Master Data Engineering with Azure Synapse and PySpark

Course Description

This training is designed for data professionals who want to master Azure Synapse Analytics, PySpark, and modern data engineering techniques. The course provides a comprehensive understanding of Azure Synapse's architecture, integration with Azure Data Lake, and advanced data processing using PySpark. Participants will learn to build scalable ETL/ELT pipelines, optimize performance, and implement security and governance best practices. The training also covers emerging trends like Delta Lake, real-time data processing, and machine learning integration, ensuring participants are equipped to tackle real-world data engineering challenges.

Course Duration: 40 hours

Pre-requisites

To get the most out of this training, participants should have:

- Basic knowledge of cloud computing (preferably Azure).
- Familiarity with Python programming.
- Understanding of SQL and relational databases.
- Experience with data processing concepts (e.g., ETL, data pipelines).
- Basic exposure to Apache Spark or PySpark is helpful but not required.

Training Outline :

Module 1: Azure Synapse & Data Lake Concepts

- **Introduction to Azure Synapse Analytics**
    - Overview of Azure Synapse Architecture
    - Understanding Synapse Workspaces and Capabilities
    - Integration with Azure Data Lake Storage (ADLS)
    - Synapse Analytics Use Cases: Data Warehousing vs. Big Data Analytics

- **Azure Data Lake Storage (ADLS)**
    - Overview of Hierarchical File System
    - Storage Tiers: Hot, Cool, and Archive - Scenarios and Cost Considerations
    - Security Features: Access Control, Encryption, and Shared Access Signatures
    - Best Practices for Data Partitioning in ADLS
- **Linked Services and Datasets**
    - Configuring Linked Services for Data Sources
    - Defining and Managing Datasets in Synapse
    - Strategies for Data Integration with External Data Sources
    - Handling Data Schema Drift in Synapse Pipelines

Module 2: Python & PySpark Fundamentals

- **PySpark Basics**
    - Introduction to SparkContext and SparkSession
    - Configuring Spark Jobs for Synapse
    - Working with Spark Configurations and Environment Variables
    - Using PySpark in Synapse Notebooks
- **DataFrames**
    - Creating DataFrames from CSV, JSON, and Parquet Files
    - Transformations and Actions on DataFrames
    - Querying DataFrames with PySpark APIs
    - Handling Missing Data in DataFrames
- **RDDs (Resilient Distributed Datasets)**
    - When to Use RDDs vs. DataFrames
    - Transformations and Actions in RDDs
    - Optimizing RDD Operations
    - Converting Between RDDs and DataFrames
- **Spark SQL**
    - Writing SQL Queries in Spark

- Interacting with Hive Tables using Spark SQL
- Performance Optimization for Spark SQL Queries
- Advanced SQL Features in Spark: Window Functions and CTEs

Module 3: Data Engineering

- **Data Ingestion**
  - Loading Data from Azure Data Lake Storage
  - Extracting Data from Blob Storage and External Databases
  - Handling Real-time Data Ingestion with Event Hubs or Kafka
  - Strategies for Ingesting Semi-structured and Unstructured Data
- **Data Transformation**
  - Data Cleaning Techniques with PySpark
  - Aggregation and Grouping Datasets
  - Advanced Joins and Window Functions
  - Implementing Business Logic in PySpark Workflows
- **Data Pipelines**
  - Building ETL Pipelines using Synapse Pipelines
  - Creating and Managing Notebooks for ELT Pipelines
  - Automating Data Pipelines with Triggers and Alerts
  - Deploying and Monitoring Data Pipelines

Module 4: Optimization & Performance

- **Partitioning and Bucketing**
  - Partitioning Strategies for Large Datasets
  - Bucketing for Faster Querying
  - Choosing Partitioning Keys and Bucketing Columns
- **Caching and Broadcasting**
  - Cache Management in PySpark
  - Using Broadcast Variables for Optimized Joins

    - Identifying and Avoiding Skewed Data
- **Monitoring and Debugging**
    - Utilizing Azure Synapse Monitoring Tools
    - Exploring Spark UI for Debugging and Performance Insights
    - Troubleshooting Common PySpark and Synapse Issues
    - Best Practices for Logging and Metrics in Synapse Workflows

Module 5: Security & Governance

- **Access Control**
    - Implementing Role-based Access Control (RBAC)
    - Managing Access Policies for ADLS
    - Integrating Azure Active Directory for Authentication
- **Data Masking & Encryption**
    - Masking Sensitive Data in Synapse Workloads
    - Encrypting Data at Rest and in Transit
    - Using Azure Key Vault for Secrets Management
- **Auditing and Compliance**
    - Logging Access and Activities in Synapse
    - Ensuring Regulatory Compliance for Data Workflows
    - Implementing Data Governance with Azure Purview

Module 6: Advanced Topics

- **Delta Lake**
    - Introduction to Delta Lake
    - Implementing ACID Transactions in Synapse
    - Schema Evolution and Version Control
    - Time Travel Queries in Delta Lake
- **Machine Learning Integration**
    - Overview of Synapse ML Capabilities

    - Building and Deploying Machine Learning Models
    - Feature Engineering with PySpark and Synapse
    - Deploying Predictive Models in Synapse Workflows
- **Real-time Data Processing**
    - Streaming Data Pipelines with Spark Structured Streaming
    - Analyzing Streaming Data with Synapse Analytics
    - Best Practices for Streaming Data Reliability
    - Integrating Real-time Data with Event Hubs and Synapse Pipelines