

PySpark Application Development and Performance Tuning

Duration : 7 Days

Content Coverage :

Module 1: Fundamentals of PySpark

- A Brief Primer on PySpark
- Brief Introduction to Spark
- Apache Spark Stack
- Spark Execution Process
- Newest Capabilities of PySpark
- Cloning GitHub Repository

Module 2: Resilient Distributed Datasets

- Resilient Distributed Datasets
- Creating RDDs
- Schema of an RDD
- Understanding Lazy Execution
- Introducing Transformations
 - .map(...)
 - .filter(...)
 - .flatMap(...)
 - .distinct(...)
 - .sample(...)
 - .join(...)
 - .repartition(...)
- **Projects**
 - Project 1: Count Data Project (ingestion of dataset, preprocessing, and exploratory data analysis using map, filter, flatMap, distinct, join, and repartition)

- Project 2: Weather Temperature Crunch (ingestion of dataset, preprocessing, and exploratory data analysis on streaming data using transformations)

Module 3: Resilient Distributed Datasets and Actions

- Introducing Actions
 - .collect(...)
 - .reduce(...) and .reduceByKey(...)
 - .count()
 - .foreach(...)
 - .aggregate(...) and .aggregateByKey(...)
 - .coalesce(...)
 - .combineByKey(...)
 - .histogram(...)
 - .sortBy(...)
 - Saving Data
 - Descriptive Statistics
- **Projects**
 - Project 3: 10 Tasks in Students/Professor University Datasets (ingestion, preprocessing, and applying RDD actions)
 - Project 4: 8 Tasks in Customer Data Datasets (ingestion, preprocessing, and applying RDD actions)
 - Project 5: Movie Ratings

Module 4: DataFrames and Transformations

- Creating DataFrames
- Specifying Schema of a DataFrame
- Interacting with DataFrames
- The .agg(...) Transformation
- The .sql(...) Transformation
- Creating Temporary Tables
- Joining Two DataFrames
- Performing Statistical Transformations

- The .distinct(...) Transformation
- **Projects**
 - Project 6: Company Mega Data (transformations, column logic, aggregation, and exploratory data analysis)
 - Project 7: University Data (end-to-end PySpark execution of insights delivery)

Module 5: Collaborative Filtering and Techniques

- Collaborative Filtering
- Utility Matrix
- Explicit and Implicit Ratings
- Expected Results
- Dataset Preparation
- Joining DataFrames
- Train and Test Data
- ALS Model
- Optimization: Hyperparameter Tuning and Cross-Validation
- Best Model Evaluation
- **Project**
 - Project 8: IMDB Rating Project (optimization logics with extensive PySpark techniques)

Module 6: Spark Streaming

- Introduction to Spark Streaming
- Spark Streaming with RDD
- Spark Streaming Context
- Reading Data in Spark Streaming
- Cluster Restart in Spark Streaming
- RDD Transformations in Spark Streaming
- DataFrame Display in Spark Streaming
- DataFrame Aggregation in Spark Streaming
- **Project**

- Project 9: Streaming Crunch Dataset (orchestration of a streaming pipeline project for live data ingestion)

Module 7: Spark ETL and Capstone Project

- Introduction to ETL
- ETL Pipeline
- Dataset Preparation
- Preprocessing, Extraction, Transformation
- Loading Data and Cleaning
- RDS Networking
- Downloading PostgreSQL
- Configuration and Execution
- **Project**
 - Project 10: Completion of Capstone Project (full end-to-end project for Streaming Crunch Dataset applying all PySpark concepts)

Module 8 : Spark Architecture

- RDDs
- DataFrames and Datasets
- Lazy Evaluation
- Pipelining

Module 9 : Data Sources and Formats

- Available Formats Overview
- Impact on Performance
- The Small Files Problem

Module 10 : Inferring Schemas

- The Cost of Inference
- Mitigating Tactics

Module 11 : Dealing With Skewed Data

- Recognizing Skew
- Mitigating Tactics

Module 12 : Catalyst and Tungsten Overview

- Catalyst Overview
- Tungsten Overview

Module 13: Mitigating Spark Shuffles

- Denormalization
- Broadcast Joins
- Map-Side Operations
- Sort Merge Joins

Module 14 : Partitioned and Bucketed Tables

- Partitioned Tables
- Bucketed Tables
- Impact on Performance

Module 15 : Improving Join Performance

- Skewed Joins
- Bucketed Joins
- Incremental Joins

Module 16 : PySpark Overhead and UDFs

- PySpark Overhead
- Scalar UDFs
- Vector UDFs using Apache Arrow
- Scala UDFs

Module 17 : Caching Data for Reuse

- Caching Options
- Impact on Performance
- Caching Pitfalls