Faysal Aziz

Python for Spark

Using Colab and Azure Databricks

FA-0390 · 5 day

Software Development · DevOps, Cloud & Infrastructure · Data & Analytics

---

Introduction

Welcome to our in-depth 5 day course, meticulously crafted to balance foundational Python programming with the advanced capabilities of Apache Spark for big data processing. Through this course, learners will acquire a robust understanding of Python, essential for data manipulation and analysis, and then transition into mastering Spark within the context of big data and cluster computing using Azure Databricks.

Learning outcomes

By the end of this course, participants will be able to:

1. Write effective Python code, laying a strong foundation for data analysis and Spark programming.
2. Utilize Pandas for sophisticated data manipulation and Matplotlib for insightful data visualizations.
3. Understand and implement Object-Oriented Programming (OOP) concepts in Python to structure code efficiently.
4. Master the architecture of Spark, perform complex data transformations, and build robust data aggregation pipelines.
5. Develop PySpark applications, handling batch and streaming data, and employing advanced techniques for real-time processing in the context of big data.

Prerequisites

Participants are expected to have:

- Basic knowledge of programming concepts.
- Understanding in data analysis, data processing, and the principles of big data.
- Access to a computer with a stable internet connection for using Azure Databricks and other online resources.

Training outline

1. Python Programming for Data Analysis

Python Basics

1. Syntax and Semantics
2. Variables, Data Types, and Control Structures
3. Functions, Modules, and Python Standard Library

Data Analysis with Pandas

1. DataFrames and Series
2. Data Loading, Cleaning, and Preparation
3. Data Aggregation, Group Operations, and Time Series Analysis

Data Visualization with Matplotlib

1. Basic Plotting: Line, Bar, Histograms
2. Customizing Plots: Labels, Colors, and Styles
3. Advanced Visualizations: Interactive Plots and Geographical Data

Object-Oriented Programming in Python

1. Classes, Objects, Attributes, and Methods
2. Inheritance, Polymorphism, and Encapsulation
3. Special Methods, Decorators, and Exception Handling

2. Apache Spark for Big Data Processing

Introduction to Big Data and Spark

1. Big Data Characteristics and Challenges
2. Overview of the Spark Ecosystem and its Comparison with Hadoop
3. Understanding Spark's Place in the Big Data Landscape

Spark Basics and Architecture

1. SparkContext and SparkSession
2. Resilient Distributed Datasets (RDDs) and DataFrames
3. Spark Transformations, Actions, and Lazy Evaluation

Data Processing with Spark on Azure Databricks

1. Azure Databricks Setup and Cluster Management

2. Detailed Exploration of Spark Transformations and Actions
3. DataFrames Operations, Spark SQL for Complex Queries
4. Aggregation Functions and Window Functions for Advanced Data Analysis

Real-Time Data Processing with Spark Streaming

1. Fundamentals of Stream Processing and Comparison with Batch Processing
2. DStreams and Structured Streaming Concepts
3. Stateful Operations, Window Functions, and Handling Late Data

This course is designed to provide a blend of theoretical knowledge and practical application, ensuring that participants gain hands-on experience with the tools and technologies critical in the fields of data analysis and big data processing. Get ready to embark on this comprehensive learning journey that bridges Python programming with the expansive capabilities of Apache Spark!

© 2026 Faysal Aziz · Kuala Lumpur