PoC Airflow

## Objetivo General

Desarrollar una solución de ETL usando herramientas de Google Cloud para la disponibilización de datos en un warehouse apropiado.

## Objetivos específicos

* Configurar y desplegar airflow utilizando Google Cloud Composer.
* Diseñar flujos de datos que satisfagan diferentes requerimientos del cliente.
* Desarrollar diferentes funcionalidades que permitan implementar las soluciones diseñadas, utilizando buenas prácticas de programación, alineando los procesos con los recursos disponibles.
* Integrar las distintas herramientas que ofrece Google Cloud para la implementación de flujos de trabajo automatizados.

## Entrega

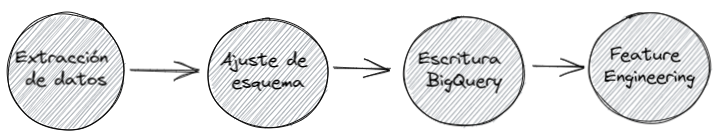
* Google Cloud Composer funcional con un DAG que permita ejecutar el siguiente proceso:
  1. Descargar datos de [vuelos](#_ol8gqkjd12sx)  (airlines, airports, flights) y dejarlos en un bucket de Google Cloud Storage en forma de parquet y todos los campos como string.
  2. Leer datos del proceso anterior, realizar un casteo de esquema para los tipos adecuados y escribir nuevamente en una ruta diferente.
  3. Crear (o asegurarse que existen) las tablas de airlines, airports y flights en BigQuery.
  4. Escribir los datos procesados del bucket de Google Cloud Storage a BigQuery.
  5. Generar una vista en BigQuery, que tenga solo 1 tabla integrando las 3 tablas ingestadas.

## Restricciones:

* Airflow nunca debe cargar datos en RAM, si es necesario hacerlo, se deben utilizar herramientas externas como Cloud Functions, Google Kubernetes Engine u otros.

## Extras:

* Adicionar pruebas unitarias a las diferentes funciones y clases implementadas en el código (debe tener un coverage de mínimo el 80%).
* Para un desafío adicional, realizar el “a” en un Dag independiente con trigger de tiempo y las etapas “b” - “e” deben ejecutarse al momento de llegar los objetos al bucket, utilizando una notificación de OBJECT\_FINALIZE.



## Recursos

* Datos de vuelos: <https://github.com/dpinedaj/airflights-kaggle-data>
* Apache Airflow: <https://airflow.apache.org/docs/apache-airflow/stable/index.html>
* Cloud Functions: <https://cloud.google.com/functions/docs>
* Cloud Composer: <https://cloud.google.com/composer/docs>
* Coverage: <https://coverage.readthedocs.io/en/coverage-4.0.3/config.html>
* Pytest: <https://docs.pytest.org/en/7.1.x/>