# APACHE HADOOP Administrator

**Prerequisites**

Prior knowledge of Apache Hadoop is not required. Unix/Linux administration knowledge will be helpful.

**Associated Certification(s):**

Upon completion of the course, attendees can go for CCAH or HDP Administrator. Certification is a great differentiator; it helps establish you as a leader in the field, providing employers and customers with tangible evidence of your skills and expertise.

**Course Objectives**

This 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. From installation and configuration through load balancing and tuning. This training course is the best preparation for the real-world challenges faced by Hadoop administrators.

**Course Content**

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

* The internals of YARN, MapReduce, and HDFS
* Determining the correct hardware and infrastructure for your cluster
* Proper cluster configuration and deployment to integrate with the data center
* How to load data into the cluster from dynamically-generated files using Flume and from RDBMS using Sqoop
* Configuring the FairScheduler to provide service-level agreements for multiple users of a cluster
* Best practices for preparing and maintaining Apache Hadoop in production
* Troubleshooting, diagnosing, tuning, and solving Hadoop issues

**Course Outline**

Introduction  
The Case for Apache Hadoop

* Why Hadoop?
* Core Hadoop Components
* Fundamental Concepts

HDFS

* HDFS Features
* Writing and Reading Files
* NameNode Memory Considerations
* Overview of HDFS Security> Using the Namenode Web UI
* Using the Hadoop File Shell

Getting Data into HDFS

* Ingesting Data from External Sources with
* Flume
* Ingesting Data from Relational Databases with Sqoop
* Best Practices for Importing Data

YARN and MapReduce

* What Is MapReduce?
* Basic MapReduce Concepts
* YARN Cluster Architecture
* Resource Allocation
* Failure Recovery
* Using the YARN Web UI
* MapReduce Version 1

Planning Your Hadoop Cluster

* General Planning Considerations
* Choosing the Right Hardware
* Network Considerations
* Configuring Nodes
* Planning for Cluster Management

Hadoop Installation and Initial Configuration

* Deployment Types
* Installing Hadoop
* Specifying the Hadoop Configuration
* Performing Initial HDFS Configuration
* Performing Initial YARN and MapReduce Configuration
* Hadoop Logging

Installing and Configuring Hive, Impala, and Pig

* Hive
* Impala
* Pig

Hadoop Clients

* What is a Hadoop Client?
* Installing and Configuring Hadoop Clients
* Installing and Configuring Hue
* Hue Authentication and Authorization

Cloudera Manager / APACHE Ambari

* The Motivation for Cloudera Manager /Apache Ambari
* Cloudera Manager/ Apache Ambari Features
* Express and Enterprise Versions
* Cloudera Manager / Apache Ambari Topology
* Installing Cloudera Manager / Apache Ambari
* Installing Hadoop Using Cloudera Manager / Apache Ambari
* Performing Basic Administration Tasks Using Cloudera Manager / Apache Ambari

Advanced Cluster Configuration

* Configuring Hadoop Ports
* Explicitly Including and Excluding Hosts
* Configuring HDFS for Rack Awareness
* Configuring HDFS High Availability

Hadoop Security

* Why Hadoop Security Is Important
* Hadoop’s Security System Concepts
* What Kerberos Is and How it Works

Cluster Maintenance

* Checking HDFS Status
* Copying Data Between Clusters
* Adding and Removing Cluster Nodes
* Rebalancing the Cluster
* Cluster Upgrading

Cluster Monitoring and Troubleshooting

* General System Monitoring
* Monitoring Hadoop Clusters
* Common Troubleshooting Hadoop Clusters
* Common Misconfigurations

Conclusion