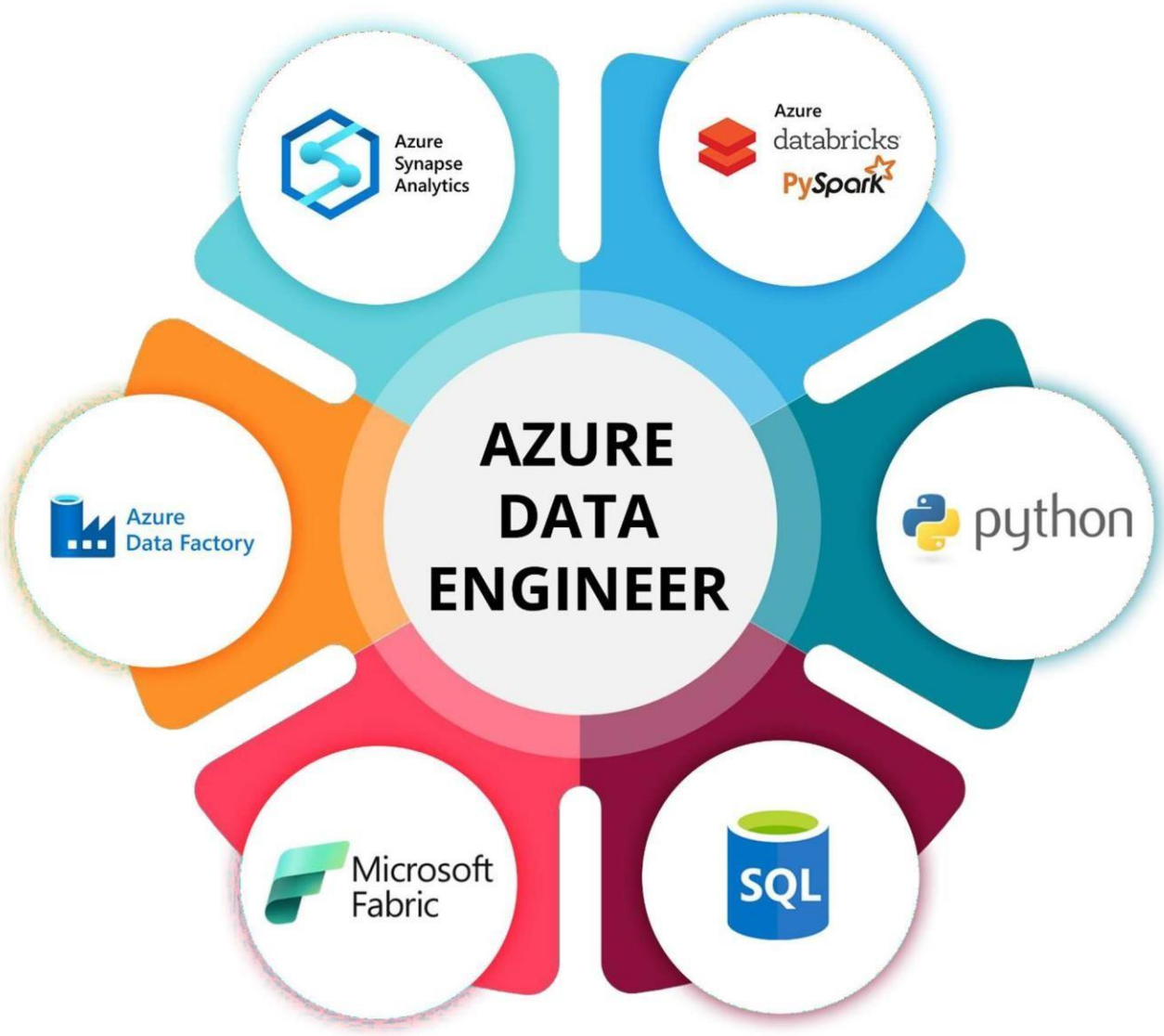
Azure Data Engineering Course

(Azure Data Factory + Azure Databricks + Azure Synapse Analytics + Microsoft Fabric Course)



Overview of Cloud

1) **Basics of Cloud computing**
    1. What is Cloud?
    2. Types of Cloud deployment models
        A. Private Cloud
        B. Public Cloud
        C. Hybrid Cloud
    3. Types of Cloud Services
        A. IaaS – Infrastructure as a Service
        B. PaaS – Platform as a Service
        C. SaaS – Software as a Service

2) **Cloud computing Platforms / Vendors**
    1. Azure
    2. AWS – Amazon Web Services
    3. GCP – Google Cloud Platform etc

3) **Introduction to Azure**

4) **Azure Portal Walkthrough**
    1. What is Subscription?
    2. What is a Resource Group?
    3. What is a Resource?

5) **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
    6. Function App
    7. Logic Apps

6) **Introduction to BigData**
    1. What is Data?
    2. What is BigData?
    3. Data Sources of Big Data?
    4. Characteristics of BigData

5. Variety, Velocity, Volume, Veracity, Value
6. Types of Data
    A. Structured Data
    B. Semi-structured Data
    C. Unstructured Data

7) Python Basics

1. Variables
2. DataTypes
3. Operators
4. Collections
5. Functions
6. Packagea and Modules

8) Basics of SQL

1. DQL Commands (select)
2. DDL commands (create, alter, drop , truncate)
3. DML Commands (insert , update, delete, merge)
4. Joins
5. Window functions
6. Aggregate functions

9) Over View of Azure Storage Accounts

1. Types of strorage accounts
2. Blob storage
3. Access Tiers
4. Data Replication Policies
5. Azure Data Lake Storage Gen2

10) Azure Key Vault

1. Introduction to Key Vault
2. Keys, Secrets, Certificates
3. Creating and configuring Key Vault

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. General
    B. Source
    C. Sink
    D. Mapping
    E. Settings
    F. User Properties
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. Azure Function
8. Lookup, Stored Procedure
9. Get Metadata, Delete
10. Execute Pipeline
11. Validation, Fail
12. Iteration & Conditionals
    A. Filter
    B. ForEach
    C. If Condition
    D. Switch
    E. Until

6) **What is a Trigger?**

1. Types
    A. Schedule
    B. Tumbling window
    C. 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) **Source control**
   1. Git configuration
   2. ARM Template
      A. Export / Import
   3. Azure Devops Repos

9) **Global parameters**

10) **Credentials**

11) **Monitoring ADF Jobs**

12) **Alerts**

13) **Send Failure Notifications using Logic Apps**

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

15) **Azure Devops**
   1. Repos

16) **SDLC**

17) **Agile Methodology**

18) **ADF Interview Questions**

19) **ADF Resume Preparation**

20) **End TO End ADF Project**

21) **ADF Exercises**

1. Create variables using set variable activity
2. How to use if condition using if condition activity
3. Iterating files using for loop activity
4. Creating linked services, Data sets
5. Copy activity - blob to blob
6. Copy activity - blob to azure SQL
7. Copy activity - pattern matching files copy
8. Copy activity - copy the filtered file formats
9. Copy activity - copy multiple files from blob to another blob
10. Copy activity - Delete source files after copy activity
11. Copy activity - using parameterized data sets
12. Copy activity - convert one file format to another file format
13. Copy activity - add additional columns to the source columns
14. Copy activity - filter files and copy from one blob to another
15. Delete the files from blob with more than 100KB
16. How to use getmetdata activity
17. Bulk copy tables and files
18. How to integrate keyvault in ADF
19. How to set up integration run time
20. Copy data from on premises to azure cloud
21. How to use databricks activity activity and pass paraemeters to it
22. How to use scheduling trigger
23. How to use tumbling window trigger
24. How to use event based trigger
25. How to use with Activity
26. How to use Until Activity
27. Dataflows - select the rows
28. Dataflows - Filter the rows
29. Dataflows - join Transformations
30. Dataflows - union Transformations

31. Dataflows - look up Transformations
32. Dataflows - window functions transformations
33. Dataflows - pivot, unpivot transformations
34. Dataflows - Alter rows transformations
35. Dataflows - Removing Duplicates transformations
36. How to pass parameters to the pipeline
37. How to create alerts and rules
38. How to set global parameters
39. How to import and export ARM templates
40. How to integrate ADF with Devops
41. How to use Azure devops Repos
42. How to send mail notifications using logic apps
43. How to monitor the pipelines
44. How to debug the pipelines
45. How to schedule pipeline using triggers
46. How to create trigger dependency
47. How to one pipeline in another pipeline

Azure Databricks

1) Introduction to BigData

1. What is Data?
2. What is Database?
3. What is BigData?
4. What are the challenges of BigData?
5. Why Traditional Databases Doesn't handle Bigdata

2) Introduction to Hadoop

1. What is Hadoop?
2. How Hadoop overcome bigdata challenges
3. Hadoop Architecture
4. Hadoop Daemons
5. HDFS
6. YARN
7. MapReduce

3) Introduction to Spark

1. Spark Architecture
2. Spark internals
3. Spark RDD
4. Spark DataFrame
5. Spark Streaming

4) Introduction To Databricks

1. What is Databricks?
2. Databricks Architecture
3. Working in Databricks workspace
4. Workign with Databricks notebook

5) Working with Databricsks FileSystem - DBFS

1. What is DBFS?
2. DBFS commands - mkdirs , cp , mv , head, put, rm , rmdir
3. How to handle multiple files in DBFS
4. How to process the files in DBFS
5. How to archive the files in DBFS

6) Databricks -Sparck Core

1. RDD Programming
2. Operations on RDD
3. Transformations- Narrow
4. Transformations -Wide
5. Actions
6. Loading Data and Saving Data
7. Key Value Pair RDD
8. Broadcast variables

7) Databricks - Spark-SQL- DataFrames

1. Creating Data Frames
2. DataFrames internal execution
3. Transformations using DataFrame API
4. Actions using DataFrame API
5. User-defined functions in Spark SQL

8) Databricks- Handle multiple file formats

1. CSV Data
2. JSON Data
3. parquet files
4. Excel files
5. ORC file format

9) Databricks utilities

1. credentials utility
2. FilSystem utility
3. Notebook utility
4. secrets utility
5. widgets utility

10) Databricks Cluster Management

1. Creating and configuring clusters
2. Managing Clusters
3. Displaying clusters
4. Starting a cluster
5. Terminating a cluster
6. Delete a cluster
7. Cluster Information
8. Cluster logs
9. Types of Clusters
10. All pupose clusters
11. Job cluster
12. Clusters Mode
13. Standard
14. High Concurrency
15. Autoscalling
16. Databricks runtime versions

11) Databricks – Batch Processing

1. Historical Data load
2. Incremental Data load
3. Date Transformations
4. Aggregations
5. Join Operations

6. window functions
7. union operations

12) Introduction to Azure

1. Azure Portal Walkthrough
2. What is Subscription?
3. What is a Resource Group?
4. What is a Resource?
5. Overview of Azure Resources / Services
6. Azure Data bricks
7. BLOB Storage, Data Lake Storage Gen2
8. Azure SQL Server, SQL Database
9. Key Vault

13) Databricks Integration with

1. Blob strorage storage
2. Azure Datalake storage gen2
3. Azure SQL Database
4. Synapse
5. Azure Keyvault

14) Databricks – Streaming API

1. What is streaming?
2. Process streaming using Pyspark API
3. Handling bad records
4. Stream data into Gen2lake
5. Load the data into Tables

15) Databricks – Lakehouse (Delta Lake)

1. Difference between Data lake and Delta Lake
2. Introduction to Deltalake
3. Features of DeltaLake
4. How to create delta table
5. How to DML operations in Delta Table
6. Merge statements
7. Handling SCD Type1 and Type2
8. Handling Data Deduplication in delta tables
9. Handling streaming Data in Delta lake

16) **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)

17) **Workflows in Databricks**

1. Introduction to workflows
2. Create, run and manage Databricks jobs
3. Schedule Databricks jobs
4. Monitor Databricks Jobs

18) **Azure DevOps – Repos**

1. What are DevOps Repos
2. Integrate databricks notebooks with Repos
3. Commit, Sync notebooks to and from Repos

19) **SDLC and Agile methodology**

20) **End to End Data Migration Project from On Premises to Cloud.**

21) **Interview Questions**

22) **Mock Interviews**

Azure Synapse

1) **Introduction & Overview**

1. Azure Synapse Analytics Overview
2. Anzure Synapse Analytics Architecture
3. Create Azure Free Account for Synapse

2) **Overview of pools in Synapse Analytics**

1. Dedicated SQL pools
2. Serverless SQL pool
3. Apache Spark pools
4. Data Explorer pools

3) Using Azure Synapse Analytics to Query Data Lake

1. Creating Azure Synapse Analytics Workspace
2. Uploading Sample Data into Data Lake Storage
3. Exploring Azure Synapse Workspace and Studio
4. Querying a Data Lake Store using serverless SQL pools in Azure Synapse Analytics
5. Creating a View for CSV Data with a Serverless SQL Pool

4) Azure Storage Account Integration with Azure Synapse

1. Copy multiple files from blob to blob using wildcard file options
2. Copy multuple folders from blob to blob using dataset parameters
3. Get File Names from Folder Dynamically and copy latest file from folder

5) Azure Synapse Triggers

1. Schedule Trigger in Azure Synapse
2. Event Based Trigger in Azure Synapse

6) Azure SQL Database integration with Azure Synapse

1. Azure SQL Databases - Introduction ___Relational databases in Azure
2. Copy data from SQL Database to ADLS Gen2 using table, query and stored procedure
3. Overwrite and Append Modes in Copy Activity in Azure Synapse
4. Use Foreach loop activity to copy multiple Tables- Step by Step Explanation

7) Incremental Load to Azure Synapse in Azure Synapse

1. Incremental Load or Delta load from SQL to blob Storage in Azure Synapse
2. Multi Table Incremental Load or Delta load from SQL to to Azure Synapse
3. Incrementally copy new and changed files based on Last Modified Date

8) Logging and Notification and keyvault integration ___Azure Logic Apps

1. Log Pipeline Executions to SQL Table using Azure Synapse
2. Custom Email Notifications and keyvault integration with Linked Service
3. Send Error notification with logic app
4. Use Foreach loop activity to copy multiple Tables with pipeline logs logic and notifications

9) Deep dive into Copy Activity in Azure Synapse

1. Load data from on premise sql server to Azure Synapse
2. Copy Data from sql server to to Azure Synapse with polybase & Bulk Insert
3. Copy Data from on-premise File System to Azure Synapse
4. Loop through REST API copy data 2 ADLS Gen2 Linked Service Parameters

10) Data Flows Introduction

1. Azure Data Flows Introduction
2. Setup Integration Runtime for Data Flows
3. Basics of SQL Joins for Azure Data FlowsServerless SQL Pool Demo
4. Joins in Azure DataFlowsDedicated SQL Pool Demo
5. Difference Between Join vs.Lookup Transformation& Merge Functionality Spark Pool Demo
6. Dataflows – select the rows
7. Dataflows – Filter the rows
8. Dataflows – join Transformations
9. Dataflows – union Transformations
10. Dataflows – look up Transformations
11. Dataflows – window functions transformations
12. Dataflows – pivot, unpivot transformations
13. Dataflows – Alter rows transformations
14. Dataflows – Removing Duplicates transformations

11) Spark Pool Introduction in Azure Synapse

1. Spark Introduction and components
2. Spark Architecture
3. Create notbook and notebook option and create notebook in different langauges
4. MSSparkUtils for file system
5. MSSparkUtils for for creating notebook parameters
6. Magic commands and calling one synapse notebook from another and returning output of synapse notebook
7. Configure keyvault in azure synapse notebook
8. Different ways to connect to ADLSGen2 from synapse notebook
9. Different ways to connect to Blob from synapse notebook
10. Different ways to connect to Azure SQL Database from synapse notebook
11. Different ways to connect to on premise SQL Server from synapse notebook

12. Optimization while Reading and writing CSV files from Azure Synapse
13. Reading and writing parquet files from Azure Synapse
14. Reading and writing JSON files from Azure Synapse
15. Reading and writing avro and orc files from Azure Synapse
16. Reading and writing EXCEL files from Azure Synapse
17. Different ways to create RDD in synapse notebook
18. Different ways to create dataframes in synapse notebook
19. When to use repartition and coalesce
20. Joins in Synapse Notebook
21. Broadcast Joins in Synapse Notebook and configuration of spark for optimization
22. what is catalyst optimiser and skewness issue in spark
23. Optimization techniques in pyspark
24. Implementing SCD1 in Synapse Notebook
25. Implementing SCD2 in Synapse Notebook
26. Executing synapse notebooks from synapse pipleines with input and output parameters

12) **Project:** End to End DataMigration using Synapse Analytics

Microsoft Fabric Course

1) **Introduction to Big Data**

1. What is Data?
2. What is a Database?
3. What is Big Data?
4. Challenges of Big Data
5. Why Traditional Databases Cannot Handle Big Data?

2) **Introduction to Microsoft Fabric**

1. What is Microsoft Fabric?
2. How to Enable Fabric in Your Organization?
3. Fabric Workspace Structure
4. Big Data Analytics with Microsoft Fabric
5. Microsoft Fabric – A Unified Platform
6. Advantages of Microsoft Fabric

3) Introduction to Fabric Components

1. One-Lake
2. Real-time Intelligence
3. Data Factory
4. Fabric Data Science
5. Fabric Data Engineering
6. Data Activator
