Apache Spark Programming with Databricks

Code:	ASPD
Length:	2 days
URL:	View Online

This course uses a case study driven approach to explore the fundamentals of Spark Programming with Databricks, including Spark architecture, the DataFrame API, Structured Streaming, and query optimization. You will start by visualizing and applying Spark architecture concepts in example scenarios. Then, you will explore and preprocess datasets by applying a variety of DataFrame transformations and actions. After ingesting data from various file formats, you will apply these preprocessing steps and write them to Delta tables. The case study then expands to stream from Delta in an analytics use case that demonstrates core Structured Streaming concepts. Lastly, you will explore the Spark UI and how query optimization, partitioning, and caching affect performance.

Skills Gained

- Define the major components of Spark architecture and execution hierarchy
- Describe how DataFrames are built, transformed, and evaluated in Spark
- Apply the DataFrame API to explore, preprocess, join, and ingest data in Spark
- Apply the Structured Streaming API to perform analytics on streaming data
- Navigate the Spark UI and describe how the catalyst optimizer, partitioning, and caching affect Spark's execution performance

Who Can Benefit

- SQL analyst
- Data engineer
- Data scientist
- Machine learning engineer
- Data architect

Prerequisites

- Familiarity with basic SQL concepts (select, filter, groupby, join, etc)
- Beginner programming experience with Python or Scala (syntax, conditions, loops, functions)

ExitCertified® Corporation and iMVP® are registered trademarks of ExitCertified ULC and ExitCertified Corporation and Tech Data Corporation, respectively
Copyright ©2020 Tech Data Corporation and ExitCertified ULC & ExitCertified Corporation.
All Rights Reserved.