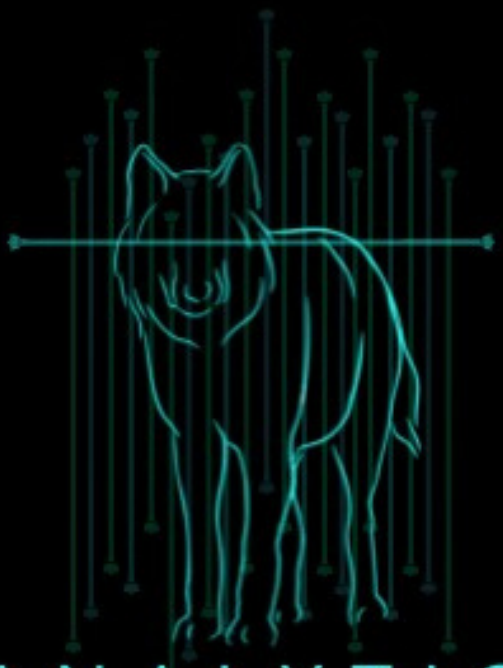


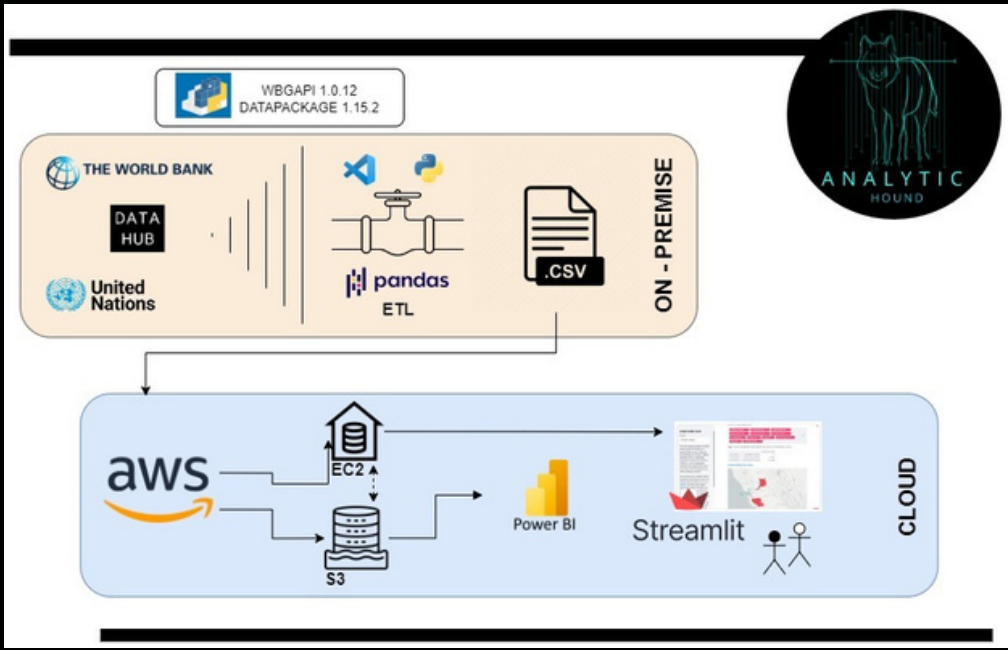
ONG - **Soy Henry**

Reporte Semana #2

Data Process

Indice

Workflow



Fuentes

Para el desarrollo de los requerimientos decidimos prescindir de los datos brindados por el cliente. Los motivos principales fueron que encontramos datos de mayor calidad, variedad, frecuencia temporal y en mejor estado. En detrimento de dicha base de datos, escogimos las siguientes fuentes:

- La página oficial de data de el World Bank:
- https://data.worldbank.org/
- La página oficial de data de United Nations: https://www.un.org/development/desa/pd/data-landing-page
- Otra fuente oficial de data de UN: https://data.un.org/

Carga incremental de datos:

El aplicativo de la carga incremental de datos hace referencia a la actualización de los mismos en la medida que las fuentes principales sean modificadas a lo largo del tiempo.

En el contexto del análisis migratorio del proyecto en cuestión, dicha actualización es esperable que suceda con una periodicidad promedio de tipo semestral/anual.

Para la correcta implementación de la carga incremental, se remitió a la documentación oficial destinadas a desarrolladores dentro de las bases de datos seleccionadas Con el soporte de las mismas, se establece el pipeline correspondiente a dicha tarea.

Como puede observarse en los links adjuntos, las siguientes librerías de python fueron utilizadas para la extracción directa de datos de las distintas fuentes:

- WBGAPI 1.0.12 | https://pypi.org/project/wbgapi/
- DATAPACKAGE 1.15.2 | https://pypi.org/project/datapackage/
    * *Nota: datahub presenta un funcionamiento con datos de tipo "paquete" permitiendo su manejo con la librería mencionada.*

En resumen, el código utilizado para esta tarea presenta el siguiente orden de ejecución:

1. Importación de librerías
2. Extracción de datos en forma directa mediante librerías de python.
3. Guardado de las mismas en variables definidas que luego son materia prima del proceso de ETL que sucederá inmediatamente luego de finalizada la carga.

ETL: Informe completo.

Origen de los datos

Los datos que se utilizan en este proyecto provienen de múltiples archivos CSV que contienen información sobre diversas variables de todos los países y regiones. Dichas variables son estudiadas según su influencia en la migración positiva y negativa de cada país. Los datos fueron extraídos de sitios web mundiales, los cuales ponen a disposición dicha información.

Proceso de transformación

Una vez que se importaron los datos del archivo CSV, se realizaron varias transformaciones para prepararlos para su uso. Estas incluyeron:

- Renombrar columnas para mayor claridad.
- Completar los datos faltantes con criterio.
- Revisar y eliminar filas duplicadas.
- Eliminar columnas innecesarias.
- Eliminar datos correspondientes a años previos al 2000.
- Reorganización de las columnas para uniformidad.

Para llevar a cabo estas transformaciones, se diseñaron dos pipelines que agrupan las diversas transformaciones que se deben aplicar a los distintos datasets. El lenguaje utilizado es Python, y las bibliotecas principales fueron Pandas, wbgapi, datapackage, y scikit learn.

Destino final

Los datos transformados se exportaron a un archivo CSV para su posterior utilización y evaluación para los modelos de aprendizaje. Todo el pipeline fue entregado al equipo de ingeniería para que puedan levantarlo en la nube.

Conclusiones

El proceso de ETL fue exitoso, tanto en la recolección como en la transformación de los datos provenientes de diferentes fuentes. Las transformaciones realizadas permiten que los datos se analicen de manera efectiva para obtener información valiosa sobre el efecto que distintas variables tienen sobre la migración.

Features Dictionary

Categoría	Feature	Definición
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

Categoría	Feature	Definición
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

Categoría	Feature	Definición
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

Categoría	Feature	Definición
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

Stack Elegido

Python

Lenguaje de alto nivel de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código, se utiliza para desarrollar aplicaciones de todo tipo. Lenguaje de programación multiparadigma. Soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma. Administrado por Python Software Foundation, posee una licencia de código abierto, denominada Python Software Foundation License. Se clasifica constantemente como uno de los lenguajes de programación más popular.

Visual studio code (VSC):

Editor de código fuente desarrollado por Microsoft para Windows, Linux, macOS y Web. Incluye soporte para la depuración, control integrado de Git, resaltado de sintaxis, finalización inteligente de código, entre otras funciones idóneas para la construcción de este proyecto.

Librerías dentro de VSC

Pandas: librería escrita para el lenguaje de programación de python diseñada con el fin de permitir la manipulación y análisis de datos. Ofrece estructuras de datos y operaciones para la manipulación de tablas numéricas y series temporales.
Las librerías de WBGAPI 1.0.12 y DATAPACKAGE 1.15.2 se describen en la página 4, “Carga incremental de datos”.

Amazon Web Services (AWS)

Colección de servicios de computación en la nube pública (también llamados servicios web) que en conjunto forman una plataforma de computación en la nube, ofrecidas a través de Internet por Amazon.com.

Stack Elegido

Microsoft Power BI

Software de visualización de datos interactivo desarrollado por Microsoft con un enfoque principal en la inteligencia empresarial.

Streamlit

Biblioteca Python de código abierto que facilita la creación de aplicaciones web personalizadas para el aprendizaje automático y la ciencia de datos

Estructura de datos implementada (DW, DL)

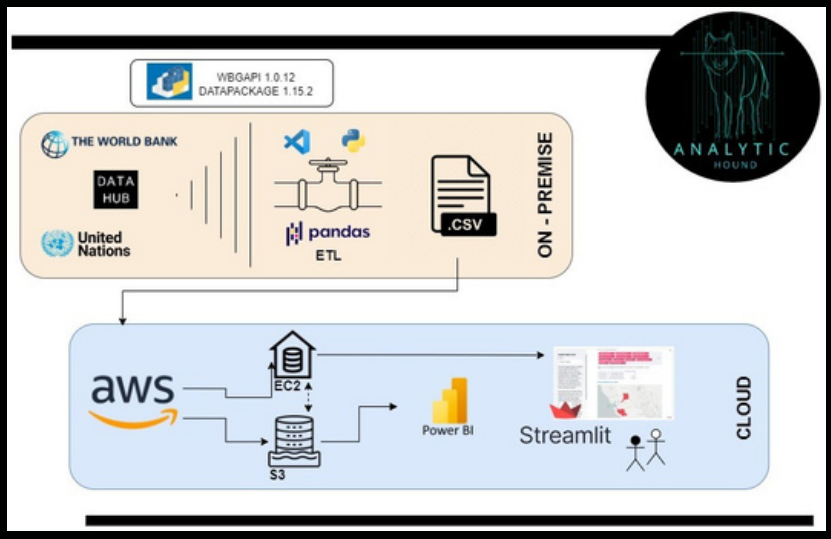
Considerando los requerimientos del cliente, en este caso el departamento de Engineering determinó que la opción más eficaz consiste en montar un DL y un DW en los servicios Cloud de Amazon Web Services.
Entre los motivos más importantes para inclinarnos por AWS se encuentra el hecho de que en una misma plataforma ofrece un DL llamado S3 y un DW llamado EC2.

- EC2 permite a los usuarios alquilar computadores virtuales en los cuales pueden ejecutar sus propias aplicaciones. Permite el despliegue escalable de aplicaciones proveyendo un servicio Web. Este servicio lo utilizamos para hostear streamlit.
- S3 proporciona almacenamiento de objetos escalables que se organizan en cubos (buckets). Cada objeto se identifica mediante una clave única asignada por el usuario. Este servicio lo utilizamos para almacenar archivos sin procesar y que a su vez nos sirva de backup para la información en EC2.

El comienzo del trabajo fue crear una instancia en EC2 y subir archivos de forma manual desde Visual Studio Code para ver su comportamiento. Luego creamos un bucket en S3 para poder almacenar la información y lo testeamos subiendo archivos de manera manual desde AWS.
Posteriormente creamos un rol para la conexión ya que necesitaríamos mover información de uno hacia el otro. Es importante recalcar que esta conexión no impide acceder individualmente a cualquiera de los dos, ni tampoco impide poder borrar datos uno sin afectar al otro.
Inicialmente teníamos la idea de automatizar el enlace entre GitHub y el bucket, pero posteriormente fue descartada por una opción más rápida y sencilla. Modificamos el pipeline de ETL para enviar directamente el csv final con toda la data limpia al bucket, para luego pasarla automáticamente al datawarehouse. El siguiente paso a implementar es la configuración del CRON, el cual ejecutará de forma mensual el comando que actualizará el archivo que contiene la información con la cual realizamos nuestros análisis. De esta forma tendremos todo el proceso automatizado que se ejecutará una vez por mes.

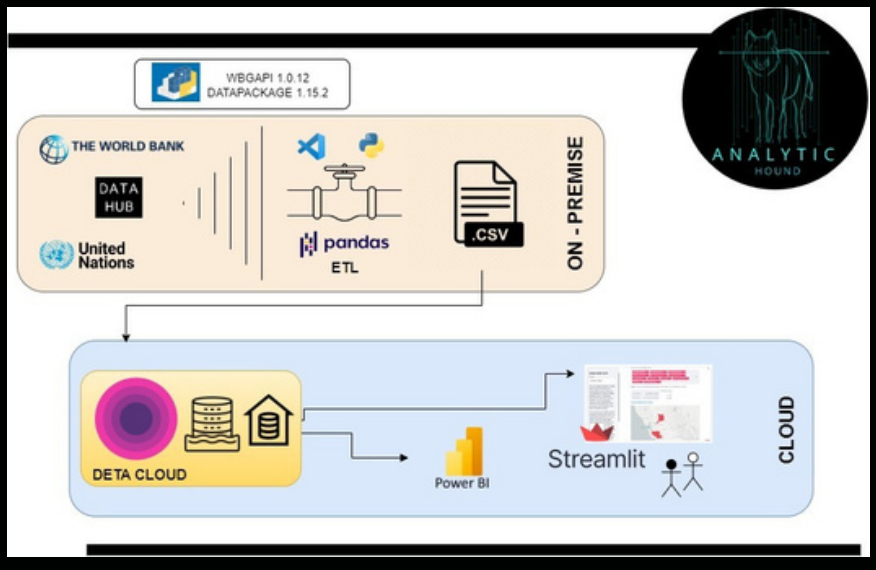
Solución de automatización del DW

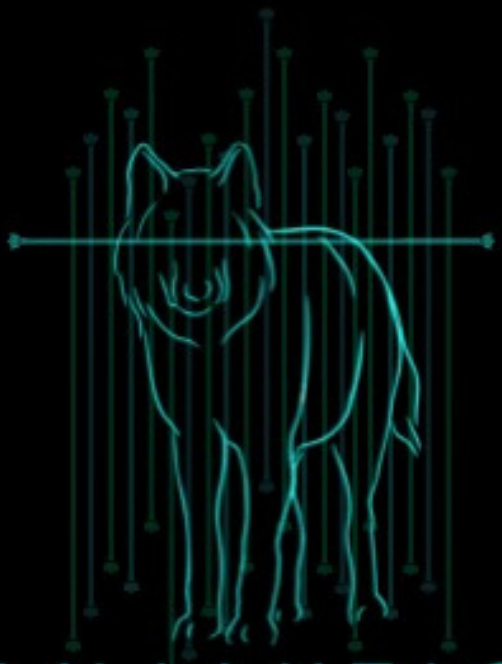
En este apartado mostramos todo el proceso que detallamos anteriormente, de forma gráfica.



Asimismo en el siguiente diagrama detallamos el workflow que satisface el requerimiento realizado por el PO al final del Sprint #1. Dicha modificación de los entregables consistía en considerar un plan alternativo de cumplimiento de los requerimientos en caso de que por alguna eventualidad la opción primaria no pudiera ser implementada.

En esta alternativa decidimos utilizar el servicio cloud de DETA para el almacenamiento de datos y para el deploy de streamlit. Esta versión funciona como un PLAN B el cual ya ha sido testeado e implementado.





ANALYTIC

HOUND