ETL Proyecto:

Interpretado por:

- Antonio Cárdenas
- Luis Ángel Garcia
- Juliana Toro Serrano

Análisis Integrado de Tabaquismo y Cáncer de Pulmón Mediante Tecnologías de Procesamiento de Datos

Introducción

El cáncer de pulmón es una de las principales causas de muerte a nivel mundial, y su estrecha relación con el consumo de tabaco ha sido ampliamente demostrada por la literatura médica. En este proyecto se desarrolló un sistema completo de análisis que permite estudiar el impacto del tabaquismo sobre la incidencia y mortalidad por cáncer de pulmón a través de datos integrados provenientes de diversas fuentes. Para ello, se aplicaron herramientas como Kafka, PostgreSQL, Airflow, Docker, y visualizaciones con Streamlit.

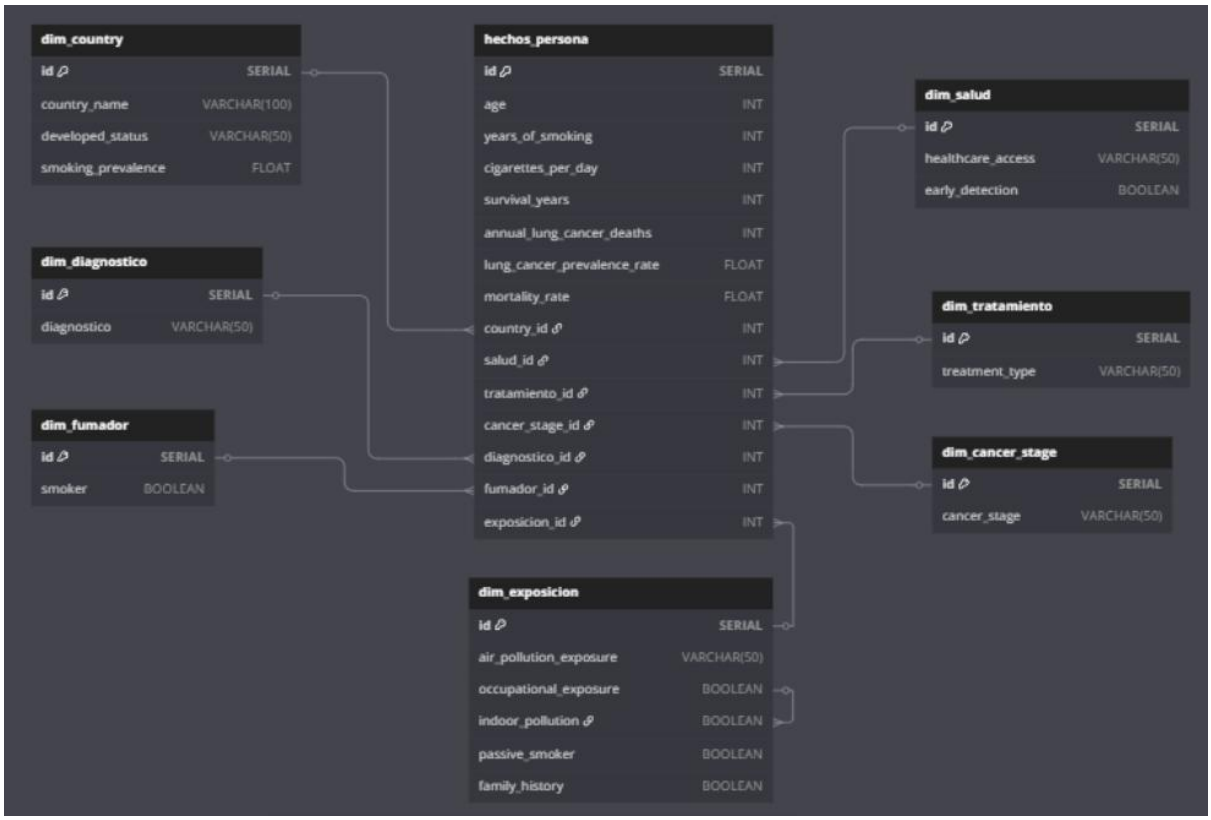
1. Extracción de Datos desde la OMS

La primera fase del proyecto consistió en la recolección de datos relevantes sobre la prevalencia del tabaquismo a través de la API pública de la Organización Mundial de la Salud (OMS).

Este endpoint proporciona el indicador **WHOSIS_000001**, que representa el porcentaje estimado de adultos fumadores por país y por año. A partir de esta fuente, se diseñó una función en Python que realiza lo siguiente:

- Establece conexión con la API mediante una solicitud HTTP.
- Filtra los registros obtenidos para conservar solo aquellos con campos completos y cuya fecha sea posterior a 2018.
- Extrae los valores claves: `SpatialDim` (código del país) y `Value` (porcentaje de fumadores adultos).
- Almacena estos datos limpios en un archivo CSV para su posterior uso.

2. Diseño del Modelo Dimensional



Se diseñó un modelo de datos basado en el paradigma de almacenamiento dimensional, ampliamente utilizado en sistemas analíticos. Este modelo permite estructurar la información en torno a una tabla de hechos principal (`hechos_persona`) y diversas tablas de dimensiones (`dim_country`, `dim_salud`, `dim_tratamiento`, `dim_diagnostico`, `dim_estadio`, `dim_exposicion_ambiental`, `dim_fumador`).

La tabla de hechos almacena registros individuales de personas diagnosticadas con cáncer de pulmón, incluyendo variables como años de consumo, cantidad diaria de cigarrillos, tasas de prevalencia y de mortalidad. Las tablas dimensionales contienen información contextual que enriquece el análisis y permite segmentar los datos por país, estado de salud, tipo de tratamiento recibido, y otras variables clínicas y ambientales.

Esta arquitectura nos facilitó la consulta, exploración y análisis de los datos, al mismo tiempo que nos permitió escalar el sistema para incluir más variables y casos en el futuro.

3. Implementación de un Sistema de Comunicación en Tiempo Real con Kafka

Una de las innovaciones del proyecto es la implementación de un canal de comunicación basado en Kafka, para nuestro procesamiento de flujo en tiempo real. Se desarrollaron dos scripts principales:

3.1 Productor (`producer.py`)

Este componente:

- Se conecta a la base de datos PostgreSQL.
- Ejecuta una consulta SQL que calcula el promedio de variables relevantes (años fumando, cigarrillos por día, tasas de prevalencia y mortalidad) por país.
- De entre los países disponibles, selecciona uno al azar.
- Envía estos datos estructurados como JSON al tópico `lung_cancer_metrics` de Kafka.

El objetivo del productor es generar mensajes periódicos que representen indicadores clave para cada país, simulando así un flujo continuo de métricas agregadas.

3.2 Consumidor (`consumer.py`)

Este componente:

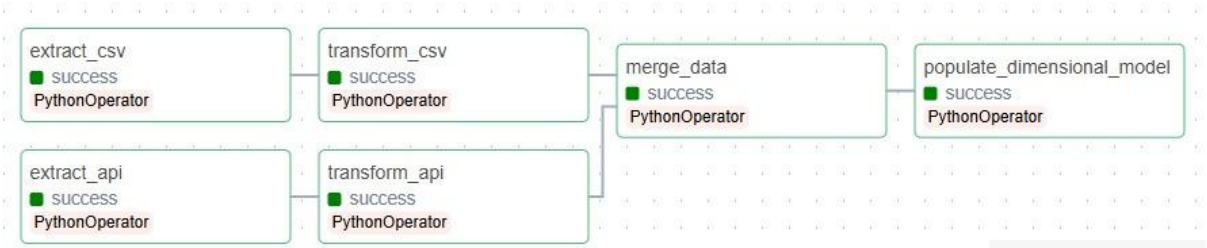
- Se suscribe al tópico `lung_cancer_metrics`.
- Recibe y decodifica los mensajes JSON.
- Muestra por consola la información recibida, incluyendo país, condición de desarrollo, y valores promedio de tabaquismo y cáncer.

El consumidor actúa como receptor de las métricas en tiempo real, lo cual permite monitorear de forma continua las estadísticas más relevantes del sistema.

4. Orquestación y Contenerización con Docker y Airflow

Para facilitar la ejecución, despliegue y mantenimiento del entorno, se utilizó **Docker**, permitiendo levantar todos los servicios necesarios en contenedores independientes: PostgreSQL, Apache Kafka, Airflow y la aplicación de Streamlit. Esto nos permitio replicar fácilmente el sistema en cualquier máquina.

Además, se integró **Apache Airflow** para programar y monitorizar las tareas ETL (Extracción, Transformación y Carga) del sistema. Este orquestador asegura que las dependencias entre procesos se ejecuten en el orden correcto y de forma automatizada.



5. Visualización de Resultados con Streamlit

Para democratizar el acceso a la información y facilitar la exploración de los resultados, se implementó un tablero interactivo usando Streamlit. En esta herramienta se visualiza:

- Mapas por país con tasas de prevalencia de tabaquismo.
- Comparaciones entre países desarrollados y en vías de desarrollo.
- Gráficos de dispersión y líneas de tendencia de tabaquismo vs. mortalidad.
- Filtros por variables demográficas y clínicas.

6. Impacto y Aplicaciones del Proyecto

- **Investigación epidemiológica:** permite correlacionar factores de riesgo y resultados clínicos.
- **Toma de decisiones en salud pública:** apoya a gobiernos y ONGs en el diseño de campañas de prevención basadas en datos reales.
- **Modelos predictivos:** sienta las bases para entrenar modelos de machine learning que predigan incidencia y mortalidad por cáncer de pulmón.
- **Educación:** sirve como ejemplo práctico para enseñar conceptos de bases de datos, procesamiento en tiempo real, y análisis de datos en salud.

Conclusión

Este proyecto permitió conectar distintas herramientas para analizar cómo el tabaquismo influye en el cáncer de pulmón. Al integrar datos confiables y automatizar su procesamiento, logramos construir una base sólida para futuras investigaciones y visualizaciones claras. Es una muestra de cómo se puede ayudar a entender mejor problemas de salud pública.