Práctica 1 – Data Warehouse, ETL y SQL analítico (4 puntos)

Enunciado

Diseñar, implementar y explotar un DW para Airbnb.

Fuente de datos base: https://insideairbnb.com/get-the-data/

Tarea 1 (0,5+0,5 puntos): Diseño del modelo multidimensional utilizando las fases del modelo de Kimball.

1. Determinar el proceso de negocio a analizar.
2. Establecer el nivel de detalle o granularidad (de la tabla de hechos).
3. Elegir las dimensiones.
4. Determinar los hechos medibles.

Más concretamente, se debe incluir:

- Una explicación de las suposiciones realizadas sobre los requisitos de explotación y los atributos a considerar. Debe considerarse aquí las necesidades de consulta. Para la justificación de las dimensiones y sus atributos considerados se debe incluir ejemplos de consultas analíticas en donde se usarían.
- Un diagrama, siguiendo la notación DFM vista en clase, que refleje los hechos, dimensiones y atributos seleccionados.
    - Se recomienda seleccionar nombres descriptivos para los atributos, evitando el uso de abreviaturas. No es necesario incluir explicaciones detalladas de los atributos, salvo cuando sean necesarias para entender su significado o funcionamiento.
    - Acompañando al diagrama, debe incluirse una justificación de las decisiones más relevantes tomadas, y en particular las relacionadas con:
        - Dimensiones que tengan particularidades relevantes: dimensiones degeneradas, dimensiones que cambian en el tiempo y la estrategia seleccionada para representarlas (y la justificación de la elección de dicha estrategia), etc.
        - Medidas/métricas seleccionadas con su significado. Se debe incluir información sobre el carácter aditivo, semiaditivo o no aditivo de cada una de las métricas.

Tarea 2 (0,5 punto): Diseño del modelo lógico e implementación del DW en Oracle

- Proponer el diagrama del modelo lógico para el diseño realizado (normalmente será un modelo sencillo, en estrella o en copo de nieve), incluyendo una breve explicación de las decisiones tomadas y crear las tablas resultantes en Oracle.
- Crear las tablas necesarias en Oracle.

Tarea 3 (1 punto): Diseño e implementación del ETL

- Diseñar e implementar el proceso ETL (Extract, Transform, Load) completo en Apache Hop, que permita extraer, depurar/limpiar, transformar y cargar los datos reales de Airbnb en el Data Warehouse creado en la tarea 2. El proceso deberá automatizar la carga desde los ficheros fuente hasta las tablas destino en Oracle, garantizando la calidad, consistencia y trazabilidad de los datos.
- Deberán especificarse los datasets de origen utilizados y documentar las principales transformaciones realizadas (limpieza, conversión de tipos, homogeneización de categorías, generación de claves sustitutas, etc.). Además, es necesario describir cómo se tratan valores nulos o inconsistentes, claves foráneas faltantes, formatos de fecha y moneda, posibles duplicados...
- El fichero principal de ejecución debe llamarse: etl_airbnb_main.hpl y debe ser ejecutable en Hop sin necesidad de editar las rutas internas. El entregable final debe incluir una carpeta datasets (o enlace a los ficheros fuente utilizados).
- El proyecto de Hop debe incluir al menos: Un job principal (.hpl) que coordine la ejecución completa del proceso ETL con varias transformaciones específicas para cada tabla. El pipeline debe ser ejecutable de forma completa desde un único job principal.
- Todas las rutas de ficheros y conexiones deben configurarse de manera relativa (no absolutas) para facilitar su ejecución en otro entorno. La conexión a Oracle debe configurarse mediante variables de entorno de Hop (${ORACLE_HOST}, ${ORACLE_USER}, etc.), documentadas en el informe.
- Todas las transformaciones deben estar claramente nombradas y documentadas dentro del proyecto Hop. También deben utilizarse Hop variables para parametrizar rutas, nombres de archivos y credenciales, facilitando la reutilización.
- No es requisito mínimo, pero se valorará positivamente la inclusión de al menos una de las siguientes funcionalidades avanzadas: Gestión de dimensiones lentamente cambiantes (SCD tipo 1 o 2), control de carga incremental (detección de nuevos registros o cambios) o mecanismo de logging o control de errores mediante variables o salidas de Hop.

Tarea 3 (1,5 punto): Explotación del DW

- Se deben plantear y resolver tres consultas analíticas relevantes sobre el DW implementado. Para cada consulta, debe incluirse una breve descripción de la consulta (qué hace, y por qué es relevante o para qué podría aplicarse), y la sentencia SQL que resuelve dicha consulta sobre el esquema seleccionado. Las consultas deben de ser realistas y diferentes (utilizando diferentes tipos de consultas analíticas).
- Se debe crear un informe en Power BI que muestre visualizaciones interactivas de los datos, incluyendo alguna métrica DAX.

Comentarios:

- El enunciado es intencionadamente abierto, a fin de permitir múltiples interpretaciones. Se valorará fundamentalmente la consistencia entre las suposiciones realizadas y el modelo generado, al no tener conocimiento experto del dominio.

- Se espera que se incluyan al menos unas 4-5 dimensiones, de forma que el alcance del ejercicio sea suficiente para valorar los conocimientos del tema.

Grupos

La práctica se realizará en grupos de cuatro estudiantes.

Fechas de entregas y defensa

Las entregas se realizarán a través del campus virtual.

Primera entrega (sólo tarea 1): 24 de octubre de 2025, hasta las 23:59 (0,5 puntos)

Segunda entrega (tarea 1 revisada + tareas 2, 3 y 4): 14 de noviembre de 2025, hasta las 23:59 (3,5 puntos, que se otorgarán después de la defensa).

Defensa: durante último día de clase de prácticas de la asignatura, es decir, 16/18 de diciembre de 2025. La defensa será presencial, y permitirá comprobar que los estudiantes han adquirido correctamente los fundamentos del diseño de DW y ETL utilizados, así como su explotación con SQL analítico.

Entregables

Primera entrega: un documento pdf que refleje el trabajo realizado en la primera tarea (descripción del entorno analítico, diagrama y justificaciones).

Segunda entrega: un fichero comprimido que incluya

- Un documento pdf con el detalle del trabajo realizado, e incluya lo solicitado en cada tarea (descripción del entorno analítico, diagramas, documentación de los procesos de ETL, enunciados de consultas, capturas de pantallas, justificaciones, etc.).
- Fichero .hpl con el pipeline de Hop implementado.
- Scripts complementarios que hayan sido utilizados (creación de tablas, resolución de consultas, preprocesado de datos, …).