Implementing an Azure Data Engineer

Introduction to Azure

- Introduction to Azure Cloud
- What is difference between Azure Cloud and On-Premises
- What is Subscriptions and Resource Groups
- Different offerings of Cloud IaaS, PaaS and SaaS

Introduction to Storage

- Azure Storage
  - Azure Blob
  - Table
  - Message
  - Queue
- Azure Data Lake Store Gen I & Gen II
  - What is Data Lake
  - Data Lake vs. Hadoop
  - Blob Storage vs. Data Lake
  - Hierarchical Namespace
  - Ingestion through different tools i.e.; Azure Data Explorer, AzCopy, Azure CLI, Powershell

Azure Data Factory

- Introduction to Azure Data Factory
- Creation of Linked Services, Datasets, Pipelines
- Creation of Integration Runtime and different types
- Slowly Changing Dimensions
- Design and implement a Type 1 slowly changing dimension with mapping data flows
- Debug data factory pipelines
- Activities
  - Copy
  - Data flow
  - Stored Procedure
  - Lookup
  - ForEach
  - Get Metadata
  - Filter Activity
  - Spark
  - U-SQL

- o Databricks Notebooks
- o Web
- o If Condition
- o Delete
- Sending mails through Logic Apps
- Few more Activities ......
- Dataset and Pipeline Parameterization
- Monitor -- Azure and Visually
- Setup Alerts from Azure Data Factory

Realize Integrated Analytical Solutions with Azure Synapse Analytics

- Introduction
- What is Azure Synapse Analytics
- How Azure Synapse Analytics works
- When to use Azure Synapse Analytics
- Create Azure Synapse Analytics workspace
- Exercise - Create and manage Azure Synapse Analytics workspace
- Describe Azure Synapse Analytics SQL
- Explain Apache Spark in Azure Synapse Analytics
- Exercise - Create pools in Azure Synapse Analytics
- Orchestrate data integration with Azure Synapse pipelines
- Exercise-Identifying Azure Synapse pipeline components
- Visualize your analytics with Power BI
- Understand hybrid transactional analytical processing with Azure Synapse Link
- Use Azure Synapse Studio
- Understand the Azure Synapse Analytical processes
- Explore the Data hub, Develop hub, Integrate hub
- Explore the Monitor hub, Manage hub
- Describe a modern data warehouse
- Define a modern data warehouse architecture
- Exercise - Identify modern data warehouse architecture components
- Design ingestion patterns for a modern data warehouse
- Understand data storage for a modern data warehouse
- Understand file formats and structure for a modern data warehouse
- Prepare and transform data with Azure Synapse Analytics
- Serve data for analysis with Azure Synapse Analytics

Azure Synapse Analytics

Introduction
- Why Warehouse in cloud
- Traditional vs. Modern Warehouse architecture
- What is Synapse Analytics Service

- Create Dedicated SQL Pool and Spark Pool
- Create Azure Synapse Analytics Studio Workspace
- Analyze Data using Dedicated SQL Pool and Spark Pool
- Analyze Data using Apache Spark Notebook
- Analyze Data using Serverless SQL Pool
- Azure Synapse Benefits

Azure Databricks

- Spark Basics
- Why Spark is difficult? Why Databricks Evolved?
- Why Databricks in Cloud? Introduction to Azure Databricks
- Demo
- Provision Databricks, Clusters and workbook
- Mount Data Lake to Databricks DBFS
- Explore, Analyze, Clean, Transform and Load Data in Databricks
- Azure Databricks Clusters
- Azure Databricks other Important Components
- Databricks - Monitoring
- How to create Cluster
- How to work with Databricks File System
- How to create notebooks and Integrate with ADF
- How to import and export the Notebooks
- How to connect to blob, SQL DB from Databricks
- How to read data files from Azure Blob and Azure Data Lake Store
  - Using Scala, R, Python, Spark SQL Language
- Creating Data Frames
- Converting Data Frames into Temporary Table or Temporary View
- Incremental and Full Load with Azure SQL Data Warehouse
- Understand the architecture of Azure Databricks spark cluster
- Understand the architecture of spark job
- Read data in CSV format
- Read data in JSON format
- Read data in Parquet format
- Read data stored in tables and views
- Write data
- Describe a DataFrame
- Use common DataFrame methods
- Use the display function
- Exercise: Distinct articles
- Describe the difference between eager and lazy execution
- Describe the fundamentals of how the Catalyst Optimizer works
- Define and identify actions and transformations

- Describe the column class
- Work with column expressions
- Perform date and time manipulation
- Use aggregate functions
- Exercise: Deduplication of data
- Describe the Azure Databricks platform architecture
- Perform data protection
- Describe Azure key vault and Databricks security scopes
- Secure access with Azure IAM and authentication
- Describe security
- Exercise: Access Azure Storage with key vault-backed secrets
- Describe the open source Delta Lake
- Exercise: Work with basic Delta Lake functionality
- Describe how Azure Databricks manages Delta Lake
- Exercise: Use the Delta Lake Time Machine and perform optimization
- Describe Azure Databricks structured streaming
- Perform stream processing using structured streaming
- Process data from spark structured streaming
- Describe bronze, silver, and gold architecture
- Perform batch and stream processing
- Schedule Databricks jobs in a data factory pipeline
- Pass parameters into and out of Databricks jobs in data factory
- Integrate with Azure Synapse Analytics
- Understand workspace administration best practices
- List security best practices
- Describe tools and integration best practices
- Explain Databricks runtime best practices
- Understand cluster best practices

Introduction to Azure SQL Database

- Introduction to Azure SQL Database
- Why choosing SQL Server in Azure
- Azure IaaS vs. PaaS database offerings
- IaaS vs. Managed Instance
- SQL Server PaaS deployment options
- Demo - Azure Single Database
- Purchasing models and Service Tier
- Elastic Database Pool
  - Introduction
  - Azure Elastic Database
  - Demo - Azure Elastic Database
- Managed Instance Database
  - Introduction
  - Azure Managed Instance Database
  - Difference between on-premises and managed instance

- Migration options for Managed Instance
- Service tiers for Managed Instance
- Demo - Managed Instance

- **Azure Database Security**
  - Introduction
  - Azure Database and Managed Instance Security options
  - Encrypting Data at Rest and Motion
  - High Availability vs. Disaster Recovery
  - Azure SQL Database High Availability and Disaster Recovery options

Azure SQL Database Scaling

Just Enough Python for Apache Spark

- Getting Started With Python
- Variables and Data Types
- Conditionals and Loops
- Methods, Functions, Packages
- Collections and Classes
- Pandas