Extracting Sustainability Data via APIs and Integration with Azure

Capstone Project

Yahya Kocakale
ID: 2303099

1. Introduction

My data engineering capstone project aims to develop a robust data engineering solution that provides the Company with critical sustainability data for their consulting engagements. During the data sourcing phase, I identified relevant open data APIs from the World Bank and Eurostat.

We sourced data from the World Bank Indicators API and Eurostat Statistics API. The next step was to load this data into the company's Azure platform using Azure Data Factory (ADF). I created a single pipeline comprising four Copy Data activities and four Data Flows, which saved data from each KPI to a separate table in Azure SQL Database.

After deploying the data to the Azure SQL database, I utilized a Databricks notebook for comprehensive data analytics and subsequently developed a Power BI dashboard using Power BI Desktop. The data analytics revealed a distinct upward trend in renewable energy consumption alongside a significant decrease in CO2 emissions across various regions globally. The results from the ARIMA model indicated a negative correlation between CO2 emissions and renewable energy consumption, suggesting that as renewable energy usage increases, CO2 emissions tend to decrease.

Through this capstone project, I aim to contribute significantly to the global sustainability movement by equipping organizations with the necessary data and analytical tools to make informed, impactful decisions.

2. Key Performance Indicators for Sustainability

To measure the effectiveness and progress of sustainability efforts, we identified four essential Key Performance Indicators (KPIs):

- Renewable Energy Consumption (% of total final energy consumption)
- CO2 Emissions (metric tons per capita)
- Renewable Energy Production (kWh)
- CO2 emissions (kg per 2015 US$ of GDP)

3. Setting up Azure Data Factory

This capstone project focuses on utilizing ADF to extract data from the World Bank and Eurostat APIs, transform it, and store it in an Azure SQL Database for further analysis.

3.1. Creating the Linked Services

Authentication is set to anonymous since a token is not necessary for accessing the Eurostat and World Bank APIs. With this, the linked service was successfully created.

Additionally, an Azure SQL Database linked service was necessary to "sink" the data from the source API into the SQL Database.

3.2. Creating Datasets in Azure Data Factory

The next step involves creating two new datasets in the "Author" section of Azure Data Factory. The first dataset we create is the source dataset, which will connect to the source API linked service. We choose REST for this. We import the schema from the connection and use the GET request method since this is an API call.

3.3. Pipeline

The configuration of the pipeline focuses on the sink tab, where the SQL database is specified as the target location for the data. Upon completion of these configurations, the data pipeline is prepared for the data ingestion process.

3.4. Data Flow

Within the Azure Data Factory environment, the Data Flow activity provides a robust mechanism for data manipulation, particularly suited for data cleaning tasks.

4. Creating Dashboards

4.1. Creating Dashboards in Power BI

The initial stage involves establishing a connection between Power BI and the Azure SQL Database, which serves as the repository for the imported data via APIs.

The final step entails the compilation of the created visualizations into cohesive reports. These reports provide a holistic perspective of the data, enabling users to gain a comprehensive understanding of the underlying trends and insights.

5. Statistical Analysis

After loading data from APIs using Azure Data Factory into an Azure SQL Database, I utilized Databricks for statistical analysis.

5.1. Granger Causality Analysis

A Granger causality test is employed to assess whether past values of renewable energy consumption can statistically predict future values of CO2 emissions per GDP.

Prior to the test, the Austrian data series were extended using the data retrieved from the Azure SQL database. The Augmented Dickey-Fuller (ADF) test indicated non-stationarity in both

series. To address this, differencing was applied, and the ADF test was re-run. After performing second-order differencing, one of the series achieved stationarity. Consequently, both time series were differenced twice.

The analysis suggests that historical renewable energy consumption data may be helpful in predicting future CO2 emissions per GDP in Austria.

5.2. Predicting CO2 Emissions

Having established a potential causal relationship between renewable energy consumption and CO2 emissions per GDP, we can proceed with time series regression to predict future CO2 emissions.
Although the ACF and PACF visuals suggested a model like ARIMAX(13, 2, 0), the best model we found using trial and error method and MSE metric was ARIMAX(2, 2, 1). Since we can make the series stationary by differencing twice of 2, we took d = 2.

6. Summary

This project showed how different Microsoft Azure services can work together to handle sustainability data. We used Azure Data Factory to move and clean data from several public websites (APIs) like Eurostat and the World Bank. The data ended up in a secure Azure SQL Database.

Then, we built interactive dashboards with Power BI to make the data easier to understand. These dashboards used charts and graphs to show trends in renewable energy use and carbon emissions for different countries. This information helps people see how these factors are changing over time and how different countries compare.