Course duration

- 4 days

Course Benefits

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

Course Outline

1. The Cloudera Enterprise Data Hub
    1. Cloudera Enterprise Data Hub
    2. CDH Overview
    3. Cloudera Manager Overview
    4. Hadoop Administrator Responsibilities
2. Installing Cloudera Manager and CDH
    1. Cluster Installation Overview
    2. Cloudera Manager Installation
    3. CDH Installation
    4. CDH Cluster Services
3. Configuring a Cloudera Cluster
    1. Overview
    2. Configuration Settings
    3. Modifying Service Configurations
    4. Configuration Files
    5. Managing Role Instances
    6. Adding New Services
    7. Adding and Removing Hosts

4. Hadoop Distributed File System
    1. Overview
    2. HDFS Topology and Roles
    3. Edit Logs and Checkpointing
    4. HDFS Performance and Fault Tolerance
    5. HDFS and Hadoop Security Overview
    6. Web User Interfaces for HDFS
    7. Using the HDFS Command Line Interface
    8. Other Command Line Utilities
5. HDFS Data Ingest
    1. Data Ingest Overview
    2. File Formats
    3. Ingesting Data using File Transfer or REST Interfaces
    4. Importing Data from Relational Databases with Apache Sqoop
    5. Ingesting Data From External Sources with Apache Flume
    6. Best Practices for Importing Data
6. Hive and Impala
    1. Apache Hive
    2. Apache Impala
7. YARN and MapReduce
    1. YARN Overview
    2. Running Applications on YARN
    3. Viewing YARN Applications
    4. YARN Application Logs
    5. MapReduce Applications
    6. YARN Memory and CPU Settings
8. Apache Spark
    1. Spark Overview
    2. Spark Applications
    3. How Spark Applications Run on YARN
    4. Monitoring Spark Applications
9. Planning Your Cluster
    1. General Planning Considerations
    2. Choosing the Right Hardware
    3. Network Considerations
    4. Virtualization Options
    5. Cloud Deployment Options
    6. Configuring Nodes
10. Advanced Cluster Configuration
    1. Configuring Service Ports
    2. Tuning HDFS and MapReduce
    3. Enabling HDFS High Availability
11. Managing Resources
    1. Configuring cgroups with Static Service Pools
    2. The Fair Scheduler
    3. Configuring Dynamic Resource Pools
    4. Impala Query Scheduling

12. Cluster Maintenance
    1. Checking HDFS Status
    2. Copying Data Between Clusters
    3. Rebalancing Data in HDFS
    4. HDFS Directory Snapshots
    5. Upgrading a Cluster
13. Monitoring Clusters
    1. Cloudera Manager Monitoring Features
    2. Health Tests
    3. Events and Alerts
    4. Charts and Reports
    5. Monitoring Recommendations
14. Cluster Troubleshooting
    1. Overview
    2. Troubleshooting Tools
    3. Misconfiguration Examples
    4. Essential Points
15. Installing and Managing Hue
    1. Overview
    2. Managing and Configuring Hue
    3. Hue Authentication and Authorization
16. Security
    1. Hadoop Security Concepts
    2. Hadoop Authentication Using Kerberos
    3. Hadoop Authorization
    4. Hadoop Encryption
    5. Securing a Hadoop Cluster
17. Apache Kudu
    1. Kudu Overview
    2. Architecture
    3. Installation and Configuration
    4. Monitoring and Management Tools
18. Apache Kafka
    1. What Is Apache Kafka?
    2. Apache Kafka Overview
    3. Apache Kafka Cluster Architecture
    4. Apache Kafka Command Line Tools
    5. Using Kafka with Flume
19. Object Storage in the Cloud
    1. Object Storage
    2. Connecting Hadoop to Object Storag

Class Materials

Each student will receive a comprehensive set of materials, including course notes and all the class examples.

Class Prerequisites

Experience in the following *is required* for this Hadoop class:

- Basic Linux experience.