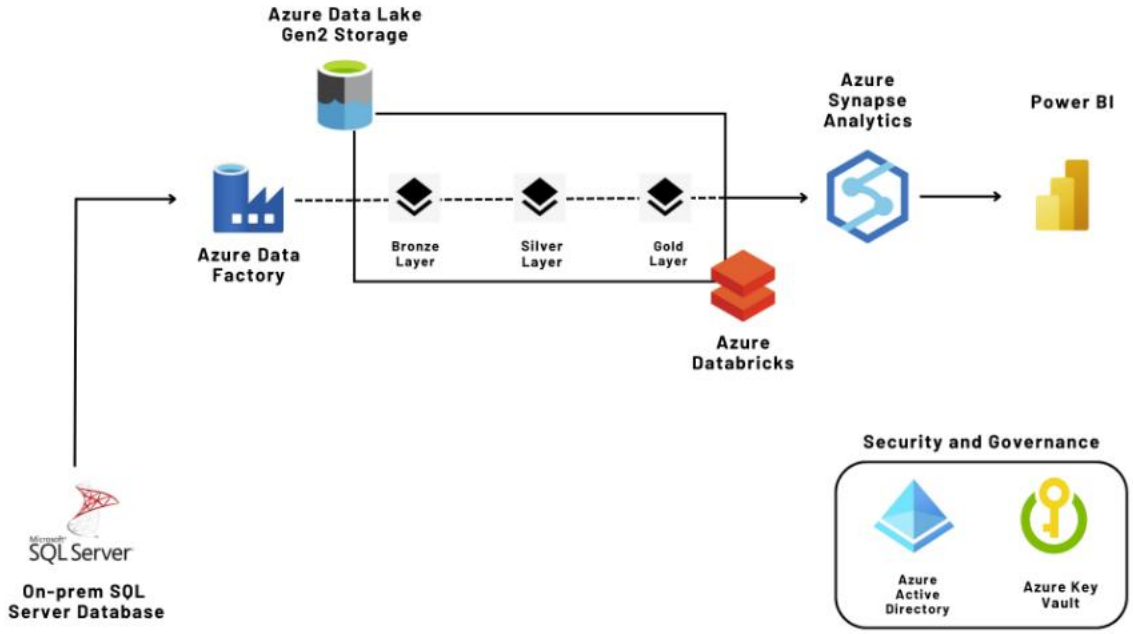
CAPSTONE PROJECT 3

Instructions:

Update the project on GitHub repo
Allowed time (A week for phase-1 and 2weeks for phase-2)

In this project we are going to create an end-to-end data pipeline that moves the data from an On-Prem SQL database and following steps such as Data Ingestion, Data Transformation, Data Loading, Data Governance and finally Data Reporting using Microsoft Power BI.

Data Architecture



Requirements:

- Create free account on Azure
- SSMS & MySQL Server
- Download dataset AdventureWorks DW 2017 from the link https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver17&tabs=ssms
- Free account on Azure Databricks
- SQL Agent job & schedule.

Tasks (PHASE 1)

1. Restore adventureWorks2017 on MySQL server as the DW (onprem) based on the above architecture
2. Confirm all the tables and records was populated correctly with SELECT statement on some of the tables
3. Develop the data dictionary for five (5) of the tables with columns such as field Name, description, DB_Field Name, Foreign Key, Foreign key table, Data type, and comments. Store the dictionary in excel format.
4. Using stored procedure, develop the following on the DW
   Total sales by year, Top N customers by Sales, Sales by Product Category for a given year, Monthly sales trend for a specific Product, create a stored procedure that returns customers who has purchased more than once and sales by Territory with Date Range.
5. Implement agent job for the stored procedure sales trend and Top N customers by Sales to run every morning by 7am using Agent jobs in SSMS.
6. Load the Data (Stored Procedures) into power BI and provide the visuals.

Tasks (PHASE 2)

Create an end-to-end data pipeline that moves the data from an On-Prem SQL database and following steps such as Data Ingestion, Data Transformation, Data Loading, Data Governance and finally Data Reporting using Microsoft Power BI.

The procedures include:

1. Data Ingestion - Create a data ingestion pipeline to extract data from on-premises SQL Server Database using Azure Data Factory

2. Data Storage - Create a centralized repository to store data from SQL Server Database into Azure Data Lake Gen 2 storage.

3. Data Transformation - Create ETL job to extract the data, do simple transformations and load the clean data using Azure Databricks.

4. Data Governance - Create Azure key vaults and Active Directory to monitor and govern the whole project using Azure roles.

5. Data Analytics - Create data integration pipeline with Power BI using Azure Synapse Analytics to create powerful visualizations.

Microsoft Azure Resources:

Azure Data Lake Gen2 Storage
Azure Data Factory
Azure Databricks
Azure Synapse Analytics
Azure Key vault (Unity Catalog)
Azure Active Directory

Implementation steps:

Step 1 - Create a data integration link service to connect SQL server with Azure Data Factory Since, the database is in an on-premises SQL Server, Microsoft Azure needs a way to detect the stored data and able to interact with it.

Step 2 - Mounting the database to perform Data Transformation using Azure Databricks To do any kind of transformations, we need some compute power to do perform them. In Azure databricks, the 'compute' option gives us the capability to fire spark clusters and perform data transformations.

Step 3 - Connecting Azure Data Factory with Azure Databricks to create data pipelines for data transformations. The data is ever increasing, and we need a way to automate the data ingestion and transformation processes as much as possible.

Step 4 - Load the data to Azure Synapse Analytics for further big data analytics Azure Synapse Analytics is built on top of Azure Data Factory, so many options can be found in the Synapse Analytics. In Azure Synapse Analytics, we can create databases which is not available in ADF.

Step 5 - Connecting PowerBI to Azure Synapse Analytics to create interactive visualizations The PowerBI desktop will be used, and the data source will be marked as Azure Synapse Analytics SQL views.