Comprehensive Data Engineering with Python and Azure Databricks

Day 01: Python Basics & Control Flow

Topics

- Python Basics
- Variables and Data Types
- Input/Output Operations
- Basic Operators
- Conditional Statements
- Loops (for, while)
- Control Flow Exercises

Day 02: Functions and Collections in Python

Topics

- Defining and Calling Functions
- Arguments and Return Values
- Lists, Tuples, Sets, Dictionaries
- Python Functions, Python Collection Objects
- Iteration and Comprehension Exercises

Day 03: NumPy and Pandas Series

Topics

- Introduction to NumPy Arrays
- Mathematical Operations
- Creating Pandas Series
- Indexing and Filtering
- Descriptive Statistics

Day 04: Pandas DataFrames and Visualization

Topics

- Creating and Modifying DataFrames
- Filtering, Grouping, Aggregating
- Handling Missing Data
- Plotting with Matplotlib
- Customizing Charts
- Statistical Plots with Seaborn

Day 05: Transition to PySpark

Topics

- Python Code Evaluation
- Writing Efficient Functions
- Setting Up PySpark
- RDDs vs DataFrames
- Running Simple Transformations
- Practical Comparison with Pandas

Day 06: SQL Fundamentals and Filtering

Topics

- SQL Syntax and Datatypes
- SELECT Statements
- WHERE Clause for Filtering
- LIKE, IN, BETWEEN
- Practice with Sample Tables
- Filtering and Pattern Matching

Day 07: SQL Sorting and Joins

Topics

- ORDER BY Clause
- Sorting Multiple Columns
- INNER JOIN, LEFT JOIN, RIGHT JOIN

- Joining Multiple Tables
- Join Conditions and Filtering Joins

Day 08: Aggregations and SQL Best Practices

Topics

- COUNT, SUM, AVG, MIN, MAX
- GROUP BY with Aggregates
- HAVING Clause
- NULL Handling in Aggregations
- Nested Aggregates
- SQL Coding Best Practices

Day 09: SQL + Python Integration and Practice

Topics

- Writing Advanced SQL Queries
- Multi-table Joins + Aggregations
- Query Optimization Hints
- Reading SQL Data in Python
- Executing SQL from Python

Day 10: Python Practice & Getting Started with Azure Databricks

Topics

- Python Recap with Advanced Practice
- Introduction to Azure Databricks
- Databricks Workspace Overview
- Cluster Creation and Management
- Uploading and Exploring Datasets

Day 11: Understanding Databricks Runtime

Overview of Databricks Runtime Architecture

- Runtime Versions and Compatibility
- Performance and Cost Considerations
- Selecting the Right Runtime (Standard, ML, Photon, etc.)

Deep Dive into Databricks Components

Introduction to Jobs and Tasks

- Job Clusters vs Interactive Clusters
- Workflows and Pipelines in Databricks
- Job Monitoring and Logs
- Task Dependencies and Alerting
- Using Tags for Resource Management

Scheduling and Automation

- Job Scheduler Basics
- Time-based Scheduling and Triggers
- Using Parameters and Task Values
- Notebook-based vs SQL-based Tasks
- Alerts and Notifications
- Hands-on Lab: Creating a Scheduled Workflow

Day 12: Hands-on with Workflow Creation and Management

Creating Multi-task Jobs in the UI

- Using Databricks Workflows with Notebooks
- Debugging Failed Runs
- Dependency Trees and Retry Logic
- Lab: Create a Realistic Multi-stage Workflow

Monitoring and Troubleshooting Jobs

- Viewing Job Histories
- Interpreting Logs and Metrics
- Error Resolution Techniques

- Lab: Analyze Failed Jobs and Fix Issues

Landmark Architecture on Azure

Overview of Current Reference Architectures

- Role of Databricks in the Modern Azure Stack
- Integration with Azure Data Factory, ADLS, Synapse
- Data Lakehouse Concepts
- Governance, Security, and Networking in Azure

Day 13

Introduction to Delta Lake

What is Delta Lake?

- Role of Delta in Azure Databricks
- Evolution: From Parquet to Delta
- Delta Lake Architecture Overview
- Key Benefits: Reliability, Performance, and Scalability

Batch vs. Incremental Data Processing

Definition and Use Cases

- Comparing Batch, Micro-batch, and Streaming
- When to use Incremental Data Loads
- Real-world examples in Azure Data Lake / Databricks
- Demo: Loading Batch and Incremental Data

ACID Transactions in Delta

- What are ACID properties?
- How Delta Lake ensures ACID compliance
- Transaction Logs and Checkpoints
- Demo: Handling Concurrent Writes and Read Isolation

Delta Lake Reliability Features

- Schema Enforcement and Evolution
- Data Versioning (Time Travel)
- File Compaction (Optimize Command)
- Handling Deletes, Updates, and Merges (MERGE INTO)
- Demo: Using Delta for Reliable Ingestion & Transformations

Implementing Delta Workflows and Medallion Architecture

- Use Case: ETL Pipeline with Bronze → Silver → Gold Layers
- Load Raw Data to Bronze
- Clean and Transform to Silver
- Aggregate and Publish to Gold
- ACID transaction behaviour during processing

Day 14: Performance Tuning and Best Practices

- OPTIMIZE, ZORDER, and Vacuum
- Partitioning Strategy
- File Sizes and Auto Compaction
- Delta vs. Parquet Performance Comparison
- Demo: Optimizing a Delta Table for Performance

Monitoring and Data Lineage

- Delta Lake Tables in Unity Catalog
- Auditing and Monitoring Delta Tables
- Reading Delta Logs and Transaction History
- Integration with Azure Purview for Lineage
- Troubleshooting common Delta failures

Day 15: Databricks Visualization and Dashboard

Topics

Introduction to Data Visualization in Databricks

- Importance of Visualization in Analytics
- Overview of Databricks SQL for Visualization
- Visual Layer vs. Query Layer
- Tour of Databricks SQL Editor and Dashboard Interface
- Best Practices for Designing Dashboards

Writing Queries for Visualizations

- Writing SQL Queries for Data Exploration
- Filtering and Aggregation for Charts
- Handling NULLs and Outliers
- Formatting Query Results for Charts
- Lab: Build a Base Query for Revenue by Region

Visualization Types and Use Cases

- Bar Charts, Line Charts, Area Charts
- Pie and Donut Charts
- Scatter Plots and Bubble Charts
- Choosing the Right Chart Type Based on Data
- Lab: Create Visualizations for Business KPIs

Creating Interactive Dashboards

- Adding and Arranging Visualizations
- Dashboard Filters and Global Parameters
- Time Filters and Refresh Controls
- Adding Text, Titles, and Descriptions
- Lab: Build a Complete Sales Dashboard with Filters

Evaluation and Task

- Practical Evaluation based on Content Covered
- Post Test