ESCUELA DE
INGENIERÍA EN CIENCIAS Y SISTEMAS
FACULTAD DE INGENIERÍA
UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

ECYS | DTT
DESARROLLO DE TRANSFERENCIA TECNOLÓGICA

Día, Fecha	Lunes, 29/07/2024
Hora de inicio:	17:20

Seminario de Sistemas 2 [A]

Jose Fernando Alvarez Morales

LABORATORIO

Seminario de Sistemas 2 'A'

Agenda

- Avisos
- Clase 2
  - Tablas de dimensión y hecho
  - Datawarehouse
  - ETL
- Tarea 1
- Hoja de Trabajo 1

Avisos Generales

- Lectura Nuevo Programa
- DTT
- Notas en el Excel

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 prosee un identificador único (**llave subrogada**) que lo une a la tabla de hechos

Llave subrogada

Es un identificador única que se asigna a cada registro de una tabla de dimensión

Son siempre de tipo numérico.
Preferiblemente, un entero autoincremental

Esta clave, generalmente, no tiene ningún sentido específico de negocio

Llave subrogada

Facilita el particionamiento eficiente de los datos físicos.

Crear una separación de modelos multidimensionales para facilitar el control de cambios.

Mejorar el rendimiento de operaciones.

Tablas de dimensión

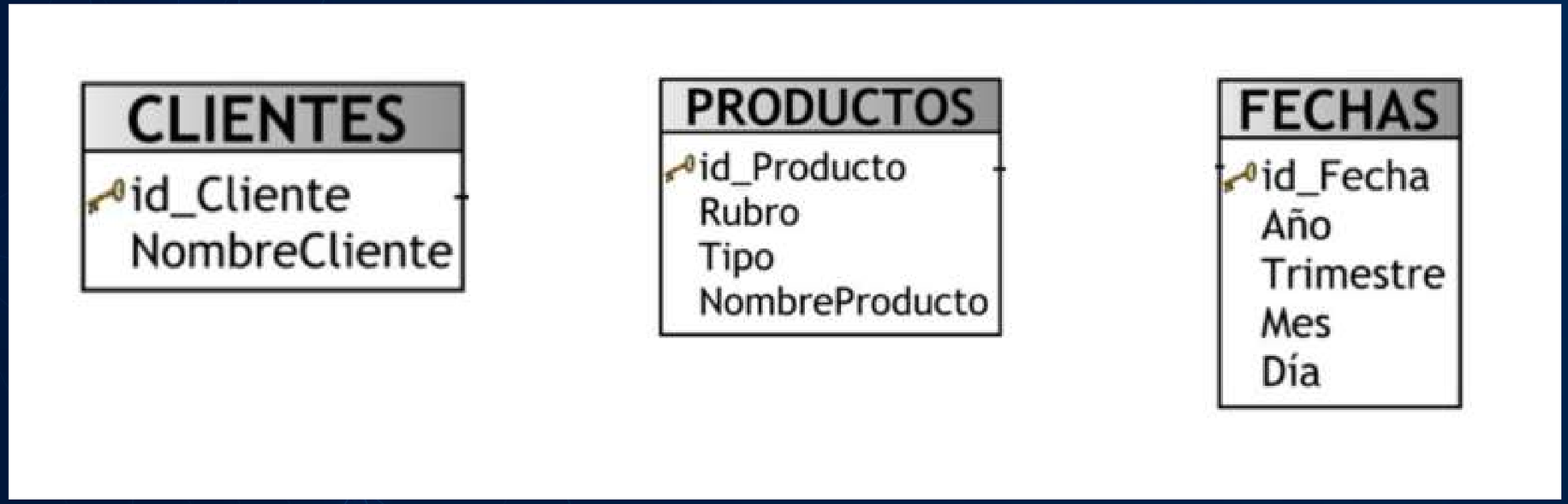


Tabla de Hechos

- Hechos utilizados por los analistas de negocio para la toma de decisiones.
- Indicadores del negocio como ventas, pedidos, reclamos, entre otros.
- Cada registro de esta tabla posee una clave primaria que se compone por claves primarias (subrogadas).
- Es importante resaltar que la tabla de hechos idealmente debe almacenar solo valores numéricos.

Tabla de Hechos - Llaves Subrogadas

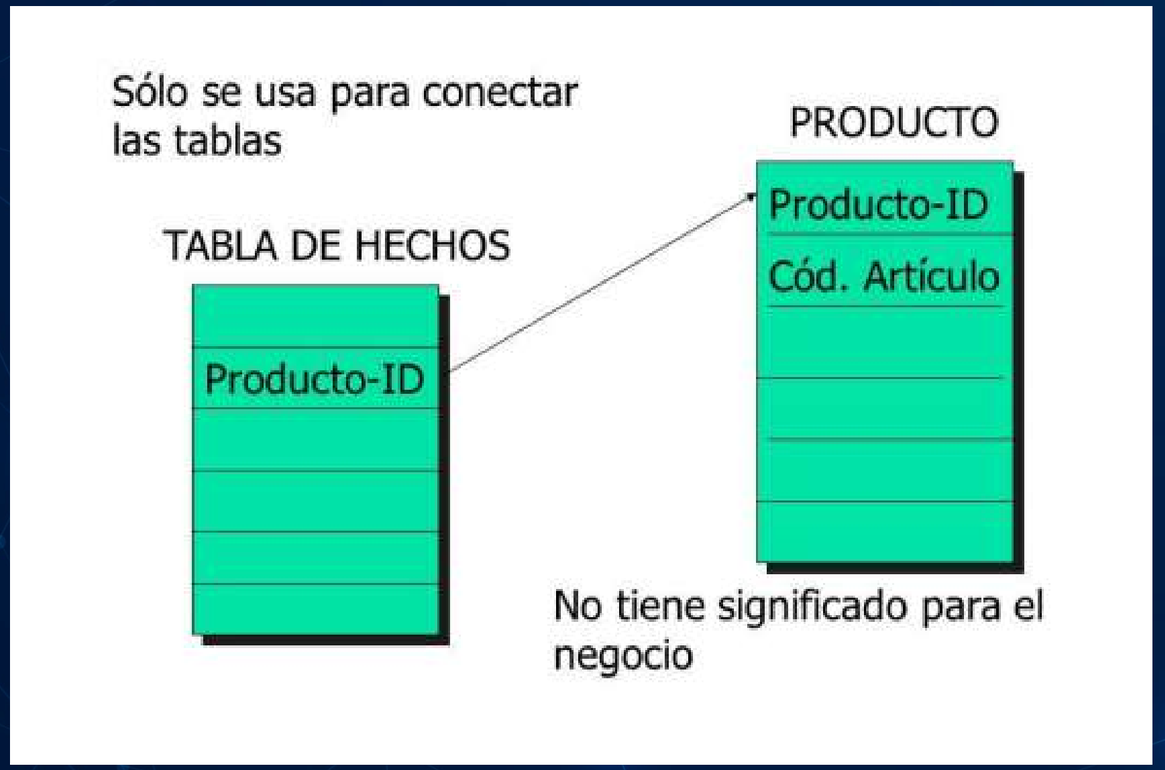


Tabla de Hechos

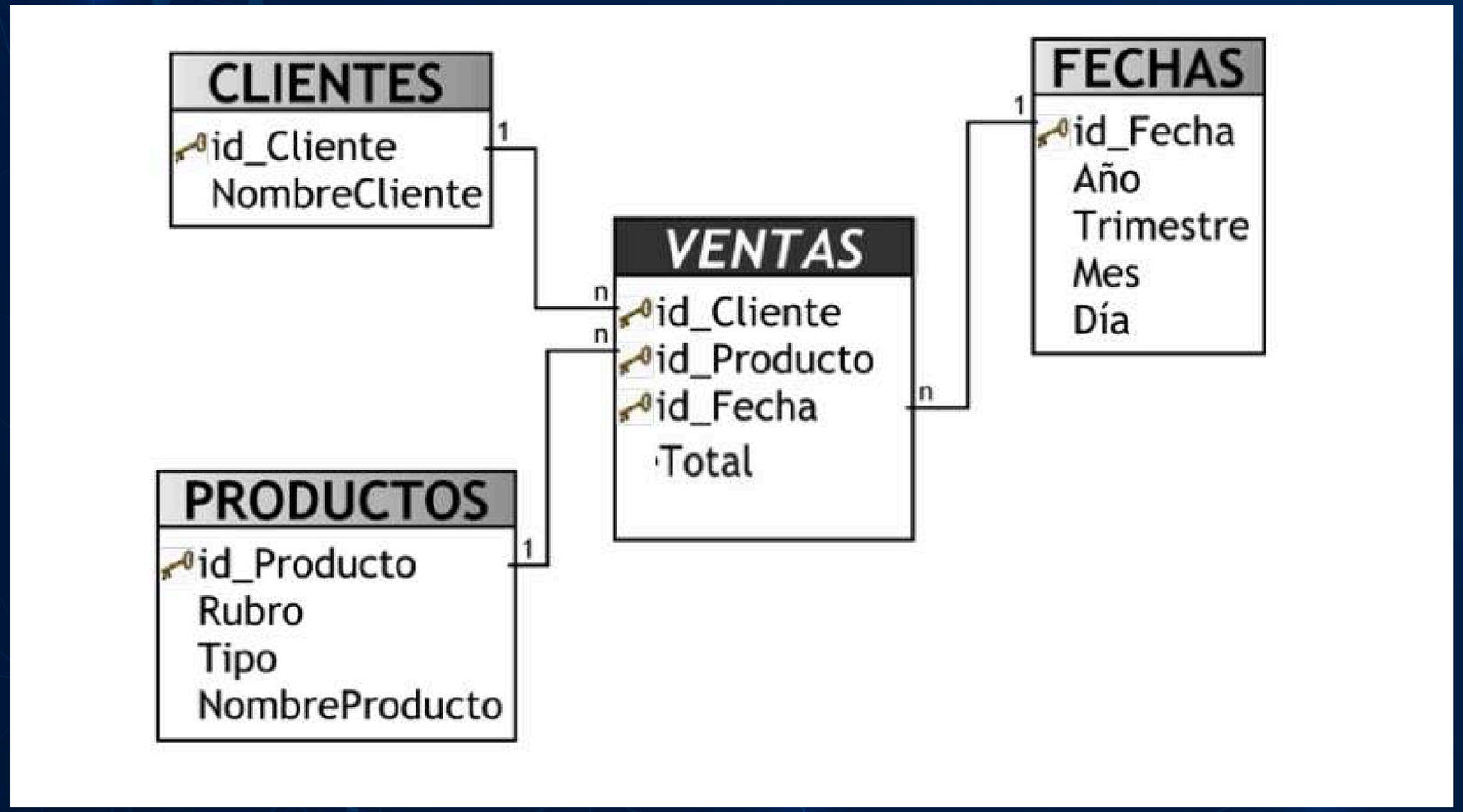
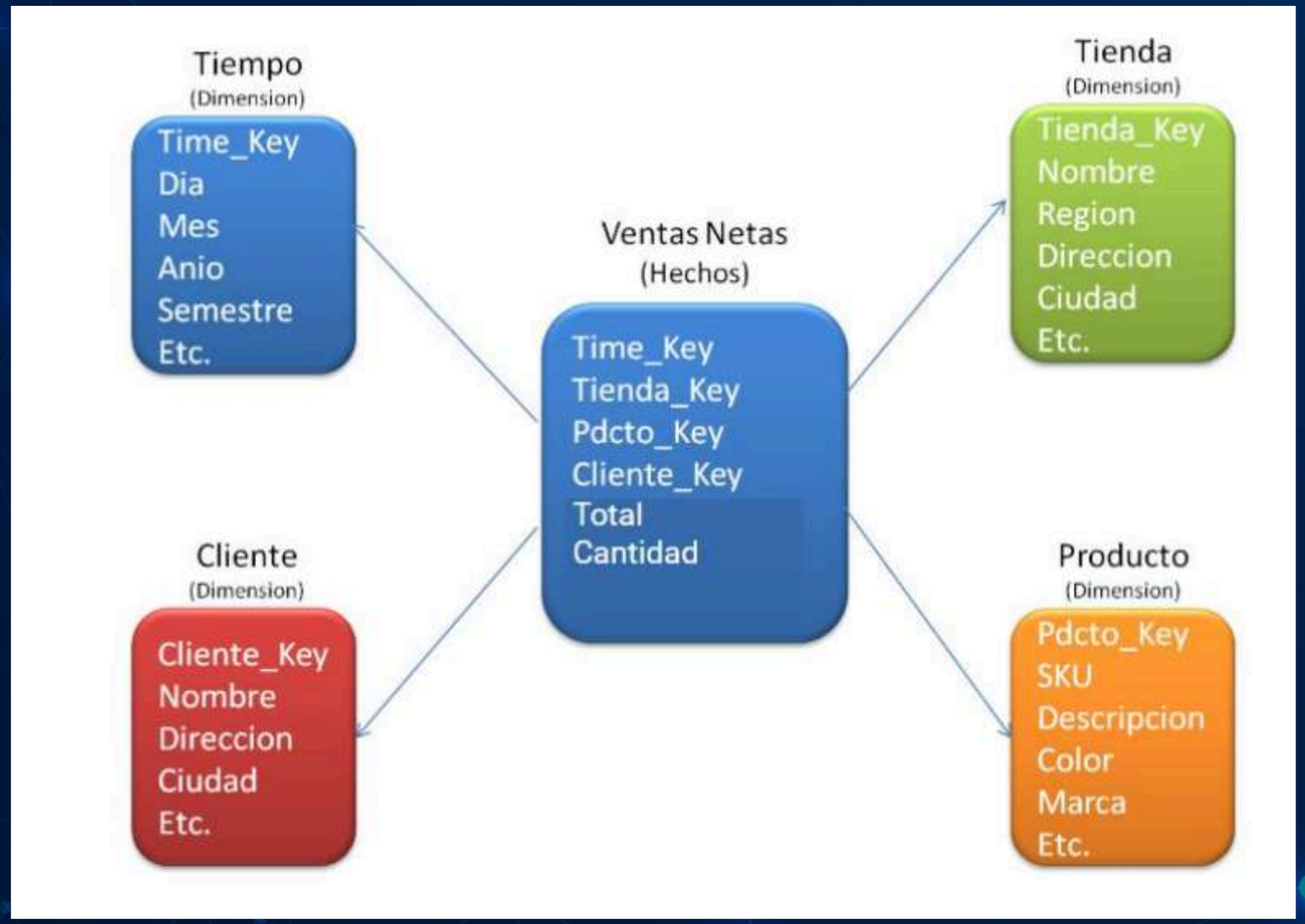


Tabla de Hechos



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.

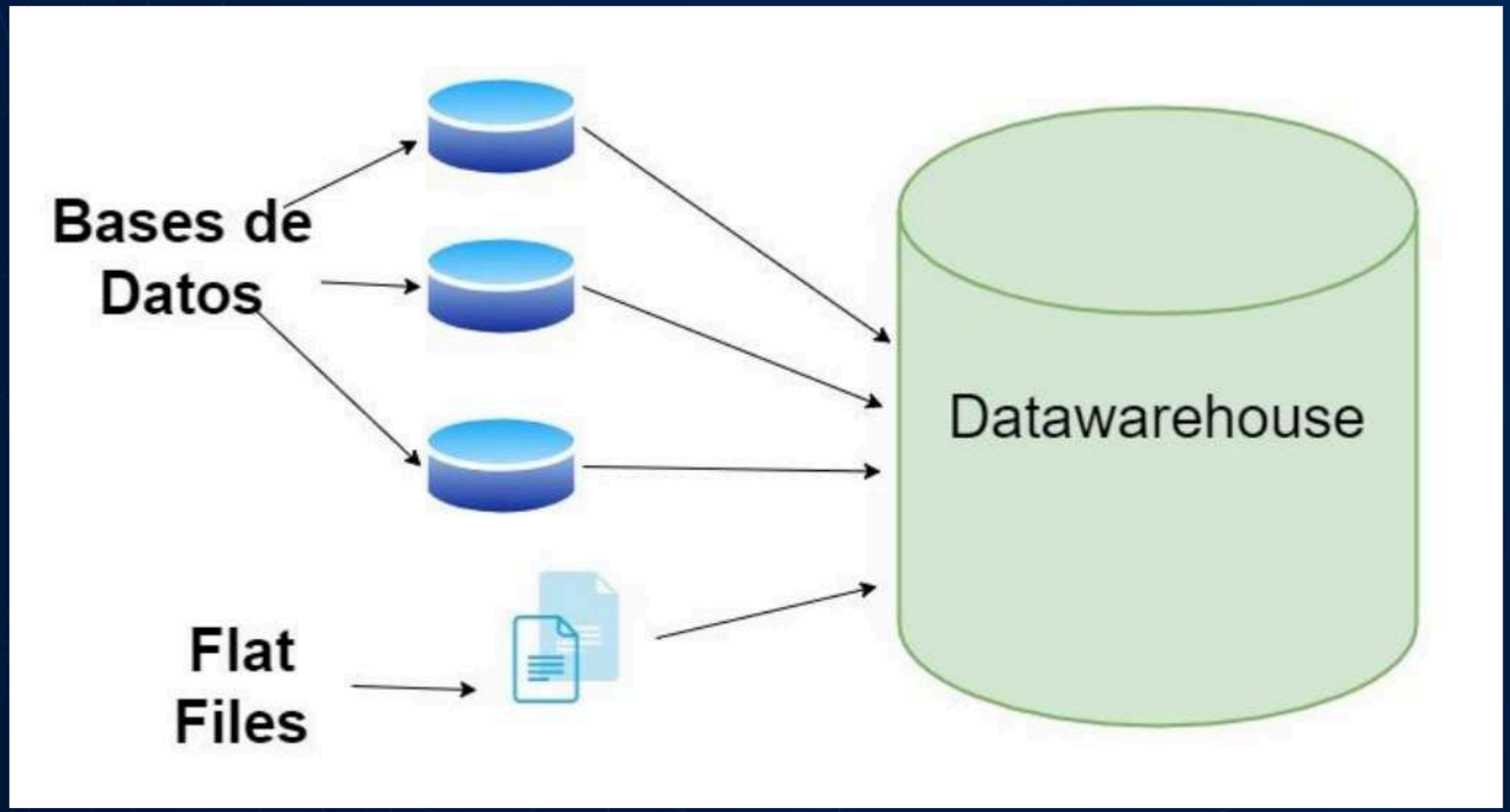
Datawarehouse

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

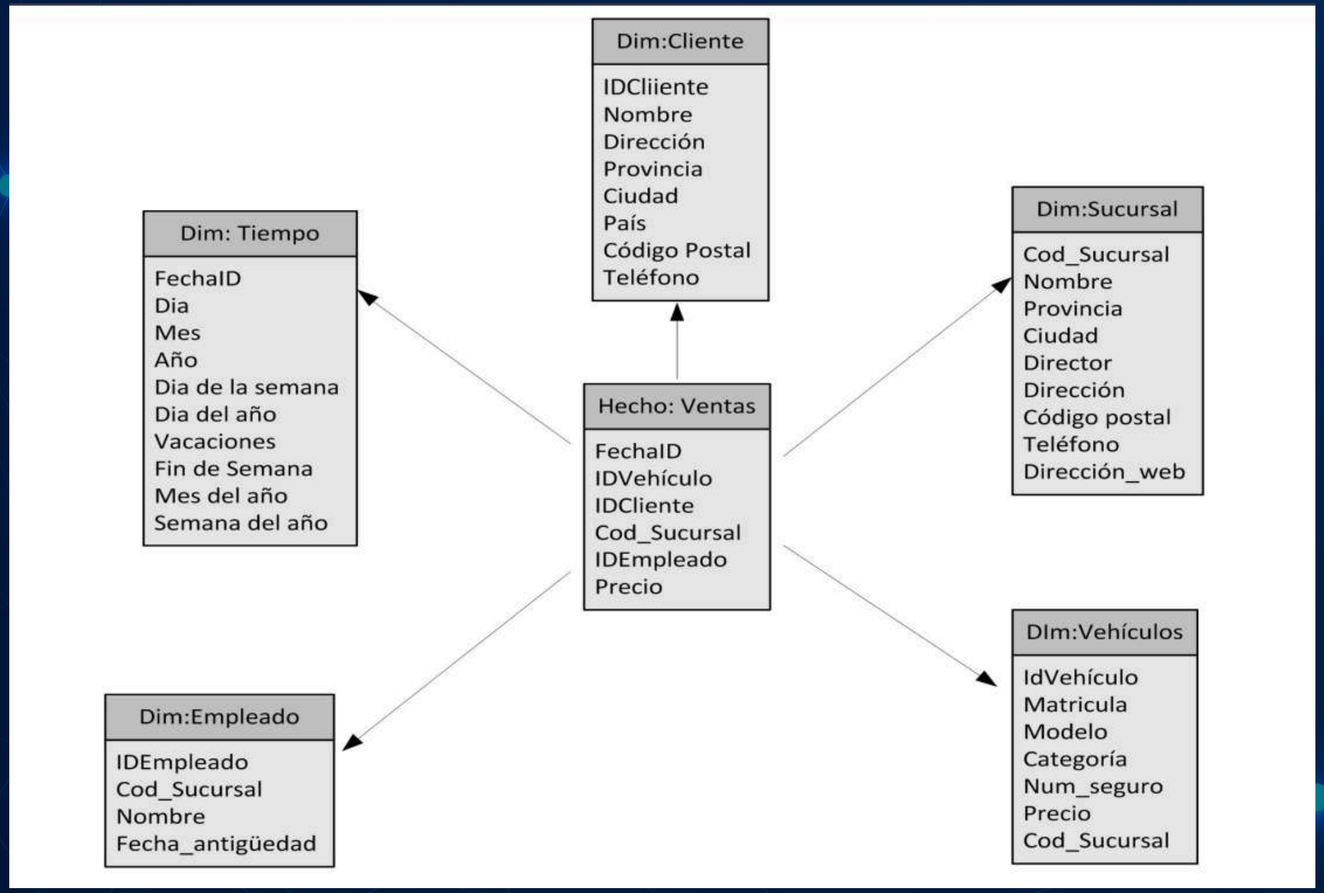
Datawarehouse

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 Datawarehose



Representación gráfica de un Datawarehose



¿Cuándo me interesa implementar un DataWarehouse

- Si necesito integrar muchas diferentes fuentes 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 adminstrar

Ventajas

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

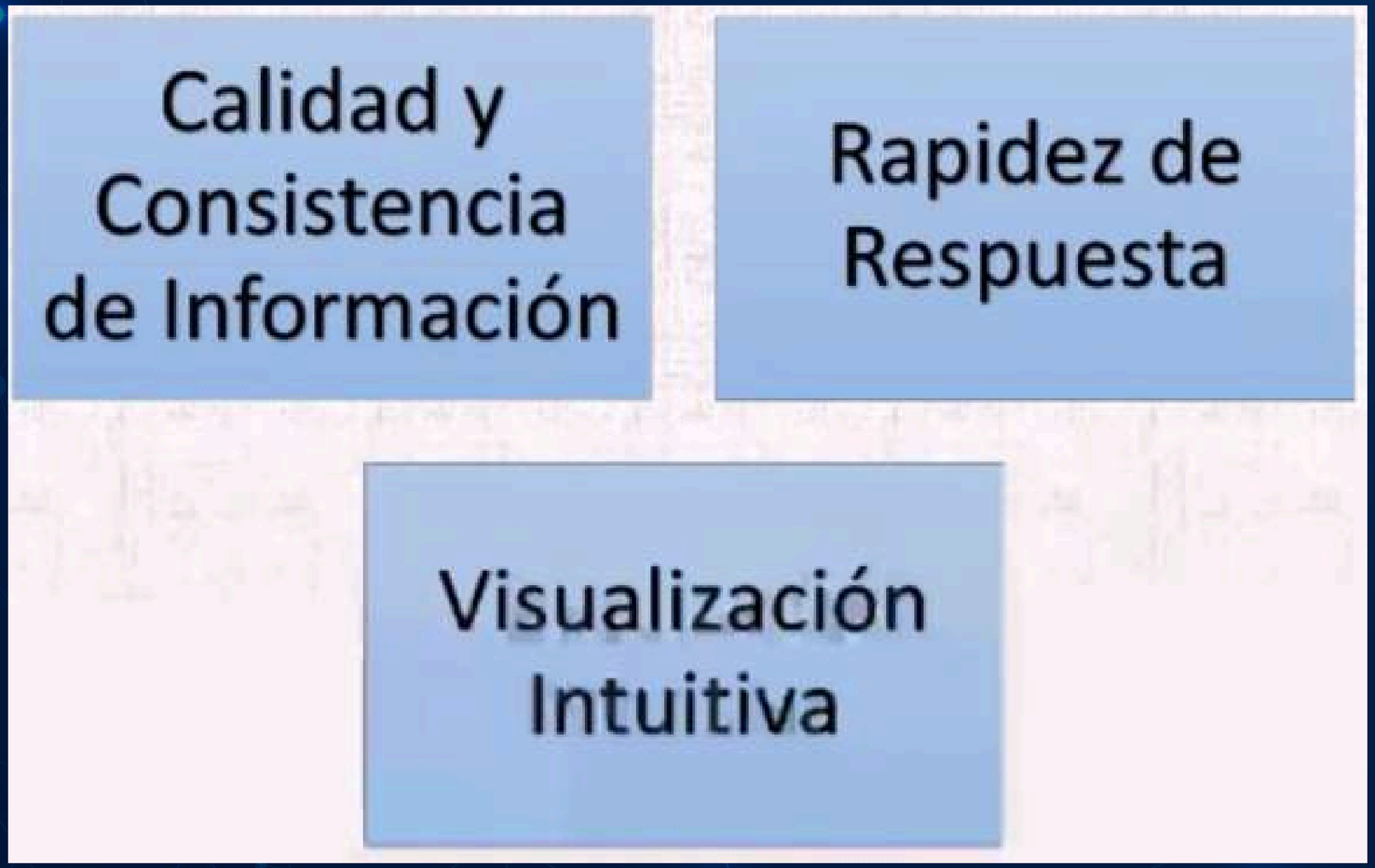
Desventajas

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

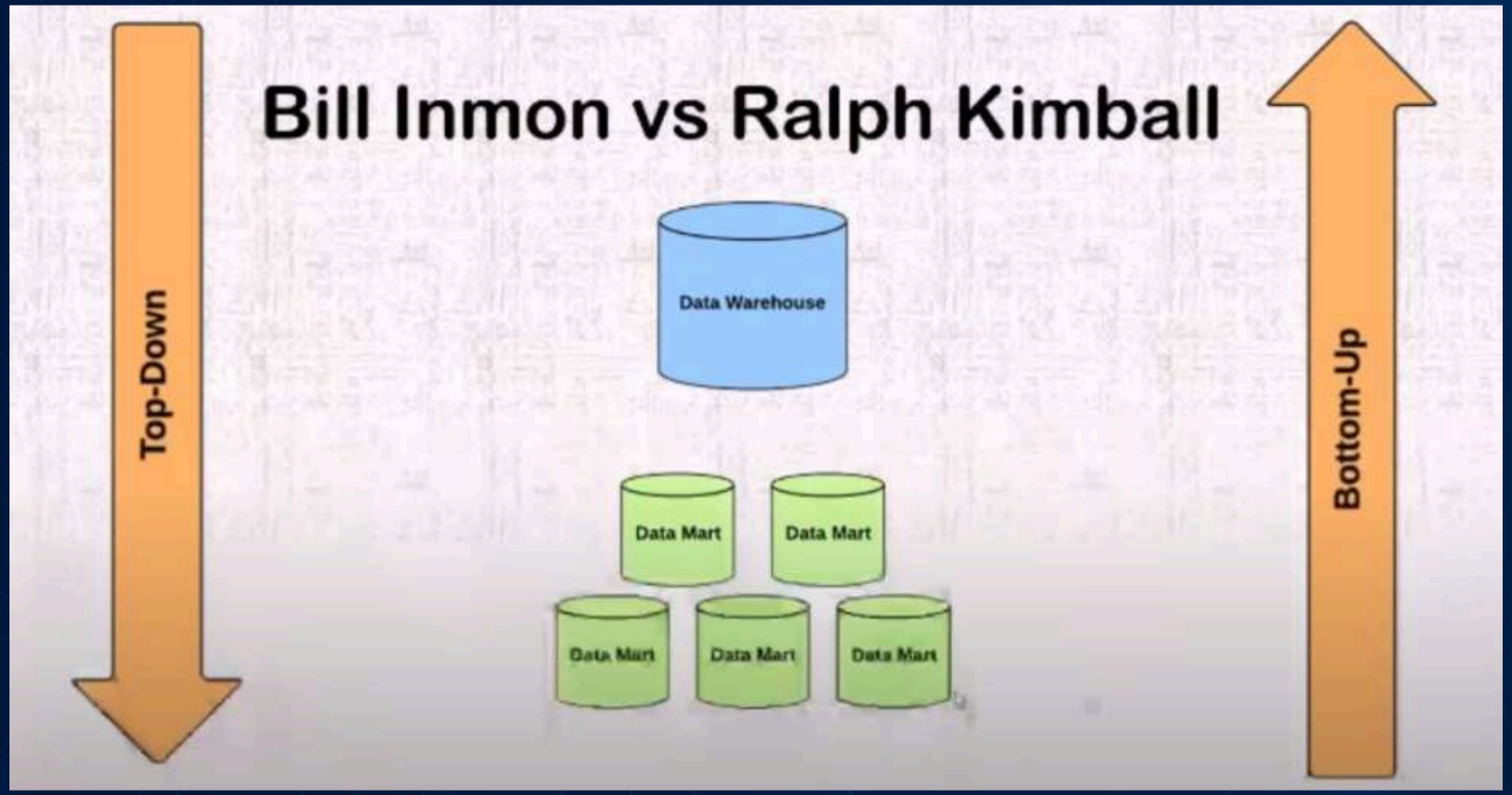
Aplicaciones

Predicción de Mercado	Análisis de Comportamiento	Modelado de Costos y Presupuestos
• Predecir el flujo de un mercado con información histórica y detección de patrones a través del tiempo.	• Estudiar y clasificar el comportamiento de los clientes y negocios de acuerdo a parámetros específicos.	• Utilizando funciones de agregación y agrupamientos, se pueden analizar los costos de operación para hacer mejoras en el negocio.

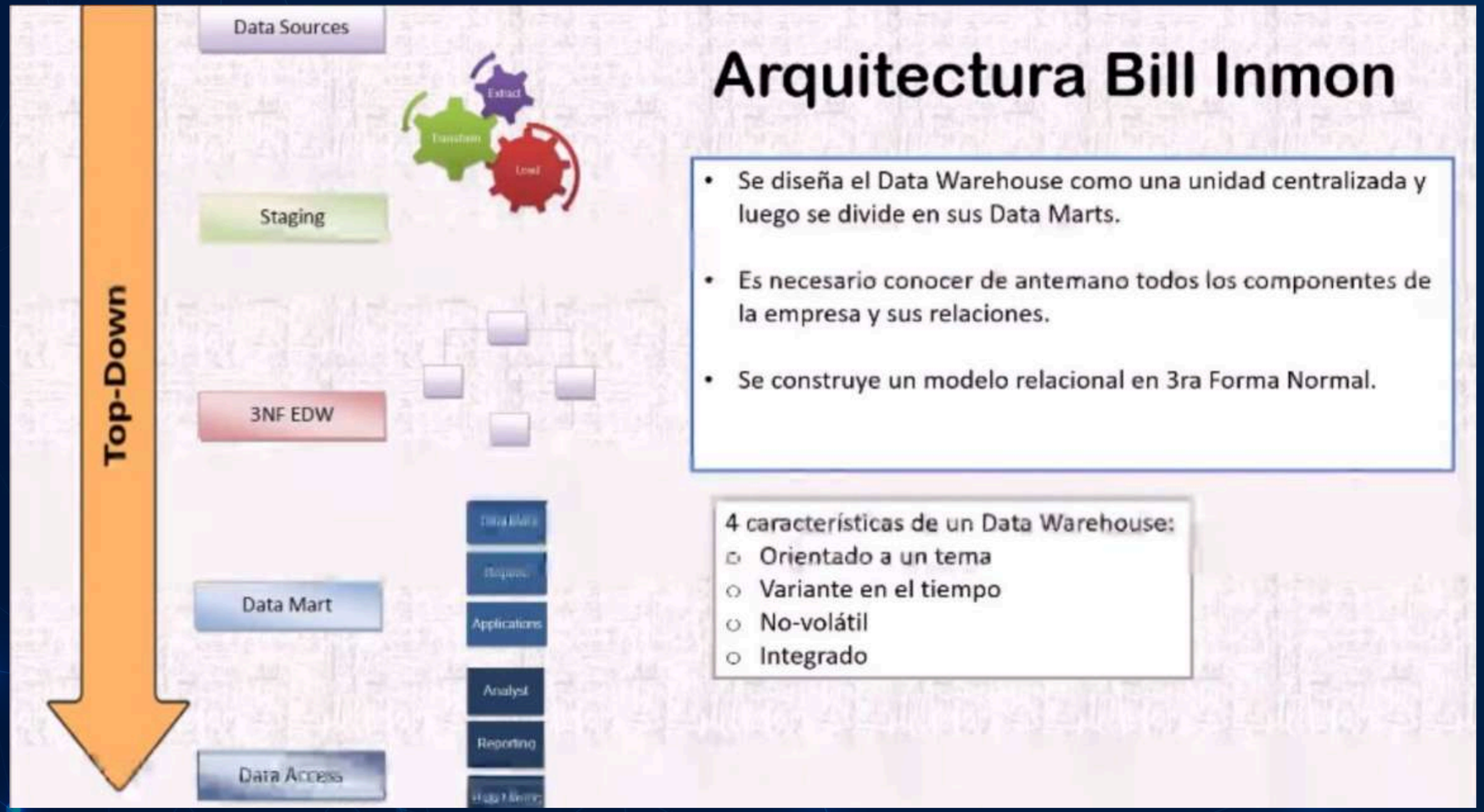
Beneficios


Calidad y Consistencia de Información
Rapidez de Respuesta
Visualización Intuitiva

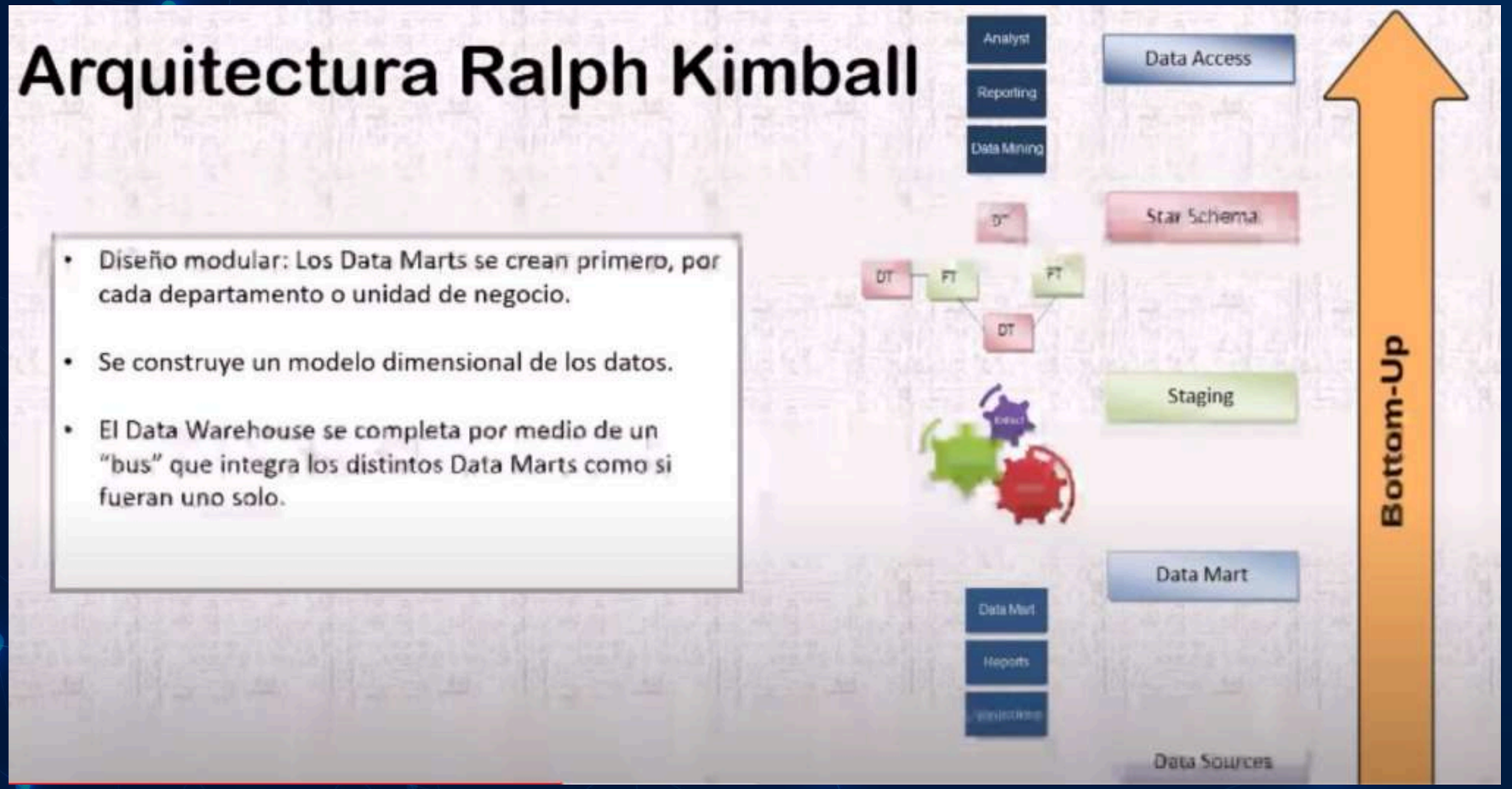
Arquitecturas



Arquitectura Bill Inmon



Arquitectura Ralph Kimball



Arquitectura Ralph Kimball

Bill Inmon (Top – Down)	Ralph Kimball (Bottom – Up)
Mayor tiempo en diseño previo, se necesita conocer toda la estructura de la empresa y sus procesos.	Modular, cada departamento crea su propio Data Mart de forma independiente.
Modelo relacional: 3ra Forma Normal. Parecido a Bases transaccionales. Poco eficiente para análisis.	Modelo dimensional: Esquema estrella, diseñado para análisis de datos y facilidad de lectura.
Datos actualizados de forma continua e integrada.	Datos se actualizan independientes en cada Data Mart, de forma asíncrona.
Proceso de carga y transformación de datos unificado.	Cada Data Mart se encarga de su carga y transformación de datos.
Mantenimiento y escalabilidad complejos.	Mantenimiento depende de cada departamento y se puede escalar agregando Data Marts.

Procesos 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 una Datawarehouse.

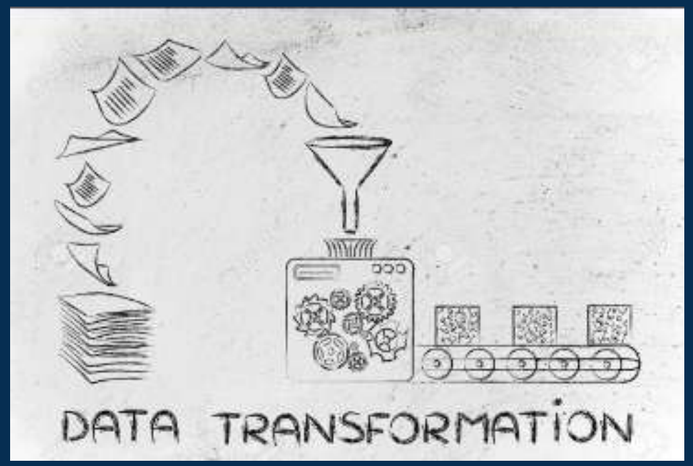
Extract - Extracción

- Es el primer 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

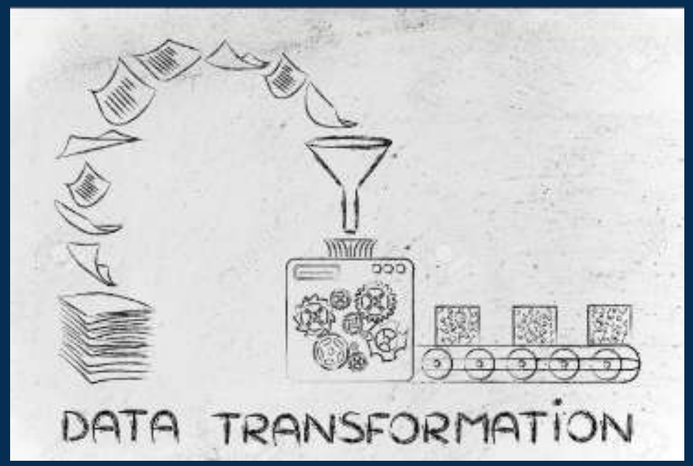
- El volumen de datos extraídos, así como el intervalo de tiempo entre extracciones, depende de las necesidades y requisitos del negocio

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:
  - Validación
  - Cálculos
  - Filtrado
  - Remoción 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

- 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

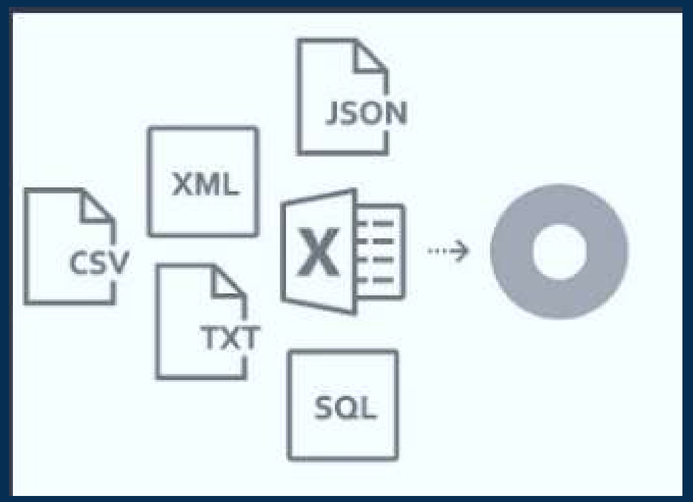
Desventajas

- 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 formas
- 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.

Desafios del Proceso ETL



- Procesamietno de datos en timepo real.
- Aumentar la velocidad del procesamiento de datos.
- Integración de nuevas fuentes de datos.

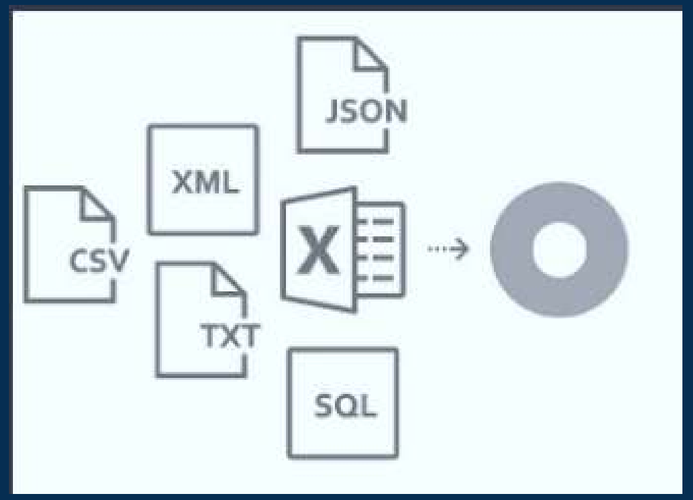
Procesamiento de datos en tiempo real.

- Cada dia 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.



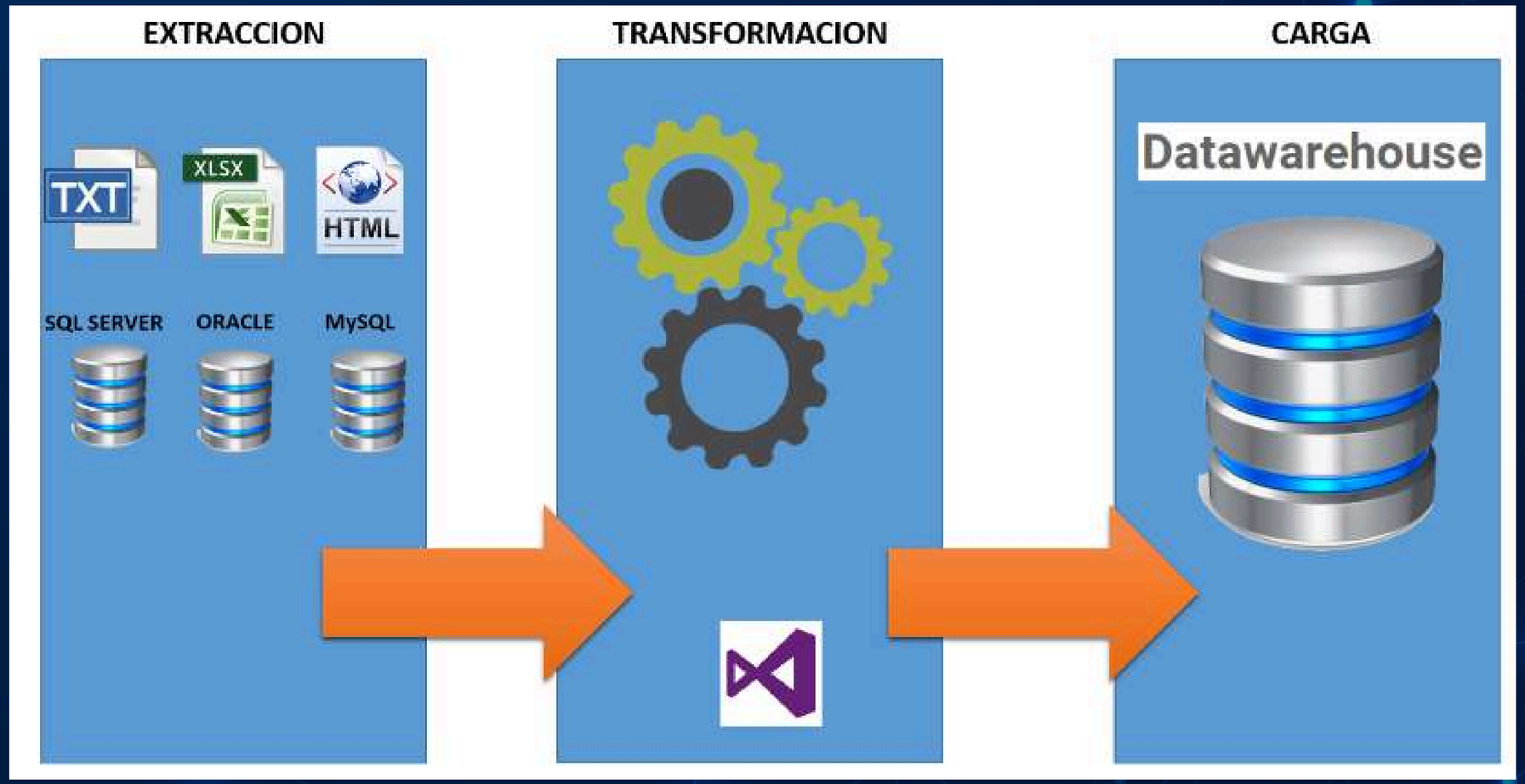
- El proceso de ETL necesita evolucionar para soportar nuevas fuentes de datos en cualquier momento.

Herramientas de ETL

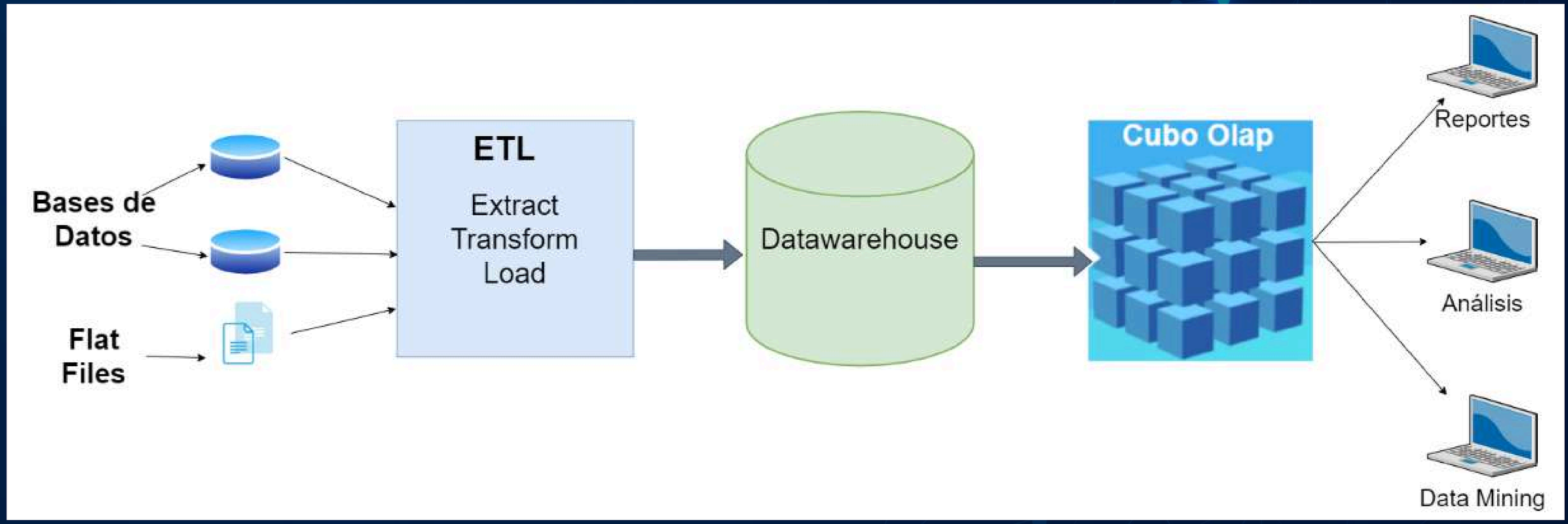
Hoy en día las principales herramientas para realizar ETL son:

- 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?



Thank You

Dudas?

Tarea 1

Ingresar a https://www.kaggle.com/datasets
Escoger un DataSet con más de 1000 datos y hacer la limpieza de los datos o transformación

-