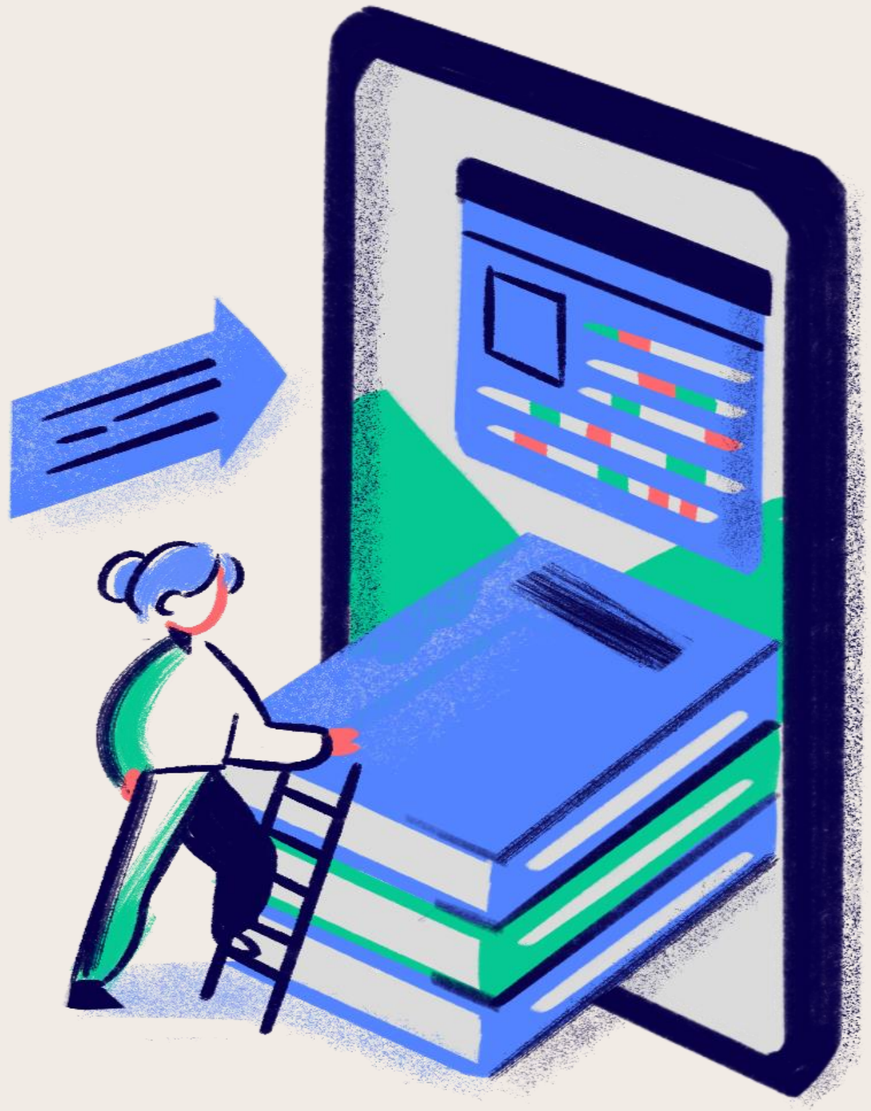
TEMA II

PROCESO ETL

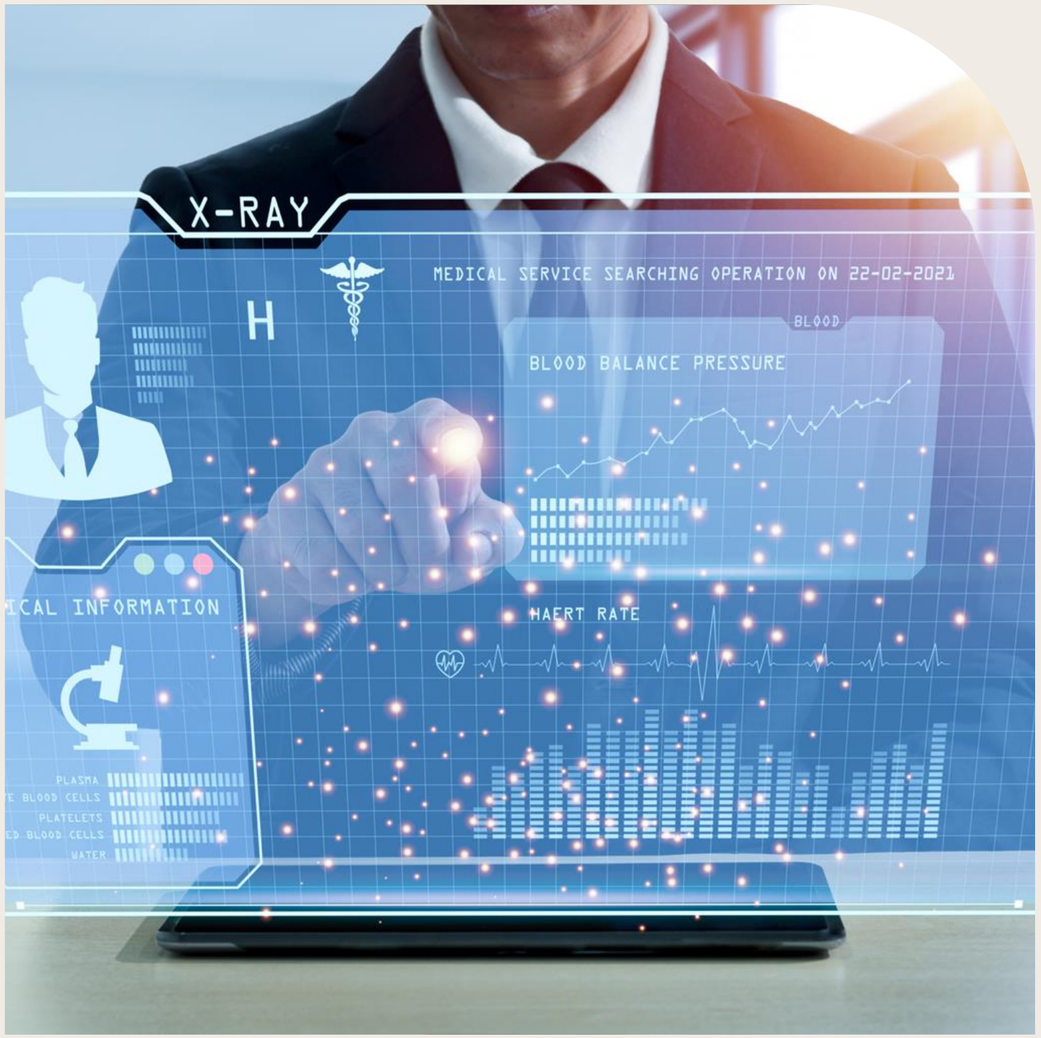
DISEÑO DE BASE DE DATOS II



ING. CARINA COZZOLINO

AGENDA

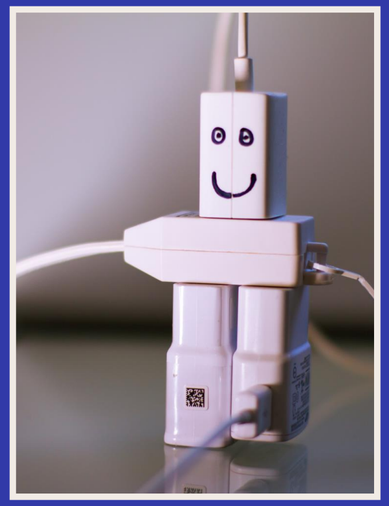
- Dimensiones Especiales
  - Calendario
  - Role Playing
  - Junk
  - Snwoflake
- Tipos de Tablas de Hecho
- Proceso de Transformación de Datos
  - Extracción
  - Transformación
  - Carga
  - Componentes
- Ejemplo Práctico
- Práctica



DIMENSIONES ESPECIALES

CALENDARIO

ROLE PLAYING



JUNK

SNOWFLAKE

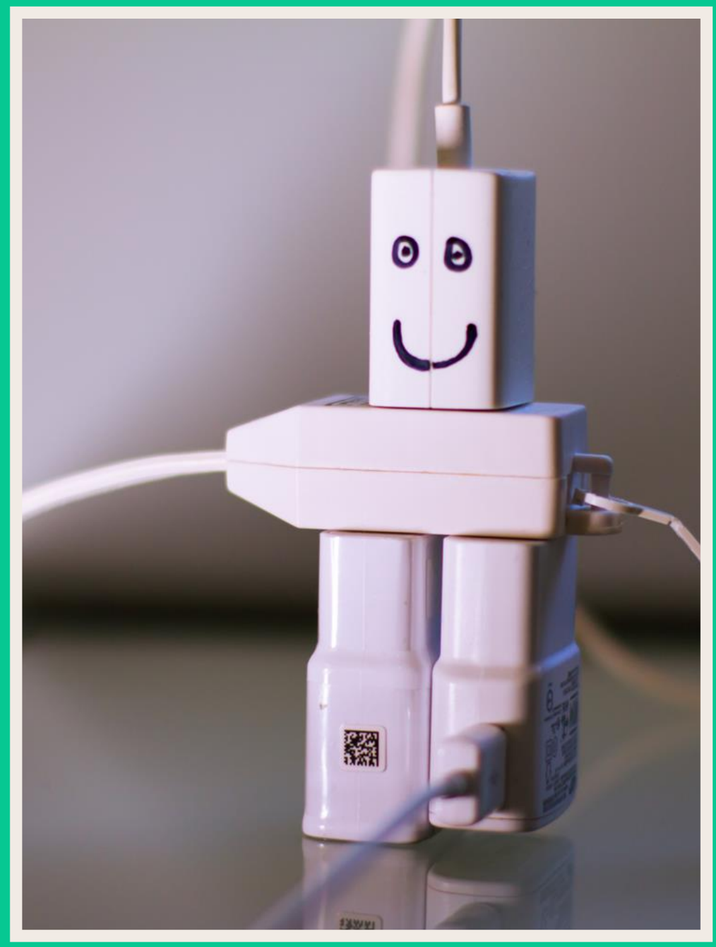
DIMENSIÓN CALENDARIO

- **Existe en todos los Data Warehouse.**
- Las Tablas de Hechos se suelen analizar a través del tiempo.
- **Evita el cálculo** de este tipo de datos
- Se pueden agregar tantos datos como sean necesarios (número de semana, periodo fiscal, etc.).
- La **Llave Primaria** es un número **entero** con el formato adecuado para representar la fecha correspondiente.

ID	Fecha	Día	Mes	Año	Semestre
01022001	01/02/2001	1	2	2001	1
23042005	23/04/2005	23	4	2005	1
16092010	16/09/2010	16	9	2010	2

DIMENSIÓN ROLE PLAYING

- Una dimensión que puede ser referenciada varias veces por la tabla de Hechos, pero de forma independiente.
- Se crean distintas vistas de la misma tabla (cada vista toma un rol diferente)



ID	KEY	Pedido	Fecha-Pedido	Fecha-Envio	Fecha-Pago
18	1	8237834	01022001	01042005	01052005
19	2	8384589	16052007	16062007	16072007
20	3	3494503	31122015	20022016	20042016

ID	Fecha	Día	Mes	Año	Semestre
01022001	1/2/2001	1	2	2001	1
16052007	16/5/2007	16	5	2007	1

ID	Fecha	Día	Mes	Año	Semestre
01042005	1/4/2005	1	4	2005	1
16062007	16/6/2007	16	6	2007	1

ID	Fecha	Día	Mes	Año	Semestre
01052005	1/5/2005	1	5	2005	1
16072007	16/7/2007	16	7	2007	2

DIMENSIÓN JUNK

- Se utilizan para evitar dimensiones con una sola columna, o con muy pocas filas.
- NO ES UN PRODUCTO CARTESIANO de todos los atributos
- Son combinaciones que se pueden aplicar en el mundo real.

ID	Key	Pedido	Fecha-Pedido	Estado-Confirmado
18	1	8237834	01032005	1
19	2	8384589	16052007	1
20	3	3494503	31120015	3

ID	Estado	Confirmado
1	Tránsito	Confirmado
2	Tránsito	Pendiente
3	Enviado	Confirmado

DIMENSIÓN SNOWFLAKE

- Dentro de un Data Warehouse pueden darse relaciones jerárquicas.
- Cuando una dimensión es normalizada, se obtienen dos tablas relacionadas por una llave foránea.
- Según el caso puede ser una buena alternativa, pero deben evitarse ya que impacta en el rendimiento de las consultas.

ID	Nombre	Edad	Categoría	Mercado
1	Juan Perez	Adulto Joven	Web	1
2	José Gomez	Adulto Joven	Partner	2
3	Martín Gacía	Adulto Mayor	Partner	3

ID	Mercado
1	LATAM
2	USA
3	EU

TABLAS DE HECHO - TIPO

TRANSACCIONAL

- Cada event es una fila
- Las fias solo exisnten si una medida ocurrió
- Permiten mayor nivel de filtrado

SNAPSHOT PERIÓDICO

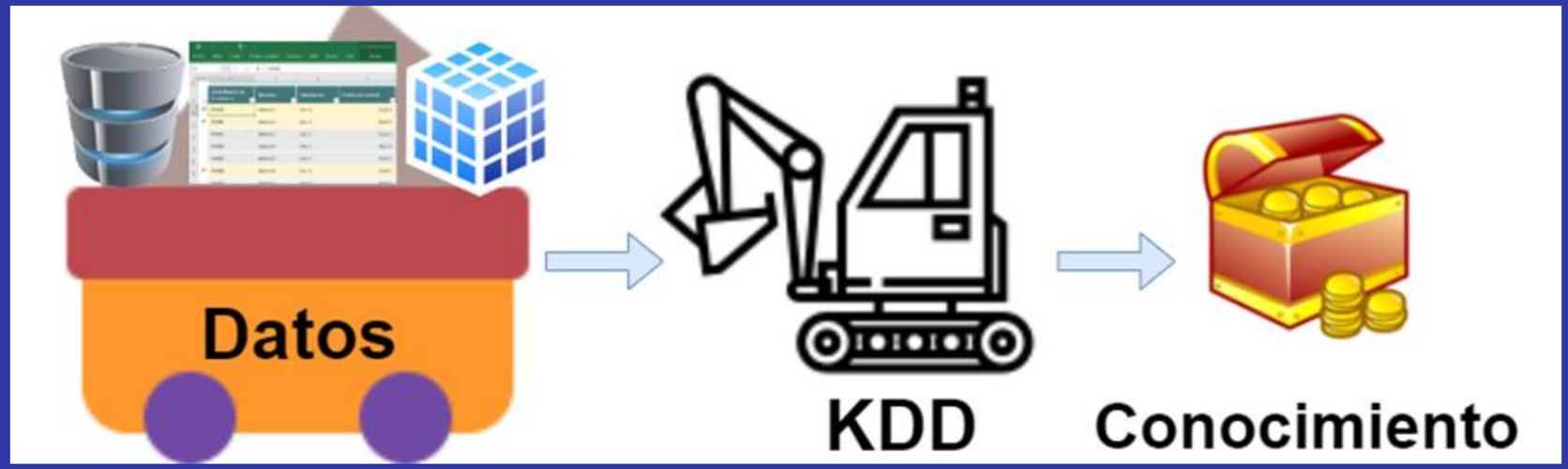
- Sumariza varias medidas basdo en un periodo de tiempo.
- Cad fila muestra EL GRANO en el tiempo, no la transacción.
- Una fila es insertada incluso si no hubo transacciones (agrega fila en 0).

- **FACTLES**
- Sirven para analizar lo que NO PASÓ
- Las filas no tienen medidas
- Reistra un conjunto de dimensiones en un punto del tiempo.

PROCESO ETL

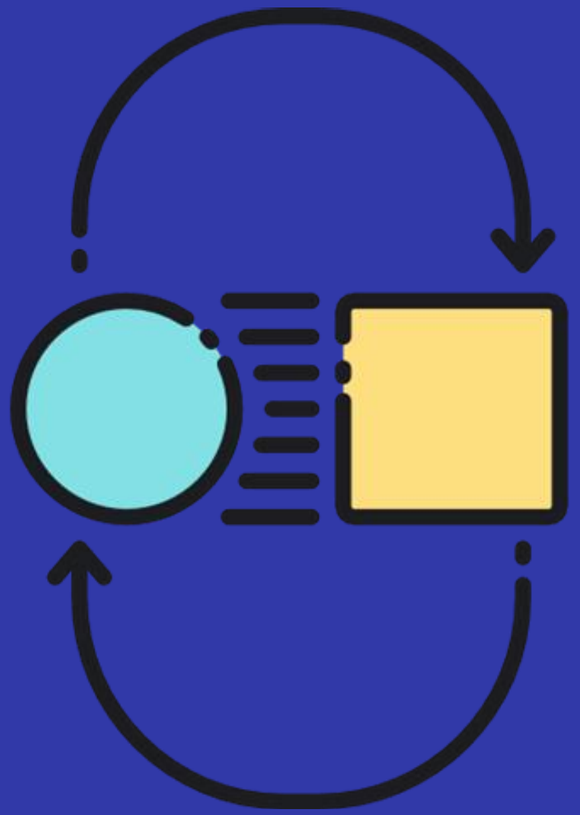
- Proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos, limpiarlos y cargarlos en un almacén de datos para apoyar un proceso de negocio.
- Se aplican criterios de calidad y consistencia para que los datos pudan unirse en un solo almacén.

EXTRACCIÓN DE DATOS



- Determina qué fuentes de datos se van a procesar.
- La velocidad y el orden de extracción de dicha información tienen gran impacto en todo el proceso de integración.
- Durante la extracción de los datos de la fuente original, eleproceso de ETL realiza un análisis y limpieza de todos los datos, que ayuda a diferenciarlos.
- El lenguaje de consulta estructurado (SQL) permite gestionar y extraer las partes de una base de datos en forma de informes.
- Se realizan variedad de acciones sobre las tablas y filas de datos específicas.

TRANSFORMACIÓN



- Se realiza la transformación de los datos, se corrigen y se resuelven todas las diferencias que puedan contener los datos para su mejor clasificación.
- Se lleva a cabo a través de un conjunto de reglas que proporcionan el orden y la claridad con los datos que van a ser integrados en la base de datos y que varían según los criterios de cada compañía.

Por medio de una validación, eliminación de duplicados, codificación y filtrado en el formato deseado, permite conocer cuáles datos tiene alguna diferencia para ver si se omiten o se hacen a un lado para un análisis más profundo.

CARGA

El proceso de Carga de Datos en una base de datos destino, que es un Almacén de Datos o Repositorio centralizado (en la nube o físicamente en una instalación).

COMPONENTES DEL PROCESO ETL

El proceso de ETL permite ahorrar tiempo en la extracción y preparación de datos para las empresas.

Los componentes de un proceso de ETL incluyen:

- COMPATIBILIDAD: al cargar nuevos datos, se actualizan los existentes según parámetros establecidos previamente.
- AUDITORÍA Y REGISTRO: es necesario un registro detallado de los datos para que garantice la precisión de los datos y eliminar los errores.
- MANEJO DE MÚLTIPLES FORMATOS: al extraer los datos de diversas fuentes, debe manejar una gran variedad de datos.
- TOLERACIA A FALLAS: deben recuperarse ante cualquier problema que ocurra en el proceso, y asegurar que se desplacen los datos sin dificultad.
- SOPORTE DE NOTIFICACIONES: debe incluir un sistema de notificaciones que dé aviso cuando se presenta algún problema.
- ACTUALIZACIONES: para la toma de decisiones en tiempo real, la actualización de datos debe ser fluída y óptima.
- ESCALABILIDAD: debe tenerse en cuenta la expansión de la empresa en cuanto a datos y procesos. Se debe prever almacenamiento y procesamiento.
- PRECISIÓN: los datos deben grantizar una óptima carga y un flujo preciso de información, que refleje la veracidad en cada etapa del porceso.

EJEMPLO

La situación actual de la empresa es que n cuenta con los siguientes reportes, que estarían siendo necesarios:

a. Clasificación de los productos (**familia**)

b. Distribución de los clientes por zona de venta (**región**, **provincia**, **comuna**, **sucursal**).

c. Tipo de clientes que están prefiriendo los productos (**edad**, **estado civil**).

d. Relación entre las ventas ($) por vendedor y la cantidad de horas de capacitación que reciben (**tipo de capacitación** y **horas de capacitación**).

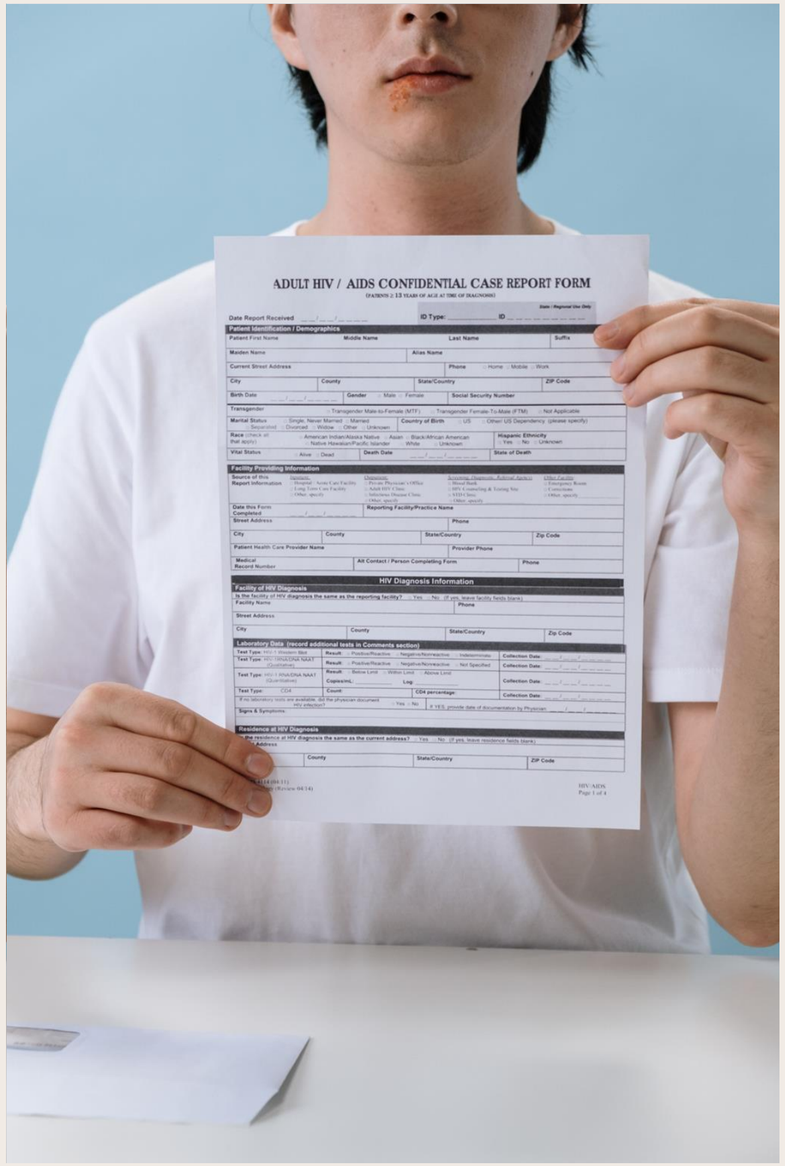
e. Ventas mensuales y anuales.

REQUERIMIENTOS

- Generar un modelo multidimensional del tipo estrella, a partir de los reportes requeridos por la empresa:
    a. Crear las tablas de Dimensiones:
        - Productos
        - Clientes
        - Sucursal
        - Vendedor
    b. Crear la tabla de Hechos o medidas correspondientes en el Data Warehouse con la siguiente información:
        - Cantidad de Ventas
        - Monto de Venta
        - Mes
        - Año
        - ID de cada Dimensión.

2.Implementar un ETL en Integration Services, traspasar los registro desde la Base de Datos Transaccional al Data Warehouse

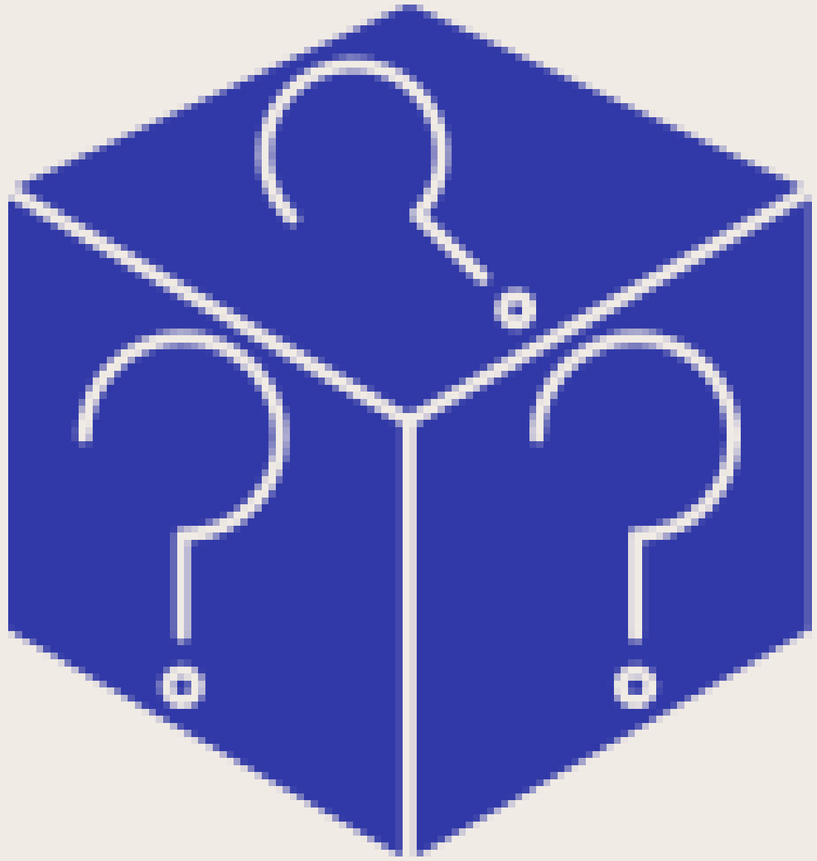
TRABAJO PRÁCTICO N°2



- Crear en SQL Server la Base de Datos del Data Warehouse, a partir del modelo multidimensional diseñado en el Trabajo Práctico N° 1
- Desarrollar al menos 10 requerimientos a responder con el Data Warehouse.
- Generar los ETLs necesarios para traspasar los datos desde la Base de Datos Operacional al Data Warehouse.

FECHA DE ENTREGA: 4 de Octubre

CONSULTAS



ING. CARINA COZZOLINO