Propuesta de Diseño de pipeline ETL para el procesamiento de datos de ventas bajo arquitectura Medallion para Voltaje S.R.L.

Alumno: Toledo, Alvaro Julian
Legajo: 52721
Cátedra: Práctica supervisada

La presente propuesta describe el diseño de un pipeline ETL (Extracción, Transformación y Carga) para preparar los datos de ventas de la empresa para su análisis. Como insumo se utilizarán las ventas correspondientes al período enero–diciembre de 2025, provistas por la empresa como fuente de datos.

Los artículos incluidos en las ventas contarán previamente con una clasificación por tipo y categoría obtenida mediante agentes desarrollados en **Python**. A partir de ese resultado, este trabajo se enfocará únicamente en el procesamiento de los datos dentro del pipeline ETL.

Alcance del pipeline ETL

El pipeline ETL tendrá como objetivo organizar y preparar los datos de ventas para su uso analítico. El flujo completo será orquestado mediante **Apache Airflow**, permitiendo definir, ejecutar y monitorear cada etapa del proceso de forma controlada.

El pipeline procesará la totalidad de las ventas del año 2025.

Objetivo general

Diseñar un pipeline ETL que permita ingerir, transformar y almacenar los datos de ventas de la empresa correspondientes al año 2025, dejándolos preparados para su posterior análisis.

Objetivos específicos

- Ingerir los datos de ventas provistos por la empresa mediante **Airbyte** y almacenarlos sin modificaciones en una base de datos PostgreSQL como capa Bronze.

- Aplicar transformaciones sobre los datos ingeridos mediante **dbt** y lógica desarrollada en **Python**, conformando la capa Silver.

- Cargar los datos transformados en un data warehouse PostgreSQL que represente la capa Gold.

- Orquestar el flujo completo del pipeline utilizando **Apache Airflow**.

- Ejecutar la infraestructura del pipeline mediante **Docker**, asegurando portabilidad y reproducibilidad.

- Dejar los datos finales preparados para su uso en algoritmos de minería de datos y análisis posteriores.

Ingesta de datos (Bronze)

La etapa de ingesta consistirá en cargar los datos de ventas en una base de datos PostgreSQL utilizada como punto de entrada del pipeline. La ingesta se realizará mediante **Airbyte**, ejecutado dentro de un entorno **Docker** de manera local, permitiendo extraer los datos desde la fuente provista por la empresa y cargarlos sin aplicar transformaciones.

Esta etapa corresponderá a la capa Bronze de la arquitectura Medallion, ya que su función es conservar los datos en su estado original como referencia del origen.

Transformación de datos (Silver)

En la capa Silver se trabajará sobre los datos provenientes de la capa Bronze. Las transformaciones se realizarán utilizando **dbt**, complementado con lógica en **Python** cuando sea necesario, para limpiar, validar y generar nuevos datos a partir de los existentes.

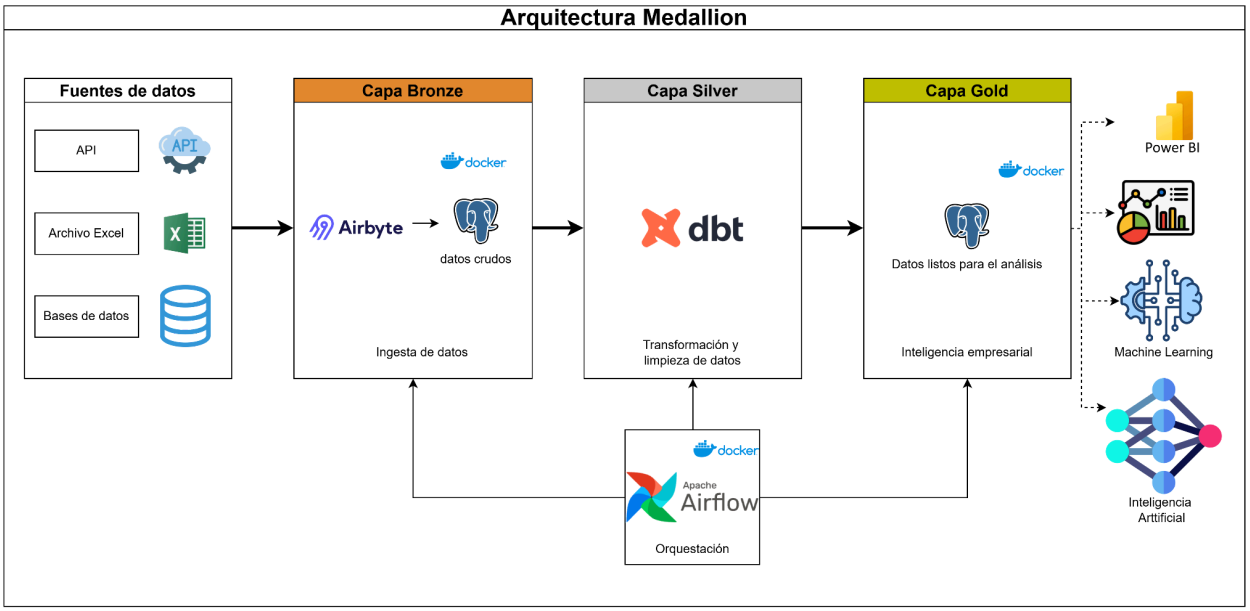
Esta capa corresponderá a Silver porque los datos dejan de ser crudos y pasan a estar procesados y enriquecidos, aunque todavía no sean el resultado final de análisis.

Carga en el data warehouse (Gold)

Los datos transformados se cargarán en una segunda base de datos PostgreSQL que funcionará como data warehouse. Esta base representará la capa Gold de la arquitectura Medallion y será alimentada a partir de los modelos finales definidos en dbt.

Esta etapa será Gold porque contendrá los datos finales, consolidados y listos para ser utilizados directamente en análisis y en la aplicación de algoritmos de minería de datos como Apriori, FP-Growth o Eclat. El desarrollo de estos análisis se abordará en una etapa posterior del proyecto.

Arquitectura de la solución propuesta



Desarrollo

El proyecto **voltaje_etl_pipeline** implementa un pipeline de datos completo utilizando Docker Compose, Apache Airflow y dbt, siguiendo una arquitectura Medallion. El objetivo principal es construir, de forma reproducible y automatizada, distintas capas de datos dentro de un Data Warehouse en PostgreSQL, sin necesidad de ejecutar comandos manuales dentro de los contenedores. El pipeline está pensado para trabajar sobre datos de ventas y clientes obtenidos del dataset generado anteriormente en el proyecto Synthetic_Sales_Generator, permitiendo distintos usos según el nivel de transformación de la información.

La infraestructura se levanta mediante un único archivo *docker-compose.yaml*, que define todos los servicios necesarios. Se utilizan dos instancias de PostgreSQL: una para la metadata de Airflow y otra para el Data Warehouse. Esta última expone el puerto hacia el host y se inicializa de forma automática con los schemas necesarios: bronze, silver, gold, datamining y silver_datamining. Además, se incluye Redis como sistema de colas para el ejecutor Celery de Airflow. Todos los servicios de Airflow utilizan una misma imagen custom, lo que simplifica el mantenimiento y evita inconsistencias entre componentes.

La imagen de Airflow se construye a partir de una imagen oficial y agrega dbt dentro de un entorno virtual aislado. Esto permite que Airflow y dbt convivan sin conflictos de dependencias. El diseño evita modificar el entrypoint original de Airflow, lo que asegura una inicialización correcta del sistema y previene errores comunes relacionados con la carga de módulos. Durante el arranque, un servicio de inicialización ejecuta automáticamente las migraciones de Airflow y crea el usuario administrador, dejando el entorno listo para su uso.

El Data Warehouse se alimenta siguiendo el enfoque Medallion. La capa **Bronze** contiene los datos crudos cargados por Airbyte desde sistemas externos y no es responsabilidad de dbt. A partir de esta capa, dbt construye las capas **Silver**, **Gold** y **Datamining**. En Silver se realizan tareas de limpieza básica, tipado de columnas y normalización mínima, sin agregar métricas. Gold contiene tablas agregadas orientadas a reporting y dashboards, mientras

que Datamining genera una tabla plana pensada específicamente para análisis de canastas de compra.

La orquestación del pipeline se realiza mediante DAGs de Airflow. El DAG principal llamado voltaje_medallion_etl.py verifica primero que las tablas Bronze existan en el Data Warehouse antes de ejecutar cualquier transformación. Una vez confirmada la disponibilidad de los datos, se ejecutan de forma ordenada los modelos de dbt para Silver, Datamining y Gold, junto con tests de calidad en la capa final. De esta manera, cada ejecución garantiza que las capas superiores se construyan siempre a partir de datos consistentes y actualizados.

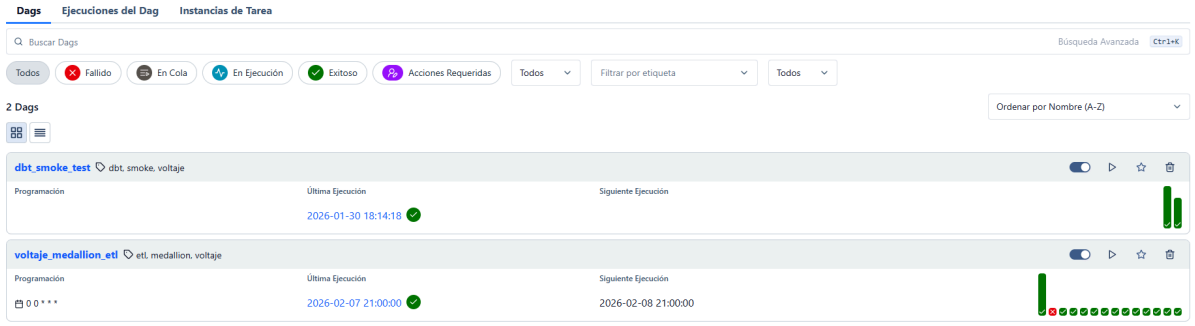


Imagen 1: DAGs generado

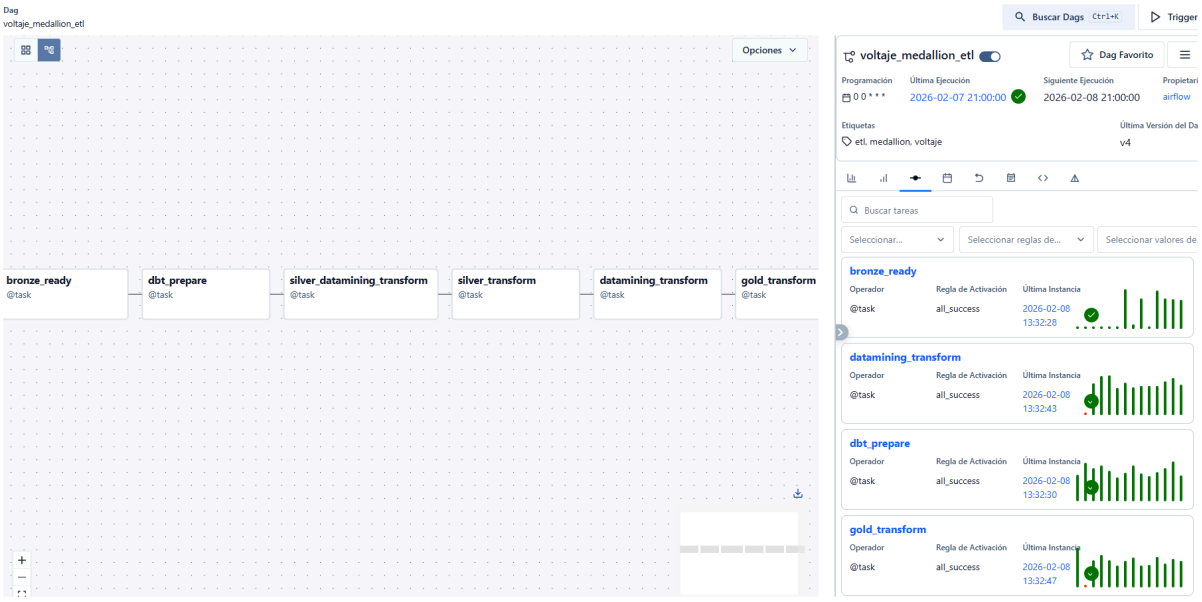


Imagen 2: Vista de DAG ejecutado desde Airflow UI

Resultados finales

Como resultado del proyecto, se obtiene un **Data Warehouse en PostgreSQL completamente estructurado**, con múltiples capas de datos claramente separadas según su nivel de procesamiento. La capa Bronze contiene los datos tal como fueron cargados por Airbyte, incluyendo posibles metadatos técnicos. La capa Silver ofrece tablas limpias y tipadas, listas para ser reutilizadas sin problemas de calidad básica.

La capa Gold produce tablas agregadas que permiten analizar ventas diarias, mensuales, comportamiento de clientes, artículos más vendidos y facturación, entre otros indicadores. Estas tablas están pensadas para ser consumidas directamente por herramientas de visualización o reporting, sin necesidad de transformaciones adicionales. Cada ejecución del pipeline reconstruye estas tablas para reflejar siempre el estado más reciente de los datos.

Por su parte, la capa Datamining genera una tabla plana con una fila por transacción y producto, sin precios ni datos de clientes. Este diseño permite que la tabla sea utilizada directamente desde notebooks para aplicar algoritmos de Market Basket Analysis, como Apriori o FP-Growth. En conjunto, el pipeline produce un entorno de datos listo tanto para análisis descriptivo como para análisis más avanzados.

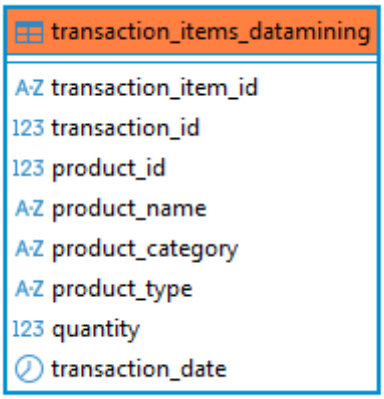


Imagen 3: Dataset generado para análisis en Schema Datamining

Conclusiones

El proyecto **voltaje_etl_pipeline** demuestra cómo es posible construir un pipeline de datos completo, ordenado y reproducible utilizando herramientas ampliamente adoptadas en la industria. La combinación de Docker, Airflow y dbt permite automatizar todo el proceso de transformación de datos, desde la verificación de la disponibilidad de las fuentes hasta la generación de capas finales listas para análisis.

Además, el uso de la arquitectura Medallion facilita la comprensión y el mantenimiento del Data Warehouse, ya que cada capa tiene un propósito claro y bien definido. El proyecto deja un entorno preparado para escalar, agregar nuevas fuentes o incorporar nuevos modelos, y constituye una base sólida para prácticas académicas relacionadas con ingeniería de datos, orquestación de pipelines y modelado analítico.