Faysal Aziz

Apache Spark

Training on Azure Databricks - 3-Day Course

FA-0101 · 3-Day

Software Development · Data & Analytics

Introduction

This course is designed to equip participants with a comprehensive understanding of Apache Spark, leveraging Azure Databricks as the primary platform. Aimed at SQL-experienced learners, the training will cover Python basics, Spark with PySpark, and advanced PySpark applications in machine learning. The course is structured to provide both theoretical insights and practical lab experiences.

Learning outcomes

- Gain a foundational understanding of Python programming.
- Develop proficiency in using Apache Spark with PySpark.
- Learn to implement machine learning algorithms with PySpark.
- Acquire hands-on experience with Azure Databricks.

Prerequisites

- Strong understanding of SQL.
- Basic familiarity with data processing and data analytics concepts.
- An Azure account for Databricks lab exercises (optional but recommended).

Training outline

1. Day 1: Introduction to Python

Objective: Introduce Python programming, focusing on aspects relevant to Spark and data processing.

Python Basics

1. Data Types and Variables
2. Control Structures (if, for, while)
3. Functions and Modules
4. Working with Collections (Lists, Tuples, Dictionaries)

Data Handling in Python

1. Reading and Writing Files
2. Basic Data Processing
3. Introduction to Pandas and DataFrames

Lab Session: Python Basics

1. Hands-on exercises in Azure Databricks using Python.

2. Day 2: Spark with PySpark

Objective: Dive into Apache Spark fundamentals and its implementation using PySpark.

Understanding Apache Spark

1. Spark Architecture and Components
2. RDDs (Resilient Distributed Datasets)
3. SparkSQL and DataFrames

PySpark Basics

1. Setting up PySpark in Databricks
2. RDD Operations
3. DataFrames and SQL Operations

Lab Session: PySpark Basics

1. Executing Spark Transformations and Actions
2. SQL Querying with DataFrames

3. Day 3: Advanced PySpark and Machine Learning

Objective: Explore advanced PySpark features and machine learning implementations.

Advanced PySpark Techniques

1. Aggregations and Joins
2. Spark Streaming Basics
3. Working with Multiple Data Sources

Machine Learning with PySpark

1. Introduction to MLlib
2. Building and Evaluating Machine Learning Models
3. ML Pipelines in Spark

Lab Session: Machine Learning with PySpark

1. Implementing a basic machine learning project
2. Model training and evaluation

© 2026 Faysal Aziz · Kuala Lumpur

https://faysalaziz.com/courses/apache-spark/