



Maestría en Inteligencia artificial y ciencia de datos

Arquitectura Analítica.

Implementación Data Mart

Torres Polanco, Cristhian; Londoño, Miguel; Mendoza, Sergio

Informe – Construcción del Data Mart de Vuelos (IATA)

Análisis del caso

La base de datos **IATA.db** es un sistema transaccional que registra la operación de vuelos comerciales, integrando información sobre aerolíneas, aviones, modelos, aeropuertos, ciudades, usuarios y los itinerarios asociados a cada vuelo. Su estructura relacional permite almacenar cada evento del proceso, desde la programación del itinerario hasta la compra del pasaje, de forma detallada, asegurando integridad referencial mediante claves foráneas entre las tablas. En esencia, sirve como la fuente operativa desde la cual se nutre el DataMart analítico.

La idea del proyecto es base a un entorno analítico, creando un DataMart que permitiera comparar años, analizar destinos más frecuentes o ver qué aerolíneas generan más ingresos.

1. El modelo adopta una arquitectura en estrella, con una única tabla de hechos (fact_vuelos) que centraliza los indicadores de negocio, y varias tablas dimensión (dim) que describen los distintos ejes de análisis. Tabla de hechos:
 - a. fact_vuelos concentra la información de cada vuelo realizado, incluyendo los identificadores de avión, usuario, itinerario, aeropuertos de origen y destino, fechas de salida y llegada (convertidas en claves de tiempo), y el costo asociado. Se creó mediante la unión de las tablas vuelos e itinerarios y sumando información como ciudades, modelos y aerolíneas.

b. Dimensiones incluidas:

- dim_aerolinea: identifica la compañía aérea responsable del vuelo. Se creó a partir de la tabla aerolineas
- dim_ciudad: agrupa los registros por ciudad, tanto de usuarios como de aeropuertos. Se creó a partir de la tabla ciudades
- dim_aeropuerto: describe los aeropuertos de origen y destino. Se creó a partir de la tabla aeropuertos.
- dim_avion: detalla los aviones, su aerolínea y su modelo. Se creó a partir de la tabla aviones.
- dim_modelo: define el tipo de aeronave. Se creó a partir de la tabla modelos
- dim_usuario: representa al pasajero con atributos personales y ciudad de residencia. Se creó a partir de la tabla usuarios
- dim_calendar: permite analizar los hechos por fechas, con granularidad diaria (año, mes, día). Se creó a partir de la fecha mínima y máxima de la tabla itinerarios.

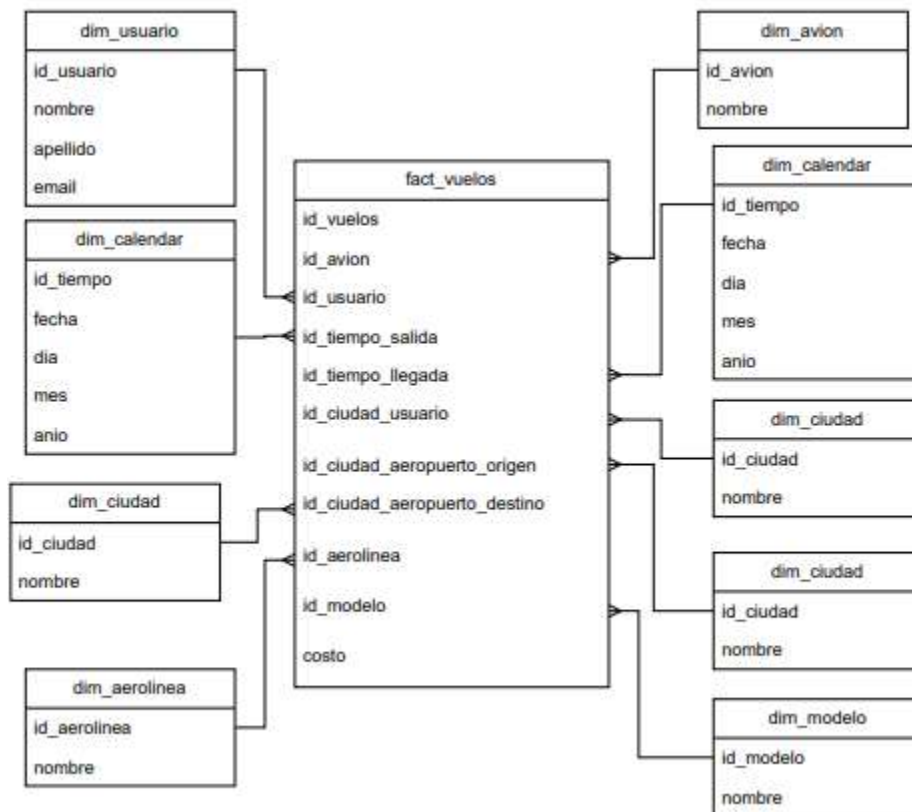


Ilustración 1- Diagrama DataMart

2. Proceso ETL del Data Mart

La presente ETL se encuentra en el siguiente Repositorio de GITHUB. En el archivo README se puede evidenciar el proceso y la estructura a detalle:

Repo: <https://github.com/cristhianalextores/DataMart>

- a. Extracción: Se toman los datos desde la base transaccional IATA.db, obteniendo información de tablas como vuelos, itinerarios, aviones, usuarios, aerolíneas y ciudades.
- b. Transformación: Se limpian y reorganizan los datos para crear las dimensiones del modelo estrella.
 - Se generan claves de tiempo en formato YYYYMMDD para facilitar el análisis temporal.
 - Se unifican las relaciones entre entidades (por ejemplo, vuelos con itinerarios y usuarios).
- c. Carga: Se crean las tablas dimensión en el Data Mart: dim_aerolinea, dim_ciudad, dim_aeropuerto, dim_avion, dim_modelo, dim_usuario y dim_calendar.
Se llena la tabla de hechos fact_vuelos, que integra las claves de las dimensiones y los indicadores de negocio (como el costo del vuelo).

Alcance

La IATA quiere analizar en qué medida la pandemia de COVID-19 afectó el transporte aéreo de pasajeros durante el año 2020. Los requerimientos de análisis son los siguientes:

1. ¿Cuál aerolínea realizó el mayor número de vuelos a la ciudad de Roma en el año 2019 y cuál en el año 2020?

R/ En el 2019 fue Avianca con 11 vuelos y, en el 2020, no se encontraron vuelos a Roma.

Consulta:

```

11  qry = """
12      SELECT
13          dim_calendar.anio,
14          dim_aerolinea.nombre,
15          COUNT(*) AS Numero_Vuelos,
16          ROW_NUMBER() OVER (
17              PARTITION BY dim_calendar.anio
18              ORDER BY COUNT(*) DESC
19          ) AS rn
20      FROM fact_vuelos
21      INNER JOIN dim_avion ON fact_vuelos.id_avion = dim_avion.id_avion
22      INNER JOIN dim_aerolinea ON fact_vuelos.id_aerolinea = dim_aerolinea.id_aerolinea
23      INNER JOIN dim_calendar ON fact_vuelos.id_tiempo_salida = dim_calendar.id_tiempo
24      INNER JOIN dim_ciudad ON fact_vuelos.id_ciudad_aeropuerto_destino = dim_ciudad.id_ciudad
25      WHERE dim_ciudad.nombre = 'Roma'
26      AND dim_calendar.anio IN (2019, 2020)
27      GROUP BY dim_calendar.anio, dim_aerolinea.nombre
28      """
29
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS
anio  nombre  Numero_Vuelos  rn
0  2019  Avianca  11  1
1  2019  Wingo  4  2
(Datamart) PS C:\ProyectosAI\DatamartLab2\src>

```

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.

R/ en el siguiente Cuadro vemos los valores recaudados por Aerolinea el primer semestre de los años 2019 y 2020

Año	Aerolinea	Total_Primer_Semestre
2019	Avianca	\$ 90.089.500,00
2019	Latam	\$ 81.830.300,00
2019	Wingo	\$ 46.104.600,00
2020	Latam	\$ 14.530.000,00
2020	Avianca	\$ 9.890.000,00
2020	Wingo	\$ 6.000.000,00

Consulta:

```

37  qry = """
38      SELECT
39          dim_calendar.anio,
40          dim_aerolinea.nombre AS Aerolinea,
41          SUM(fact_vuelos.costos) AS Total_Semestre
42      FROM fact_vuelos
43      LEFT JOIN dim_avion ON fact_vuelos.id_avion = dim_avion.id_avion
44      LEFT JOIN dim_aerolinea ON fact_vuelos.id_aerolinea = dim_aerolinea.id_aerolinea
45      LEFT JOIN dim_calendar ON fact_vuelos.id_tiempo_salida = dim_calendar.id_tiempo
46      LEFT JOIN dim_ciudad ON fact_vuelos.id_ciudad = dim_ciudad.id_ciudad
47      WHERE dim_calendar.mes BETWEEN 1 AND 6
48      AND dim_calendar.anio IN (2019, 2020)
49      GROUP BY dim_calendar.anio, dim_aerolinea.nombre
50      ORDER BY dim_calendar.anio, Total_Semestre DESC
51      ; """
52
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS
anio  Aerolinea  Total_Semestre
0  2019  Avianca  90089500
1  2019  Latam  81830300
2  2019  Wingo  46104600
3  2020  Latam  14530000
4  2020  Avianca  9890000
5  2020  Wingo  6000000
(Datamart) PS C:\ProyectosAI\DatamartLab2\src>

```

3. ¿Cuál modelo de avión realizó el mayor número de vuelos en el año 2019 y cuál en el año 2020?

R/ El modelo que mayor número de vuelos tuvo fue Airbus 320 tanto para 2019 con 70 vuelos, como para el 2020 con 24 vuelos.

Consulta:

```
37 qry = ""
38 SELECT
39     anio,
40     nombre AS Modelo_Avion,
41     Numero_Vuelos
42 FROM (
43     SELECT
44         dim_calendar.anio,
45         dim_modelo.nombre,
46         COUNT(*) AS Numero_Vuelos,
47         ROW_NUMBER() OVER (
48             PARTITION BY dim_calendar.anio
49             ORDER BY COUNT(*) DESC
50         ) AS rn
51     FROM fact_vuelos
52     LEFT JOIN dim_avion ON fact_vuelos.id_avion = dim_avion.id_avion
53     LEFT JOIN dim_aerolinea ON fact_vuelos.id_aerolinea = dim_aerolinea.id_aerolinea
54     LEFT JOIN dim_calendar ON fact_vuelos.id_tiempo_salida = dim_calendar.id_tiempo
55 )
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

La BD ya existe: ../data/IATA.db
La BD ya existe: ../data/DataMart.db

	anio	Modelo_Avion	Numero_Vuelos
0	2019	Airbus 320	70
1	2020	Airbus 320	24

(Datamart) PS C:\ProyectosAI\DatamartLab2\src> cd .\src\

4. ¿Cuál fue la ciudad cuyos habitantes viajaron más en el año 2019 y cuál en el año 2020?

R/ Medellín fue la ciudad que tuvo más viajeros; en el 2019 tuvo 36 vuelos, para el 2020 tuvo 12 vuelos.

Consulta:

```
rc > consultas.py > ...
37 qry =
38 SELECT
39     anio,
40     nombre AS Ciudad,
41     Numero_Vuelos
42 FROM (
43     SELECT
44         dim_calendar.anio,
45         dim_ciudad.nombre,
46         COUNT(*) AS Numero_Vuelos,
47         ROW_NUMBER() OVER (
48             PARTITION BY dim_calendar.anio
49             ORDER BY COUNT(*) DESC
50         ) AS rn
51     FROM fact_vuelos
52     LEFT JOIN dim_avion ON fact_vuelos.id_avion = dim_avion.id_avion
53     LEFT JOIN dim_aerolinea ON fact_vuelos.id_aerolinea = dim_aerolinea.id_aerolinea
54     LEFT JOIN dim_calendar ON fact_vuelos.id_tiempo_salida = dim_calendar.id_tiempo
55     LEFT JOIN dim_ciudad ON fact_vuelos.id_ciudad = dim_ciudad.id_ciudad
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
La BD ya existe: ../data/IATA.db
(Datamart) PS C:\ProyectosAI\DatamartLab2\src> python.exe consultas.py
La BD ya existe: ../data/IATA.db
La BD ya existe: ../data/DataMart.db
  anio  Ciudad  Numero_Vuelos
0 2019  Medellín    36
1 2020  Medellín    12
(Datamart) PS C:\ProyectosAI\DatamartLab2\src>
```