FLUJOS MIGRATORIOS EN ARGENTINA (1960 – 2022)



Sprint 1: Entendimiento de situación, alcance de propuesta y solución tecnológica para implementación de pipeline

**I. Entendimiento de la situación propuesta**

Los flujos migratorios, el movimiento de personas de un lugar a otro, son causados ​​por una compleja interacción de varios factores. Estos factores se pueden clasificar en términos generales en factores de empuje y factores de atracción. Es importante tener en cuenta que la migración es un fenómeno multifacético, y las decisiones individuales de migrar están influenciadas por una combinación de estos factores en lugar de uno solo. Aquí hay una descripción general de las principales causas de los flujos migratorios:

|  |  |
| --- | --- |
| **Factores de presión** | **Factores de atracción** |
| **Razones económicas:** La falta de oportunidades laborales, los salarios bajos y las malas condiciones económicas en el país de origen pueden empujar a las personas a buscar mejores perspectivas económicas en otros lugares. | **Mejores oportunidades económicas:** la promesa de salarios más altos, mejores perspectivas laborales y mejores niveles de vida en un país de destino puede atraer a inmigrantes que buscan un progreso económico. |
| **Conflicto y violencia:** los conflictos armados, las guerras civiles, la inestabilidad política y la violencia pueden obligar a las personas a huir de sus hogares en busca de seguridad y protección. | **Educación y capacitación:** el acceso a instalaciones de educación y capacitación de calidad puede atraer a inmigrantes que desean mejorar sus habilidades y conocimientos. |
| **Factores ambientales:** los desastres naturales como huracanes, terremotos, sequías y otros desafíos ambientales pueden hacer que las áreas sean inhabitables, lo que lleva a las personas a migrar a regiones más seguras. | **Estabilidad política y seguridad:** los países con sistemas políticos estables, estado de derecho y bajas tasas de criminalidad pueden ser atractivos para las personas que buscan seguridad y protección. |
| **Falta de servicios:** el acceso limitado a la atención médica, la educación y otros servicios esenciales puede llevar a las personas y las familias a buscar mejores condiciones de vida en otros lugares. | **Reunificación familiar:** el deseo de reunirse con miembros de la familia que ya han migrado puede ser un fuerte factor de atracción para muchos migrantes. |
| **Factores políticos:** los regímenes represivos, la falta de libertades políticas y los abusos de los derechos humanos pueden motivar a las personas a abandonar sus países de origen en busca de asilo político. | **Vínculos culturales y sociales:** las comunidades de inmigrantes de la misma región o antecedentes culturales pueden crear un entorno de apoyo que aliente a otros a unirse a ellos. |
|  | **Mejor calidad de vida:** el acceso a mejores servicios de salud, servicios sociales, infraestructura y calidad de vida en general puede ser atractivo para los migrantes que buscan un mejor nivel de vida. |
|  | **Políticas de inmigración abiertas:** los países con políticas de inmigración más indulgentes y actitudes acogedoras hacia los inmigrantes pueden atraer a personas que buscan un nuevo hogar. |
|  | **Turismo y Trabajo Temporal:** Algunos migrantes se mudan temporalmente por trabajo de temporada o turismo, pero pueden decidir quedarse más tiempo o permanentemente debido a las experiencias positivas que tienen. |

Fuente: Simpson, Nicole B. Demographic and economic determinants of migration. IZA World of Labor. https://wol.iza.org/articles/demographic-and-economic-determinants-of-migration/long.

The push-pull factors of migration-geography.Tutor2u. https://www.tutor2u.net/geography/reference/the-push-pull-factors-of-migration.

Kyaing Kyaing Thet. Pull and Push Factors of Migration: A Case Study in the Urban Area of Monywa Township, Myanmar. World of Statistics. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.worldofstatistics.org/files/2014/03/Pull-and-Push-Factors-of-Migration-Thet.pdf.

Es importante reconocer que estos factores están interconectados y pueden influir en los flujos migratorios de diferentes maneras. Además, la percepción de estos factores puede variar de persona a persona, dando lugar a una amplia gama de experiencias y motivaciones migratorias.

**II. Objetivos: específicos del trabajo y del grupo**

**II.1 General:**

Implementar un pipeline que permita explorar, analizar y predecir los flujos migratorios de los próximos años en Argentina.

**II.2 Específicos del trabajo**

Diseñar un pipeline que permita monitorear el ciclo completo de los datos de flujos de migración en Argentina del año 1960 al 2021.

Desarrollar procesos automatizados en carga incremental de datos y procesamientos intermedios para el análisis estadístico, visualizaciones y machine learning.

Producir una base de datos para realizar análisis estadístico, visualizaciones y machine Learning, y automatizar su carga.

Comprender el comportamiento o mecánica de los flujos migratorios.

Brindar KPIs útiles para entender el fenómeno y tener una idea de control para tomar decisiones.

**II.3 Específicos del grupo**

Aprender desde 0 tecnologías para implementar y diseñar pipelines.

Fortalecer las habilidades técnicas de automatización de subprocesos en pipeline.

Fortalecer las habilidades blandas y de comunicación bajo presión.

Afianzar y perfeccionar conocimientos aprendidos en la carrera de Data Science.

**III. Alcance**

Diseñar un pipeline usando tecnología cloud computing que permita automatizar el ciclo de dato de los flujos migratorios en argentina del año 1960 al 2022 para explicar el fenómeno de desplazamiento de personas, comprender, crear métricas de monitoreo y KPIs que permitan tomar decisiones a un gobierno.

Acciones a realizar:

* Carga inicial en ambiente cloud.
* Automatización de carga incremental y limpieza de datos en base de datos Google Cloud Storage.
* Automatización del envío de datos desde nuestro Datalake hacía nuestro Datawherehouse en BigQuery.
* Análisis de datos (en ambiente cloud).
* Machine learning (en ambiente cloud).
* Visualización de datos (en ambiente cloud).

**IV. Almacenamiento de código**

Repositorio de Github: <https://github.com/RomanAichino/PF_Migraciones_Grupo_2>

**V. Solución propuesta:**

La solución esta diseñada para ser implementada usando como principal plataforma Google Cloud Computing.

**V1. Stack tecnológico**

**Programación**:

Python y Jupyter: para los scripts en donde descargamos los archivos csvs que contienen nuestros datos, para posteriormente el análisis de los mismos, crear funciones que automaticen la carga incremental, el ETL y el envío de datos hacia BigQuery en forma de tabla, y para nuestro modelo predictivo de Machine Learning.

Librerías:

Pandas: hemos utilizado Pandas tanto en el EDA, como el ETL y en el modelo de Machine Learning, ya que al trabajar con datos la forma más completa y sencilla de hacerlo es a partir de dataframes.

Scikit-Learn: es una de las bibliotecas mas populares para el aprendizaje automático con Python, y en nuestro caso, hemos utilizado las funcionalidades de regresión lineal y manejo de valores faltante para construir un modelo capaz de predecir la migración neta.

Mathplotlib: esta es una librería utilizada para la creación de graficos, decidimos elegirla porque todos la conocíamos y el uso que le dimos fue durante el EDA.

**Visualización**:

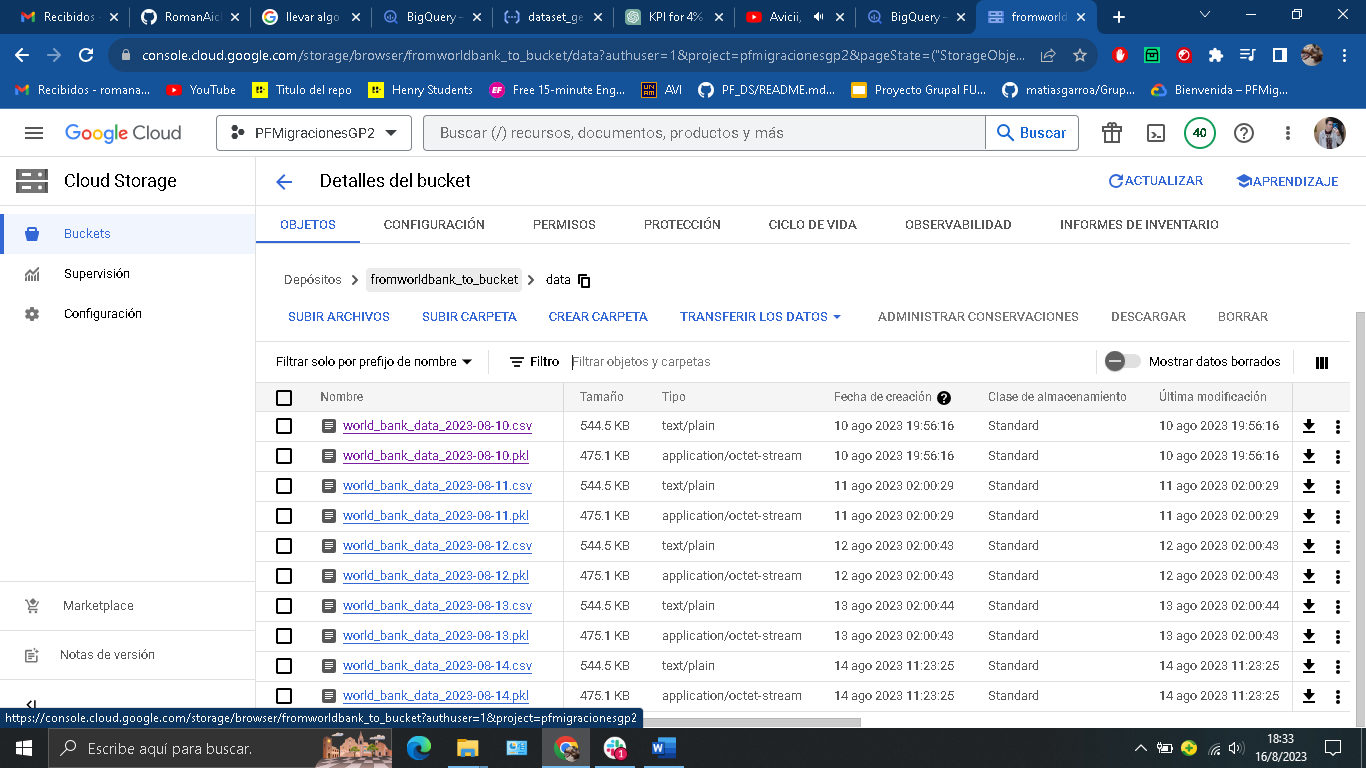
Jupyter: para el EDA.

Power BI: para un análisis más exhaustivo de los datos, la creación de KPIs y el dashboard.

**Cloud** **Computing**:

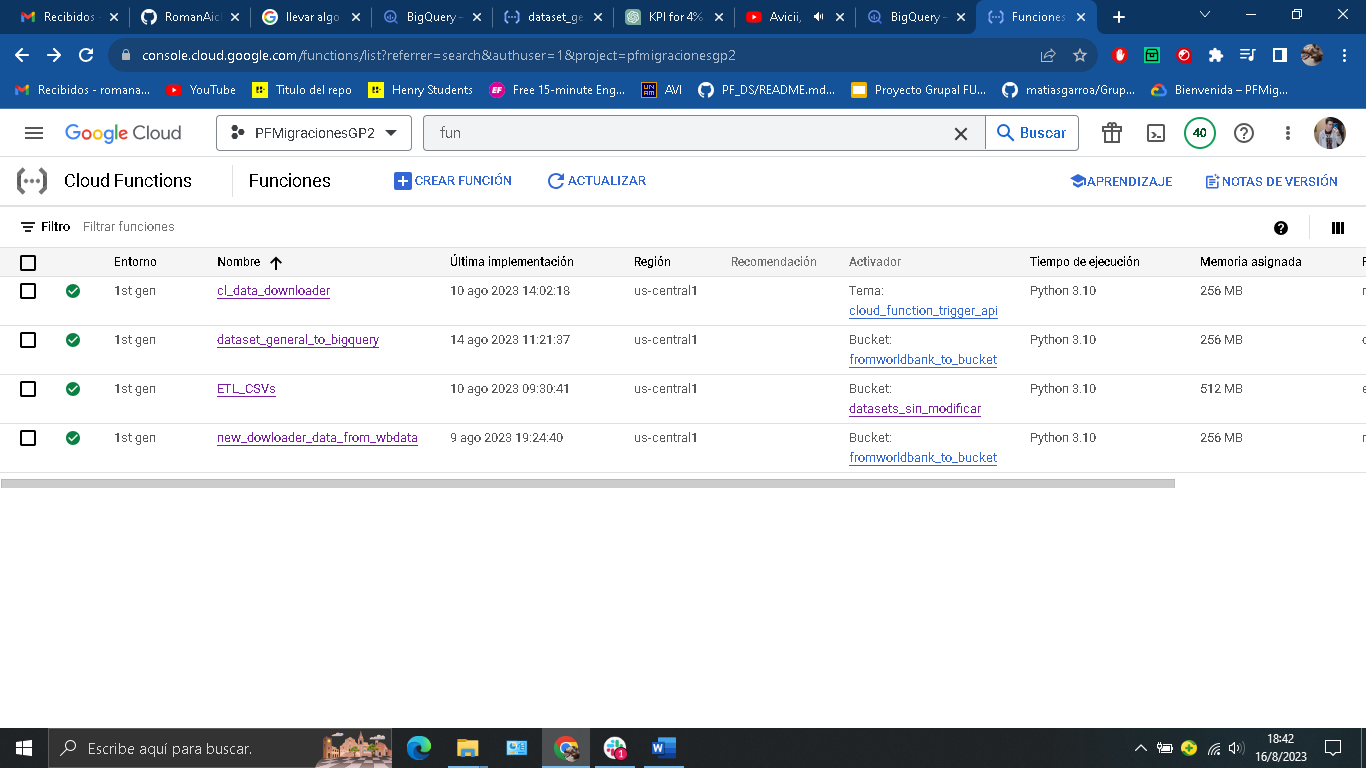
Google Cloud Storage:

Esta tecnología la utilizamos para nuestro Datalake, aquí llegan los datos en bruto para ser almacenados luego de realizarse el ETL.



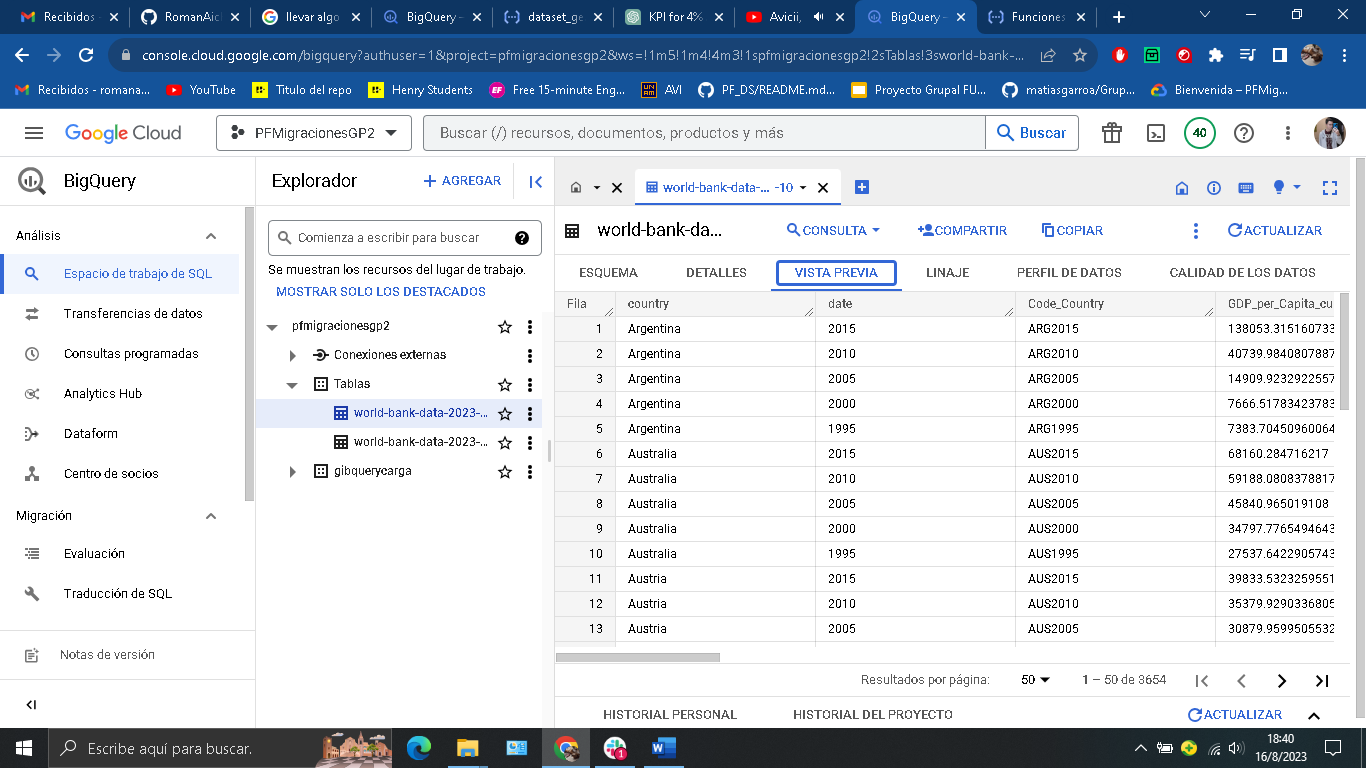
Cloud Functions:

En esta plataforma hemos creado las funciones que automatizan el ciclo completo del dato, en donde cada una hacen el trabajo de un pipeline, empezando por una primera función que realiza la extracción de datos a través de APIs, luego les realiza el ETL y finalmente lo almacena en un bucket de Google Cloud Storage; al terminar eso, inmediatamente de ejecuta la segunda función con la tarea de importar los datos del Datalake, y cargarlos en BigQuery en forma de tabla.



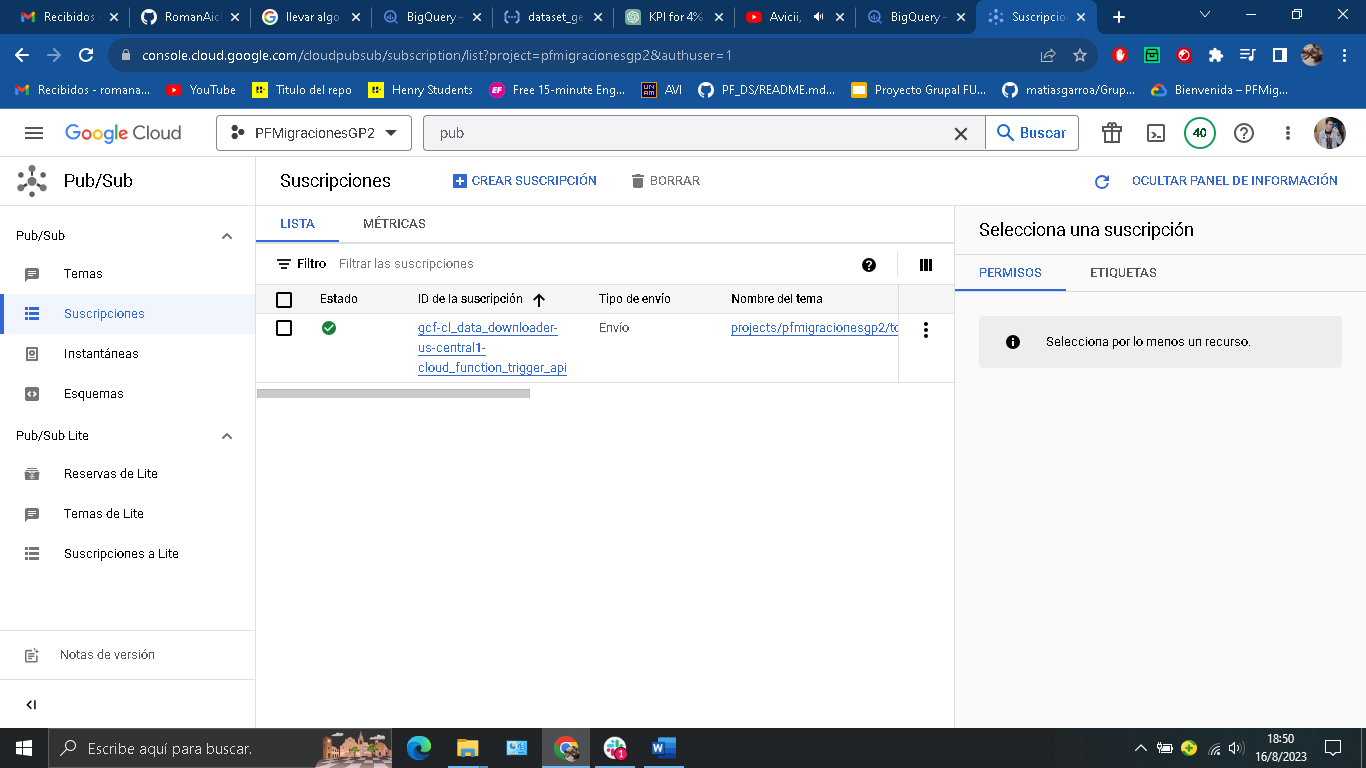
BigQuery:

Aquí, como nombre anteriormente, es donde almacenamos nuestros datos finales, en tablas que mensualmente se actualizarán. BigQuery es lo que vamos a utilizar como Datawherehouse, ya que además de tener una gran simpleza a la hora de aprender a usarlo, también cuenta con SQL como lenguaje de consulta, lo que lo hace sencillo para las personas que estudiamos Ciencia de datos.



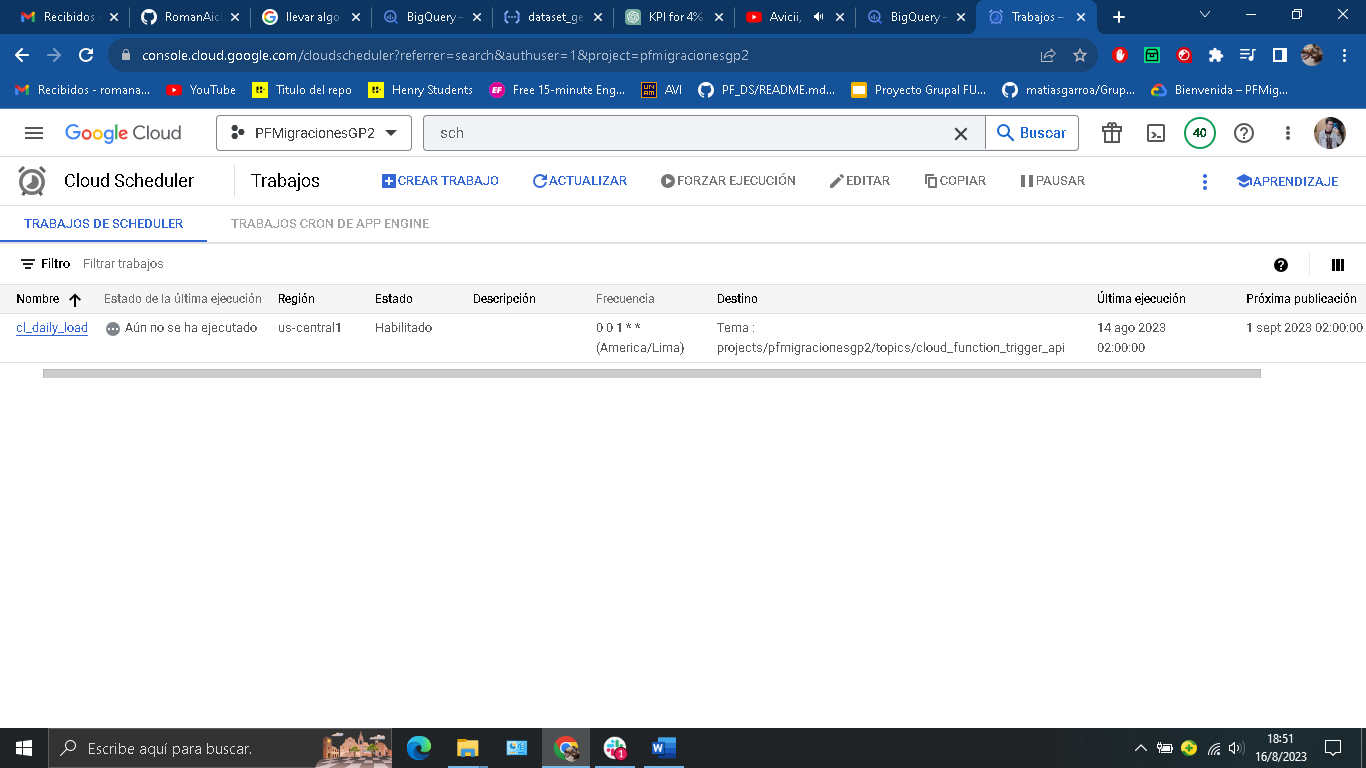
Pub Sub:

Lo utilizamos para establecer una comunicación entre los microservicios.



Cloud Scheduler:

Con este servicio configuramos la activación de la primera función de Cloud Functions, haciendo que se ejecute el primer día de cada mes ya que no sabemos con exactitud cuando habrá actualizaciones de los datos en las páginas, pero si sabemos que suele ser una actualización por año.



**V2. Metodología de trabajo**

Para el trabajo nos dividimos en fases, en donde hacíamos las cosas algunas veces por separados y otras todos juntos.

Fase 1: Indagación en los datos para decidir a qué país o región íbamos a enfocarnos.

Fase 2: Comenzamos a bajar manualmente los datos para ver que teníamos y analizar que podíamos hacer con ellos.

Fase 3: Empezamos a indagar en Google Cloud Platform, para ver que herramientas tenía y decidir cuales íbamos a usar durante el transcurso del trabajo.

Fase 4: Determinamos los datasets a utilizar y comenzamos a realizar el ETL de los mismos en Cloud Functions.

Fase 5: Determinamos métricas e iniciamos los KPIs, aunque más adelante serían modificados.

Fase 6: Automatización de la carga incremental.

Fase 7: Automatización del traspaso de datos al Datawherehouse.

Fase 8: Finalizar los KPIs y volcarlos en un dashboard interactivo.

Fase 9: Creación de un modelo de Machine Learning que pueda predecir las futuras olas migratorias.

Fase 10: Almacenar todo en un repositorio de Github.

**V3. Roles y responsabilidades**

Roman Aichino: Data Engineer

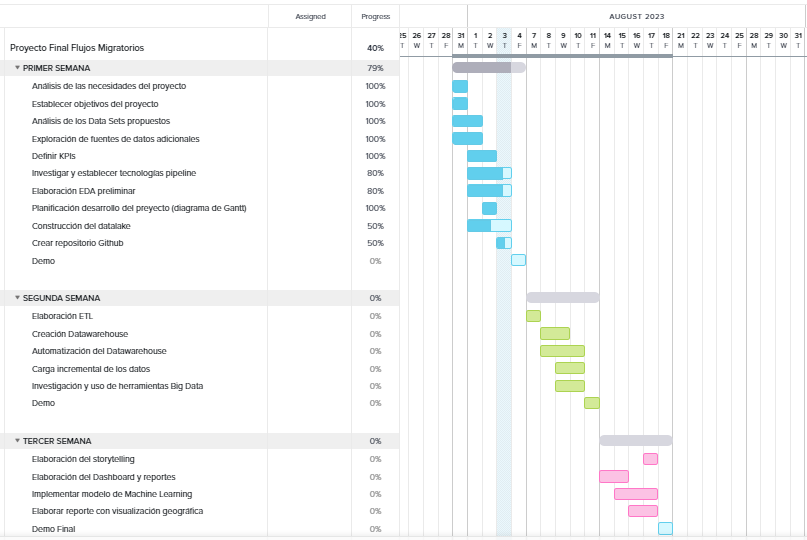
Santiago Babsia: Data Engineer – Machine Learning

Max Jeffer: Data Analytics – Data Engineer

Gabriel Guiffrida: Data Analytics – Machine Learning

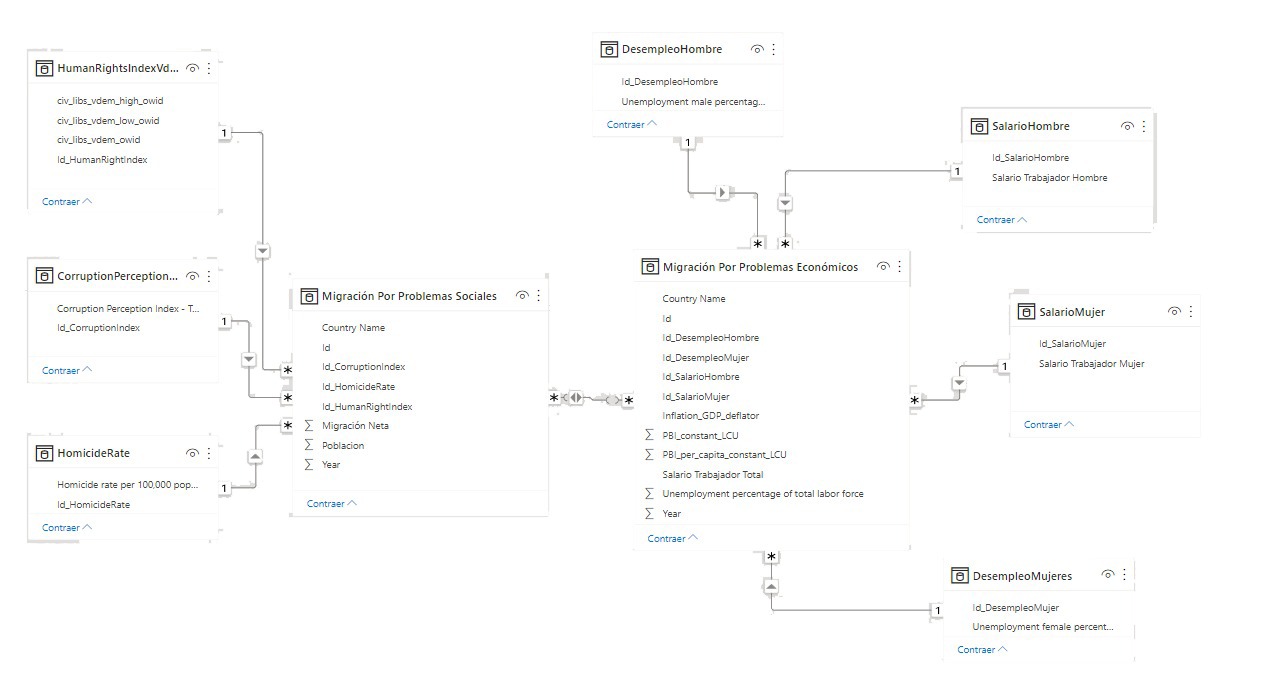
Monserrat Castillo: Data Analytics

**V4. Cronograma de Gantt**



Sprint 2: Diseño de modelo de tablas relacionales, automatización de carga, descarga y actualización de base de datos en ambiente cloud, automatización de normalización de base de datos en ambiente cloud.

I. Diseño de modelo ER (Entidad Relación)



II. Automatización pipeline

Automatización de carga, descarga y actualización de base de datos en ambiente cloud:

Permítanme presentarles la fascinante técnica de carga incremental de datos, un proceso que optimiza el acceso a información actualizada y valiosa. A través de la utilización de diversas APIs, estamos en condiciones de extraer los datos que necesitamos. Este proceso está cuidadosamente programado mediante Google Cloud Function, lo cual nos brinda un flujo continuo de información relevante.

Imagínense, cada mes, este proceso se activa de manera automatizada gracias a Google Schedule. Esto significa que siempre estamos al tanto de los datos más recientes, lo que a su vez nos permite tomar decisiones basadas en información actualizada y precisa.

Una vez que los datos son obtenidos, son dirigidos de forma automática hacia un almacenamiento seguro: nuestro bucket. La gestión de este bucket también está programada a través de una función en Google Cloud, la cual crea las tablas necesarias con los datos preparados y listos para ser utilizados en Google BigQuery.

Este proceso no solo nos brinda datos siempre actualizados, sino que también optimiza nuestra eficiencia al automatizar cada paso del camino, desde la extracción inicial hasta la preparación y almacenamiento en BigQuery. Aprovechamos la tecnología para hacer que la gestión de datos sea más inteligente y eficaz.

lll. Diccionario de datos

country: país al que se refiere la fila completa.

date: fecha.

Code\_Country: código del país al que refiere la fila completa.

GDP\_per\_Capita\_current\_US\_dollars:

Inflation\_consumer\_Prices\_Annual:

Inflation\_GDP\_Deflator\_Annual:

GDP\_Constant\_LCU:

GDP\_per\_Capita\_Constant\_LCU:

Personal\_Remittances\_Received\_of\_GDP:

Population\_Total:

Unemployment\_Total\_of\_Total\_Labor\_Force:

Net\_Migration:

Unemployment\_Male\_of\_Male\_Labor\_Force\_Nat\_Estimate:

Unemployment\_Female\_of\_Female\_Labor\_Force\_Modeled\_ILO\_Estimat:

Wage\_and\_Salaried\_Workers\_Total\_of\_Total\_Employment\_Modeled\_ILO\_Estimate:

Wage\_and\_Salaried\_Workers\_Male\_of\_Male\_Employment\_Modeled\_ILO\_Estimate:

Wage\_and\_Salaried\_Workers\_Female\_of\_Female\_Employment\_Modeled\_ILO\_Estimate:

International\_Migrant\_Stock\_of\_Population:

Sprint 3: Análisis de datos, machine learning y visualización de datos para diseño de reportes, dashboard, y verificar funcionalidad.

I. Diseño de reportes

I.1 métricas de monitoreo

I.2 KPIs

II. Dashboards

lll. Implementación modelos machine Learning

En este modelo estamos interesados en comprender cómo los cambios en la tasa de desempleo pueden afectar la migración neta en Argentina. Para abordar esta pregunta, se ha utilizado la regresión lineal, una metodología que modela la relación entre dos variables, en este caso, la tasa de desempleo y la migración neta, mediante una línea recta

El Proceso:

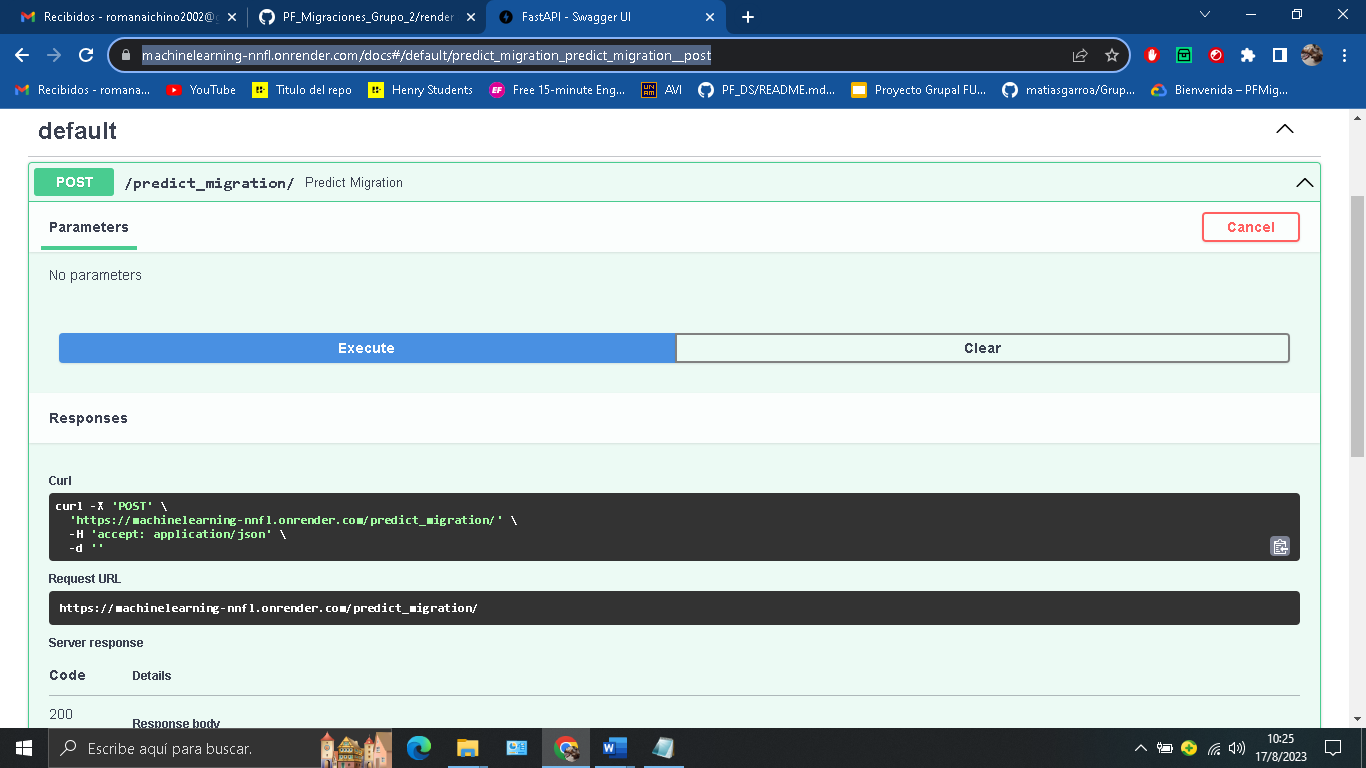
Carga y Preprocesamiento de Datos: Utilizamos Pandas para cargar los datos desde el archivo 'datos.csv'. Luego, seleccionamos solo los datos relacionados con Argentina. La tasa de desempleo es nuestra variable independiente, y utilizamos SimpleImputer para manejar los valores faltantes mediante la imputación de la media.

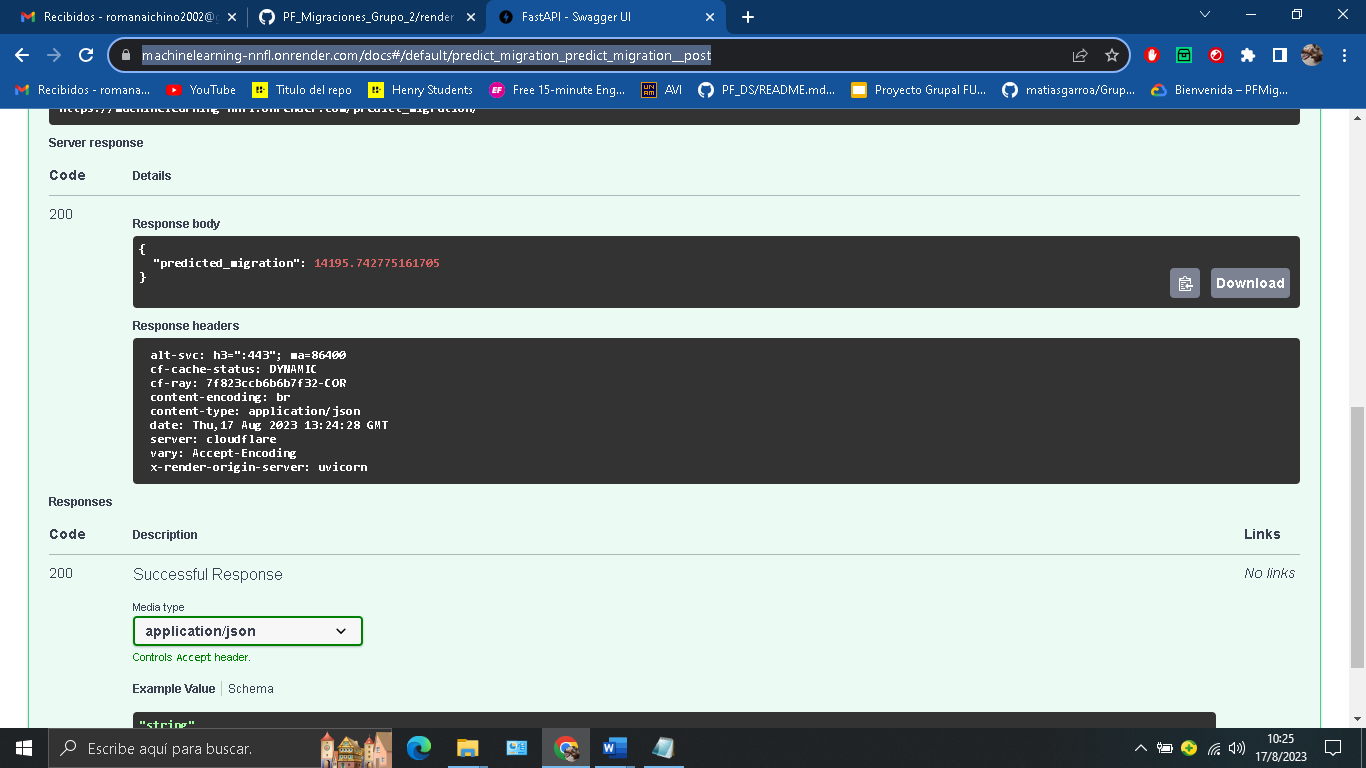
Creación del Modelo de Regresión Lineal: Hemos aprovechado la funcionalidad de regresión lineal de scikit-learn. Este modelo matemático se adapta a los datos de Argentina, capturando la relación entre la tasa de desempleo y la migración neta.

Predicción y Resultados: Supongamos que queremos predecir la migración neta en 2023. Introducimos una tasa de desempleo proyectada y dejamos que el modelo haga el trabajo. La migración neta esperada para 2023 se almacena en ‘predicted\_migration\_2023.’

El modelo esta subido a render:

Link: <https://machinelearning-nnfl.onrender.com/docs#/default/predict_migration_predict_migration__post>





CONCLUSIONES

RECOMENDACIONES