SPRINTZEAL
Empowering Professionals

BIG Data Hadoop and Analyst Certification

Course Agenda

Total: 42 Hours of Training

Introduction:

This course will enable an Analyst to work on Big Data and Hadoop which takes into consideration the on-going demands of the industry to process and analyse data at high speeds. This training course will give you the right skills to deploy various tools and techniques to be a Hadoop Analyst working with Big Data.

Lesson 1: Course Introduction

- Course Introduction
- Accessing Practice Lab

Lesson 2: Introduction to Big Data 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 Distribution
- Key Takeaways
- Knowledge Check

Lesson 3: 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 Implementation
- HDFS Component File System Namespace
- Data Block Split
- Data Replication Topology
- HDFS Command Line
- YARN Introduction
- YARN Use Case
- YARN and its Architecture
- Resource Manager
- How Resource Manger 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
- Hadoop Architecture ,Distributed Storage (HDFS) and YARN

Lesson 4: 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
- Apache Sqoop

- Apache Flume
- FLUME Model
- Scalability in FLUME
- Components in FLUME 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
- Key Takeaways
- Knowledge Check
- Data Ingestion into Big Data Systems and ETL

Lesson 5: Distributed Processing – MapReduce Framework and Pig

- Distribute Processing MapReduce Framework and Pig
- Distribute Processing MapReduce
- Word Count Example
- Map Execution Phases
- Map Execution Distributed Two Node Environment
- MapReduce Jobs
- Hadoop MapReduce Job Work Interaction
- Setting Up 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 to Pig
- Pig Data Model
- Pig Interactive Modes
- Pig Operations
- Various Relations Performed by Developers
- Demo: Analysing Web Log Data Using MapReduce
- Demo: Analysing Sales Data and Solving KPIs using PIG
- Apache Pig
- Demo: Wordcount
- Key takeaways
- Knowledge Check
- Distibuted Processing- MapReduce Framework Pig

Lesson 6: 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 Filteration
- Demo: Real World Problem

- Demo : Data Representation and Import using Hive
- Key Takeaways
- Knowledge Check
- Apache Hive

Lesson 7: NoSQL Database -HBase

- NoSQL Database -HBase
- NoSQL Introduction
- Demo: YARN Turning
- HBase Overview
- HBase Architecture
- Data Model
- Connecting to HBase
- HBase Shell
- Key Takeaways
- Knowledge Check
- NoSQL Databases- HBase

Lesson 8: Basics of Functional Programming and Scala

- Basics of Functional Programming and Scala
- Introduction to Scala
- Demo: Scala Introduction
- Functional Programming
- Programming with Scala
- Basic Literals and Arithmetic Operators
- Logical Operators
- Type Inference Classes Objects and Functions in Scala
- Demo: Type Inference Functions Anonymous Functions 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
- Basics of Functional Programming and Scala

Lesson 9: 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 Shells
- Demo: Setting up Execution Environment in IDE
- Demo: Spark Web UI
- Key Takeaways
- Knowledge Check
- Apache Spark Next Generation Big Data Framework

Lesson 10: Spark Core Processing RDD

- Processing RDD
- Introduction to Spark RDD
- RDD in Spark
- Creating Spark RDD
- Pair RDD
- RDD Operations
- Demo: Spark Transformation Detailed Exploration Using Scala
- Demo: Spark Action Detailed Exploration Using Scala
- Caching and Persistence
- Storage Levels
- Linage and DAG
- Need for DAG
- Debugging in Spark
- Partitioning in Spark
- Scheduling in Spark
- Shuffling in Spark
- Sort Shuffle

- Aggregating Data with pair RDD
- Demo: Spark Application with Data Written Back to HDFS
- 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
- Spark Core Processing RDD

Lesson 11: Spark SQL Processing Data Frames

- Spark SQL Processing Data Frames
- Spark SQL Introduction
- Spark SQL Architecture
- DataFrames
- Demo: Handling Various Data Formats
- Demo: Implement Various DataFrames Operations
- Demo: UDF and UDAF
- Interoperations with RDDs
- Demo: Process DataFrame Using SQL Query
- RDD vs DataFrames vs Dataset
- Processing DataFrames
- Key Takeaways
- Knowledge Check
- Spark SQL Processing DataFrames

Lesson 12: Spark MLLib Modelling Big Data with Spark

- Spark MLLib Modelling Big Data with Spark
- Roles 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 MLlib

- MLlib Pipelines
- Key Takeaways
- Knowledge Check
- Spark MLLib Modelling Big Data with Spark

Lesson 13: Stream Processing Frameworks and Spark Streaming

- Stream Processing Frameworks and Spark Streaming
- Streaming Overview
- Real Time Processing of Big Data
- Data Processing Architecture
- Data Processing Architecture
- Demo: Real Time Data Processing
- Spark Streaming
- Demo: Writing Spark Streaming Application
- Introduction to DStreams
- Design Patter for using ForreachRDD
- State Operations
- Windowing Operations
- Join Operations stream-dataset Join
- Demo: Windowing of Real – Time Data Processing
- Streaming Sources
- Demo: Processing Twitter Streaming Processing 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
- Spark Streaming
- Key Takeaways
- Knowledge Check
- Stream Processing Frameworks and Spark Streaming

Lesson 14: Spark GraphX

- Spark GraphX
- Introduction to Spark

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

Practice Projects

- Car Insurance Analysis
- Transactional Data Analysis