TPG01 – 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, desde la ingesta 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 según los grupos establecidos para la materia de no más de tres integrantes.
- La entrega de este TP consiste en publicar un documento resumiendo lo realizado en función de lo que se especifica más abajo, además de los componentes desarrollados.
- Además, cada grupo deberá exponer en clase una síntesis del trabajo realizado con una duración máxima de 10'.
- Las fechas de publicación y presentación serán indicadas en la plataforma
- **Incluyan en los archivos a entregar la lista de los integrantes.**
- La evaluación se realizará según la rúbrica descrita más abajo.
- Los integrantes de cada grupo obtendrán la misma calificación.
- Los docentes evalúan 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. No copien textos externos, si fuera necesario, citen la fuente.

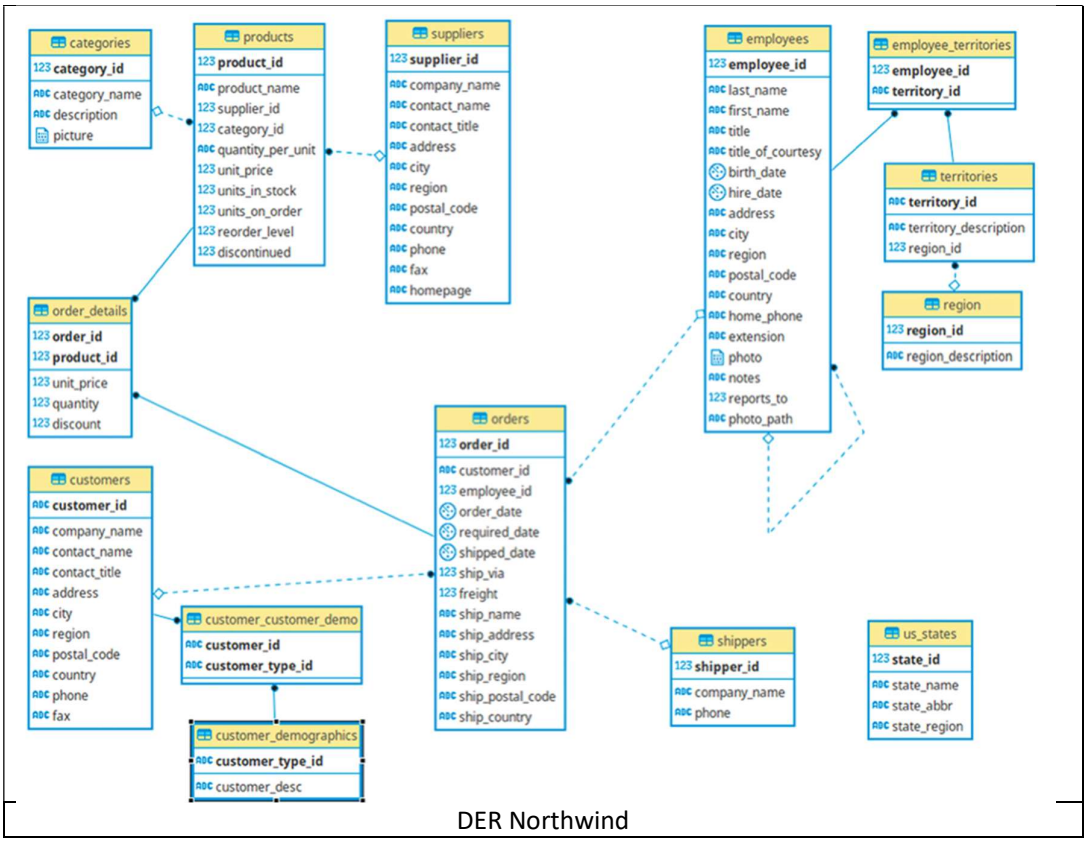
Contexto general

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.
2. Ingesta2: corresponde al mismo conjunto datos pero para una actualización posterior.

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 los datos persistidos en el DWA. Se deben incluir los controles de calidad necesarios, la memoria institucional, el enriquecimiento de los datos y gestionar la metadata.

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



DER Northwind

Descripción detallada

Desarrollar el flujo de datos de un DWA.

Los materiales se encuentran publicados en:

https://drive.google.com/drive/folders/1_515B1dDwWc1fc1zZZefGkb9i02YR-bG?usp=sharing

Se pide:

Adquisición de una entidad de datos

1) Analizar las tablas Ingesta1.
2) Comparar la estructura de tablas y el modelo de entidad relación. Adecuar si fuera necesario. Verificar y crear las FOREIGN-KEYS necesarias. Analizar integridad referencial.
3) Analizar la tabla de países.
4) Crear un área temporal y persistir todas entidades tal cual se encuentran.
5) Crear el soporte para la Metadata y utilizarlo para describir las entidades.

Crear al DWA

6) Definir y crear el modelo 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.
8) Registrar el diseño en la Metadata.

Carga inicial del DWA

9) Procesar **Ingesta1**:
   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.
   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: *outliers*, datos faltantes, valores que no respetan los formatos, etc.
   c) Ingestar los datos de Ingesta1 en el DWA definido. Las datos se deben insertar desde los .CSV entregados. Actualizar todas las capas. Siempre y cuando se superen los umbrales de calidad.
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) Se debe considerar además la capa de Memoria para persistir la historia de los campos que han sido modificados.
   d) Se debe considerar además actualizar la capa de Enriquecimiento para persistir los datos derivados que se vean afectados.
   e) Desarrollar y ejecutar los scripts correspondientes para actualizar el DWA con los nuevos datos.
   f) Actualizar el DQM si fuera necesario.
   g) Actualizar la Metadata si fuera necesario.
11) Publicación:
   a) Publicar un producto de datos resultante del DWA para un caso de negocio particular y un período dado si corresponde. Considerar altas, bajas y modificaciones. Tener en cuenta el orden de prevalencia para las actualizaciones.
   b) Desarrollar y ejecutar el script correspondiente. Dejar huella en el DQM.
12) Explotación
   a) Desarrollar y publicar un tablero para la visualización del producto de datos desarrollado. Dejar huella en el DQM.
   b) Desarrollar y publicar un tablero de visualización que permita navegar por los datos persistidos en el DWA. Dejar huella en el DQM.

Recomendaciones:

13) 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) TMP_ para temporales
   b) DWA_ para el Datawarehouse
   c) DQM_ para el Data Quality Mart
   d) DWM_ para la memoria
   e) MET_ para la metadata
   f) DP_ para los productos de datos
14) En https://en.wikiversity.org/wiki/Database_Examples/Northwind/SQLite tienen algunas ayudas para crear las tablas.
15) En todo control 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.
16) 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.
17) No es necesario pasar al DWA todos los atributos de las entidades originales, decidan cuáles son importantes y justifiquen.
18) Sean prolijos y explícitos al codificar los scripts y documenten en el mismo fuente.

19) 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 SQL estándar.
20) Se prefiere un trabajo simple, completo y bien hecho.

Resultado esperado:

Informe y presentación exponiendo:

1. Entrega de un informe y/o presentación con un resumen de lo realizado.
2. Se deben incluir como anexos todos los scripts desarrollados y los DER correspondientes.
3. Entrega como .ZIP de la base resultante con todos los componente (.db, .sql, etc. y los tableros) para verificación de autoría si fuera necesario.
4. Se recomienda usar SQLite pero no es obligatorio, pueden usar cualquier base SQL.

Rúbrica

Atributo	**Valores**
#Grupo	• 3 - #Grupo
Demora	• Por cada semana de demora en la entrega, a partir de la mañana siguiente: -1
Presentación e Informe	Evaluación general de: claridad, compleción, 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
Adquisición	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
Modelado	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
Ingesta inicial	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

Atributo	Valores
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
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
Gestión de Metadata	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
Gestión de memoria	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
Publicación y desarrollo de tableros	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
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.