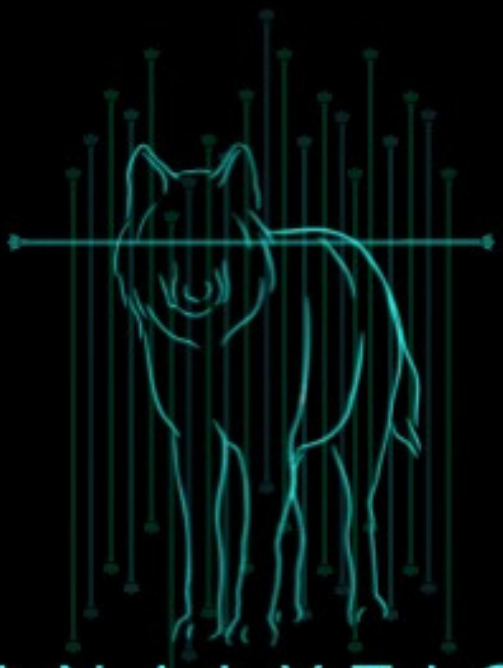


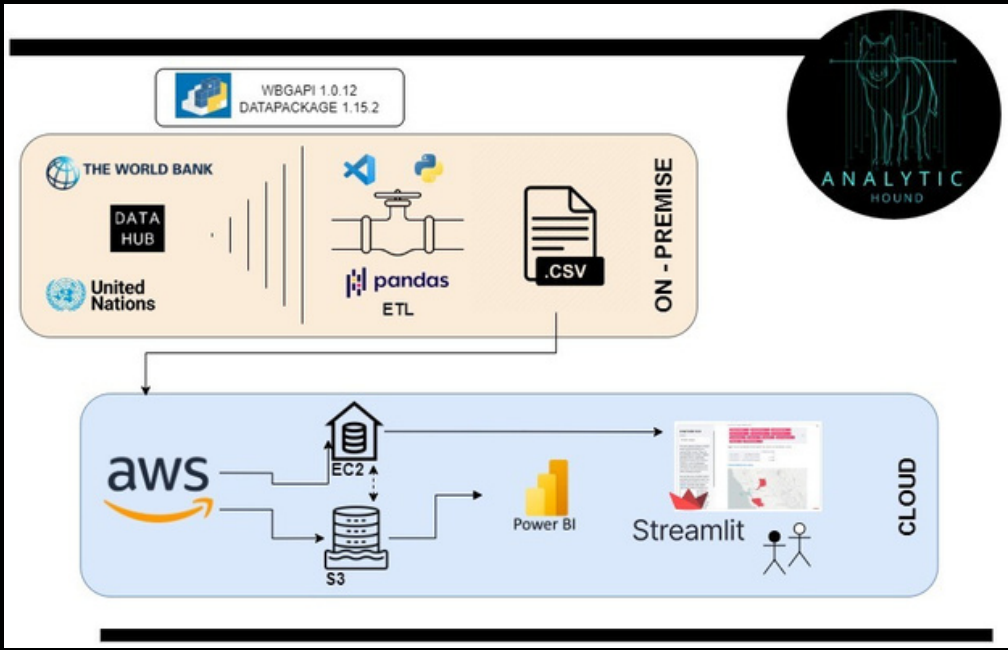
ONG - **Soy Henry**

Report Week #2

Data Process

Milestones - Index

Workflow



Sources:

For the development of the requirements we decided to dispense with the data provided by the client. The main reasons were that we found data of higher quality, variety, temporal frequency and in better condition. To the detriment of said database, we chose the following sources:

- The official data page of the World Bank:
  - https://data.worldbank.org/
- United Nations official data landing page:
  - https://www.un.org/development/desa/pd/data-landing-page
- Another official UN data source:
  - https://data.un.org/

INCREMENTAL DATA LOAD:

The incremental data load application refers to updating the data as the main sources are modified over time.
In the context of the migratory analysis of the project in question, said update is expected to occur with an average periodicity of the semi-annual/annual type.
For the correct implementation of the incremental load, the official documentation for developers within the selected databases was referred. With their support, the pipeline corresponding to said task is established.
As can be seen in the attached links, the following python libraries were used for the direct extraction of data from the different sources:

- WBGAPI 1.0.12 | https://pypi.org/project/wbgapi/
- DATAPACKAGE 1.15.2 | https://pypi.org/project/datapackage/

** Note: datahub works with data of the "package" type, allowing it to be handled with the aforementioned library.*

In short, the code used for this task has the following order of execution:

1. Library import
2. Data extraction directly through python libraries.
3. Saving them in defined variables that are later raw material for the ETL process that will happen immediately after the load is finished.

ETL: FULL REPORT

Data origin

The data used in this project comes from multiple CSV files containing information on various variables from all countries and regions. These variables are studied according to their influence on the positive and negative migration of each country. The data was extracted from world websites, which make such information available.

Transformation proccesses

Once the data from the CSV file was imported, several transformations were performed to prepare it for use. These included:

- Rename columns for clarity.
- Fill in the missing data judiciously.
- Review and remove duplicate rows.
- Remove unnecessary columns.
- Eliminate data corresponding to years prior to 2000.
- Rearrangement of columns for uniformity.

To carry out these transformations, two pipelines were designed that group the various transformations that must be applied to the different datasets. The language used is Python, and the main libraries were Pandas, wbgapi, datapackage, and scikit learn.

Final destination

The transformed data was exported to a CSV file for further use and evaluation for the learning models. The entire pipeline was handed over to the engineering team so they can lift it up in the cloud.

Conclusions

The ETL process was successful, both in the collection and in the transformation of data from different sources. The transformations performed allow the data to be effectively analyzed to obtain valuable information about the effect that different variables have on migration.

Features Dictionary

Category	Feature	Definition
Economy	gdp	Gross Domestic Product (GDP) total gross value added by all resident producers in the economy of each country in U$D.
	gdp_growth	GDP growth: annual percentage growth rate of GDP.
	cons_expen	Final consumption expenditure: SUM of household final consumption expenditure (private consumption) and general government final consumption expenditure.
	gni_capita	GNI per capita (atlas method): gross national income (converted to dollars using the world bank atlas method) divided by the midyear population.
	gross_savings	Gross savings: calculated as gross national income - total consumption.
	consumer_price	Consumer price index: cost of the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly.

Features Dictionary

Category	Feature	Definition
People	gover_exp	Government expenditure on education: General government expenditure on education expressed as a percentage of GDP.
	unemploy	Unemployment, total: Share of the labor force that is without work but available for and seeking employment
Environment	elect	Access to electricity: percentage of population with access to electricity. Electrificacion data is collected from industry, national surveys, and international sources.
	basic_sanitation	People using at least basic sanitation services in urban areas: improved sanitation facilities that are not shared with other households. This indicator encompasses both people using basic sanitation services as well as those using safely managed sanitation services.
	pop_density	Population Density : amount of people per square meter of land area

Features Dictionary

Category	Feature	Definition
Poverty	population_ below	Population below $1.90 a day: percentage of the population living on less than $1.90 a day at 2011 international prices. As a result of revisions in PPP exchange rates, poverty rates for individual countries cannot be compared with poverty rates reported in earlier editions.
	maternal_mortality	Maternal mortality ratio: number of women who die from pregnancy-related causes while pregnant or within 42 days of pregnancy termination per 100,000 live births. The data are estimated with a regression model using information on the proportion of maternal deaths among non-AIDS deaths in women ages 15-49, fertility, birth attendants, and GDP.
	tuberculosis	Incidence of tuberculosis: estimated number of new and relapse tuberculosis cases arising in a given year, expressed as the rate per 100,000 population. All forms of TB are included, including cases in people living with HIV. Estimates for all years are recalculated as new information becomes available and techniques are refined, so they may differ from those published previously.

Features Dictionary

Category	Feature	Definition
Poverty	contributing_f	Contributing family workers and own-account workers, female: workers who hold "self-employment jobs" as own-account workers in a market-oriented establishment operated by a related person living in the same household.
	contributing_m	Contributing family workers and own-account workers, male: workers who hold "self-employment jobs" as own-account workers in a market-oriented establishment operated by a related person living in the same household.
State	profit_tax	Profit tax: amount of taxes on profits paid by the business.
	mobilesubs	Mobile cellular subscriptions: subscriptions to a public mobile telephone service that provide access to the PSTN using cellular technology. The indicator includes (and is split into) the number of postpaid subscriptions, and the number of active prepaid accounts.
	gdp_percapita	GDP per capita
	povertygap	Poverty gap

CHOSEN STACK:

Python

A high-level interpreted programming language whose philosophy emphasizes the readability of its code, it is used to develop applications of all kinds. Multiparadigm programming language. It partially supports object orientation, imperative programming and, to a lesser extent, functional programming. It is an interpreted, dynamic and cross-platform language. Managed by the Python Software Foundation, it is licensed under an open source license, called the Python Software Foundation License. It consistently ranks as one of the most popular programming languages.

Visual studio code (VSC):

Source code editor developed by Microsoft for Windows, Linux, macOS and Web. It includes support for debugging, integrated Git control, syntax highlighting, smart code completion, among other features suitable for building this project.

Libraries within VSC

Pandas: A library written for the python programming language designed to allow data manipulation and analysis. It offers data structures and operations for manipulating number tables and time series.

Amazon Web Services (AWS)

A collection of public cloud computing services (also called web services) that together form a cloud computing platform, delivered over the Internet by Amazon.com.

CHOSEN STACK:

Microsoft Power BI

Interactive data visualization software developed by Microsoft with a primary focus on business intelligence.

Streamlit

Open source Python library that makes it easy to build custom web applications for machine learning and data science.

Implemented data structure (DW, DL)

Considering the client's requirements, in this case the Engineering department determined that the most efficient option is to mount a DL and a DW in the Amazon Web Services Cloud services.
Among the most important reasons for choosing AWS is the fact that on the same platform it offers a DL called S3 and a DW called EC2.

- EC2 allows users to rent virtual computers on which they can run their own applications. Allows scalable deployment of applications by providing a Web service. We use this service to host streamlit.
- S3 provides scalable object storage that is organized into buckets. Each object is identified by a unique key assigned by the user. We use this service to store raw files and which in turn serves as a backup for the information in EC2.

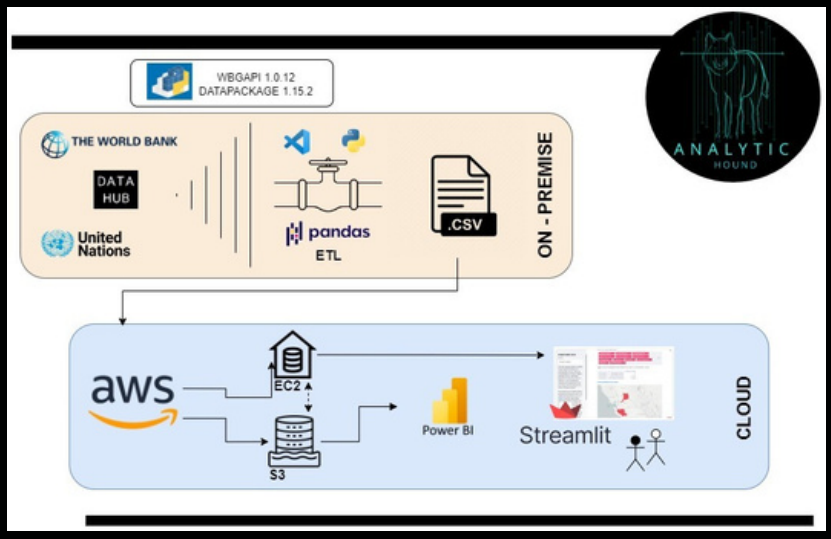
The start of the job was to create an instance on EC2 and manually upload files from Visual Studio Code to see its behavior. Then we create a bucket in S3 to store the information and we test it by manually uploading files from AWS.

Later we create a role for the connection since we would need to move information from one to the other. It is important to emphasize that this connection does not prevent individual access to either one, nor does it prevent one from being able to delete data without affecting the other.
We initially had the idea of automating the link between GitHub and the bucket, but it was later dropped in favor of a faster and easier option. We modified the ETL pipeline to directly send the final csv with all the clean data to the bucket, to then automatically pass it to the data warehouse.

The next step to implement is the CRON configuration, which will execute the command that will update the file containing the information with which we perform our analysis on a monthly basis. In this way we will have the entire automated process that will be executed once a month.

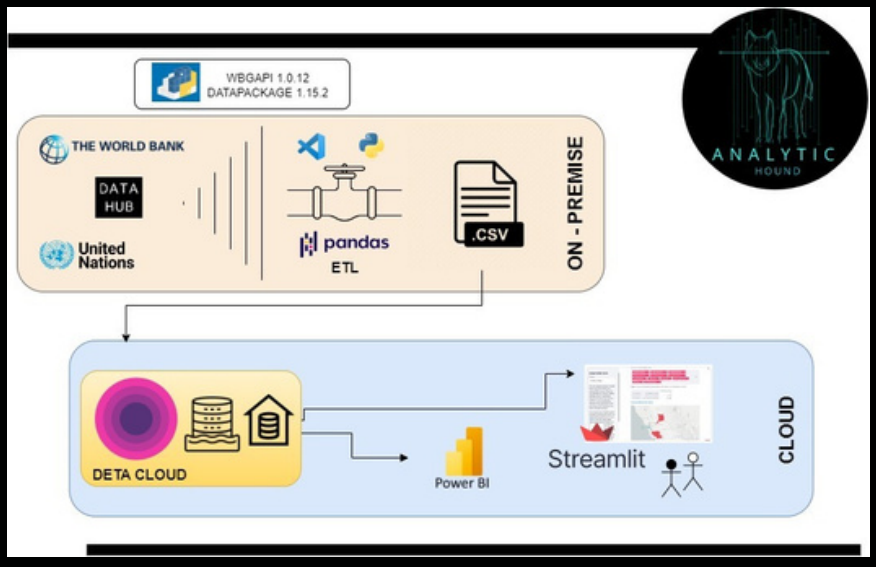
DW Automation Solution

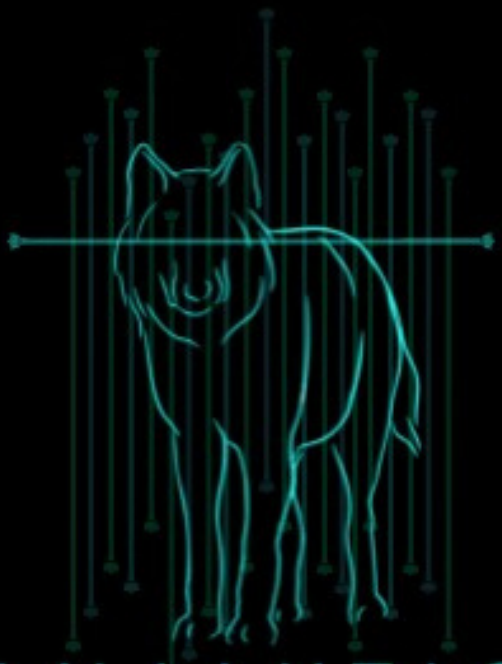
In this section we show the entire process that we detailed previously, graphically.



Likewise, in the following diagram we detail the workflow that satisfies the requirement made by the PO at the end of Sprint #1. Said modification of the deliverables consisted of considering an alternative plan for compliance with the requirements in the event that for some eventuality the primary option could not be implemented.

In this alternative we decided to use the DETA cloud service for data storage and for deploying streamlit. This version works as a PLAN B which has already been tested and implemented.





ANALYTIC

HOUND