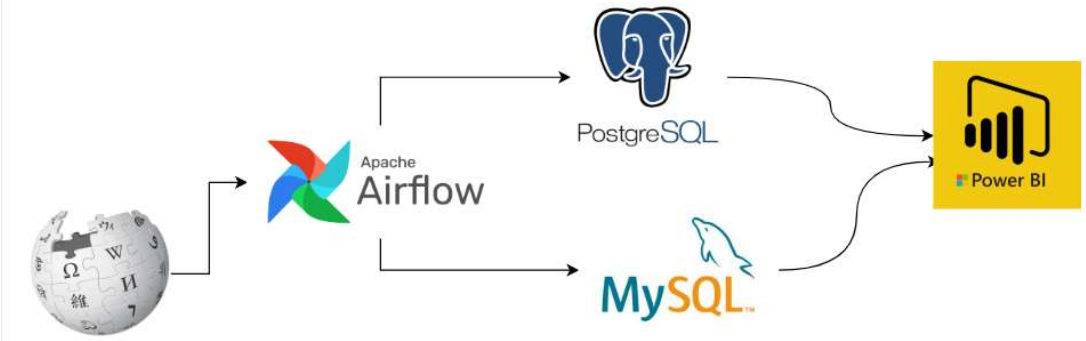
Data Engineering Project: Football Stadiums Data Extraction and Visualization.

Overview

This project aims to extract a list of football stadiums from Wikipedia, clean and transform the data, load it into a PostgreSQL or MySQL database, perform data manipulation to answer key questions, and visualize the results in Power BI.

Tools Required

- Apache Airflow
- Python
- PostgreSQL or MySQL
- Pandas
- Power BI



Instructions

Step 1: Environment Setup

1. Install Required Packages:
2. Set Up PostgreSQL or MySQL:
    - Install the database of your choice.
    - Create a new database for the project.

Step 2: Create Airflow DAG

Step 3: Extract Data from Wikipedia (List of association football stadiums by capacity).

Step 4: Clean and Transform Data

Step 5: Load Data into Database (PostgreSQL / MySQL database).

Step 6: Data Manipulation

Step 7: Load Data into Power BI

Step 8: Visualization in Power BI

1. Load the exported data.
2. Create visualizations based on the answers to the key questions.
3. Key Questions to Visualize:
    - Total number of stadiums by country.
    - Average stadium capacity per continent.
    - Distribution of stadiums built over the years.
    - Largest stadiums by capacity comparison.
    - Count of stadiums.
    - Map visualization.