APACHE SPARK

1: Introduction to Apache Spark
1.1.What is Big Data?
- 1.1.1. Definition and Characteristics
- 1.1.2. Challenges of Big Data

1.2.Introduction to Spark
- 1.2.1. Spark vs. Hadoop
- 1.2.2. Spark's Role in Big Data

2: Spark Fundamentals
2.1.Resilient Distributed Datasets (RDDs)
- 2.1.1. RDD Introduction
- 2.1.2. Creating and Transforming RDDs

2.2.Spark's Execution Model
- 2.2.1. Spark Driver and Executors
- 2.2.2. Stages and Tasks

2.3. Spark Clusters
- 2.3.1. Standalone vs. Cluster Managers
- 2.3.2. Cluster Configuration

3: Spark Data Processing
3.1.Loading Data into Spark
- 3.1.1. Reading from Various Data Sources
- 3.1.2. Streaming Data into Spark

3.2.Data Transformation
- 3.2.1. Map, Filter, and Reduce Operations
- 3.2.2. Joins and Aggregations

3.3.Caching and Persistence

- 3.3.1. In-Memory Data Storage
- 3.3.2. Cache Management

4: Advanced Spark Concepts

4.1. Spark SQL

- 4.1.1. Introduction to Spark SQL
- 4.1.2. Running SQL Queries in Spark

4.2.Streaming with Spark

- 4.2.1. Spark Streaming Overview
- 4.2.2. Processing Real-time Data

4.3.Machine Learning with Spark MLlib

- 4.3.1. MLlib Introduction
- 4.3.2. Building ML Pipelines

5: Spark Ecosystem

5.1.Graph Processing with GraphX

- 5.1.1. GraphX Introduction
- 5.1.2. Graph Algorithms

5.2.Cluster Computing with Spark Cluster Managers

- 5.2.1. Working with YARN and Mesos
- 5.2.2. Cluster Resource Management

6: Spark Development and Deployment

6.1.Spark Development Environment

- 6.1.1. Setting Up Spark Locally
- 6.1.2. Integrated Development Environments (IDEs)

6.2.Packaging and Deploying Spark Applications

- 6.2.1. Building JAR Files
- 6.2.2. Deploying on Spark Clusters

7: Spark Administration and Optimization

7.1.Cluster Monitoring and Management

- 7.1.1. Monitoring Spark Cluster
- 7.1.2. Scaling and Managing Resources

7.2.Performance Tuning

- 7.2.1. Identifying Bottlenecks
- 7.2.2. Optimization Techniques

8: Real-World Spark Use Cases

8.1.Real-World Applications

- 8.1.1. Spark in Industry
- 8.1.2. Case Studies

8.2.Best Practices

- 8.2.1. Data Pipeline Design
- 8.2.2. Scalability and Fault Tolerance

9: Final Project and Course Review

9.1.Project Proposal and Planning

- 9.1.1. Identifying a Real-World Problem
- 9.1.2. Designing a Spark Solution

9.2.Implementation and Presentation

- 9.2.1. Building and Deploying the Solution
- 9.2.2. Final Project Presentation