

Analytics with Python and Azure Databricks

Python data workflows, visualisation and an introduction to distributed processing

FA-0579 · 3 days

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

Develop Python analytics workflows and explore PySpark batch and streaming examples in Azure Databricks.

Introduction

This three-day course is intended for participants with existing programming knowledge who want to apply Python to analytics. It combines a focused Python refresher with NumPy, pandas, visualisation and introductory distributed processing in Azure Databricks.

Hands-on exercises use prepared datasets and a configured learning environment. Object-oriented techniques, Flask integration and advanced Spark topics are covered at an introductory or demonstration level where appropriate. The course develops a small analytics workflow rather than promising mastery of Python, production web deployment or a full streaming platform.

Learning outcomes

The course teaches participants to:

- Use Python functions, files, exceptions and selected object-oriented patterns in data workflows.
- Manipulate arrays and tabular data with NumPy and pandas.
- Create and refine data visualisations using Matplotlib.
- Build a small Flask example for presenting an analytics result.
- Use PySpark DataFrames, SQL, aggregations and window operations in a prepared Databricks workspace.

- Compare batch processing and Structured Streaming and identify the role of checkpoints, data sources and model pipelines.

Prerequisites

- Good working knowledge of a programming language; prior Python experience is helpful.
- Familiarity with basic algebra and statistics.
- A suitable local Python environment or access to Google Colab.
- Access to an approved Azure Databricks learning workspace with suitable permissions and compute; resource use and charges should be agreed separately.
- A browser and internet connection. A local Flask example is sufficient; a remote production server is not required.

Training outline

1. Day 1 – Python Foundations for Analytics

Module 1 – Programming Refresher

Introduction to Python

- Syntax, data types, variables
- Control structures: loops, conditionals
- Math operations and functions

Functions in Python

- Defining and calling functions
- Scope and lifetime, arguments, and return values

Basic I/O Operations

- Reading from and writing to files
- Handling different file formats (CSV, JSON)

Error and Exception Handling

- Try, except, finally blocks
- Custom exceptions for robust error management

Module 2 – Reusable Python and NumPy

Object-Oriented Programming (OOP)

- Classes, objects, inheritance, polymorphism

- Encapsulation, abstraction, special methods

Data Handling with NumPy

- Array operations, indexing, slicing
- Basic linear algebra and statistical operations

Data Manipulation

- Recursion
- Data flattening
- Data Conversion

2. Day 2 – Tabular Analytics and Presentation

Module 3 – pandas Data Workflows

Data Analysis with pandas

- DataFrame and Series, data wrangling
- Advanced operations: merging, joining, and concatenating

Module 4 – Visualisation and a Flask Demonstration

Data Visualization with Matplotlib

- Create and refine selected plots; compare appropriate chart types.
- Improve labels, scales and presentation for communicating analytical findings.

Introduction to Flask for Web Applications

- Setting up a Flask environment
- Routing, templates, and form handling

Module 5 – Files, Debugging and Testing

Working with the OS and File System

- File system navigation, path manipulation
- Reading and writing files, directory management
- Debug a selected data-transformation function and write meaningful checks for expected outputs and error handling.
- Apply object-oriented organisation where it makes a small analytics workflow clearer rather than adding unnecessary complexity.

3. Day 3 – PySpark and Azure Databricks

Module 6 – Workspace and Batch Processing

Data Processing with PySpark

- Use a prepared Azure Databricks workspace; compare serverless and classic compute, permissions and relevant runtime constraints.
- DataFrame operations, Spark SQL

Debugging and Testing

- Running test scenarios
- Use supported notebook commands and inspect execution results.

Data Processing with Spark on Azure Databricks

- Transformations, joins and an introductory discussion of partitioning and clustering.
- Advanced data analysis: aggregation, window functions

Module 7 – Streaming Concepts and Advanced Demonstrations

Fundamentals of Stream Processing

- Stream vs. batch processing
- Processing modes, event-time concepts and checkpointing in a selected demonstration.
- Compare joins and data-layout concepts such as bucketing; verify support for the chosen runtime and data format.
- Structured Streaming basics using DataFrame operations.
- Discuss multiple data sources and integration constraints.

Module 8 – Machine-Learning and Integration Overview

Machine Learning with PySpark

- Introduction to MLlib
- Demonstrate a small model example and basic evaluation.
- Overview of Spark ML pipelines.
- Kafka integration concepts and a prepared demonstration where the environment supports it.
- Review a small analytics workflow and explain remaining engineering, data-quality and operational requirements.
- Discuss runtime and compute compatibility before applying demonstrated features to a different workspace.

