O&B ACADEMY

COURSE SYLLABUS

BIGDATA-202

Apache Spark

BIGDATA-202

Apache Spark

OVERVIEW

Skill Level	:	Intermediate
Suitable for	:	It is suitable for a variety of job positions and roles within the field of data engineering, data science, and big data analytics. This includes data engineers, analysts, and data architects
Duration	:	10 Days

Apache Spark is a versatile and powerful framework for processing and analyzing large-scale data. Its in-memory processing, distributed computing capabilities, and diverse libraries make it a valuable tool for a wide range of data processing and analysis tasks, spanning from batch processing to real-time streaming and machine learning.

PREQUISITES

- BIGDATA-101 — Classic Hadoop
- BIGDATA-103 - Python for Data Engineers
- SQL-102 — PostgreSQL Training
- DEVOPS-101 — Docker and Kubernetes Fundamentals

LEARNING OUTCOMES

- Learning Spark will familiarize you with the concept of in-memory data processing and its advantages. You'll learn how Spark leverages memory to speed up computations and iterative algorithms, resulting in significant performance improvements.

- You'll learn techniques to optimize Spark jobs for efficiency, such as data partitioning, caching, and leveraging built-in optimizations. This knowledge is crucial for ensuring optimal performance in real-world scenarios.
- You'll understand how Spark integrates with other big data tools and ecosystems, like Hadoop, cloud platforms, databases, and data warehouses. This knowledge is essential for building end-to-end data pipelines.

COURSE OUTLINE

Day 1

- Introduction to Spark
- Spark vs X
  - Spark vs Classic MapReduce
  - Spark vs Hive
  - Spark vs Pig
  - Spark vs Sqoop
  - Spark vs Flink
- Modern Big Data Stacks
  - Airflow + Spark + S3 + Kubernetes
  - AWS Redshift
  - GCP BigQuery
  - Azure Synapse
  - Databricks
  - Snowflake
- Architecture
  - Driver
  - Executor
  - Nodes
  - RDD
  - Storage
  - Lifecycle
- Working Group Formation

Day 2

- Spark SQL
  - SparkSession
  - DataFrames
  - Local Files
  - JDBC/ODBC Server
  - CSV Files
  - JSON Files

Day 3

- Parquet Files
- ORC Files
- Hive Tables
- Caching
- Join Hints

Day 4

- Spark Structured Streaming
  - DataFrame
  - State Store
  - Sinks
  - Triggers
  - Checkpointing

Day 5

- Spark MLLib
  - Data Sources
  - Pipelines
  - Feature Extraction
  - Classification and Regression
  - Clustering
  - Collaborative Filtering
  - Pattern Mining
  - Model Selection and Tuning
  - Linear Methods

Day 6

- Spark GraphX
  - Property Graph
  - Operators
  - Pregel API
  - Graph Builders
  - Vertex and Edge RDDs

Day 7

- Spark Submit
- Spark Standalone
- Spark on HDFS/YARN
- Spark on S3/Kubernetes
  - Spark Operator
  - Spark Submit via Airflow
- Spark on AWS EMR Serverless

Day 8

- Zeppelin on Scala Spark
- Jupyter on Pyspark

Day 9-10

- Deployment
- Sample Application
- Individual and Group Work
- Presentations
- Final Exam

O&B ACADEMY

Engineering for the Real World

Enquiries

☎ +63 2 5322 2307

✉ training-sales@orangeandbronze.com