Desarrollo de un pipeline de datos tipo ELT y un data warehouse escalable

1. Contexto del Proyecto

El equipo directivo ha solicitado el diseño de una **arquitectura de datos escalable y automatizada**, capaz de recopilar, almacenar y transformar grandes volúmenes de información provenientes de distintas fuentes, con el objetivo de **optimizar la toma de decisiones estratégicas**.

Este proyecto se enfoca en el análisis de datos del mercado de hospedaje (Airbnb NYC), permitiendo identificar patrones de precios, demanda, comportamiento de hosts y distribución geográfica de las propiedades.

En esta primera etapa, se define la **arquitectura base del pipeline ELT**, que servirá como fundamento para las fases de ingesta, transformación, orquestación y análisis.

Objetivo del Pipeline ELT

Diseñar e implementar un **pipeline ELT robusto**, desplegado en la nube, que permita:

- Ingerir datos desde archivos planos y fuentes externas.
- Almacenar los datos de forma estructurada y versionada.
- Transformar la información para su análisis.
- Automatizar el flujo mediante orquestación y CI/CD.
- Garantizar calidad, trazabilidad y escalabilidad.

Descripción General del Pipeline ELT

El pipeline sigue un enfoque **ELT (Extract, Load, Transform)**, donde los datos se cargan primero en su forma original y posteriormente se transforman dentro del ecosistema de datos.

🔹 Extract

- Los datos se extraen desde:
    - Archivos CSV (AB_NYC.csv).

    - APIs externas

- La extracción se realiza mediante **scripts en Python**, ejecutados en contenedores Docker.

🔹 Load

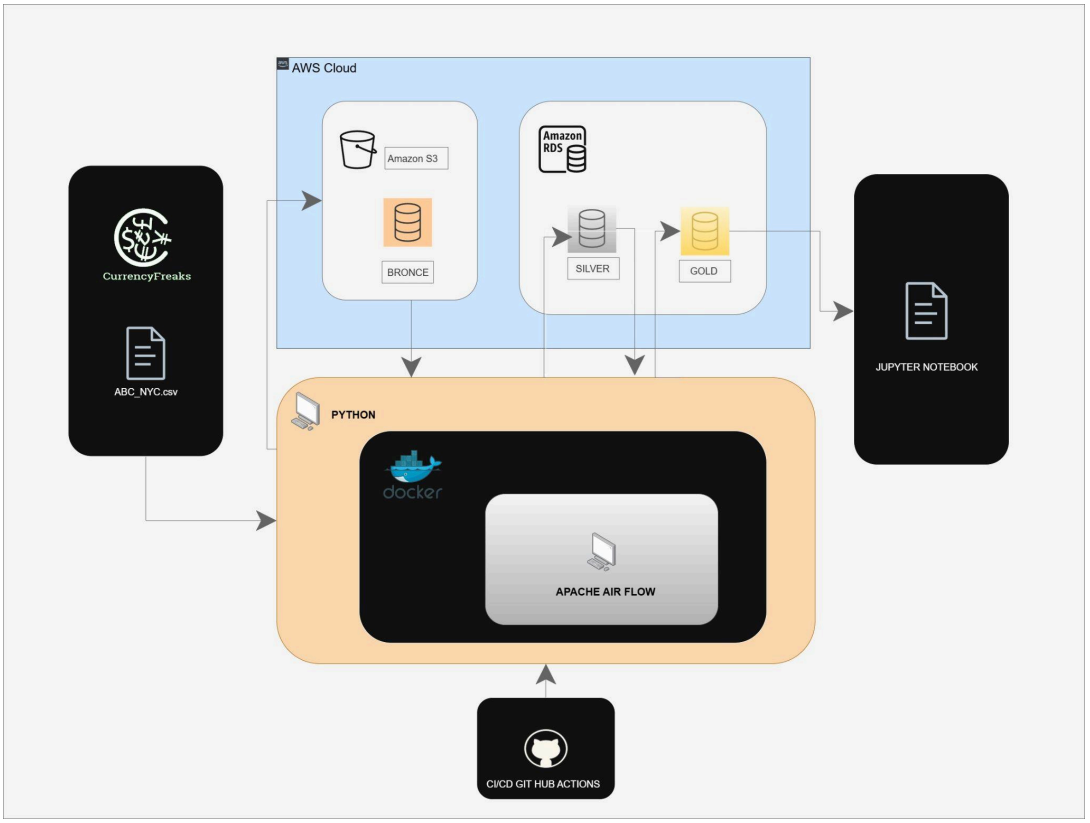
- Los datos extraídos se almacenan sin transformaciones en **Amazon S3**, correspondiente a la **capa Bronze (Raw)**.
- Se mantiene el formato original para garantizar trazabilidad y reprocesamiento.

🔹 Transform

- A partir de la capa Bronze, los datos son transformados y cargados en una base de datos relacional en **Amazon RDS**, donde se estructuran en:

    - **Silver Layer**: datos limpios y normalizados.
    - **Gold Layer**: datos agregados y optimizados para análisis.

Arquitectura General del Pipeline



Componentes principales

Bronze Layer (Raw)

- Servicio: **Amazon S3**
- Contenido:
    - Datos crudos
    - Sin transformaciones
    - Versionados por fecha
- Propósito:
    - Preservar el dato original
      Permitir reprocesamiento

Silver Layer

- Servicio: **Amazon RDS**
- Contenido:
    - Datos limpios
    - Tipos de datos corregidos
    - Eliminación de duplicados
- Propósito:
    - Preparar datos confiables

Gold Layer

- Servicio: **Amazon RDS**
- Contenido:
    - Métricas de negocio
      Agregaciones
    - Modelos analíticos
- Propósito:
    - Consumo analítico y reporting

Procesamiento y Orquestación

- **Python + Docker**
    - Ejecuta scripts de extracción y validación.
- **Apache Airflow**
    - Orquesta el pipeline completo.
    - Define dependencias entre tareas.
    - Maneja reintentos y fallos.

CI/CD

- **GitHub Actions**
    - Automatiza:
        - Tests
        - Build de imágenes Docker
        - Despliegue del pipeline
- Garantiza calidad continua del código.

Consumo de Datos

- **Jupyter Notebook**
    - Exploración de datos
    - Análisis ad-hoc
    - Validaciones finales

Definición de Capas del Data Warehouse

Capa	Tecnología	Función
Bronze	Amazon S3	Datos crudos
Silver	Amazon RDS	Datos limpios
Gold	Amazon RDS	Datos analíticos

Cada capa cumple un rol específico dentro del ciclo de vida del dato, permitiendo un flujo ordenado y escalable.

Justificación de Herramientas

Python

- Lenguaje estándar en Data Engineering.
- Amplio ecosistema de librerías.

Docker

- Portabilidad y reproducibilidad.
  Facilita despliegue en la nube.

AWS (S3 + RDS)

- Alta disponibilidad.
- Escalabilidad automática.
- Integración nativa entre servicios.

Apache Airflow

- Orquestación robusta.
- Manejo de dependencias.
- Visibilidad del pipeline.

GitHub Actions

- CI/CD integrado al repositorio.
- Automatización sin infraestructura adicional.

Relación Preguntas de Negocio – Datos

Pregunta	Fuente
Zonas más rentables	AB_NYC.csv
Precio promedio por barrio	AB_NYC.csv
Tipos de alojamiento más demandados	AB_NYC.csv
Hosts con mejor desempeño	AB_NYC.csv