## Ítem 1: Descripción general de la solución al caso de uso

Para este caso de uso se utilizará un Bucket en Google Cloud Storage que contenga las 3 fuentes de datos, estos archivos se ingestarán en tablas de Google Bigquery para posteriormente realizar una query sobre ellas y así poder generar una tabla final con los datos necesarios para el modelo de Machine Learning. Los pipelines serán construidos en código Python usando Apache Airflow y se utilizará GitHub como sistema de control de versiones git. A continuación, se procede a describir un poco más cada punto.

Link del repositorio GitHub: <https://github.com/cesar-munsav/exercise-meli>

## Ítem 2: Sources

El caso de uso contempla 3 fuentes de datos que serán utilizadas para generar un dataset que será utilizado por un modelo de Machine Learning. Para realizar esta tarea se plantean los siguientes supuestos para estos sources:

1. Existirá un proyecto default en GCP que cuente con un Bucket en Google Cloud Storage donde se puedan almacenar estas 3 fuentes de datos. El pipeline va a considerar que los 3 archivos ya se encuentran almacenados en este Bucket.
2. Tal como se menciona en el enunciado, existirá 1 archivo “csv” delimitado por “,” (pays) y 2 archivos “json lines” (prints y taps) como fuentes de datos.

## Ítem 3: Ingesta Sources en Bigquery

Dado que el caso de uso involucra procesar la información que existe en los archivos de origen de forma conjunta, se propone llevar la información a un formato tabular en Bigquery para posteriormente realizar querys sobre estas tablas y entregar como producto final una tabla con la información que se necesita para el modelo. Para realizar esta ingesta desde Google Cloud Storage hacia Bigquery se plantean los siguientes supuestos:

1. Se construirá un pipeline en código Python usando la librería de Apache Airflow para construir un dag que contenga este workflow. El dag contempla generar 3 task para poder ingestar cada uno de los archivos en una tabla distinta en Bigquery.

## Ítem 4: Creación de Tabla final con el producto

Tras generar las 3 tablas en Bigquery con la información de los sources, se procederá a generar una tabla final en Bigquery que contenga la información necesaria para el modelo. Para realizar esta tarea se plantean los siguientes supuestos:

1. Existirá un proyecto default en GCP que contengan en Bigquery las tablas que son productos de datos.
2. Se creará un dag que contemple una task que actúe como sensor y espere a que el primer dag creado finalice (el que ingesta los sources), también cuenta con una task que ejecuta una query en Bigquery y su resultado se almacena en la tabla final de Bigquery que contiene la data para el modelo.

## Ítem 5: Control de versiones

Se utilizará GitHub por conveniencia como sistema de control de versiones git, se contarán con los ambientes “dev”, “uat” y “prd”, estos ambientes se usan como variables dentro del código y se hacen referencia a ellas al momento de ingestar la data y crear el producto final.

## Ítem 5: Diagramas y Dashboard

De forma adicional se construye un diagrama de arquitectura con el todo el flujo, este diagrama se encuentra en el repositorio (pipeline diagram), también se agrega un dashboard en Lookerstudio con un análisis global de la data final del modelo (bi-analytics).