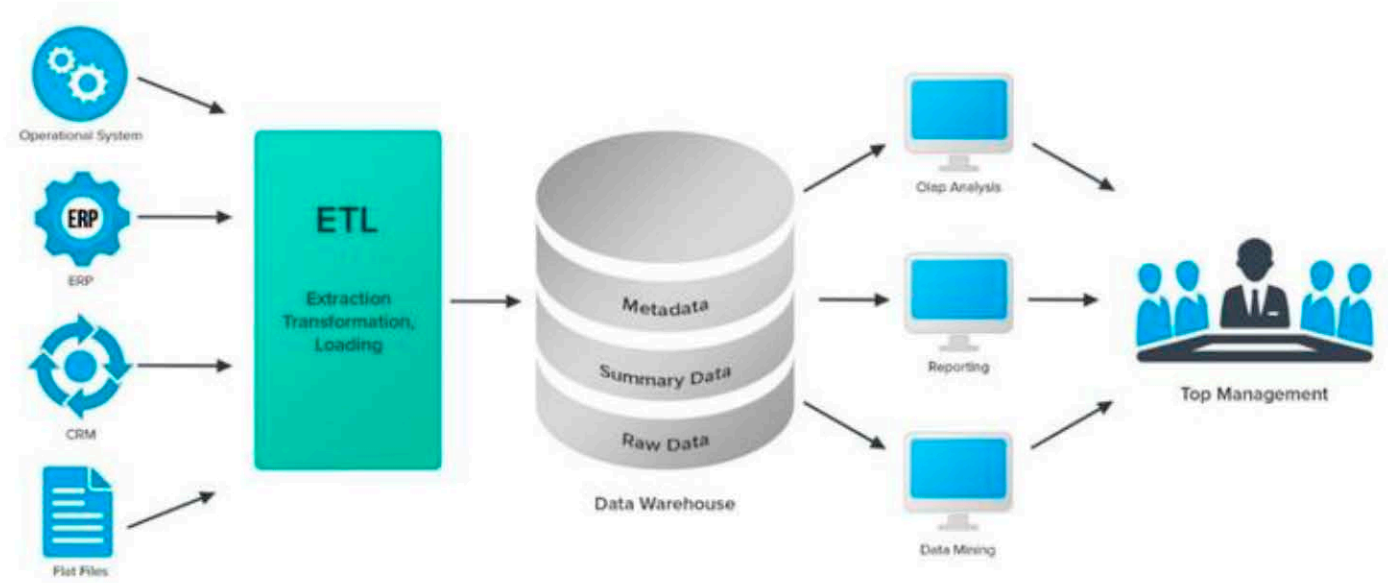
FACULTAD DE INGENIERÍA
ESCUELA DE COMPUTACIÓN



Datawarehouse y Minería de Datos

Guía #2: Herramientas ETL - Parte 1

¿Qué es el Datawarehouse?

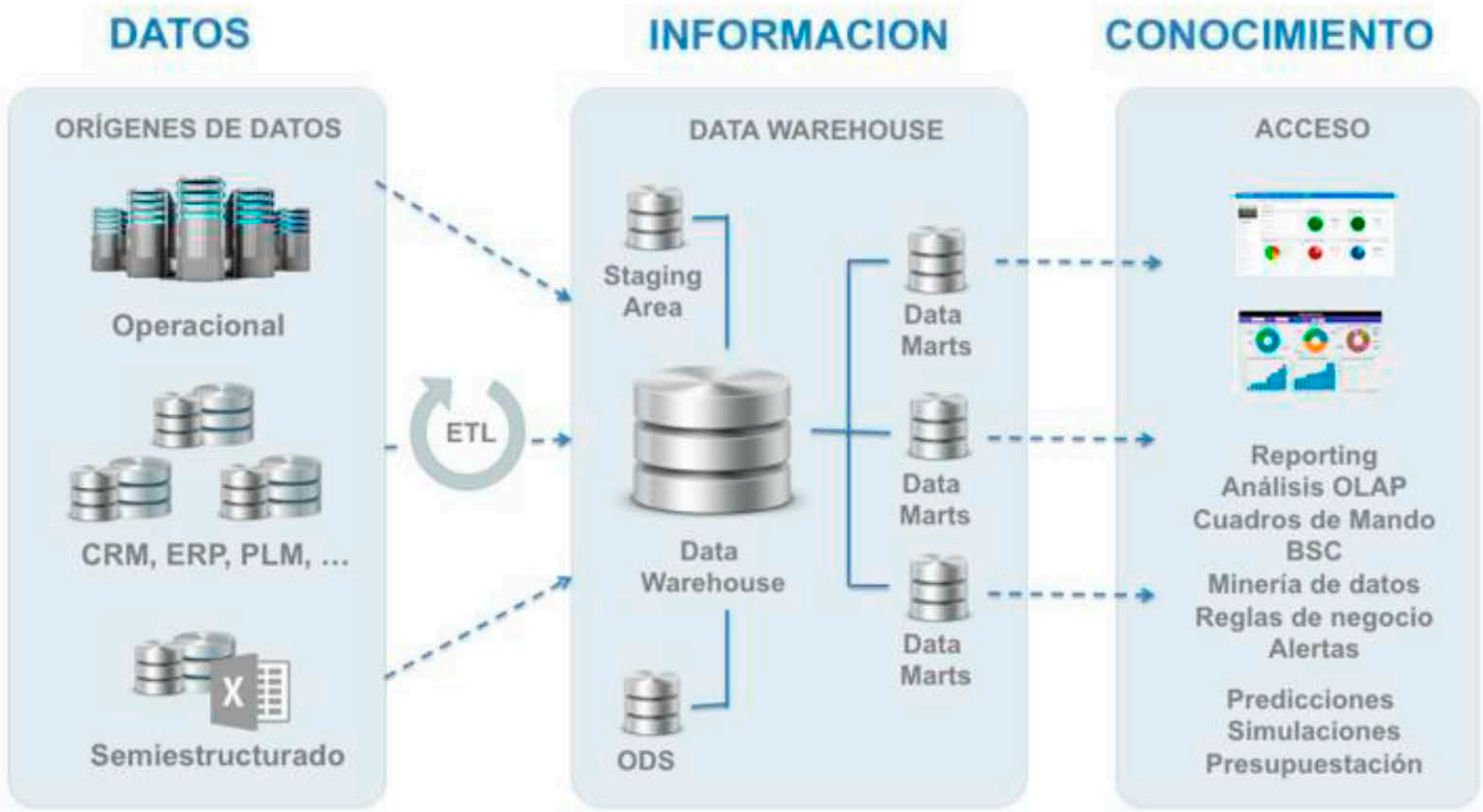
Un Data Warehouse es un gran almacén de datos e información que, además, recoge todos aquellos que son realmente necesarios para la realización de análisis e informe relacionado con el **Business Intelligence** (BI). Una parte fundamental en la toma de decisiones de las grandes empresas a la hora de establecer objetivos, establecer normativas y plantear riesgos.

En todos los casos, el data warehouse cumple la función de guardar la información de interés para la empresa con fines prácticos y estratégicos. Sus estructuras determinan el grado de usabilidad, aunque sobre todo ayudan a saber hasta qué punto es más fácil orientarse a diferentes ramas de un mismo negocio en base a los datos obtenidos.

¿Para que me sirve el Datawarehouse?

- Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
- Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén.
- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Simplifica dentro de la empresa la implantación de sistemas de gestión integral de la relación con el cliente.
- Supone una optimización tecnológica y económica en entornos de centro de información, estadística o de generación de informes.

Estructura del Datawarehouse



¿Qué es un ETL?

El termino ETL significa en español: Extracción, Transformación y Carga de Datos. Cada una de sus fases de explica a continuación:

1. En la **extracción** se obtienen los datos de las fuentes de origen mediante descarga de ficheros planos de texto, o facilitados por el cliente, y luego se cargan en el repositorio (ODS) en tablas intermedias, que contienen los datos sin la estructura final del modelo.

2. En la **transformación** se adecúa la información. En este proceso es típico duplicar tablas que contienen la información correcta y la creación de nuevos campos o nuevas tablas con datos agregados y/o calculados. Por ejemplo para agrupar información por criterios geográficos, temporales, o de estructura jerárquica o comercial que serán útiles para el análisis.

3. **Carga de datos**, donde una vez reorganizada la información, la cargamos en las tablas definitivas de nuestro/s repositorio/s de datos: datawarehouse (corporativo) y/o datamart (departamental). Nuevamente se duplican las tablas que contienen la información correcta

¿Qué es un Datamart?

El *DataMart* es un sistema orientado a la consulta, cuya distribución interna de los datos es clara y no hay dudas al respecto, estando éstos estructurados en modelos ***dimensionales de estrella o copo de nieve***.

Sin embargo, no sé puede decir lo mismo del *Datawarehouse*, para el que hay diferentes enfoques en cuanto a sus características y funciones. En este sentido, y haciendo alusión al principio de esta entrada donde comentaba que existen diferentes tipos de arquitecturas, es aquí donde tiene lugar un debate abierto desde la década de los 90 sobre las bases del *DataWarehouse*.

Existen otros enfoques en cuanto a la estructura interna y construcción del *DataWarehouse*, pero los más importantes son los de **Bill Inmon y Ralph Kimball.**

Modelo tipo Estrella

Los esquemas en estrella están compuestos por una tabla de hechos y varias tablas de dimensiones alrededor de ella, adquiriendo la característica **forma de estrella** que da nombre al esquema:

- La **tabla de hechos** contiene los índices y las métricas sobre los eventos de una empresa que se deben registrar de forma continua (el volumen de ventas, por ejemplo).
- Las **tablas de dimensiones** contienen atributos que describen los datos de la tabla de hechos. Se trata así de un conjunto de datos de referencia para los eventos registrados en la tabla de hechos.

Modelo tipo Estrella

En este tipo de esquema, la tabla de hechos está conectada por relaciones de clave externa con todas sus tablas de dimensiones, pero no estas entre sí. La siguiente imagen muestra una representación simplificada de esta estructura de datos:

Comencemos con nuestra práctica