Trabajo Final

Data Enginner

Alumno: Bustamante, Matias Iván

Profesor: Pineyro, Federico

Año: 2025

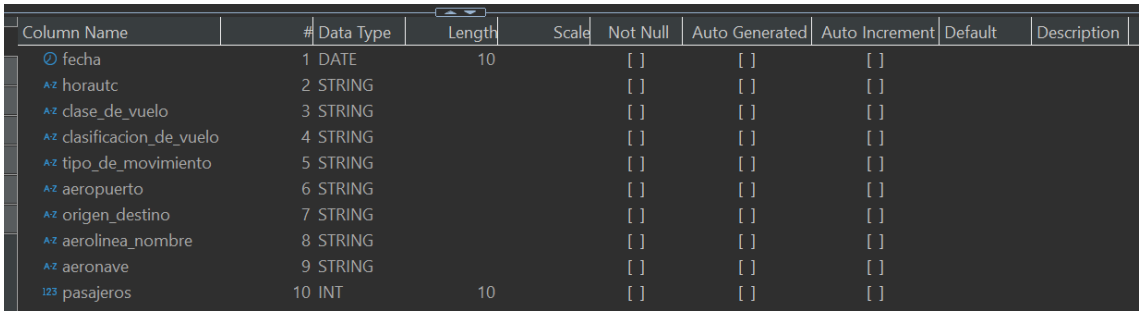
Ejercicio 1:

Aviación Civil

La Administración Nacional de Aviación Civil necesita una serie de informes para elevar al ministerio de transporte acerca de los aterrizajes y despegues en todo el territorio Argentino, como puede ser: cuales aviones son los que más volaron, cuántos pasajeros volaron, ciudades de partidas y aterrizajes entre fechas determinadas, etc. Usted como data engineer deberá realizar un pipeline con esta información, automatizarlo yrealizar los análisis de datos solicitados que permita responder las preguntas de negocio, y hacer sus recomendaciones con respecto al estado actual.

1. Datasets:
    a. https://data-engineer-edvai-public.s3.amazonaws.com/2021-informe-ministerio.csv
    b. https://data-engineer-edvai-public.s3.amazonaws.com/202206-informe-ministerio.csv
    c. https://data-engineer-edvai-public.s3.amazonaws.com/aeropuertos_detalle.csv

2. Estructura de Tabla
    a. Tabla Vuelos:



Column Name	#	Data Type	Length	Scale	Not Null	Auto Generated	Auto Increment	Default	Description
fecha	1	DATE	10		[ ]	[ ]	[ ]		
horautc	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		[ ]	[ ]	[ ]		

Figura 1. Tabla vuelos

b. Tabla detalle de vuelos

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			[ ]	[ ]	[ ]		

Figura 2. Tabla detalle de vuelos

3. Diagrama de orquestación airflow

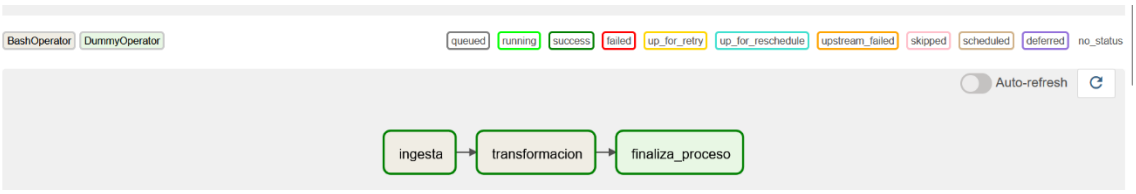


Figura 3. DAG airflow

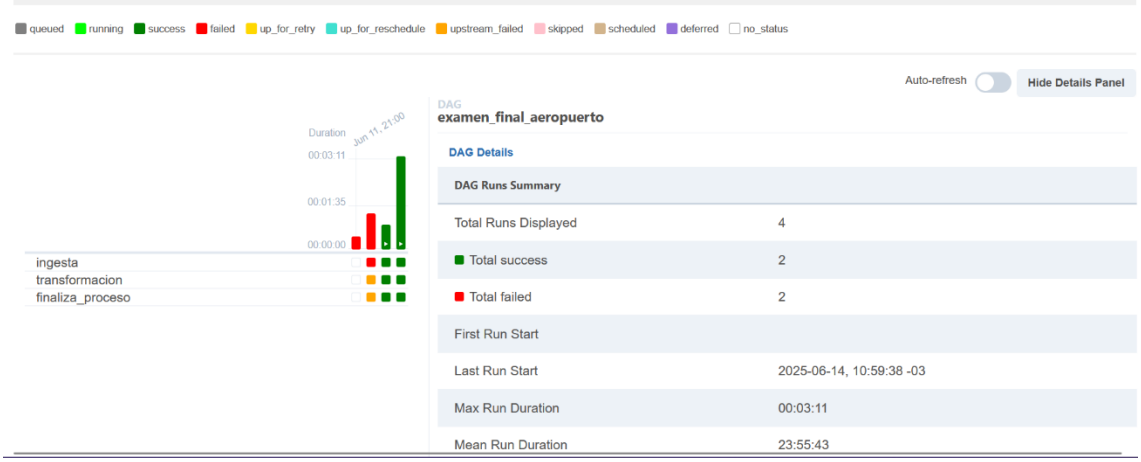


Figura 4. Vista de Grilla

4. Github: https://github.com/Matias Bustamante/TrabajoFinalDE/tree/main/Aeropuerto

5. Tabla de Vuelo

```
CREATE DATABASE aeropuerto
CREATE TABLE aeropuerto.vuelos
(
    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
)
```

Figura 5. Creación de Tabla vuelo

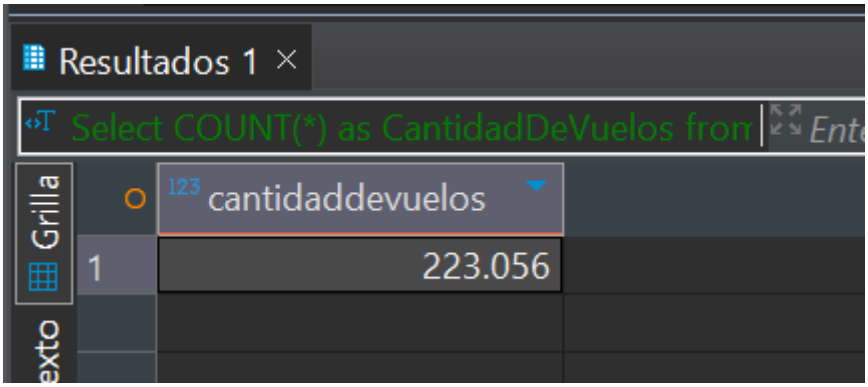
Tabla de Detalle de vuelo

```
CREATE TABLE aeropuerto.detalle_vuelo
(
    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
)
```

Figura 6. Creación de Tabla detalle_vuelo

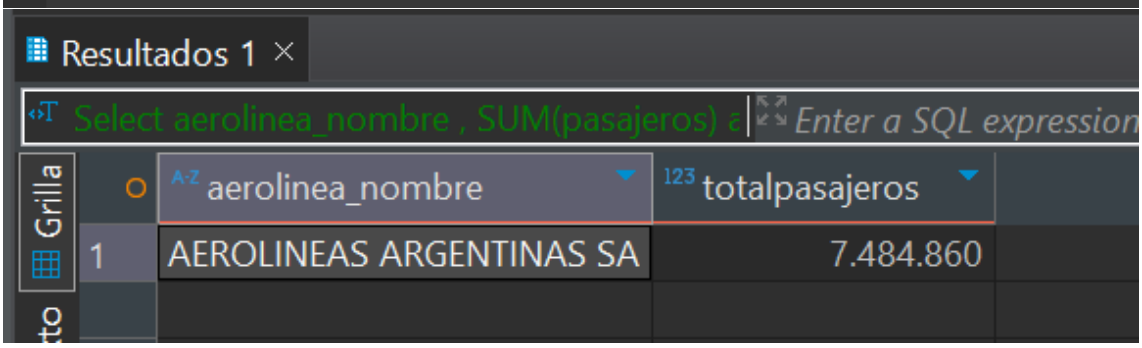
6. Determinar la cantidad de vuelos entre las fechas 01/12/2021 y 31/01/2022. Mostrar consulta y Resultado de la query

```
--Determinar la cantidad de vuelos entre las fechas 01/12/2021 y 31/12/2022

Select COUNT(*) as CantidadDeVuelos from aeropuerto.vuelos v
where fecha between '2021-12-01' and '2022-12-31'
```



7. Cantidad de pasajeros que viajaron en Aerolíneas Argentinas entre el 01/01/2021 y 30/06/2022. Mostrar consulta y Resultado de la query.

```
--Cantidad de pasajeros que viajaron en aerolineas argentinas entre el 01/01/2021 30/06/2022

Select aerolinea_nombre , SUM(pasajeros) as TotalPasajeros from aeropuerto.vuelos
where aerolinea_nombre like '%AEROLINEAS ARGENTINAS SA%' and
fecha between '2021-01-01' and '2022-06-30'
group by aerolinea_nombre
```



8. Mostrar fecha, hora, código aeropuerto salida, ciudad de salida, código de aeropuerto de arribo, ciudad de arribo, y cantidad de pasajeros de cada vuelo, entre el 01/01/2022y el 30/06/2022 ordenados por fecha de manera descendiente. Mostrar consulta y Resultado de la query.

```sql
--Mostrar fecha, hora, código aeropuerto salida, ciudad de salida, código de aeropuerto
--de arribo, ciudad de arribo, y cantidad de pasajeros de cada vuelo, entre el 01/01/2022
--y el 30/06/2022 ordenados por fecha de manera descendiente

Select v.fecha,
v.horautc ,
v.aeropuerto as codigo_salida,
a1.denominacion as ciudad_salida,
v.origen_destino as codigo_arribo,
a2.denominacion as ciudad_arribo,
v.pasajeros  as cantidadPasajero
from aeropuerto.vuelos as v
join aeropuerto.detalle_vuelo a1
on v.aeropuerto =a1.iata
join aeropuerto.detalle_vuelo a2
on v.origen_destino =a2.iata
where v.fecha between '2022-01-01' and '2022-06-30'
order by v.fecha desc
```

Resultados 1

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

	fecha	horautc	codigo_salida	ciudad_salida	codigo_arribo	ciudad_arribo	cantidadpasajero
1	2022-06-30	20:09	ROS	ROSARIO/ISLAS MALVINAS	EZE	EZEIZA/MINISTRO PISTARINI	[NULL]
2	2022-06-30	20:02	CRR	CERES	RES	RESISTENCIA	[NULL]
3	2022-06-30	19:58	CRR	CERES	CRR	CERES	0
4	2022-06-30	19:42	CRR	CERES	CRR	CERES	0
5	2022-06-30	19:38	CRR	CERES	CRR	CERES	0
6	2022-06-30	19:36	FDO	SAN FERNANDO	FDO	SAN FERNANDO	0
7	2022-06-30	19:21	ROS	ROSARIO/ISLAS MALVINAS	FDO	SAN FERNANDO	0

9. Cuales son las 10 aerolíneas que más pasajeros llevaron entre el 01/01/2021 y el 30/06/2022 exceptuando aquellas aerolíneas que no tengan nombre. Mostrar consulta y Visualización.

```sql
--Cuales son las 10 aerolíneas que más pasajeros llevaron entre el 01/01/2021 y el
--30/06/2022 exceptuando aquellas aerolíneas que no tengan nombre.

Select
t.aerolinea,
t.cantidadPasajero,
t.ranking
FROM (
Select
RANK() over(order by SUM(v.pasajeros) desc) as ranking,
v.aerolinea_nombre as aerolinea,
SUM(v.pasajeros) as cantidadPasajero
FROM aeropuerto.vuelos as v
where v.aerolinea_nombre <>"0" and v.fecha between '2021-01-01' and '2022-06-30'
group by v.aerolinea_nombre   ) as t
where t.ranking<=10
```

Resultados 1

Select t.aerolinea, t.cantidadPasajero, t.ranking F | Enter a SQL expression to filter res

Grilla | Texto

	aerolinea	cantidadpasajero	ranking
1	AEROLINEAS ARGENTINAS SA	7.484.860	1
2	JETSMART AIRLINES S.A.	1.511.650	2
3	FB LÍNEAS AÉREAS - FLYBONDI	1.482.473	3
4	AMERICAN JET S.A.	25.789	4
5	L.A.D.E.	15.074	5
6	BAIRES FLY SA	4.960	6
7	LADE	3.895	7
8	FUERZA AEREA ARGENTINA	3.855	8
9	FUERZA AEREA ARGENTINA (FA	3.138	9
10	FLYING AMERICA SA	2.839	10

10. Cuáles son las 10 aeronaves más utilizadas entre el 01/01/2021 y el 30/06/22 que despegaron desde la Ciudad autónoma de Buenos Aires o de Buenos Aires,exceptuando aquellas aeronaves que no cuentan con nombre. Mostrar consulta y Visualización

```
--Cuales son las 10 aeronaves más utilizadas entre el 01/01/2021 y el 30/06/22 que
--despegaron desde la Ciudad autónoma de Buenos Aires o de Buenos Aires,
--exceptuando aquellas aeronaves que no cuentan con nombre.
Select
t.aeronave ,
t.Utilizacion,
t.ranking
FROM (
Select
RANK() OVER(order by COUNT(*) desc) as ranking,
v.aeronave as aeronave,
count(*) as Utilizacion
from aeropuerto.vuelos v
join aeropuerto.detalle_vuelo as dv
on v.aeropuerto=dv.iata and dv.provincia  like '%BUENOS AIRES%'
WHERE v.aeronave <>"0" and v.fecha between '2021-01-01' and '2022-06-30'
group by v.aeronave
order by Utilizacion desc ) as t
where t.ranking <=10
```

	aeronave	utilizacion	ranking
1	PA-PA-28-181	4.865	1
2	EMB-ERJ190100IGW	4.005	2
3	LJ-60	2.455	3
4	CE-152	2.246	4
5	LJ-45	1.931	5
6	RO-R-44RAVEN-II	1.867	6
7	BO-737-800	1.808	7
8	CE-150-L	1.588	8
9	RO-R-44	1.442	9
10	CE-150-J	1.366	10

11. Qué datos externos agregaría en este dataset que mejoraría el análisis de los datos

    Los datos externos que sería recomendable agregar serían los siguientes:

    - Cantidad de vuelos con retrasos y cancelados.
    - Información financiera de cada una de las aerolíneas.
    - Precios promedio de pasajes por ruta y categoría (económica, bussiness, primera clase).
    - Tiempo promedio de atención en mostradores y seguridad (check-in, migraciones, seguridad).
    - Tiempo de conexión o escala entre vuelos.
    - Frecuencia y puntualidad por aerolineas y destinos.
    - Capacidad y uso de puertas de embarques y pistas.
    - Carga transportada (tonelada por vuelos o rutas).
    - Demanda histórica y proyección de pasajeros por rutas.
    - Gasto promedio por pasajeros en tiendas, gastronomía y servicios dentro del aeropuerto.
    - índice de ocupación hotelera cercana al aeropuerto.
    - Ingresos por concesiones (comercio, estacionamiento, publicidad)
    - Condiciones climáticas históricas y en tiempo real por aeropuerto.
    - Índices de calidad del aire en la zona del aeropuerto.

- Conectividad terrestre: frecuencia de colectivos, trenes, taxis, uber, entre otros.
- Distancia y tiempo estimado a punto clave de la ciudad.
- Perfil de los pasajeros (edad, género, nacionalidad, motivo del viaje).
- Volumen de pasajeros corporativos vs turistas
- Comparación de tarifas aeroportuarias.
- Índice de satisfacción de los pasajeros.

12. Elabore sus conclusiones y recomendaciones sobre este proyecto.

Análisis y recomendaciones:

El análisis de los datos de vuelos entre el 01/01/2022 y el 30/06/2022 revela una significativa cantidad de operaciones registradas con cero (0) pasajeros. Confirmamos que estos registros corresponden a **vuelos de naturaleza privada**.
Los vuelos privados, aunque no transporten pasajeros comerciales, representan una parte activa y relevante de la utilización de la infraestructura aeroportuaria y del espacio aéreo. Su alta frecuencia (como se observa en los datos) tiene implicaciones directas en la gestión del tráfico, la asignación de slots, el uso de servicios terrestres y, potencialmente, en la generación de ingresos indirectos para la Administración.
Se sugiere a la Administración implementar las siguientes acciones para optimizar la gestión y maximizar el valor de la operación de vuelos privados:

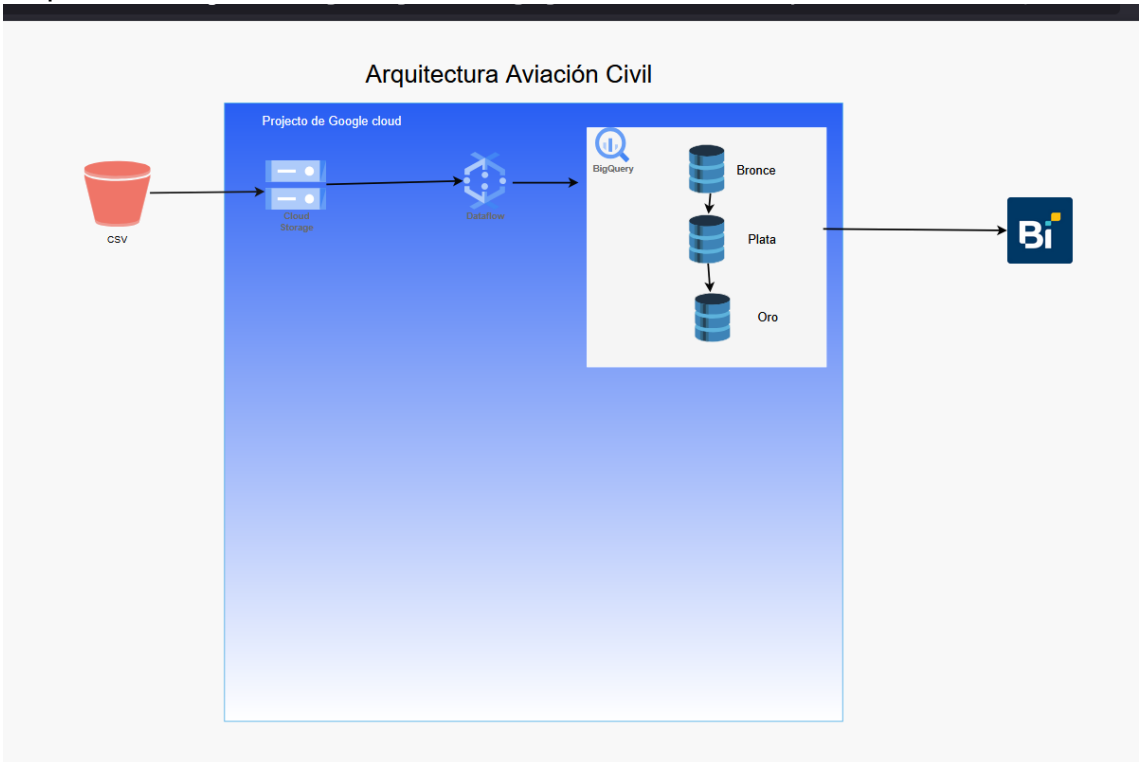
1. Revisar y ajustar la estructura de tarifas y cargos asociados a los vuelos privados (ejemplo: tasa de aterrizajes, estacionamiento, servicios específicos para aviación ejecutiva). Asegurar que reflejen el valor del servicio prestado y el uso de la infraestructura.
2. Evaluar la demanda y ofertas de servicios premium para la aviación privada (Salones VIP, servicios de conserjería, despacho aduanero) que puedan generar ingresos adicionales y mejorar la experiencia de los usuarios.
3. Considerar las necesidades de los vuelos privados en la planificación a largo plazo de la infraestructura aeroportuaria, incluyendo plataformas de estacionamiento y facilidades de mantenimiento especificas para este segmento.

Por otra parte, analizando las aeronaves que más han sido utilizada se encuentran las siguientes: **PA-PA-28-181, EMB-ERJ190100IGW.** La aeronave PA-PA-28-181 se trata de un avión utilitario y de entrenamiento, esta aeronave es considerado critico por la importancia en la formación aeronáutica y otras actividades que no involucra el traslado de pasajero comerciales. La recomendación detenida a la

Administración para la optimización de esta familia de aeronave son las siguiente:

1. Establecer un programa de mantenimiento preventivo y predictivo más riguroso. Esto conlleva a implementar mecanismo de monitoreo continuo de cada componente críticos.
2. Analizar si la demanda de entrenamiento y usos utilitarios justifica la adquisición de unidades adicionales.
3. Reforzar los protocolos de seguridad operacional. Esto incluye la revisión periódica de los procedimientos de emergencia y la formación continua del personal de mantenimiento de los instructores.
4. Realizar un análisis detallado del costo total de propiedad y operación considerando su alta utilización. Esto incluirá costo de combustible, mantenimiento, seguros, depreciación y capacitación.
5. Comparar este análisis con modelos alternativos que podría cumplir funciones similares evaluando la eficiencia a largo plazo y la posibilidad de diversificar la flota si se identifican opciones más ventajosas.
6. Establecer programa de implementación de regímenes de mantenimiento predictivo como puede ser análisis de sensores que permitan predecir fallas antes de que ocurran.
7. Asegurar un stock estratégico de repuestos críticos específicos para el EMB-ERJ190100IGW. Dada su alta utilización, la falta de una pieza podría generar demoras o cancelaciones costosa. Establecer acuerdos con proveedores para garantizar la entrega rápida.
8. Ajustar la programación de vuelos para maximizar la rentabilidad, considerando la demanda estacional, los horarios preferidos por los pasajeros y la optimización de los tiempos de escala para reducir costos de personal y combustible.
9. Asegurar los estándares de confort, limpieza y servicio se mantengan altos para fomentar la lealtad de los clientes.

13. Arquitectura alternativa



Ejercicio 2:

Alquiler de automóviles

Una de las empresas líderes en alquileres de automóviles solicita una serie de dashboards y reportes para poder basar sus decisiones en datos. Entre los indicadores mencionados se encuentran total de alquileres, segmentación por tipo de combustible, lugar, marca y modelo de automóvil, valoración de cada alquiler, etc. Como Data Engineer debe crear y automatizar el pipeline para tener como resultado los datos listos para ser visualizados y responder las preguntas de negocio.

1. Crear en hive una database car_rental_db y dentro una tabla llamada car_rental_analytics.

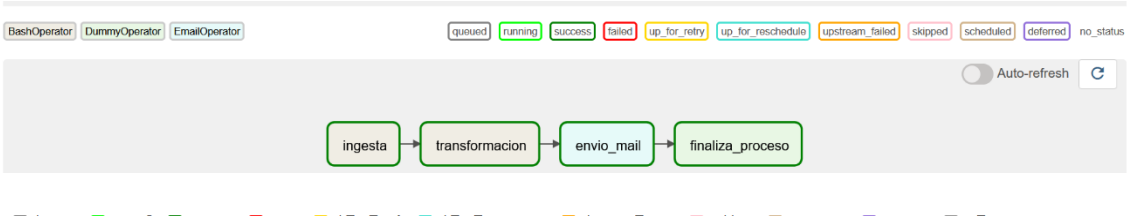
```
CREATE DATABASE car_rental

create table car_rental.car_rental_analytics
(
    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
)
```

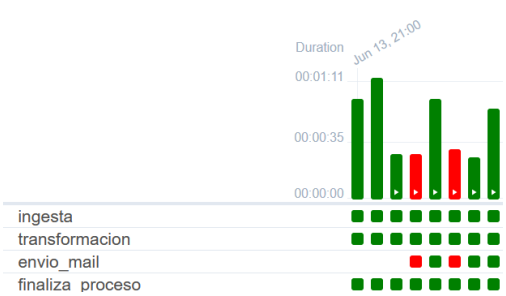
2. Crear script para el ingest de estos dos files
   Aquí va la url de github

3. Crear un script para tomar el archivo desde HDFS y hacer las siguientes transformaciones:
   a. En donde se necesario, modificar los nombres de las columnas. Evitar espacios y puntos (reemplazar por guion _). Evitar nombres de columnas largos.
   b. Redondear los float de rating y castear a int.
   c. Joinear ambos files.
   d. Eliminar los registros con rating nulo.
   e. Cambiar mayúscula por minúscula en “fuelType”.
   f. Excluir el estado de Texas.
   g. Finalmente insertar en Hive el resultado.

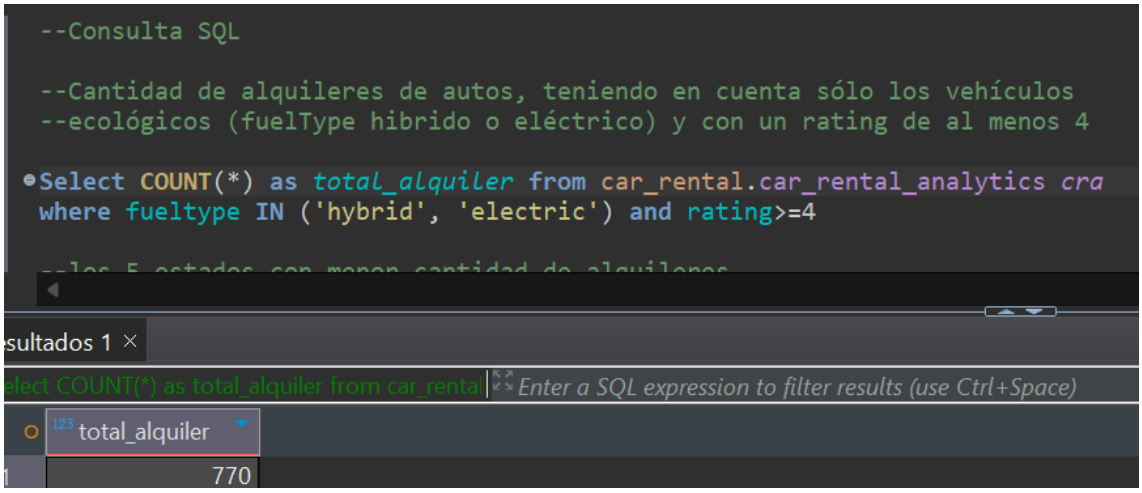
   Aquí va el github de la transformación.

4. Realizar un proceso automático en airflow que orqueste los pipelines creados en los puntos anteriores. Crear dos tareas:
    a. Un DAG padre que ingeste los archivos y luego llame al DAG hijo.
    b. Un DAG hijo que procese la información la información y lo cargue en Hive.





5. Por medio de consulta SQL al data warehouse, mostrar:
    a. Cantidad de alquileres de autos, teniendo en cuenta solo los vehículos ecológicos (fuelType hibrido o electrico) y con un rating de al menos 4.



    b. Los 5 estados con menor cantidad de alquileres.

```sql
--los 5 estados con menor cantidad de alquileres

Select
t.ranking,
t.state_name,
t.total_alquiler
FROM (
Select
ROW_NUMBER() OVER(order by COUNT(*) asc) as ranking,
state_name as state_name,
Count(*) as total_alquiler
from car_rental.car_rental_analytics cra
group by state_name) as t
where t.ranking <=5
```

esultados 1 ×

elect t.ranking, t.state_name, t.total_alquiler FR(Enter a SQL expression to filter results (use Ctrl+Space

	ranking	state_name	total_alquiler
1	1	MT	1
2	2	NH	3
3	3	WV	3
4	4	MS	4
5	5	AR	4

c. Los 10 modelos junto con su marca de autos más rentados.

```sql
--los 10 modelos (junto con su marca) de autos más rentados
Select
t.ranking,
t.marca,
t.modelo,
t.precio
FROM
(
Select
ROW_NUMBER() OVER(order by SUM(rate_daily) DESC ) as ranking,
make as marca,
model as modelo,
SUM(rate_daily) as precio
from car_rental.car_rental_analytics cra
group by make, model ) as t
where t.ranking <=10
```

Resultados 1

Select t.ranking, t.marca, t.modelo, t.precio FRO | Enter a SQL expression to filter results (use Ctrl+Spac

	ranking	marca	modelo	precio
1	1	Tesla	Model 3	36.867
2	2	Tesla	Model X	19.848
3	3	Tesla	Model S	16.521
4	4	Chevrolet	Corvette	11.982
5	5	Ford	Mustang	10.182
6	6	Jeep	Wrangler	8.451
7	7	Lamborghini	Huracan	8.348
8	8	Porsche	911	7.895
9	9	BMW	i8	7.683
10	10	Maserati	Ghibli	6.423

d. Mostrar por año, cuantos alquileres se hicieron, teniendo en cuenta automóviles fabricados desde 2010 a 2015.

```sql
--Mostrar por año, cuántos alquileres se hicieron, teniendo en cuenta automóviles
--fabricados desde 2010 a 2015

Select year, Count(*) as total_alquiler FROM car_rental.car_rental_analytics cra
where year between 2010 and 2015
group by year
```

ultados 1

ect year, Count(*) as total_alquiler FROM car_r | Enter a SQL expression to filter results (use Ctrl+Space)

	year	total_alquiler
	2.010	144
	2.011	200
	2.012	225
	2.013	305
	2.014	382
	2.015	532

e. Las 5 ciudades con más alquileres de vehículos ecológicos (fuelType hibrido o electrico).

```
--las 5 ciudades con más alquileres de vehículos ecológicos (fuelType hibrido o electrico)
Select
t.ranking ,
t.city,
t.total_alquiler
FROM (
Select
row_number() OVER(order by COUNT(*) desc) as ranking,
city,
Count(*) as total_alquiler
from car_rental.car_rental_analytics cra
where fuelType in ('hybrid', 'electric')
group by city ) as t
where t.ranking <=5
```

sultados 1 ×

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

	ranking	city	total_alquiler
1	1	San Diego	44
2	2	Las Vegas	34
3	3	Portland	20
4	4	Phoenix	17
5	5	San Jose	15

f. El promedio de review, segmentando por tipo de combustible.

```
--el promedio de reviews, segmentando por tipo de combustible

Select fuelType as TipoCombustible, AVG(reviewcount ) as revisiones from car_rental.car_rental_analytics cra
group by fueltype
```

esultados 1 ×

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

	tipocombustible	revisiones
1	[NULL]	21,0491803279
2	diesel	17,5
3	electric	28,3394833948
4	gasoline	31,9270236613
5	hybrid	34,8733624454