UNIVERSIDAD
AUSTRAL

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

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

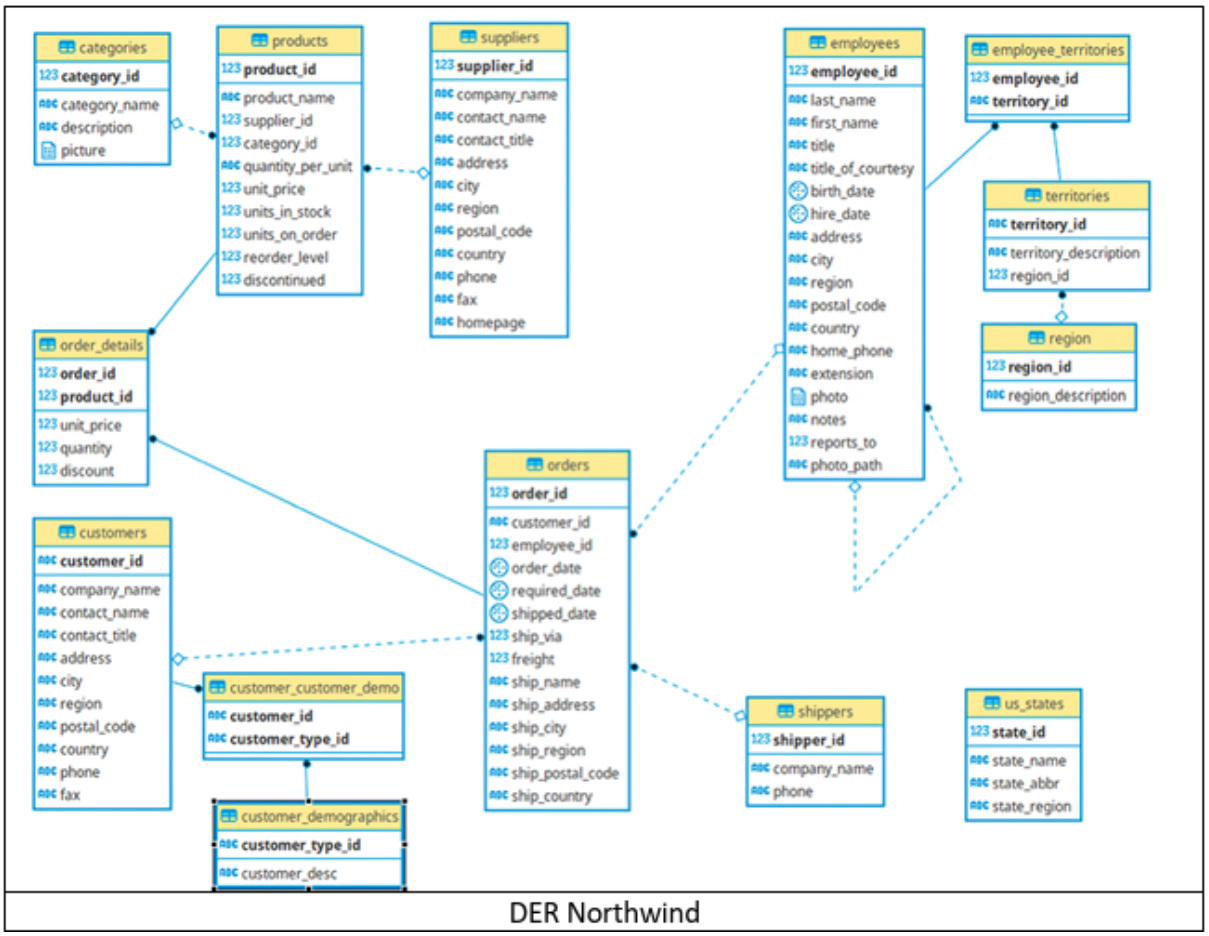
Se publicarán dos conjuntos de *datasets* provenientes de una base de datos transaccional y de otras fuentes secundarias.

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.



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

6) 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).
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.
9) 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.
    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 las tablas temporales creadas. 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) 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) Se debe considerar además la capa de Memoria para persistir la historia de los campos que han sido modificados.
f) Se debe considerar además actualizar la capa de Enriquecimiento para persistir los datos derivados que se vean afectados.
g) Desarrollar y ejecutar los scripts correspondientes para actualizar el DWA con los nuevos datos.
h) Actualizar el DQM si fuera necesario.
i) Actualizar la Metadata si fuera necesario.

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.

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.
21) Lo que no esté especificado y sea necesario para el trabajo, decídanlo y justifíquenlo.
22) Se recomienda usar SQLite pero no es obligatorio, pueden usar cualquier base SQL. SI usan SQLite se recomienda utilizar también SQLiteStudio.
23) 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 se utilizaron para construir el *pipeline* sólo Python y SQL guardando los resultas en un base SQLite.

Finalizando, el flujo del ciclo de trabajo se desarrolla en la “Figura 1”.

Figura 1

Pipeline de trabajo y detalle de actividades

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” –. Como aclaración adicional, si se quisiera volver a ejecutar todo el ciclo inicial (no una actualización incremental), se debiera ejecutar, en primer lugar, el *script* “98*_drop_all_tables*.py”.

Todo el código fuente así como los tableros y la base resultante de procesar ingesta1 e ingesta2 están disponibles en https://github.com/georgsmeinung/dwa-lite.

Adquisición

Para iniciar el ciclo completo de carga inicial, es necesario ejecutar el *script “*00a*_run_start_pipeline*.py”, y, para realizar la actualización incremental, es necesario ejecutar el *script “*00b*_run_update_pipeline*.py”.

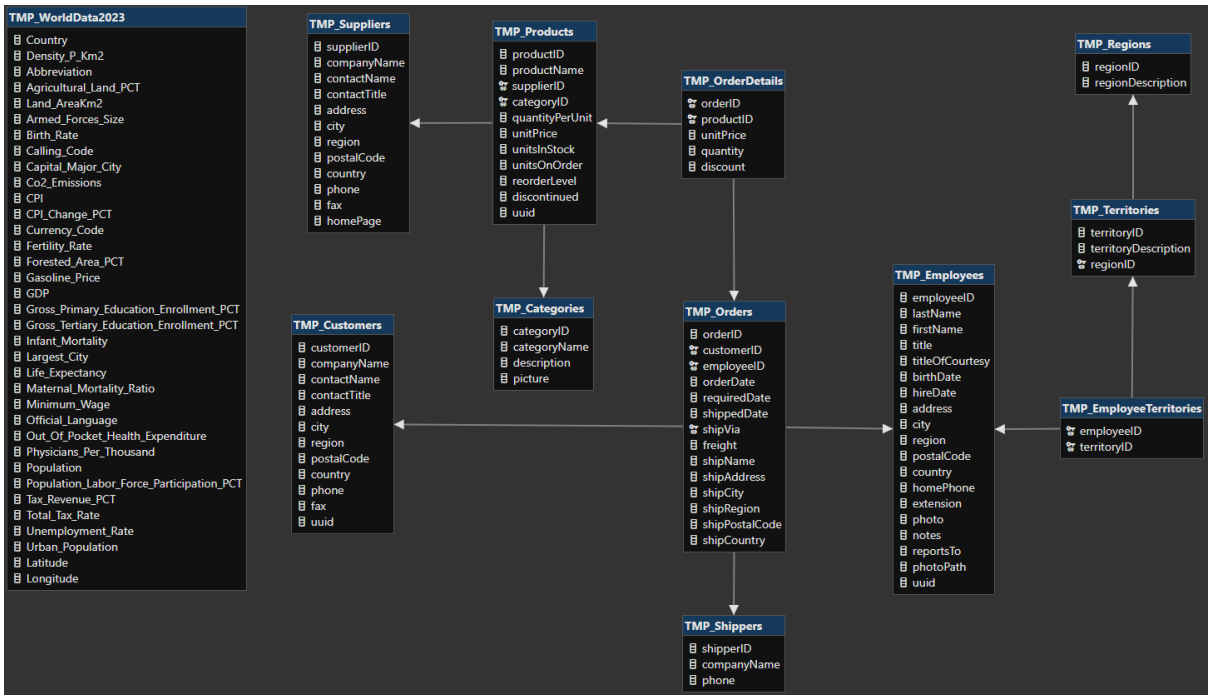
En el inicio del ciclo de este proceso, se crean las distintas tablas temporales, una por cada archivo .csv con datos de origen, determinando tanto las claves primarias, así como las foráneas donde correspondiera. Asimismo, se crean las tablas soporte para la *metadata* y se deja registro allí de este 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 2”, el diagrama de entidad-relación.

Figura 2

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.

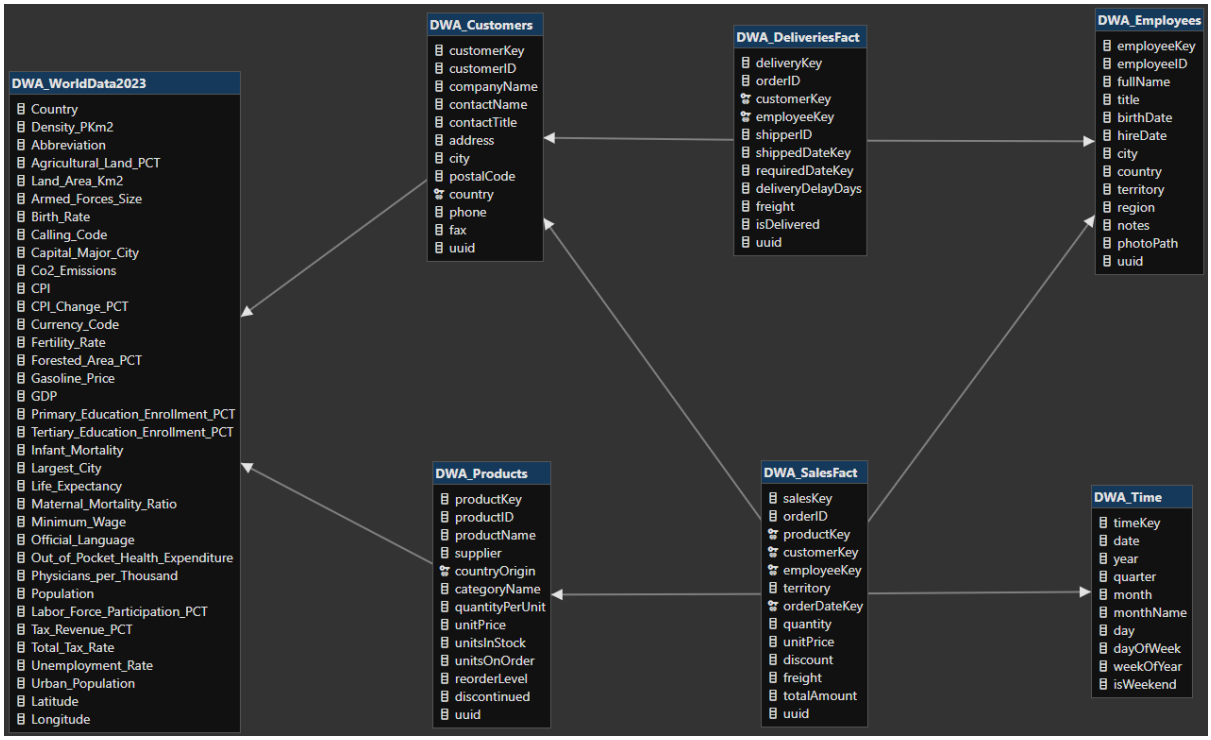
Ingeniería

En primer lugar, durante esta etapa, se crea la tabla DWA_*Time*, la cual permite asociar las claves de fecha a las tablas que necesitaran este dato durante el modelo del DWA. A partir de la tabla de órdenes, del ambiente anterior, es posible obtener las fechas de interés para la creación de esta tabla, así como también la suma de una proyección para los próximos dieciocho meses. Asimismo, como fuera mencionado inicialmente, se crearon el resto de las tablas del *Data Warehouse* Analítico. Por lo tanto, el siguiente paso consiste en el traslado de los datos, ya modificados (si correspondiera), del ámbito de pruebas a esta sección.

Dentro del DWA, al contar con todas sus tablas alimentadas, es posible obtener el modelo relacional de este, el cual puede apreciarse en la “Figura 3”.

Figura 3

Modelo entidad-relación del Data Warehouse Analítico (DWA)



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

Posteriormente, a partir de la tabla de fechas extendidas, es posible asignar las claves de estas fechas a las tablas de hechos, denominadas “ventas” y “entregas”. De esta actividad surgen nuevas registraciones dentro de la validación de calidad y la *metadata* del proyecto.

Una vez realizadas las asignaciones, se actualizan las tablas de memoria del modelo, que servirán para el registro de actualizaciones que puedan generarse a partir de los datos contenidos en “Ingesta2”. Sobre este proceso existe un nuevo registro en la *metadata*.

Se han realizado controles de calidad a las distintas tablas, detectando principalmente valores nulos y duplicados. Se ha decidido considerar umbrales relativos de aceptación, permitiendo mayor flexibilidad en tablas de menor importancia, como puede ser STG_*WorldData*2023, la cual, a pesar de contener una cantidad importante de valores nulos, no afectan al ciclo de trabajo ni a los productos finales o a los tableros publicados. En el caso de las tablas de hechos, no se encontraron valores que representen, al menos, el 5% del total con estas observaciones.

Al ejecutar el *pipeline* de actualización de información, se toman las tablas de “Ingesta2”, se procesan, de igual forma que en el ciclo de iniciación, y se cargan en las tablas TMP, con su correspondiente copia en STG. Se revisa que las nuevas fechas se incluyan en la tabla DWA_*Time* y, en caso de que así no fuera, se incorporan las novedades a esta. Posteriormente, debido a que no hay datos que pudieran generar conflictos, se traslada al DWA la actualización. Como se realizó en el primer *pipeline*, se asignan las claves de fechas a las tablas de hechos, para incorporar este dato a las novedades, y, esto deja registro en la *metadata* y en los controles de calidad. En este momento se actualizan, en las tablas de memoria, las siguientes: DWA_*Customers*, DWA_*Employees*, DWA_*Products*, DWA_*WorldData*2023, DWA_*SalesFact* y DWA_*DeliveriesFact*. Nuevamente, al igual que en el ciclo de carga inicial, se deja registro de calidad y en la *metadata* del proyecto.

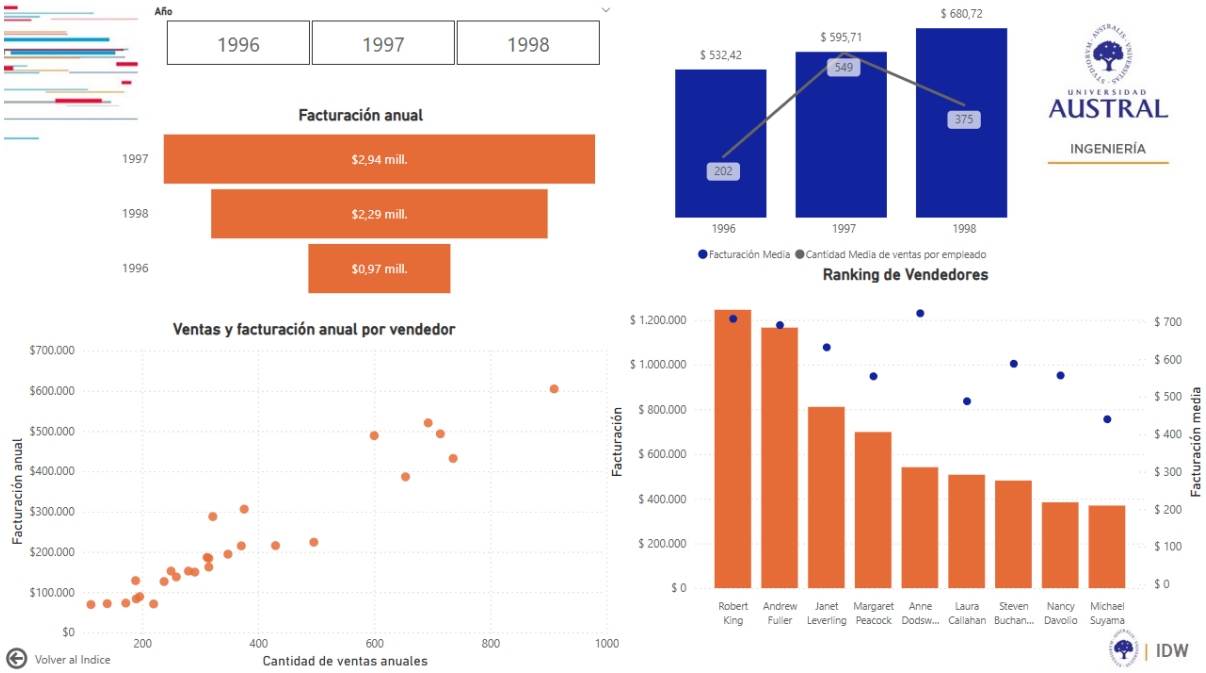
Publicación

En la etapa de adquisición se realizó la creación de todas las tablas, a excepción de aquellas que se encontraron en el ambiente de STG y la tabla DWA_*Time*. Entre las tablas que se crearon, se incluyeron las de los cuatro productos que se publicaron: Ventas consolidadas por producto por mes, Ranking de clientes por facturación, Desempeño de los empleados por año y Productos con devoluciones o cancelaciones.

A partir de combinaciones de las tablas del ambiente de DWA, se alimentaron los cuatro productos, obteniendo nuevos datos a partir de las tablas base anteriores. Luego de ello, nuevamente, se actualizaron el *Data Quality Mart* y la *metadata*.

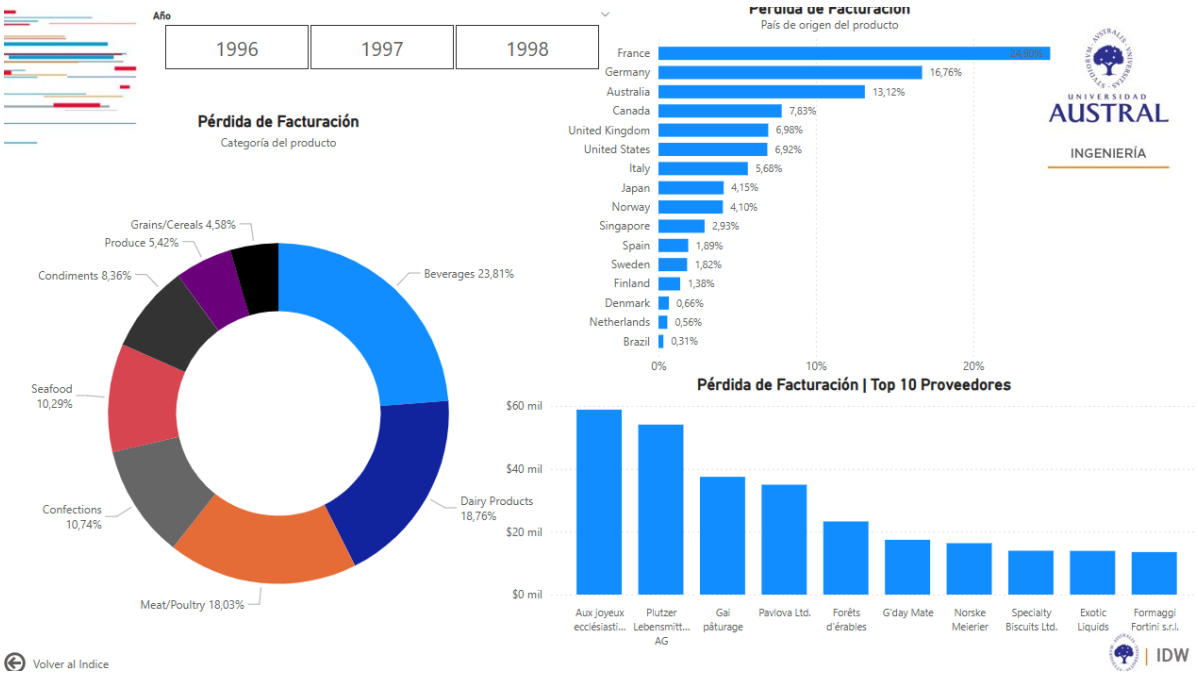
Para cada uno de los productos de datos creados, se le generó un tablero dentro del archivo “Tablero.pbix”, que es parte del presente proyecto. Este cuenta con un índice de los productos y el tablero particular de cada uno de ellos. Sobre cada uno se pueden obtener diversas conclusiones:

- Desempeño de los empleados:



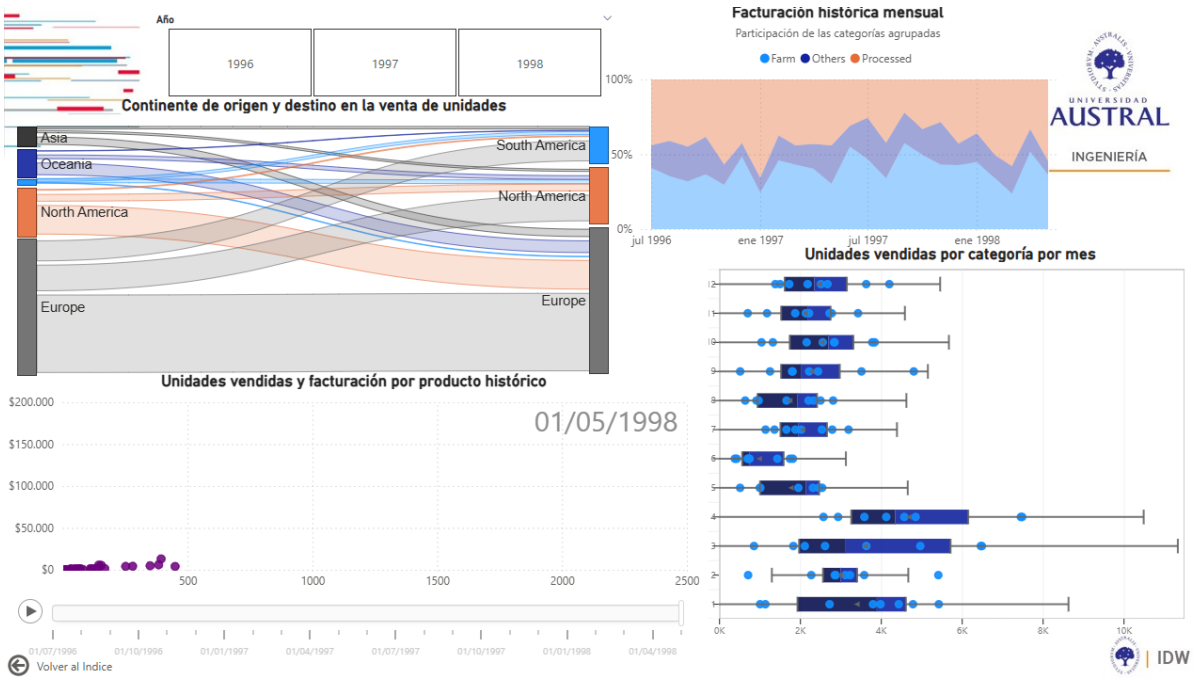
  - Hay una relación positiva y cercana a lineal entre la cantidad de ventas realizadas y la facturación generada.
  - Durante el año 1997 la facturación fue mayor que la de 1996 y 1998, pero, durante el último año, con una menor cantidad de ventas se ha logrado una facturación media mayor que en los anteriores.
  - Se destaca a la vendedora Anne Dodsworth, la cual, no siendo aquella con la mayor facturación, tiene una mayor facturación media por venta, es decir que factura más que sus compañeros con menor cantidad de ventas.

- Productos con devoluciones:



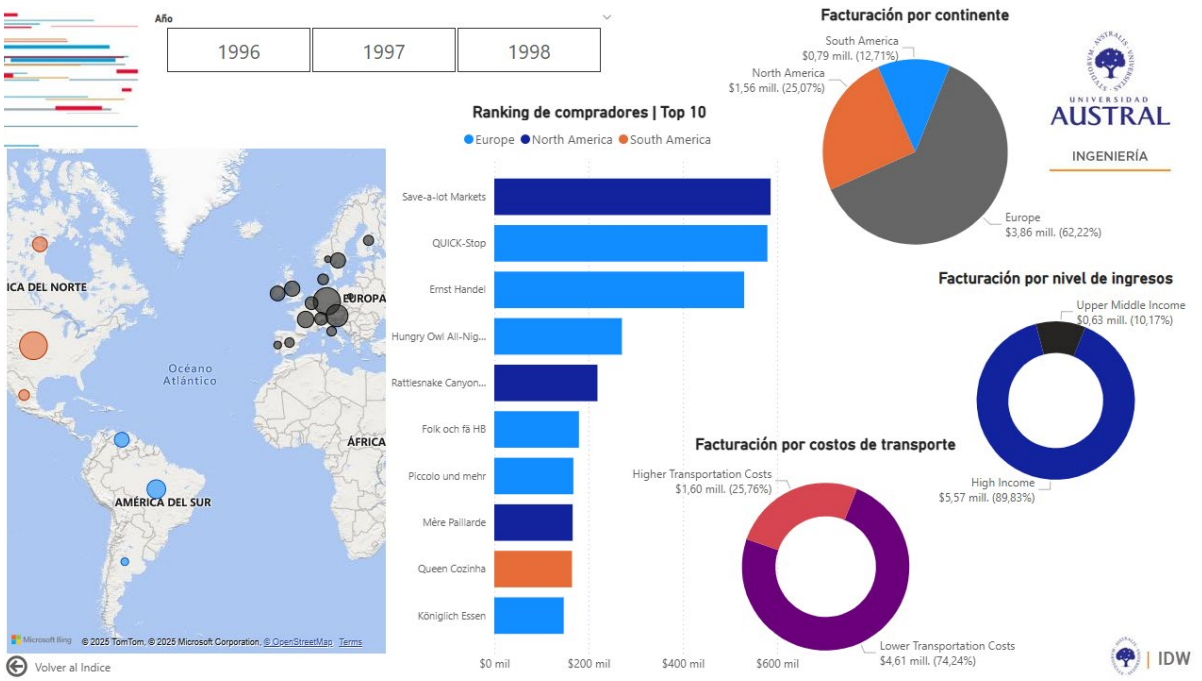
- Las bebidas y productos lácteos representan más del 40% de la pérdida de facturación de la empresa en todo el período.
- Los países de origen de los productos que mayor pérdida de facturación generan son, mayormente, europeos. Francia lidera esa posición, seguido de productos de origen alemán y, en tercer lugar, aquellos provenientes de Australia.
- Entre los proveedores de esos productos, aquellos provenientes de Francia, se encuentran en primer y tercer lugar, y, de productos alemanes, se ocupa el segundo lugar de pérdida de facturación por proveedor.

- Ventas por producto por mes:



- Agrupando las categorías de productos, puede observarse cierta estacionalidad donde aumenta relativamente la facturación de productos de granja (carne, aves, productos lácteos, granos, cereales) y disminuye, relativamente, la venta de procesados, lo que incluye las bebidas, dulces y manufacturados. La facturación por venta de productos marítimos y condimentos se mantiene casi constante a lo largo del período.
- Los productos provienen principalmente del continente europeo, seguido por los norteamericanos, producidos en Oceanía, Asia y, en último lugar, en Sudamérica. Los destinos de venta son, principalmente, Europa, donde se vende más del 60% de las unidades totales que vende esta empresa, seguido por América del Norte y del Sur, donde se venden, aproximadamente, el 25% y 15% de los restantes productos.
- Existe gran dispersión en las cantidades vendidas por categoría durante el primer cuatrimestre. Puede observarse una caída de las unidades vendidas a fin de este y hasta la mitad del año, con estabilización entre unas 3.000 unidades, en promedio, vendidas por mes.

- Ranking de clientes:



- En concordancia con lo anteriormente expuesto, la mayor facturación proviene de ventas al continente europeo, seguido por los países del norte de América, completando el total facturado por el sur de este mismo continente.
- No obstante lo expuesto, el principal comprador corresponde a una empresa de Estados Unidos. Si bien, luego, la cantidad de empresas europeas aumentan la participación de este continente en la facturación total, es destacable que el principal cliente no corresponde a esos países.
- A partir de los datos de PBI y costo de combustible en los distintos países, es posible observar que, como podría suponerse a priori, la mayor facturación proviene de países de ingresos altos por sobre aquellos de medios-altos, y de países con menores costos de transporte internos que de aquellos que, relativamente, presentan costos de combustible más elevados. Sobre estos conceptos, es importante destacar que, el criterio utilizado en sus casos fue:
    - Para ingresos: se tomó como medida principal el Producto Bruto Interno (PBI) per cápita de cada uno de los países y, con datos de Banco Mundial para 2023 – en coincidencia con la tabla de *World Data* –, se los clasificó con su mismo criterio, es decir, por un umbral de PBI per cápita.
    - Para costos de transporte interno: de los países participantes del producto, aquellos que tuvieran un precio de combustible mayor a

la mediana se consideraron costos elevados y, aquellos por debajo de la mediana, con menores costos.

Por último, se generó el “Tablero_DQM.pbix”, para obtener detalles acerca de las estadísticas y observaciones, el cual se acompaña a este proyecto.

Referencias

Banco Mundial Blogs (1 de julio de 2024). *Clasificación de países del Banco Mundial por nivel de ingreso correspondiente a 2024-25*. https://blogs.worldbank.org/es/opendata/clasificacion-de-paises-del-banco-mundial-por-nivel-de-ingreso-2024-25#:~:text=La%20clasificaci%C3%B3n%20de%20ingresos%20del,disponible%20de%20la%20capacidad%20econ%C3%B3mica.

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/

Cancelas, M. y Nicolau, J. (22 de abril de 2025). *dwa-lite*[Repositorio] (Github). https://github.com/georgsmeinung/dwa-lite

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