

Actividad 5 Ejercicio Caso IATA

Alumnos:

Restrepo Rosero Juan

Piñeros Castro Carlos

Doria Atencia Joel

Docente:

Sierra Galvis Martín Vladimir Alonso

Gestión de datos

Maestría en ciencia de datos

Universidad Javeriana de Cali

Noviembre, 2024



Pontificia Universidad
JAVERIANA
Cali

Contexto

Uno de los métodos de transporte más utilizados por los seres humanos es el transporte aéreo. Aunque el número de vuelos diarios se vio afectado por la pandemia de la COVID-19, es bien sabido que son muchos los aviones que diariamente se movilizan de un sitio a otro llevando pasajeros. Es por ello que, para tener un control de estos movimientos, se han creado diversas organizaciones internacionales que llevan un registro de aerolíneas, aviones, itinerarios, equipajes, entre otras cosas. Una de ellas es la Asociación Internacional de Transporte Aéreo, IATA por sus siglas en inglés. Esta asociación surge en el año 1919 en los Países Bajos y estaba conformada por 57 miembros de 31 países pertenecientes, principalmente a Europa y Norteamérica. Años más tarde, en 1945, fue relanzada en Cuba. Actualmente, incluye 290 aerolíneas de 120 países. Sus misiones principales comprenden la promoción de la seguridad, la protección y la fiabilidad en el transporte aéreo y el cuidado del medio ambiente. Todo esto en pro del beneficio económico de sus accionistas. Adicional a esto la IATA representa a las 290 aerolíneas que la conforman, simplifica los procesos y reduce los costos que aumentan su flujo financiero, y actúa como un puente para asegurar el movimiento de las personas alrededor del mundo.

Caso

La Asociación Internacional de Transporte Aéreo, IATA, posee una base de datos donde tiene almacenada la información recolectada de todos los vuelos hechos a nivel global a partir del año 1945. El Modelo Relacional que la IATA utilizó para la construcción de su Base de Datos Relacional fue el siguiente:

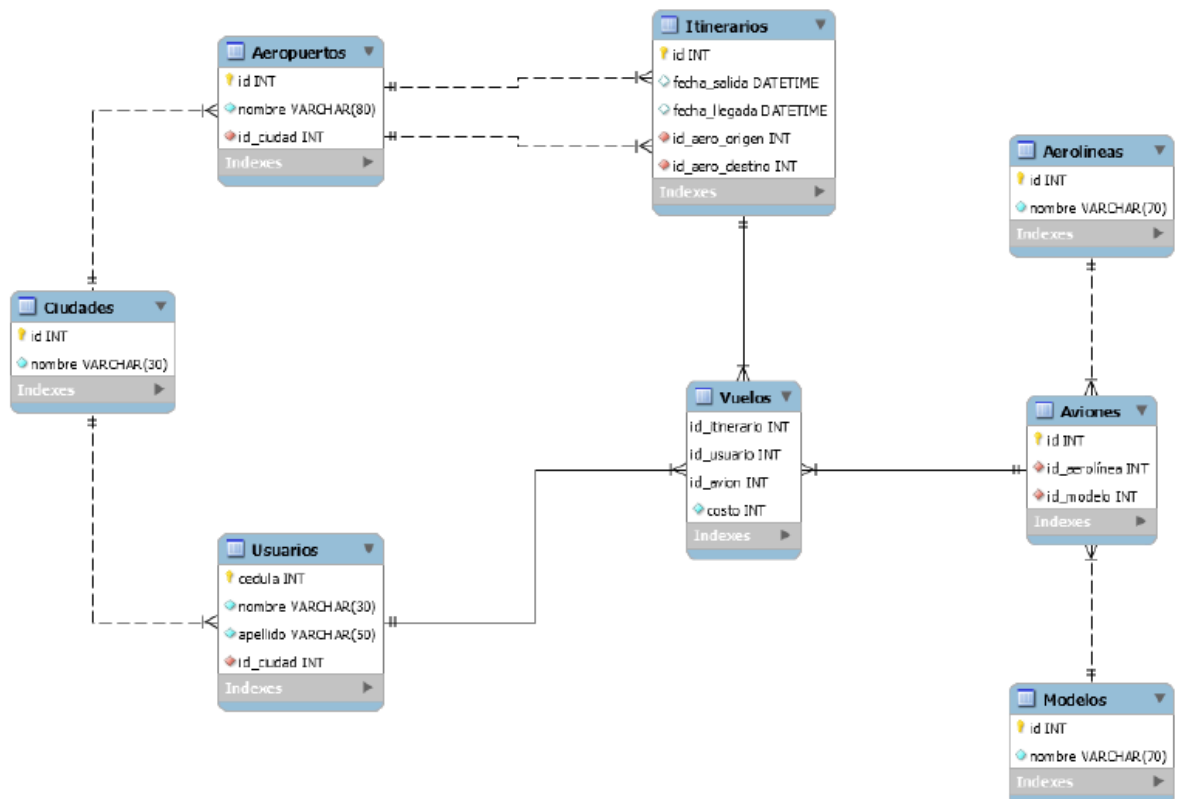


Fig 1: Base de Datos de la IATA

Alcance

En este Modelo Relacional se encuentran almacenados los siguientes datos:

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.

La IATA, utilizando los datos almacenados en su base de datos relacional, busca abordar diversas necesidades de análisis relacionadas con los vuelos. En particular, desea evaluar el impacto de la pandemia de COVID-19 en el transporte aéreo de pasajeros durante el año 2020. Los análisis a realizar incluyen 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?
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?

Para este fin, la IATA contrata a un grupo de científicos de datos, expertos en la gestión de datos en bases de datos analíticas, para que implemente, a partir de su base de datos relacional, un Modelo Estrella que permita dar respuesta a los requerimientos de análisis. Para responder estos requerimientos y dar completa solución al caso deben tener en cuenta los siguientes puntos:

Análisis Crítico

El análisis crítico implica examinar tanto la estructura de los datos existentes como la pertinencia de la información almacenada para satisfacer los requerimientos planteados. La base de datos relacional de la IATA contiene información relacionada con usuarios, aeropuertos, aviones, itinerarios, modelos, ciudades, aerolíneas y vuelos.

Ventajas del Enfoque Relacional

- **Integridad de Datos:** La estructura relacional asegura la precisión y consistencia mediante el uso de claves primarias y foráneas.
- **Facilidad de Acceso:** Facilita consultas complejas a través de SQL, lo que permite obtener información específica y detallada de manera eficiente.

Desventajas del Enfoque Relacional

- **Rendimiento:** El procesamiento de grandes volúmenes de datos puede ser lento, especialmente cuando las consultas requieren unir múltiples tablas.
- **Escalabilidad:** Escalar bases de datos relacionales puede ser complicado, especialmente frente al creciente volumen de datos en la industria de la aviación.

Proceso de Construcción del Modelo Multidimensional

Para cumplir con los requerimientos analíticos de la IATA, se adoptará un **Modelo Estrella**. Este enfoque es ideal para análisis OLAP (Procesamiento Analítico en Línea), ya que ofrece una estructura eficiente y optimizada para realizar consultas analíticas complejas.

Un modelo estrella es una estructura de base de datos diseñada para consultas analíticas rápidas [1]. Consiste en una tabla de hechos central, que contiene las medidas numéricas (como el costo de los vuelos), rodeada de tablas de dimensiones, que contienen atributos descriptivos (como la fecha, la aerolínea, la ciudad origen y ciudad destino)

Pasos para la Construcción del Modelo Estrella

1. Identificación de Hechos y Dimensiones

- **Tabla de Hechos:** Se centra en el registro de vuelos, incluyendo métricas como costos.
- **Dimensiones:**
 - **Dimensión Avión:** Incluye atributos como modelo y aerolínea.
 - **Dimensión Itinerario:** Detalla año, semestre y ciudad.

2. Diseño del Esquema Estrella

- **Tabla de Hechos:** Almacena claves foráneas de las dimensiones junto con las medidas, como el costo total.
- **Tablas de Dimensiones:** Cada tabla de dimensión describe atributos detallados que caracterizan las dimensiones correspondientes.

Tabla de Hechos (Vuelos_usuario):

- ID_Itinerario
- fecha_salida
- id_aerolinea
- id_modelo
- id_ciudad_origen
- id_ciudad_destino
- id_usuario
- id_itinerario
- Costo

Dimensión Aerolinea:

- ID_Aerolinea
- Nombre

Dimensión Ciudad origen:

- ID_ciudad
- Nombre

Dimensión Ciudad destino:

- ID_ciudad

- Nombre

Dimensión Fecha_vuelo:

- fecha_salida
- año
- mes
- día
- semestre

Dimensión Modelos:

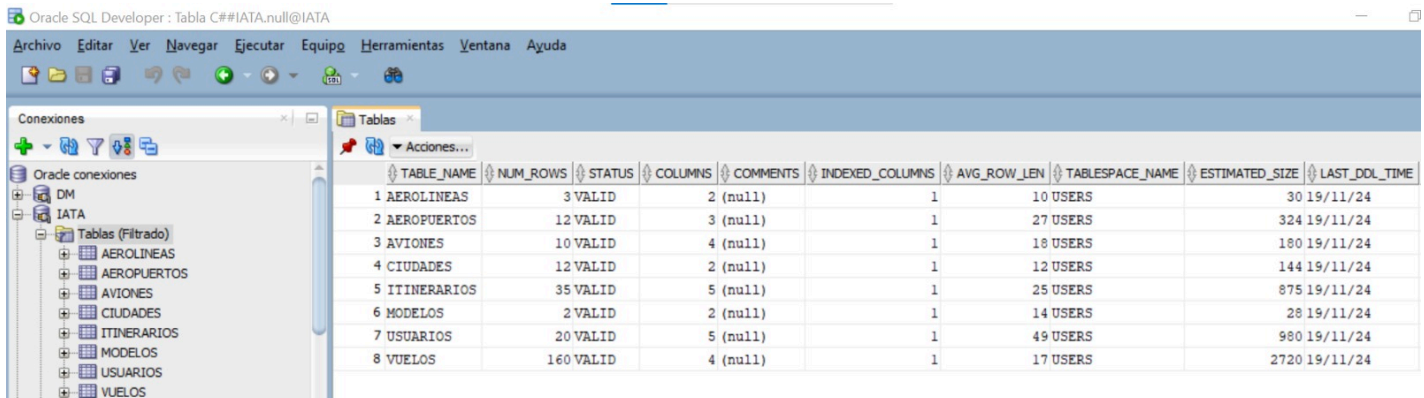
- ID_modelo
- Nombre

Dimension usuarios:

- id_usuario
- Nombre
- Apellido
- Ciudad Residencia

Carga de los datos en Oracle SQL Developer

Inicialmente, se realiza la carga de la base de datos original con sql developer:



TABLE_NAME	NUM_ROWS	STATUS	COLUMNS	COMMENTS	INDEXED_COLUMNS	AVG_ROW_LEN	TABLESPACE_NAME	ESTIMATED_SIZE	LAST_DDL_TIME
1 AEROLINEAS	3	VALID	2 (null)		1	10	USERS	30	19/11/24
2 AEROPUERTOS	12	VALID	3 (null)		1	27	USERS	324	19/11/24
3 AVIONES	10	VALID	4 (null)		1	18	USERS	180	19/11/24
4 CIUDADES	12	VALID	2 (null)		1	12	USERS	144	19/11/24
5 ITINERARIOS	35	VALID	5 (null)		1	25	USERS	875	19/11/24
6 MODELOS	2	VALID	2 (null)		1	14	USERS	28	19/11/24
7 USUARIOS	20	VALID	5 (null)		1	49	USERS	980	19/11/24
8 VUELOS	160	VALID	4 (null)		1	17	USERS	2720	19/11/24

Fig 2: Base datos IATA en Oracle SQL Developer

Data Mart físico en Oracle

Un data mart físico es una base de datos optimizada para consultas analíticas, diseñada específicamente para responder preguntas de negocio de manera rápida y eficiente. A diferencia de una base de datos transaccional, un data mart se enfoca en almacenar y organizar datos históricos para análisis, en lugar de procesar transacciones en tiempo real [2]. Para el caso actual, se construyó el data mart físico en Oracle como se muestra en la fig. 3:

TABLE_NAME	NUM_ROWS	STATUS	COLUMNS	COMMENTS	INDEXED_COLUMNS	AVG_ROW_LEN	TABLESPACE_NAME
1 AEROLINEAS	3	VALID	2 (null)		1	10	USERS
2 CIUDAD_DESTINO	12	VALID	2 (null)		1	12	USERS
3 CIUDAD_ORIGEN	12	VALID	2 (null)		1	12	USERS
4 CIUDAD_ORIGEN	12	VALID	2 (null)		2	12	USERS
5 FECHA_VUELO	35	VALID	5 (null)		1	18	USERS
6 MODELOS	2	VALID	2 (null)		1	14	USERS
7 USUARIOS	20	VALID	4 (null)		1	25	USERS
8 VUELOS_USUARIO	160	VALID	8 (null)		0	34	USERS

Fig 3: Data mart físico en Oracle SQL Developer

Cubo OLAP con Pentaho

Un cubo OLAP (Online Analytical Processing) es una estructura de datos multidimensional que permite a los usuarios realizar análisis complejos y rápidos sobre grandes volúmenes de datos [3]. Los cubos OLAP facilitan la exploración de datos desde diferentes perspectivas, lo que los convierte en una herramienta esencial para la toma de decisiones basadas en datos. Se creó el cubo OLAP para el análisis utilizando la herramienta Schema Workbench de Pentaho:

Attribute	
name	DM_cube2
description	
caption	
cache	<input checked="" type="checkbox"/>
enabled	<input checked="" type="checkbox"/>
visible	<input checked="" type="checkbox"/>

Fig 4: Cubo OLAP

Procesos ETL con Data-Integration

Se utiliza la herramienta Spoon para llevar a cabo los procesos de carga en el Data Mart, comenzando con la configuración de la conexión a las bases de datos. Posteriormente, se ejecutan los procesos de transformación, iniciando con la opción "Input - Table Input". Así, se implementa el proceso ETL para transferir los datos al Data Mart.

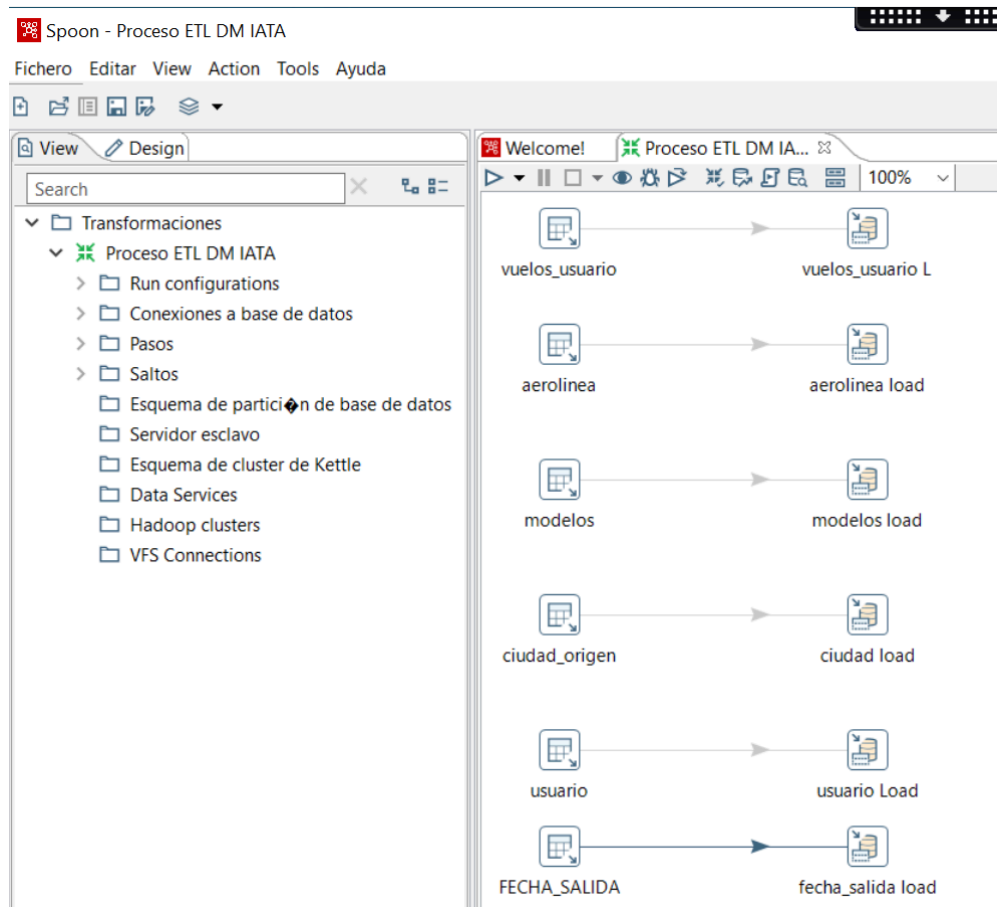


Fig 5: Proceso de ETL del data mart

Visualizando el Data Mart y el Cubo OLAP con LinceBI

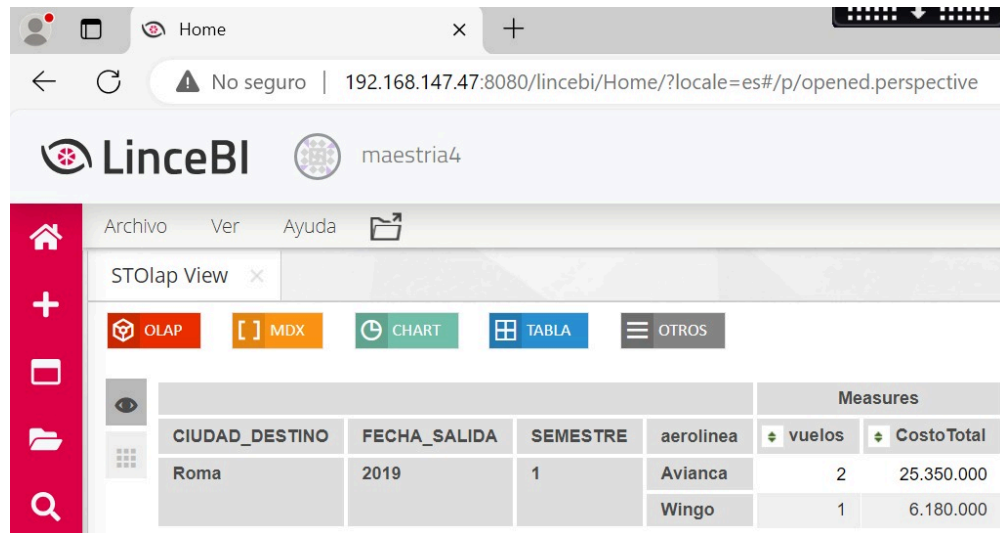
LinceBI es una plataforma de Business Intelligence (BI) de código abierto diseñada para ofrecer una solución completa y escalable para el análisis de datos. Basada en Pentaho, LinceBI proporciona un conjunto de herramientas y funcionalidades que permiten a las empresas extraer valor de sus datos y tomar decisiones más informadas [4].

LinceBI ofrece una gran flexibilidad para adaptarse a las necesidades específicas de cada organización, permitiendo una personalización profunda de los dashboards, informes y análisis.

Utilizando esta herramienta, se responden los requerimientos de análisis planteados en el caso:

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

En el primer semestre de 2019, Avianca tuvo el mayor número de vuelos a Roma. En el 2020 no se registraron vuelos a Roma:



Home

No seguro | 192.168.147.47:8080/lincebi/Home/?locale=es#/p/opened.perspective

LinceBI maestría4

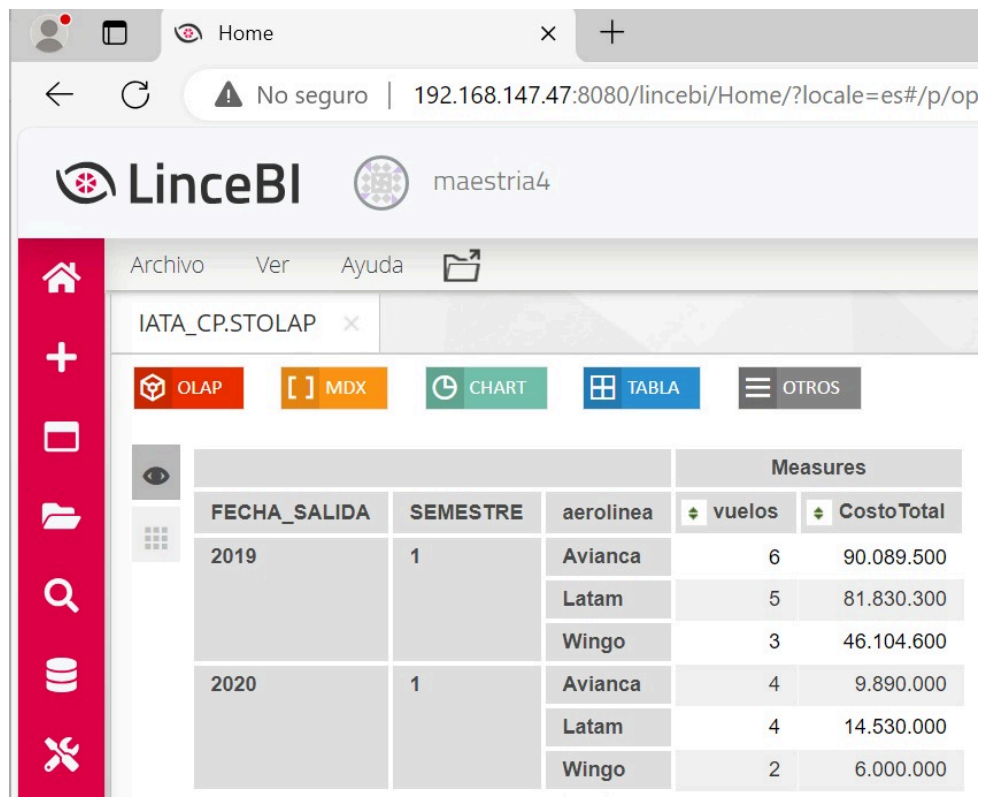
Archivo Ver Ayuda

STOLap View

OLAP MDX CHART TABLA OTROS

				Measures	
CIUDAD_DESTINO	FECHA_SALIDA	SEMESTRE	aerolinea	vuelos	CostoTotal
Roma	2019	1	Avianca	2	25.350.000
			Wingo	1	6.180.000

Total de dinero recaudado por vuelos de cada aerolínea en el primer semestre del año 2019 y en el primer semestre del año 2020.



Home

No seguro | 192.168.147.47:8080/lincebi/Home/?locale=es#/p/op

LinceBI maestría4

Archivo Ver Ayuda

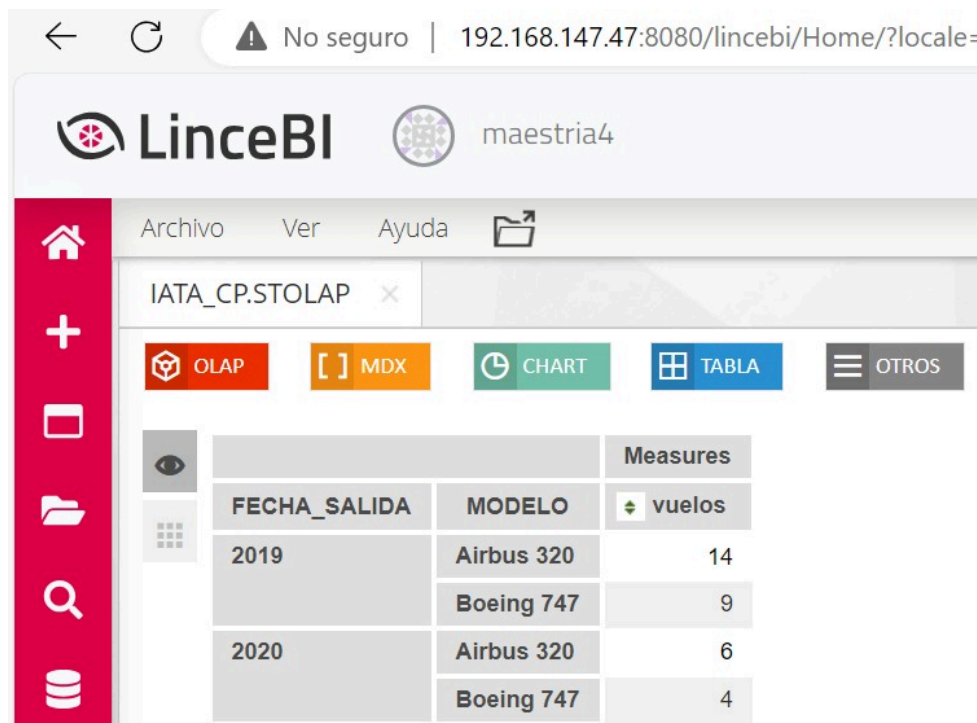
IATA_CP.STOLAP

OLAP MDX CHART TABLA OTROS

			Measures	
FECHA_SALIDA	SEMESTRE	aerolinea	vuelos	CostoTotal
2019	1	Avianca	6	90.089.500
		Latam	5	81.830.300
		Wingo	3	46.104.600
2020	1	Avianca	4	9.890.000
		Latam	4	14.530.000
		Wingo	2	6.000.000

En el primer semestre de 2019, Avianca fue la aerolínea que generó mayores ingresos, seguida de Latam y Wingo. Todas las aerolíneas experimentaron una disminución considerable en sus ingresos durante el primer semestre de 2020. Esta reducción podría atribuirse a diversos factores como la pandemia de COVID-19, que afectó significativamente al sector de la aviación. A pesar de la caída generalizada, Avianca continúa siendo la aerolínea con mayores ingresos en el primer semestre de 2020, aunque con una diferencia menor respecto a Latam.

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

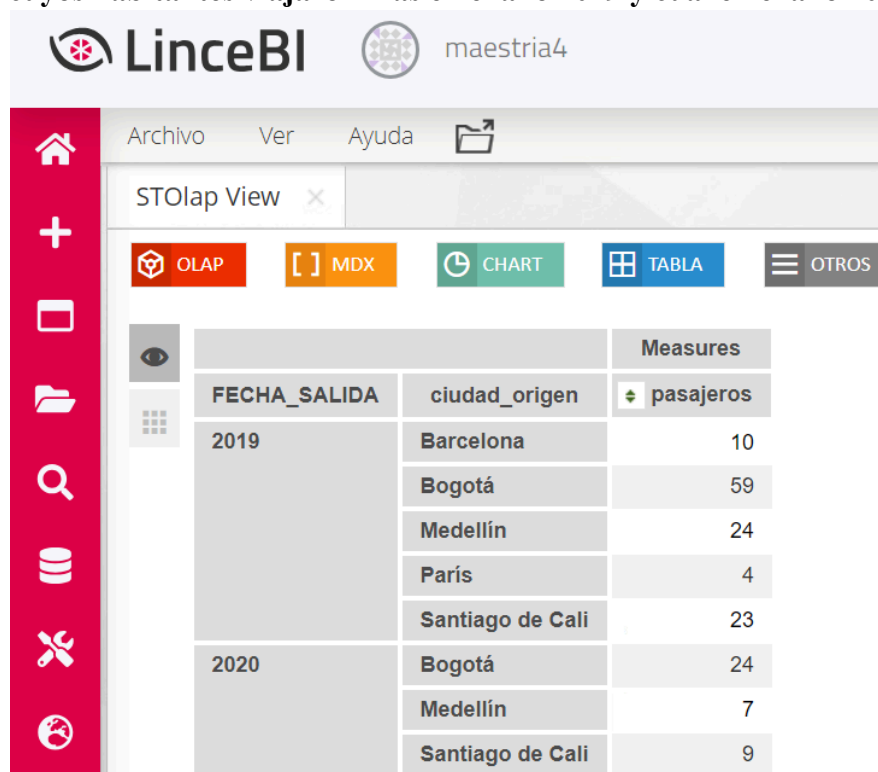


The screenshot shows the LinceBI web application interface. The browser address bar displays "192.168.147.47:8080/lincebi/Home/?locale=". The application header includes the LinceBI logo and the user "maestria4". A sidebar on the left contains navigation icons. The main content area shows a table for "IATA_CP.STOLAP" with columns "FECHA_SALIDA", "MODELO", and "vuelos". The table data is as follows:

FECHA_SALIDA	MODELO	vuelos
2019	Airbus 320	14
	Boeing 747	9
2020	Airbus 320	6
	Boeing 747	4

El modelo de avión que realizó el mayor número de vuelos en 2019 fue el **Airbus 320** con un total de 14 vuelos. Al igual que en 2019, el modelo de avión que realizó el mayor número de vuelos en 2020 fue el **Airbus 320** con un total de 6 vuelos.

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



The screenshot shows the LinceBI web application interface. The browser address bar displays "192.168.147.47:8080/lincebi/Home/?locale=". The application header includes the LinceBI logo and the user "maestria4". A sidebar on the left contains navigation icons. The main content area shows a table for "STOLap View" with columns "FECHA_SALIDA", "ciudad_origen", and "pasajeros". The table data is as follows:

FECHA_SALIDA	ciudad_origen	pasajeros
2019	Barcelona	10
	Bogotá	59
	Medellín	24
	París	4
	Santiago de Cali	23
2020	Bogotá	24
	Medellín	7
	Santiago de Cali	9

La ciudad de **Bogotá** fue la que registró el mayor número de pasajeros saliendo, con un total de 59. Esto

indica que Bogotá fue el principal punto de origen de los vuelos en ese año según los datos proporcionados. Al igual que en 2019, **Bogotá** volvió a ser la ciudad con el mayor número de pasajeros saliendo, con un total de 24.

Conclusiones:

- La creación de un modelo estrella en Oracle implica un diseño cuidadoso y la ejecución de comandos SQL precisos para construir un modelo sólido y eficiente para los análisis requeridos.
- La calidad de los datos de origen tiene un impacto directo en la calidad de los datos del data mart. Es fundamental implementar procesos de limpieza y transformación de datos robustos
- El rendimiento del proceso ETL es un factor crítico, especialmente para grandes volúmenes de datos. Es necesario optimizar los procesos y utilizar las herramientas adecuadas para garantizar un tiempo de ejecución aceptable.
- El data mart requiere un mantenimiento continuo para garantizar la actualización de los datos y la adaptación a los cambios en el negocio.
- Es fundamental implementar medidas de seguridad adecuadas para proteger los datos del data mart, como control de acceso, encriptación y auditoría.
- Spoon ofrece una amplia gama de transformaciones y componentes que permiten realizar procesos ETL complejos y personalizados
- Pentaho ofrece una versión comunitaria gratuita, lo que lo hace accesible para proyectos de menor escala.

Referencias:

- [1] Wikipedia, "Esquema en estrella," Wikipedia, 30-may-2024. [En línea]. Disponible: https://es.wikipedia.org/wiki/Esquema_en_estrella. [Accedido: 25-nov-2024].
- [2] IBM, "¿Qué es un data mart?," IBM, 2024. [En línea]. Disponible: <https://www.ibm.com/co-es/cloud/learn/data-mart>. [Accedido: 25-nov-2024].
- [3] IBM, "¿Qué es OLAP (procesamiento analítico en línea)?," IBM, 2024. [En línea]. Disponible: <https://www.ibm.com/co-es/cloud/learn/olap>. [Accedido: 25-nov-2024].
- [4] Redacción Byte TI, "Dade2 Cloud y StrateBI ofrecen LinceBI: Solución Analítica basada en Open Source," Revista Byte TI, 6 abril, 2021. [En línea]. Disponible: <https://revistabyte.es/actualidad-ti/dade2-cloud-y-stratebi-ofrecen-lincebi/>. [Accedido: 25-nov-2024].
- [5] peter-myers, «Descripción de un esquema de estrella e importancia para Power BI - Power BI». Accedido: 25 de noviembre de 2024. [En línea]. Disponible en: <https://learn.microsoft.com/es-es/power-bi/guidance/star-schema>
- [6] PriskeyJeronika-MS, «Introducción a los cubos OLAP para análisis avanzados». Accedido: 25 de noviembre de 2024. [En línea]. Disponible en: <https://learn.microsoft.com/es-es/system-center/scsm/olap-cubes-overview?view=sc-sm-2025>
- [7] «¿Qué es el OLAP? - Explicación del procesamiento analítico en línea - AWS». Accedido: 25 de noviembre de 2024. [En línea]. Disponible en: <https://aws.amazon.com/es/what-is/olap/>

[8] «¿Qué es ETL? - Explicación de extracción, transformación y carga (ETL) - AWS», Amazon Web Services, Inc. Accedido: 25 de noviembre de 2024. [En línea]. Disponible en: <https://aws.amazon.com/es/what-is/etl/>