Cloudera Administrator Training for Apache Hadoop

32 hours

Course objectives:

Cloudera University's four-day administrator training 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.

Through instructor-led discussion and interactive, hands-on exercises, participants will navigate the Hadoop ecosystem, learning topics such as:

- Learn about the topology of a typical Cloudera cluster and the role the major CDH components play in the cluster
- How to install Cloudera Manager and CDH
- How to use Cloudera Manager to create, configure, deploy, and monitor a CDH cluster
- What tools Cloudera provides to ingest data from outside sources into a cluster
- How to configure cluster components for optimal performance?
- What routine tasks are necessary to maintain a cluster, including updating to a new version of CDH
- Learn about detecting, troubleshooting, and repairing problems
- Key Cloudera security features

Who Should Attend:

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

Required Skills:

- Basic Linux experience
- Prior knowledge of Apache Hadoop is not required.

Course Contents:

The Cloudera Enterprise Data Hub

- Cloudera Enterprise Data Hub
- CDH Overview
- Cloudera Manager Overview
- Hadoop Administrator Responsibilities

Installing Cloudera Manager and CDH

- Cluster Installation Overview
- Cloudera Manager Installation
- CDH Installation
- CDH Cluster Services

Configuring a Cloudera Cluster

- Overview
- Configuration Settings
- Modifying Service Configurations
- Configuration Files
- Managing Role Instances

- Adding New Services
- Adding and Removing Hosts

Hadoop Distributed File System

- Overview
- HDFS Topology and Roles
- Edit Logs and Checkpointing
- HDFS Performance and Fault Tolerance
- HDFS and Hadoop Security Overview
- Web User Interfaces for HDFS
- Using the HDFS Command Line Interface
- Other Command Line Utilities

HDFS Data Ingest

- Data Ingest Overview
- File Formats
- Ingesting Data using File Transfer or REST Interfaces
- Importing Data from Relational Databases with Apache Sqoop
- Ingesting Data From External Sources with Apache Flume
- Best Practices for Importing Data

Hive and Impala

- Apache Hive
- Apache Impala

YARN and MapReduce

- YARN Overview
- Running Applications on YARN
- Viewing YARN Applications
- YARN Application Logs
- MapReduce Applications
- YARN Memory and CPU Settings

Apache Spark

- Spark Overview
- Spark Applications
- How Spark Applications Run on YARN
- Monitoring Spark Applications

Planning Your Cluster

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

Advanced Cluster Configuration

- Configuring Service Ports
- Tuning HDFS and MapReduce
- Enabling HDFS High Availability

Managing Resources

- Configuring cgroups with Static Service Pools
- The Fair Scheduler
- Configuring Dynamic Resource Pools
- Impala Query Scheduling

Cluster Maintenance

- Checking HDFS Status
- Copying Data Between Clusters
- Rebalancing Data in HDFS
- HDFS Directory Snapshots
- Upgrading a Cluster

Monitoring Clusters

- Cloudera Manager Monitoring Features
- Health Tests
- Events and Alerts
- Charts and Reports
- Monitoring Recommendations

Cluster Troubleshooting

- Overview
- Troubleshooting Tools
- Misconfiguration Examples

Installing and Managing Hue

- Overview
- Managing and Configuring Hue
- Hue Authentication and Authorization

Security

- Hadoop Security Concepts
- Hadoop Authentication Using Kerberos
- Hadoop Authorization
- Hadoop Encryption
- Securing a Hadoop Cluster

Apache Kudu

- Kudu Overview
- Architecture
- Installation and Configuration
- Monitoring and Management Tools

Apache Kafka

- What Is Apache Kafka?
- Apache Kafka Overview
- Apache Kafka Cluster Architecture
- Apache Kafka Command Line Tools
- Using Kafka with Flume

Object Storage in the Cloud

- Object Storage
- Connecting Hadoop to Object Storage