



Blended Learning Outline for Cloudera Administrator Training for Apache Hadoop (171103c)

This training course is for system administrators and others responsible for managing Apache Hadoop clusters in production or development environments.

Prerequisites

This course is best suited to systems administrators and IT managers who have basic Linux experience. Prior knowledge of Apache Hadoop is not required.

Format

Participants enrolling in blended learning will be provided with:

- OnDemand access of Cloudera Administrator Training.
 - o Including 20-hours of cloud-based lab access
 - o Access will be available one week prior to the first live session through the Friday following the final live session
- Five three hour live-virtual sessions with a senior Cloudera Instructor.
 - o Live-virtual sessions will be focused on demonstrating labs and covering select topics from the weekly lessons. The live sessions allow time for students to ask questions, but assume participants have already completed the Ondemand lessons for that particular week.

Week 1:

Introduction

In this session students will ensure they have access to the courseware materials, can connect to the lab environment and be given details about the structure of the upcoming sessions.

Week 2:

The Case for Apache Hadoop

- Why Apache Hadoop?
- Fundamental Concepts
- Core Hadoop Components
- Essential Points

Hadoop Cluster Installation

- Rationale for a Cluster Management Solution
- Cloudera Manager Features
- Cloudera Manager Installation
- CDH Installation
- Essential Points

The Hadoop Distributed File System (HDFS)

- HDFS Features
- Writing and Reading Files
- NameNode Memory Considerations
- Overview of HDFS Security
- Web UIs for HDFS
- Using the Hadoop File Shell
- More Storage Technologies
- Essential Points

Week 3:

MapReduce and Spark on YARN

- The Role of Computational Frameworks
- YARN: The Cluster ResourceManager
- MapReduce Concepts
- Apache Spark Concepts
- Running Computational Frameworks on YARN
- Exploring YARN Applications Through the Web UIs and the Shell
- YARN Application Logs
- Essential Points

Hadoop Configuration and Daemon Logs

- Managing Configurations Using Cloudera Manager
- Locating Configurations and Applying Configuration Changes
- Managing Role Instances and Adding Services
- Configuring the HDFS Service
- Configuring Hadoop Daemon Logs
- Configuring the YARN Service
- Essential Points

Getting Data Into HDFS

- Ingesting Data From External Sources With Flume
- Ingesting Data From Relational Databases With Sqoop
- REST Interfaces
- Best Practices for Importing Data
- Essential Points

Week 4:

Planning Your Hadoop Cluster

- General Planning Considerations
- Choosing the Right Hardware
- Virtualization Options
- Cloud Deployment Options
- Network Considerations
- Configuring Nodes
- Essential Points

Installing and Configuring Hive, Impala, Pig, and Search

- Apache Hive
- Apache Impala (incubating)
- Apache Pig
- Cloudera Search
- Essential Points

Hadoop Clients Including Hue

- What Are Hadoop Clients?
- Installing and Configuring Hadoop Clients
- Installing and Configuring Hue
- Hue Authentication and Authorization
- Oozie Workflows
- Essential Points

Advanced Cluster Configuration

- Advanced Configuration Parameters
- Configuring Hadoop Ports
- Configuring HDFS for Rack Awareness
- Configuring HDFS High Availability
- Essential Points

Week 5:

Hadoop Security

- Why Hadoop Security Is Important
- Hadoop's Security System Concepts
- What Kerberos Is and How it Works
- Securing a Hadoop Cluster with Kerberos
- Other Security Topics

- Essential Points

Managing Resources

- Configuring cgroups with Static Service Pools
- The Fair Scheduler
- Configuring Dynamic Resource Pools
- YARN Memory and CPU Settings
- Impala Query Scheduling
- Essential Points

Cluster Maintenance

- Checking HDFS Status
- Copying Data Between Clusters
- Adding and Removing Cluster Nodes
- Rebalancing the Cluster
- Directory Snapshots
- Cluster Upgrading
- Essential Points

Cluster Monitoring and Troubleshooting

- Cloudera Manager Monitoring Features
- Monitoring Hadoop Clusters
- Troubleshooting Hadoop Clusters
- Common Misconfigurations
- Essential Points

Conclusion