Big Data, Hadoop, and Spark Basics

Learning Objectives

- Describe Big Data, its impact, processing methods and tools, and use cases.
- Describe Hadoop architecture, ecosystem, practices, and applications, including Distributed File System (HDFS), HBase, Spark, and MapReduce.
- Describe Spark programming basics, including parallel programming basics, for DataFrames, data sets, and SparkSQL.
- Describe how Spark uses RDDs, creates data sets, and uses Catalyst and Tungsten to optimize SparkSQL.
- Apply Apache Spark development and runtime environment options.

Syllabus

Module 1: What is Big Data?

- Module Introduction and Learning Objectives
- What is Big Data?
- Impact of Big Data
- Parallel Processing and Scaleability
- Big Data Tools and Ecosystem
- Open Source and Big Data
- Beyond the Hype
- Big Data Use Cases
- Summary & Highlights
- Practice Quiz: Introduction to Big Data
- Graded Quiz: Introduction to Big Data

Module 2: Introduction to the Hadoop Ecosystem

- Module Introduction and Learning Objectives
- Introduction to Hadoop
- Intro to MapReduce
- Hadoop Ecosystem
- HDFS
- HIVE
- HBASE
- Hands-on Lab: MapReduce
- Summary & Highlights
- Practice Quiz: Introduction to Hadoop
- Graded Quiz: Introduction to Hadoop

Module 3: Apache Spark

- Module Introduction and Learning Objectives
- Why use Apache Spark?
- Functional Programming Basics
- Parallel Programming using Resilient Distributed Datasets
- Scale out / Data Parallelism in Apache Spark
- Dataframes and SparkSQL
- Hands-on Lab: Getting started with Spark on Watson Studio
- Hands-on Lab: Practical examples with PySpark
- Summary & Highlights
- Practice Quiz: Introduction to Apache Spark
- Graded Quiz: Introduction to Apache Spark

Module 4: Data-Frames and SparkSQL

- Module Introduction and Learning Objectives
- RDDs in Parallel Programming and Spark
- Data-frames and Datasets
- Catalyst and Tungsten
- ETL with Data-frames
- Hands-on Lab: Jupyter Notebook for Introduction to Data-Frames
- Hands-on Lab: Introduction to Data-frames
- Real-world usage of SparkSQL
- Hands-on Lab: Jupyter Notebook for Introduction to SparkSQL
- Hands-on Lab: Introduction to SparkSQL
- Summary & Highlights
- Practice Quiz: Introduction to Data-Frames & SparkSQL
- Graded Quiz: Introduction to Data-Frames & SparkSQL

Module 5: Development and Runtime Environment options

- Module Introduction and Learning Objectives
- Apache Spark architecture
- Overview of Apache Spark Cluster Modes
- How to Run an Apache Spark Application
- Summary & Highlights
- Practice Quiz: Spark Architecture
- Graded Quiz: Spark Architecture
- Using Apache Spark on IBM Cloud
- Hands-on Lab: Getting started with Spark on IBM Cloud
- Setting Apache Spark Configuration
- Running Spark on Kubernetes
- Hands-on Lab: Spark on Kubernetes
- Summary & Highlights
- Practice Quiz: Spark Runtime Environments
- Graded Quiz: Spark Runtime Environments

Module 6: Monitoring & Tuning

- Module Introduction and Learning Objectives
- The Apache Spark User Interface
- Monitoring Application Progress
- Debugging Apache Spark Application Issues
- Understanding Memory resources
- Understanding Processor resources
- Hands-on Lab: Monitoring and Performance tuning
- Summary & Highlights
- Practice Quiz: Introduction to Monitoring & Tuning
- Graded Quiz: Introduction to Monitoring & Tuning

Final Exam

- Instructions for the Final Exam
- Graded Quiz: Final Quiz
- Congrats & Next Steps
- Team & Acknowledgements