Unidad 3

Recopilación. Almacenes de datos

Modelizado de Minería de Datos - Q12025

Las fases del proceso de extracción de conocimiento

Para extraer conocimiento útil de los datos, es fundamental comenzar con una recopilación e integración efectiva. La complejidad de esta tarea varía desde el análisis de simples archivos hasta la gestión de grandes volúmenes de datos provenientes de diversas fuentes. Este capítulo se enfoca en las metodologías y tecnologías necesarias para esta etapa crucial.

De la Recolección a la Preparación para el Análisis

Existen tres conceptos fundamentales en la fase inicial de la minería de datos:

- la recopilación de datos
- la integración
- la preparación

Aunque pueden sonar similares, cada uno tiene un papel único y crucial para asegurar que el conocimiento que extraigamos sea de alta calidad

Recopilación de Datos: La Estrategia

La recopilación de datos es la fase de planificación y estrategia. No se trata solo de copiar datos, sino de tomar decisiones inteligentes sobre el "qué" y el "dónde".

- Decidir las Fuentes
- Organización y Mantenimiento
- Niveles de Detalle y Vistas

Integración de Datos: La Tecnología

La integración de datos es la fase tecnológica que permite que los datos recopilados de diversas fuentes se unan y se vuelvan útiles.

- Data Warehousing
- OLAP (On-Line Analytical Processing)

OLAP: El Cubo de la Inteligencia de Negocio 🧊

OLAP (On-Line Analytical Processing) es una tecnología y una metodología de análisis de datos que permite a los usuarios explorar y analizar grandes volúmenes de datos desde múltiples perspectivas.

A diferencia de las bases de datos transaccionales, que están optimizadas para el registro de operaciones diarias, los sistemas OLAP están diseñados para el análisis interactivo y la toma de decisiones.

Su metáfora más común es la de un "cubo de datos".

OLAP: algunas claves

Análisis Multidimensional:

- Las herramientas OLAP permiten a los usuarios consultar grandes cantidades de datos desde diferentes perspectivas, o dimensiones, para obtener información valiosa.

Cálculos Rápidos

- Los sistemas OLAP están optimizados para operaciones de lectura intensiva, lo que permite calcular datos agregados (como ventas diarias o mensuales) de forma rápida y eficiente.

Toma de Decisiones:

- Al proporcionar información rápida y comprensible, las herramientas OLAP ayudan a las empresas a tomar decisiones basadas en datos, predecir tendencias y optimizar la asignación de recursos.

OLAP: algunas herramientas

Azure Analysis Services:

- Es una solución de análisis basada en la nube que proporciona funcionalidades de modelado y procesamiento OLAP, y se integra con Power BI para visualizar los datos.

Microsoft SQL Server Analysis Services (SSAS):

- Una herramienta tradicional de Microsoft para crear y administrar bases de datos OLAP, utilizando cubos para el procesamiento analítico en línea (MOLAP) y modelado multidimensional.

Amazon Redshift:

- Un almacén de datos en la nube diseñado para el análisis de grandes volúmenes de datos, optimizado para consultas analíticas rápidas y escenarios OLAP, según Amazon Web Services.

OLAP vs. Minería de Datos

Es crucial entender la diferencia entre OLAP y la minería de datos.

- OLAP es Deductivo: Su objetivo es confirmar una hipótesis o responder a una pregunta predefinida. El analista ya sabe qué busca.
- La Minería de Datos es Inductiva: Su objetivo es descubrir patrones o reglas ocultas que el analista no conoce de antemano.

Ambas tecnologías son complementarias. El análisis OLAP puede identificar una tendencia y la minería de datos puede descubrir la razón detrás de esa tendencia.

Juntos, ofrecen una visión completa del negocio.

OLTP y OLAP

Visita sugerida:
https://aws.amazon.com/es/what-is/olap/
https://aws.amazon.com/es/compare/the-difference-between-olap-and-oltp/

OLTP y **OLAP** son dos enfoques para gestionar bases de datos. OLTP se centra en transacciones en tiempo real, mientras que OLAP se centra en analizar datos

	OLTP	**OLAP**
Objetivo	Procesar transacciones comerciales	Analizar datos para tomas de decisiones
Operaciones	Insertar, actualizar, eliminar	Consultas analíticas complejas
Actualizaciones	Real time	Programadas
Copias de seguridad	Muy frecuentes	Menos frecuentes
Ejemplos	MySQL, MariaDB, PostgreSQL	AWS RedShift, Google BigQuery

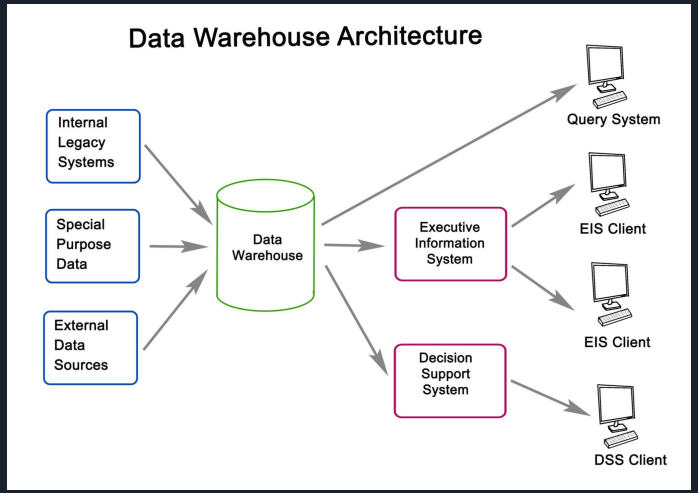
Datawarehouse

Los almacenes de datos (data warehouses) proporcionan metodologías y tecnología para recopilar e integrar datos históricos para análisis y extracción de conocimiento. Están diseñados para grandes volúmenes de datos estructurados, pero son útiles para conjuntos de datos más pequeños.

Datawarehouse: características clave

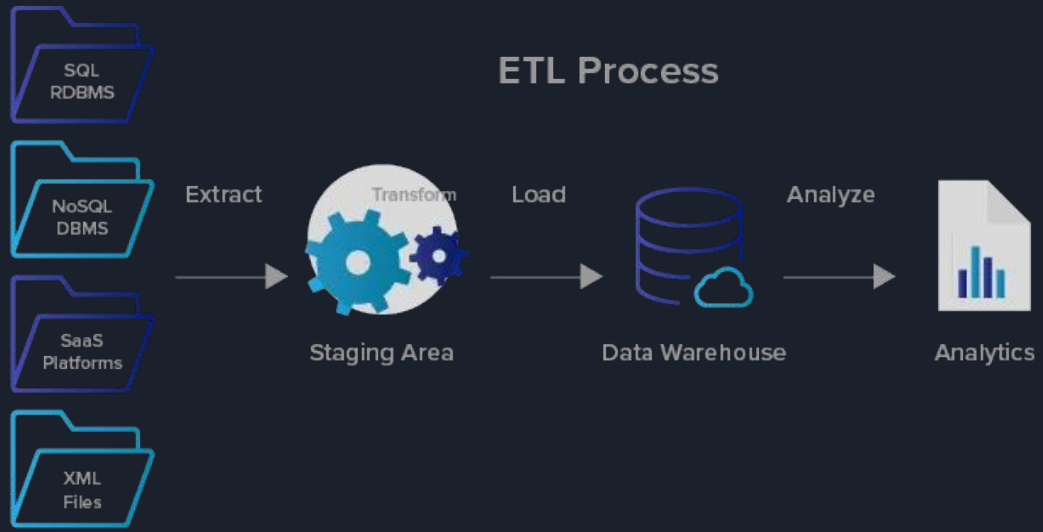
A diferencia de las bases de datos operacionales (las que gestionan las transacciones diarias), un Data Warehouse tiene características distintivas que lo hacen ideal para el análisis:

- Orientado por tema
- Integrado
- No volátil
- Variante en el tiempo



Carga y mantenimiento del almacén de datos

La carga y mantenimiento de un almacén de datos es una tarea delicada y requiere mucho esfuerzo. Se utiliza un sistema especializado llamado ETL (Extraction, Transformation, Load).

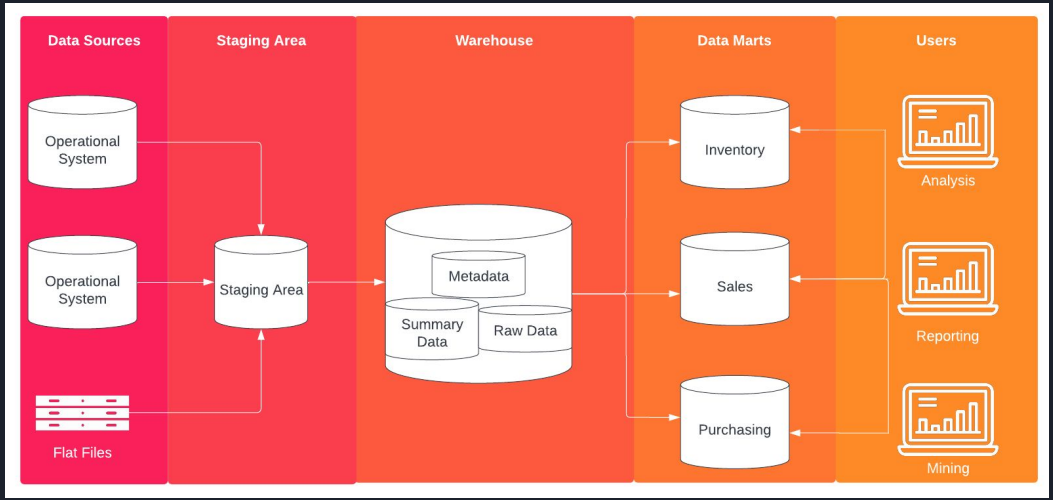


Tareas Clave de un Sistema ETL

- Lectura de Datos Transaccionales
- Incorporación de Datos Externos
- Creación de Claves Primarias
- Integración de Datos
- Obtención de Agregaciones
- Limpieza y Transformación
- Identificación de Cambios
- Planificación de Carga y Mantenimiento

El Repositorio de Datos Intermedio: La "Lona en Blanco" del ETL

- Este repositorio, a veces llamado "área de staging", es un espacio de trabajo temporal que sirve como una lona en blanco

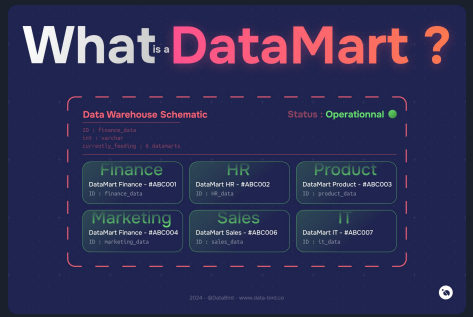


Datamarts

Un datamart es una versión específica del almacén de datos (data warehouse) centrados en un tema o un área de negocio dentro de una organización.

- Enfoque en un tema único
- Tamaño reducido
- Acceso para usuarios finales

Los datamarts no son un reemplazo de los data warehouses. Más bien, son complementarios



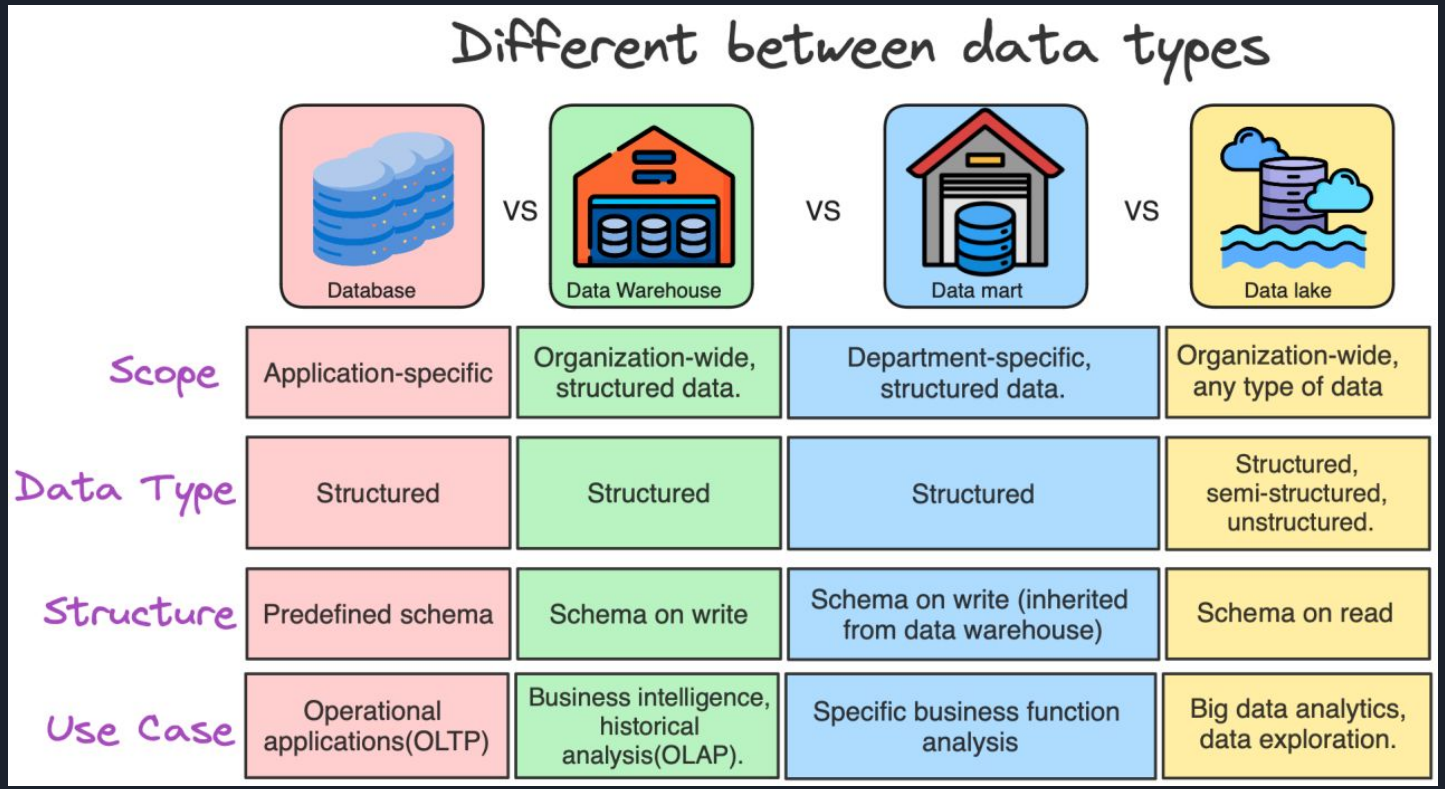
Data Lake: El Repositorio del Big Data en su Estado Crudo

Un data lake es un repositorio centralizado de datos que permite almacenar, procesar y proteger grandes cantidades de información. Los datos se pueden guardar en su formato original, sin necesidad de organizarlos o procesarlos previamente

Datalake

Característica	Data Lake	Data Warehouse
Tipos de datos	Crudos, de todo tipo.	Estructurados, limpios.
Estructura	Flexible (schema-on-read)	Rígida (schema-on-write)
Propósito	Exploración, descubrimiento, análisis avanzado	Reportes, Business Intelligence, análisis predefinido.
Usuarios	Científicos de datos, analistas avanzados	Usuarios de negocio, analistas de BI
Proceso ETL / ELT	ELT	ETL

Resumen



https://www.linkedin.com/pulse/differences-between-database-vs-data-warehouse-mart-lake-nguyen-tuan-ejo7c