

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:

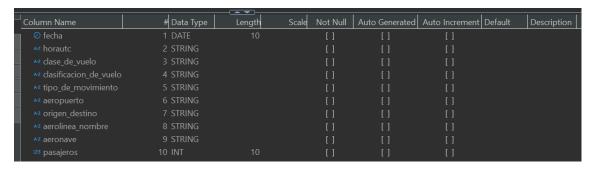


Figura 1. Tabla vuelos

b. Tabla detalle de vuelos

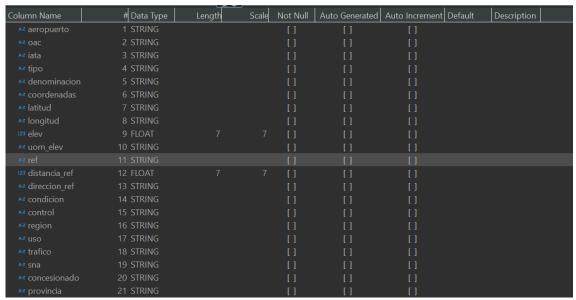


Figura 2. Tabla detalle de vuelos

3. Diagrama de orquestación airflow

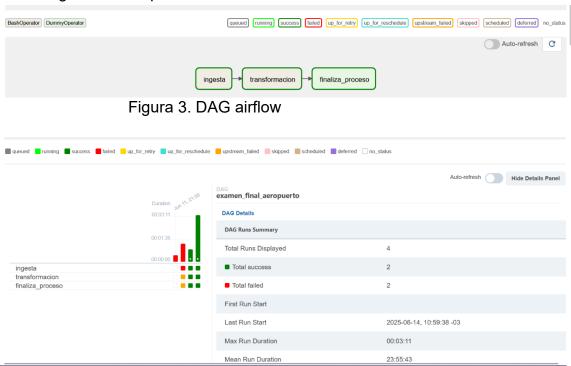


Figura 4. Vista de Grilla

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
```

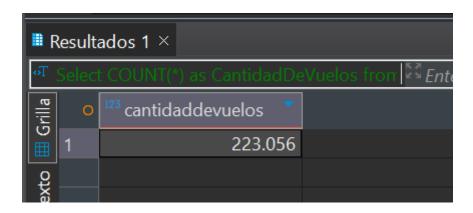
Figura 6. Creación de Tabla detalle_vuelo

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

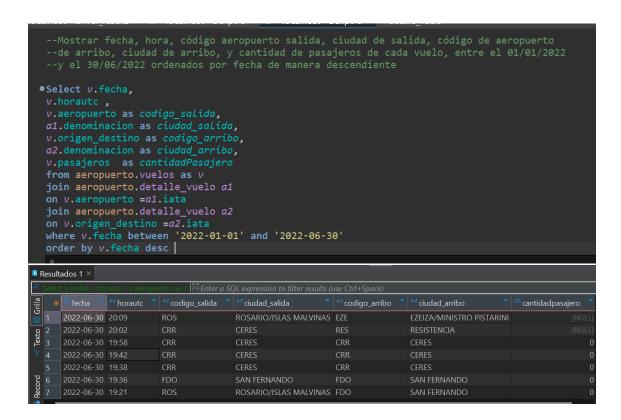
where fecha between '2021-12-01' and '2022-12-31'
```



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.



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.



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.

```
--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 ×											
оΤ	$^{\bullet T}$ Select Laerolinea, LcantidadPasajero, Lranking F $^{\frac{\kappa}{2}}$ Enter a SQL expression to filter res											
Grilla	0	^{A-Z} aerolinea	¹²³ cantidad pasajero	123 ranking								
	1	AEROLINEAS ARGENTINAS SA	7.484.860	1								
2	2	JETSMART AIRLINES S.A.	1.511.650	2								
Texto	3	FB LÍNEAS AÉREAS - FLYBONDI	1.482.473	3								
Ê	4	AMERICAN JET S.A.	25.789	4								
	5	L.A.D.E.	15.074	5								
н	6	BAIRES FLY SA	4.960	6								
н	7	LADE	3.895	7								
н	8	FUERZA AEREA ARGENTINA	3.855	8								
$\lceil \rfloor$	9	FUERZA AEREA ARGENTINA (FA	3.138	9								
<u>l</u>	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.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
```

■ Resultados 1 ×										
«Т	^{◦T} Select t.aeronave , t.Utilizacion, t.ranking FROM Enter a SQL ex									
Grilla	0	^{A-Z} aeronave	123 utilizacion	123 ranking						
	1	PA-PA-28-181	4.865	1						
얼	2	EMB-ERJ190100IGW	4.005	2						
Texto	3	LJ-60	2.455	3						
Ê	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						
ord										

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 (checkin, 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:

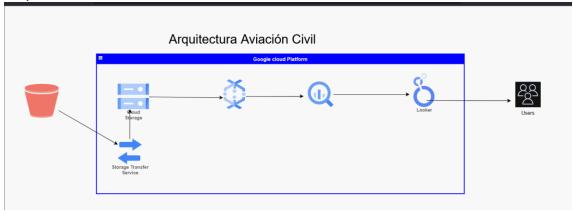
- 1. Revisar y ajustar la estructura de tarifas y cargos asociados a los vuelos privados (ejemplo: tasa de aterrizajes, estacionamiento, servicios específicos para aviación ejecutiva). Asegurar que reflejen el valor del servicio prestado y el uso de la infraestructura.
- Evaluar la demanda y ofertas 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.
- 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.
- 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.
- 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.
- 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.

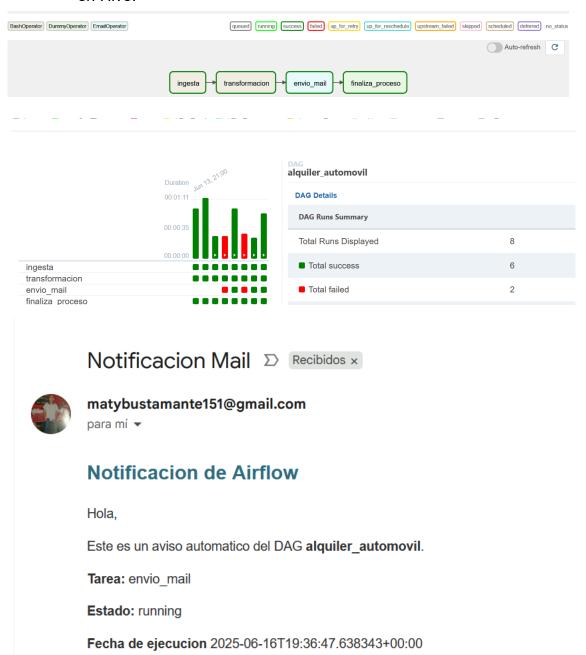
```
CREATE DATABASE car_rental

•create table car_rental.car_rental_analytics

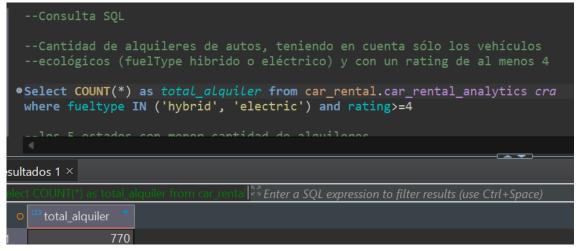
(
    fuelType string,
    rating int,
    renterTripsTaken int,
    reviewCount int,
    city string,
    state_name string,
    owner_id int,
    rate_daily int,
    make string,
    model string,
    year int
)
```

- Crear script para el ingest de estos dos files https://github.com/Matias-Bustamante/TrabajoFinalDE/tree/main/Alquileres/ingest
- 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.
 - h. Github:https://github.com/Matias Bustamante/TrabajoFinalDE/tree/main/Alquileres/transform

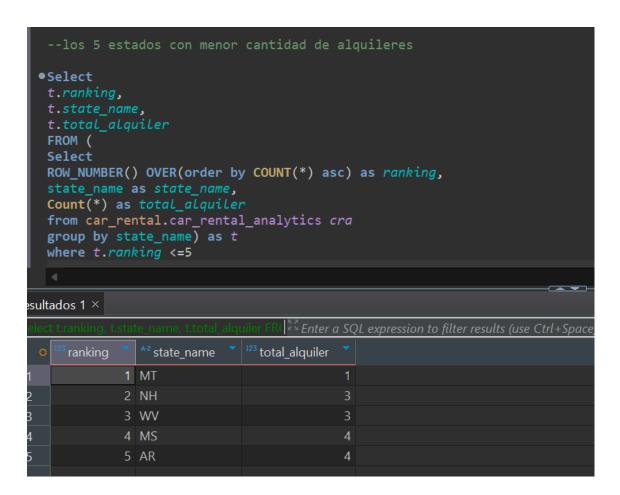
- 4. Realizar un proceso automático en airflow que orqueste los pipelines creados en los puntos anteriores. Crear dos tareas:
 - a. Un DAG padre que ingeste los archivos y luego llame al DAG hijo.
 - b. Un DAG hijo que procese la información la información y lo cargue en Hive.



- 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.



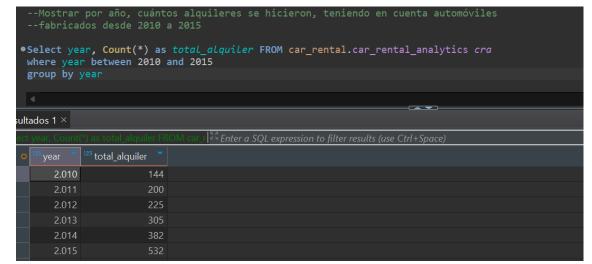
b. Los 5 estados con menor cantidad de alquileres.



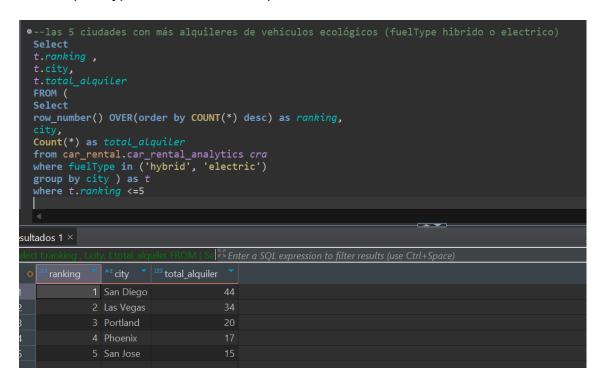
c. Los 10 modelos junto con su marca de autos más rentados.

```
●--los 10 modelos (junto con su marca) de autos más rentados
    Select
    t.ranking,
    t.modelo,
    FROM
    Select
    ROW_NUMBER() OVER(order by SUM(rate_daily) DESC ) as ranking,
    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 ×
                     <sup>A-Z</sup> marca
                                  <sup>A-Z</sup> modelo
                                                  <sup>123</sup> precio
        ranking
                                  Model 3
                  1 Tesla
                                                        36.867
                  2 Tesla
                                  Model X
                                                        19.848
                  3 Tesla
                                  Model S
                                                        16.521
                  4 Chevrolet
                                  Corvette
                                                        11.982
                  5 Ford
                                  Mustang
                                                        10.182
                                  Wrangler
                                                         8.451
                  6 Jeep
                  7 Lamborghini Huracan
                                                         8.348
                  8 Porsche
                                  911
                  9 BMW
                                                         7.683
                 10 Maserati
                                  Ghibli
                                                         6.423
```

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



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



El promedio de review, segmentando por tipo de combustible.



Conclusiones y Recomendaciones

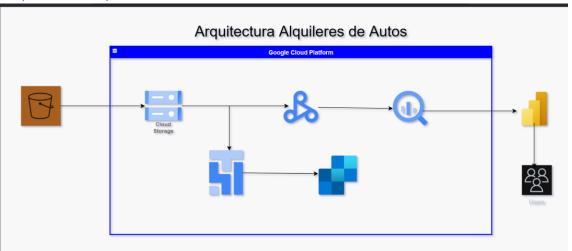
El alquiler de autos por tipos de combustible ha ido incrementando año tras años sobre todo en los autos a combustión por gasolina, exceptuando en el año 2020 producto de la pandemia del covid-19 hubo una disminución de alquileres de autos en general. Por otra parte, los autos híbridos son aquellos que tienen una mayor reseña por parte de los clientes, pero en términos de volumen de autos alquilados no son los que más son consumidos por los distintos clientes, unos de los factores influyentes en los alquileres es el precio y por este motivo los autos a gasolina son más económico en comparación con autos híbridos. También se puede observar que el perfil de consumidores esta direccionado con el cuidado del medio ambiente y disminución de CO2.

Los autos de tipos combustibles representan el 82% del volumen total de alquileres, mientras que los autos ecológicos representan el 15,7% esto indica que existe una alta preferencia de los clientes por los autos a gasolina o bien que no existe suficiente demanda de los autos ecológicos. Por otra existe una correlación entre las ciudades con más alquileres y las ciudades con más puntuaciones.

Recomendaciones

- Identificar barreras: Identificar la causa de porcentaje de alquileres ecológico bajo. ¿Falta de disponibilidad en la flota, falta de demanda de los clientes o precios altos?
- Promoción y educación: Si la demanda es un problema, educar a los clientes sobre los beneficios de los vehículos ecológicos (ahorro de combustible, experiencia de conducción, impacto ambiental) y ofrecer incentivos especiales para su alquiler.
- Expansión de flota: Si la baja es por disponibilidad, considerar expandir la flota de vehículos ecológicos, especialmente en ciudades donde la demanda podría ser mayor (por ejemplo, ciudades con conciencia ambiental).
- Estudiar las mejores prácticas: Analizar qué factores contribuyen a la alta satisfacción del cliente en Las Vegas, Portland y San Diego. ¿Es el servicio al cliente, la calidad de los vehículos, la facilidad del proceso de alquiler? Implementar estas mejores prácticas en otras ubicaciones.
- Marketing dirigido: Utilizar los altos ratings como un punto de venta clave en la publicidad para estas ciudades, destacando la satisfacción del cliente.

7. Propuesta de Arquitectura



Link de Dashboard Alguiler de autos: https://acortar.link/dd6W95