Apache Spark Programming with Databricks Training

Duration: **2 Day**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Overview

This practical course provides an accessible starting point for professionals who want to learn Apache Spark programming using Databricks and develop skills for large-scale data processing.

The programme consists of **four modules of approximately four hours each**. Participants begin with Apache Spark's distributed architecture before progressing to data processing with the DataFrame API, ETL development, real-time processing with Structured Streaming, and the optimisation of Spark workloads on Databricks.

Throughout the course, **Apache Spark, Python, the Spark DataFrame API, Structured Streaming, Delta Lake, and Databricks** are brought together in practical data engineering scenarios. Core concepts are reinforced through demonstrations and hands-on labs, allowing participants to apply their knowledge to realistic data processing requirements.

The later stages of the programme explore production-oriented topics including Lakehouse architecture, Delta Lake, Spark performance optimisation, and the monitoring of Spark workloads within Databricks.

Prerequisites

Participants should have:

- Familiarity with Python programming
- An understanding of data types, lists, dictionaries, and variables
- Experience with functions, loops, and conditional statements
- Knowledge of exception handling
- Familiarity with accessing classes and using third-party Python libraries
- Basic SQL knowledge
- The ability to write queries using SELECT, WHERE, GROUP BY, ORDER BY, LIMIT, and JOIN

Who Should Attend

This course is particularly suitable for:

- Data engineers and data scientists who want to strengthen their Apache Spark programming skills
- Developers looking to use Apache Spark and Delta Lake within Databricks
- Professionals involved in large-scale data processing and real-time analytics

Apache Spark Programming with Databricks Training

Duration: **2 Day**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

What You Will Learn

Introduction to Apache Spark

Participants develop an understanding of Apache Spark's core concepts and distributed architecture, with a focus on its practical role in large-scale data processing.

The module introduces Spark programming frameworks and the DataFrame API while developing the skills required to read, write, and transform data through Python-based Spark workflows.

Developing Applications with Apache Spark

Participants learn how to build scalable data processing applications with Apache Spark. Using the Spark DataFrame API, they develop ETL pipelines and work with distributed data transformations.

The course covers grouping, aggregation, joins, set operations, and window functions together with complex data types such as arrays, maps, and structs. Participants also explore good practices for improving Spark application performance.

Stream Processing and Analysis with Apache Spark

This part of the course introduces the principles of processing real-time and continuously arriving data.

Participants develop applications with the **Spark Structured Streaming API** and explore techniques including stream aggregation and window analysis. The module develops the practical skills required to create scalable and fault-tolerant streaming applications for dynamic data environments.

Monitoring and Optimising Apache Spark Workloads on Databricks

This module explores scalable data workflows through **Lakehouse Architecture** and **Medallion Architecture** concepts.

Participants examine data governance, access control, and lineage with Unity Catalog while exploring how Delta Lake can support reliable, ACID-compliant data pipelines.

Spark optimisation methods including partitioning, caching, and query tuning are also considered alongside performance monitoring, troubleshooting, and good practices for efficient data engineering and analytics workloads.

Apache Spark Programming with Databricks Training

Duration: **2 Day**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Outline

Module 1: Introduction to Apache Spark

- Spark Runtime Architecture
- Exploring Apache Spark architecture in Databricks
- Introduction to Spark DataFrames and SQL
- Reading and writing data with DataFrames
- Distributed system programming fundamentals
- Basic ETL with the DataFrame API
- Flight data ETL using the DataFrame API
- Analysing transaction data with DataFrames

Module 2: Developing Applications with Apache Spark

- DataFrame API basics
- Demo (Optional): Basic ETL with the DataFrame API
- Grouping and aggregating data
- Demo: Grouping and aggregating data
- Lab: Grouping and aggregating e-commerce data
- Relational operations
- Demo: Data relational operations in Apache Spark
- Working with complex data
- Demo: Working with complex data types in Apache Spark
- Lab: Working with complex data types in e-commerce data

Module 3: Stream Processing and Analysis with Apache Spark

- Introduction to stream processing
- Spark Structured Streaming
- Demo: Introduction to Spark Structured Streaming
- Lab: Introduction to Spark Structured Streaming
- Advanced stream processing and analysis
- Demo: Window aggregation in Spark Structured Streaming
- Lab: Window aggregation in Spark Structured Streaming

Module 4: Monitoring and Optimising Apache Spark Workloads on Databricks

- Apache Spark and Databricks
- Using Apache Spark with Delta Lake
- Demo: Introduction to Delta Lake
- Lab: Introduction to Delta Lake
- Optimising Apache Spark
- Demo: Optimising Apache Spark
- Lab: Optimising Apache Spark