DVS Hadoop Administration Content

Objective 1.1 – Big Data

What is Big Data?

Where Big Data is coming from?

What are Big Data use cases?

How Data is growing?

What are 3 V's of Big Data?

What are the challenges in Big Data Storage & Access?

Objective 1.2 – Hadoop

Why Hadoop?

What is Hadoop?

What is Hadoop History?

What are Hadoop distributions?

What are Hadoop components?

Hadoop Architecture

Objective 2.1 – HDFS

Understanding File System

Understanding Hadoop Distributed File System (HDFS)

HDFS Replication

HDFS Components

– NameNode

– DataNode

– Secondary NameNode

HDFS Features

HDFS Design Assumptions

Formatting NameNode

Communication between Nodes in a Cluster

How Metadata is maintained in Hadoop?

Types of Metadata

What is HDFS Block Report?

Check Pointing Mechanism

Metadata Memory Allocation

Anatomy of a File Write into & Read from HDFS

HDFS Block Replication Strategy

How to deal with Data Corruption?

HDFS Rebalancing & Space Reclamation

File Systems supported by Hadoop

Compression Formats supported by Hadoop

Lab:

- Understanding Hadoop Installation Prerequisites
- Building Hadoop Nodes
- Installation of Hadoop 1x (Pseudo Mode)
- Installation of Hadoop 1x (Distribution Mode)
- Commission and Decommission of nodes
- Understanding VERSION, FSImage, Editlog
- Hadoop Admin Commands (FSCK & Block Scanner Report)
- HDFS Replication (by XML file, by Host, by individual file)
- Increase & Decrease Replication
- Hadoop Rack Awareness

– Default Hadoop Settings

Objective 2.2 – Map Reduce

Map Reduce Introduction

How Map Reduce works?

Communication between JobTracker and TaskTracker

Anatomy of a Map Reduce Job Submission

Hadoop Schedulers

– FIFO Scheduler

– Fair Scheduler

– Capacity Scheduler

Lab:

– Setting up Mappers & Reducers

– Setting up Fair Scheduler

– Setting up Capacity Scheduler

– Setting up topology

– Setting up Logs and Logging mechanism

Objective 3.1 – Hadoop 2.X

Hadoop 2.X Architecture

What is Edge/Gateway/Connecting Node?

What is Zookeeper?

Difference between Hadoop 1.X and Hadoop 2.X

Understand the architecture of YARN

Understand the components of the YARN ResourceManager

Demonstrate the relationship between NodeManagers and ApplicationMaster

Demonstrate the relationship between ResourceManager and ApplicationMaster

Explain the relationship between Containers and ApplicationMasters

Job Flow in YARN

Namenode High Availability

– Using Shared Edits

– Using Zookeeper Quorum

Lab:

– Namenode High Availability using NFS Shared Edits & Zookeepers

– Namenode High Availability using Journal Nodes & Zookeepers

– Resource Manager High Availability

Objective 3.2 –Cluster Planning

Understanding Hardware Components

– Master Hardware

– Slave Hardware

– CPU, I/O, Network

Plan your cluster growth

Managing Users & Groups

Cluster sharing across multiple use cases

Objective 4.1 – Install and Configure Cloudera

Understand the minimum hardware and software requirements

Understand the Cloudera Architecture

Understand how to install CDH using Cloudera Manager

Understand complete deployment layout

Understand how to configure and manage different services

Understand different configuration parameters

Lab:

– Cloudera Cluster Installation

– Cloudera Manager Walkthrough

Objective 5.1 – Install and Configure Ambari & Hortonworks

Understand the minimum hardware and software requirements

Understand the Ambari Architecture

Understand how to install Ambari & Hortonworks

Understand complete deployment layout

Understand how to configure and manage different services

Understand different configuration parameters

Lab:

– Hortonworks Cluster Installation

– Hortonworks Ambari Walkthrough

Objective 6.1 – MapR Distribution

Understand the minimum hardware and software requirements

Understand the MCS Architecture

Understand various services configured in MapR

Objective 7.1 – Monitor and Administering Hadoop Clusters

Monitor using the CM or Ambari UI

Back up and recover Hadoop data

Use Hadoop snapshots

Objective 7.2 – Hadoop Security

Understand security concepts

Understanding & Configuring Hadoop ACLs

Understanding & Configuring Kerberos

Understanding & Configuring Knox & Ranger

Objective 8 – Hadoop Ecosystem Tools

Introduction to Sqoop

Introduction to Flume

Introduction to Pig

Introduction to Hive

Introduction to Hbase

Introduction to Oozie

Objective 9 – Other Admin Concepts

Hadoop Cluster Backup

Hadoop Cluster Upgrade

OS & Hadoop Patching

Objective 10 – Performance Tuning

Hadoop Performance Turning from OS Level

Hadoop Performance Turning from HDFS Level (Storage Layer)

Hadoop Performance Turning from MR/YARN Level (Processing Layer)

Objective 11 – Real Time Concepts

Day to Day Admin Activities

Frequently Occurring Issues

Roles and Responsibilities