HADOOP OPERATIONS
Course Code: 7316

This course will cover cluster planning, installation, administration, resource management, and monitoring.

Apache Hadoop is an open source software project that enables distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with very high degree of fault tolerance. In this learning path you will learn about cluster planning, installation and administration, resource management and monitoring and logging.

What You’ll Learn
• Designing Hadoop Clusters
• Hadoop in the Cloud
• Deploying Hadoop Clusters
• Hadoop Cluster Availability
• Securing Hadoop Clusters
• Operating Hadoop Clusters
• Stabilizing Hadoop Clusters
• Capacity Management for Hadoop Clusters
• Performance Tuning of Hadoop Clusters
• Cloudera Manager and Hadoop Clusters

Who Needs to Attend
Developers interested in expanding their knowledge of Hadoop from the operations perspective.

Prerequisites
Apache Hadoop is an open source software for affordable supercomputing; it provides the distributed file system and the parallel processing required to run a massive computing cluster. This learning path provides an explanation and demonstration of the most popular components in the Hadoop ecosystem. It defines and describes theory and architecture, while also providing instruction on installation, configuration, usage, and low-level use cases for the Hadoop ecosystem. This learning path can be used to help prepare for the Cloudera Certified Developer for Hadoop, HDP Certified Developer, Cloudera Certified Administrator for Hadoop, or Hadoop 2.0 Administrator Certification exam.
**HADOOP OPERATIONS**

Course Code: 7316

<table>
<thead>
<tr>
<th>On-Demand</th>
<th>$1,695 USD</th>
</tr>
</thead>
</table>

On-Demand Outline

Designing Hadoop Clusters

Big Data Engineering
  • Defining Supercomputing
  • Examining Engineering Teams
  • Exploring Big Data Solutions

Principles of Hadoop Clusters
  • Examining Axioms of Supercomputing
  • Exploring Design Principles for Hadoop
  • Examining Additional Design Principles

Architecture of a Hadoop Cluster
  • Examining Hadoop Cluster Architecture
  • Scaling Hadoop Architectures

Network for the Hadoop Cluster
  • Examining Network Clusters

Hardware for the Hadoop Cluster
  • Examining Hardware Responsibilities
  • Exploring Master Server Best Practices
  • Examining Data Server Recommendations

Operating Systems for the Hadoop Cluster
  • Exploring Operating Systems Best Practice
  • Examining Hostnames and DNS Recommendations

Storage for the Hadoop Cluster
  • Examining Storage Options
  • Calculating Storage Amounts
  • Evaluating Storage Options
Deployment of an Admin Server
  • Planning a Deployment
  • Setting Up Flash Drives
  • Setting Up Kickstart Files
  • Setting Up Network Installer

Hadoop in the Cloud

Amazon Web Services
  • Examining Cloud Computing
  • Examining Amazon Web Services
  • Examining AWS EC2

Setup of AWS
  • Examining AWS Credentials
  • Creating an AWS Account
  • Examining AWS Access Keys
  • Examining Identification and Access Management
  • Setting up Identification and Access Management

AWS System Security
  • Exploring SSH Keys

AWS S3 and EC2
  • Setting Up S3
  • Provisioning a Micro EC2

Setup of AWS Cluster
  • Configuring Hadoop for AWS
  • Creating an EC2 Baseline Server
  • Creating an Amazon Machine Image
  • Creating an Amazon Cluster
  • Exploring the AWS Command Line Interface

Moving Data
  • Using the AWS Command Line Interface
  • Moving Data into AWS

Elastic MapReduce
  • Examining Hadoop Cloud Implementations
  • Examining AWS Elastic MapReduce
  • Examining EMR and End-users
  • Setting Up EMR Clusters
  • Running EMR Jobs
  • Running EMR Jobs with Hue
  • Running EMR Jobs with the Command Line Interface

Deploying Hadoop Clusters

Configuration Management Tools
• Examining Configuration Management Tools
• Simulating Configuration Management Tools

Create Configuration Items
• Building Images for Baseline Servers
• Building Images for Data Servers
• Building Images for Master Servers

Setup a CM Environment
• Provisioning Admin Servers

Deploy a Hadoop Cluster
• Exploring Cluster Architecture
• Provisioning Hadoop Clusters
• Deploying Support Tools
• Starting and Stopping Hadoop Clusters
• Configuring Hadoop Clusters
• Configuring Logging
• Building Client Servers
• Configuring MySQL Databases
• Building Hadoop Clients
• Configuring MySQL Databases
• Building Hadoop Clients
• Configuring Hive Daemons
• Validating Flume, Sqoop, HDFS, and MapReduce
• Validating Hive and Pig
• Configuring HCatalog Daemons
• Configuring Oozie
• Configuring Hue

Hadoop Cluster Availability

Availability of Hadoop
• Defining Hadoop Fault Tolerance
• Examining NameNode Reliability
• Exploring Checkpoint Node
• Testing NameNode Failure
• Examining NameNode Recovery
• Swapping NameNodes
• Examining DataNode Reliability
• Testing DataNode Reliability
• Examining DataNode Recovery
• Exploring DataNode Replications

High Availability for HDFS
• Recovering Missing Data Blocks
• Defining HDFS High Availability
• Configuring for High Availability
• Setting up NameNode High Availability
• Examining High Availability Auto Failovers
• Creating High Availability Auto Failovers

YARN Containers
• Examining YARN Task Reliability
• Examining YARN Containers
• Testing YARN Container Reliability

YARN Jobs
Examining YARN Job Reliability
Testing Application Reliability

High Availability for YARN
Examining YARN High Availability
Setting Up High Availability for ResourceManagers

Securing Hadoop Clusters

Hadoop Security
• Examining Security Risks

Network Security
• Locking Down Networks
• Implementing Security Groups

Kerberos
• Examining Kerberos
• Creating Kerberos Diagrams
• Preparing for Kerberos Installation
• Installing Kerberos
• Configuring Kerberos

Services Security
• Examining Hadoop and Kerberos
• Configuring HDFS for Kerberos
• Configuring YARN for Kerberos
• Examining Hive with Kerberos
• Configuring Hive for Kerberos
• Examining Pig, Sqoop, Oozie with Kerberos
• Configuring Pig and HTTPFS for Kerberos
• Configuring Oozie for Kerberos
• Configuring Hue for Kerberos
• Examining Flume and Kerberos

User Security
• Managing User Security
• Managing User Access
• Creating Access Control Lists

Data Security
• Examining Data in Motion
• Encrypting Data in Motion
• Encrypting Data at Rest
• Examining Hadoop Security
• Monitoring Hadoop Security

Operating Hadoop Clusters

Hadoop Operations
• Managing Hadoop Service Levels
• Deploying Hadoop Releases
• Examining Hadoop Change Management

Racks Awareness for Hadoop
• Examining Rack Awareness
• Installing Rack Awareness

File System Management for HDFS
• Starting and Stopping a Hadoop Cluster
• Writing Init Scripts
• Administering HDFS
• Managing HDFS
• Setting Quotas
• Installing Trash

DataNode Management for HDFS
• Managing HDFS DataNodes
• Replacing a DataNode
• Managing HDFS Scaling
• Adding DataNodes

Balancing a Hadoop Cluster
• Managing Hadoop Balancing
• Balancing Hadoop Clusters

Backup and Recovery for HDFS
• Managing HDFS Backup and Recovery
• Copying Data

Managing Jobs
• Examining MapReduce Job Management
• Performing MapReduce Job Management

Upgrades for a Hadoop Cluster
• Managing Hadoop Upgrades

Stabilizing Hadoop Clusters

Hadoop Stability
• Exploring Event Management
• Exploring Incident Management
• Exploring Problem Management
• Examining Ganglia
• Examining Ganglia Functionality
• Installing Ganglia
• Examining Hadoop Metrics2
• Install Hadoop Metrics2 for Ganglia
• Exploring Ganglia
• Using Ganglia
• Examining Nagios
• Installing Nagios
• Nagios Contact Records
• Nagios Push
• Using Nagios Commands
• Using Nagios
• Using Hadoop Metrics2 for Nagios
• Examining Hadoop Logs
• Configure Logging for Jobs
• Configuring log4j for Hadoop
• Configuring JobHistoryServer logs
• Configuring Hadoop Logs
• Exploring Problem Management Lifecycle
• Examining Problem Management Best Practices
• Examining Common Problems
• Performing Root Cause Analysis

Capacity Management for Hadoop Clusters

Capacity Management
• Examining Capacity Management
• Examining Capacity Strategies

HDFS Capacity
• Examining Schedulers
• Setting HDFS Quotas

YARN Capacity
• Examining MRv2
• Exploring Fair Schedulers
• Examining Fair Scheduler Algorithms
• Examining Scheduler Behaviors
• Monitoring Fair Share
• Examining Single Resource Fairness
• Balancing Resources
• Examining Single Resource Fairness Configurations
• Configuring Single Resource Fairness
• Examining Minimum Resources
• Configuring Minimum Resources
• Examining Preemption
• Configuring Preemption

Service Performance
• Examining Dominant Resource Fairness
• Writing Service Levels

Performance Tuning of Hadoop Clusters

Performance Tuning Hadoop Clusters
• Managing Performance Tuning
• Examining Best Practices for Performance Tuning

Performance Tuning Networks
• Examining Best Practices for Network Tuning
• Installing Compression

Performance Tuning Servers
• Examining Operating System Tune Up Options
• Examining Java Tune Up Options
• Examining Input and Output Tune Up Options

Performance Tuning Memory
• Optimizing Memory for Daemons
• Optimizing Memory for YARN
• Optimizing Memory for Containers
• Tuning Memory for Hadoop Clusters

Performance Tuning HDFS
• Examining Tune Up Options for HDFS
• Examining HDFS Data Blocks
• Testing Data Blocks
• Performance Tuning HDFS

Performance Tuning YARN
• Examining Tune UP Options for YARN
• Configure Speculative Execution
• Examining MapReduce Tune Up Options
• Performance Tuning MapReduce
• Examining Benchmarking
• Examining Best Practices for Benchmarking
• Stress Testing and Benchmarking Hadoop Clusters

Modeling Applications
• Examining Applications Modeling

Cloudera Manager and Hadoop Clusters

Cluster Management Tools
- Defining Cluster Management
- Examining Cluster Management Tools

Cloudera Manager Introduction
- Examining Cloudera Manager
- Installing Cloudera Manager
- Deploying Clusters
- Installing Hadoop with Cloudera Manager

Cloudera Manager Administration
- Exploring Cloudera Manager Admin Console
- Exploring Cloudera Manager Architecture
- Performing Cluster Management
- Managing Services
- Managing Hosts with Cloudera Manager
- Setting Cloudera Manager for High Availability
- Managing Resources
- Monitoring with Cloudera Manager
- Diagnosing with Cloudera Manager
- Improving Performance
- Installing and Configuring Impala
- Installing and Configuring Sentry
- Using Hive for Sentry Administration
- Using Cloudera Manager for Administration

Manage Data with Hue
- Configuring Hue with MySQL
- Importing Data with Hue
- Running Hive Jobs with Hue
- Editing Oozie Workflows with Hue

Visit us at www.globalknowledge.com or call us at 1-866-716-6688.

Date created: 7/9/2019 2:42:21 AM
Copyright © 2019 Global Knowledge Training LLC. All Rights Reserved.