ **Module 4**
  - ✓ Securing your Hadoop Cluster
  - ✓ Backup and Recovery
- ✓ **Module 5**
  - ✓ Hadoop 2.0 New Features
  - ✓ HDFS High Availability
- ✓ **Module 6**
  - ✓ Quorum Journal Manager (QJM)
  - ✓ Hadoop 2.0 - YARN
- ✓ **Module 7**
  - ✓ Oozie Workflow Scheduler
  - ✓ Hive and Hbase Administration
- ✓ **Module 8**
  - ✓ **Hadoop Cluster Case Study**
  - ✓ **Hadoop Implementation**

Topics of the Day

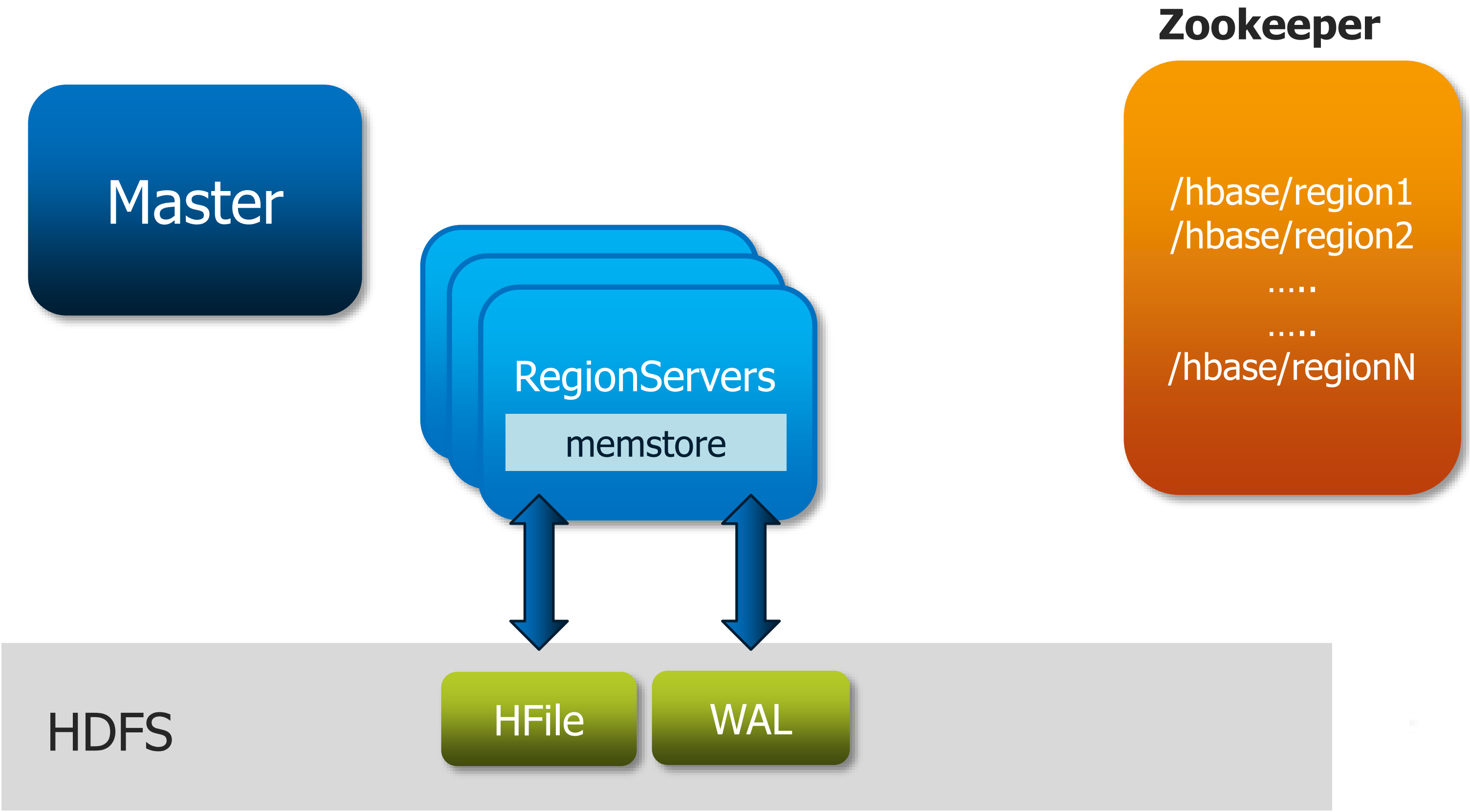
- Let’s Revise
- PIG setup and Configuration
- SQOOP – Hadoop and RDBMS
- Hadoop Cluster – A typical Use Case
- Hadoop Performance and Tuning
- HDFS High Availability (HDFS HA)
- Cloudera Distribution of Hadoop
- Cloudera Manager

http://www.pvsd.k12.ca.us/site/default.aspx?PageID=129

Let’s Revise - Hive



Let’s Revise - HBase



Hadoop Cluster: Configure PIG

Pig is an **open-source high-level dataflow system**.

It provides a simple language for queries and data manipulation **Pig Latin**, that is compiled into map-reduce jobs that are run on Hadoop.

Why is it Important?

- ✓ Companies like Yahoo, Google and Microsoft are collecting enormous data sets in the form of click streams, search logs, and web crawls.
- ✓ Some form of ad-hoc processing and analysis of all of this information is required.

Use Cases Where Pig Is Used…

edureka!

Processing of **Web Logs**

Data processing for search platforms

Support for **Ad Hoc queries** across large datasets.

Quick Prototyping of algorithms for processing large datasets.

PIG Configuration

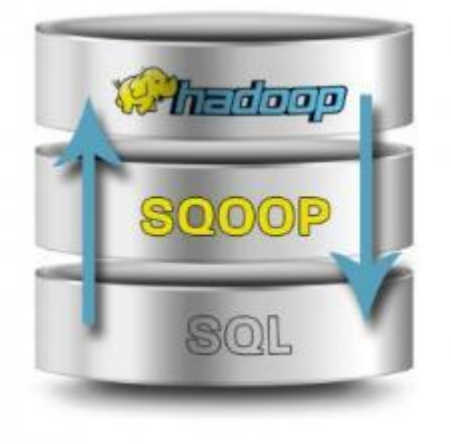
Install and Configure PIG

Hadoop Cluster: Data Loading

SQOOP

Apache Sqoop (TM) is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases.

- ✓ Imports individual tables or entire databases to HDFS.
- ✓ Generates Java classes to allow you to interact with your imported data.
- ✓ Provides the ability to import from SQL databases straight into your Hive data warehouse.

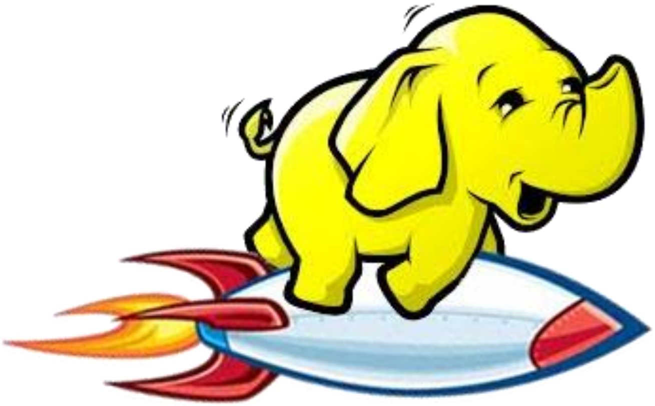


Hadoop Cluster: Real Word Implementation

Real Word Implementation

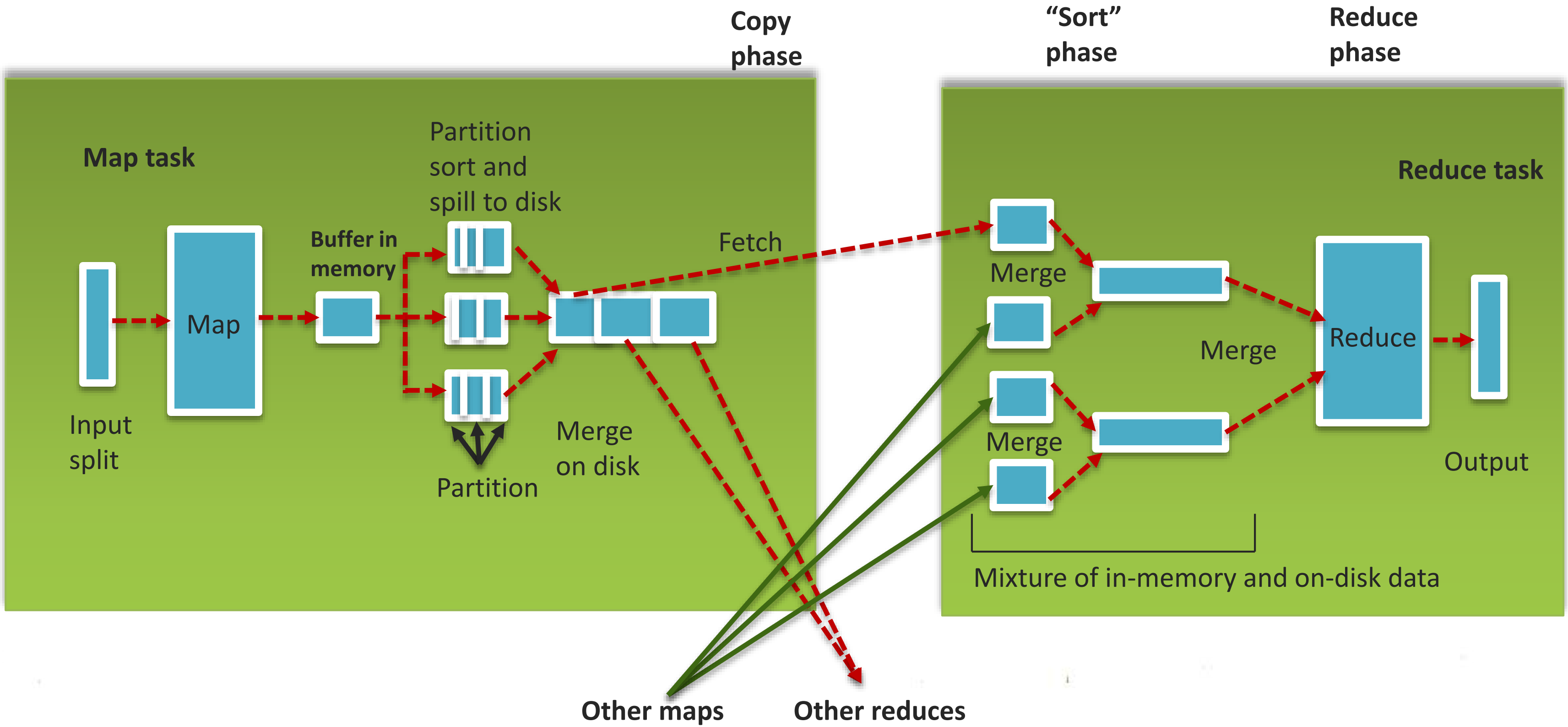
A typical Use Case

Hadoop Performance

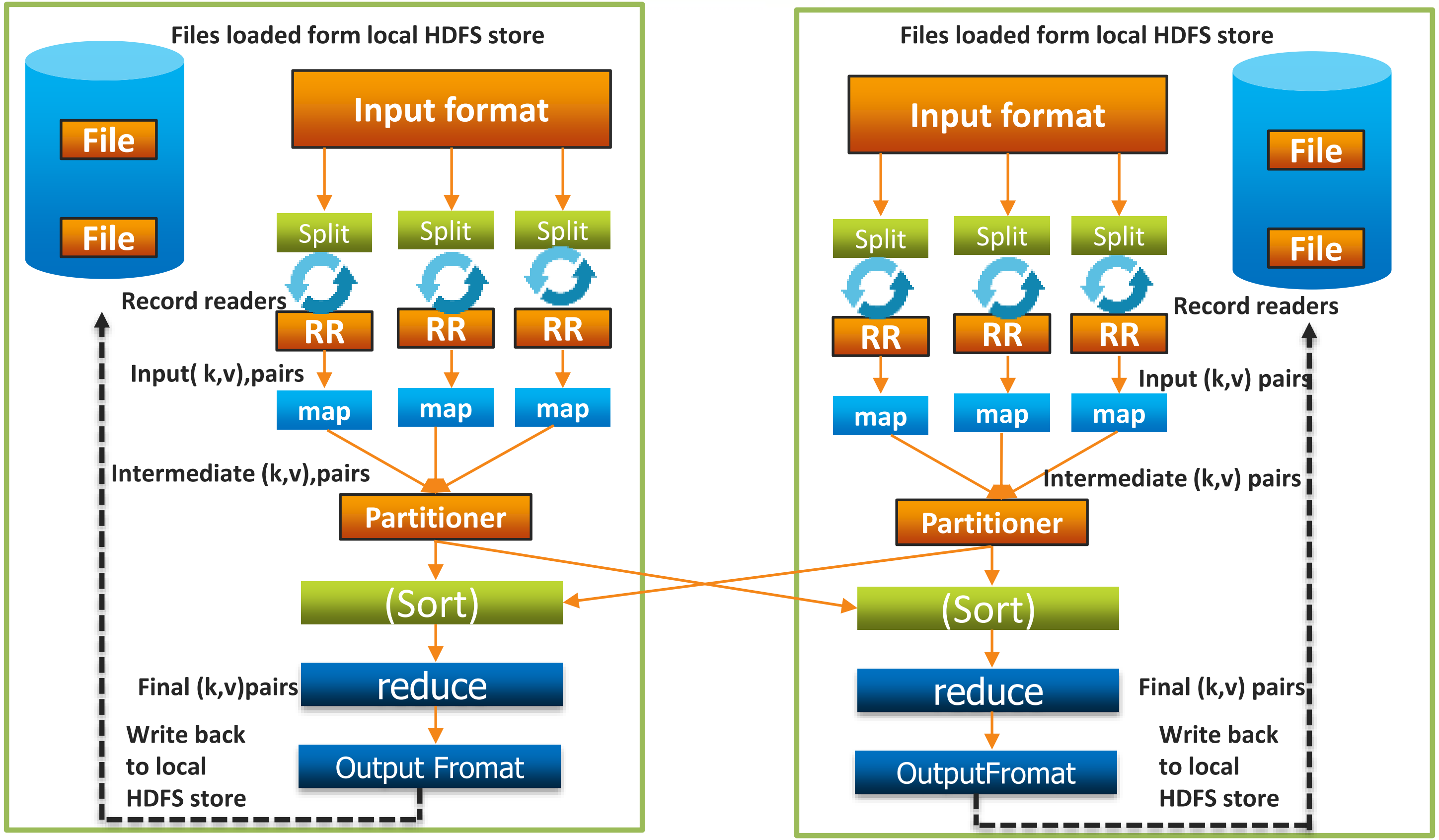


MapReduce execution

edureka!



MapReduce execution (Contd.)



* RR-Record Readers

MapReduce execution (Contd.)

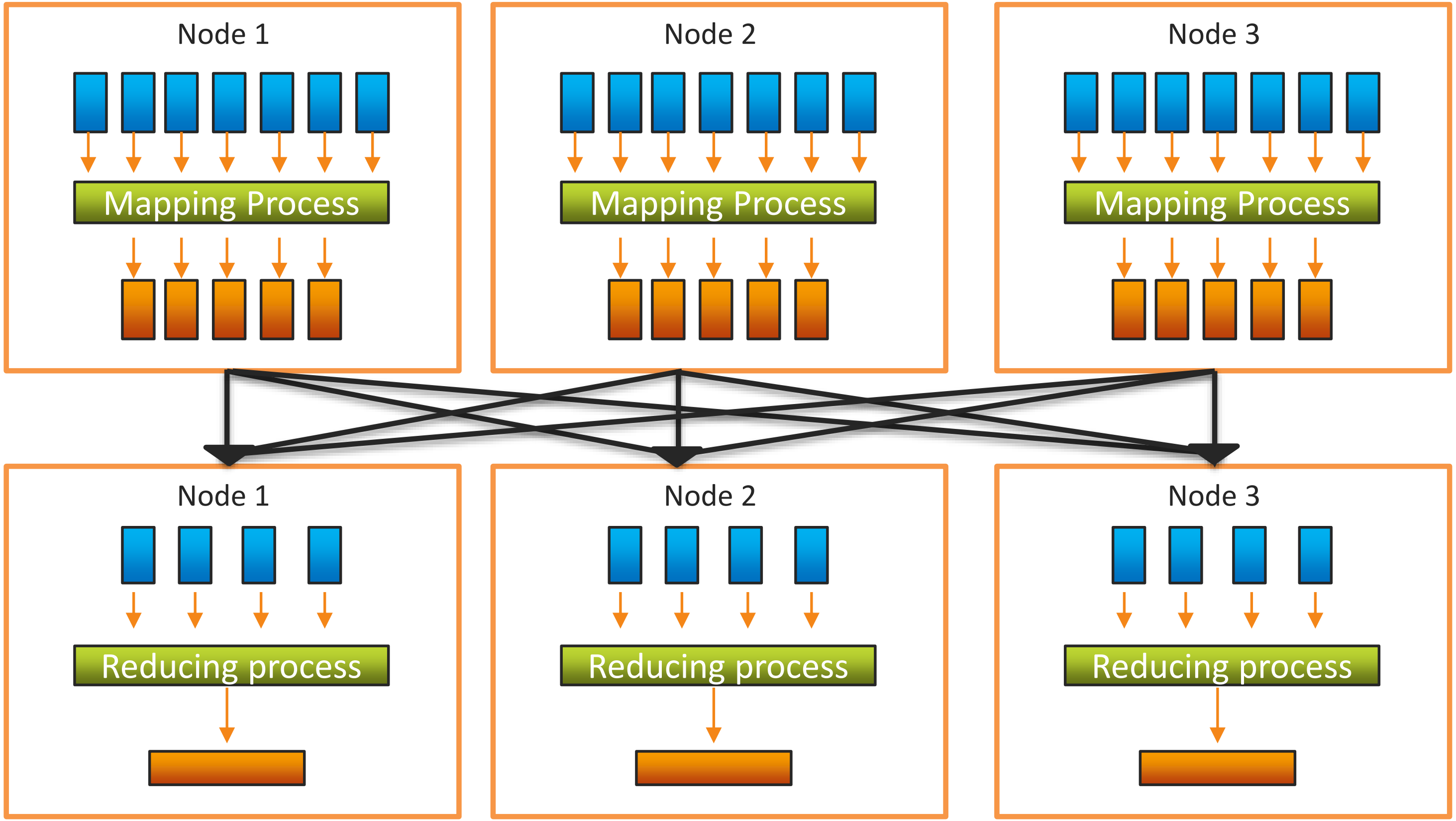
Pre loaded local input data

Intermediate data from mappers

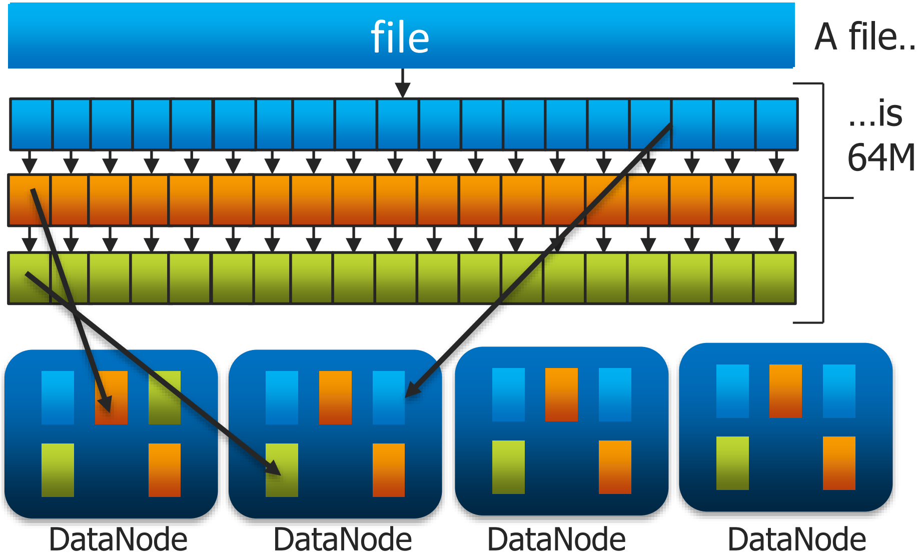
Values exchanges by shuffled process

Reducing process generate output

Outputs stored locally



Storage Considerations



Storage Considerations (Contd.)

HDFS Block Size

Property ‘dfs.block.size’ in hdfs-site.xml (default: 64 MB)

128 MB or even 256 MB in real cluster implementations to ease Memory Pressure on. NameNode and to provide more data to Mappers to work upon

I/O buffer size

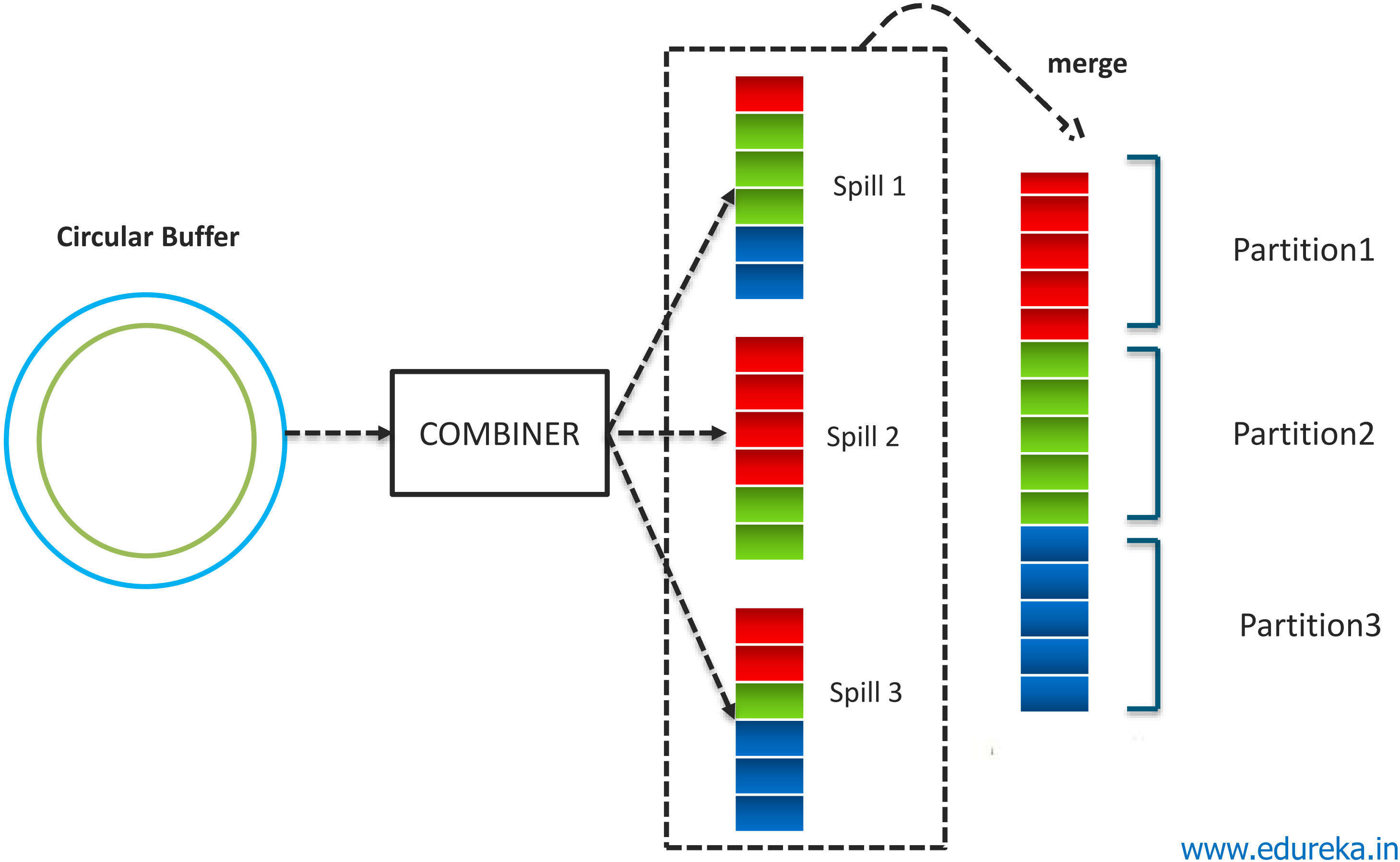
Property ‘io.file.buffer.size’ in core-site.xml (default: 4 KB)

Performance benefits with 128 KB

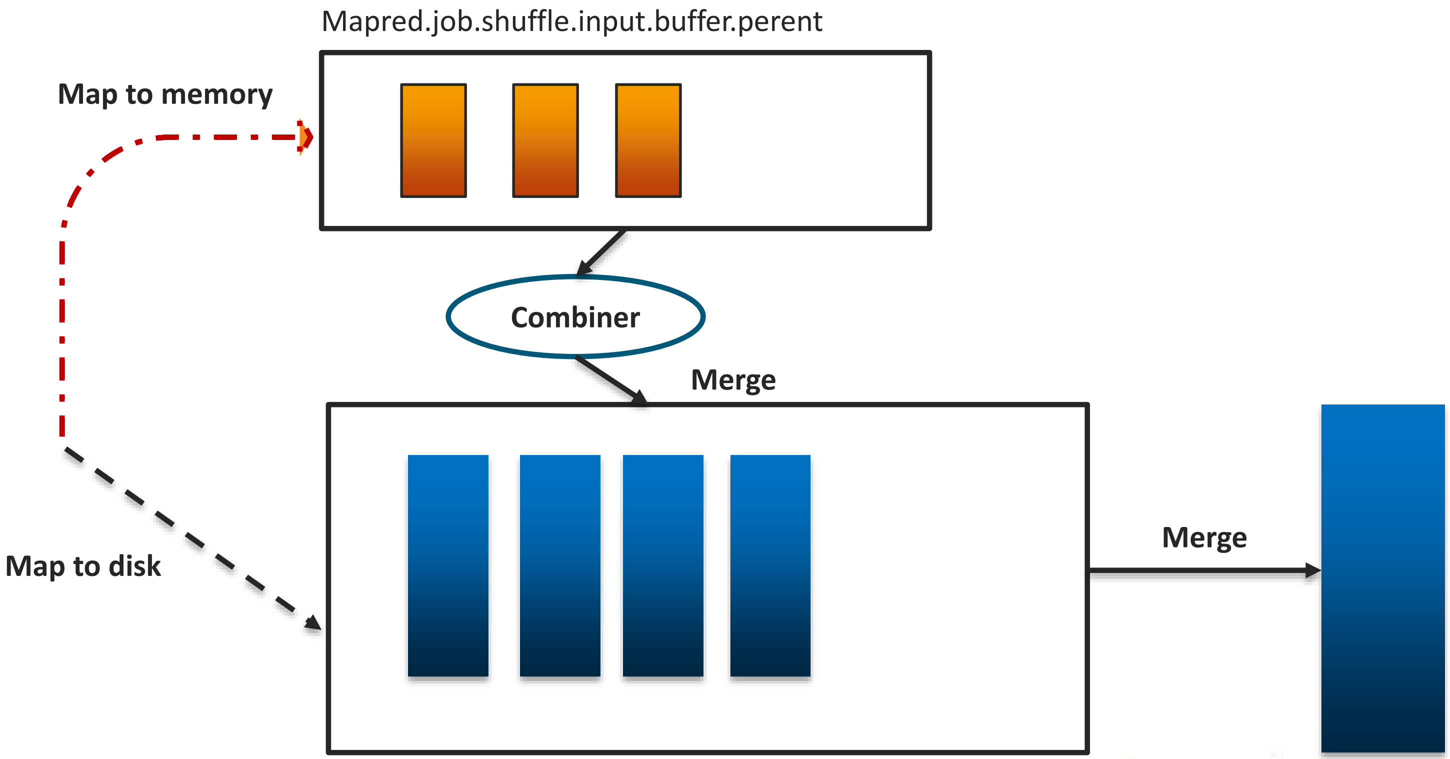
Reserved storage space

Property ‘dfs.datanode.du.reserved’ to reserve storage for non-HDFS usage as by default DataNode try to use all the available storage volumes.

Map Data


merge
Circular Buffer
COMBINER
Spill 1
Spill 2
Spill 3
Partition1
Partition2
Partition3

Reduce Data


Mapred.job.shuffle.input.buffer.perent
Map to memory
Combiner
Merge
Map to disk
Merge

Important Parameters - CPU

CPU-related parameters :

mapred.tasktracker.map
reduce.tasks.maximum

- ✓ These two parameters are the most relative ones to CPU utilization.
- ✓ The default value of both parameters is 2.
- ✓ Properly increasing their values according to your cluster condition increases the CPU utilization and therefore improves the performance.

For example, assume each node of the cluster has 4 CPUs supporting simultaneous multi-threading, and each CPU has 2 cores; then the total number of daemons should be no more than 4x2x2=16. Considering DN and TT would take 2 slots, there are at most 14 slots for map/reduce tasks, so the best value is 7 for both parameters.

Important Parameters - Disk

edureka!

Disk I/O-related parameters:

- ✓ **mapred.compress.map.output**
- ✓ **mapred.output.compress**
- ✓ **mapred.map.output.compression.codec**

io.sort.mb parameter:

This parameter sets the buffer size for map-side sorting, in units of MB, 100 by default. The greater the value, the fewer spills to the disk, thus reducing I/O times on the map side. Notice that increasing this value increases memory required by each map task.

Important Parameters – Disk (Contd.)

mapred.job.reduce.input.buffer.percent

This parameter sets the percentage of memory (relative to the maximum heap size) to retain map outputs during the reduce phase. When the shuffle is concluded, any remaining map outputs in memory must consume less than this threshold before the reduce phase can begin, 0 by default. The greater this value is, the less merge on the disk, thus reducing I/O times on the local disk during the reduce phase.

Hadoop 2.0: HDFS High Availability

DEMO

HDFS HA setup and Configuration

Annie's Question

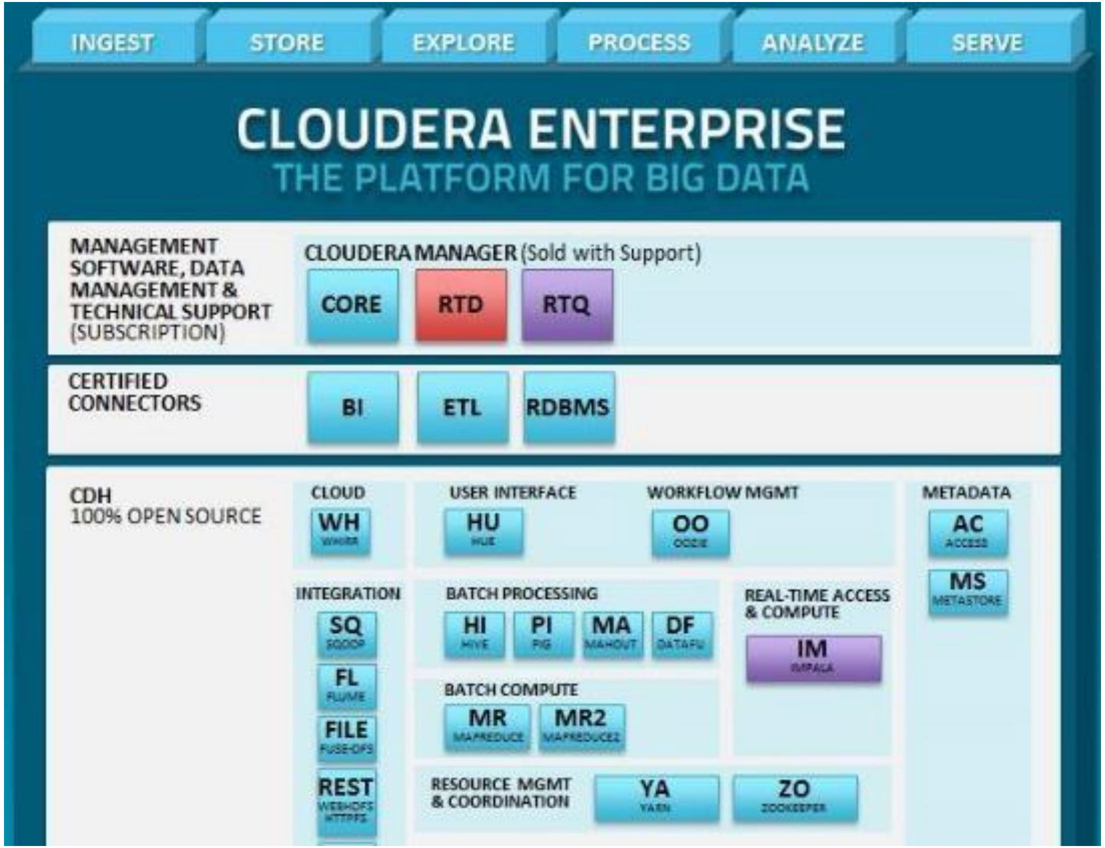
We can enable Name Node High Availability in hdfs-site.xml with the help of:

a) 'dfs.nameservices'
b) 'dfs.nameID'
c) 'dfs.namenode'

Annie’s Answer

Answer: ‘dfs.nameservices’

Cloudera Distribution of Hadoop



http://www.theregister.co.uk/2012/10/24/cloudera_hadoop_impala_real_time_query/

Cloudera Manager

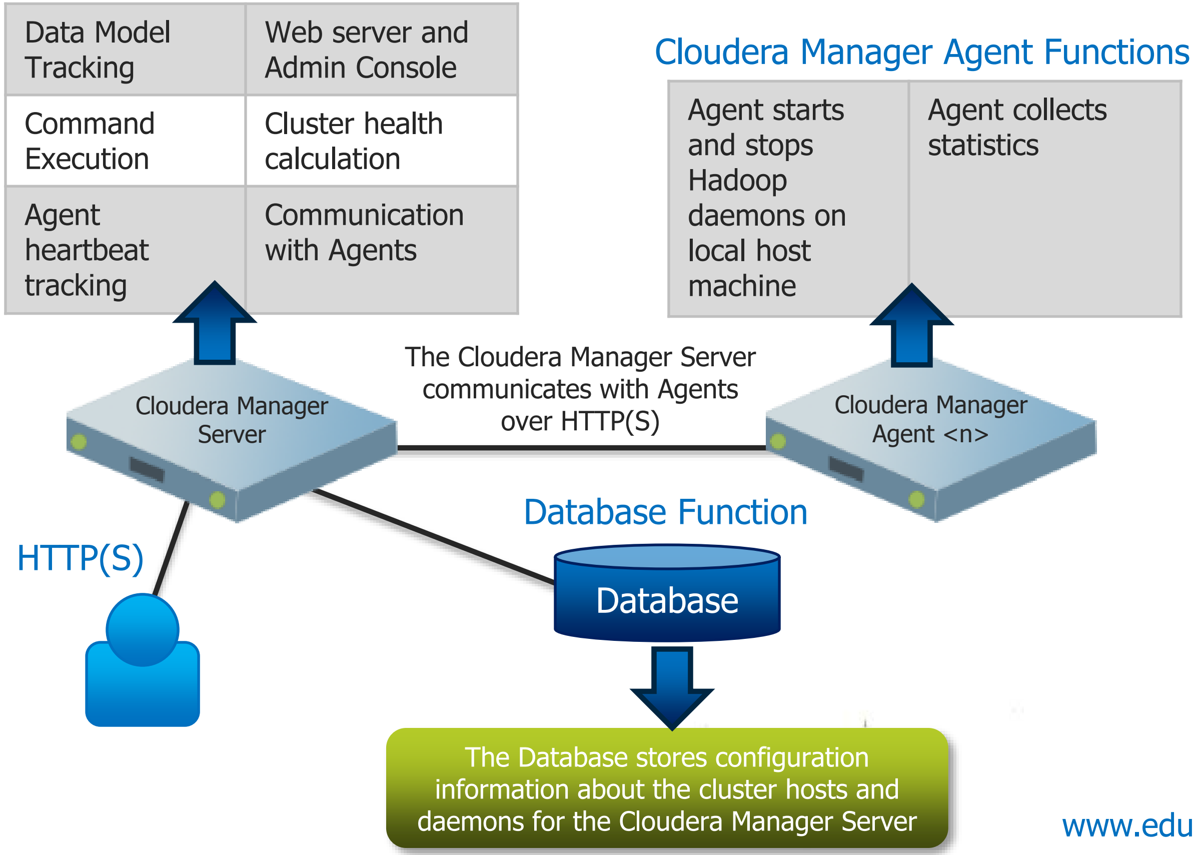
- ✓ Cloudera Inc. provides Apache Hadoop-based software, support and services to data driven enterprises.

- ✓ Cloudera's open-source Apache Hadoop distribution, CDH (Cloudera Distribution Including Apache Hadoop), targets enterprise-class deployments of that technology.

- ✓ Makes Hadoop Administration simple and straightforward, at any scale.

- ✓ Easy deployment and central operation of the complete Hadoop stack
  - ✓ Manage
  - ✓ Monitor
  - ✓ Diagnose
  - ✓ Integrate

Cloudera Manager Functions



Cloudera Manager Server Functions

Data Model Tracking	Web server and Admin Console
Command Execution	Cluster health calculation
Agent heartbeat tracking	Communication with Agents

Cloudera Manager Agent Functions

Agent starts and stops Hadoop daemons on local host machine	Agent collects statistics

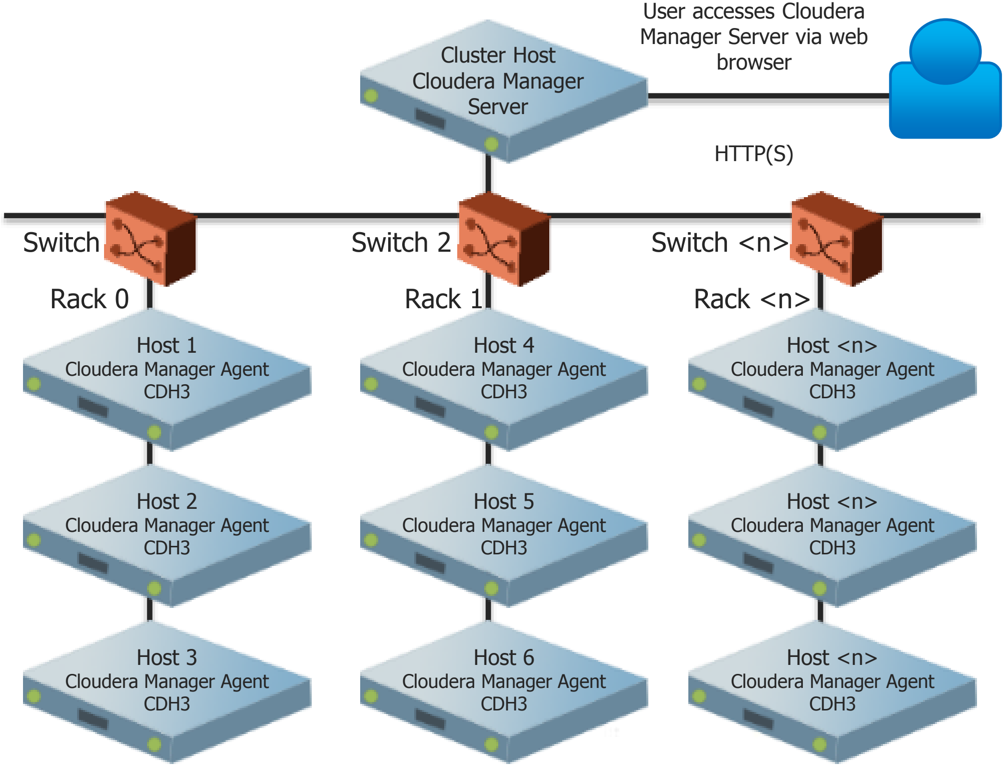
The Cloudera Manager Server communicates with Agents over HTTP(S)

Database Function

The Database stores configuration information about the cluster hosts and daemons for the Cloudera Manager Server

Cloudera Manager Functions

edureka!



User accesses Cloudera Manager Server via web browser
Cluster Host Cloudera Manager Server
HTTP(S)
Switch
Switch 2
Switch <n>
Rack 0
Rack 1
Rack <n>
Host 1 Cloudera Manager Agent CDH3
Host 2 Cloudera Manager Agent CDH3
Host 3 Cloudera Manager Agent CDH3
Host 4 Cloudera Manager Agent CDH3
Host 5 Cloudera Manager Agent CDH3
Host 6 Cloudera Manager Agent CDH3
Host <n> Cloudera Manager Agent CDH3
Host <n> Cloudera Manager Agent CDH3
Host <n> Cloudera Manager Agent CDH3

Further Reading

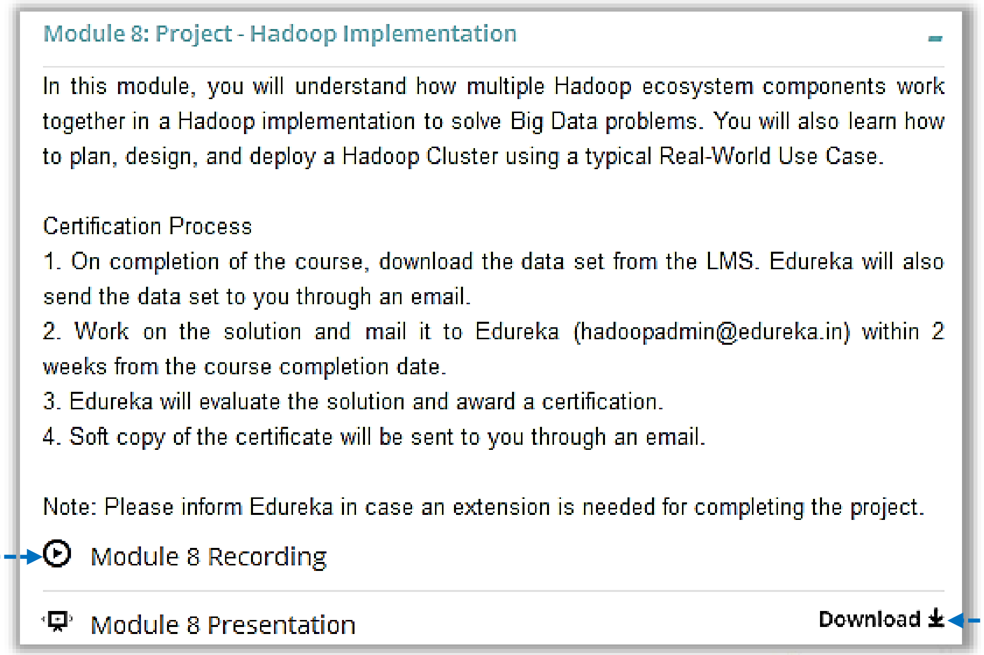
- ✓ http://www.cloudera.com/content/cloudera-content/cloudera-docs/CDH4/4.2.0/CDH4-Installation-Guide/CDH4-Installation-Guide.html
- ✓ https://docs.google.com/file/d/0Bx6N95pJhrROblJiaEJ0dHpwVmc/edit
- ✓ http://www.michael-noll.com/blog/2011/04/09/benchmarking-and-stress-testing-an-hadoop-cluster-with-terasort-testdfsio-nnbench-mrbench/

Assignments

Tasks for you

- **Attempt the following Assignments using the concepts discuss in the class:**
  - Complete the Course Project and Certification Project.

What’s Within the LMS?



Module 8: Project - Hadoop Implementation

In this module, you will understand how multiple Hadoop ecosystem components work together in a Hadoop implementation to solve Big Data problems. You will also learn how to plan, design, and deploy a Hadoop Cluster using a typical Real-World Use Case.

Certification Process

1. On completion of the course, download the data set from the LMS. Edureka will also send the data set to you through an email.
2. Work on the solution and mail it to Edureka (hadoopadmin@edureka.in) within 2 weeks from the course completion date.
3. Edureka will evaluate the solution and award a certification.
4. Soft copy of the certificate will be sent to you through an email.

Note: Please inform Edureka in case an extension is needed for completing the project.

Module 8 Recording

Module 8 Presentation

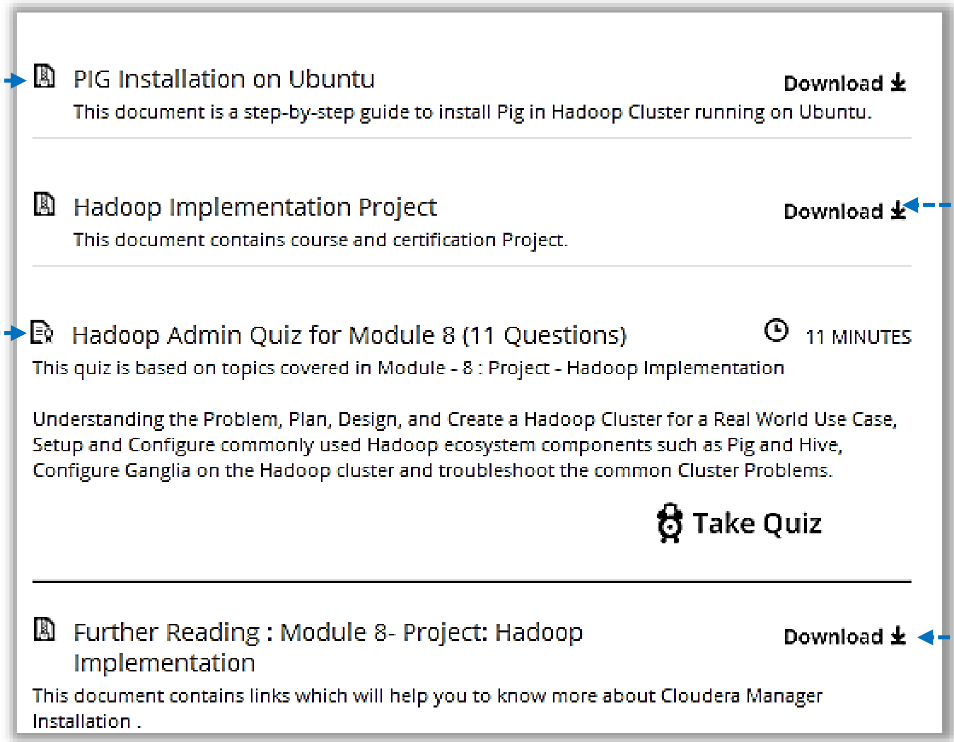
Recording of the Class

Presentation

What’s Within the LMS?

Installation Guide

Quiz



PIG Installation on Ubuntu — Download

This document is a step-by-step guide to install Pig in Hadoop Cluster running on Ubuntu.

Hadoop Implementation Project — Download

This document contains course and certification Project.

Hadoop Admin Quiz for Module 8 (11 Questions) — 11 MINUTES

This quiz is based on topics covered in Module - 8 : Project - Hadoop Implementation

Understanding the Problem, Plan, Design, and Create a Hadoop Cluster for a Real World Use Case, Setup and Configure commonly used Hadoop ecosystem components such as Pig and Hive, Configure Ganglia on the Hadoop cluster and troubleshoot the common Cluster Problems.

Take Quiz

Further Reading : Module 8- Project: Hadoop Implementation — Download

This document contains links which will help you to know more about Cloudera Manager Installation .

Implementation Project

Further Reading

edureka!

Thank You