TP – Flujo de datos en un DWA

El objetivo de este trabajo práctico es desarrollar todas las capas de datos y ejecutar los procesos correspondientes del flujo *end-to-end* en un DWA (*Data Warehouse* Analítico), desde la adquisición hasta la publicación y la explotación.

El material básico para la elaboración del presente trabajo se encuentra publicado en la plataforma del curso, además de lo expuesto en clase.

Aclaraciones

- Este trabajo debe elaborarse por equipos.
- La **cantidad optima** de integrantes por equipo será publicada en la plataforma.
- Los equipos con más integrantes que la cantidad optima serán penalizados.
- La entrega de este TP consiste en **entregar un documento en la plataforma** resumiendo lo realizado según se especifica más abajo. Además se deben entregar todos los componentes desarrollados.
- En la última clase cada grupo deberá **exponer una síntesis de los aspectos relevante del trabajo realizado** con una duración máxima de 10'. Excepcionalmente, podría reemplazarse con un video.
- Las fechas de publicación y presentación serán indicadas en la plataforma
- **Incluyan en los archivos a entregar la lista de los integrantes. Se recomienda considerar una carátula en donde se identifique el grado/posgrado, la cohorte, la materia, el título del informe, los integrantes del equipo y la fecha.**
- La evaluación se realizará según la **rúbrica** descrita más abajo.
- Los integrantes de cada equipo obtendrán la **misma calificación**.
- Los docentes evaluarán el trabajo realizado por lo que se manifiesta en la presentación y en los documentos entregados, por lo tanto se recomienda una elaboración cuidada y comentada. El contenido debe transmitir las tareas realizadas con la especificidad suficiente para comprenderlas pero sin entrar en detalles irrelevantes. Es deseable que se comenten sintéticamente los problemas o contratiempos que hayan enfrentado.
- No copien textos externos, si fuera necesario, citen la fuente.

Contexto general

Se publicarán dos conjuntos de datasets provenientes de una base de datos transaccional y de otras fuentes secundarias.

1. **Ingesta1**: corresponde a los datos de una ingesta inicial para alimentar un DWA vacío. Los datos fueron obtenidos de un sistema transaccional persistidos en un modelo relacional tradicional.
2. **Ingesta2**: corresponde a un subconjunto de la misma entidad de datos que se utilizará para una actualización posterior. Además se entrega una tabla externa para incluir en el DWA.

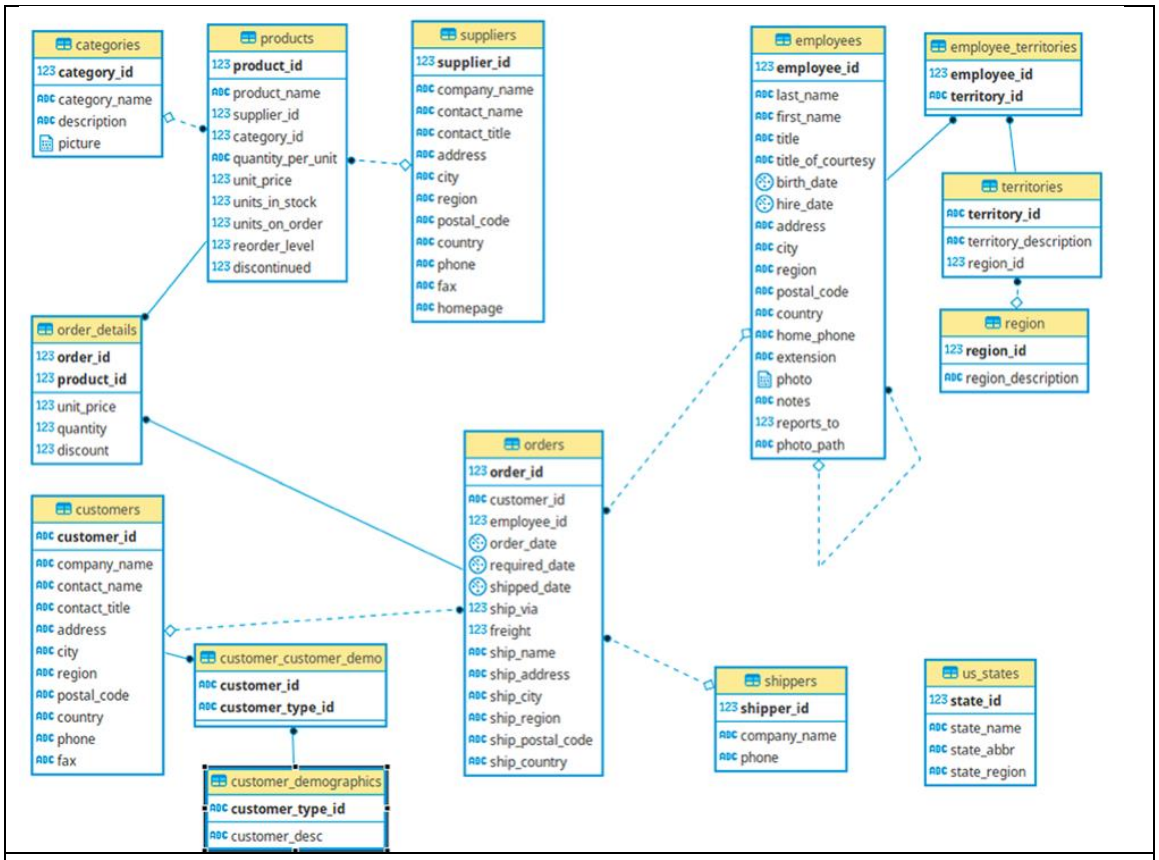
Se deberán desarrollar todas las capas y procesos necesarios para implementar el flujo de datos dentro del DWA para proveer de información a la organización. El objetivo final es desarrollar un tablero de visualización a partir de un Producto de Datos obtenido de los datos persistidos en el DWA.

El desarrollo debe incluir los **controles de calidad** necesarios y su persistencia, la **memoria** institucional, el **enriquecimiento** de los datos, la gestión de la **metadata**, la publicación de **productos** de datos y de alguna **aplicación** de datos.

Toda la implementación se debe desarrollar en una base de datos relacional utilizando comandos SQL estándares. Se recomienda SQLite (con SQLLiteStudio) y PBI, pero se puede utilizar cualquier base relacional estándar y cualquier software de tableros.

Descripción detallada

A modo ilustrativo pero no exhaustivo, en la siguiente imagen se muestra el DER de la base transaccional.



DER Northwind

Desarrollar el flujo de datos de un DWA en etapas como se indica a continuación.

Se pide:

- Desarrollar scripts para cada tarea.
- Llevar un inventario de scripts en una tabla con identificación.
- Toda ejecución de un scripts debe quedar registrada en una tabla de LOG con inicio, fin y resultado.
- Los controles de calidad de cada campo/tabla deben quedar persistidos en el DQM.
- También debe persistirse un perfilado (totales de control) de cada tabla validada en el DQM.
- Todo proceso de transformación o copia de datos debe persistir una huella en el DQM.

Etapa 1 - **Adquisición**

1) Analizar las tablas (.CSV) incluidas en **Ingesta1**.
2) Comparar la estructura de las tablas y el modelo de entidad relación. Adecuar si fuera necesario.
3) Crear un área temporal y persistir los datos de los .CSV:
   a) Crear un juego de tablas TXT con todos los campos en formato TEXT o equivalente.
   b) Crear un juego de tablas TMP con los tipos de datos correspondientes al DER. Incluir claves primarias
   c) Bajar los CSV a estas tablas TXT.

d) Validar cada campo de cada tabla para verificar si su contenido es compatible con el formato, de forma que pueda se convertida en su totalidad en su correspondiente TMP. Desarrollar y ejecutar los scripts de validación sobre TXT.
e) Validad que no va a fallar la clave primaria.
f) Realizar un perfilado de los datos y una validación básica (Outliers, faltantes, etc.).
g) Si la validación da bien, pasar las tablas TXT a las tablas TMP.
h) Desarrollar y ejecutar scripts de validación de la integridad referencial sobre las tablas TMP.

4) Crear las tablas DQM para que soporten lo anterior.

Etapa 2 – Ingeniería

5) Crear el soporte para la Metadata y utilizarlo para describir las entidades.
6) Definir y crear el Modelo Dimensional del DWA y documentarlo en la Metadata. Debe incluir una capa de Memoria y una de Enriquecimiento (datos derivados).
7) Diseñar y crear el DQM para poder persistir los procesos ejecutados sobre el DWA, los descriptivos de cada entidad procesada y los indicadores de calidad. Documentar el diseño en la Metadata.
8) Realizar la carga inicial del DWA con los datos que se seleccionen de las tablas recibidas y procesadas.
    a) Definir los controles de calidad de **ingesta** para cada tabla, los datos que se persistirán en el DQM y los indicadores y límites para aceptar o rechazar los datasets. Realizar y ejecutar los scripts correspondientes. Tener en cuenta: *outliers*, datos faltantes, valores que no respetan los formatos, etc.
    b) Definir los controles de calidad de **integración** para el conjunto de tablas, los datos que se persistirán en el DQM y los indicadores y límites para aceptar o rechazar los datasets. Realizar y ejecutar los scripts correspondientes. Tener en cuenta: la integridad referencial e indicadores de comparación.
    c) Ingestar los datos de Ingesta1 en el DWA definido. Las datos se deben insertar desde las tablas TMP validadas. Actualizar todas las capas. Siempre y cuando se superen los umbrales de calidad. Tener en cuenta el orden de prevalencia.

Etapa 3 - Actualización

9) Considerar también la tabla de países (World-Data-2023) y vincularla con las tablas que correspondan. Modificar todos los componentes afectados (DQM, DWA, Metadata).
10) Actualización:
    a) Persistir en área temporal las tablas entregadas como Ingesta2.
    b) Repetir los pasos definidos para Ingesta1 que sean adecuados para Ingesta2.
    c) Considerar altas, bajas y modificaciones. Tener en cuenta el orden de prevalencia para las actualizaciones.
    d) Si hubiera errores se debe decidir si se cancela toda la actualización, se procesa en parte o en su totalidad. Lo que suceda debe quedar registrado en el DQM.
    e) Para el caso del presente TP es válido corregir el CSV si este tuviera un error y volver a realizar la ingesta hasta que pase los controles.
    f) Se debe considerar además la **capa de Memoria** para persistir la historia de los campos que han sido modificados.
    g) Se debe considerar además actualizar la **capa de Enriquecimiento** para persistir los datos derivados que se vean afectados.
    h) Desarrollar y ejecutar los scripts correspondientes para actualizar el DWA con los nuevos datos.
    i) Actualizar el DQM.
    j) Actualizar la Metadata.

Etapa 4 - Publicación

11) Publicar un producto de datos resultante del DWA para un caso de negocio particular y un período dado si corresponde.
    a) Desarrollar y ejecutar los scripts necesarios.
    b) Dejar huella en el DQM.
    c) Dejar huella en la Metadata de ser necesario.
12) Explotación
    a) Desarrollar y publicar un tablero para la visualización del producto de datos desarrollado. Dejar huella en el DQM y en Metadata de ser necesario.
    b) Desarrollar y publicar un tablero de visualización que permita navegar por los datos persistidos en el DQM. Dejar huella en el DQM y en Metadata de ser necesario.

Entrega final

13) Para la entrega final se puede corregir lo entregado en las etapas anteriores.
14) Entregar una presentación o un documento que describa lo realizado.
15) Entregar todos los componentes desarrollados.

Recomendaciones

16) Se puede utilizar un único esquema de base de datos para todas las capas. Se recomienda identificar las distintas capas con un prefijo, por ejemplo:
    a) TXT_ para las temporales sin formato.
    b) TMP_ para temporales para la validación de ingesta.
    c) ING_ para la capa temporal a ingestar.
    d) DWA_ para el Datawarehouse.
    e) DQM_ para el Data Quality Mart.
    f) DWM_ para la memoria.
    g) MET_ para la metadata.
    h) DPxx_ para los productos de datos.
17) En https://en.wikiversity.org/wiki/Database_Examples/Northwind/SQLite tienen algunas ayudas para crear las tablas.
18) En todo control de calidad se deben detectar los errores, faltantes o inconsistencias y describir el proceso que se llevaría adelante para corregirlos. Los indicadores de calidad deberán permitir decidir si la entidad se procesa o no, completa o parcialmente.
19) El DQM debe persistir los indicadores que sirvan para determinar la calidad de los datos procesados y una estadística que permita describir cuantitativamente al conjunto.
20) No es necesario pasar al DWA todos los atributos de las entidades originales, decidan cuáles son importantes y justifiquen. Lo mismo vale para las capas de Memoria y Enriquecimiento.
21) Sean prolijos y explícitos al codificar los scripts y documenten en el mismo fuente.
22) Este es un TP para una materia de DW, por lo tanto el foco debe estar puesto en los conceptos fundamentales de esta disciplina. El uso de la BD es solo una herramienta para gestionar el DWA. Existen múltiples herramientas para realizar los procesos solicitados, pero en este caso se pide realizarlos utilizando solo SQL estándar.
23) Se prefiere un trabajo simple, que cubra todos los aspectos pero no necesariamente exhaustivo en los detalles y, por supuesto, bien hecho.
24) Lo que no esté especificado y sea necesario para el trabajo, decídanlo y justifíquenlo.

Resultado esperado

Informe y presentación exponiendo:

1. Entrega de un informe y/o presentación (.PDF/.PPTX) con un resumen de lo realizado. Esto permitirá evaluar el resultado sin necesariamente abrir ningún entorno de base de datos.
2. Se deben incluir como anexos todos los scripts desarrollados, los DER y estructuras correspondientes.
3. Entregar como .ZIP la base resultante con todos los componentes (.db, .sql, etc. y los tableros) para verificación de autoría si fuera necesario.
4. Entregar el tablero desarrollado (por ejemplo, Tablero.PBIX).
5. En la presentación en clase deberán ejecutar los tableros desarrollados.
6. Salvo el informe/presentación que debe ser publicado en el aula virtual, los demás objetos pueden ser publicados en un drive con libre acceso.

Rúbrica

#	Atributo	Valores
1	Grupo	• "Cantidad Optima" - #Grupo
2	Demora	• Por cada semana de demora en la entrega, a partir de la mañana siguiente: -1
3	Presentación e Informe	Evaluación general de: claridad, completión, corrección, síntesis y consistencia. • No realizó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buena presentación, clara, completa, correcta: +2 • Supera lo esperado: +3
4	Adquisición	Evaluación general de: claridad, completión, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3
5	Modelado	Evaluación general de: claridad, completión, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3
6	Ingesta inicial	Evaluación general de: claridad, completión, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3
7	Actualización	Evaluación general de: claridad, completión, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3
8	Gestión de Calidad	Evaluación general de: claridad, completión, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3

#	Atributo	Valores
9	Gestión de Metadata	Evaluación general de: claridad, compleción, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3
10	Gestión de Memoria	Evaluación general de: claridad, compleción, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3
11	Gestión de Enriquecimiento	Evaluación general de: claridad, compleción, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3
12	Publicación y desarrollo de producto de datos.	Evaluación general de: claridad, compleción, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3
13	Publicación y desarrollo de aplicación de datos.	Evaluación general de: claridad, compleción, corrección, síntesis y consistencia. • Si no entregó o está muy incompleto, con faltas graves o no se puede ejecutar: 0 • Confuso, desprolijo, poco o muy detallado, incompleto: +1 • Buen informe, claro, completo, correcto: +2 • Supera lo esperado: +3
14	Extras	A criterio de los profesores. Se puede sumar 1 o 2 puntos si se considera que hay algún aspecto que amerite ser calificado positivamente y no está evaluado en los otros atributos.

Para la nota final:

- Se suman todos los puntos de cada grupo.
- Se divide la suma de los puntos de cada grupo por el máximo de puntos de todos los grupos, se multiplica por 10 y de su redondeo resulta la nota final.
- Si la nota final >= 6 aprueban.
- Si la nota final < 6 se devuelven para su corrección, pero comienza a descontar la penalización por Demora.