

EXAMEN FINAL

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 y realizar los análisis de datos solicitados que permita responder las preguntas de negocio, y hacer sus recomendaciones con respecto al estado actual.

Listado de vuelos realizados:

<https://datos.gob.ar/lv/dataset/transporte-aterrizajes-despegues-procesados-por-administracion-nacional-aviacion-civil-anac>

Listado de detalles de aeropuertos de Argentina:

<https://datos.transporte.gob.ar/dataset/lista-aeropuertos>

TAREAS

1. Hacer ingest de los siguientes files relacionados con transporte aéreo de Argentina :

2021: <https://data-engineer-edvai-public.s3.amazonaws.com/2021-informe-ministerio.csv>

2022: <https://data-engineer-edvai-public.s3.amazonaws.com/202206-informe-ministerio.csv>

Aeropuertos_detalles:

https://data-engineer-edvai-public.s3.amazonaws.com/aeropuertos_detalle.csv

2. Crear 2 tablas en el datawarehouse, una para los vuelos realizados en 2021 y 2022 (2021-informe-ministerio.csv y 202206-informe-ministerio) y otra tabla para el detalle de los aeropuertos (aeropuertos_detalle.csv)

Schema Tabla 1:

campos	tipo
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	integer

Schema Tabla 2:

Campo	Tipo
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

3. Realizar un proceso automático orquestado por airflow que ingeste los archivos previamente mencionados entre las fechas 01/01/2021 y 30/06/2022 en las dos columnas creadas.

Los archivos 202206-informe-ministerio.csv y 202206-informe-ministerio.csv → en la tabla aeropuerto_tabla

El archivo aeropuertos_detalle.csv → en la tabla aeropuerto_detalles_tabla

4. Realizar las siguientes transformaciones en los pipelines de datos:
- Eliminar la columna inhab ya que no se utilizará para el análisis
 - Eliminar la columna fir ya que no se utilizará para el análisis
 - Eliminar la columna “calidad del dato” ya que no se utilizará para el análisis
 - Filtrar los vuelos internacionales ya que solamente se analizarán los vuelos domésticos
 - En el campo pasajeros si se encuentran campos en Null convertirlos en 0 (cero)
 - En el campo distancia_ref si se encuentran campos en Null convertirlos en 0 (cero)
5. Mostrar mediante una impresión de pantalla, que los tipos de campos de las tablas sean los solicitados en el datawarehouse (ej: fecha date, aeronave string, pasajeros integer, etc.)
6. Determinar la cantidad de vuelos entre las fechas 01/12/2021 y 31/01/2022. Mostrar consulta y Resultado de la query
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 01/01/2022

y 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
10. 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. Mostrar consulta y Visualización
11. Qué datos externos agregaría en este dataset que mejoraría el análisis de los datos
12. Elabore sus conclusiones y recomendaciones sobre este proyecto.
13. Proponer una arquitectura alternativa para este proceso ya sea con herramientas on premise o cloud (Sí aplica)

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, con estos campos:

campos	tipo
fuelType	string

rating	integer
renterTripsTaken	integer
reviewCount	integer
city	string
state_name	string
owner_id	integer
rate_daily	integer
make	string
model	string
year	integer

2. Crear script para el ingest de estos dos files

<https://data-engineer-edvai-public.s3.amazonaws.com/CarRentalData.csv>

<https://data-engineer-edvai-public.s3.amazonaws.com/georef-united-states-of-america-state.csv>

Sugerencia: descargar el segundo archivo con un comando similar al abajo mencionado, ya que al tener caracteres como ‘&’ falla si no se le asignan comillas. Adicionalmente, el parámetro -O permite asignarle un nombre más legible al archivo descargado

```
wget -P ruta_destino -O ruta_destino/nombre_archivo.csv ruta_al_archivo
```

Info del dataset: <https://www.kaggle.com/datasets/kushleshkumar/cornell-car-rental-dataset>

3. Crear un script para tomar el archivo desde HDFS y hacer las siguientes transformaciones:

- En donde sea necesario, modificar los nombres de las columnas. Evitar espacios y puntos (reemplazar por _). Evitar nombres de columna largos
- Redondear los float de 'rating' y castear a int.
- Joinear ambos files
- Eliminar los registros con rating nulo
- Cambiar mayúsculas por minúsculas en 'fuelType'
- Excluir el estado Texas

Finalmente insertar en Hive el resultado

4. Realizar un proceso automático en Airflow que orqueste los pipelines creados en los puntos anteriores. Crear dos tareas:

- a. Un DAG padre que ingente los archivos y luego llame al DAG hijo
- b. Un DAG hijo que procese la información y la cargue en Hive

5. Por medio de consultas SQL al data-warehouse, mostrar:

- a. Cantidad de alquileres de autos, teniendo en cuenta sólo los vehículos ecológicos (fuelType híbrido o eléctrico) y con un rating de al menos 4.
- b. los 5 estados con menor cantidad de alquileres (mostrar query y visualización)
- c. los 10 modelos (junto con su marca) de autos más rentados (mostrar query y visualización)
- d. Mostrar por año, cuántos alquileres se hicieron, teniendo en cuenta automóviles fabricados desde 2010 a 2015
- e. las 5 ciudades con más alquileres de vehículos ecológicos (fuelType híbrido o eléctrico)
- f. el promedio de reviews, segmentando por tipo de combustible

6. Elabore sus conclusiones y recomendaciones sobre este proyecto.

7. Proponer una arquitectura alternativa para este proceso ya sea con herramientas on premise o cloud (Si aplica)