Juan Sebastian Burgos Ormeño

INFORME DE PRUEBA TECNICA DATA ENGINEER

Objetivo

Esta es una prueba diseñada para los candidatos al cargo de Data Engineer en Hunty. El objetivo de la prueba es evaluar la capacidad del candidato en aprender a usar las herramientas típicas que en el día a día se usan en el equipo de Data así como también validar conocimientos previos en las mismas.

El desarrollo de la prueba requiere de parte del candidato crear un proyecto en Google Cloud con el que pueda hacer las pruebas que le permitan ejecutar los desafios aqí planteados. La siguiente documentación le será bastante útil.

Desarrollo Primera Parte

El desarrollo de la primera parte de esta prueba técnica requiere un conocimiento previo de arquitectura de GCP como de desarrollo en Python, para lograr el objetivo de esta primera parte se realizaron las siguientes configuraciones:

Lo primero que se debe hacer, con el fin de poder utilizar los USD 300 que obsequia google para usar GCP, es crear una cuenta de correo Gmail, la evidencia de este paso la encontrará en la ***Ilustración 1***

Ilustración 1

El paso siguiente corresponde a la activación del periodo de prueba del GCP free, lo anterior nos permitirá disponer de la gran mayoría de herramientas que ofrece google en su nube, lo anterior con el fin de tener la posibilidad de comparar el mencionado GCP con otras nubes tales como AWS y Azure.

INFORME DE PRUEBA TECNICA DATA ENGINEER

Este punto nos brinda las herramientas para desarrollar tanto la parte 1 como la 2 de la presente prueba técnica y lo podrán evidenciar en la ***Ilustración 2***.

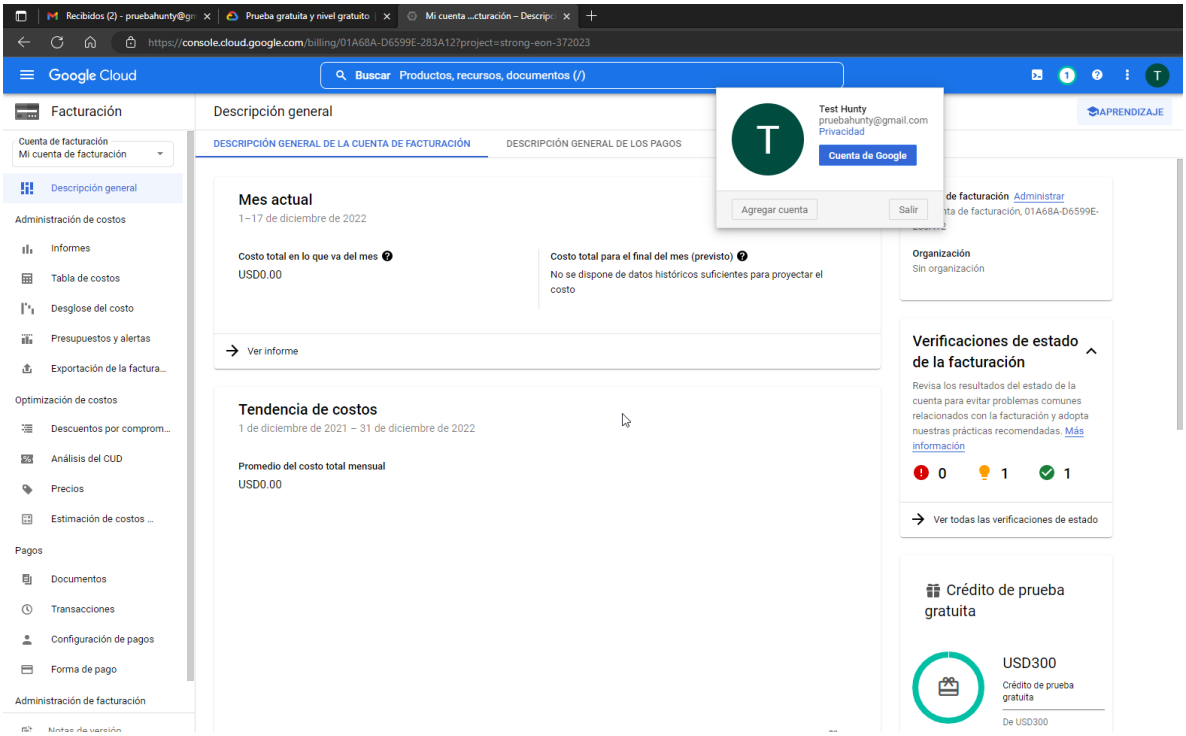

Ilustración 2

Una vez se cuenta con el acceso a las herramientas y la facturación configurado, lo que sigue por realizar es la creación del proyecto donde se albergarán todas las soluciones del presente proceso, para este caso se crea un proyecto llamado **PruebaHunty** y lo podrá evidenciar en la ***Ilustración 3***.

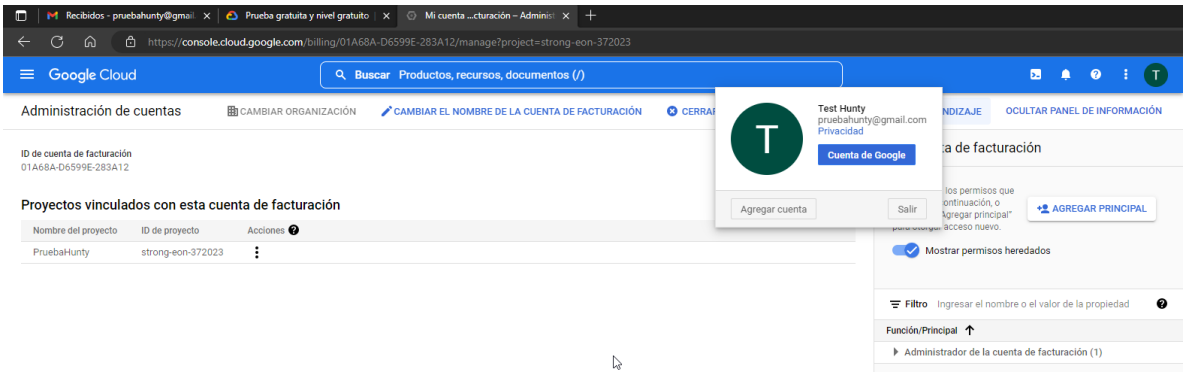

Ilustración 3

INFORME DE PRUEBA TECNICA DATA ENGINEER

Ya teniendo creado el proyecto, el paso a seguir es crear la cuenta de servicio que permitirá tener acceso desde el desarrollo hacia los diferentes componentes de GCP mediante la creación de una Api Key, teniendo muy en cuenta los roles que se le deben asignar. Ver ***Ilustración 4***.

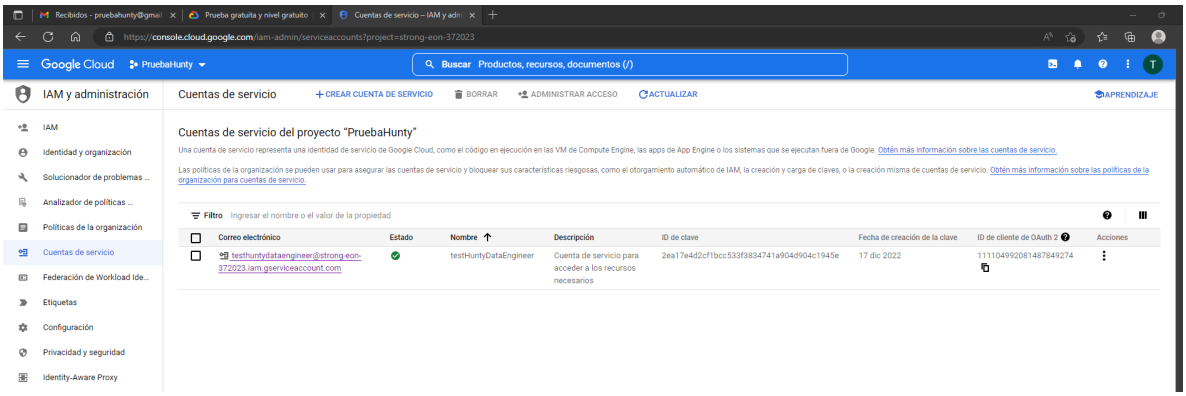

Ilustración 4

Para la elaboración de la primera parte de esta prueba técnica se solicita almacenar 2 archivos json en un bucket para posteriormente ser consumidos desde Python mediante un script. La evidencia de la creación de dicho bucket (pruebaHunty) la podrán encontrar en la ***Ilustracion 5.***

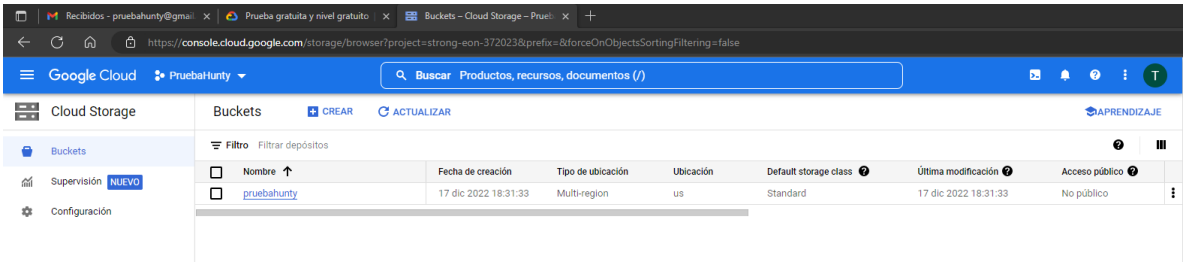

Ilustración 5

Luego de tener listo el bucket, se procede a cargar los archivos json (después de validar su estructura ya que contiene errores) como se muestra en la ***Ilustración 6***.

INFORME DE PRUEBA TECNICA DATA ENGINEER

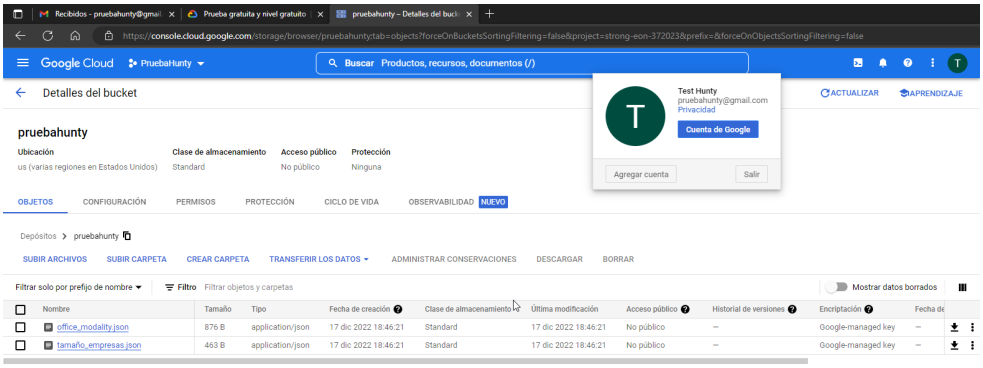

Ilustración 6

Posteriormente, y para poder utilizar los Google Sheet se debe dirigir a la biblioteca de API que tiene GCP y habilitar Google Drive API y Google Sheet API y asociarles la cuenta de servicio creada anteriormente. Estas evidencias las encontrarán en las ***Ilustraciones 7, 8, 9 y 10***.

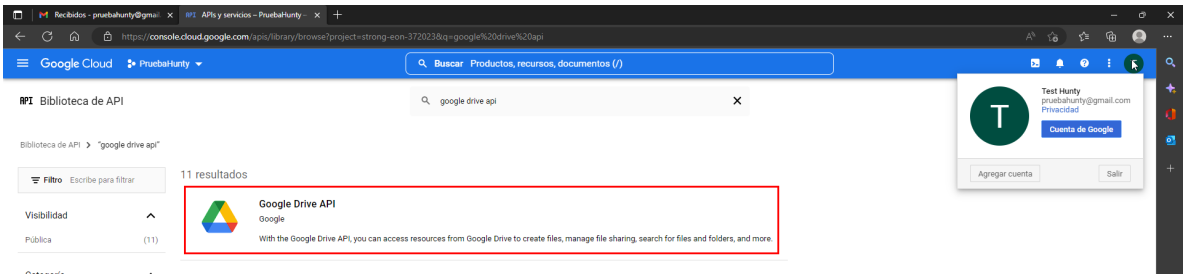

Ilustración 7

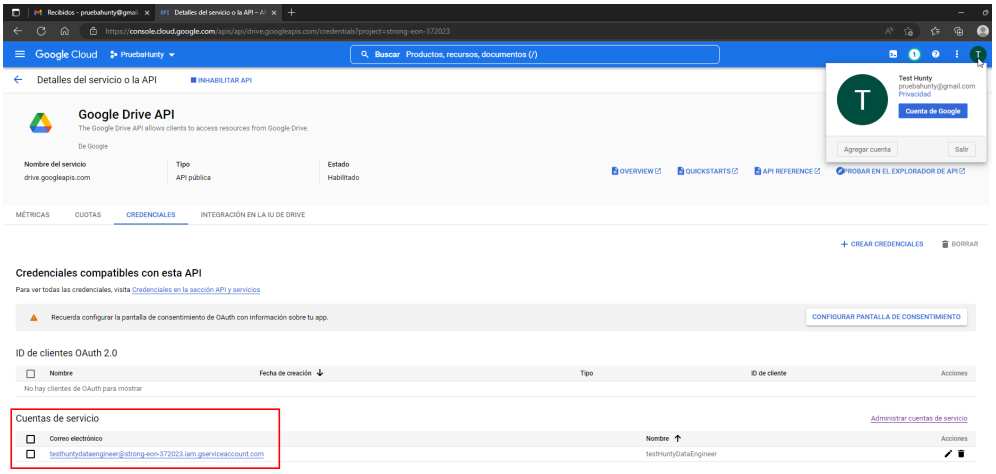

Ilustración 8

INFORME DE PRUEBA TECNICA DATA ENGINEER

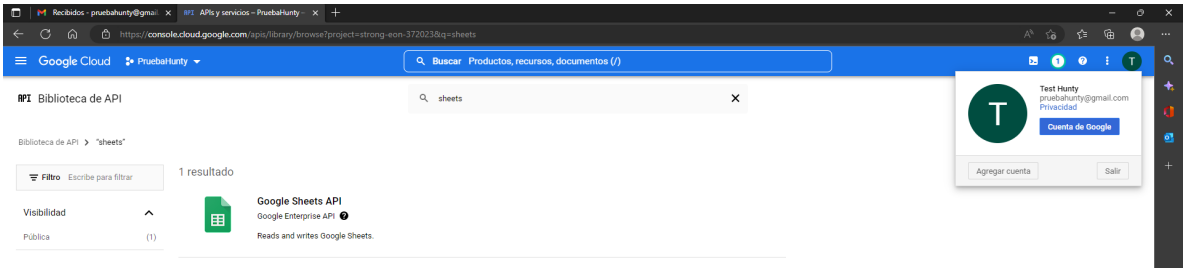

Ilustración 9

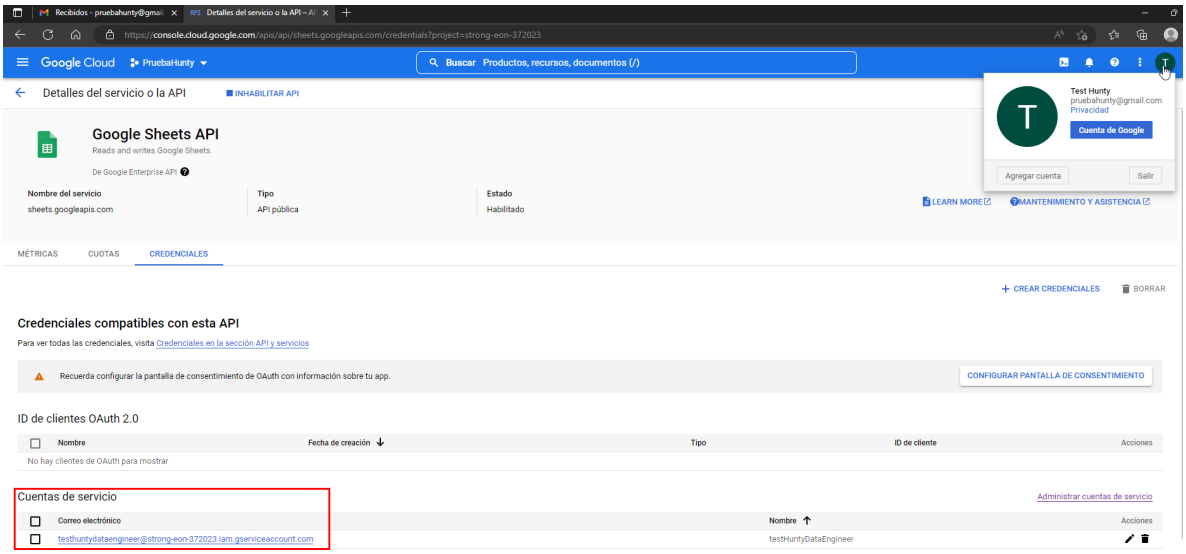

Ilustración 10

INFORME DE PRUEBA TECNICA DATA ENGINEER

Desarrollo Primera Parte

Para el desarrollo de esta segunda parte, se debe crear una instancia en el motor de bases de datos PostgreSQL desde GCP. Para lo anterior se debe seleccionar la opción SQL y posteriormente el motor mencionado, en caso de no tener habilitada la opción API de Compute Engine GCP brindará la opción, esto se muestra en la ***Ilustración 11***.

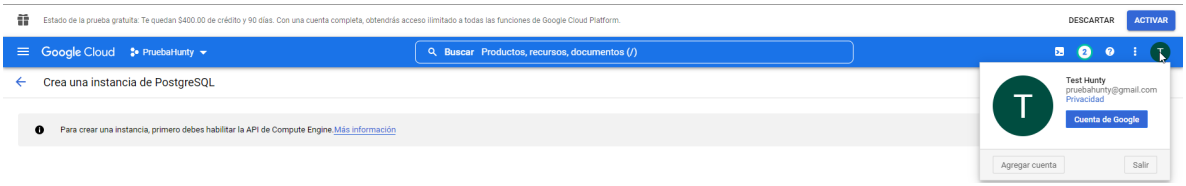

Ilustración 11

Una vez habilitada la API de Compute Engine se procede a la creación de la instancia como se muestra en la ***Ilustración 12***.

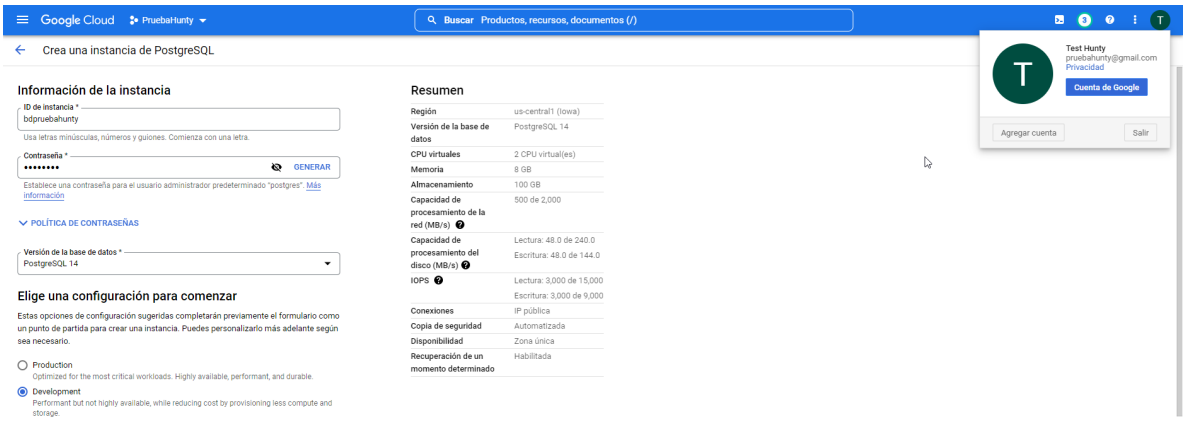

Ilustración 12

Ahora procedemos a crear la base de datos en PostgreSQL, esto se muestra en la ***Ilustración 13***.

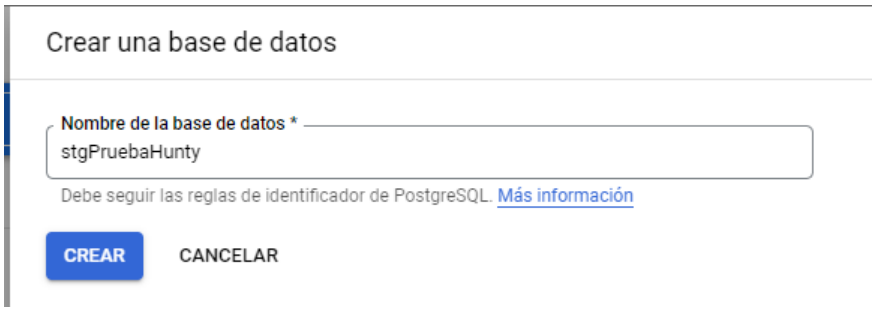

Ilustración 13

Los pasos anteriores que tratan sobre la creación de la instancia en PostgreSQL, sin embargo, y de acuerdo al enunciado del ejercicio, deseo brindar una propuesta de solución para el punto 2 que consiste en lo siguiente:

Para el desarrollo de dicho punto se sugiere realizar el cargue de la información cruda en un dataset creado en BigQuery con prefijo ***raw*** (aclarando que se presentarán algunas transformaciones que nos permitirán el condicionamiento en los Queries). Este paso se realizar con el objetivo de mantener data original para auditar posteriormente o si en algún momento cambian las reglas de negocio (todo negocio evoluciona).

Posteriormente se crea un dataset en BigQuery con prefijo ***mst*** donde almacenará la información de cada tabla ya pulida con las reglas de negocio.

Por último y para finalizar el ejercicio, se crea la vista con las condiciones establecidas en el ejercicio cuyas tablas apuntaran a los objetos creados en el dataset con prefijo ***mst.*** Se presenta la estructura de BigQuery en la ***ilustración 14***.

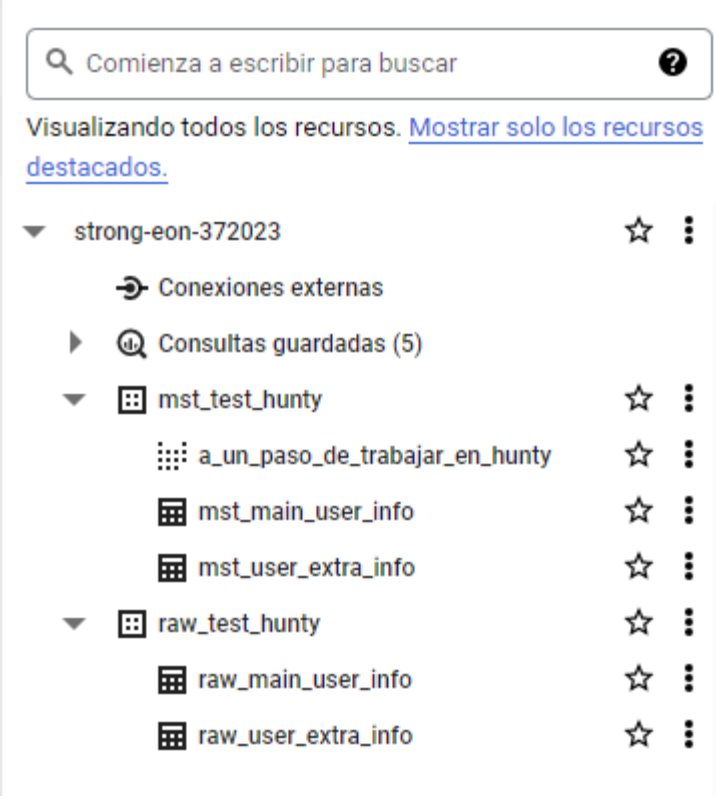

Ilustración 14

INFORME DE PRUEBA TECNICA DATA ENGINEER

A continuación, se relacionan los queries utilizados para la creación de los objetos en la capa MST de la solución en BigQuery.

CAPA MST

```sql
CREATE OR REPLACE TABLE `strong-eon-372023.mst_test_hunty.mst_main_user_info`
AS
  SELECT
    user_id
    ,first_name
    ,INITCAP(LOWER(last_name)) last_name
    ,Phone
    ,CAST(CAST(load_date AS TIMESTAMP) AS DATE) load_date
  FROM
    `strong-eon-372023.raw_test_hunty.raw_main_user_info`
```

Ilustración 15: mst_main_user_info

```sql
CREATE OR REPLACE TABLE `strong-eon-372023.mst_test_hunty.mst_user_extra_info`
AS
  SELECT
    user_id
    ,vacancy_area_id
    ,location_change_city_ids
    ,available_time_week_id
    ,vacancy_area_custom
    ,change_city
    ,years
    ,months
  FROM
    `strong-eon-372023.raw_test_hunty.raw_user_extra_info`
  WHERE
    recurrencia_item > 1
    AND (CAST(CASE WHEN vacancy_area_id = '' THEN '0' ELSE vacancy_area_id END AS FLOAT64) >= 2
    OR (CAST(CASE WHEN vacancy_area_id = '' THEN '0' ELSE vacancy_area_id END AS FLOAT64) < 2 AND employment_status = '0'))
```

Ilustración 16: mst_user_extra_info

```sql
CREATE VIEW `strong-eon-372023.mst_test_hunty.a_un_paso_de_trabajar_en_hunty`
AS
  SELECT
    T2.user_id
    ,CASE WHEN T2.first_name = '' THEN 'None' ELSE T2.first_name END first_name
    ,CASE WHEN T2.last_name = '' THEN 'None' ELSE T2.last_name END last_name
    ,CASE WHEN T2.Phone = '' THEN 'None' ELSE T2.Phone END Phone
    ,T2.load_date
    ,CASE WHEN T1.vacancy_area_id = '' THEN 'None' ELSE T1.vacancy_area_id END vacancy_area_id
    ,CASE WHEN T1.location_change_city_ids = '' THEN 'None' ELSE T1.location_change_city_ids END location_change_city_ids
    ,CASE WHEN T1.available_time_week_id = '' THEN 'None' ELSE T1.available_time_week_id END available_time_week_id
    ,CASE WHEN T1.vacancy_area_custom = '' THEN 'None' ELSE T1.vacancy_area_custom END vacancy_area_custom
    ,CASE WHEN T1.change_city = '' THEN 'None' ELSE T1.change_city END change_city
    ,T1.years
    ,T1.months
  FROM
    `strong-eon-372023.mst_test_hunty.mst_user_extra_info` T1 INNER JOIN `strong-eon-372023.mst_test_hunty.mst_main_user_info` T2
    ON T1.user_id = T2.user_id
```

Ilustración 17: mst_test_hunty.a_un_paso_de_trabajar_en_hunty

Juan Sebastian Burgos Ormeño

INFORME DE PRUEBA TECNICA DATA ENGINEER

Ya para finalizar solo queda decir:

Gracias por la oportunidad.

INFORME DE PRUEBA TECNICA DATA ENGINEER

Evidencia de los resultados Parte 1:

Para poder ver los resultados de la ejecución del script, fue necesario compartir dichos archivos a una cuenta personal con la que se observaran los datos cargados de acuerdo a la solicitud del ejercicio.

En la ***Ilustración 11*** encontrarán los datos cargados del archivo json tamaño_empresas.json

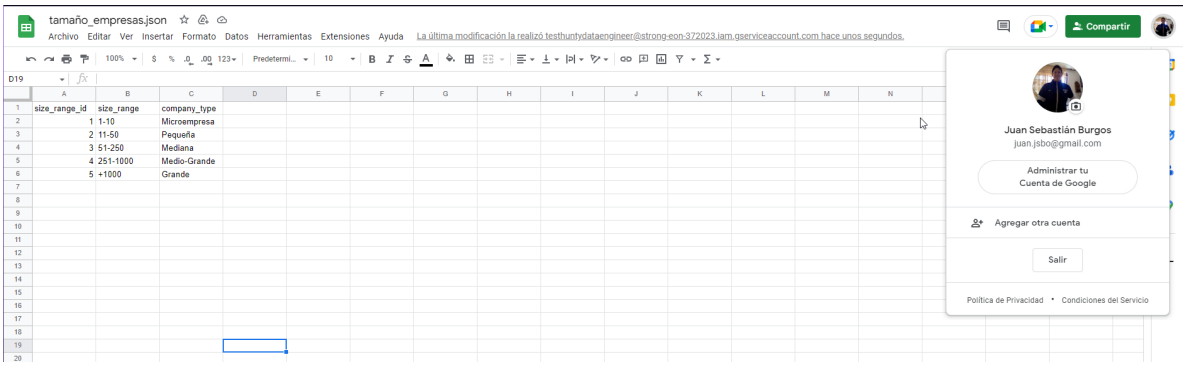

Ilustración 18

En la ***Ilustración 12*** encontrarán los datos cargados del archivo json office_modality.json.

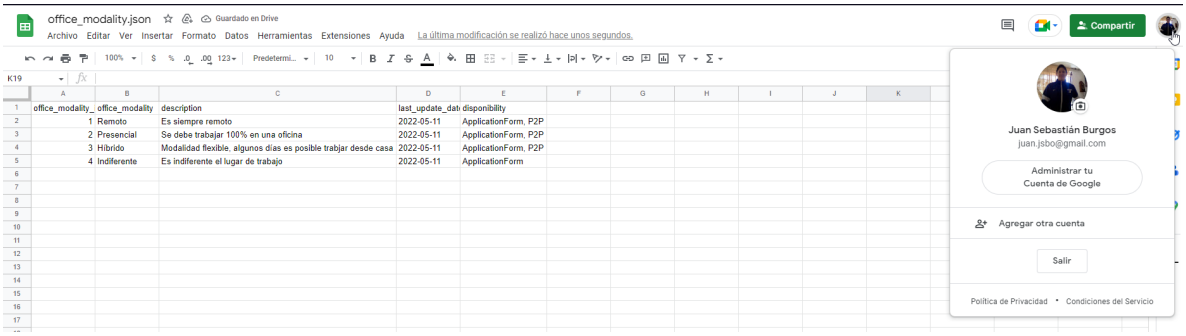

Ilustración 19

Por último, se adjunta evidencia de archivos compartidos por medio del script en la Ilustración 13

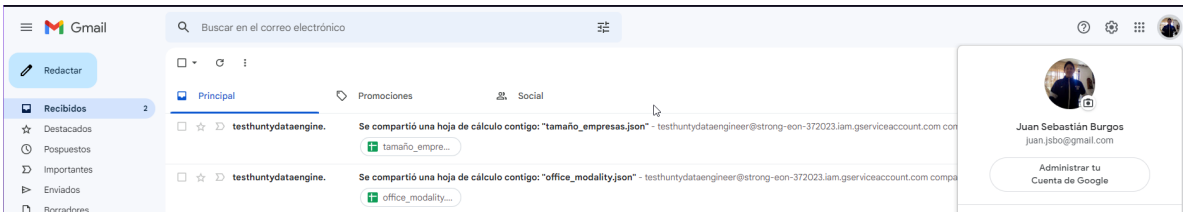

Ilustración 20

Juan Sebastian Burgos Ormeño

INFORME DE PRUEBA TECNICA DATA ENGINEER

Para finalizar el resumen de la primera parte de la prueba técnica, se adiciona el link del GitHub donde encontrarán el script generador de esta primera parte (PruebaTecnicaHunty\parte1):