Course Outline

Cloudera Administrator Training for Apache Hadoop (HADOOP-ADMIN)

Course Summary

Description

This course for Apache Hadoop provides participants with a comprehensive understanding of all the steps necessary to operate and maintain a Hadoop cluster using Cloudera Manager. From installation and configuration through load balancing and tuning, Cloudera's training course is the best preparation for the real-world challenges faced by Hadoop administrators.

Objectives

At the end of this course, students will be able to:

- Understand Cloudera Manager features that make managing your clusters easier, such as aggregated logging, configuration management, resource management, reports, alerts, and service management
- Configure and deploying production-scale clusters that provide key Hadoop-related services, including YARN, HDFS, Impala, Hive, Spark, Kudu, and Kafka
- Determine the correct hardware and infrastructure for your cluster
- Use proper cluster configuration and deployment to integrate with the data center
- Ingesting, storing, and accessing data in HDFS, Kudu, and cloud object stores such as Amazon S3
- How to load file-based and streaming data into the cluster using Kafka and Flume
- Know how to configure automatic resource management to ensure service-level agreements are met for multiple users of a cluster
- Understand best practices for preparing, tuning, and maintaining a production cluster
- Understand troubleshooting, diagnosing, and solving cluster issues

Topics

- The Cloudera Enterprise Data Hub
- Installing Cloudera Manager and CDH
- Configuring a Cloudera Cluster
- HDFS Data Ingest
- YARN and MapReduce
- Apache Spark
- Planning Your Cluster
- Advanced Cluster Configuration
- Managing Resources
- Cluster Maintenance
- Monitoring Clusters
- Cluster Troubleshooting
- Installing and Managing Hue
- Security
- Apache Kudu
- Apache Kafka
- Object Storage in the Cloud

Audience

This course is best suited to systems administrators and IT managers who have basic Linux experience.

Prerequisites

Students should have basic Linux experience. Prior knowledge of Apache Hadoop is not required.

Duration

Four days

Course Outline

Cloudera Administrator Training for Apache Hadoop (HADOOP-ADMIN)

Course Outline

I. The Cloudera Enterprise Data Hub
- A. Cloudera Enterprise Data Hub
- B. CDH Overview
- C. Cloudera Manager Overview
- D. Hadoop Administrator Responsibilities

II. Installing Cloudera Manager and CDH
- A. Cluster Installation Overview
- B. Cloudera Manager Installation
- C. CDH Installation
- D. CDH Cluster Services

III. Configuring a Cloudera Cluster
- A. Overview
- B. Configuration Settings
- C. Modifying Service Configurations
- D. Configuration Files
- E. Managing Role Instances
- F. Adding New Services
- G. Adding and Removing Hosts
- H. Hadoop Distributed File System
- I. Overview
- J. HDFS Topology and Roles
- K. Edit Logs and Checkpointing
- L. HDFS Performance and Fault Tolerance
- M. HDFS and Hadoop Security Overview
- N. Web User Interfaces for HDFS
- O. Using the HDFS Command Line Interface
- P. Other Command Line Utilities

IV. HDFS Data Ingest
- A. Data Ingest Overview
- B. File Formats
- C. Ingesting Data using File Transfer or REST Interfaces
- D. Importing Data from Relational Databases with Apache Sqoop
- E. Ingesting Data From External Sources with Apache Flume
- F. Best Practices for Importing Data
- G. Hive and Impala
- H. Apache Hive
- I. Apache Impala

V. YARN and MapReduce
- A. YARN Overview
- B. Running Applications on YARN
- C. Viewing YARN Applications
- D. YARN Application Logs
- E. MapReduce Applications
- F. YARN Memory and CPU Settings

VI. Apache Spark
- A. Spark Overview
- B. Spark Applications
- C. How Spark Applications Run on YARN
- D. Monitoring Spark Applications

VII. Planning Your Cluster
- A. General Planning Considerations
- B. Choosing the Right Hardware
- C. Network Considerations
- D. Virtualization Options
- E. Cloud Deployment Options
- F. Configuring Nodes

VIII. Advanced Cluster Configuration
- A. Configuring Service Ports
- B. Tuning HDFS and MapReduce
- C. Enabling HDFS High Availability

IX. Managing Resources
- A. Configuring cgroups with Static Service Pools
- B. The Fair Scheduler
- C. Configuring Dynamic Resource Pools
- D. Impala Query Scheduling

X. Cluster Maintenance
- A. Checking HDFS Status
- B. Copying Data Between Clusters
- C. Rebalancing Data in HDFS
- D. HDFS Directory Snapshots
- E. Upgrading a Cluster

XI. Monitoring Clusters
- A. Cloudera Manager Monitoring Features
- B. Health Tests
- C. Events and Alerts
- D. Charts and Reports
- E. Monitoring Recommendations

Course Outline

Cloudera Administrator Training for Apache Hadoop (HADOOP-ADMIN)

Course Outline

XII. Cluster Troubleshooting

A. Overview
B. Troubleshooting Tools
C. Misconfiguration Examples
D. Essential Points

XIII. Installing and Managing Hue

A. Overview
B. Managing and Configuring Hue
C. Hue Authentication and Authorization

XIV. Security

A. Hadoop Security Concepts
B. Hadoop Authentication Using Kerberos
C. Hadoop Authorization
D. Hadoop Encryption
E. Securing a Hadoop Cluster

XV. Apache Kudu

A. Kudu Overview
B. Architecture
C. Installation and Configuration
D. Monitoring and Management Tools

XVI. Apache Kafka

A. What Is Apache Kafka?
B. Apache Kafka Overview
C. Apache Kafka Cluster Architecture
D. Apache Kafka Command Line Tools
E. Using Kafka with Flume

XVII. Object Storage in the Cloud

A. Object Storage
B. Connecting Hadoop to Object Storage