

# 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:

- a. *aeropuerto\_tabla*
- b. *aeropuerto\_detalle\_tabla*

A continuación, los screen shots de ambas tablas:

### *aeropuerto\_tabla*

Column Name	#	Data Type	Length	Scale	Not Null	Auto Generated	Auto Increment	Default	Description
fecha	1	DATE	10		[ ]	[ ]	[ ]		
hora_aut	2	STRING			[ ]	[ ]	[ ]		
clase_de_vuelo	3	STRING			[ ]	[ ]	[ ]		
clasificacion_de_vuelo	4	STRING			[ ]	[ ]	[ ]		
tipo_de_movimiento	5	STRING			[ ]	[ ]	[ ]		
aeropuerto	6	STRING			[ ]	[ ]	[ ]		
origen_destino	7	STRING			[ ]	[ ]	[ ]		
aerolinea_nombre	8	STRING			[ ]	[ ]	[ ]		
aeronave	9	STRING			[ ]	[ ]	[ ]		
pasajeros	10	INT	10		[ ]	[ ]	[ ]		

### *aeropuerto\_detalle\_tabla*

Column Name	#	Data Type	Length	Scale	Not Null	Auto Generated	Auto Increment	Default	Description
aeropuerto	1	STRING			[ ]	[ ]	[ ]		
oac	2	STRING			[ ]	[ ]	[ ]		
iata	3	STRING			[ ]	[ ]	[ ]		
tipo	4	STRING			[ ]	[ ]	[ ]		
denominacion	5	STRING			[ ]	[ ]	[ ]		
coordenadas	6	STRING			[ ]	[ ]	[ ]		
latitud	7	STRING			[ ]	[ ]	[ ]		
longitud	8	STRING			[ ]	[ ]	[ ]		
elev	9	FLOAT	7	7	[ ]	[ ]	[ ]		
uom_elev	10	STRING			[ ]	[ ]	[ ]		
ref	11	STRING			[ ]	[ ]	[ ]		
distancia_ref	12	FLOAT	7	7	[ ]	[ ]	[ ]		
direccion_ref	13	STRING			[ ]	[ ]	[ ]		
condicion	14	STRING			[ ]	[ ]	[ ]		
control	15	STRING			[ ]	[ ]	[ ]		
region	16	STRING			[ ]	[ ]	[ ]		
uso	17	STRING			[ ]	[ ]	[ ]		
trafico	18	STRING			[ ]	[ ]	[ ]		
sna	19	STRING			[ ]	[ ]	[ ]		
concesionado	20	STRING			[ ]	[ ]	[ ]		
provincia	21	STRING			[ ]	[ ]	[ ]		

### 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

```
1 from datetime import datetime, timedelta
2 from airflow import DAG
3 from airflow.operators.bash import BashOperator
4 from airflow.operators.python import PythonOperator
5 import subprocess
6
7 # Default arguments for the DAG
8 default_args = {
9     'owner': 'airflow',
10     'depends_on_past': False,
11     'start_date': datetime(2025, 6, 20),
12     'email_on_failure': False,
13     'email_on_retry': False,
14     # 'retries': 1,
15     # 'retry_delay': timedelta(minutes=5),
16 }
17
18 # Let's define the DAG
19 dag = DAG(
20     dag_id='exercise_1_ingest_pyspark_dag',
21     default_args=default_args,
22     description='DAG that runs shell script then PySpark job',
23     schedule_interval='@daily',
24     catchup=False,
25     tags=['spark', 'shell', 'ingest', 'pipeline'],
26 )
27
28 def run_shell_script():
29     result = subprocess.run(['/bin/bash', '/home/hadoop/scripts/job1.sh'],
30                             capture_output=True, text=True)
31     print(f"Return code: {result.returncode}")
32     print(f"Output: {result.stdout}")
33     if result.stderr:
34         print(f"Error: {result.stderr}")
35     if result.returncode != 0:
36         raise Exception(f"Script failed with return code {result.returncode}")
37
38 run_script_task = PythonOperator(
39     task_id='run_job1_script',
40     python_callable=run_shell_script,
41     dag=dag,
42 )
43
44 run_pyspark_job = BashOperator(
45     task_id='run_pyspark_job',
46     bash_command='sshpass -p "edvai" ssh hadoop@localhost /home/hadoop/spark/bin/spark-submit --files /home/hadoop/hive/conf/hive-site.xml /home/hadoop/scripts/pyspark_shell.py', # Update this path
```

Este DAG contiene dos procesos

- Un primer proceso corre el archivo .sh que tiene como propósito la ingesta de los archivos y almacenado en HDFS
- 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**

```
# Add Hadoop environment setup at the top of job1.sh
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64 # Adjust path as needed
export HADOOP_HOME=/home/hadoop/hadoop # Adjust path as needed
export PATH=$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PATH

# Variables
FILE_1_URL="https://data-engineer-edvai-public.s3.amazonaws.com/2021-informe-ministerio.csv"
FILE_2_URL="https://data-engineer-edvai-public.s3.amazonaws.com/202206-informe-ministerio.csv"
FILE_3_URL="https://data-engineer-edvai-public.s3.amazonaws.com/aeropuertos_detalle.csv"

# Downloading file number 1 from site
echo "Downloading File Number 1"
wget -O /home/hadoop/landing/informe_1.csv $FILE_1_URL

# Check if first file was downloaded successfully
if [ $? -eq 0 ]; then
    echo "File number 1 successfully downloaded"
else
    echo "Error downloading File number 1"
    exit 1
fi

# Downloading file number 2 from site
echo "Downloading File Number 2"

wget -O /home/hadoop/landing/informe_2.csv $FILE_2_URL

# Check if second file was downloaded successfully
if [ $? -eq 0 ]; then
    echo "File number 2 successfully downloaded"
else
    echo "Error downloading File number 2"
    exit 1
fi

# Downloading file number 2 from site
echo "Downloading File Number 2"

wget -O /home/hadoop/landing/aeropuerto.csv $FILE_3_URL

# Check if third file was downloaded successfully
if [ $? -eq 0 ]; then
```

```

    echo "Error downloading File number 2"
    exit 1
fi

# Downloading file number 2 from site
echo "Downloading File Number 2"

wget -O /home/hadoop/landing/aeropuerto.csv $FILE_3_URL

# Check if third file was downloaded successfully
if [ $? -eq 0 ]; then
    echo "File number 3 successfully downloaded"
else
    echo "Error downloading File number 3"
    exit 1
fi

echo "Moving files to ingest directory"

echo "Sending files to HDFS..."

hdfs dfs -put /home/hadoop/landing/aeropuerto.csv /ingest
hdfs dfs -put /home/hadoop/landing/informe_1.csv /ingest
hdfs dfs -put /home/hadoop/landing/informe_2.csv /ingest

if [ $? -eq 0 ]; then
    echo "Files successfully moved to HDFS!!"
else
    echo "Error moving files to HDFS"
    exit 1
fi

```

#### **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

```

from pyspark.sql.functions import to_date
from pyspark.sql.functions import coalesce,col
import sys
from pyspark.sql import SparkSession

def main():

    #This works
    spark = SparkSession.builder.master("spark://localhost:7077").appName("HiveLocalConnection").config("hive.metastore.uris", "thrift://localhost:9083").enableHiveSupport().getOrCreate()

    # spark = SparkSession.builder.master("spark://localhost:7077").getOrCreate()

    try:
        print("Starting the Spark session...")

        #PARTE CERO - AEROPUERTOS DETALLES
        print("Processing aeropuerto.csv...")
        df = spark.read.option("header","true").csv("hdfs://172.17.0.2:9000/ingest/aeropuerto.csv",sep=',')
        cols_drop = ['fin','inhab']
        df_select = df.drop(*cols_drop)
        df_select = df.select.withColumn('distancia_ref',df_select['distancia_ref'].cast('float'))
        df_select = df_select.withColumn('elev',df_select['elev'].cast('float'))
        df_select = df_select.na.fill(0,subset=['distancia_ref'])
        df_select = df_select.na.fill(0,subset=['elev'])
        new_col_names = ['aeropuerto','oac','iata','tipo','denominacion','coordenadas','latitud','longitud','elev','uom_elev','ref','distancia_ref','direccion_ref','condicion','control','region','uso','trafico']
        df_select = df_select.toDF(*new_col_names)
        df_select.write.mode('append').insertInto('airport_trips_data.aeropuerto_detalle_tabla')

        print("File aeropuerto.csv processed successfully.")

        #PARTE 1 - INFORME 1
        print("Processing informe_1.csv...")
        df2 = spark.read.option("header","true").csv("hdfs://172.17.0.2:9000/ingest/informe_1.csv",sep=',')
        cols_drop = ['Calidad dato']
        df2_select = df2.drop(*cols_drop)
        df2_select = df2_select.withColumn('Fecha',coalesce(
            to_date(col('fecha'),'MM/dd/yyyy'),
            to_date(col('fecha'),'dd/MM/yyyy'),
            to_date(col('fecha'),'yyyy-MM-dd')))
        df2_select = df2_select.withColumn('Pasajeros',df2_select['Pasajeros'].cast('int'))
        df2_select = df2_select.na.fill(0,subset=['Pasajeros'])
        new_col_names = ['fecha','horaUTC','clase_de_vuelo','clasificacion_de_vuelo','tipo_de_movimiento','aeropuerto','origen_destino','aerolinea_nombre','aeronave','pasajeros']
        df2_select = df2_select.toDF(*new_col_names)
        df2_select = df2_select.filter(df2_select['clasificacion_de_vuelo'].isin(["Domestico","Doméstico"]))
        df2_select.write.mode('append').insertInto('airport_trips_data.aeropuerto_tabla')

```

## 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`

```

tripdata
tripsdb
Time taken: 0.412 seconds, Fetched: 3 row(s)
hive> create database airport_trips_data;
OK
Time taken: 0.575 seconds
hive> show databases;
OK
airport_trips_data
default
tripdata
tripsdb
Time taken: 0.027 seconds, Fetched: 4 row(s)
hive>

```

## Creación de la tabla aeropuerto\_tabla

```
OK
Time taken: 0.055 seconds
hive> create table if not exists aeropuerto_tabla (fecha DATE,horaUTC STRING,clase_de_vuelo STRING,clasificacion_de_vuelo STRING,
tipo_de_movimiento STRING,aeropuerto STRING, origen_destino STRING, aerolinea_nombre STRING, aeronave STRING,
pasajeros INT);
OK
Time taken: 0.361 seconds
hive> show tables;
OK
aeropuerto_tabla
Time taken: 0.04 seconds, Fetched: 1 row(s)
hive> describe aeropuerto_tabla;
OK
fecha                date
horautc              string
clase_de_vuelo       string
clasificacion_de_vuelo string
tipo_de_movimiento   string
aeropuerto           string
origen_destino       string
aerolinea_nombre     string
aeronave             string
pasajeros            int
Time taken: 0.066 seconds, Fetched: 10 row(s)
hive> |
```

## Creación de la tabla aeropuerto\_detalle\_tabla;

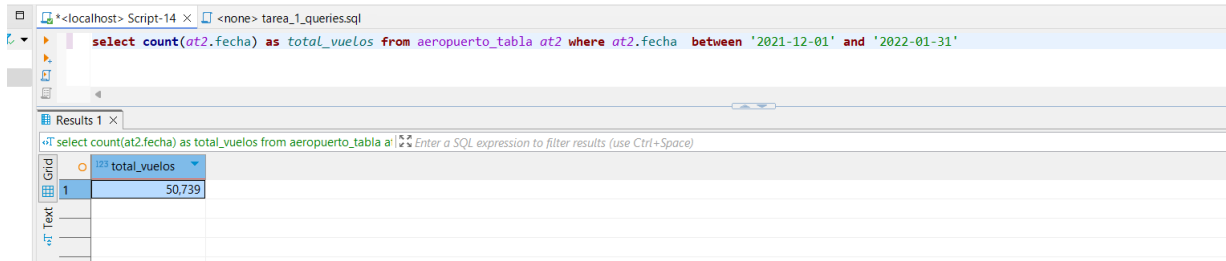
```
pasajeros            int
Time taken: 0.066 seconds, Fetched: 10 row(s)
hive> create table if not exists aeropuerto_detalle_tabla (aeropuerto STRING, oac STRING, Display all 574 possibilities
? (y or n)
hive> create table if not exists aeropuerto_detalle_tabla (aeropuerto STRING, oac STRING, iata STRING,tipo STRING,denom
inacion STRING, coordenadas STRING, latitud STRING, longitud STRING, elev FLOAT, uom_elev STRING, ref STRING, distancia_
ref FLOAT, direccion_ref STRING, condicion STRING, control STRING, region STRING, uso Display all 574 possibilities? (y
or n)
hive> create table if not exists aeropuerto_detalle_tabla (aeropuerto STRING, oac STRING, iata STRING,tipo STRING,denom
inacion STRING, coordenadas STRING, latitud STRING, longitud STRING, elev FLOAT, uom_elev STRING, ref STRING, distancia_
ref FLOAT, direccion_ref STRING, condicion STRING, control STRING, region STRING, uso STRING, trafico STRING, sna Displa
y all 574 possibilities? (y or n)
hive> create table if not exists aeropuerto_detalle_tabla (aeropuerto STRING, oac STRING, iata STRING,tipo STRING,denom
inacion STRING, coordenadas STRING, latitud STRING, longitud STRING, elev FLOAT, uom_elev STRING, ref STRING, distancia_
ref FLOAT, direccion_ref STRING, condicion STRING, control STRING, region STRING, uso STRING, trafico STRING, sna STRING
, concesionado STRING, Display all 574 possibilities? (y or n)
hive> create table if not exists aeropuerto_detalle_tabla (aeropuerto STRING, oac STRING, iata STRING,tipo STRING,denom
inacion STRING, coordenadas STRING, latitud STRING, longitud STRING, elev FLOAT, uom_elev STRING, ref STRING, distancia_
ref FLOAT, direccion_ref STRING, condicion STRING, control STRING, region STRING, uso STRING, trafico STRING, sna STRING
, concesionado STRING, provincia STRING);
OK
Time taken: 0.138 seconds
hive> show tables;
OK
aeropuerto_detalle_tabla
```

```
hive
hive> show tables;
OK
aeropuerto_detalle_tabla
aeropuerto_tabla
Time taken: 0.039 seconds, Fetched: 2 row(s)
hive> describe aeropuerto_detalle_tabla;
OK
aeropuerto      string
oac              string
iata             string
tipo            string
denominacion     string
coordenadas     string
latitud         string
longitud        string
elev            float
uom_elev        string
ref             string
distancia_ref   float
direccion_ref   string
condicion       string
control         string
region          string
uso             string
trafico         string
sna             string
concesionado    string
provincia       string
Time taken: 0.071 seconds, Fetched: 21 row(s)
hive> |
```

hive> create table if not exists aeropuerto\_detalle\_tabla (aeropuerto string, oac string, Display all

## 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***



The screenshot shows a SQL query editor with the following query:

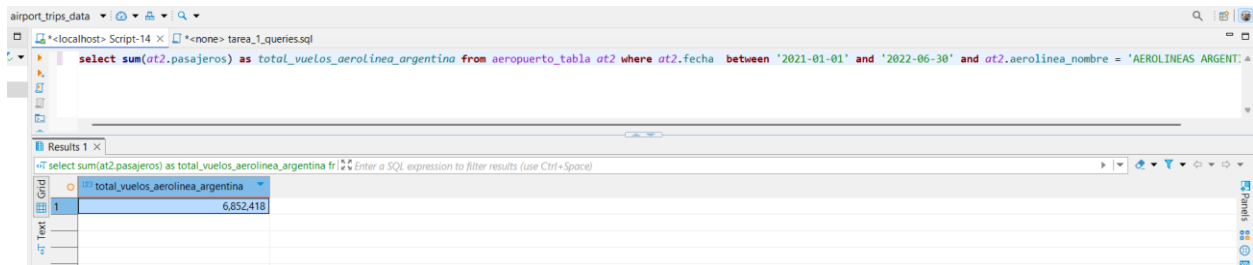
```
select count(at2.fecha) as total_vuelos from aeropuerto_tabla at2 where at2.fecha between '2021-12-01' and '2022-01-31'
```

The results are displayed in a grid:

	total_vuelos
1	50,739

## 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***



The screenshot shows a SQL query editor with the following query:

```
select sum(at2.pasajeros) as total_vuelos_aerolinea_argentina from aeropuerto_tabla at2 where at2.fecha between '2021-01-01' and '2022-06-30' and at2.aerolinea_nombre = 'AEROLINEAS ARGENTINA'
```

The results are displayed in a grid:

	total_vuelos_aerolinea_argentina
1	6,852,418

## 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.



Report: trips\_data

```

--select at2.fecha,at2.hora,at2.aeropuerto as aeropuerto_salida,q1.provincia as ciudad_salida,at2.pasajeros,at2.origen_destino as aeropuerto_arribo,q2.provincia as ciudad_arribo from
--join (select at2.aeropuerto,adt.provincia from aeropuerto_tabla at2
--join aeropuerto_detalle_tabla adt on at2.aeropuerto = adt.aeropuerto
--group by at2.aeropuerto,adt.provincia) q1
--on at2.aeropuerto = q1.aeropuerto
--join (select at2.origen_destino ,adt.provincia from aeropuerto_tabla at2
--join aeropuerto_detalle_tabla adt on at2.origen_destino = adt.aeropuerto
--group by at2.origen_destino ,adt.provincia) q2
--on at2.origen_destino = q2.origen_destino
--where at2.fecha between '2021-01-01' and '2022-06-30' and at2.tipo_de_movimiento = 'Despegue' and at2.pasajeros>0
--order by at2.fecha desc

```

Results 1 x

Enter a SQL expression to filter results (use Ctrl+Space)

	fecha	hora	aeropuerto_salida	ciudad_salida	pasajeros	aeropuerto_arribo	ciudad_arribo
1	2022-06-30	23:39	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	87	SAL	SALTA
2	2022-06-30	23:47	CBA	CÓRDOBA	90	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES
3	2022-06-30	00:44	CBA	CÓRDOBA	86	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES
4	2022-06-30	00:47	ROS	SANTA FÉ	11	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES
5	2022-06-30	00:50	SAL	SALTA	57	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES
6	2022-06-30	01:02	SVO	SANTA FÉ	3	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES
7	2022-06-30	01:12	BAR	RÍO NEGRO	84	EZE	BUENOS AIRES
8	2022-06-30	01:56	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	73	GRA	TIERRA DEL FUEGO ANTÁRTIDA E ISLAS DEL ATLÁNTICO SUR
9	2022-06-30	02:10	SDE	SANTIAGO DEL ESTERO	45	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES
10	2022-06-30	07:08	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	15	NEU	NEUQUÉN
11	2022-06-30	07:33	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	62	GRA	TIERRA DEL FUEGO ANTÁRTIDA E ISLAS DEL ATLÁNTICO SUR
12	2022-06-30	08:03	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	86	ECA	SANTA CRUZ
13	2022-06-30	08:24	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	88	NEU	NEUQUÉN
14	2022-06-30	08:29	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	15	BCA	BUENOS AIRES
15	2022-06-30	08:38	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	86	BAR	RÍO NEGRO
16	2022-06-30	08:52	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	47	JUJ	JUJUY
17	2022-06-30	09:05	AER	CIUDAD AUTÓNOMA DE BUENOS AIRES	91	DOZ	MENDOZA
18	2022-06-30	10:22	EZE	BUENOS AIRES	79	TRE	CHIRIQUIT

## 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.

Report: trips\_data

```

--select at2.aerolinea_nombre,sum(at2.pasajeros) as numero_pasejeros from aeropuerto_tabla at2 where at2.aerolinea_nombre is not NULL and at2.aerolinea_nombre not in ('0') and at2.fecha between '2021-01-01' and '2022-06-30'
--group by at2.aerolinea_nombre order by numero_pasejeros desc LIMIT 10

```

Results 1 x

Enter a SQL expression to filter results (use Ctrl+Space)

	aerolinea_nombre	numero_pasejeros
1	AEROLINEAS ARGENTINAS SA	6,852,418
2	JETSMART AIRLINES SA	1,378,001
3	IB LINEAS AEREAS - FLYBONDI	1,364,305
4	AMERICAN JET SA	24,177
5	L.A.D.E.	14,437
6	BAIRES FLY SA	4,664
7	FUERZA AEREA ARGENTINA	3,730
8	LADE	3,191
9	FUERZA AEREA ARGENTINA (FAA)	3,066
10	FLYING AMERICA SA	2,732

## 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.**

The screenshot shows a SQL query in a database interface. The query is as follows:

```
select at2.aeronave, count(at2.aeronave) as numero_vuelos from aeropuerto_tabla at2
join aeropuerto_detalle at1 on at2.aeropuerto = at1.aeropuerto
where at2.fecha between '2021-01-01' and '2022-06-30' and at2.provincia in ('BUENOS AIRES', 'CIUDAD AUTÓNOMA DE BUENOS AIRES') and at2.tipo_de_movimiento = 'Despegue' and at2.aeronave not in ('0')
GROUP BY at2.aeronave order by numero_vuelos desc LIMIT 10
```

The results table shows the top 10 aircraft by flight count:

Rank	aeronave	numero_vuelos
1	EMB-ER190100KJW	11,587
2	CE-150-L	7,486
3	CE-152	7,437
4	CE-150-M	5,703
5	AIB-A320-232	4,839
6	BO-737-800	4,150
7	CE-150-G	2,722
8	CE-150-J	2,665
9	BO-737-800	2,468
10	PA-PA-28-181	2,257

## Punto Número 11

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

- Número de vuelo.** 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.
- 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.
- Otro dato pudiera ser **horas de vuelo de la aeronave** y **rutinas de mantenimiento preventivo y/o correctivo**.
- 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.
- 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.
- 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:

- a. Google Cloud Storage para almacenar la data
- b. Google Dataproc
- c. 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**

```
WARNING: All illegal access operations will be denied in a future release
hive> show databases;
OK
airport_trips_data
car_rental_db
default
tripdata
Time taken: 0.997 seconds, Fetched: 4 row(s)
hive>
```

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

```
FAILED: SemanticException [Error 10072]: Database does not exist: car_rental_db
hive> use car_rental_db;
OK
Time taken: 0.055 seconds
hive> show tables;
OK
car_rental_analytics
Time taken: 0.14 seconds, Fetched: 1 row(s)
hive>
```

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

```
hive> describe car_rental_analytics;
OK
fueltype          string
rating            int
rentertripstaken  int
reviewcount       int
city              string
state_name        string
owner_id          int
rate_daily        int
make              string
model             string
year              int
Time taken: 0.056 seconds, Fetched: 11 row(s)
hive>
```

### 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**.

```
>_ job2_final.sh X
Tarea_2 >>_ job2_final.sh
1  #!/bin/bash
2
3  # Add Hadoop environment setup at the top of job1.sh
4  export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64 # Adjust path as needed
5  export HADOOP_HOME=/home/hadoop/hadoop # Adjust path as needed
6  export PATH=$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PATH
7
8  # Variables
9  FILE_1_URL="https://data-engineer-edvai-public.s3.amazonaws.com/CarRentalData.csv"
10 FILE_2_URL="https://data-engineer-edvai-public.s3.amazonaws.com/georef-united-states-of-america-state.csv"
11
12 # Downloading file number 1 from site
13 echo "Downloading File Number 1"
14 wget -O /home/hadoop/landing/cars_data.csv $FILE_1_URL
15
16 # Check if first file was downloaded successfully
17 if [ $? -eq 0 ]; then
18     echo "File number 1 successfully downloaded"
19 else
20     echo "Error downloading File number 1"
21     exit 1
22 fi
23
24 # Downloading file number 2 from site
25 echo "Downloading File Number 2"
26
27 wget -O /home/hadoop/landing/geo_data.csv $FILE_2_URL
28
29 # Check if second file was downloaded successfully
30 if [ $? -eq 0 ]; then
31     echo "File number 2 successfully downloaded"
32 else
33     echo "Error downloading File number 2"
34     exit 1
35 fi
36
37 echo "Moving files to ingest directory"
38
39
40
41 echo "Sending files to HDFS..."
42
43 hdfs dfs -put /home/hadoop/landing/cars_data.csv /ingest
44 hdfs dfs -put /home/hadoop/landing/geo_data.csv /ingest
45
46
47 if [ $? -eq 0 ]; then
48     echo "Files successfully moved to HDFS!!"
49 else
50     echo "Error moving files to HDFS"
51     exit 1
52 fi
53
```

### 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**

```

1 from pyspark.sql import SparkSession
2 from pyspark.sql.functions import round
3 from pyspark.sql.functions import lower, col
4 import sys
5
6 def main():
7     spark = SparkSession.builder.master("spark://localhost:7077").appName("HiveLocalConnection").config("hive.metastore.uris", "thrift://localhost:9083").enableHiveSupport().getOrCreate()
8     # spark = SparkSession.builder.master("spark://localhost:7077").getOrCreate()
9     try:
10         print("Starting the Spark session...")
11         print("Processing cars data csv and geo data csv files...")
12
13         df = spark.read.option("header", "true").csv("hdfs://172.17.0.2:9000/ingest/cars_data.csv", sep=',')
14         df_geo = spark.read.option("header", "true").csv("hdfs://172.17.0.2:9000/ingest/geo_data.csv", sep=',')
15
16         print("File cars_data.csv and geo_data.csv read successfully.")
17
18         print("Processing data...")
19
20         cols_to_drop = ['location.latitude', 'location.country', 'location.longitude', 'vehicle.type']
21         df_select = df.drop(*cols_to_drop)
22         df_geo_select = df_geo.select("United States Postal Service state abbreviation", "Official Name State")
23         df_geo_select = df_geo_select.withColumnRenamed("United States Postal Service state abbreviation", "state_code").withColumnRenamed("Official Name State", "state_name")
24
25         new_col_names = ['fueltype', 'rating', 'rentertripstaken', 'reviewcount', 'city', 'state_code', 'owner_id', 'rate_daily', 'make', 'model', 'year']
26         df_select = df_select.toDF(*new_col_names)
27
28
29         print("Data processing completed. Merging dataframes...")
30         merged_df = df_select.join(df_geo_select, "state_code", "left")
31         merged_df = merged_df.filter(merged_df.state_name != "Texas")
32         merged_df = merged_df.withColumn("fueltype", lower(col("fueltype")))
33         merged_df = merged_df.na.drop(subset=['rating'])
34         merged_df = merged_df.withColumn("rating", merged_df['rating'].cast('float'))
35         merged_df = merged_df.withColumn("rentertripstaken", merged_df['rentertripstaken'].cast('int'))
36         merged_df = merged_df.withColumn("owner_id", merged_df['owner_id'].cast('int'))
37         merged_df = merged_df.withColumn("rate_daily", merged_df['rate_daily'].cast('int'))
38         merged_df = merged_df.withColumn("reviewcount", merged_df['reviewcount'].cast('int'))
39         merged_df = merged_df.withColumn("year", merged_df['year'].cast('int'))
40         merged_df = merged_df.withColumn("rating", round(merged_df['rating'], 0).cast('int'))
41         col_to_drop = ['state_code']
42         merged_df = merged_df.drop(*col_to_drop)
43         column_order = ['fueltype', 'rating', 'rentertripstaken', 'reviewcount', 'city', 'state_name', 'owner_id', 'rate_daily', 'make', 'model', 'year']
44         merged_df = merged_df.select(*column_order)
45         print("Dataframes merged successfully.")
46
47         print("Writing the merged dataframe to the car_rental_analytics table...")
48
49         merged_df.write.mode('append').insertInto('car_rental_db.car_rental_analytics')
50
51         print("Data written successfully to the car_rental_analytics table.")
52
53     except Exception as e:
54         print(f"An error occurred: {e}", str(e.args))
55         sys.exit(1)

```

## 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”***

```

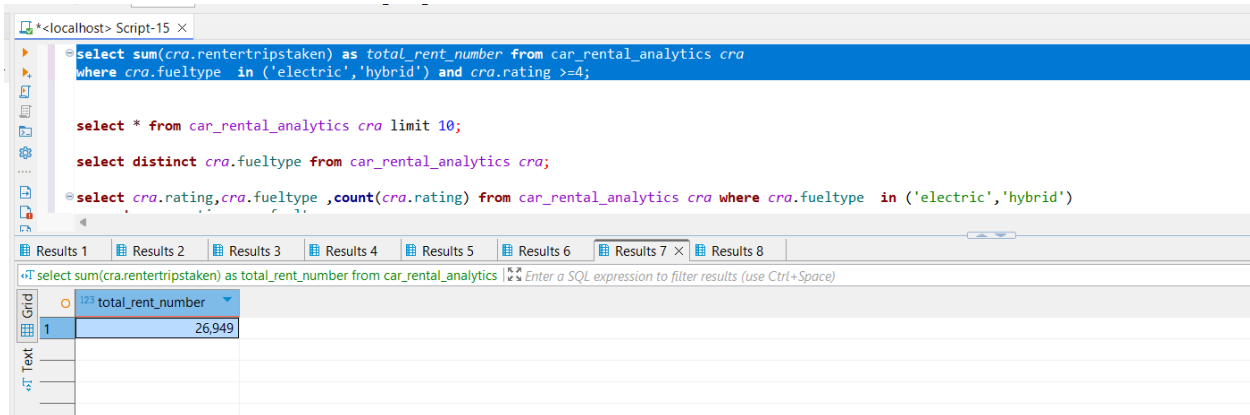
Tarea_2 > dag_exercise2.py > ...
1  from datetime import datetime, timedelta
2  from airflow import DAG
3  from airflow.operators.bash import BashOperator
4  from airflow.operators.python import PythonOperator
5  from airflow.operators.trigger_dagrun import TriggerDagRunOperator
6  import subprocess
7
8  # Default arguments for the DAG
9  default_args = {
10     'owner': 'airflow',
11     'depends_on_past': False,
12     'start_date': datetime(2025, 6, 20),
13     'email_on_failure': False,
14     'email_on_retry': False,
15     'retries': 1,
16     'retry_delay': timedelta(minutes=1),
17 }
18
19 # Let's define the DAG
20 dag_ingest = DAG(
21     dag_id='exercise2-dag-edvai',
22     default_args=default_args,
23     description='DAG that runs shell script then trigger another DAG',
24     schedule_interval='@daily',
25     catchup=False,
26     tags=['spark', 'shell', 'trigger', 'son_dagcat'],
27 )
28
29 def run_shell_script():
30     result = subprocess.run(['/bin/bash', '/home/hadoop/scripts/job2.sh'],
31                             capture_output=True, text=True)
32     print(f"Return code: {result.returncode}")
33     print(f"Output: {result.stdout}")
34     if result.stderr:
35         print(f"Error: {result.stderr}")
36     if result.returncode != 0:
37         raise Exception(f"Script failed with return code {result.returncode}")
38
39 run_script_task = PythonOperator(
40     task_id='run_job2_script',
41     python_callable=run_shell_script,

```

### **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.



```
select sum(cra.rentertripstaken) as total_rent_number from car_rental_analytics cra
where cra.fueltype in ('electric','hybrid') and cra.rating >=4;

select * from car_rental_analytics cra limit 10;

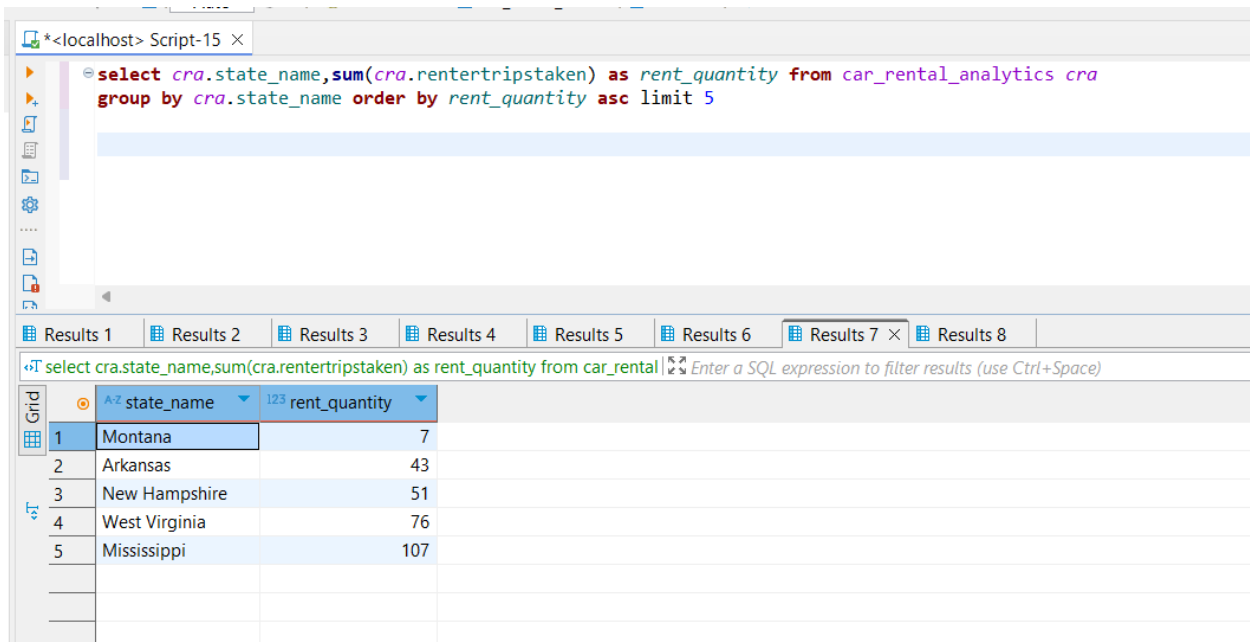
select distinct cra.fueltype from car_rental_analytics cra;

select cra.rating,cra.fueltype ,count(cra.rating) from car_rental_analytics cra where cra.fueltype in ('electric','hybrid')
```

	total_rent_number
1	26,949

### **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”

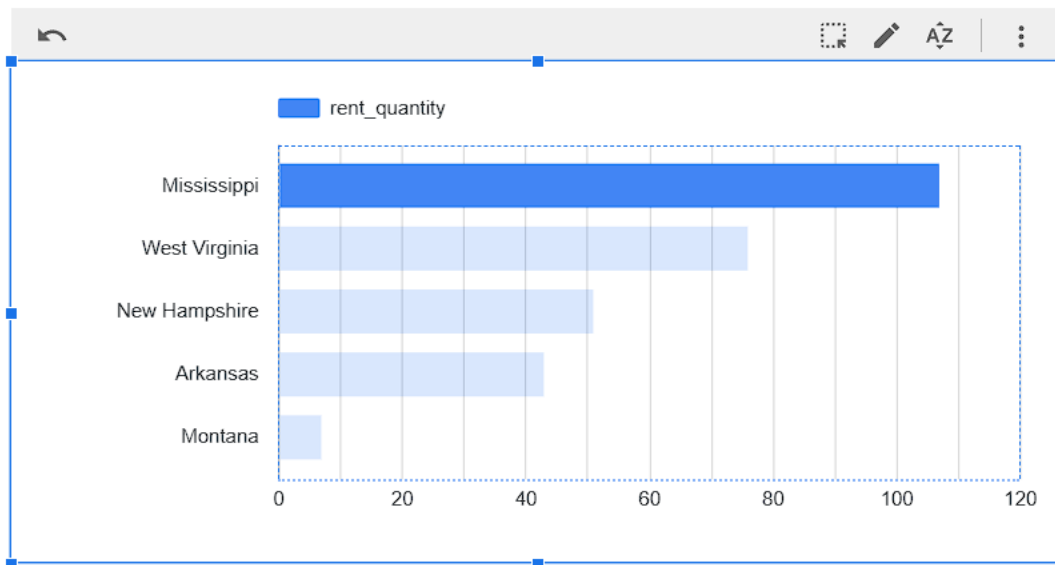


```
select cra.state_name,sum(cra.rentertripstaken) as rent_quantity from car_rental_analytics cra
group by cra.state_name order by rent_quantity asc limit 5
```

	state_name	rent_quantity
1	Montana	7
2	Arkansas	43
3	New Hampshire	51
4	West Virginia	76
5	Mississippi	107



## Gráfica de Looker



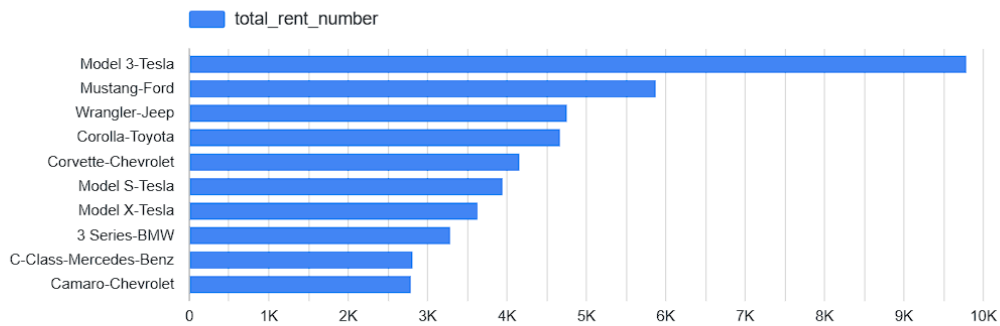
## Punto Número 5.c

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

```
select concat_ws("-",cra.model,cra.make) as model_make,sum(cra.rentertripstaken) as total_rent_number from car_rental_analytics cra group by cra.model,cra.make order by total_rent_number desc limit 10
```

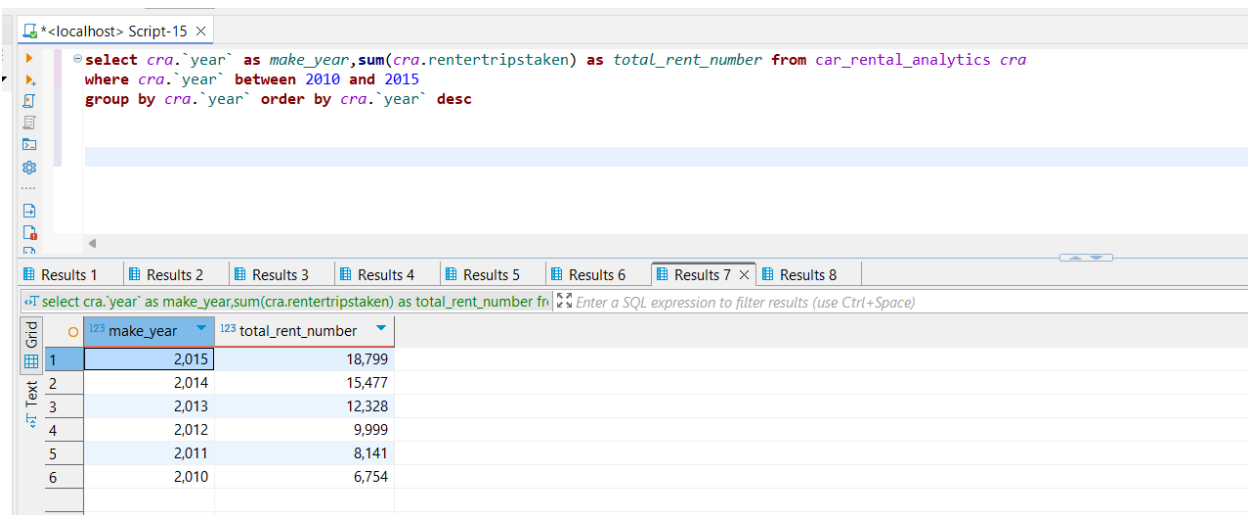
model_make	total_rent_number
Model 3-Tesla	9,794
Mustang-Ford	5,882
Wrangler-Jeep	4,762
Corolla-Toyota	4,676
Corvette-Chevrolet	4,164
Model S-Tesla	3,952
Model X-Tesla	3,638
3 Series-BMW	3,293
C-Class-Mercedes-Benz	2,818
Camaro-Chevrolet	2,797

## Gráfica de Looker



### 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



### 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

localhost Script-15

```
select cra.city,sum(cra.rentertripstaken) as total_rent_number from car_rental_analytics cra
where cra.fueltype in ('electric','hybrid') group by cra.city order by total_rent_number desc limit 5
```

Results 1 Results 2 Results 3 Results 4 Results 5 Results 6 Results 7 Results 8

select cra.city,sum(cra.rentertripstaken) as total\_rent\_number from car\_rental\_analytics cra

	A-z city	total_rent_number
1	San Diego	1,793
2	Las Vegas	1,551
3	Los Angeles	1,075
4	San Francisco	1,058
5	Portland	928

### Punto Número 5.f

--query 6

```
select cra.fueltype,avg(cra.rating) as avg_rating from car_rental_analytics cra where cra.fueltype is not null
group by cra.fueltype order by avg_rating desc
```

Results 1 Results 2 Results 3 Results 4 Results 5 Results 6

select cra.fueltype,avg(cra.rating) as avg\_rating from car\_rental\_analytics cra

	fueltype	avg_rating
1	hybrid	4.9912663755
2	electric	4.9870848708
3	diesel	4.9827586207
4	gasoline	4.9805728518

Refresh Save Cancel Export data 200 4 4 row(s) fetched - 3.578s (0.014s fetch), on 2025-06-22 at 19:22:50

Save: Save 'localhost' queries\_car\_rental\_table' changes... COT en Writable Smart Insert 23:10:949 Sel: 0 | 0

### Punto Número 6

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

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

The screenshot shows the Google Cloud Dataprep interface. At the top, the breadcrumb navigation reads 'Creating a Data Transformation Pipeline with Cloud Dataprep'. The main content area has a green checkmark icon and the text 'Verify if the Cloud Dataprep jobs output the data to BigQuery' and 'Assessment Completed!'. Below this, a message says 'Congratulations! You've successfully explored your ecommerce dataset and created a data pipeline.' On the left, there's a sidebar with a 'Set up console' button, a timer at '00:23:11', and a warning about account security. On the right, a 'Lab instructions and tasks' sidebar lists seven tasks, with the last one, 'Task 7. Running Cloud', highlighted in orange.

a. ¿Para qué se utiliza Dataprep?

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

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

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.

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.

**e. ¿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.

**f. ¿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.

**g. ¿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.

**h. ¿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.

**i. ¿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.

**j. ¿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.

- a. Para almacenamiento de datos se puede usar Google Cloud Storage.
- b. Para ingestar datos se puede usar el servicio de BigQuery DataTransfer Service.
- c. Para proceder los datos se puede usar Google Cloud Dataproc
- d. Una herramienta para BI, se puede usar looker.
- e. Finalmente una herramienta para podemos simple se puede usar BigQueryML