Azure Data Engineer

Course Content

Duration: 3 months

Fees: 25,000/-

Sub Course : 01

databricks & pyspark

Duration: 1.5 months

Fees: 12,000/-

Sub Course : 02

ADF (Azure Data Factory)

Duration: 1.5 months

Fees: 8,000/-

Sub Course : 03

PySpark

Duration: 5 weeks

Fees: 8,000/-

1) Introduction to Azure Data Engineering

a) Fundamentals of Data Engineering
b) Azure Data Services Overview
c) Role of Azure Data Engineer in Real-world Projects

2) Introduction to Azure

a) Azure Portal Walkthrough

1. Creating Azure Free Account
2. What is Subscription?
2. What is a Resource Group?
3. What is a Resource?

b) Overview of Azure Resources / Services

1. Data Factory
2. Azure Data bricks
3. BLOB Storage, Data Lake Storage Gen1 and Gen2
4. Azure SQL Server, SQL Database
5. Key Vault, Secret Scope
6. Logic Apps

c) Types of Data

1. Structured Data
2. Semi-structured Data
3. Unstructured Data

d) Azure Storage & Data Ingestion

1. Azure Blob Storage & Azure Data Lake Gen2
2. Batch vs Real-time Data ingestion
3. Hands-on: Create a Data Lake & upload raw datasets for further processing

Spark

- Spark Architectures
- Spark RDD
- Cach & persist
- Common Transformations and Actions
- Spark DataFrame, joins & Aggregate Functions
- Shared Variables & Broadcast Variables
- Accumulator
- Fault Tolerance
- spark performance optimization
- Partitioning bucketing caching, AQE

3) Python Basics

a) Variables
b) DataTypes
c) Operators
d) Collections (Arrays) – List, Tuple, Sets and Dictionary
e) Conditional Statements – If else and Nested If else and elif
f) User Defined Functions – Defining, Calling, Types of Functions, Arguments
g) String Manipulation – Basic Operations, Slicing & Functions and Methods

4) Basics of SQL

a) DQL Commands (select)
b) DDL commands (create, alter, drop , truncate)
c) DML Commands (insert , update, delete, merge)
d) Joins
e) Window functions
f) Aggregate functions
g) CTE expression
h) Set Operators

5) Azure Data Factory

1) Azure Data Factory
   1. What is Azure Data Factory?
   2. Azure Data Factory Architecture
   3. Azure Data Factory Portal UI
   4. Top-level concepts
      A. Pipelines
      B. Activities
      C. Linked services
      D. Datasets
      E. Triggers
      F. Data Flows
      G. Integration Runtimes

2) Pipeline
   1. What is a Pipeline?
   2. Create a new pipeline
   3. Organize pipelines into folders
   4. Debug pipeline
   5. Publish pipeline
   6. Parameters / Pipeline Parameters

3) Linked Service
   1. What is a Linked Service?
   2. Create a Linked Service for –
      A. BLOB
      B. SQL Database
      C. SQL Server
      D. Data Lake Storage Gen1
      E. Azure Data Lake Storage Gen2 etc
   3. Parameters / Linked Service Parameterization

4) DataSets
   1. What is a Data Set?
   2. Create a Data Set for –
      A. Avro, Binary, CSV, Excel, JSON, ORC, Parquet, XML in BLOB/ADLS Gen1/ADLS Gen2.
      B. Table in SQL Database, SQL Server, Oracle Database etc
   3. Parameters / Data Set Parameterization

5) Activities
   1. Wait
   2. Variables
      A. Create a variable
      B. Set variable
      C. Append variable
   3. Copy Data
      A. Source, Sink, Mapping & Settings
   4. Copy file(s) from one BLOB Container to another Container
      A. One file from a folder
      B. All files from a folder
      C. All files and folders recursively from a folder
   5. Copy data / file from BLOB to SQL Database / ADLS Gen2
      A. As CSV, TSV, Parquet, Avro, ORC etc.
   6. Databricks Notebook
   7. Lookup, Stored Procedure
   8. Get Metadata, Delete
   9. Execute Pipeline
   10. Validation, Fail
   11. Iteration & Conditionals
      A. Filter, ForEach, If Condition, Switch & Until

6) What is a Trigger?
   1. Types
      A. Schedule, Tumbling window & Storage Events
   2. Triggers with Parameters

7) Integration Runtime (IR)
   1. Azure AutoResolveIntegrationRuntime
   2. Azure Managed Virtual Network
   3. Self-Hosted
   4. Linked Self-Hosted

8. Git configuration
9. ARM Template
10. Azure Devops Repos
11. Global parameters
12. Monitoring ADF Jobs
13. Alerts

14) Data Flows

    1. What is Data Flow?
    2. Mapping Data Flow
    3. Data Flow Debug
    4. Transformations
        A. Filter, Aggregate, Join
        B. Conditional Split, Derived Column
        C. Exists, Union, Lookup, Sort,
        D. GroupBy, Pivot, Unpivot, Flatten etc.
        E. Flatten, parase, stringify
        F. Filter sort, alterrow,asset
        G. flowlet
    5. Validate Schema, Schema Drift
    6. Remove Duplicate Rows using Mapping Data Flows in Azure Data Factory

6) Azure Databricks

1) Introduction to Spark
    1. Spark Architecture
2) Introduction To Databricks
    1. Databricks Architecture
    2. Working with Databricks workspace, notebook, FileSystem (DBFS)
    3. DBFS commands - mkdirs , cp , mv , head, put, rm , rmdir
    4. How to process, archive & handle multiple files in DBFS
3) RDD Programming
4) Operations on RDD
    a) Transformations: Narrow & Wide
    b). Actions
5) Broadcast variables
6) Databricks- Handle multiple file formats
    1. CSV Data
    2. JSON Data
    3. parquet files
    4. Excel files
7) Databricks Cluster Management
    1. Creating and configuring clusters
    2. Managing, Starting, Terminating & Delete Clusters
8) Types of Clusters
    1. All pupose clusters, Job cluster
    3. Clusters Mode
        a. Standard
        b. High Concurrency
        c. Autoscalling
9) Databricks – Batch Processing
    1. Historical & Incremental Data load
    2. Date Transformations
    3. Aggregations
    4. Join Operations
    5. window functions
    6. union operations

10) Databricks Integration with
    1. Blob strorage storage
    2. Azure Datalake storage gen2
    3. Azure SQL Database
    4. Azure Keyvault
11) Databricks – Lakehouse (Delta Lake)
    1. Difference between Data lake and Delta Lake
    2. How to create delta table
    3. How to DML operations in Delta Table
    4. Merge statements
    5. Handling SCD Type1 and Type2
12) Delta Lake: Medallion Architecture
    1. Implement the Bronze Layer (Raw Data)
    2. Implement the Silver Layer (Cleansed & Transformed Data)
    3. Implement the Gold Layer (Curated, Business-Ready Data)
13) Workflows in Databricks
    1. Introduction to workflows
    2. Create, run and manage Databricks jobs
    3. Schedule Databricks jobs
    4. Monitor Databricks Jobs

7) PySpark

- How to read & write different file formats (csv, parquet, json)
- Creating dataframe with schema (Schema Enforcement and Evolution)
- Common dataframe operations: select, withColumn, add, modify, rename, sort, filter & drop columns.
- Dataframe functions (show, display, count, limit, collect)
- Identify and Remove duplicates
- Combine dataframes
- Window functions (first, last, max, min, row_number, groupBy & aggregation)
- Identify & fill null's
- Date & timestamp functions
- How to convert data types?
- Working with data formats (StructType, Map Type, Array, JSON & AVRO) and functions to handle these formats.
- Set operators (except, exceptAll, subtract, union, union all, unionByName)
- String functions (split, substring, locate, translate)
- Partioning and bucketing
- Joins (Inner, left, right, cross join, broadcast join)
- Create temp views
- Managed and External tables
- Time travel, Vacuum & ZORDER
- SCD type 1 & 2 (Full and Incremental Load)

8) PROJECT

- END TO END Project Implementation