Azure Databricks

Duration: 7 Days **Audience: Freshers**
Methodology: Hands-On + Task and Assignments **Domain: Retail**

- This training will help participants to get started with working data on Azure cloud platform.
- Participants will be able to capture data from any source, structured and un-structured data, Flat files, Relational Tables, Stream Data, IOT devises, storage account etc.
- Participants will be able to perform ETL operations, automate them and monitor them on cloud
- Perform In depth practical hands-on task in projects on Azure Databricks and Azure Data Factory

Agenda

Day1 – Getting Started with Azure

Day 2 – Azure Data Platform – RDBMS (SQL Server and PostGre), NoSQL – Cosmo DB

Day 3 – Data Lake – Azure Storage Account – Blobs, Containers, File Share etc

Day 4 – Azure Data Factory (Pipeline, ETL Process)

Day 5 – Azure Databricks (Clusters, PySPark, Jobs) – ETL Process and Analysis

Day 6 – Azure Data Bricks and ADF Integration, Monitoring, troubleshooting, Maintenance

Day 7 – Azure Automation with ARM Templates

Describe Cloud Concepts for Data

- Identify the benefits and considerations of using cloud services
- identify the benefits of cloud computing, such as High Availability, Scalability, Elasticity,

  Agility, and Disaster Recovery
- identify the differences between Capital Expenditure (CapEx) and Operational
- Expenditure (OpEx)
- describe the consumption-based model
- Describe the differences between categories of cloud services
- describe the shared responsibility model
- describe Infrastructure-as-a-Service (IaaS),
- describe Platform-as-a-Service (PaaS)
- describe serverless computing
- describe Software-as-a-Service (SaaS)
- identify a service type based on a use case
- Describe the differences between types of cloud computing
- define cloud computing
- describe Public cloud
- describe Private cloud

- describe Hybrid cloud
- compare and contrast the three types of cloud computing.

Describe core solutions and management tools on Azure

- describe the benefits and usage of Internet of Things (IoT) Hub, IoT Central
- describe the benefits and usage of Azure Synapse Analytics, HDInsight, and Azure Databricks

describe the benefits and usage of serverless computing solutions that include Azure

- **Functions and Logic Apps for data**
- **describe the benefits and usage of Azure DevOps, GitHub, GitHub Actions, and Azure**
- **DevTest Labs**

Develop for Azure storage

- **Understanding Data Lake, Delta Lake for data processing and storage**
- **Develop solutions that use storage tables**
- **design and implement policies for tables**
- **query table storage by using code**
- **implement partitioning schemes**

Develop solutions that use Cosmos DB storage

- **create, read, update, and delete data by using appropriate APIs**
- **implement partitioning schemes**
- **set the appropriate consistency level for operations**

Develop solutions that use a relational database

- **provision and configure relational databases**
- **configure elastic pools for Azure SQL Database**
- **create, read, update, and delete data tables by using code**
- **provision and configure Azure SQL Database serverless instances**
- **provision and configure Azure SQL and Azure PostgreSQL Hyperscale instances**

Develop solutions that use blob storage

- **move items in Blob storage between storage accounts or containers**
- **set and retrieve properties and metadata**
- **implement blob leasing**
- **implement data archiving and retention**
- **implement Geo Zone Redundant Storage**

Implement access control for data

- **implement CBAC (Claims-Based Access Control) authorization**
- **implement RBAC (Role-Based Access Control) authorization**

- create shared access signatures

Implement secure data solutions

- encrypt and decrypt data at rest and in transit
- create, read, update, and delete keys, secrets, and certificates by using the KeyVault API

Design and Develop Data Processing

Ingest and transform data

- transform data by using Apache Spark
- transform data by using Transact-SQL
- transform data by using Data Factory
- transform data by using Azure Synapse Pipelines
- transform data by using Stream Analytics
- cleanse data
- split data
- shred JSON
- encode and decode data
- configure error handling for the transformation
- normalize and denormalize values
- transform data by using Scala
- perform data exploratory analysis

Design and develop a batch processing solution

- develop batch processing solutions by using Data Factory, Data Lake, Spark, Azure
- Synapse Pipelines, PolyBase, and Azure Databricks
- create data pipelines
- design and implement incremental data loads
- design and develop slowly changing dimensions
- handle security and compliance requirements
- scale resources
- configure the batch size
- design and create tests for data pipelines
- integrate Jupyter/Python notebooks into a data pipeline
- handle duplicate data
- handle missing data
- handle late-arriving data
- upsert data
- regress to a previous state
- design and configure exception handling
- configure batch retention
- design a batch processing solution

debug Spark jobs by using the Spark UI

Design and develop a stream processing solution

develop a stream processing solution by using Stream Analytics, Azure Databricks, and

- Azure Event Hubs
- process data by using Spark structured streaming
- monitor for performance and functional regressions
- design and create windowed aggregates
- handle schema drift
- process time series data
- process across partitions
- process within one partition
- configure checkpoints/watermarking during processing
- scale resources
- design and create tests for data pipelines
- optimize pipelines for analytical or transactional purposes
- handle interruptions
- design and configure exception handling
- upsert data
- replay archived stream data
- design a stream processing solution

Manage batches and pipelines

- trigger batches
- handle failed batch loads
- validate batch loads
- manage data pipelines in Data Factory/Synapse Pipelines
- schedule data pipelines in Data Factory/Synapse Pipelines
- implement version control for pipeline artifacts
- manage Spark jobs in a pipeline

Monitor and Optimize Data Storage and Data Processing

Monitor data storage and data processing

- implement logging used by Azure Monitor
- configure monitoring services
- measure performance of data movement
- monitor and update statistics about data across a system
- monitor data pipeline performance
- measure query performance
- monitor cluster performance

- understand custom logging options

schedule and monitor pipeline tests

- interpret Azure Monitor metrics and logs
- interpret a Spark directed acyclic graph (DAG)

Optimize and troubleshoot data storage and data processing

- compact small files
- rewrite user-defined functions (UDFs)
- handle skew in data
- handle data spill
- tune shuffle partitions
- find shuffling in a pipeline
- optimize resource management
- tune queries by using indexers
- tune queries by using cache
- optimize pipelines for analytical or transactional purposes
- optimize pipeline for descriptive versus analytical workloads
- troubleshoot a failed spark job
- troubleshoot a failed pipeline run