Module 01 - 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 Distributions

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 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