MODELO DE ARQUITECTURA

Empresa Farmacéutica

PROYECTO: Análisis de la Esperanza de Vida al Nacer para Factibilidad de lanzamiento de multivitamínico a nivel Global

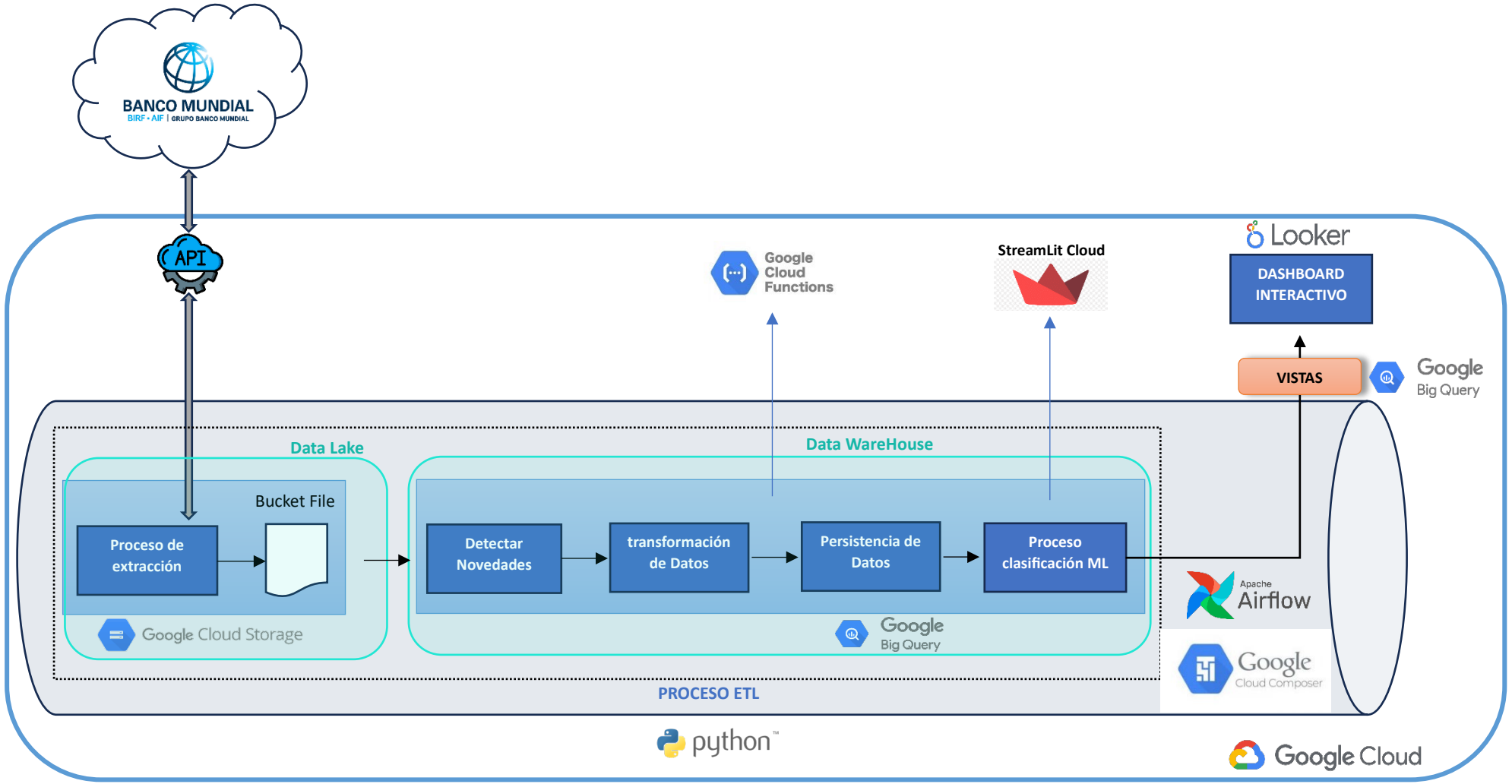
---

1. Introducción

Este documento describe la arquitectura tecnológica propuesta para el proyecto de análisis de esperanza de vida y factores determinantes en diferentes regiones del mundo. Este proyecto, encargado por una empresa farmacéutica líder, se centra en recopilar, procesar y analizar datos para respaldar decisiones estratégicas en el lanzamiento de productos.

La arquitectura tecnológica proporciona la infraestructura clave para gestionar datos a gran escala, realizar análisis avanzados y garantizar la disponibilidad de información relevante. Este modelo se enfoca en la eficiencia, escalabilidad y confiabilidad del proceso.

A lo largo de este documento, presentaremos los componentes clave de la arquitectura, incluyendo tecnologías, procesos y la infraestructura en la nube utilizada. El objetivo es proporcionar una plataforma sólida y escalable para la toma de decisiones informadas en la industria farmacéutica.


BANCO MUNDIAL
BIRF - AIF | GRUPO BANCO MUNDIAL
API
Google Cloud Functions
StreamLit Cloud
Looker
DASHBOARD INTERACTIVO
VISTAS
Google Big Query
Data Lake
Bucket File
Proceso de extracción
Google Cloud Storage
Data WareHouse
Detectar Novedades
transformación de Datos
Persistencia de Datos
Proceso clasificación ML
Google Big Query
Apache Airflow
Google Cloud Composer
PROCESO ETL
python
Google Cloud

2. Detalle del Stack

1. **Proceso Consumidor de API**: Se creará un proceso que consuma la API del banco mundial y descargue los datos de los factores seleccionados. Este proceso generara unos archivos en formato Bucket para subir a **Google Cloud Storage**, que vendría a ser como nuestro **DATA LAKE**.
2. Para poder persistir los datos en Google Big Query, se ejecutarán los siguientes procesos:
    a. **Detección de novedades**: Proceso que detectara las novedades a cargar, favoreciendo la carga incremental y evitando la carga completa de los datos.
    b. **Transformación de Datos**: Se aplicarán las transformaciones necesarias para adecuarnos a las estructuras de datos de nuestro Warehouse.
    c. **Persistencia en Warehouse**: Utilizaremos Google Big Query como Data Warehouse, en donde persistiremos los datos estructurados y modificados.

- Los puntos 1 y 2 constituyen la fase de ETL, que abarca desde la extracción de los datos hasta su transformación y, finalmente, su almacenamiento en el Warehouse. Sin embargo, para concluir el proceso de ETL, se llevará a cabo la ejecución de nuestro Modelo de Machine Learning (Modelo de Clasificación Binaria). Este modelo clasificará las distintas regiones para determinar la viabilidad del lanzamiento del producto, y se reentrenará si es necesario. Todos estos procesos serán encadenados por medio de DAGs haciendo uso de Google DataFlow.

3. Por último, se creará un Dashboard Interactivo en Looker, haciendo uso de los datos persistidos en el WareHouse.

De esta manera, el STACK TECNOLOGICO a utilizar pertenece en su totalidad a GOOGLE CLOUD PLATFORM.

❖ Por otra parte, nuestro lenguaje principal de programación será Python.

3. Motivos de la elección Stack

Los siguientes motivos nos llevaron a la selección del STACK TECNOLOGICO indicado antes:

1. GCP tiene una gran simplificación de la curva de aprendizaje y configuración. A diferencia de algunos competidores, como AWS o Azure, GCP ofrece una experiencia más amigable para los usuarios, permitiéndonos ahorrar tiempo y recursos en la implementación de nuestras soluciones de datos y análisis. Esto se traduce en una mayor eficiencia y rapidez en el desarrollo de nuestro proyecto.
2. Posee una robusta infraestructura, escalabilidad y confiabilidad. GCP proporciona la capacidad de manejar grandes volúmenes de datos de manera eficiente y brinda herramientas avanzadas para el procesamiento, almacenamiento y visualización de datos."
3. Ofrece una integración perfecta con las herramientas de análisis y procesamiento de datos que necesitamos para nuestro proyecto, como Dataflow para el procesamiento de datos en tiempo real, BigQuery para nuestro WareHouse y Looker para crear paneles de control interactivos."
4. Dispone de amplio conjunto de servicios especializados en datos y análisis, lo que nos permite abordar los desafíos complejos de nuestro proyecto de consultoría de datos de extremo a extremo. Además, su enfoque en la seguridad de los datos nos brinda tranquilidad para proteger la información confidencial."
5. Google Cloud Platform no solo cumple con nuestras necesidades actuales, sino que también nos proporciona la flexibilidad necesaria para adaptarnos a futuras demandas y expansiones del proyecto. Su comunidad activa y soporte técnico nos brindan la asistencia requerida para garantizar el éxito de nuestro proyecto de datos.

4. Modificaciones del Stack En c/ Sprint

Sprint #2 – Data Engineering

Durante la etapa de Data Engineering, nos hemos encontrado con costos elevados en la plataforma GCP. Con el fin de bajar dichos costos a la menor expresión posible, sin perder escalabilidad y facilidad de adaptación del proceso a futuros cambios, hemos decidido realizar los siguientes cambios en el Stack:

- No usaremos DataFlow. Para orquestar nuestro Pipeline utilizaremos **Airflow** y **Google Composer**.
- Utilizaremos **Cloud Functions** para cada una de las tareas a realizar, las cuales estarán codificadas en lenguaje Python. La ventaja de los Cloud Functions es su bajo costo, dado que se cobra solo por su uso. Además, ofrece facilidades para hacer pruebas unitarias y aisladas de cada función, sin tener que correr todo el pipeline. Es rápido visualizar los logs y favorece la escalabilidad.

Sprint #3 – Data Analitycs & MLOps

Durante la etapa de Analitycs y MLOps, hemos completado nuestra arquitectura para poder dar al usuario una herramienta funcional y operativa. Los cambios que realizamos para lograrlo son:

- Agregamos una capa de Vistas en BIG QUERY para quitarle complejidad al DashBoard.
- Realizamos el deploy del modelo de Machine Learning en streamlit Cloud.
- Además, el modelo de Machine Learning fue implementado en el pipeline de datos, para que pueda ser consumido por el DashBoard.