**Logotipo, nombre de la empresa

El contenido generado por IA puede ser incorrecto.**

**Facultad de Ingeniería**

**Maestría en Ciencia de Datos – 2024/2025**

**Introducción a *Data Warehousing***

**Trabajo Práctico Final: Flujo de Trabajo**

**Integrantes:**

* Cancelas, Martín.
* Nicolau, Jorge.

Contenido

[Objetivo 3](#_Toc198886881)

[Aclaraciones 3](#_Toc198886882)

[Contexto general 3](#_Toc198886883)

[Descripción detallada 4](#_Toc198886884)

[Informe 8](#_Toc198886885)

[Adquisición 8](#_Toc198886886)

[Ingeniería 10](#_Toc198886887)

[Publicación 11](#_Toc198886888)

[Referencias 12](#_Toc198886889)

# Objetivo

Desarrollar todas las capas de datos y ejecutar los procesos correspondientes del flujo *end-to-end* en un DWA (*Data Warehouse* Analítico), 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. Los grupos de más **de tres integrantes** serán penalizados.
* La entrega de este TP consiste en **publicar un documento** resumiendo lo realizado según se especifica más abajo, además de los componentes desarrollados.
* Cada grupo deberá **exponer en clase una síntesis del trabajo** realizado con una duración máxima de 10’. 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 posgrado, 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 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. 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.

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 la gestión de la *metadata*.

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

Diagrama

El contenido generado por IA puede ser incorrecto.

# 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**

1. Analizar las tablas (.CSV) **Ingesta1**.
2. Comparar la estructura de tablas y el modelo de entidad relación. Adecuar si fuera necesario. Definir y crear las FOREIGN-KEYS necesarias para verificar la integridad referencial.
3. Considerar también la tabla de países (World-Data-2023) y vincularla con las tablas que correspondan.
4. Crear un área temporal y persistir todas entidades tal cual se encuentran en los .CSV
5. Crear el soporte para la Metadata y utilizarlo para describir las entidades.

**Ingeniería**

1. Definir y crear el modelo del DWA (Modelo Dimensional) y documentarlo en la Metadata. Debe incluir una capa de Memoria y una de Enriquecimiento (datos derivados).
2. 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.
3. Registrar el diseño en la Metadata.
4. Realizar la carga inicial del DWA con los datos que se seleccionen de las tablas recibidas y procesadas.
   1. 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.
   2. 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.
   3. Ingestar los datos de Ingesta1 en el DWA definido. Las datos se deben insertar desde las tablas temporales creadas. Actualizar todas las capas. Siempre y cuando se superen los umbrales de calidad.
5. Actualización:
   1. Persistir en área temporal las tablas entregadas como Ingesta2.
   2. Repetir los pasos definidos para Ingesta1 que sean adecuados para Ingesta2.
   3. Considerar altas, bajas y modificaciones. Tener en cuenta el orden de prevalencia para las actualizaciones.
   4. 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.
   5. Se debe considerar además la capa de Memoria para persistir la historia de los campos que han sido modificados.
   6. Se debe considerar además actualizar la capa de Enriquecimiento para persistir los datos derivados que se vean afectados.
   7. Desarrollar y ejecutar los scripts correspondientes para actualizar el DWA con los nuevos datos.
   8. Actualizar el DQM si fuera necesario.
   9. Actualizar la Metadata si fuera necesario.

**Publicación**

1. Publicar un producto de datos resultante del DWA para un caso de negocio particular y un período dado si corresponde.
   1. Desarrollar y ejecutar los scripts necesarios.
   2. Dejar huella en el DQM.
   3. Dejar huella en la Metadata de ser necesario.
2. Explotación
   1. Desarrollar y publicar un tablero para la visualización del producto de datos desarrollado. Dejar huella en el DQM y en Metadata de ser necesario.
   2. 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.

**Recomendaciones**

1. 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:
   1. TMP\_ para temporales
   2. DWA\_ para el Datawarehouse
   3. DQM\_ para el Data Quality Mart
   4. DWM\_ para la memoria
   5. MET\_ para la metadata
   6. DP\_ para los productos de datos
2. En https://en.wikiversity.org/wiki/Database\_Examples/Northwind/SQLite tienen algunas ayudas para crear las tablas.
3. 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.
4. 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.
5. No es necesario pasar al DWA todos los atributos de las entidades originales, decidan cuáles son importantes y justifiquen.
6. Sean prolijos y explícitos al codificar los scripts y documenten en el mismo fuente.
7. 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.
8. Se prefiere un trabajo simple, completo y bien hecho.
9. Lo que no esté especificado y sea necesario para el trabajo, decídanlo y justifíquenlo.
10. Se recomienda usar SQLite pero no es obligatorio, pueden usar cualquier base SQL. SI usan SQLite se recomienda utilizar también SQLiteStudio.
11. Para construir tableros se puede utilizar Power-BI Desktop u otros que conozcan (particularmente si quieren verlo en IOS).

**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 que debe ser publicado en el aula virtual, los demás objetos pueden ser publicados en un drive con libre acceso.

# Informe

El presente documento tiene por finalidad dar cuenta de lo realizado a lo largo de la obtención, análisis, tratamiento y exposición de los datos correspondientes al paquete “Ingesta1” – “Ingesta2”. Los mencionados archivos corresponden a un modelo de negocio de compra y venta de artículos comestibles y distintas bebidas.

Para el presente se han realizado distintos archivos individuales con *scripts* específicos, así como también dos *pipelines,* que permiten la ejecución del ciclo completo, tanto para la carga inicial de “Ingesta1”, así como la actualización que se recibe de “Ingesta2”. Esto permite, o bien ejecutar todo el proceso, o bien realizar un paso a paso de todas las etapas de este.

Se considera de importancia destacar que, en todos los archivos individuales, se ha incorporado la documentación correspondiente sobre su funcionalidad, requisitos y uso de cada uno de ellos. Además, se pone en conocimiento que HABLAR DE LOS LENGUAJES, INTERPRETES, ETC, QUE USAMOS????.

Respuesta (no se si hay que agregar algo más en las generalidades)

Finalizando, el flujo del ciclo de trabajo se desarrolla de la siguiente forma:

Por último, el ciclo presentado, así como los comentarios que se describen posteriormente, son válidos de igual forma tanto para la carga de los datos iniciales – “Ingesta1” –, así como para la actualización recibida posteriormente – “Ingesta2” –.

## Adquisición

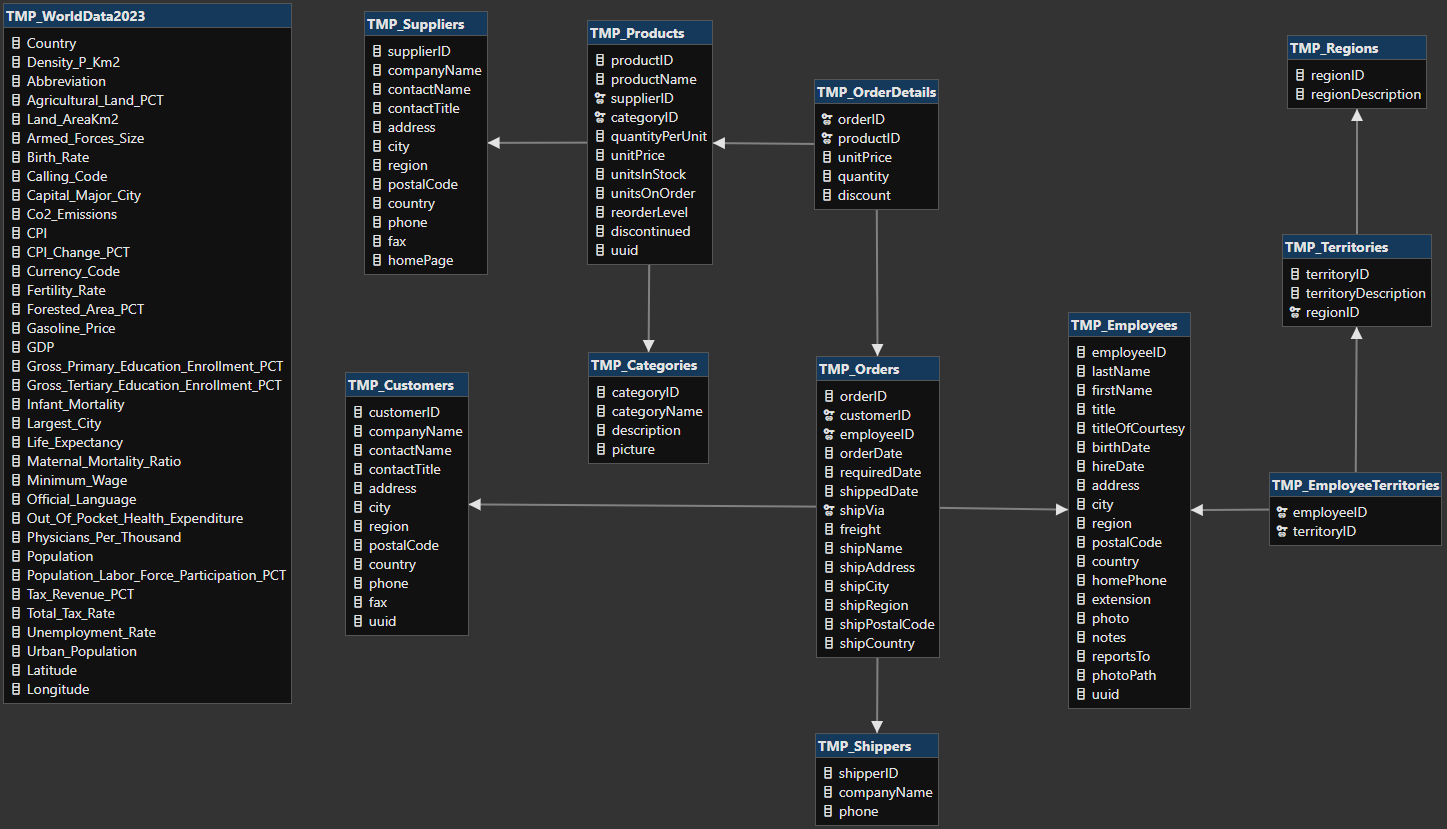
En el inicio del ciclo de este proceso, se crean las distintas tablas temporales, una por cada archivo .csv con datos de origen, determinando tanto las claves primarias, así como las foráneas donde correspondiera. Asimismo, se crean las tablas soporte para la *metadata* y se deja registro allí de este paso.

Sobre las tablas de “Ingesta1”, es posible observar que, la tabla *World*-*Data*-2023 cuenta con una estructura no adecuada en su cabecera, por lo que se ha considerado dentro del *script* de carga de datos en las tablas temporales la corrección de este formato, con el objetivo de que se obtengan todas las columnas, en principio, de forma correcta. El resto de las tablas se pueden ingresar al área temporal de forma satisfactoria, habiendo creado correctamente las tablas en la base de datos a utilizar.

Una vez creadas las tablas en el área temporal y determinadas las claves en estas, es posible visualizar, en “Figura 1”, el diagrama de entidad-relación:

**Figura 1**

*Modelo entidad-relación del área temporal (TMP)*

**

*Nota.* Es posible encontrar el modelo anteriormente presentado en https://github.com/georgsmeinung/dwa-lite/blob/main/dashboards/TMP\_ERD.html

Habiéndose construido el área temporal, creado las tablas y sus relaciones a través de claves, y alimentado este ambiente con los datos recibidos de los archivos .csv, se crea un área de trabajo *Stage* (STG) para realizar todos los ajustes que fueran necesarios antes de construir el ambiente y el modelo entidad-relación de *Data Warehousing* Analítico (DWA). En el mencionado ambiente, se verificaron y modificaron aquellos nombres de países que no eran coincidentes entre las tablas de clientes, proveedores y *WorldData*, para que se sirvan de claves a la hora de la creación del modelo de datos del DWA. Además, debido a que existen clientes que provienen del país Irlanda, y este no se encuentra en la tabla de países, se ha incorporado, así como también tres datos relevantes: PBI, precio del combustible y cantidad de habitantes, los cuales serán utilizados, finalmente, en la publicación de productos. Asimismo, al observar el precio del combustible del país Venezuela, surge la inconsistencia de que este es igual a cero, por lo que se decidió reemplazar este valor por un precio razonable obtenido de acuerdo con el litro de combustible excediendo el consumo del cupo mensual que se tiene en ese país. Por último, se realiza un control exclusivo del formato de datos en los descuentos aplicados en el detalle de las ordenes, habiéndose obtenido una respuesta favorable por parte de estos.

Una vez finalizada la etapa de adquisición de datos, se deja huella en la *metadata* y valida la calidad de estos, para dar paso a la etapa de ingeniería.

1. Analizar las tablas (.CSV) **Ingesta1**.
2. Comparar la estructura de tablas y el modelo de entidad relación. Adecuar si fuera necesario. Definir y crear las FOREIGN-KEYS necesarias para verificar la integridad referencial.
3. Considerar también la tabla de países (World-Data-2023) y vincularla con las tablas que correspondan.
4. Crear un área temporal y persistir todas entidades tal cual se encuentran en los .CSV
5. Crear el soporte para la Metadata y utilizarlo para describir las entidades.

## Ingeniería

Respuesta

1. Definir y crear el modelo del DWA (Modelo Dimensional) y documentarlo en la Metadata. Debe incluir una capa de Memoria y una de Enriquecimiento (datos derivados).
2. 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.
3. Registrar el diseño en la Metadata.
4. Realizar la carga inicial del DWA con los datos que se seleccionen de las tablas recibidas y procesadas.
   1. 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.
   2. 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.
   3. Ingestar los datos de Ingesta1 en el DWA definido. Las datos se deben insertar desde las tablas temporales creadas. Actualizar todas las capas. Siempre y cuando se superen los umbrales de calidad.
5. Actualización:
   1. Persistir en área temporal las tablas entregadas como Ingesta2.
   2. Repetir los pasos definidos para Ingesta1 que sean adecuados para Ingesta2.
   3. Considerar altas, bajas y modificaciones. Tener en cuenta el orden de prevalencia para las actualizaciones.
   4. 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.
   5. Se debe considerar además la capa de Memoria para persistir la historia de los campos que han sido modificados.
   6. Se debe considerar además actualizar la capa de Enriquecimiento para persistir los datos derivados que se vean afectados.
   7. Desarrollar y ejecutar los scripts correspondientes para actualizar el DWA con los nuevos datos.
   8. Actualizar el DQM si fuera necesario.
   9. Actualizar la Metadata si fuera necesario.

## Publicación

Respuesta

1. Publicar un producto de datos resultante del DWA para un caso de negocio particular y un período dado si corresponde.
   1. Desarrollar y ejecutar los scripts necesarios.
   2. Dejar huella en el DQM.
   3. Dejar huella en la Metadata de ser necesario.
2. Explotación
   1. Desarrollar y publicar un tablero para la visualización del producto de datos desarrollado. Dejar huella en el DQM y en Metadata de ser necesario.
   2. 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.

# Referencias

Argadoña, A. (24 de septiembre de 2020). *La responsabilidad social como actitud*. https://blog.iese.edu/antonioargandona/2020/09/24/la-responsabilidad-social-como-actitud/?utm\_source=feedburner&utm\_medium=email&utm\_campaign=Feed%3A+iese%2Fantonioargandona+%28Blog+de+Antonio+Argando%C3%B1a%29

Bloomberg Línea (5 de enero de 2024). *Precio de la gasolina en Latinoamérica: los países con el litro más caro en 2024*. https://www.bloomberglinea.com/2024/01/05/precio-de-la-gasolina-en-latinoamerica-los-paises-con-el-litro-mas-caro-en-2024/

CIA.gov. (21 de mayo de 2025). *The World Factbook. Explore All Countries - Ireland.* https://www.cia.gov/the-world-factbook/countries/ireland/

Trading Economics. (sf). *Irlanda - Precios de la gasolina*. https://es.tradingeconomics.com/ireland/gasoline-prices

TEDx Talks. (7 de mayo de 2018). *Qué es la ética de la tecnología y por qué debería importarte | Gemma Galdón-Clavell | TEDxZaragoza*[Archivo de Video]. Youtube. https://www.youtube.com/watch?v=XJDBN2j9rr4&ab\_channel=TEDxTalks

TEDx Talks. (15 de agosto de 2018). *Industria 4.0: una revolución para las personas | Beatriz Gonzalez | TEDxUDeusto* [Archivo de Video]. Youtube. https://www.youtube.com/watch?v=a0Ycxn-bZak&ab\_channel=TEDxTalks

Universidad Austral. (2025). *EMPRESA Y SUS RESPONSABILIDADES – 2da parte* [Archivo PDF]. https://campusvirtual.austral.edu.ar/pluginfile.php/3284956/mod\_resource/content/5/Clase%2005\_12%20-%20RSE%20MCD.pdf

Valenzia Plaza. (3 de noviembre de 2021). *Adela Cortina: "A una empresa que no sea ética se le van a cerrar puertas”.* https://valenciaplaza.com/adela-cortina-a-una-empresa-que-no-sea-etica-se-le-van-a-cerrar-puertas

Worldcob. (6 de mayo de 2014). *2. ¿Qué es la RSE?*[Archivo de Video]. Youtube. https://www.youtube.com/watch?v=9j9gZTo4SO0