DOCUMENTO DE DATA ENGINEERING

Empresa Farmacéutica

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

---

Contenido

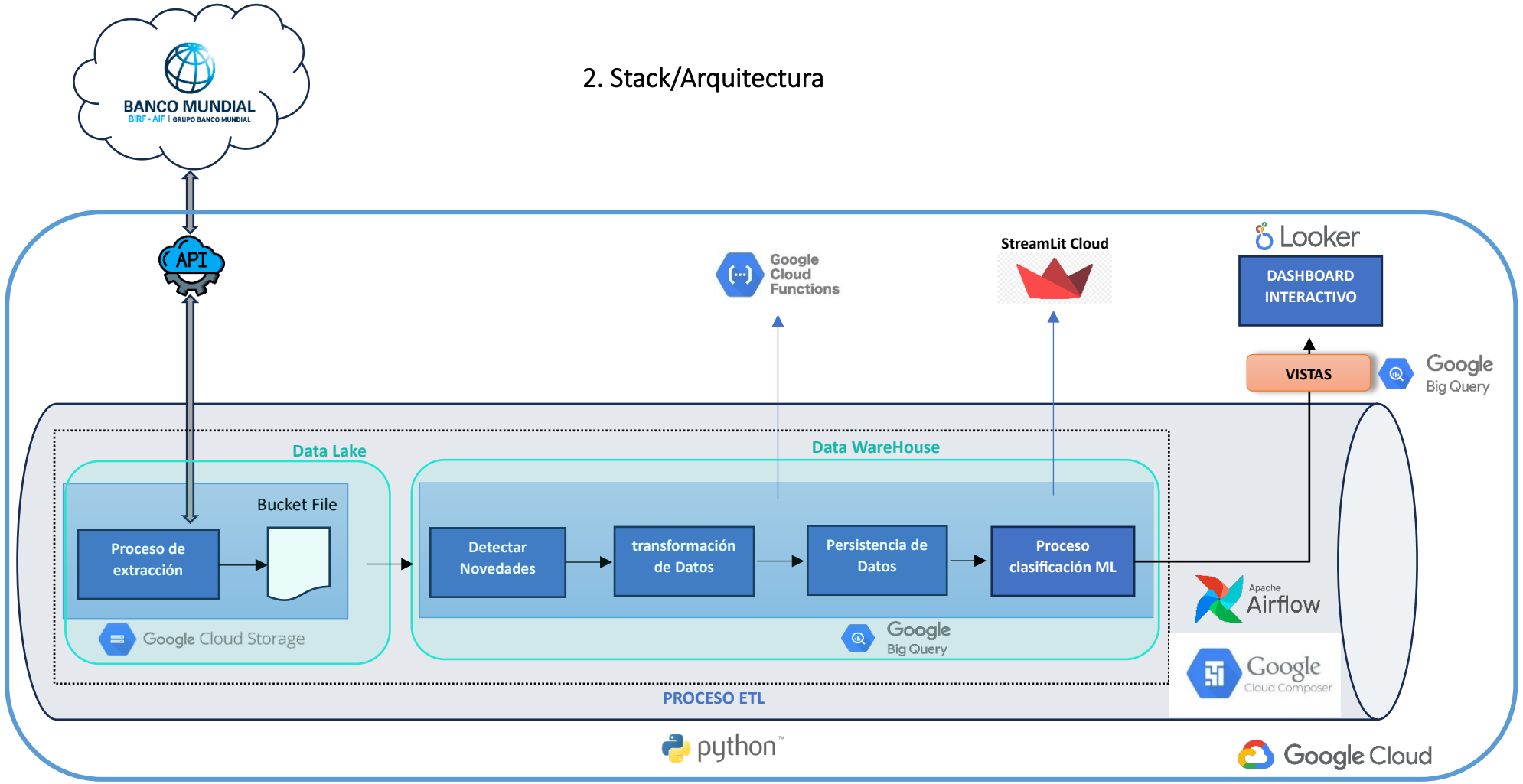
1. Introducción

Este documento presenta la infraestructura tecnológica desarrollada para automatizar el proceso ETL en el marco de un proyecto dedicado a las necesidades de una destacada empresa farmacéutica. En consonancia con nuestro éxito previo en el análisis de esperanza de vida y factores determinantes en diversas regiones del mundo, este proyecto se centra en la mejora de la extracción, transformación y carga de datos, respaldando así las operaciones críticas de la empresa.

La presente infraestructura está meticulosamente diseñada para optimizar la eficiencia, escalabilidad y confiabilidad del proceso ETL. Haciendo uso de plataformas CLOUD para evitar la infraestructura ON PREMISE dentro de la empresa.

A lo largo de este documento, exploraremos detalladamente los componentes esenciales de este pipeline de datos, abarcando tecnologías específicas, procesos clave y la infraestructura en la nube implementada. El objetivo es proporcionar una visión integral de la estructura subyacente que respalda la automatización del proceso ETL, estableciendo así una base sólida para la toma de decisiones informadas en el ámbito de la gestión de datos en la industria farmacéutica.

2. Stack/Arquitectura



3. Cloud Storage – Inicio del Proceso

EL proceso ETL inicia en Cloud Storage, en donde disponemos de **3 buckets** que serán utilizados como contenedores de almacenamientos para diferentes archivos durante toda la ejecución del Pipeline:

- **pf-henry-esperanza-parametros**:

  Este bucket será utilizado para recepcionar los archivos que serán tomados de input en el proceso. El proceso espera los siguientes archivos:

  - Parametros_Paises.csv: Archivo obligatorio. Contendrá la lista de Países (por medio de su código de 3 letras) a procesar. El proceso realizara una carga incremental de los mismos en el DW.
  - Parametros_Indicadores.csv: Archivo obligatorio. Contendrá la lista los indicadores del Banco Mundial a procesar. El proceso realizara una carga incremental de los mismos en el DW.
  - Parametros_Indicador_Rentabilidad.csv: Archivo Opcional. Contendrá la lista los indicadores del Banco Mundial a procesar. El proceso realizara una carga incremental de los mismos en el DW.

- **pf-henry-esperanza-archivos-intermedios**:

  Este bucket será utilizado para recepcionar los archivos intermedios que son generados por procesos del Pipeline y que serán utilizados posteriormente por otros procesos del mismo workflow.

- **pf-henry-esperanza-archivos-mlops**:

  Este bucket será utilizado para recepcionar los archivos utilizados y generados por el proceso de MLOps para el modelo de Clusterizacion.

- **pf-henry-esperanza-respaldos:**

  Este bucket será utilizado para almacenar el bkp de los archivos de input y los intermedios. Modificará el nombre de los mismos para agregar la fecha de proceso antes de la extensión de los mismos.

4. Cloud Functions

Todos los procesos que ejecutan en el Dag son **cloud functions** de GCP. Tienen la gran ventaja de solamente consumir recursos cuando ejecutan, bajando el costo cobrado por la plataforma. Además son muy escalables y fáciles de probar unitariamente. A continuación detallamos todas las funciones presentes en el proceso:

verificar_archivos

Es el primer Proceso del DAG. Verifica que estén en el bucket correspondiente los archivos de input.

eliminar_duplicados_parametros

Elimina los registros duplicados de los archivos de parámetros y los reescribe en el mismo bucket.

cargar_categorias

En base a los indicadores a procesar, realiza la carga de las categorías a las que pertenecen los indicadores en la tabla correspondiente de bigquery.

cargar_indicadores

Una vez cargadas las categorías, le asigna un Id de categoría a cada indicador y los carga en su tabla correspondiente de bigquery.

cargar_continente

Realiza la carga de continentes en la tabla correspondiente de bigquery.

extraer_paises

En base a los países a procesar indicados en el input, realiza una extracción de datos complementarios desde la API del banco mundial, con el objetivo de tener mas datos referentes a los países. Genera un archivo intermedio que será procesado por el siguiente proceso.

cargar_paises

Toma el archivo intermedio del proceso anterior, y realiza la carga de los países y sus datos en la tabla de Bigquery correspondiente.

extraer_datos_BM

Realiza la extraccion de los datos de los indicadores, países y años seleccionados del banco mundial de datos. Genera un archivo intermedio que será utilizado por el siguiente proceso.

transformar_columnas_a_registros_BM

Toma el archivo del proceso anterior, el cual tiene a los indicadores a nivel columna, lo que es incómodo y difícil de persistir en bigquery. Es por eso que este proceso generara un archivo intermedio con los datos del proceso anterior pero pasando los indicadores a nivel registro.

transformar_imputar_BM

Este es el proceso donde se aplican las transformaciones e imputaciones a los datos. Las decisiones de estas transformaciones e imputaciones están basadas en el EDA Exhaustivo de datos.Como resultado del proceso, se persisten los datos estructurados en la tabla correspondiente de bigquery.

cargar_indicador_rentabilidad

Realiza la carga de la tabla de los Indicadores de Rentabilidad. Estos indicadores serán utilizados para clasificar a los países, estableciendo si son o no candidatos a ser rentables para el lanzamiento del multivitamínico. Dichos valores son los que utilizara el modelo de Clasificación Binaria de Machine Learning.

generar_data_ML

Genera un archivo intermedio que será utilizado por el modelo de Machine Learning.

clusterizar_paises

Tomando el archivo del paso anterior, realiza una clusterizacion de los países, utilizando determinados indicadores, y genera un archivo intermedio con el cluster de los países con mayor posibilidad de ser rentables para el lanzamiento del multivitamínico.

imputar_rentabilidad

Tomando el archivo del paso anterior, realiza la persistencia en BigQuery de los países rentables. Para esto hace uso del Indicador de Rentabilidad presente en la tabla de País.

respaldar_archivos

Este proceso toma los archivos de input e intermedios y los mueve al bucket de respaldo a modo de backup. Anexa al nombre de los mismos la fecha de proceso.

mostrar_estadisticas

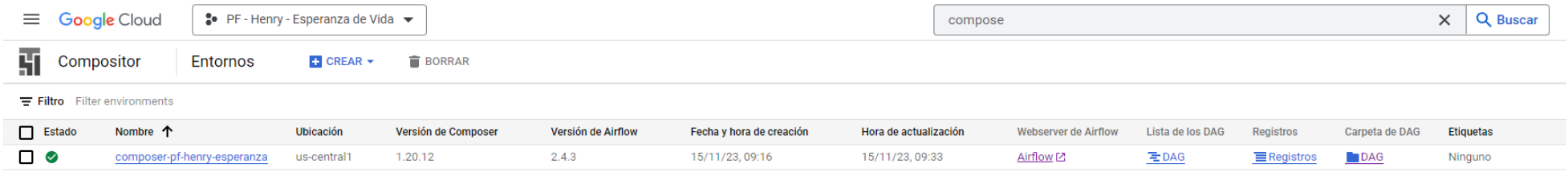
Final del proceso ETL, muestra por log una estadística de todo el impacto generado en el DW.

Cabe destacar que todos los procesos cumplen con los siguientes puntos:

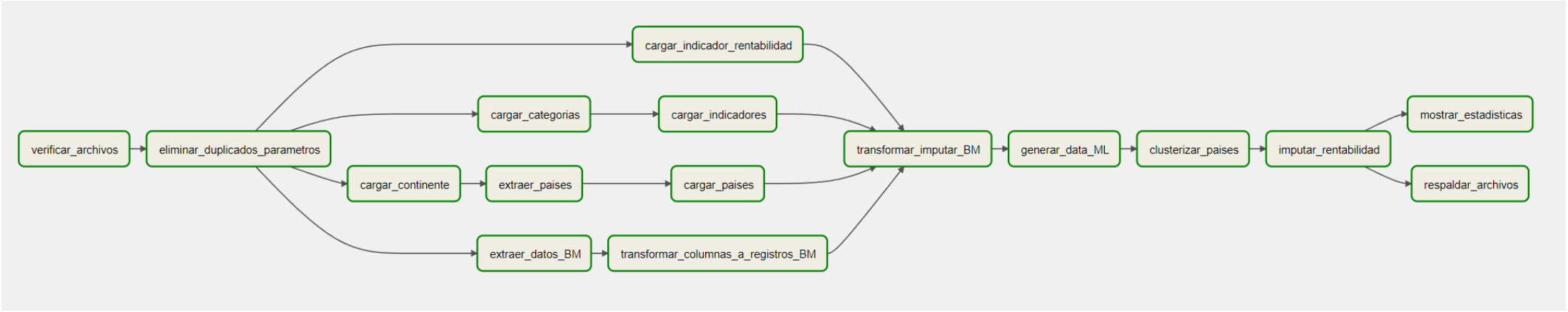
- ✓ Están optimizados para volúmenes de datos grandes
- ✓ Están preparados para cargas incrementales
- ✓ Generan datos de auditoria en una tabla independiente de bigquery, a modo de tener registro de que hizo cada proceso con las tablas del modelo.
- ✓ Genera Backups de los archivos procesados.
- ✓ Las cloud function son invocadas por medios de Bash Operators en el DAG.

5. Instancia Composer y Dag

A continuación mostramos la instancia de Composer:

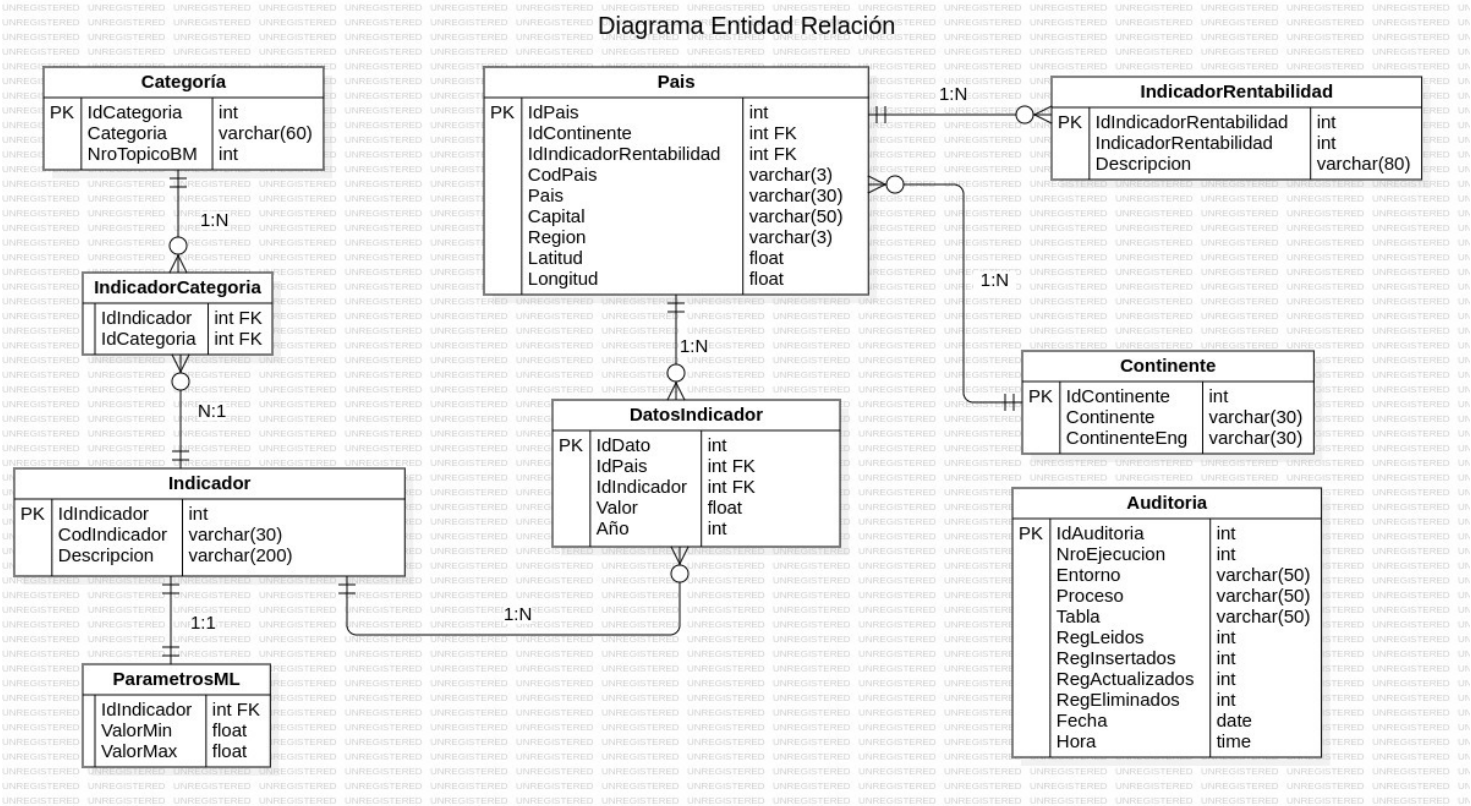


Dicha instancia tiene cargado el siguiente DAG, encargado de Orquestar todas las funciones mencionadas en el punto anterior:



6. BigQuery

Utilizamos big query como Data Warehouse, pero tratamos de simular una base de datos relacional a pesar de no serlo. Con esto nos referimos a asegurar la integridad referencial de los datos, generar claves primarias y foráneas. Todo esto es logrado a través de las funciones ejecutadas en el DAG. El modelo de entidad relación es el siguiente:



En los documentos de Diccionario de Datos y Modelo Entidad Relación se podrá encontrar mucha más información respecto a las tablas y sus campos.