Azure Data Engineer + Databricks Developer

1. *DataBricks*
2. *Spark*
3. *Pyspark*
4. *Spark SQL*
5. *Delta Lake*
6. *Azure Data Factory*
7. *Azure Synapse DW (Dedicated SQL POOL)*
8. *Azure ADF & Databricks Projects*

Azure Databricks Concepts.

1)**Azure Databricks Introduction**
A.Databricks Architecture
B.Databricks Components overview
C.Benefits for data engineers and data scientists

2) **Azure Databricks concepts**
   A. Workspace – Creation and managing workspace.
   B. Notebook – creating notebooks, calling and managing different notebooks.
   C. Library - installing libraries, managing libraries

3) **Data Management**
   A. Databricks File System. - DBFS commands copy and manage files using DBFS.
   B. Database - Creating database, tables and managing databases and tables.
   C. Table - Creating Tables, dropping tables, loading data ..
   D. Metastore - managing metadata and delta tables creation, managing delta tables.

4) **Computation Management**
   A. Cluster -- Creating Clusters , managing clusters
   B. Pool - creating pools and using pools for Auto scaling.

C. Databricks RunTime - understanding and using Databricks runtimes based on requirement.
D. Jobs - creating jobs from notebooks and assigning types of clusters for jobs.
E. Workload - monitoring jobs and managing loads.
F. Execution Context – understanding context.

5) **Databricks Advanced topics.**
A. Databricks Workflows
B. Calling one notebook into another notebook.
C. Creating global variables (widgets) and using into Azure ADF pipeline.
D. How to implement parallelism in notebooks execution.
E. Mounting azure blob storage and data lake storage accounts.
F. Integrating source code (notebooks) with GitHub
G. Calling DataBricks notebooks into Azure Data factory.
H. Databricks Clusters logs monitoring flow.

SPARK Concepts

1) **Introduction to Spark - Getting started**

A. What is Spark and what is its purpose?
B. Components of the Spark unified stack
C. Resilient Distributed Dataset (RDD)
D. Downloading and installing Spark standalone
E. Scala and Python overview
F. Launching and using Spark's Scala and Python shell ©

2) **Resilient Distributed Dataset and DataFrames**

A. Understand how to create parallelized collections and external datasets
B. Work with Resilient Distributed Dataset (RDD) operations
C. Utilize shared variables and key-value pairs

3) **Spark application programming**

A. Understand the purpose and usage of the Spark Context
B. Initialize Spark with the various programming languages
C. Describe and run some Spark examples
D. Pass functions to Spark
E. Create and run a Spark standalone application
F. Submit applications to the cluster

4) **Introduction to Spark libraries**

   A. Understand and use the various Spark libraries

5) **Spark configuration, monitoring and tuning**

   A. Understand components of the Spark cluster
   B. Configure Spark to modify the Spark properties, environmental variables, or logging properties
   C. Monitor Spark using the web UIs, metrics, and external instrumentation ,Understand performance tuning considerations

PySpark Content

- **Introduction To Pyspark**
  1) What is SparkSession
  2) How to create spark session
  3) What is SparkContext
  4) How to create SparkContext
  5) What is SQLContext
- How to Use Jupyter Notebooks & Databricks notebooks for Python Development.
- Install and configure PySpark in Local System for development.
- Introduction to Big Data and Apache Spark
- Apache Spark Framework & Execution Process.
- **Introduction To RDDs**
  1) Different Ways to Create RDD's in Pyspark.
  2) RDD Transformations
  3) RDD Actions
  4) RDD Cache & Persist
- **Introduction to DataFrame.**
  1) Different Ways to Create Data Frame's in Pyspark.
  2) Dataframe Transformations
  3) Dataframe Actions
  4) Dataframe Cache & Persist
- **Different types of Big Data File systems.**
  1) Difference between Row store format and column store format.
  2) Avro File
  3) Parquet file
  4) ORC File
- **Reading and Writing Different Types of Files using Dataframe.**

1) Csv files
2) Json files
3) Xml files
4) Excel files
5) Complex Json files
6) Avro files
7) Parquet files
8) Orc files

- **Need for Spark SQL**
- What is Spark SQL
  1) SQL Table Creation
  2) SQL Join Types
  3) SQL Nested Queries
  4) SQL DML Operations
  5) SQL Merge Scripts
  6) SQL SCD Type 2 implementation

DataFrame Transformations

1. DataFrames Metadata Transformations
   1. Adding new column
   2. Renaming a column
   3. removing a column
   4. changing data type
   5. renaming all columns
2. Dataframe Displaying functions
   1. df.show()
   2. display(df)
   3. df.collect()
3. Dataframe Data Validations
   1. Removing duplicate rows
   2. Removing null rows
   3. converting null rows to actual rows
   4. customer filters
4. Specifying Schema of a DataFrame
5. Interacting with DataFrames
6. The .agg(…) Transformation
7. The .sql(…) Transformation
8. Creating Temporary Tables
9. Joining Two DataFrames
10. Performing Statistical Transformations
11. The .distinct(…) Transformation
12. Data Processing with Spark DataFrames
13. Filtering Data
14. Aggregating Data
15. Selecting Data
16. Transforming Data
17. Presenting Data
18. Sorting DataFrames
19. Saving DataFrames
20. Pitfalls of UDFs
21. Repartitioning Data
22. Performance Tuning

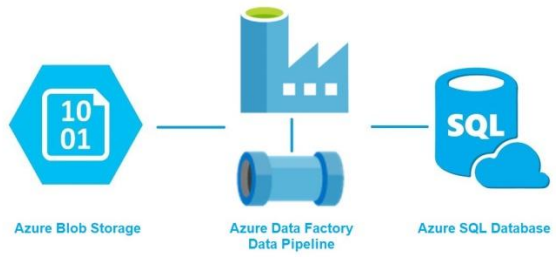
Pyspark Project with execution.

1) End to End Pyspark Projects implementation
2) Executing Pyspark Project in Databricks
3) Executing PySpark project in Azure ADF.

DELTA LAKE

1) **Delta Lake usage in Databricks.**
   A. Delta Lake Architecture
   B. Delta Lake Storage Understanding
   C. Delta lake table creation and API options
   D. Delta Lake DML Operations usage.
   E. Delta Lake partitions
   F. Delta Lake Schema Enforcement
   G. Delta Lake Schema Evolution
   H. Delta Lake Versions
   I. Delta Lake Time Travel
   J. Delta Lake Vaccum
   K. Delta Lake Merge (SCD Type 1 and SCD Type2)

Azure Data Engineer



Overview of the Microsoft Azure Platform
   A. Introduction to Azure
   B. Basics of Cloud computing
   C. Azure Infrastructure
   D. Walkthrough of Azure Portal
   E. Overview of Azure Services

Azure Data Architecture
   A. Traditional RDBMS workloads.
   B. Data Warehousing Approach
   C. Big data architectures.
   D. Transferring data to and from Azure

Azure Storage options
   A. Blob Storage
   B. ADLS Gen1 & Gen2

C. RDBMS
D. Hadoop
E. NoSQL
F. Disk

Blob Storage

A. Azure Blob Resources
B. Types of Blobs in Azure
C. Azure storage account data objects
D. Azure storage account types and Options
E. Replications in distribution
F. Secure access to an application's data
G. Azure Import/Export service
H. Storage Explorer
I. Practical section on Blob Storage

Azure Data Factory

A. Azure Data Factory Architecture
B. Creating ADF Resource and Use in azure cloud
C. Pipeline Creation and Usage Options
D. Copy Data Tool in ADF Portal, Use
E. Linked Service Creation in ADF
F. Dataset Creation, Connection Reuse
G. Staging Dataset with Azure Storage
H. ADF Pipeline Deployments
I. Pipeline Orchestration using Triggers
J. ADF Transformations and other tools integration.
K. Processing different type's files using ADF.
L. Integration Runtime
M. Monitoring ADF Jobs
N. Manage IR's and Linked Services.

Azure Data Lake Gen1 & Gen2

A. Explore the Azure Data Lake enterprise-class security features.
B. Understand storage account keys.
C. Understand shared access signatures.
D. Understand transport-level encryption with HTTPS.
E. Understand Advanced Threat Protection.
F. Control network access.
G. Differences between Gen1 & Gen2

Azure Synapse SQL DW (Dedicated SQL POOL)

A. Azure Synapse DW (Dedicated SQL POOL)?
B. Synapse DW Architecture.
C. Creating Internal table with default distribution
D. Creating external table in synapse dw
E. Loading data from databricks to azure synase dw
F. Loading data from adls gen2 to azure synapse dw
G. What is dedicated sql pool
H. data warehouse unit overview
I. Distributed table with example
J. Hash distribution with example
K. Round robin distribution with example

L. Replicate distribution with example
M. What are the types of indexes with examples
N. Clustered Index with example
O. Non-Clustered index with example
P. Clustered Column Store Index with example
Q. Heap Index with example

SPARK SQL:

1) **Introduction to Spark SQL.**
2) **Spark SQL Create database**
3) **Drop databases**
4) **Create internal table**
5) **Create external table**
6) **Create partitioned table**
7) **Create partitioned with bucketing table**
8) **SPARK DML insert, update, delete and merge operations**
9) **SPARK SQL DRL Select queries with different clauses**
10) **Spark SQL MERGE With SCD Type 1 and SCD Type 2**
11) **Spark SQL WHERE Clause, Group By Clause and Having Clauses**
12) **Spark SQL Order by, Sort By clauses**
13) **Spark SQL join types, Window , Pivot , Limit and Like**
14) **Spark SQL Grouping Sets, Rollup and Cube**
15) **Spark SQL Cultured By and Distributed By**
16) **Spark SQL Case, With and Take sample**