Big Data Hadoop and Spark Developer

simplilearn

Table of **Contents:**

Program **Overview:**

This Big Data Hadoop Certification course is designed to give you an in-depth knowledge of the big data framework using Hadoop and Spark. In this hands-on big data course, you will execute real-life, industry-based projects using Simplilearn's integrated labs.

Program **Features:**

- 74 hours of blended learning
- 22 hours of Online self-paced learning
- 52 hours of instructor-led training
- Four industry-based course-end projects
- Interactive learning with integrated labs
- 2Curriculum aligned to Cloudera CCA175 certification exam
- Training on essential big data and Hadoop ecosystem tools, and Apache Spark
- Dedicated mentoring session from faculty of industry experts

Delivery **Mode:**

Blended - Online self-paced learning and live virtual classroom

Prerequisites:

It is recommended that you have knowledge of:

- Core Java
- SQL

Target **Audience:**

- Analytics professionals
- Senior IT professionals
- Testing and mainframe professionals
- Data management professionals
- Business intelligence professionals
- Project managers
- Graduates looking to begin a career in big data analytics

Key Learning **Outcomes:**

This Big Data Hadoop and Spark Developer course will enable you to:

- Learn how to navigate the Hadoop ecosystem and understand how to optimize its use
- Ingest data using Sqoop, Flume, and Kafka.
- Implement partitioning, bucketing, and indexing in Hive
- Work with RDD in Apache Spark
- Process real-time streaming data
- Perform DataFrame operations in Spark using SQL queries
- Implement User-Defined Functions (UDF) and User-Defined Attribute Functions (UDAF) in Spark

Certification **Alignment:**

Our curriculum is aligned to Cloudera CCA175 certification exam.

Certification Details and **Criteria:**

- Completion of at least 85 percent of online self-paced learning or attendance of one live virtual classroom
- A score of at least 75 percent in course-end assessment
- Successful evaluation in at least one project

Course **Curriculum:**

Lesson 01 - Introduction to Bigdata and Hadoop

- Introduction to Big Data and Hadoop
- Introduction to Big Data
- Big Data Analytics
- What is Big Data?
- Four vs of Big Data
- Case Study Royal Bank of Scotland
- Challenges of Traditional System
- Distributed Systems
- Introduction to Hadoop
- Components of Hadoop Ecosystem Part One
- Components of Hadoop Ecosystem Part Two
- Components of Hadoop Ecosystem Part Three
- Commercial Hadoop Distributions
- Demo: Walkthrough of Simplilearn Cloudlab
- Key Takeaways
- Knowledge Check

Lesson 02 - Hadoop Architecture Distributed Storage (HDFS) and YARN

- Hadoop Architecture Distributed Storage (HDFS) and YARN
- What is HDFS
- Need for HDFS
- Regular File System vs HDFS
- Characteristics of HDFS
- HDFS Architecture and Components
- High Availability Cluster Implementations
- HDFS Component File System Namespace
- Data Block Split
- Data Replication Topology
- HDFS Command Line
- Demo: Common HDFS Commands
- Practice Project: HDFS Command Line
- Yarn Introduction
- Yarn Use Case
- Yarn and its Architecture
- Resource Manager
- How Resource Manager Operates
- Application Master
- How Yarn Runs an Application
- Tools for Yarn Developers
- Demo: Walkthrough of Cluster Part One
- Demo: Walkthrough of Cluster Part Two
- Key Takeaways
- Knowledge Check
- Practice Project: Hadoop Architecture, distributed Storage (HDFS) and Yarn

Lesson 03 - Data Ingestion into Big Data Systems and ETL

- Data Ingestion Into Big Data Systems and Etl
- Data Ingestion Overview Part One
- Data Ingestion Overview Part Two
- Apache Sqoop
- Sqoop and Its Uses
- Sqoop Processing
- Sqoop Import Process
- Sqoop Connectors
- Demo: Importing and Exporting Data from MySQL to HDFS
- Practice Project: Apache Sqoop
- Apache Flume
- Flume Model
- Scalability in Flume
- Components in Flume's Architecture
- Configuring Flume Components
- Demo: Ingest Twitter Data
- Apache Kafka
- Aggregating User Activity Using Kafka
- Kafka Data Model
- Partitions
- Apache Kafka Architecture
- Demo: Setup Kafka Cluster
- Producer Side API Example
- Consumer Side API
- Consumer Side API Example
- Kafka Connect
- Demo: Creating Sample Kafka Data Pipeline Using Producer and Consumer
- Key Takeaways
- Knowledge Check
- Practice Project: Data Ingestion Into Big Data Systems and ETL

Lesson 04 - Distributed Processing MapReduce Framework and Pig

- Distributed Processing Mapreduce Framework and Pig
- Distributed Processing in Mapreduce
- Word Count Example
- Map Execution Phases
- Map Execution Distributed Two Node Environment
- Mapreduce Jobs
- Hadoop Mapreduce Job Work Interaction
- Setting Up the Environment for Mapreduce Development
- Set of Classes
- Creating a New Project
- Advanced Mapreduce
- Data Types in Hadoop
- Output formats in Mapreduce
- Using Distributed Cache
- Joins in Mapreduce
- Replicated Join
- Introduction to Pig
- Components of Pig
- Pig Data Model
- Pig Interactive Modes
- Pig Operations
- Various Relations Performed by Developers
- Demo: Analyzing Web Log Data Using Mapreduce
- Demo: Analyzing Sales Data and Solving Kpis Using Pig
- Practice Project: Apache Pig
- Demo: Wordcount
- Key Takeaways
- Knowledge Check
- Practice Project: Distributed Processing - Mapreduce Framework and Pig

Lesson 05 - Apache Hive

- Apache Hive
- Hive SQL over Hadoop Mapreduce
- Hive Architecture
- Interfaces to Run Hive Queries
- Running Beeline from Command Line
- Hive Metastore
- Hive DDL and DML
- Creating New Table
- Data Types
- Validation of Data
- File Format Types
- Data Serialization
- Hive Table and Avro Schema
- Hive Optimization Partitioning Bucketing and Sampling
- Non-Partitioned Table
- Data Insertion
- Dynamic Partitioning in Hive
- Bucketing
- What Do Buckets Do?
- Hive Analytics UDF and UDAF
- Other Functions of Hive
- Demo: Real-time Analysis and Data Filtration
- Demo: Real-World Problem
- Demo: Data Representation and Import Using Hive
- Key Takeaways
- Knowledge Check
- Practice Project: Apache Hive

Lesson 06 - NoSQL Databases HBase

- NoSQL Databases HBase
- NoSQL Introduction
- Demo: Yarn Tuning
- Hbase Overview
- Hbase Architecture
- Data Model
- Connecting to HBase
- Practice Project: HBase Shell
- Key Takeaways
- Knowledge Check
- Practice Project: NoSQL Databases - HBase

Lesson 07 - Basics of Functional Programming and Scala

- Basics of Functional Programming and Scala
- Introduction to Scala
- Demo: Scala Installation
- Functional Programming
- Programming With Scala
- Demo: Basic Literals and Arithmetic Programming
- Demo: Logical Operators
- Type Inference Classes Objects and Functions in Scala
- Demo: Type Inference Functions Anonymous Function and Class
- Collections
- Types of Collections
- Demo: Five Types of Collections
- Demo: Operations on List
- Scala REPL
- Demo: Features of Scala REPL
- Key Takeaways
- Knowledge Check
- Practice Project: Apache Hive

Lesson 08 - Apache Spark Next-Generation Big Data Framework

- Apache Spark Next-Generation Big Data Framework
- History of Spark
- Limitations of Mapreduce in Hadoop
- Introduction to Apache Spark
- Components of Spark
- Application of In-memory Processing
- Hadoop Ecosystem vs Spark
- Advantages of Spark
- Spark Architecture
- Spark Cluster in Real World
- Demo: Running a Scala Programs in Spark Shell
- Demo: Setting Up Execution Environment in IDE
- Demo: Spark Web UI
- Key Takeaways
- Knowledge Check
- Practice Project: Apache Spark Next-Generation Big Data Framework

Lesson 09 - Spark Core Processing RDD

- Introduction to Spark RDD
- RDD in Spark
- Creating Spark RDD
- Pair RDD
- RDD Operations
- Demo: Spark Transformation Detailed Exploration Using Scala Examples
- Demo: Spark Action Detailed Exploration Using Scala
- Caching and Persistence
- Storage Levels
- Lineage and DAG
- Need for DAG
- Debugging in Spark
- Partitioning in Spark
- Scheduling in Spark
- Shuffling in Spark
- Sort Shuffle
- Aggregating Data With Paired RDD
- Demo: Spark Application With Data Written Back to HDFS and Spark UI
- Demo: Changing Spark Application Parameters
- Demo: Handling Different File Formats
- Demo: Spark RDD With Real-world Application
- Demo: Optimizing Spark Jobs
- Key Takeaways
- Knowledge Check
- Practice Project: Spark Core Processing RDD

Lesson 10 - Spark SQL Processing DataFrames

- Spark SQL Processing DataFrames
- Spark SQL Introduction
- Spark SQL Architecture
- Dataframes
- Demo: Handling Various Data Formats
- Demo: Implement Various Dataframe Operations
- Demo: UDF and UDAF
- Interoperating With RDDs
- Demo: Process Dataframe Using SQL Query
- RDD vs Dataframe vs Dataset
- Practice Project: Processing Dataframes
- Key Takeaways
- Knowledge Check
- Practice Project: Spark SQL - Processing Dataframes

Lesson 11 - Spark MLib Modelling BigData with Spark

- Spark Mlib Modeling Big Data With Spark
- Role of Data Scientist and Data Analyst in Big Data
- Analytics in Spark
- Machine Learning
- Supervised Learning
- Demo: Classification of Linear SVM
- Demo: Linear Regression With Real World Case Studies
- Unsupervised Learning
- Demo: Unsupervised Clustering K-means
- Reinforcement Learning
- Semi-supervised Learning
- Overview of Mlib
- Mlib Pipelines
- Key Takeaways
- Knowledge Check
- Practice Project: Spark Mlib - Modelling Big data With Spark

Lesson 12 - Stream Processing Frameworks and Spark Streaming

- Streaming Overview
- Real-time Processing of Big Data
- Data Processing Architectures
- Demo: Real-time Data Processing
- Spark Streaming
- Demo: Writing Spark Streaming Application
- Introduction to DStreams
- Transformations on DStreams
- Design Patterns for Using Foreachrdd
- State Operations
- Windowing Operations
- Join Operations Stream-dataset Join
- Demo: Windowing of Real-time Data Processing
- Streaming Sources
- Demo: Processing Twitter Streaming Data
- Structured Spark Streaming
- Use Case Banking Transactions
- Structured Streaming Architecture Model and Its Components
- Output Sinks
- Structured Streaming APIs
- Constructing Columns in Structured Streaming
- Windowed Operations on Event-time
- Use Cases
- Demo: Streaming Pipeline
- Practice Project: Spark Streaming
- Key Takeaways
- Knowledge Check
- Practice Project: Stream Processing Frameworks and Spark Streaming

Lesson 13 - Spark GraphX

- Spark GraphX
- Introduction to Graph
- GraphX in Spark
- GraphX Operators
- Join Operators
- GraphX Parallel System
- Algorithms in Spark
- Pregel API
- Use Case of GraphX
- Demo: GraphX Vertex Predicate
- Demo: Page Rank Algorithm
- Key Takeaways
- Knowledge Check
- Practice Project: Spark GraphX
- Project Assistance

Course End **Projects:**

The course includes four real-world, industry-based projects. The successful evaluation of one of the following projects is a part of the certification eligibility criteria:

Project 1: Analyzing Historical Insurance Claims

Use Hadoop features to predict patterns and share actionable insights for a car insurance company

This project uses New York Stock Exchange data from 2010 to 2016, captured from 500+ listed companies. The data set consists of each listed company's intraday prices and volume traded. The data is used in both machine learning and exploratory analysis projects for the purposes of automating the trading process and predicting the next trading-day winners or losers. The scope of this project is limited to exploratory data analysis.

Domain: BFSI

Project 2: Employee Review of Comment Analysis

Use Hive features for data analysis and share the actionable insights with the HR team for the purpose of taking corrective actions.

The HR team is surfing social media to gather current and ex-employee feedback and sentiments. This information will be used to derive actionable insights and take corrective actions to improve the employer-employee relationship. The data is web-scraped from Glassdoor and contains detailed reviews of 67K employees from Google, Amazon, Facebook, Apple, Microsoft, and Netflix.

Domain: Human Resources

Project 3: K-Means Clustering for Telecommunication Domain

LoudAcre Mobile is a mobile phone service provider whichthat has introduced a new open network campaign. As a part of this campaign, the company has invited users to complain about mobile phone network towers in their area if they are experiencing connectivity issues with their present mobile network. LoudAcre has collected the dataset of users who have complained.

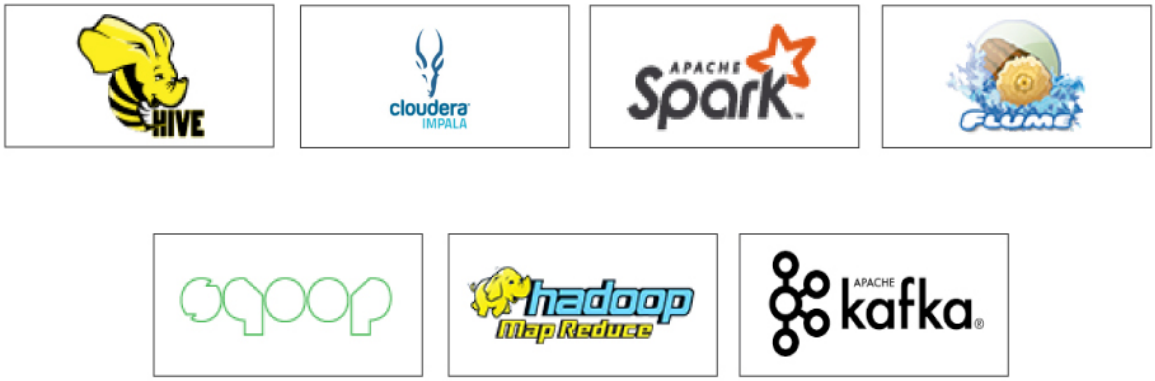
Domain: Telecommunication

Project 4: Market Analysis in Banking Domain

Our client, a Portuguese banking institution, ran a marketing campaign to convince potential customers to invest in a bank term deposit promotion. The marketing campaign pitches were delivered by phone calls. Often, however, the same customer was contacted more than once. You have to perform the marketing analysis of the data generated by this campaign, keeping in mind the redundant calls

Domain: Banking(Market Analysis)

Tools **Covered:**



Customer **Reviews:**

Hari Harasan
Technical Architect at Infosys

The session on Map reducer was really interesting, a complex topic was very well explained in an understandable manner.

Vignesh Balasubramanian
Senior Operations Professional @ IBM

I have enrolled for Big Data Hadoop Spark developer course from Simplilearn. The course was well organized, covering all the root concepts and relevant real-time experience. The trainer was well equipped to solve all the doubts during the training. Cloud lab facility and materials provided were on point.

Anusha T S
Software Developer at Zibtek

I have enrolled in Big Data Hadoop and Spark Developer from Simplilearn. I like the teaching method of the trainers. He was very helpful and knowledgeable. Overall I am very happy with Simplilearn. Their cloud labs are also very user-friendly. I would highly recommend my friends to take a course from here and upskill themselves.

Permoon Ansari
Project Manager at IBM

Gautam has been the best trainer throughout the session. He took ample time to explain the course content and ensured that the class understands the concepts. He's undoubtedly one of the best in the industry. I'm delighted to have attended his sessions.

About **Us:**

Simplilearn is a leader in digital skills training, focused on the emerging technologies that are transforming our world. Our blended learning approach drives learner engagement and is backed by the industry's highest completion rates. Partnering with professionals and companies, we identify their unique needs and provide outcome-centric solutions to help them achieve their professional goals.

For more information, please visit our website:
https://www.simplilearn.com/big-data-and-analytics/python-for-data-science-training

simplilearn | simplilearn.com

Founded in 2009, Simplilearn is one of the world's leading providers of online training for Digital Marketing, Cloud Computing, Project Management, Data Science, IT Service Management, Software Development and many other emerging technologies. Based in Bangalore, India, San Francisco, California, and Raleigh, North Carolina, Simplilearn partners with companies and individuals to address their unique needs, providing training and coaching to help working professionals meet their career goals. Simplilearn has enabled over 1 million professionals and companies across 150+ countries train, certify and upskill their employees.

Simplilearn's 400+ training courses are designed and updated by world-class industry experts. Their blended learning approach combines e-learning classes, instructor-led live virtual classrooms, applied learning projects, and 24/7 teaching assistance. More than 40 global training organizations have recognized Simplilearn as an official provider of certification training. The company has been named the 8th most influential education brand in the world by LinkedIn.

India – United States – Singapore

© 2009-2019 - Simplilearn Solutions. All Rights Reserved.
The certification names are the trademarks of their respective owners.