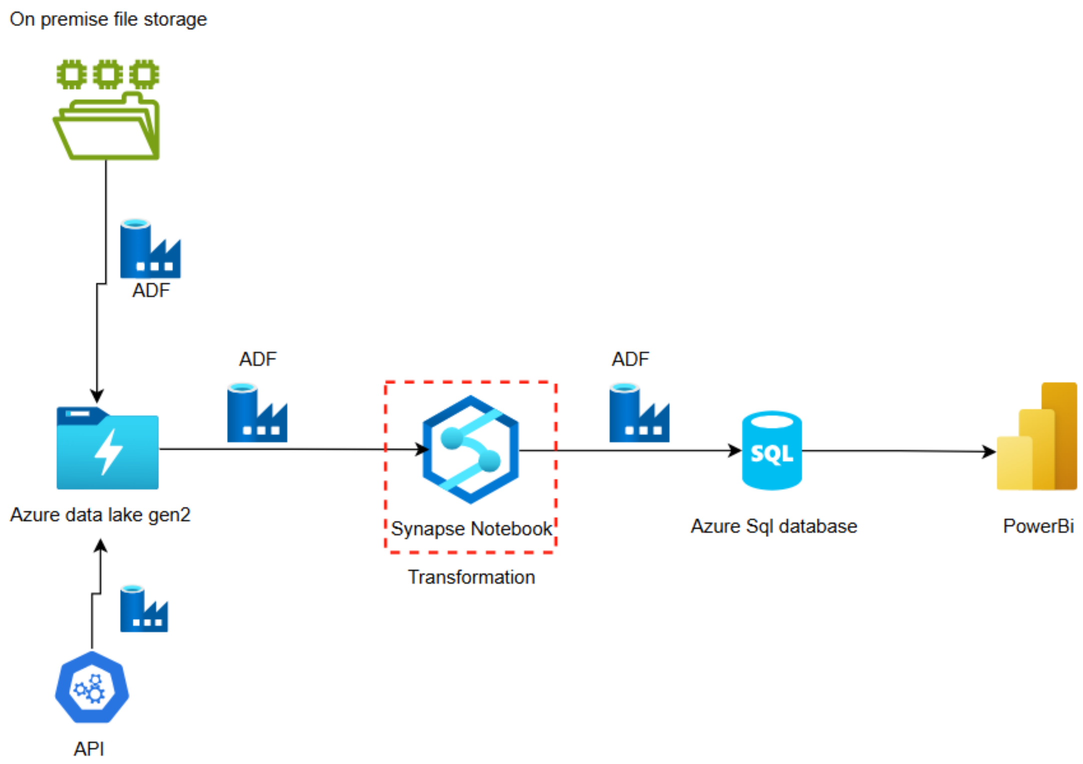
Project: OTT Movie Data ETL Pipeline using Azure



Objective:

The primary objective of this project was to combine data from multiple sources into a centralized location, making it readily available for further insights and analytics in a database. The goal was to integrate OTT movie data and user data into a unified data platform for reporting and business intelligence purposes.

Key Features:

1. **Data Sources:**
   - **OTT Movie Data (Local Server):** This data included movie release information, genre, runtime, release date, added date, IMDb score, views, and additional metadata. It was stored in CSV format on a local server.
   - **User Data (API):** This data included user information such as signup date, email, user ID, full name, and other general details. The data was initially in JSON format from a REST API and was later saved as CSV in the raw container of Azure Data Lake Gen2.
2. **Data Ingestion:**
   - **Azure Data Factory (ADF)** was used for orchestrating the ingestion of data from both sources. A self-hosted **Integration Runtime** was configured for accessing files from the

local server, while the **Azure Integration Runtime** was used for fetching user data from the REST API.

- The data was stored in **Azure Data Lake Storage Gen2**:
    - **Raw Container**: Raw files were stored directly in CSV format.
        - Local server files were saved in folders organized by date.
        - API data was saved as one CSV file per day.
    - **Refined Container**: After data transformations, the cleaned and processed data was moved to this container.

3. **Data Transformation:**
    - The data was processed and transformed using **Azure Synapse Notebooks**. The transformations included:
        - **Data Cleansing:** Handling missing values and correcting invalid data.
        - **Deduplication:** Removing duplicate entries from the datasets.
        - **Aggregations:** Summing or averaging key metrics like views or ratings.
        - **Standardizing Formats:** Ensuring uniform data formats across datasets (e.g., date formats).
        - **Joins:** Merging data from different sources (movie data with user data) based on relevant keys.
    - **Business Logic**:
        - **Movie Recommendations:** Based on users' preferred genres, recommendations were generated.
        - **Top Movies by IMDb Score:** Identified top movies within each genre based on IMDb scores.
4. **Data Storage and Access:**
    - Transformed and cleansed data was moved to an **Azure SQL Database** using ADF pipelines.
    - The data was structured to facilitate fast querying for business insights. (The structure could be normalized or denormalized, depending on reporting needs.)
    - **Power BI** was integrated with Azure SQL Database, providing business users with interactive dashboards and real-time reports.
5. **Orchestration and Automation:**
    - The entire ETL process was orchestrated using **Azure Data Factory pipelines**, with scheduled triggers set for daily execution to refresh data.

   - **Azure Key Vault** was used to securely store credentials such as API keys and connection strings, ensuring safe access to sensitive data.

6. **Performance Optimization and Error Handling:**
   - **ADF** pipelines included retry logic and error handling mechanisms to ensure reliable data ingestion and transformation processes.
   - The data in Azure Data Lake was organized by partitioning, with files grouped by date and source to improve data retrieval and processing performance.
7. **Analytics and Reporting:**
   - **Power BI** dashboards were created to visualize key business metrics, such as:
     - User engagement across genres.
     - Popular movies and their performance.
     - Movie recommendations for users based on their preferences and top-rated movies within each genre.
   - These dashboards enabled stakeholders to make data-driven decisions and gain insights from the aggregated data.

Technology Stack:

- **Azure Data Factory**: For data integration, orchestration, and pipeline management.
- **Azure Data Lake Storage Gen2**: For raw and refined data storage.
- **Azure Synapse Analytics**: For data transformation and processing using Synapse Notebooks.
- **Azure SQL Database**: For storing structured data after transformation.
- **Power BI**: For creating interactive reports and dashboards.
- **Azure Key Vault**: For managing secrets and credentials.
- **REST API**: For fetching user data.

Outcomes:

- The project successfully centralized OTT movie and user data from disparate sources into a unified data lake and SQL database.
- Automated data pipelines ensured daily updates of the datasets, keeping the analytics environment current.
- **Power BI** provided valuable insights into movie performance and user engagement, empowering business teams to make informed decisions based on real-time data.