Unogeeks

Big Data Hadoop Training Course Content

Course Duration - 45-50 Hrs., 7-8 Weeks

Course Information

Batch Options

Weekday Batch
Mon - Fri - 1.5 Hr./Day

About the Trainer

Industry Expert Trainer with 15+ Years Real Time Work Experience at Top US Based Product and Consulting Firms

Contact Us

Mobile: +91 73960 33555
Whatsapp: +91 73960 33555
Mail: Prasad@unogeeks.com
Website: Unogeeks.com

Introduction To Big Data Hadoop Training

Hadoop

Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation

What you'll learn

- HDFS and MapReduce
- Hive, Pig, Flume, Sqoop and HBase
- Spark - RDDs, Aggregating Data, Writing & Deploying Spark Apps
- Parallel processing, RDD persistence, Lib
- Kafka and It's integration with Apache Flume
- Spark Streaming, SQL, Data Frames, Scheduling & Portioning
- Master Hadoop Administration Skills
- Prepare for CCA175 certification exams and get Job Ready
- Resume & Interview preparation and Job Assistance

Course Content

Module 1: Introduction to Big Data and Hadoop

- Big Data Overview
- Big Data Analytics
- What is Big Data?
- Challenges of Traditional Systems
- Distributed systems
- Introduction to Hadoop
- Components of Hadoop Ecosystem
- Commercial Hadoop Distributions

Module 2: Understanding HDFS and MapReduce

- Introduction to MapReduce
- Introduction to HDFS
- Hadoop Distributed File System – Replications, Block Size, Secondary node, High Availability
- YARN – resource manager and node manager

Module 3: Hadoop Installation and Setup

- Architecture of Hadoop cluster
- What is High Availability and Federation?
- How to setup a production cluster?
- Various shell commands in Hadoop
- Understanding configuration files in Hadoop
- Installing a single node cluster with Cloudera Manager
- Understanding Spark, Scala, Sqoop, Pig, and Flume

Module 4: Deep Dive in MapReduce

- Learning the working mechanism of MapReduce
- Understanding the mapping and reducing stages in MR
- Various terms in MR like Input & Output Format, Partitioners, Combiners, Shuffle, and Sort

Module 5: Introduction to Hive

- Introducing Hadoop Hive
- Detailed architecture of Hive
- Comparing Hive with Pig and RDBMS
- Working with Hive Query Language
- Creation of a database, table, group by and other clauses
- Various types of Hive tables, HCatalog
- Storing the Hive Results, Hive partitioning, and Buckets

Module 6: Advanced Hive and Impala

- Indexing in Hive
- The ap Side Join in Hive
- Working with complex data types
- The Hive user-defined functions
- Introduction to Impala
- Comparing Hive with Impala
- The detailed architecture of Impala

Module 7: Introduction to Pig

- Apache Pig introduction and its various features
- Various data types and schema in Hive
- The available functions in Pig, Hive Bags, Tuples, and Fields

Module 8: Flume, Sqoop and HBase

- Apache Sqoop introduction
- Importing and exporting data
- Performance improvement with Sqoop
- Sqoop limitations
- Introduction to Flume and understanding the architecture of Flume
- What is HBase and the CAP theorem?

Module 9: Writing Spark Applications Using Scala

- Using Scala for writing Apache Spark applications
- Detailed study of Scala
- The need for Scala
- The concept of object-oriented programming
- Executing the Scala code
- Scala Classes - Getters, Setters, & Constructors
- Scala Classes - Abstract, extending objects & Overriding

Module 10: Project Use Case

- Introduction to Scala packages and imports
- The selective imports
- The Scala test classes
- Introduction to JUnit test class
- JUnit interface via JUnit 3 suite for Scala test
- Packaging of Scala applications in the directory structure
- Examples of Spark Split and Spark Scala

Module 11: Introduction to Spark

- Introduction to Spark
- Spark overcomes the drawbacks of working on MapReduce
- Understanding in-memory MapReduce
- Interactive operations on MapReduce
- Spark stack, fine vs. coarse-grained update
- Spark stack, Spark Hadoop YARN, HDFS Revision, and YARN Revision
- The overview of Spark and how it is better than Hadoop
- Deploying Spark without Hadoop
- Spark history server and Cloudera distribution

Module 12: Spark Basics

- Spark installation guide
- Spark configuration
- Memory management
- Executor memory vs. driver memory
- Working with Spark Shell
- The concept of resilient distributed datasets (RDD)
- Learning to do functional programming in Spark
- The architecture of Spark

Module 13: Working with RDDs in Spark

- Spark RDD
- Creating RDDs
- RDD partitioning
- Operations and transformation in RDD
- Deep dive into Spark RDDs
- The RDD general operations

- Read-only partitioned collection of records
- Using the concept of RDD for faster and efficient data processing
- RDD action for the collect, count, collects map, save-as-text-files, and pair RDD functions

Module 14: Aggregating Data with Pair RDDs

- Understanding the concept of key-value pair in RDDs
- Learning how Spark makes MapReduce operations faster
- Various operations of RDD
- MapReduce interactive operations
- Fine and coarse-grained update
- Spark stack

Module 15: Writing and Deploying Spark Applications

- Comparing the Spark applications with Spark Shell
- Creating a Spark application using Scala or Java
- Deploying a Spark application
- Scala built application
- Creation of the mutable list, set and set operations, list, tuple, and concatenating list
- Creating an application using SBT
- Deploying an application using Maven
- The web user interface of Spark application
- A real-world example of Spark
- Configuring of Spark

Module 16: Parallel Processing

- Learning about Spark parallel processing
- Deploying on a cluster
- Introduction to Spark partitions
- File-based partitioning of RDDs
- Understanding of HDFS and data locality
- Mastering the technique of parallel operations
- Comparing repartition and coalesce
- RDD actions

Module 17: Spark RDD Persistence

- The execution flow in Spark
- Understanding the RDD persistence overview
- Spark execution flow, and Spark terminology
- Distribution shared memory vs. RDD
- RDD limitations
- Spark shell arguments
- Distributed persistence
- RDD lineage
- Key-value pair for sorting implicit conversions like CountByKey, ReduceByKey, SortByKey

Module 18: Spark MLlib

- Introduction to Machine Learning
- Types of Machine Learning
- Introduction to MLlib
- Various ML algorithms supported by MLlib
- Linear & logistic regression, decision tree, random forest, and K-means clustering techniques

Module 19: Integrating Apache Flume and Apache Kafka

- Why Kafka and what is Kafka?
- Kafka architecture
- Kafka workflow
- Configuring Kafka cluster
- Operations
- Kafka monitoring tools
- Integrating Apache Flume and Apache Kafka

Module 20: Spark Streaming

- Introduction to Spark Streaming
- Features of Spark Streaming
- Spark Streaming workflow
- Initializing StreamingContext, discretized Streams (DStreams), input DStreams and Receivers
- Transformations & output operations on DStreams, windowed operators and why it is useful
- Important windowed operators and stateful operators

Module 21: Improving Spark Performance

- Introduction to various variables in Spark like shared variables and broadcast variables
- Learning about accumulators
- The common performance issues
- Troubleshooting the performance problems

Module 22: Spark SQL and Data Frames

- Learning about Spark SQL
- The context of SQL in Spark for providing structured data processing
- JSON support in Spark SQL
- Working with XML data
- Parquet files
- Creating Hive context
- Writing data frame to Hive
- Reading JDBC files
- Understanding the data frames in Spark
- Creating Data Frames
- Manual inferring of schema
- Working with CSV files
- Reading JDBC tables
- Data frame to JDBC
- User-defined functions in Spark SQL
- Shared variables and accumulators
- Learning to query and transform data in data frames
- Data frame provides the benefit of both Spark RDD and Spark SQL
- Deploying Hive on Spark as the execution engine

Module 23: Scheduling/Partitioning

- Learning about the scheduling and partitioning in Spark
- Hash & Range partition
- Scheduling within and around applications
- Static partitioning, dynamic sharing, and fair scheduling
- Map partition with index, the Zip, and GroupByKey
- Spark master high availability, standby masters with ZooKeeper, single-node recovery with the local file system and high order functions

Module 24: Hadoop Administration – Multi-node Cluster Setup Using Amazon EC2

- Create a 4-node Hadoop cluster setup
- Running the MapReduce Jobs on the Hadoop cluster
- Successfully running the MapReduce code
- Working with the Cloudera Manager setup

Module 25: Hadoop Administration – Cluster Configuration

- Overview of Hadoop configuration
- The importance of Hadoop configuration file
- The various parameters and values of configuration
- The HDFS parameters and MapReduce parameters
- Setting up the Hadoop environment
- The Include and Exclude configuration files
- The administration and maintenance of name node, data node directory structures, and files
- What is a File system image?
- Understanding Edit log

Module 26: Hadoop Administration – Maintenance, Monitoring and Troubleshooting

- Introduction to the checkpoint procedure, name node failure
- How to ensure the recovery procedure, Safe Mode, Metadata and Data backup,
- Various potential problems and solutions, what to look for and how to add and remove nodes

Module 27: ETL Connectivity with Hadoop Ecosystem

- How ETL tools work in Big Data industry?
- Introduction to ETL and data warehousing
- Working with prominent use cases of Big Data in ETL industry
- End-to-end ETL PoC showing Big Data integration with ETL tool

Module 28: Hadoop Application Testing

- Importance of testing
- Unit testing, Integration testing, Performance testing
- Diagnostics, Nightly QA test, Benchmark and end-to-end tests
- Functional testing, Release certification testing, Security testing
- Scalability testing, Commissioning and Decommissioning of data nodes testing
- Reliability testing, and Release testing

Module 29: Roles and Responsibilities of Hadoop Testing Professional

- Understanding the Requirement
- Preparation of the Testing Estimation
- Test Cases, Test Data, Test Bed Creation, Test Execution
- Defect Reporting, Defect Retest, Daily Status report delivery, Test completion
- ETL testing at every stage (HDFS, Hive and HBase) while loading the input (logs, files, records, etc.) using Sqoop/Flume
- Data verification, Reconciliation, User Authorization & Authentication testing (Groups, Users, Privileges, etc.),
- Reporting defects to the development team or manager and driving them to closure
- Consolidating all the defects and create defect reports
- Validating new feature and issues in Core Hadoop

Module 30: Framework Called MRUnit for Testing of MapReduce Programs

- Report defects to the development team or manager and driving them to closure
- Consolidate all the defects and create defect reports
- Responsible for creating a testing framework called MRUnit for testing of MapReduce programs

Module 31: Unit Testing

- Automation testing using the OOZIE
- Data validation using the query surge tool

Module 32: Test Execution

- Test plan for HDFS upgrade
- Test automation and result

Module 33: CCA175 Spark and Hadoop Developer Certification Exam Prep

- Explain CCA175 Spark and Hadoop Developer Certification Options
- Discuss 50+ Important CCA175 Certification Questions
- Practice CCA175 Certification questions

Module 34: Resume Preparation, Interview and Job Assistance

- Prepare Crisp Resume as Big Data Hadoop Developer
- Discuss common interview questions in Hadoop
- Explain students what jobs they should target and how