Japan Visa Analysis: Azure End-to-End Data Engineering

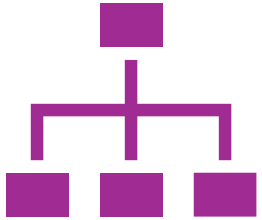
A data engineering pipeline using PySpark, Docker, and Plotly on Azure

Project Participants

Course Instructor: Ahmed Essam Azab

- Ahmed Adel Tawfik
- Mohamed Elsayed Elaraby
- Zain Khaled Alsaid
- Moustafa Mohamed Islam

Overview



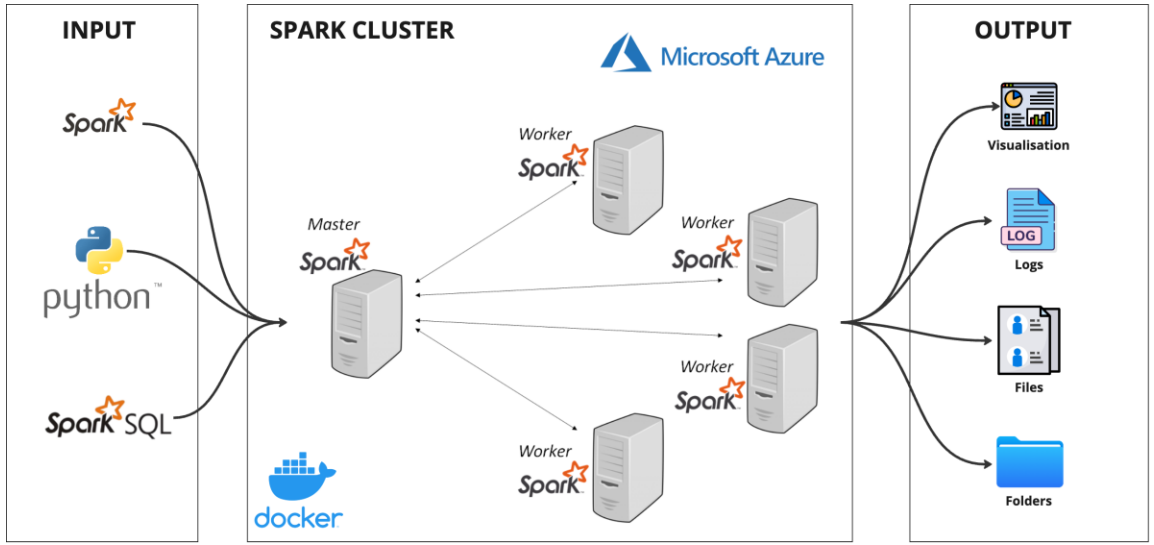
- Introduction to the project.

- Key focus on building an end-to-end data pipeline in Azure using Docker, PySpark, and Plotly.

- Goal: Visualize visa trends in Japan.

System Architecture

- The architecture leverages a Spark cluster deployed within Docker containers on Microsoft Azure.
- • It includes:
- • Input Layer: Visa data (in CSV format) is ingested using PySpark and Spark SQL scripts.
- • Processing Layer: The Spark Master node manages a distributed Spark cluster with multiple Worker nodes, performing parallel data processing.
- • Output Layer: The processed data is exported as visualizations (HTML), logs, files, and cleaned datasets, all stored for further analysis.



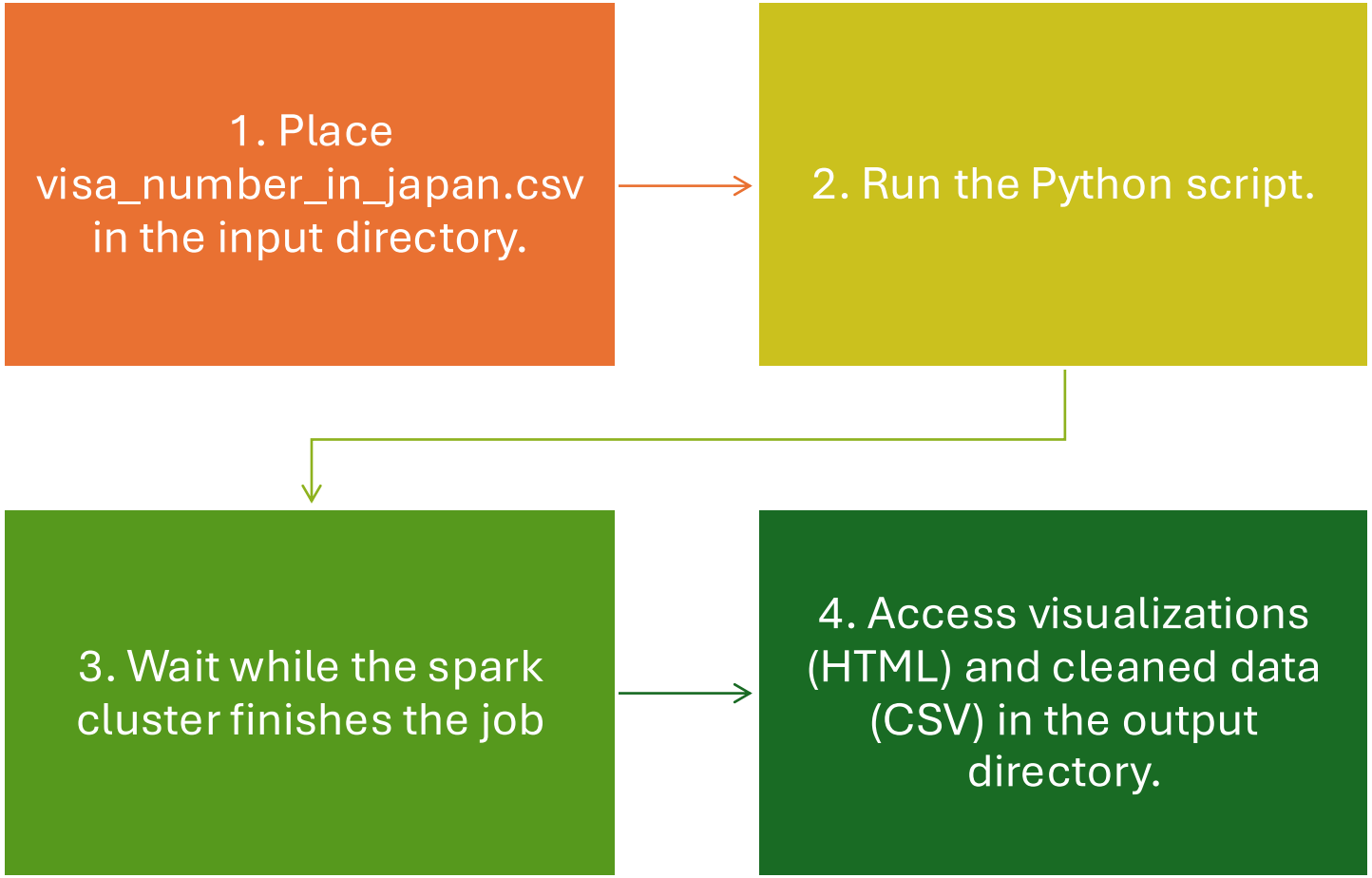
Setup & Requirements

- **Azure Account:** Necessary for deployment.

- **Docker Installation:** Required for Spark cluster.

- **Key Python Libraries:**
  - PySpark
  - Plotly Express
  - pycountry, pycountry_convert
  - fuzzywuzzy

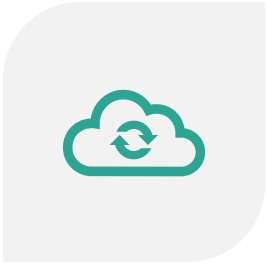
Usage


1. Place visa_number_in_japan.csv in the input directory.
2. Run the Python script.
3. Wait while the spark cluster finishes the job
4. Access visualizations (HTML) and cleaned data (CSV) in the output directory.

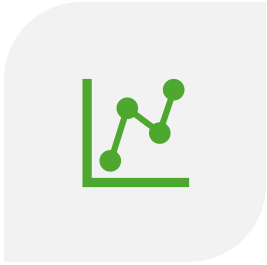
Key Features

• SPARK MASTER-WORKER CLUSTER IN DOCKER ON AZURE.

• DATA INGESTION AND CLEANING PROCESSES.



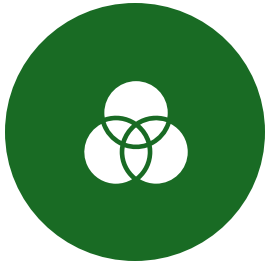
• DATA TRANSFORMATION: ADDING CONTINENT INFORMATION.



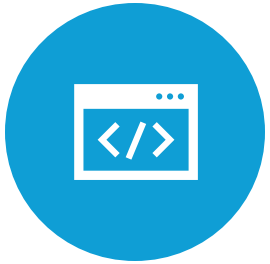
• INTERACTIVE DATA VISUALIZATIONS USING PLOTLY.

Data Cleaning

• Standardizes column names.



• Removes null columns.

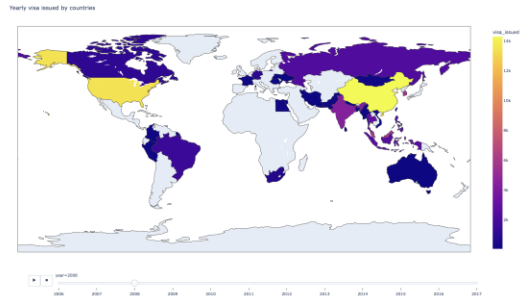


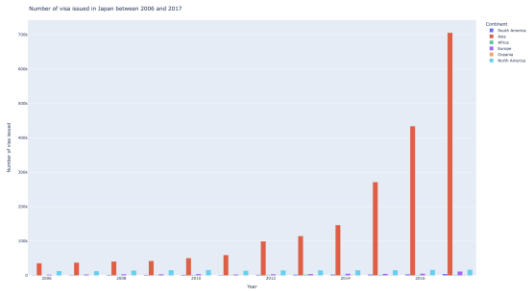
• Uses fuzzy matching for correcting country names.

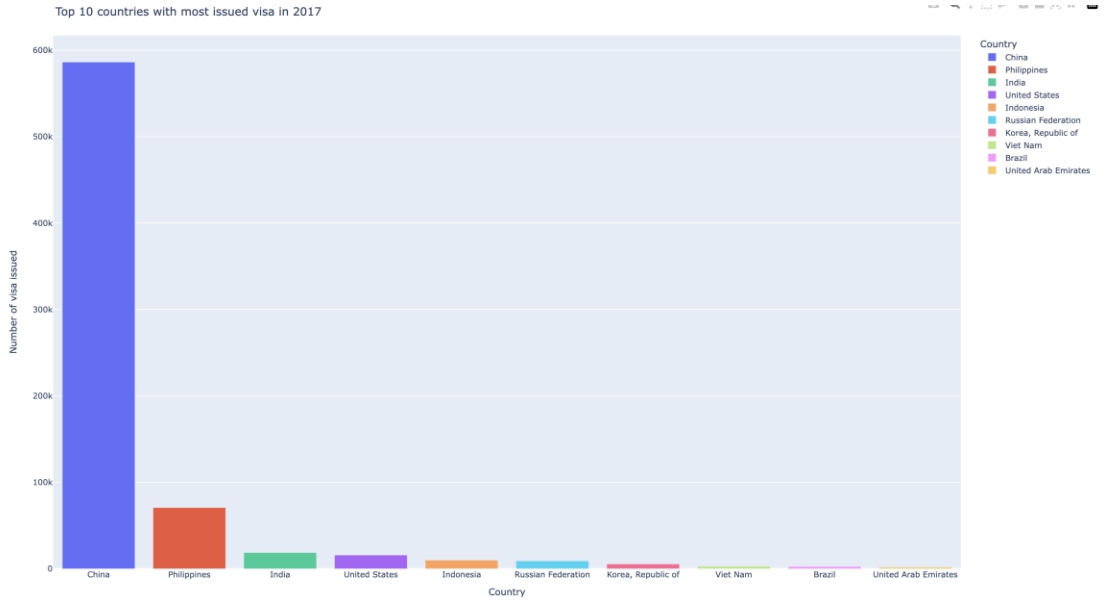
• Maps countries to continents.

Data Visualization

- • Visualizations created with Plotly Express.
- • Insights into visa trends.
- • Saved as interactive HTML files.







Notes & Considerations

- Ensure Azure and Docker are configured correctly.
- Python libraries must be updated.
- Manual country mappings can be edited in country_mapping within main.py.