PySpark Development

Duration: 5 days

Course Overview

This comprehensive PySpark course offers a hands-on, project-driven introduction to distributed data processing with Apache Spark. Beginning with core concepts, the curriculum guides learners through the Spark ecosystem, resilient distributed datasets (RDDs), and their transformations and actions. Participants gain practical skills through real-world projects such as weather analytics, university datasets, and customer profiling. The course then transitions to advanced techniques using DataFrames, SQL-like operations, collaborative filtering with ALS models, and hyperparameter tuning. Learners also explore Spark Streaming for real-time analytics and apply Spark ETL processes for efficient data extraction, transformation, and loading. By the end of the course, participants will be equipped to design scalable, end-to-end data processing pipelines using PySpark's modern capabilities.

Course Contents

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(…)
- Introducing Transformations – .filter(…)
- Introducing Transformations – .flatMap(…)
- Introducing Transformations –. distinct(…)
- Introducing Transformations – .sample(…)
- Introducing Transformations – .join(…)
- Introducing Transformations – .repartition(…)

- Project 1: Count Data Project (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying map, filter, faltmap, distinct, join and repartition)
- Project 2: Weather Temperature Crunch (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying map, filter, faltmap, distinct, join and repartition on instream data)

Module 3: Resilient Distributed Datasets and Actions

- Introducing Actions – .collect(…)
- Introducing Actions – .reduce(…) and .reduceByKey(…)
- Introducing Actions – .count()
- Introducing Actions – .foreach(…)
- Introducing Actions – .aggregate(…) and .aggregateByKey(…)
- Introducing Actions – .coalesce(…)
- Introducing Actions – .combineByKey(…)
- Introducing Actions – .histogram(…)
- Introducing Actions – .sortBy(…)
- Introducing Actions – Saving Data
- Introducing Actions – Descriptive Statistics
- Project 3: Students/Professor University Datasets (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying RDD actions.)
- Project 4: Customer Data Datasets (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying RDD actions through specified applicability)
- 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
- Project 6: CompanyMegaData (doing all the transformation logics, columunal logic and aggregation and exploratory data analysis)
- Project 7: University Data (end to end pyspark execution of insight delivery on University Data)

Module 5: Collaborative Filtering and Techniques

- Collaborative filtering
- Utility Matrix
- Explicit and Implict Rating
- Expected Results
- Dataset
- Joining Dataframe
- Train and Test Data
- ALS model
- Optimization Hyperparameter tuning and cross validation
- Best model and evaluate prediction
- Project 8: IMDB Rating project (Optimization logics focused on the project with extensive pyspark logic and clever techniques of manipulation)

Module 6: Spark Streaming

- Basics of Spark Streaming
- Creating Stream Processing Context
- Reading Real-Time Data
- Simple RDD and DataFrame Operations on Streams
- Viewing and Aggregating Stream Data

Module 7: Spark ETL

- Introduction to ETL
- Dataset
- Preprocessing, extraction, transformation
- Loading Data and cleaning