MBD

Máster en Big Data

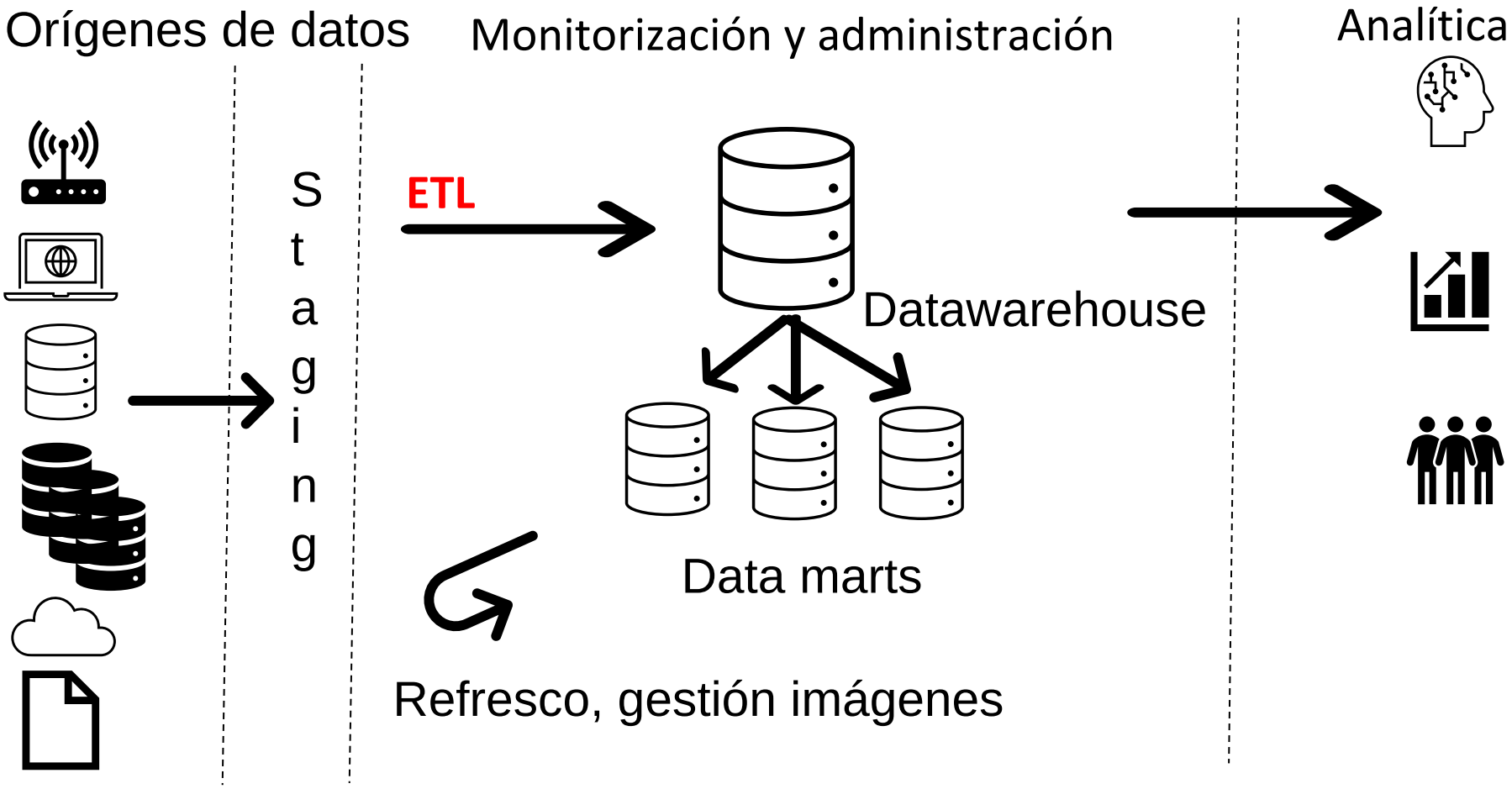
2023-2024

Tecnologías Business Intelligence

laSalle ENG
Universitat Ramon Llull

ETL asociada a almacén de datos

Arquitectura de la solución



ETL asociada a almacén de datos

Los tres pasos ETL

1) Extracción

Durante la extracción, ETL identifica los datos y los copia desde sus orígenes, de modo de poder transportar los datos al almacén de datos de destino. Los datos pueden proceder de orígenes estructurados y no estructurados, incluidos documentos, correos electrónicos, aplicaciones de negocios, bases de datos, equipos, sensores y terceros, entre otros.

2) Transformación

Dado que los datos extraídos no están procesados en su formato original, se deben asignar y transformar a fin de prepararlos para el almacén de datos final. En el proceso de transformación, ETL valida, autentica, desduplica o agrega los datos de manera que los datos resultantes sean fiables y se puedan consultar.

3) Carga

ETL traslada los datos transformados al almacén de datos de destino. Este paso puede implicar la carga inicial de todos los datos de origen o puede ser la carga de los cambios incrementales en los datos de origen. Puede cargar los datos en tiempo real o en lotes programados.

Fuente: https://www.oracle.com/es/integration/what-is-etl/

ETL asociada a almacén de datos

1) Extracción (Extract)

Identificación de fuentes de datos: Identificar las fuentes de datos que alimentarán el Data Warehouse. Esto puede incluir bases de datos transaccionales, archivos planos, servicios web, entre otros.

Conexión a fuentes: Establecer conexiones con las fuentes de datos para extraer la información necesaria. Se utilizan técnicas como consultas SQL, lectura de archivos, API, etc.

Extracción de datos: Realizar la extracción de datos de las fuentes identificadas, teniendo en cuenta aspectos como la frecuencia de actualización y la cantidad de datos a transferir.

ETL asociada a almacén de datos

2) Transformación (Transform)

Limpieza de datos: Identificar y corregir errores o inconsistencias en los datos extraídos. Esto puede incluir la eliminación de duplicados, corrección de valores nulos y normalización de datos.

Transformación de datos: Aplicar reglas de negocio y transformaciones a los datos extraídos para que se ajusten al esquema de datos del Data Warehouse. Esto implica la conversión de formatos, cálculos, derivación de nuevas variables, etc.

Enriquecimiento de datos: Agregar información adicional a los datos para mejorar su calidad y relevancia. Esto podría incluir la incorporación de datos de referencia o la combinación de datos de varias fuentes.

ETL asociada a almacén de datos

3) Carga (Load)

Diseño de estructuras de almacenamiento: Definir la estructura de las tablas y otros objetos de almacenamiento en el Data Warehouse donde se cargarán los datos transformados.

Carga de datos: Insertar los datos transformados en el Data Warehouse, siguiendo la estructura previamente definida. Esto puede implicar la actualización de datos existentes o la carga de datos nuevos.

Índices y optimización: Aplicar índices y realizar ajustes de rendimiento para garantizar un acceso eficiente a los datos en el Data Warehouse.

ETL asociada a almacén de datos

Monitorización y Mantenimiento

Monitoreo del rendimiento: Establecer mecanismos de monitorización para evaluar el rendimiento del proceso ETL y detectar posibles problemas.

Programación y automatización: Configurar programaciones automáticas para ejecutar el proceso ETL de manera regular y consistente. Monitorización asociada como proceso y calidad.

Gestión de cambios e incidencias: Gestionar y adaptar el proceso ETL a medida que cambian las fuentes de datos o los requisitos del negocio.

ETL asociada a almacén de datos

Oracle

Fuente: https://www.oracle.com/es/integration/what-is-etl/

Productos y soluciones de ETL

Conjunto de arquitectura orientada a servicios (SOA)

¿Cómo reducir la complejidad de la integración de aplicaciones? Con capacidades simplificadas de integración en la nube, móviles, locales y del Internet de las cosas, todas dentro de una única plataforma, esta solución puede proporcionar un tiempo de integración más rápido y una mayor productividad, junto con un menor costo total de propiedad (total cost of ownership, TCO). Muchas aplicaciones empresariales, como Oracle E-Business Suite, utilizan mucho este producto para orquestar flujos de datos.

GoldenGate

La transformación digital a menudo exige mover los datos desde donde se capturan hasta donde se necesitan, y GoldenGate está diseñado para simplificar este proceso. Oracle GoldenGate es una solución de replicación de datos de alta velocidad para la integración en tiempo real entre bases de datos heterogéneas ubicadas localmente, en la nube o en una base de datos autónoma. GoldenGate mejora la disponibilidad de los datos sin afectar el rendimiento del sistema, además de proporcionar acceso a los datos en tiempo real e informes operativos.

Cloud Streaming

Nuestra solución Cloud Streaming proporciona una solución duradera, escalable y totalmente gestionada para tomar y consumir flujos de datos de gran volumen en tiempo real. Utilice este servicio para mensajería, registros de aplicaciones, telemetría operativa, datos del flujo de clics web o cualquier otra instancia en la que los datos se produzcan y se procesen de forma continua y secuencial en un modelo de mensajería de publicación/suscripción. Es totalmente compatible con Spark y Kafka.

ETL asociada a almacén de datos

Capabilities (functional requirements)

- Bulk/batch data movement
- Data replication and synchronization
- Data virtualization
- Stream data integration
- Advanced data transformation
- Data API services
- Augmented data integration
- Self-service data preparation
- Metadata support
- Data governance support

Fuente Gartner: https://blogs.oracle.com/dataintegration/post/data-integration-gartner-mq2022

ETL asociada a almacén de datos

Optional capabilities

DataOps support

FinOps support

Fuente Gartner: https://blogs.oracle.com/dataintegration/post/data-integration-gartner-mq2022

MBD

Máster en Big Data

Tecnologías Business Intelligence

laSalle ENG
Universitat Ramon Llull