Project - Data Warehouse and BI Dashboard for Retail

Group : ALX_AIS4_G1e

Track : AI & Data Science

Microsoft Data Engineer

Instructor: Ahmed Azab

INTRAST skillsdynamix

Team Members

Karim Ahmed Fouad Mohamed Abouyoussef

Mohamed Abdel Karim Abdel Aziz

Khaled Tarek Abdel Aziz

Ahmed Farag Allah Abdel Aziz

Ahmed Mohamed Gaber

Karim Mohamed Abdel Aziz

Mahmoud Mohsen Mahmoud

Project Objective:

- The project is a comprehensive **data engineering solution** focused on migrating a local **SQL Server** database to the cloud using **Azure services**. It includes end-to-end processes, from data ingestion, transformation, and storage, to security, visualization, and automation.

Technologies and Tools Used:

1. **Azure Data Factory:**
   - Acts as an **ETL tool** (Extract, Transform, Load) to connect with the on-premises SQL Server database, extract the data, and move it to Azure.
2. **Azure Data Lake Gen 2:**
   - The storage solution for the project, divided into different **layers**:
     - **Bronze Layer:** Stores the raw data as it is from the SQL Server, without any transformation.
     - **Silver Layer:** Stores transformed data after basic modifications, such as renaming columns and adjusting data types.
     - **Gold Layer:** The final stage where the data is fully cleaned and ready for analysis.
3. **Azure Databricks:**
   - Used for **big data analytics** and transforming the raw data. It helps in writing and running the transformation logic using **PySpark**, **SQL**, or **Python**.
   - Supports the **Lakehouse architecture** by creating different data layers (bronze, silver, gold).
4. **Azure Synapse Analytics:**
   - Provides the environment for querying and analyzing the transformed data.
   - Helps set up a new database and tables in the cloud that mirror the original SQL Server schema.
   - Data from the **gold layer** is loaded into **Synapse Analytics** for analysis.
5. **Azure Active Directory:**
   - Manages user identities and access permissions for secure access to Azure resources.
6. **Azure Key Vault:**
   - Ensures secure storage of **sensitive credentials** such as usernames and passwords needed for connecting to databases.
   - The credentials are encrypted and retrieved when needed for secure authentication.
7. **Power BI:**
   - Used for **data visualization** and reporting. The tool connects to the transformed data stored in **Azure Synapse Analytics** and generates **dynamic reports**.
   - It creates visualizations like charts and graphs based on the analyzed data.

Detailed Project Steps:

1. Data Ingestion:

- The project starts by connecting to the local **SQL Server** database via **Azure Data Factory**.
- Data from multiple tables in SQL Server (around 6 or 7 tables) is extracted and copied to **Azure Data Lake** in the **bronze layer**.
- The bronze layer holds the data as a **backup** in its raw format without any changes.

2. Data Transformation (Using Azure Databricks):

- **Azure Databricks** is used to process and clean the data.
- Data transformation is done in two stages:
  - **Bronze to Silver Layer:** Basic transformations such as changing data types, renaming columns, and ensuring compatibility with Azure cloud storage.
  - **Silver to Gold Layer:** More complex transformations to produce the final, clean version of the data. This step ensures that the data is ready for analysis, free of errors, and in its most usable format.
- The **Lakehouse architecture** is applied, which means that data moves through different layers:
  - **Bronze Layer:** Raw data.
  - **Silver Layer:** Partially transformed data.
  - **Gold Layer:** Fully cleaned data.

3. Data Loading into Synapse Analytics:

- Once the data is transformed and cleaned in **Azure Databricks**, it is loaded into **Azure Synapse Analytics**.
- A **SQL script** is created to set up the database schema and tables in Synapse Analytics, mirroring the original on-premises database structure.
- Data from the **gold layer** is then inserted into these newly created tables, making it ready for querying and analysis.

4. Data Reporting with Power BI:

- The **Power BI** tool is used to generate reports and create visualizations from the data stored in **Azure Synapse Analytics**.
- Reports include various types of charts, graphs, and dashboards to provide insights into the data.
- These visualizations can be customized based on business needs, and they offer **real-time updates** as the data changes.

5. Automation with Pipelines:

- **Azure Data Factory Pipelines** automate the entire data flow.
- The pipeline automatically detects any new rows added to the on-premise SQL Server database and triggers the extraction, transformation, and loading (ETL) process.
- This ensures that **Power BI** reports are always up to date, as they reflect any changes made to the source data.

6. Security Management:

- **Azure Active Directory** is used to control access to different parts of the Azure environment, ensuring that only authorized users can manage resources or view sensitive data.
- **Azure Key Vault** stores critical credentials, such as database usernames and passwords, securely. These credentials are used to connect to the on-premises SQL Server and are retrieved in an encrypted format, ensuring they are not exposed directly.

Summary of the End-to-End Process:

1. **Data is ingested** from the on-premises SQL Server using **Azure Data Factory** and stored in **Azure Data Lake**.
2. **Azure Databricks** transforms the data in stages, moving it through the bronze, silver, and gold layers.
3. **Azure Synapse Analytics** is used to create databases and store the cleaned, transformed data for querying and reporting.
4. **Power BI** connects to Synapse Analytics to generate visual reports, ensuring data is always up-to-date.
5. **Automation** through **pipelines** ensures the entire ETL process happens automatically when new data is added.
6. **Security** is managed via **Azure Active Directory** and **Azure Key Vault**, ensuring that access is controlled and credentials are stored securely.