CONTENTS

Introduction to Data Engineering with Hadoop and Spark

- Course Introduction
- Overview of Data Engineering
- Introduction to Hadoop and Spark
- Setting up a development environment with Hadoop and Spark
- Hands-on: Installing Hadoop and Spark locally

Hadoop Fundamentals

- Understanding Hadoop Distributed File System (HDFS)
- Hadoop MapReduce paradigm
- Introduction to YARN (Yet Another Resource Negotiator)
- Running a simple MapReduce job | Hands-on: Working with HDFS and running a basic MapReduce job

Introduction to Apache Spark

- Overview of Apache Spark
- Spark architecture and components
- Spark RDD (Resilient Distributed Datasets)
- Transformations and Actions in Spark
- Hands-on: Writing and running Spark applications with Python

Spark DataFrames and SQL

- Introduction to Spark DataFrames
- Spark SQL for querying structured data
- Basic DataFrame operations

- Optimizations in Spark DataFrames
- Hands-on: Working with Spark DataFrames and performing SQL queries

Spark Streaming

- Real-time data processing with Spark Streaming
- Integrating Spark Streaming with external sources
- Handling stateful computations
- Hands-on: Building a simple Spark Streaming application

PySpark and Python API

- Introduction to PySpark
- Using Spark with Python
- Writing PySpark applications
- Interoperability between Spark and Python libraries
- Hands-on: Developing a PySpark application with Python

Spark MLlib for Machine Learning

- Overview of Spark MLlib
- Machine learning with Spark
- Building and training models with MLlib

Spark GraphX for Graph Processing

- Introduction to Spark GraphX
- Graph algorithms and computations

- Hands-on: Building and analyzing a graph using Spark GraphX

Spark Cluster Deployment and Optimization

- Understanding Spark Deployment Modes
- Best practices for Data Engineering with Hadoop and Spark
- Challenges and considerations in real-world scenarios
- Next steps in Data Engineering with Hadoop and Spark
- Setting up a Spark Cluster
- Monitoring and Debugging Spark Application

Data Warehousing

- Introduction to data warehousing concepts
- Comparison of traditional databases and data warehousing solutions
- Hands-on: Designing and implementing a simple data warehouse

ETL

- Understanding the ETL process in data engineering
- Tools and frameworks for ETL (e.g., Apache NiFi)
- Hands-on: Building a basic ETL pipeline

Distributed Data Storage Systems

- Overview of distributed storage systems (other than HDFS)
- Comparison of various storage options (e.g., Amazon S3, Google Cloud Storage)

CONTENTS

Data Lakes

- Integration of data lakes with data engineering pipelines
- Hands-on: Working with a data lake

Introduction to Apache Kafka in Data Engineering

- Overview of Apache Kafka
- Data Ingestion with Kafka
- Hands-on: Setting up and Using Kafka

Integrating Kafka into Data Pipelines

- Kafka Connect
- Real-time Data Processing with Kafka
- Hands-on: Building a Kafka-Powered Data Pipeline