

Agenda 06/08/2020

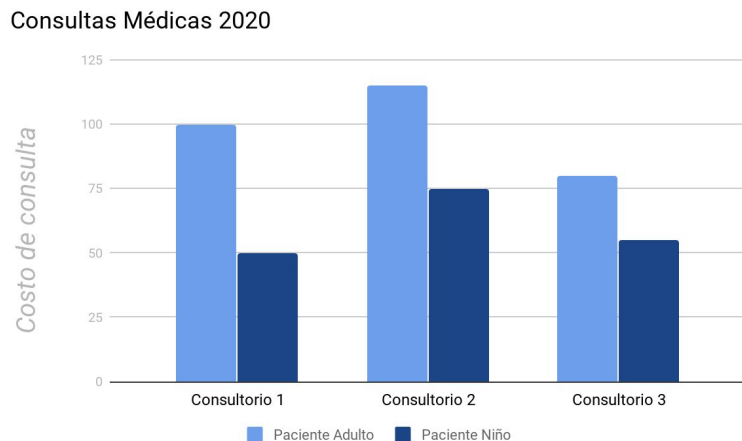
- Resolución Hoja de Trabajo 1.
- Clase 2.
 - Tablas de dimensión y hecho.
 - Datawarehouse
 - Proceso de ETL
- Tarea 1.

Resolución Hoja de Trabajo 1

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Hoja de Trabajo 1.

Identifique las dimensiones y los hechos de las siguientes gráficas.



Dimensiones:

- Consultorio
- Paciente

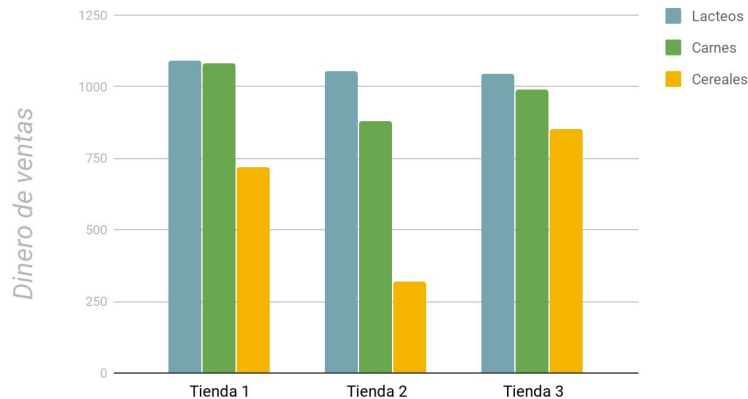
Hechos:

- Consulta o costo de consulta

Hoja de Trabajo 1.

Identifique las dimensiones y los hechos de las siguientes gráficas.

Ventas 2020



Dimensiones:

- Tienda
- Producto

Hechos:

- Ventas

Clase 2




Tablas de Dimensión

- Las tablas de dimensiones contienen atributos que describen las entidades de negocio.
- Una tabla de dimensión almacena información descriptiva sobre los valores almacenados en la tabla de hecho
- Cada tabla posee un identificador único(**llave subrogada**) que lo une a la tabla de hechos.

Tablas de Dimensión

CLIENTES
 id_Cliente
NombreCliente

PRODUCTOS
 id_Producto
Rubro
Tipo
NombreProducto


FECHAS
 id_Fecha
Año
Trimestre
Mes
Día

Tabla de Hechos

- Estas tablas contienen los hechos que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones.
- Son los indicadores del negocio, por ejemplo, ventas , pedidos, reclamos, compras, devoluciones, etc. Cada hecho representa una transacción o evento.
- Cada registro de esta tabla posee una clave primaria que está compuesta por las claves primarias(**llaves subrogadas**) de las tablas de dimensiones relacionadas a este.
- Es importante resaltar que la tabla de hechos idealmente debe almacenar solo valores numéricos.

Tabla de Hechos

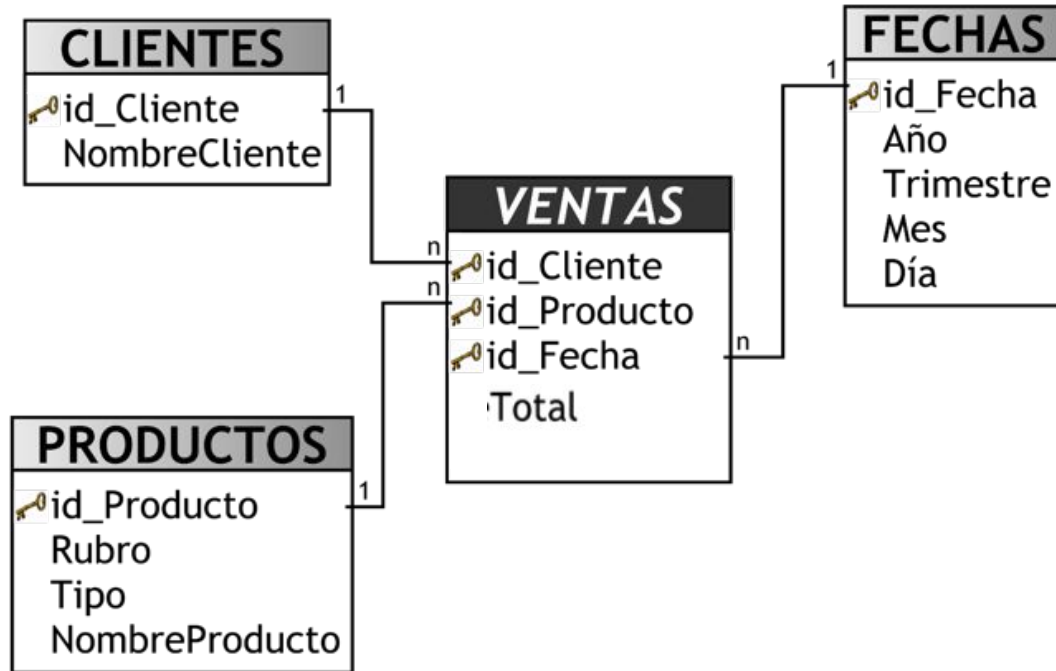
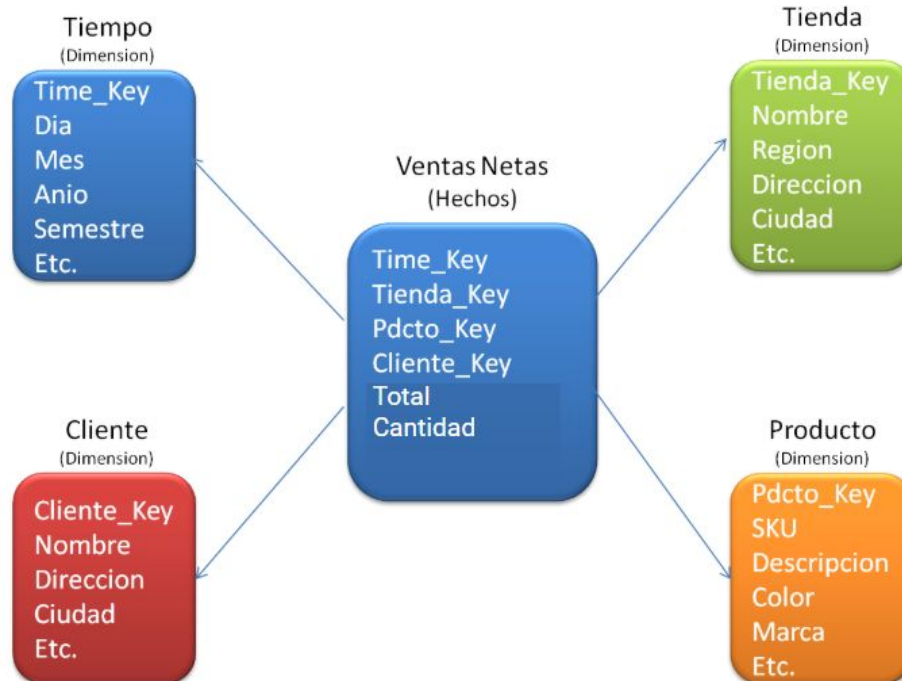


Tabla de Hechos



Datawarehouse

- Un datawarehouse es una base de datos corporativa o un almacén de datos que tiene como característica la integración y depuración de todos los datos que recogen los diversos sistemas de una empresa.

Datawarehouse

- Cuando se habla de querer implementar una solución fiable de BI(Business Intelligence) el **primer paso** es la creación de un Datawarehouse.

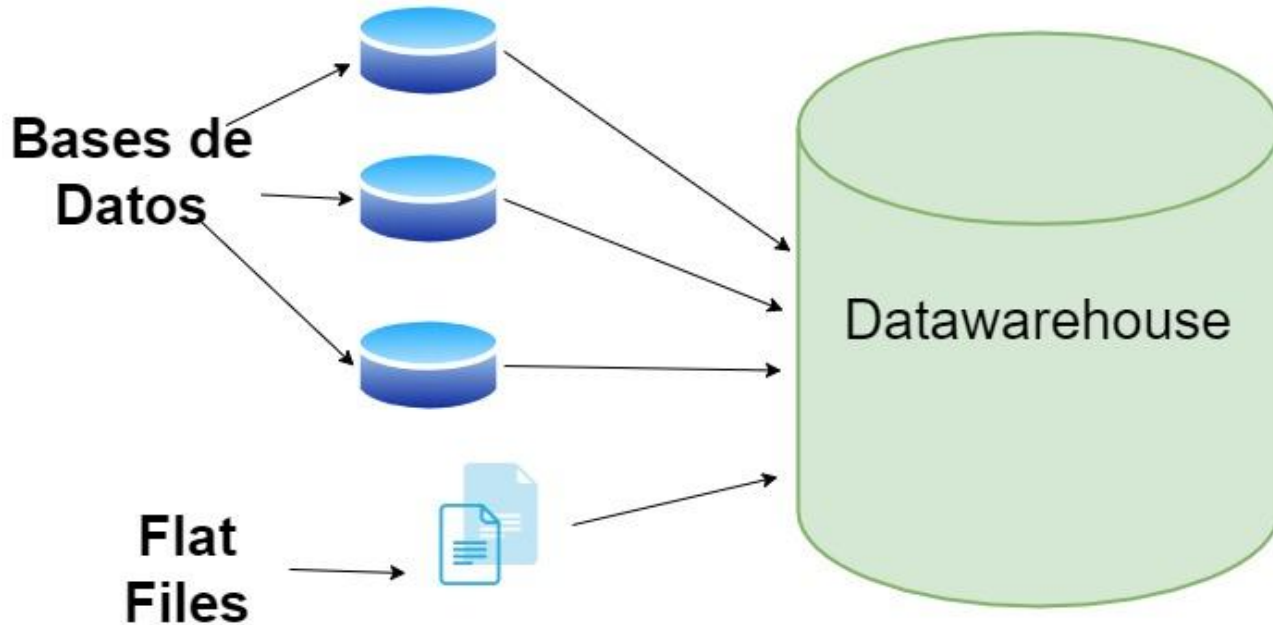
Datawarehouse

- La **función principal** de un datawarehouse es la de contener los datos necesarios o útiles para una organización o empresa y así poder utilizarlos en un futuro para extraer información ventajosa para la compañía y sus clientes.

Datawarehouse

- A diferencia de una base de datos que es un mero almacén para el ingreso de datos, un datawarehouse se encuentra especialmente estructurado para favorecer la comprensión y el análisis de los datos

Representación gráfica de un Datawarehouse.



¿Cuándo me interesa implementar un Datawarehouse?

- Si necesito integrar muchas fuentes diferentes de datos casi en tiempo real.
- Si tengo gran cantidad de datos históricos a tratar o debo mantener registros históricos, incluso si los sistemas de transacción de origen no lo hacen.
- Si necesito limpiar o mejorar la calidad de los datos para analizar.
- Si tengo riesgo de que los usuarios puedan provocar errores o pérdidas de datos durante sus consultas.

¿Cómo deben de almacenarse los datos en un Datawarehouse?

- De forma segura.
- De forma fiable.
- Fácil de recuperar
- Fácil de administrar

Ventajas y Desventajas del Datawarehouse

VENTAJAS

- Proporciona una comunicación fiable entre los departamentos.
- El acceso a la información es más rápido.
- Permite conocer en cualquier momento los buenos y malos resultados de la empresa.
- Inteligencia histórica.

DESVENTAJAS

- Requiere mucho mantenimiento transformación y limpieza.
- El coste es alto.
- El diseño es complejo.

PROCESO DE ETL

¿Qué significa ETL?

E	T	L
Extract Extracción	Transform Transformación	Load Carga

¿Qué es el proceso de ETL?

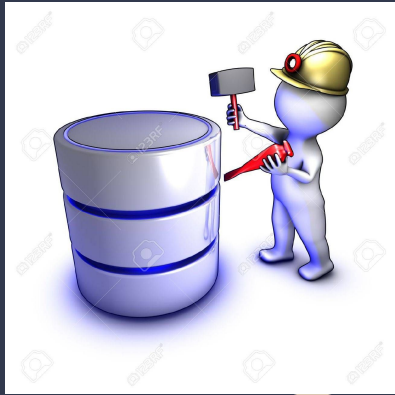
- Es un proceso mediante el cual nos permite mover datos desde múltiples fuentes (Excel, bases de datos, archivos, Internet) para integrarlos en un lugar, que se sugiere este sea un Datawarehouse.

Extract – Extracción



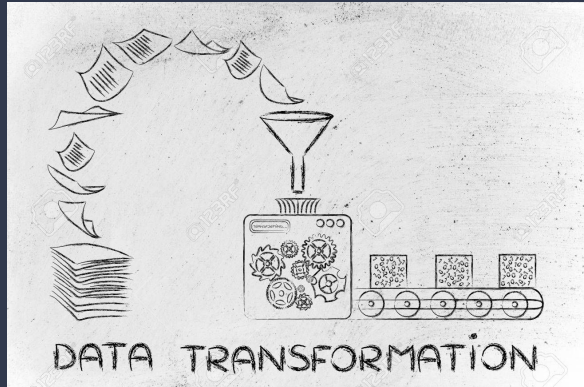
- Es la primera fase del proceso de ETL, en esta se obtiene la “materia prima” en este caso la data desde las distintas fuentes que se proporcionen.
- Esta data es con la que se trabajara en las siguientes dos fases.

Extract – Extracción



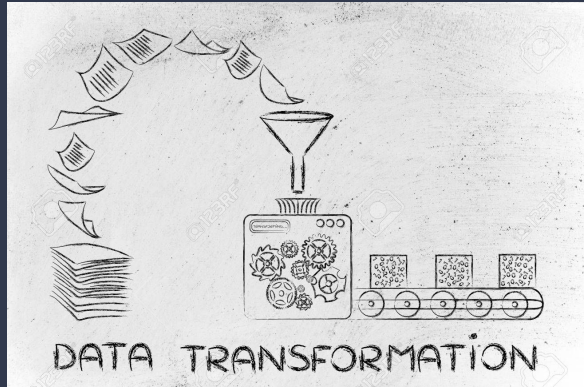
- El volumen de datos extraídos, así como el intervalo de tiempo entre extracciones, depende de las necesidades y requisitos del negocio.

Transform – Transformación



- Es la fase más crítica ya que es la que lleva más trabajo para realizar ya que la data que se trae de la **Fase 1** necesita ser limpiada, mapeada y transformada.
- Esta fase es clave ya que agrega valor y cambia los datos para que tengan sentido y puedan ser utilizados para generar informes.

Transform – Transformación



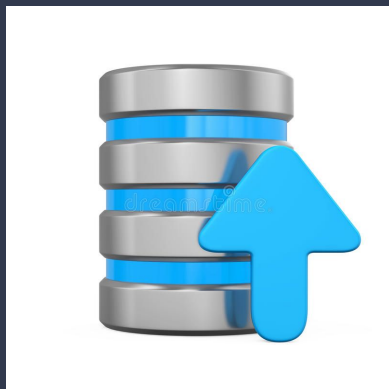
- Cuando se realiza la transformación se debe mantener la integridad de los datos al realizar operaciones como:
 - Validacion
 - Calculos
 - Filtrado
 - Remocion de duplicados.

Load – Carga



- En esta **fase 3** se llega al objetivo final que es la carga de datos en la base de datos del **Datawarehouse**.
- En caso de fallas, se deben contar con mecanismos de recuperación.

Load – Carga



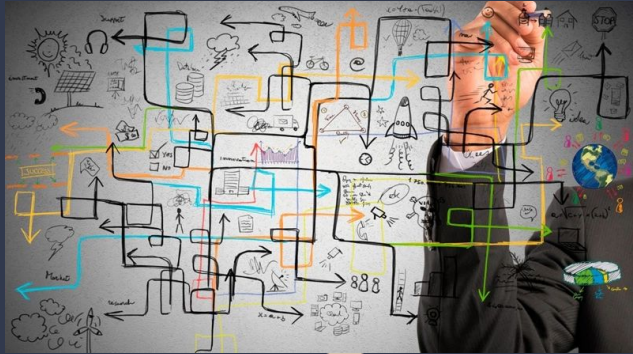
- La carga de datos debe ser de forma consistente en el Datawarehouse de destino.

Características del proceso ETL

Cada empresa llega a tener diferentes datos y necesidades distintas, pero hay características comunes en todo proceso de ETL:

- Complejidad
- Continuidad
- Criticidad

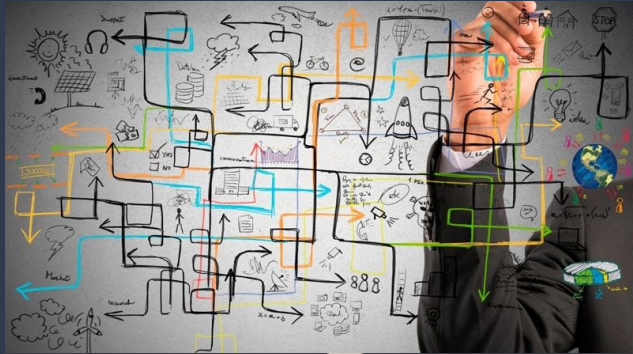
Complejidad



- Las empresas pueden contar con grandes cantidades de datos almacenados por años y generadas por distintos departamentos, repartidos en distintas fuentes como:

- Bases de Datos.
- Archivos de texto
- Flat files
- Excel
- CVS

Complejidad



- Extraer, realizar el tratamiento y consolidar toda esa información es una tarea **bastante compleja**.

Continuidad



- Para poder contar con análisis precisos, vamos a necesitar mantener el Datawarehouse constantemente actualizado ya que pueden agregarse nuevas fuentes o nuevos datos a las fuentes.

Continuidad



- Por esto, es importante que el proceso de ETL se realice cada cierto tiempo en intervalos regulares, para detectar dichos cambios, extraer los nuevos datos, transformarlos y cargarlos al Datawarehouse.

Criticidad



- Generalmente los datos que se poseen en las empresas no vienen por defecto en una forma en la cual se puedan usar para la resolución de problemas del negocio.

Criticidad



- Sin los procesos de ETL, las empresas pueden llegar a encontrarse con una cantidad de datos muy grande que **no se puede** llegar a utilizar.

Ventajas y Desventajas del proceso de ETL.

VENTAJAS

- Permite extraer y consolidar datos de múltiples fuentes.
- Permite adaptar e integrar nuevas fuentes de datos.
- Facilita el análisis y el reporte de datos de forma sencilla.

DESVENTAJAS

- Alto coste inicial.
- Se requiere un nivel avanzado de conocimientos para las herramientas.
- El mantenimiento tiene que ser constante.

Utilidades del proceso ETL

- Mover datos de una o múltiples fuentes.
- Formatear datos y realizar limpieza cuando esto sea necesario.
- Una vez alojados en el destino(Datawarehouse) se pueden analizar los datos según las necesidades de la empresa.

Desafíos del proceso ETL

- Procesamiento de datos en tiempo real.
- Aumentar la velocidad del procesamiento de datos.
- Integración de nuevas fuentes de datos.

Procesamiento de datos en tiempo real.



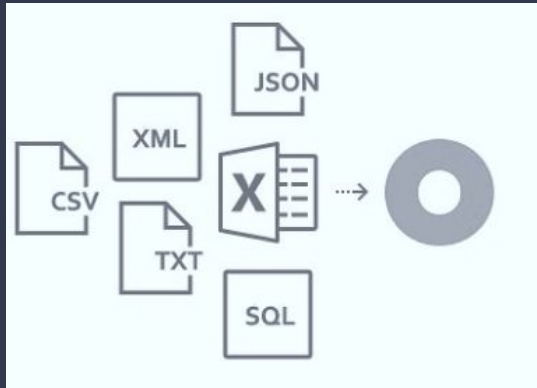
- Cada día se necesita mayor velocidad para la toma de decisiones, el proceso de ETL tiene que adecuarse para poder operar lo más cercano posible al tiempo real.

Aumentar la velocidad del procesamiento de datos.



- El aumento de cantidad de datos como de complejidad en los datos, puede llegar a dificultar la tarea de transformación.

Integración de nuevas fuentes de datos.



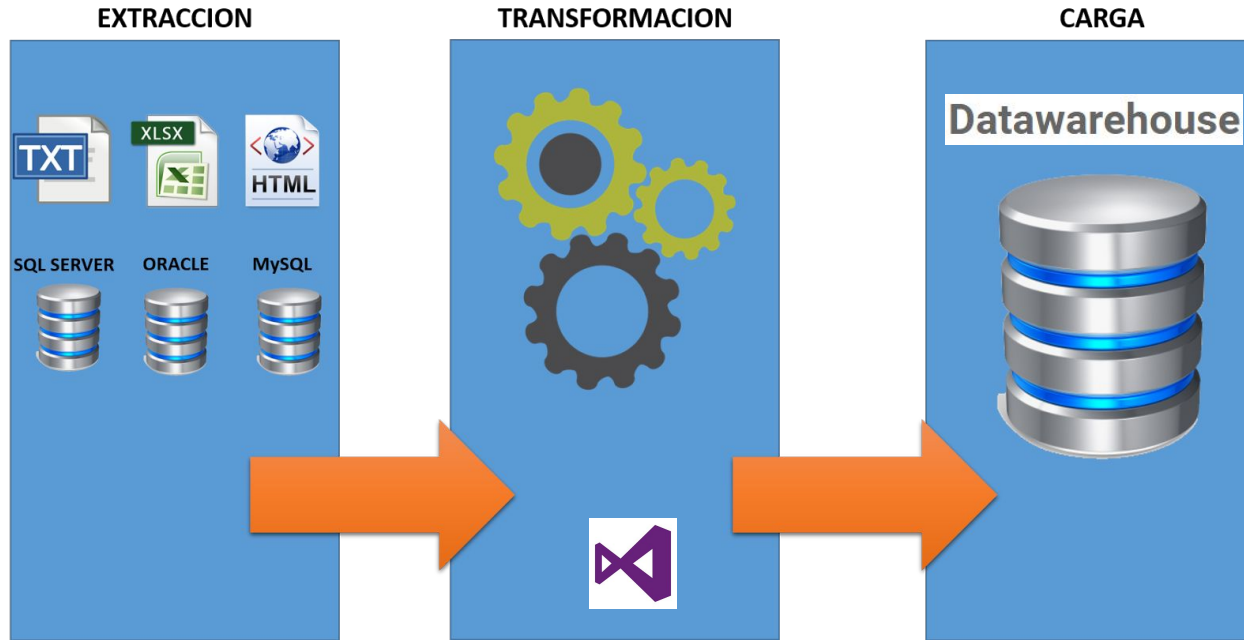
- El proceso de ETL necesita evolucionar para soportar nuevas fuentes de datos en cualquier momento.

Herramientas de ETL

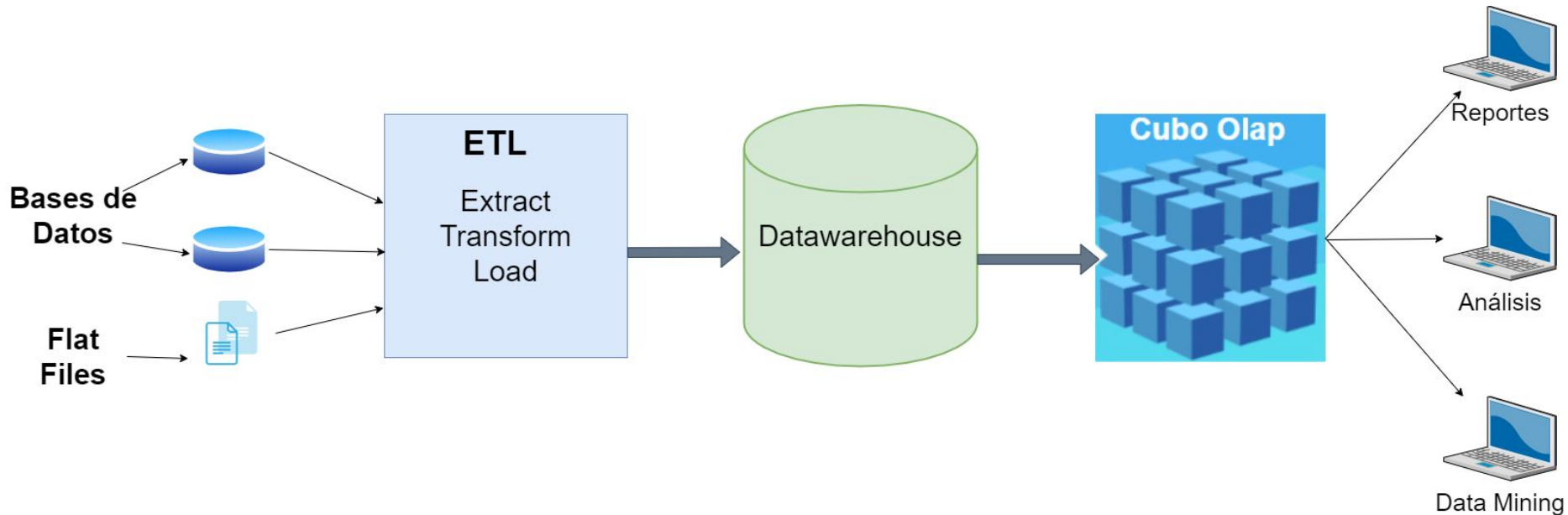
Hoy en día las principales herramientas para realizar ETL son:

- IBM InfoSphere DataStage
- Oracle Data Integrator.
- ***Microsoft SSIS.***
- Informatica PowerCenter.
- Pentaho Data Integration. (Open Source)

Representación gráfica del proceso de ETL



¿Qué hemos visto hasta ahora?



¿Dudas o Preguntas?

