MICROSOFT AZURE DATA ENGINEERING

ROAD MAP

CAREER PATH

COURSE CONTENT

LEARNING OUTCOME

ROAD MAP

WHAT IS DATA ENGINEERING?

Data Engineers are necessary to ensure that data is collected, stored, and made accessible for analysis. They are the architects behind the scenes, responsible for building, maintaining, and organizing the infrastructure that enables organizations to leverage data effectively. In essence, data engineers bridge the gap between raw data and actionable insights, making them a crucial asset in data-driven decision-making.

HOW TO BECOME A DATA ENGINEER?

Programming Skills: Learn programming languages commonly used in data engineering, such as Python, Pyspark, or Scala. Proficiency in SQL is crucial for database Management.

Database Management: Gain expertise in relational databases (e.g., PostgreSQL, MySQL) and NoSQL databases (e.g., MongoDB, Cassandra).

ETL Tools: Learn ETL tools such as Apache Nifi or Apache Airflow, which help automate data pipeline processes.

Cloud Platforms: Understand cloud computing platforms like AWS, Azure, or Google Cloud, as organizations often use these for data storage and processing.

Version Control: Use tools like Git to manage code and collaborate effectively.

Data Warehousing: Explore data warehousing solutions.

Good to Have Big Data Technologies: Familiarize yourself with big data technologies like Hadoop, Spark, and Apache Kafka, as they are integral to data engineering.

CAREER PATH

DATA ENGINEER CAREER PATH

Junior Data Engineer: Entry-level position focusing on learning the basics of data engineering. Knowledge in SQL querying and development. Basic knowledge in Visualization tools (PowerBI, Tableau etc.) Knowledge of Data Management and Data Warehousing concepts

Data Engineer: Strong knowledge in Python, SQL, and data visualization/exploration tools, Maintaining ETL processes, Responsible for building and maintaining data pipelines.

Senior Data Engineer: Involves more complex pipeline architecture and mentoring junior engineers.

Data Engineering Manager: Overseeing a team of data engineers and managing larger-scale projects.

Solution Architect: Designing an organization's overall data infrastructure and architecture.

COURSE CONTENT

Introduction to Azure **Week-1**

- 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
- Creation of Virtual Machine

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

Introduction to Azure SQL Database **Week-2**

- 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
- Azure Database vs. Azure Data Warehouse
- 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
    - RTO vs. RPO
    - Azure SQL Database High Availability and Disaster Recovery options
    - Azure SQL Database Scaling
- Installation of SQL Server 2016 and above in Virtual Machine
- Creation of External Table or PolyBase in On-Premise SQL Server
    - Creation of Master Key
    - Creation of Database Scoped Credential
    - Creation of External Data Source
    - Creation of External File Format
    - Creation of External Table
- Creation of External Table or PolyBase in Azure SQL Data Warehouse
    - Creation of Master Key
    - Creation of Database Scoped Credential
    - Creation of External Data Source
    - Creation of External File Format
    - Creation of External Table
- Different Distribution or Shredding Patterns
    - ROUND ROBIN
    - HASH
    - REPLICATION
- Cross Query Databases in Azure SQL Database
    - Creation of Master Key
    - Creation of Database Scoped Credential
    - Creation of External Data Source
    - Creation of External Table
- Creation of Elastic Pools in Azure SQL Server between Databases

Data Warehouse Internals and Architecture **Week-3**

- Introduction
- Azure Synapse MPP Architecture
- Storage and Sharding patterns
- Data Distribution and Distributing Keys
- Data Types and Table Types
- Partitioning
- Data Warehouse Concepts
- Dimensions and Facts
- Types of Dimensions and Facts
- Different types of Schemas in Data Warehouse
- Relationship types in Data Warehouse
- Best Practices for Fact and Dimension tables
- Demo - Analyze Data distribution before migration to Azure Synapse

Azure Data Factory **Week-4 & 5**

- 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
- Understand the Azure SSIS Integration Runtime
- Set-up Azure SSIS Integration Runtime
- Run SSIS Package in Azure Data Factory
- Migrate SSIS Packages to Azure Data Factory
- Integrate SQL Server Integration Services Packages within Azure Data Factory
- Activities
    - Copy, Data flow, Stored Procedure, Lookup, ForEach, Get Metadata, Filter Activity
    - Spark
    - U-SQL
    - Databricks Notebooks
    - Web
    - If Condition, Delete
- Data Flows
    - Derived Column, Join, Filter, Exists, Conditional split, Lookup, Exists, Select
    - Aggregate, Rank, Sort, Alter Row
- Dynamic Queries in ADF
- 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 **Week-6**

- 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 Event Hub, IoT Hub and Azure Stream Analytics **Week-7**

- Introduction to Azure Event Hub, IoT Hub and Stream Analytics
- Azure Stream Analytics Job
- Azure Stream Analytics Components
- Azure Stream Analytics Job
- Batching Streaming using Azure Event Hub
- Real Time Streaming using Azure IoT Hub
- Types of Window Functions
    - Tumbling Window
    - Hoping Window
    - Sliding Window
    - Session Window

Azure Databricks **Week-8 & 9**

- Spark Basics
- Why Spark is difficult? Why Databricks Evolved?
- Why Databricks in Cloud? Introduction to Azure Databricks
- 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 Spark 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
- Spark Functions ,Key-Value Pairs , Aggregate Functions
- Working with Aggregate Functions
- Joins in Spark
- 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
- Work with Time Windows
- Process data from Event Hubs with 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

Azure Delta Lake **Week-10**

- Overview of Azure Delta Lake
- Data Lakehouse Architecture
- Read and Write to Delta Lake
- Updates and Deletes on Delta Lake
- Merge/Upsert to Delta Lake
- History, Time Travel, Vaccum
- Delta Lake Transaction Log
- Convert Parquet to Delta

LEARNING OUTCOME

This course typically focuses on knowledge needed to design, implement, and manage data solutions on the Microsoft Azure platform. Below are common learning outcomes from the successful completion of the course.

- Understanding Azure Data Architecture
- Data Ingestion and Integration
- Data Storage Solutions
- Data Transformation and Processing
- Automating Data Pipelines
- Data Visualization and Analytics
- Data Governance and Lineage
- Real-Time Analytics
- Data Modeling and Schema Design
- Managing and Orchestrating Data Workflows
- Cost Management and Optimization