**RETAIL REVENUE PROJECT**

**Business Requirements**

The objective of this project was to **calculate Total Revenue and Total Purchases on a daily basis** and to analyse **Total Sales by each Product Category** using the Azure ecosystem and Power BI.

**BRONZE LAYER (Data Ingestion from API to Data Lake Storage)**

**AZURE:**

1. **Azure Data Lake Setup:**
   * Created a new **Azure Data Lake Storage (ADLS Gen2)** account and configured a container with folders named **Bronze**, **Silver**, and **Gold** to follow the **Medallion Architecture** approach.
2. **Synapse Workspace Creation:**
   * Established a new **Azure Synapse Analytics workspace** to centralize data ingestion, transformation (PySpark), and SQL query operations within a unified environment.

**SYNAPSE:**

1. **Source Data Identification:**
   * The raw data source was an **API endpoint** providing data in **JSON format**.
2. **Data Ingestion Pipeline:**
   * Created a new **pipeline** in Azure Synapse and implemented a **Copy Activity** to extract data from the API and load it into the **Bronze folder** of ADLS in **Parquet format** for efficient processing and storage.
3. **Initial Dataframe Creation:**
   * Developed a **PySpark notebook** in Synapse to import the Parquet file from the Bronze folder and created a **Bronze DataFrame** for preliminary data reading and inspection.

**SILVER LAYER (Data Cleaning and Transformation)**

**SYNAPSE:**

1. **Data Preparation:**
   * Performed cleaning and transformation to create the **Silver Layer dataset** for further analysis and aggregation.
2. **Business Logic Filtering:**
   * Filtered the dataset to include only records with **Event Type = 'Purchase'** in alignment with business requirements.
3. **Null Value Handling:**
   * Removed null values from critical fields — *Customer ID* and *Amount* — using the dropna() function to ensure data accuracy.
4. **Column Transformations:**
   * Created a new column **[Event Date]** by converting *event\_timestamp* to a proper **Date** format.
   * Standardized **Payment Method** values to **lowercase** for consistency and accurate grouping.
   * Converted the **Amount** column to **float** data type to handle decimal precision.
5. **Column Selection:**
   * Selected only the relevant columns for analysis:  
     *event\_id, customer\_id, event\_date, product\_id, product\_category, payment\_method, amount, location.*
6. **Silver Dataset Creation:**
   * Saved the cleaned and transformed data using  
     df\_silver.write.mode("overwrite").parquet()  
     to generate the final **Silver dataset** in ADLS.

**GOLD LAYER (Business Aggregation and Output Generation)**

**SYNAPSE:**

1. **Loading Silver Data:**
   * Imported the Silver dataset into a DataFrame using spark.read.parquet() for further transformation.
2. **Daily Revenue and Purchases:**
   * Grouped data by **event\_date** and calculated:
     + **Total Revenue** → *sum of amount*
     + **Total Purchases** → *count (\*)*  
       to display **daily revenue and purchase trends**.
3. **Sales by Product Category:**
   * Grouped data by **product\_category**, calculated the **sum of amount** as *Total Sales*, and displayed results in **descending order** using orderBy().
4. **Gold Dataset Storage:**
   * Saved both aggregated outputs as **two distinct tables** for easy accessibility and reporting.

**POWER BI DASHBOARD (Visualization and Insights)**

1. **Data Export and Preparation:**
   * Converted the final Gold tables into **SQL tables** and exported them in **CSV format** for Power BI integration.
2. **DAX Calculation for Daily Purchases:**
   * Created a calculated column in Power BI to display the **count of each Event Date** (Total Purchases per day) using the following DAX expression:
   * EventDateCount =
   * CALCULATE(
   * COUNTROWS('SQL script 1'),
   * ALLEXCEPT('SQL script 1', 'SQL script 1'[event\_date])
   * )
3. **Dashboard Development:**
   * Designed a **Power BI Dashboard** delivering actionable business insights, featuring:
     + **Total Sales** (Card Visual)
     + **Total Sales by Date**
     + **Total Sales by Location**
     + **Total Sales by Category**
     + **Total Sales by Payment Method**

**SUMMARY:**

The **Retail\_Revenue Project** was designed to calculate daily total revenue, total purchases, and total sales by category using Azure’s **Medallion Architecture** (Bronze–Silver–Gold). Raw JSON data was ingested from an API into **Azure Data Lake Storage** via **Synapse Pipelines** and stored in **Parquet format** (Bronze Layer). The data was then cleaned and transformed in **Synapse Notebooks** using **PySpark**, where purchases were filtered, null values removed, and key columns standardized (Silver Layer). Aggregations were performed to compute daily revenue and category-wise sales, with the results saved as Gold datasets. Finally, the processed data was visualized in **Power BI**, featuring insights such as total sales, daily trends, location-wise performance, and payment method analysis — providing a complete end-to-end data analytics solution for business decision-making.