Investigación de ETL vrs ELT y OLTP vrs OLAP

Pablo Santizo
Carne: 24000134

Maestría en Ciencia de Datos
Instituto en Investigación de Operaciones
Universidad Galileo

07 de febrero de 2024

Índice

1. ETL vrs ELT

1.1. ETL

ETL es el proceso de extracción, transformación y carga (por sus siglas en inglés), el cual consiste en un conjunto de procedimientos en la canalización de datos. En este proceso se recopilan datos brutos de sus fuentes (extract), se limpian y se agregan los datos (transform) y se guardan en una base de datos (load), para que pueda ser analizada.

1.1.1. Extraer (extract):

En esta etapa se recopilan los datos a partir de sus fuentes. Estos datos se llevarán, en última instancia hacia filas y columnas de una base de datos. Tradicionalmente, estos datos se extraían de archivos Excel y de sistemas de bases de datos de gestión, dado que estás eran las principales fuentes de información. No obstante, con el aumento de las aplicaciones, la mayoría de las empresas ahora obtienen información valiosa de las mismas, como, por ejemplo, Facebook para aumentar el rendimiento de la publicidad, de Google Analytics para los sitios web, entre otros.

1.1.2. Transformar (transform):

En esta etapa, se toman los datos que se recopilaron y se le realizan cambios (transformaciones) antes de ser guardadas en las bases de datos. Existen diferentes transformaciones como las siguientes:

- **Limpieza** de datos: se identifican datos sospechosos y se corrigen o eliminan.
- **Enriquecimiento** de datos: se agrega nueva información más allá de los datos en bruto, como por ejemplo realizar cálculos sobre dicha información.

Por lo que en esta etapa se busca que los datos se preparen para que tengan la forma y el formato correcto, para que los analistas puedan utilizarla con rapidez y puedan extraer la información necesaria, optimizando el tiempo que gastarían si los tuvieran que limpiar.

1.1.3. Carga (load):

En esta etapa se toman los datos de la etapa de transformación y se guardan en un almacén de información, como una base de datos relacional, un almacén NoSQL, en un almacén de datos (data warehouse) o un lago de datos (data lake), con el fin de que los datos estén listos para su análisis. Existen diferentes tipos de cargas de información, como la carga completa, carga incremental por lotes y carga de flujo incremental, cada uno con su pros y contras.

Independientemente de la arquitectura que se elige existen desafíos en esta etapa que se deben de considerar como, por ejemplo:

- **Inserción de datos:** por ejemplo, el orden la inserción puede afectar el resultado final si las tablas no tienen la misma llave foránea, por lo que se deberían de introducir primero los datos coincidentes para evitar omisión de datos.
- **Cambios de esquema:** derivados de la evolución de los negocios y sus necesidades, podrían resultar en actualización de dichos esquemas y conducir a pérdidas de horas de trabajo y consecuencias negativas para el sistema.
- **Calidad de datos:** los datos sospechosos que tienen formatos que eluden la validación de datos en la extracción e información. Por lo que se necesita un monitoreo adicional para asegurar la calidad de datos en las bases de datos.

Por lo que el proceso ETL recopila los datos de varias fuentes y se guarda en un lugar de preparación, posteriormente sufren grandes transformaciones y por último se guardan en un almacén de datos.

1.2. ELT

Este proceso es similar al ETL, en el sentido que también es un proceso de extracción, transformación y carga de datos. No obstante, la diferencia consiste en que los datos se recopilan, luego se cargan todos en un lago de datos (data lake), para posteriormente aplicar transformaciones y moverlos a un almacén de datos.

1.3. Ventajas y desventajas

- **Estructura de los datos en el almacén:** Los procesos ETL solo almacenan datos estructurados, también conocidos como relacionales. Por su parte, el proceso ELT almacena todos los tipos de datos estructurados tal y como aparecen en los datos de origen.

  Por lo que se considera una ventaja que el proceso ELT puede almacenar datos de cualquier estructura. Por su parte la tecnología ETL representa una desventaja al poseer mayor restricción en la estructura.

- **Volúmenes de datos:** ETL generalmente opera en los rangos de megabytes o gigabytes, mientras que ELT trabaja con órdenes de volúmenes de datos (petabytes o terabytes) de mayor magnitud.

  Por lo que ELT presenta una ventaja en el volumen de información que puede guardar.

- **Velocidad de carga de información:** Dado que la tecnología ETL necesita transformar los datos antes de cargarlos, todo el proceso experimenta más tiempo ELT. Las transformaciones pueden llevar mucho tiempo, especialmente si requieren un mayor tratamiento, como en consultas agregadas complejas o limpieza de datos, como la transformación de datos no estructurados en datos estructurados.

  Esto le da una ventaja a la tecnología ELT teniendo canalizaciones más rápidas. Por su parte la tecnología ETL presenta una desventaja al necesitar mayor tiempo para cargar la información.

- **Flexibilidad:** Dado que la estructura de los datos de ETL se especifica de antemano, colocando restricciones sobre la información, esto lo hace menos flexible. Por su parte los datos en ELT se cargan primero y se transforman después, proveyendo mayor flexibilidad.

  Esto le da una ventaja en flexibilidad a la tecnología ELT, dado que si algún algoritmo necesita mayor información, solo se debe de hacer la consulta seleccionando otros campos, lo que no sería posible en el proceso ETL (desventaja).

- **Requerimientos para almacenar:** ELT necesita mayor capacidad de almacenamiento que ETL, dado que estos se guardan sin transformar. Lo que provee a ETL una ventaja en costos de almacenamiento, no obstante, por el uso de las nubes como servicio de espacio, la brecha de este costo empieza a disminuir.

  Por lo que la tecnología ETL posee una ventaja en costos de almacenamiento y la tecnología ELT una desventaja.

- **Madurez:** La tecnología ETL ha estado en funcionamiento en varias décadas, proveyendo un mayor desarrollo y una arquitectura más solida que la tecnología ELT, por lo que esta primera tecnología tiene mas herramientas disponible y mayor seguridad.

  Lo anterior le brinda una importante ventaja a la tecnología ETL al tener mayor desarrollo en sus herramientas y seguridad, siendo una desventaja para la tecnología ELT al tener un importante riesgo en la seguridad, debido a la necesidad de mayor tiempo de desarrollar estas características de la tecnología.

1.4. ¿Qué tecnología recomendaría en cada caso?

ETL es el mejor para el análisis rápido en entornos de datos pequeños y medianos, donde los datos de entrada y las transacciones de datos están bien controlados y no cambian constantemente, por lo que no se necesita flexibilidad.

Por su parte **ELT** es mejor para trabajar con datos semiestructurados o no estructurados, en entornos de datos masivos, y donde los requisitos de la operación son cambiantes, aprovechando su flexibilidad.

2. OLTP vrs OLAP

2.1. OLTP

Es el procesamiento de transacciones en línea (OLTP por sus siglas en inglés), estas transacciones se realizan en tiempo real en sistemas de bases de datos. Esta tecnología se enfoca en el registro y el eficiente procesamiento de las transacciones individuales.

Los sistemas OLTP generalmente están orientados a los clientes y se utiliza para el procesamiento de las transacciones y realizar consultas por parte los empleados, clientes y personas de tecnologías de la información.

Este tipo de sistema gestiona los datos actuales que suelen ser demasiado detallados, con el propósito de que sean utilizados fácilmente para la toma de decisiones. Además, se centra principalmente en los datos actuales dentro de la empresa y no hace referencia a datos históricos o diferentes a los de la organización.

Los sistemas OLTP suelen adoptar un modelo de datos entidad-relación, y una orientación a aplicaciones. Asimismo, los patrones de acceso de estos sistemas consisten en transacciones atómicas cortas.

2.2. OLAP

Esta tecnología se define como procesamiento analítico en línea (OLAP por sus siglas en inglés), y regularmente se utiliza para realizar análisis de volúmenes grandes de información, considerando diferentes perspectivas. Estos tipos de análisis se utilizan para determinar tendencias, patrones, relaciones, entre otros.

Generalmente un sistema OLAP esta orientado al mercado y se utiliza para análisis de datos por parte de los trabajadores, gerentes, ejecutivos y analistas. Asimismo, estos sistemas gestionan grandes cantidades de datos históricos y proporciona facilidades para resumirlos, agregarlos, almacenarlos y gestionarlos.

Por lo que las características detalladas, hacen que los datos sean utilizados de forma más fácil en una plataforma de toma de decisiones. Además, los sistemas OLAP adopta un diseño de bases de datos orientado a temas y sus esquemas de base de datos abarca varias versiones debido al proceso evolutivo de la organización.

De la misma forma, un sistema OLAP maneja información que se origina en diferentes organizaciones y se integra la información en varios almacenes de datos. Asimismo, por su gran volumen de información, los datos de estos sistemas se almacenan en múltiples medios de almacenamiento.

2.3. Ventajas y desventajas

Ventajas:

Entre las ventajas de la tecnología **OLAP** se encuentra el análisis multidimensional, que brinda la capacidad de analizar datos desde distintas dimensiones, permitiendo identificar patrones y tendencias. Además, de una consulta más flexible por parte de los usuarios y también poder agregar datos de una forma más eficientes para poder generar informes.

En el caso de la tecnología **OLTP** destacan en sus ventajas que se pueden manejar transacciones múltiples de forma simultanea. Asimismo, que en este tipo de sistema es más fácil garantizar la precision de los datos y que su tiempo de respuesta es muy rápido, lo que es requerido en entornos donde el tiempo es valioso.

Desventajas:

Entre las desventajas de la tecnología **OLAP** se encuentra el hecho que mantener y configurar estos sistemas, suelen ser más complejo. Asimismo, que la demora de actualización de estos sistemas puede demorar más tiempo y que no es un sistema optimo para transacciones que requieren procesamiento en tiempo real.

En el caso de la tecnología **OLTP** se tiene poca capacidad de análisis o consultas con grandes volúmenes de información. Además, que sus esquema más simple no permite realizar un análisis complejo y que esta tecnología puede necesitar más recursos de almacenamiento y desarrollar capacidad de procesamiento para poder optimizar las transacciones en tiempo real.

2.4. ¿Qué tecnología recomendaría en cada caso?

Los sistemas **OLAP** están diseñados para procesar grandes cantidades de datos rápidamente. Esto se logra regularmente a través del procesamiento distribuido y una arquitectura que está orientada a columnas. Bases de datos de nube recientes son sistemas con esta tecnología, las tres soluciones son: Amazon Redshift, Google BigQuery y Snowflake. Por lo que los sistemas OLAP son utilizadas muy comúnmente por equipos de análisis y data science por su velocidad, estabilidad y bajo costo mantenimiento.

Los sistemas **OLTP** están diseñados para manejar grandes cantidades de datos transaccionales originados por varios usuarios. Por lo general, esto lleva a cabo la forma de una base de datos orientada a filas. Muchas bases de datos tradicionales de esta tecnología son: Postgres, MySQL, entre otros.

Referencias

[1] Ralph Kimball, *The Data Warehouse Toolkit*, Wiley, (2002)

[2] Matt Palmer, *Understanding ETL*, O'REYLLY, (2023)

[3] Keboola, *A Guite to ETL vs ELT data pipelines*,(2022)

[4] B.Manuvannan, *A Novel Approach of Data warehouse OLTP and OLAP Tecnology for Supporting Management prospective*,International Journal of Advanced Research in Computer and Communication Engineering,(2017)