Copy Activity Explanation

The Copy Activity in Azure Data Factory (ADF) is a core operation that moves and transforms data between data sources and destinations. It is a key part of many data workflows, facilitating Extract, Transform, Load (ETL) operations

So this project demonstrates how to ingest population data by age group for European countries from Azure Blob Storage(Data source) to Azure Data Lake(Destination) using Azure Data Factory (ADF). The objective is to prepare this data for analysis, which will aid the data science team in making predictions related to COVID-19 and its impact on different age demographics.

Requirements

- Data Source: Population data file containing age group demographics, sourced from the European Union's Eurostat website.
- Data Destination: Azure Data Lake Storage, specifically the "raw" container.

Data Description

The population data includes:

- Country Codes: Identifiers for each European country (e.g., B for Belgium, D for Denmark).
- Age Group Percentages: Distribution of population by age groups from 2008 to 2019.
- Missing Data Indicators: Columns may include placeholders (e.g., 'P' for provisional data) that require cleaning.

Structure

1. Data Upload

- Upload the downloaded population data file into Azure Blob Storage within the "Covid reporting" storage account, under the "population" container.

2. Copy Activity Configuration
- Linked Services: Define connections for Azure Blob Storage and Azure Data Lake.
- Datasets: Create datasets representing the structure of the source (GZIP file) and target (TSV file).
- Copy Activity: Configure the copy activity to:
- Extract data from the GZIP file in Blob Storage.
- Convert and save it as a TSV file in the Data Lake.

3. Pipeline Creation
- Wrap the copy activity, linked services, and datasets in a pipeline to allow execution.

Implementation Steps

1. Set Up Linked Services:
- Configure connection strings for both Azure Blob Storage and Azure Data Lake.

2. Create Datasets:
- Define the source dataset for the GZIP file.
- Define the target dataset for the TSV file.

3. Build the Copy Activity:
- Specify file types and extraction rules in the copy activity settings.

4. Create the Pipeline:
- Integrate all components and prepare for execution.

Data Cleaning (Future Steps)

Post-ingestion, data cleaning will be necessary to handle missing values and provisional data, ensuring the dataset is ready for analysis.

Conclusion

This project serves as a foundational example of using Azure Data Factory for data ingestion, setting the stage for further data processing and analysis.