Arquitecturas Analíticas
UAO – 2025

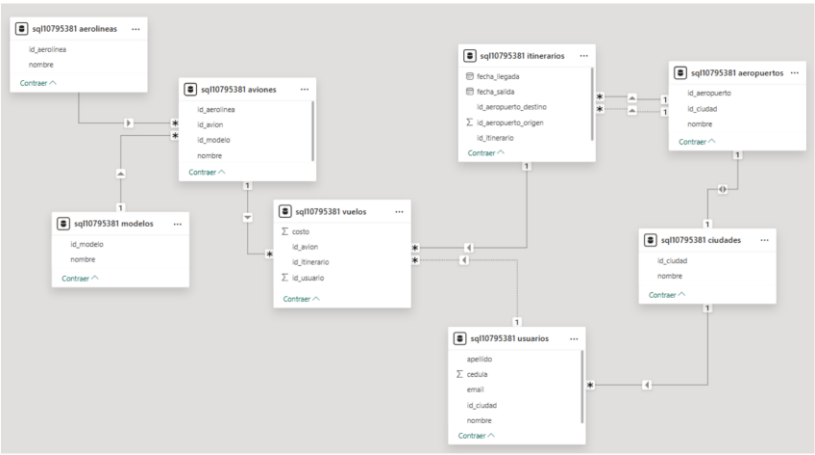
Taller 1 - Modelado multidimensional en Arquitecturas BI
Jorge Andres Jaramillo Neme

Solución:

Se parte de la información del negocio y de las preguntas asociadas a las necesidades que tienen, para poder "traducir" dicha información en el modelo estrella propuesto, que facilite la consulta rápida y filtrado de la información requerida, como también facilitar la construcción de gráficas que faciliten su interpretación:

1. ¿Cuál aerolínea realizó el mayor número de vuelos a la ciudad de Roma en el año 2019 y cuál en el año 2020?
2. Total de dinero recaudado por vuelos de cada aerolínea en el primer semestre del año 2019 y en el primer semestre del año 2020.
3. ¿Cuál modelo de avión realizó el mayor número de vuelos en el año 2019 y cuál en el año 2020?
4. ¿Cuál fue la ciudad cuyos habitantes viajaron más en el año 2019 y cuál en el año 2020?

Adicionalmente, cito la información dada sobre las tablas actuales del modelo OLTP:



1. Los usuarios y su respectiva ciudad de origen.
2. Los aeropuertos y la ciudad donde se encuentran ubicados.
3. Los aviones, sus modelos y la aerolínea a la que pertenece cada uno.

4. Los itinerarios, incluyendo fecha de salida, fecha de llegada, ciudad de origen del vuelo y ciudad de destino del vuelo.
5. Los vuelos, incluyendo itinerarios de cada vuelo, aviones implicados en cada vuelo, usuarios que tomaron determinado vuelo y costo de los diferentes vuelos registrados.

Con esto podemos notar lo siguiente inmerso en las preguntas:

- Mencionen fechas que pueden ser años, por lo tanto, debemos de tener una dimensión relacionada a esto.
- Podemos notar que toda gira entorno a los vuelos que representa la información transaccional clave.
- Podemos notar que, en el modelo relacional, hay bastantes dependencias entre por ejemplo vuelo -> aviones -> aerolíneas, lo que puede llevar a una reorganización alrededor del vuelo en el modelo estrella.

En síntesis, IATA desea cuantificar el impacto de COVID-19 en 2020 frente a 2019. Las preguntas requieren conteos de vuelos y sumas de dinero organizados por aerolínea, modelo de avión, ciudades y periodos (año/semestre):

1. Aerolínea con más vuelos hacia Roma en 2019 y en 2020.
2. Recaudo (suma de costo) por aerolínea en 1er semestre de 2019 y de 2020.
3. Modelo de avión con más vuelos en 2019 y en 2020.
4. Ciudad de residencia cuyos habitantes viajaron más en 2019 y en 2020.

Estas preguntas implican un hecho transaccional (vuelo tomado por un usuario) y dimensiones de tiempo, aerolínea/modelo (vía avión), ciudad de destino/origen y usuario.

Tablas y relaciones relevantes del esquema relacional:

- vuelos (costo, id_itinerario, id_avion, id_usuario): hecho transaccional base.
- itinerarios (fecha_salida, fecha_llegada, id_aeropuerto_origen, id_aeropuerto_destino): aporta fecha y rutas.
- aviones (id_avion, id_aerolinea, id_modelo): enlaza aerolínea y modelo.
- aerolineas (id_aerolinea, nombre).
- modelos (id_modelo, nombre).

- aeropuertos (id_aeropuerto, id_ciudad) → ciudades (id_ciudad, nombre): para origen/destino.
- usuarios (cedula, nombre, apellido, email, id_ciudad): ciudad de residencia del pasajero.

La clave primaria de vuelos es (id_itinerario, id_avion, id_usuario), lo que semánticamente equivale a "un pasajero X viajó en el itinerario Y en el avión Z con un costo dado".

Con esta información base vamos a proponer el diseño del modelo estrella:

Definición de la tabla de hechos y sus dimensiones

A partir del análisis de las tablas disponibles en la base de datos transaccional, se identificó que la entidad vuelos representa el evento central que almacena las transacciones de cada viaje realizado por un usuario, con su respectivo costo, avión e itinerario.
Esta tabla será la base para definir la tabla de hechos, complementada con dimensiones que describen los distintos contextos del vuelo: tiempo, avión, usuario y ciudades involucradas.

Tabla de hechos: fact_vuelos
Esta tabla almacena los valores numéricos que serán analizados (medidas), junto con las claves que permiten relacionarla con las dimensiones.
Campos propuestos:

Campo	Descripción
id_tiempo	Identificador de la fecha del vuelo (derivado de itinerarios.fecha_salida).
id_avion	Identificador del avión utilizado.
id_usuario	Identificador del pasajero.
id_ciudad_origen	Ciudad de origen del vuelo.
id_ciudad_destino	Ciudad de destino del vuelo.
costo	Valor pagado por el pasajero.
vuelo_cnt	Valor constante de 1, usado para conteos de vuelos.

Esta estructura permite calcular tanto totales de dinero como cantidades de vuelos, agrupando por cualquiera de las dimensiones.

Dimensión de tiempo: dim_tiempo
Permite analizar los vuelos a lo largo del tiempo y realizar comparaciones entre periodos (años, meses, semestres).

Campos:

Campo	Descripción
id_tiempo	Identificador único de la fecha.
fecha	Fecha completa del vuelo.
anio	Año (por ejemplo, 2019 o 2020).
mes	Mes numérico (1–12).
semestre	Semestre (1 o 2).

La información se deriva directamente del campo fecha_salida en la tabla itinerarios.

Dimensión de avión: dim_avion
Integra en una sola tabla los datos de los aviones, junto con los atributos de aerolínea y modelo.

Para simplificar las consultas, se copian los nombres de aerolínea y modelo directamente dentro de esta dimensión.

Campo	Descripción
id_avion	Identificador único del avión.
nombre_avion	Nombre o código del avión.
nombre_aerolinea	Nombre de la aerolínea propietaria del avión.
nombre_modelo	Modelo del avión (por ejemplo, Airbus 320, Boeing 747).

Esta dimensión permite responder preguntas relacionadas con la aerolínea o el modelo del avión sin depender de otras tablas adicionales.

Arquitecturas Analíticas
UAO – 2025

Dimensión de ciudad: dim_ciudad

Contiene el catálogo de ciudades que se utilizan tanto como origen como destino en los vuelos.
Durante la carga de datos, esta tabla se usará dos veces en la tabla de hechos: una para el campo id_ciudad_origen y otra para id_ciudad_destino.

Campos:

Campo	Descripción
id_ciudad	Identificador de la ciudad.
nombre_ciudad	Nombre de la ciudad.

Esta dimensión permite identificar destinos específicos como Roma o Bogotá y realizar comparaciones por ciudad.

Dimensión de usuario: dim_usuario

Contiene la información básica de cada pasajero.
Se agrega directamente el nombre de la ciudad de residencia para no depender de otras dimensiones.

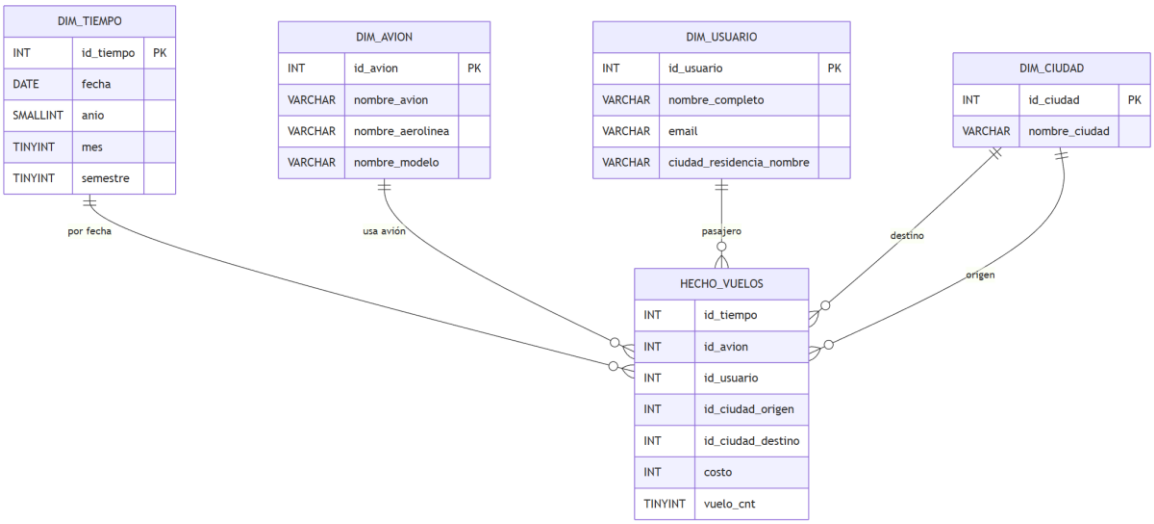
Campos:

Campo	Descripción
id_usuario	Identificador único (cédula).
nombre_completo	Concatenación del nombre y apellido.
email	Correo electrónico del usuario.
ciudad_residencia_nombre	Nombre de la ciudad donde reside el pasajero.

Esta dimensión permite analizar la cantidad de vuelos realizados por habitantes de distintas ciudades.

Arquitecturas Analíticas
UAO – 2025

Con la información dada se muestra como quedaría el diagrama del modelo estrella:



Modelo OLAP propuesto

Lógica para crear las tablas del modelo estrella en el ETL

Para hacer el ETL vamos a tener en cuenta lo siguiente:

Para construir este modelo, los datos se extraen de las tablas originales y se transforman de la siguiente manera:
Tabla fact_vuelos:

Se alimenta desde vuelos, relacionando cada registro con su itinerario, avión, usuario y ciudades de origen y destino.
El campo costo se copia directamente, y se agrega el campo vuelo_cnt con valor 1 para facilitar los conteos.

Dimensión de tiempo (dim_tiempo):
Se genera a partir de las fechas de salida (fecha_salida) en itinerarios. De cada fecha se derivan el año, mes y semestre.

Arquitecturas Analíticas
UAO – 2025

Dimensión de avión (dim_avion):
Se construye uniendo las tablas aviones, aerolineas y modelos.
Se copian los nombres correspondientes en una sola tabla.

Dimensión de ciudad (dim_ciudad):
Se carga directamente desde la tabla ciudades, que ya contiene los nombres únicos.

Dimensión de usuario (dim_usuario):
Se toma de la tabla usuarios y se complementa con el nombre de la ciudad, consultado desde ciudades.

Tecnologías para el desarrollo

Por temas del ejercicio, se realiza el proceso de ETL usando Python, hacia el modelo estrella implementado en DuckDB, la cual es una DB ligera y muy potente para temas analíticos. Puede trabajar directamente con archivos parquet y para consulta es buena.

Finalmente, los gráficos se realizan con streamlit y plotly para facilitar. En el readme del repositorio se detalle como correrse.

Implementación del proceso ETL

Para la construcción del Data Mart se desarrolló un proceso ETL (Extract, Transform, Load) que permitió extraer la información desde la base de datos relacional de la IATA, transformarla según la estructura dimensional definida, y cargarla finalmente en el modelo estrella alojado en DuckDB.

La extracción se realizó conectando a la base de datos MySQL proporcionada por el docente, utilizando las credenciales y el esquema relacional original. Desde Python se estableció la conexión con la librería mysql.connector o mediante SQLAlchemy, según disponibilidad, garantizando la lectura directa de las tablas fuente (aerolineas, ciudades, aeropuertos, aviones, modelos, usuarios, itinerarios y vuelos).

Durante la fase de transformación se aplicaron los pasos necesarios para alinear los campos del modelo relacional al modelo dimensional. Esto implicó:

- Normalizar y renombrar campos para coincidir con las claves y nombres de las dimensiones del modelo estrella (por ejemplo, id_avion, id_ciudad_destino, id_ciudad_origen, id_usuario, id_tiempo).
- Derivar atributos temporales como año, mes y semestre a partir de las fechas de salida en la tabla de itinerarios.
- Combinar las tablas relacionales mediante joins controlados para conformar las dimensiones finales (dim_avion, dim_ciudad, dim_usuario, dim_tiempo) y la tabla de hechos (hecho_vuelos), la cual consolida las métricas de análisis (principalmente el costo del vuelo y el conteo de vuelos).

Finalmente, los datos transformados se cargaron al modelo estrella dentro de una base DuckDB, generando archivos .parquet y una base persistente (iata_star.duckdb). DuckDB fue elegido por su facilidad de integración con Python, su desempeño en consultas analíticas locales y su soporte nativo para SQL de tipo OLAP, sin requerir un servidor adicional.

Visualización y análisis de resultados

Para responder los requerimientos de análisis de la IATA se desarrolló una interfaz interactiva con Streamlit, conectada directamente al Data Mart en DuckDB. Esta aplicación permite ejecutar las consultas analíticas sobre las dimensiones del modelo, generando visualizaciones dinámicas que facilitan la interpretación de resultados.
Las consultas implementadas fueron:

1. Aerolínea con mayor número de vuelos a Roma en 2019 y 2020.
   Se construyó una consulta filtrando la dimensión destino (dim_ciudad) por la ciudad de Roma y agrupando por aerolínea y año.
   El resultado se presenta en un gráfico de barras comparativo por año.
2. Recaudo total por aerolínea en el primer semestre de 2019 y 2020.
   El cálculo se realizó sumando el campo costo en la tabla de hechos, filtrando por semestre 1.
   Se utiliza un gráfico tipo donut (torta con centro vacío) que muestra la proporción del recaudo por aerolínea y el valor total formateado como dinero.
3. Modelo de avión con mayor número de vuelos por año.
   Se agruparon los registros de la tabla de hechos por id_avion y anio relacionando el modelo desde la dimensión correspondiente.

   Los resultados se presentan mediante barras verticales que comparan el volumen de vuelos por tipo de avión y año.
4. Ciudad cuyos habitantes viajaron más en 2019 y 2020.
   Se cruzó la tabla de hechos con la dimensión de usuario para identificar la ciudad de residencia y se contabilizó el número total de vuelos por año. El resultado se visualiza con un gráfico de torta, mostrando la participación proporcional de las ciudades más viajeras.

Cada gráfico incluye además su respectiva tabla de datos en formato st.dataframe, lo que permite observar tanto los valores absolutos como las proporciones.

Con esta interfaz, el Data Mart se convierte en una herramienta exploratoria que permite realizar comparaciones por año y por dimensión de análisis, evidenciando de forma clara las variaciones de actividad aérea antes y durante el año 2020, afectado por la pandemia del COVID-19.

Resultados del análisis

A partir de la implementación del modelo estrella, el proceso ETL y la consulta sobre el Data Mart en DuckDB, se obtuvieron los siguientes resultados para los requerimientos analíticos definidos por la IATA:

Pregunta 1: Aerolínea con mayor número de vuelos a Roma (2019–2020)

Del análisis de los registros de vuelos hacia la ciudad de **Roma**, se identificó que:

- En el **año 2019**, la aerolínea **Avianca** realizó el **mayor número de vuelos** con un total de **11 vuelos** registrados con destino a Roma vs Wingo con 4.
- En el **año 2020**, **no se registraron vuelos** con destino a esta ciudad dentro del conjunto de datos disponible, lo cual refleja la **disminución drástica del tráfico aéreo internacional** durante el periodo inicial de la pandemia de COVID-19, o falta de data disponible para una interpretación completa.

Arquitecturas Analíticas
UAO – 2025

Pregunta 2: Total de dinero recaudado por vuelos de cada aerolínea en el primer semestre de 2019 y 2020

El cálculo del recaudo total se basó en la suma de los costos asociados a cada vuelo registrado en el primer semestre de cada año. Los resultados obtenidos fueron los siguientes:

Año	**Aerolínea**	**Recaudo total**
2019	Avianca	**$90,089,500**
2019	Latam	**$81,830,300**
2019	Wingo	**$46,104,600**
2020	Latam	**$14,530,000**
2020	Avianca	**$9,890,000**
2020	Wingo	**$6,000,000**

Durante el primer semestre de 2019, **Avianca** lideró en ingresos, seguida por **Latam** y **Wingo**.
En el primer semestre de 2020 se observa una caída significativa del recaudo, con reducciones superiores al 80 %, destacándose Latam como la que mantuvo la mayor operación residual durante ese periodo.

Pregunta 3: Modelo de avión con mayor número de vuelos por año
El análisis de los vuelos agrupados por modelo de avión permitió identificar que:

Año	**Modelo de avión**	**Total de vuelos**
2019	**Airbus 320**	**70 vuelos**
2020	**Airbus 320**	**24 vuelos**

El **Airbus A320** se consolidó como el modelo más utilizado en ambos años, lo que evidencia su papel como aeronave predominante en rutas de corto y mediano alcance dentro de la flota global de las aerolíneas analizadas.

A pesar de la reducción general de vuelos en 2020, este modelo mantuvo su posición como el principal en operación.

Pregunta 4: Ciudad de residencia con más viajeros (2019–2020)

Al analizar la ciudad de residencia de los usuarios registrados en los vuelos:

Año	**Ciudad**	**Total de viajes**
2019	**Medellín**	**36 viajes**
2020	**Medellín**	**12 viajes**

Los pasajeros residentes en **Medellín** fueron los que más volaron en ambos años, aunque la cantidad total de viajes **disminuyó a un tercio** durante 2020, lo que nuevamente confirma el impacto global del confinamiento sobre la movilidad aérea.