Azure Databricks

Prerequisites of the Training

* Microsoft Azure Fundamentals
* Azure Storage Account
* Basics of SQL Server [Tables/Queries]

**Reference Links**

* <https://learn.microsoft.com/en-us/training/modules/explore-relational-data-offerings/>
* <https://learn.microsoft.com/en-us/training/modules/explore-provision-deploy-relational-database-offerings-azure/>
* https://learn.microsoft.com/en-us/training/modules/explore-provision-deploy-non-relational-data-services-azure/

**Day 1:**

Configuring Azure Databricks

• Configure Databricks cluster

• Understanding Databricks File System (DBFS)

• Mount Azure Data Lake Gen2 to DBFS

Extracting and Processing Data

• Extract data from connected sources

• Creating Spark tables

• Run PySpark / Scala code to apply transformations

• Monitor Spark jobs

Building DataFrames

• Reading from multiple file formats (csv, json etc.)

• Creating and applying schemas

• Analyzing and cleaning data

• Apply common transformations (selection, aggregation, filtration, joins etc.)

• Understanding partitions

• Using coalesce and repartition functions

Serverless SQL Pools

• Understand features of Serverless SQL Pool (like OPENROWSET, auto schema inference etc.)

• Configure Serverless SQL Pool

**Day 2:**

• Setup external tables and views

• Using SQL and PySpark together

• Analyze data lake contents (Parquet & CSV parsers) using SQL queries (OPENROWSET)

Delta Lake

• What is Delta Lake?

• Understanding Delta Architecture ‘

• Delta Lake scenarios/use cases

• Table creation and metadata

• How delta lake works internally?

• Copy/clone operations

• Performance comparison with Parquet format

• Schema validation and evolution

• Table delete, update and merge operations

• Creating table constraints

• Time Travel

• Concurrency control