# ***Ejercicio 1***

#### **Punto Número 2**

De acuerdo con el enunciado de la tarea, se requería crear dos tablas en el datawarehouse de HIVE. Se creó en HIVE una base de datos llamada “***airport\_trips\_data***”. En esta base de datos se crearon las siguientes tablas:

1. ***aeropuerto\_tabla***
2. ***aeropuerto\_detalles\_tabla***

A continuación, los screen shots de ambas tablas:

***aeropuerto\_tabla***

A screenshot of a computer

AI-generated content may be incorrect.

***aeropuerto\_detalles\_tabla***

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 3**

El proceso de ingesta se hizo de forma automática por medio de Apache Airflow. Para esto, se creo un DAG (en formato archivo .py). A continuación, un screen shot del DAG

A black screen with white text

AI-generated content may be incorrect.

Este DAG contiene dos procesos

1. Un primer proceso corre el archivo .sh que tiene como propósito la ingesta de los archivos y almacenado en HDFS
2. El segundo proceso, corre un archivo .py que contiene las instrucciones de PySpark para procesar la información

Para mayor detalle en la carpeta ***Tarea\_1*** está el archivo para su revisión. El archivo se llama “***exercise\_1\_dag\_final.py***”

A continuación, se presenta el archivo .sh que hace el proceso de ingesta desde la página web y almacena los archivos en el HDFS. El archivo se llama ***job1\_final.sh***. Para mayor detalle ver los archivos dentro de la carpeta ***Tarea\_1***

A screen shot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

#### **Punto Número 4**

Las instrucciones de transformación de la data se encuentran dentro del archivo ***pyspark\_shell\_final1.py.*** A continuación, un screen shot del archivo. En este script se trabajó lo requerido en el punto 4. Para mayor detalle en la carpeta **Tarea\_1** está el archivo para su revisión

A screen shot of a computer error

AI-generated content may be incorrect.

#### **Punto Número 5**

A continuación, screen shots de la creación de la base de datos y la descripción de las tablas.

##### ***Creación de la base de datos airport\_trips\_data***

A black rectangular object with a gray border

AI-generated content may be incorrect.

##### ***Creación de la tabla aeropuerto\_tabla***

A screenshot of a computer

AI-generated content may be incorrect.

##### ***Creación de la tabla aeropuerto\_detalles\_tabla;***

A screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 6**

A continuación, screen shots de los queries requeridos en el trabajo 1. Para mayor detalle dentro de la carpeta Tarea\_1 se encontrarán los archivos de soporte empleados. ***El query fue modificado de acuerdo con lo requerido. El problema inicial fue que, durante el proceso de transformación, el formato de la fecha no era el correcto***

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 7**

A continuación, screen shots de los queries requeridos en el trabajo 1. Para mayor detalle dentro de la carpeta Tarea\_1 se encontrarán los archivos de soporte empleados. ***El query fue modificado de acuerdo con lo requerido. El problema inicial fue que, durante el proceso de transformación, el formato de la fecha no era el correcto***

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 8**

A continuación, screen shots de los queries requeridos en el trabajo 1. Para mayor detalle dentro de la carpeta Tarea\_1 se encontrarán los archivos de soporte empleados. El ordenamiento se hizo en base a la fecha de forma descendiente.

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 9**

A continuación, screen shots de los queries requeridos en el trabajo 1. Para mayor detalle dentro de la carpeta Tarea\_1 se encontrarán los archivos de soporte empleados.

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 10**

A continuación, screen shots de los queries requeridos en el trabajo 1. Para mayor detalle dentro de la carpeta Tarea\_1 se encontrarán los archivos de soporte empleados. ***Se modifico el query de acuerdo a los requerimientos.***

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 11**

Personalmente considero que el dataset está bueno para el propósito que se quiere. Yo agregaría los siguientes datos:

1. **Número de vuelto**. La tabla ya contiene los datos de despegue y aterrizaje por nombre de aerolínea, fecha de salida y arribo, número de pasajeros y aeronave, pero es complicado determinar si son el mismo vuelo. Esto ayudaría a mejorar un poco los analíticos que se quieren extraer.
2. Otro dato que se pudiera colocar es el **tiempo teórico del vuelo**, por ejemplo, si fuera aerolíneas argentinas podría determinar razones de eficiencia. Tiempo real vs tiempo teórico.
3. Otro dato pudiera ser **horas de vuelo de la aeronave** y **rutinas de mantenimiento preventivo y/o correctivo**.
4. Finalmente, alguna meta data, como **comentarios** de los cuales podemos extraer ciertas palabras para clasificar vuelos o determinar patrones

#### **Punto Número 12**

#### De acuerdo con la información extraída de los queries:

#### Desde inicios de diciembre 2021 hasta finales de enero 2022, ha habido una cantidad total de un poco más de 50K vuelos.

#### Por otro lado, desde la fecha de enero 2021 hasta finales de junio 2022. Aerolíneas Argentinas ha tenido un poco más de 6.8MM de pasajeros. Los rangos de tiempo no son comparables respecto al resultado presentado en el punto anterior.

1. La aerolínea Aerolíneas Argentinas en el mismo periodo movilizó la mayor cantidad de pasajeros, con más de 6.8MM de pasajeros. Le siguió la aerolínea JET SMART con un poco más de 1.3MM de pasajeros.
2. En el mismo periodo el avión marca modelo Embraer generó mayor cantidad de vuelos, con 11,587 desde la provincia de Buenos Aires.

#### **Punto Número 13**

Yo considero que el dinamismo de la data que se extrajo para este trabajo es recomendable pensar en una arquitectura en la nube. La data tiene un poco más de 500K registros, eso fue en dos archivos en casi dos años de data. Dada la cantidad de data, mejorar considerar una arquitectura en la nube de la siguiente forma:

1. Google Cloud Storage para almacenar la data
2. Google Dataproc
3. Google BigQuery

Todo orquestrado por Google Cloud Composer. Respecto a una herramienta para manejar calidad de data conectaría a Google Big Query una herramienta como Great Expectations (GX Cloud)

# ***Ejercicio 2***

#### **Punto 1**

De acuerdo con lo requerido en esta tarea se ha creado una base de datos en HIVE llamada ***car\_rental\_db***

A black screen with white text

AI-generated content may be incorrect.

Y se ha creado la tabla ***car\_rental\_analytics***

A black screen with white text

AI-generated content may be incorrect.

A continuación, se presenta la descripción de la tabla, campos y tipos de campos

A computer screen shot of a black screen

AI-generated content may be incorrect.

#### **Punto 2**

Se anexa una imagen de pantalla del archivo .sh que se empleó para extraer los archivos. El archivo se llama ***job2\_final.sh***. Para mayor detalle ver el contenido de la carpeta ***Tarea\_2***.

A screenshot of a computer program

AI-generated content may be incorrect.

#### **Punto 3**

A continuación, se presenta una imagen de pantalla del código en ***pyspark*** realizado para poder transformar la data de acuerdo con los requerimientos de la tarea. El archivo se llama ***exercise\_2\_pyspark\_shell\_final.py.*** Para mayor detalle ver el contenido de la carpeta ***Tarea\_2***

A computer screen shot of a computer code

AI-generated content may be incorrect.

#### **Punto 4**

Se anexa una imagen del DAG que corre el proceso de ingesta y corre el llamado al DAG hijo. Para mayor detalle ver los contenidos de la carpeta **Tarea\_2.** El DAG principal se encuentra en el archivo “***dag\_exercise2.py***”, mientras el DAG hijo se encuentra en el archivo “***dag-hijo-final.py***”

A screen shot of a computer program

AI-generated content may be incorrect.

#### **Punto Número 5.a**

***Para ejecutar este punto y todos los puntos descritos más abajo, vamos a asumir que el campo de “rentertripstaken” es la cantidad de alquileres que tuvieron los vehículos. En la información dispuesta en Kaggle no había información de una meta data que nos indicara que a que corresponde cada columna.***

Para este ejercicio se pide que se coloque el número de alquileres totales. Asumo que el campo de “rentertripstaken” es la cantidad de alquileres para un vehículo.

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 5.b**

En este query se pedía obtener los estados con la menor cantidad de alquileres. Asumo que la cantidad de alquileres es representada por la columna “rentertripstaken”

A screenshot of a computer

AI-generated content may be incorrect.

**Gráfica de Looker**

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 5.c**

En este query se requería tener los 10 modelos junto con su marca más rentados. No se hace distinción por tipo de fuel.

A screenshot of a computer

AI-generated content may be incorrect.

**Gráfica de Looker**

A graph with blue bars

AI-generated content may be incorrect.

#### **Punto Número 5.d**

El query requería que se obtuviera por año la cantidad de vehículos alquilados. Como no existe una fecha de alquiler o periodo en el que se hizo el alquiler se usará la columna de “make year”. Asumo que el campo “rentertripstaken” es la cantidad de alquileres de un vehículo en particular

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 5.e**

En este query se pide seleccionar las 5 ciudades con mayor número de alquileres de vehículos que fueran híbridos o eléctricos

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 5.f**

A screenshot of a computer

AI-generated content may be incorrect.

#### **Punto Número 6**

De la información extraída de los queries podemos determinar los siguiente:

1. Hay un poco más de 26K de vehículos híbridos y/o eléctricos que tuvieron un rating por encima de 4.
2. Montana y Arkansas fueron las ciudades donde menos alquileres hubo.
3. El Tesla Model 3, es el vehículo que más alquilaron. Un total de superior a 9,700 veces
4. Los vehículos fabricados en el 2015 fueron los vehículos más alquilados con un poco más de 18K veces
5. San Diego y Las Vegas fueron las ciudades con mayor alquiler de vehículos híbridos y eléctricos
6. Finalmente, los clientes fueron bastante positivos con los alquileres de los vehículos eléctricos e híbridos

#### **Punto Número 7**

El Proyecto podemos decir que es un proyecto de ciencia de datos, con poco volumen de data. Un poco menos de 5,000 registros. Basado, creo que lo más conveniente es conserva la arquitectura on premise. Ingesta de datos de HDFS, trabajo de los datos con pyspark y almacenar la data en HIVE. Como orquestrado nos podemos quedar con Apache Airflow.

#### **Google Skill Boost - LAB**

A screenshot of a computer

AI-generated content may be incorrect.

#### **Preguntas**

1. ***¿Para qué se utiliza Dataprep?***

En base a lo visto en el LAB, Google Cloud DataPrep es un servicio para explorar visualmente, limpiar y preparar data para análisis y procesos de machine Learning. Permite a los usuarios en un esquema de poco código requerido, explorar la data, identificar patrones y potenciales problemas de calidad de data.

1. ***¿Qué cosas que se pueden realizar con Dataprep?***

Se pueden realizar las siguientes actividades:

* Explorar data
* Limpiar data
* Transformar data
* Preparar la información para posterior análisis, como reportes y o procesos de Machine Learning.

1. ***¿Por qué otras herramientas lo podrías reemplazar? ¿Por qué?***

Si entiendo bien la pregunta, por ejemplo, Google Cloud DataPrep, podría fácilmente sustituir herramientas como Tableau, Microsoft SQL Server, IBM Cognos, MicroStrategy entre otras herramientas. Razón del porque haría esto es porque esta herramienta tiene el potencial de combinar varias herramientas dentro de ella dentro de un mismo ecosistema Google Cloud Platform.

1. ***¿Cuáles son los casos de uso más comunes de DataPrep en GCP?***

Se había mencionado anteriormente, entre los posibles usos podemos listar:

* Explorar data de forma visual. Ver tipos data, patrones, anomalías en la data y potenciales problemas en la data. La herramienta te brinda estadísticos y visualizaciones para poder entender inicialmente la data.
* Limpiar la data, como por ejemplo corregir errores, lidiar con valores nulos y resolver inconsistencias en la data. Remover duplicados, cambiar de tipo de datos entre otros.
* Enriquecer la data. Como por ejemplo campos calculados, extraer información adicional de otras fuentes.
* Generar reportes a partir de una data limpia.

1. ***¿Cómo se carga los datos en DataPrep de GCP?***

En DataPrep se puede elegir la data a conectar mediante la creación de un dataset. Cuando se está definiendo el dataset nos podemos conectar a Google Cloud Storage o bien podemos conectarnos directamente a una tabla de BigQuery o subir archivos desde la computadora. Una vez que se crea el dataset, se proceder a crear un DataFlow.

1. ***¿Qué tipo de datos se pueden preparar en DataPrep en GCP?***

Basado en lo visto en el LAB, se pueden preparar datos números y datos categóricos.

1. ***¿Qué pasos se pueden seguir para limpiar y transformar datos en DataPrep de GCP?***

Una vez creado el dataset y se define el dataflow con un “recipe”, DataPrep genera un dashboard como un limitado número de registros del dataset (sampling). En este dashboard se presenta indicadores de distribución, valores nulos o faltantes y los tipos de datos. Seleccionando las columnas de interés, se puede definir reglas para limpiar, asignar, borrar y transformar datos. Por lo general se generan recomendaciones generadas por inteligencia artificial sobre que hacer con la columna seleccionada.

1. ***¿Cómo se pueden automatizar tareas de preparación de datos en DataPrep de GCP?***

Por lo general, en el “recipe” creado están definidas todas las reglas. Adicionalmente se puede especificar la frecuencia de corrida del flujo de datos.

1. ***¿Qué tipo de visualizaciones se pueden crear en DataPrep de GCP?***

Por lo que vi en el LAB no se pueden hacer visualizaciones directamente en DataPrep, pero se puede usar otra herramienta como Google Looker para generar gráficos o reportes tipo Tablau.

1. ***¿Cómo se puede garantizar la calidad de datos en DataPrep de GCP?***

En los “recipe” se puede limpiar la data y corregir errores, lidiar con valores nulos y resolver inconsistencias en la data. Remover duplicados, cambiar de tipo de datos entre otros.

#### **Arquitectura.**

En base a lo requerido, esta podría ser la recomendación.

1. Para almacenamiento de datos se puede usar Google Cloud Storage.
2. Para ingestar datos se puede usar el servicio de BigQuery DataTransfer Service.
3. Para proceder los datos se puede usar Google Cloud Dataproc
4. Una herramienta para BI, se puede usar looker.
5. Finalmente una herramienta para podemos simple se puede usar BigQueryML