Visualización de datos

Extracción, transformación y carga de datos

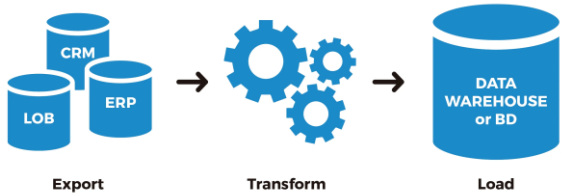
Contenido

Generalidades

¿Qué es ETL?
¿Qué características tiene un proceso ETL?
¿Para qué sirve hacer ETL?
¿Por qué es importante realizar ETL?

E(xtract) T(ransformation) L(oad)

ETL es un tipo de integración de datos que hace referencia a los tres pasos (extraer, transformar, cargar) que se utilizan para mezclar datos de múltiples fuentes.



Extraer, cargar, transformar (ELT) es un enfoque alterno pero relacionado diseñado para canalizar el procesamiento a la base de datos para mejorar el desempeño.

Características del proceso ETL

- Complejidad. Las empresas se pueden encontrar con grandes cantidades de datos almacenadas durante muchos años y generadas por sus distintos departamentos (financiero, ingeniería, marketing, ventas, etc.).

- Continuidad. Además, para realizar análisis precisos, es necesario mantener el Data Warehouse constantemente actualizado. Por esto, es importante que el proceso ETL se realice a intervalos regulares, detectando cambios en la información contenida en las fuentes, extraer los nuevos datos, transformarlos y cargarlos en el almacén de datos.

- Criticidad. Generalmente, ninguno de los datos que poseen las empresas viene por defecto en una forma que esté lista para usar para resolver los problemas de su negocio. Sin los procesos ETL, las empresas se encontrarían con una gran cantidad de datos que no pueden utilizar.

¿Para qué sirve hacer ETL?

- Migración de datos de una aplicación a otra (ej: cloud)
- Replicación de datos para copias de seguridad o análisis de redundancia
- Depositar los datos en un almacén de datos para ingerir, clasificar y transformarlos en business intelligence
- Recolección y fusión de datos desde proveedores o partners externos.
- Integración de nuevas fuentes de datos como social media, videos, dispositivos conectados a internet de las cosas, entre otras.

¿Por qué es importante un proceso ETL?

- Cuando se utiliza con un almacén de datos empresarial (datos en reposo), ETL provee profundo contenido histórico para la empresa.
- Proporcionando una vista consolidada, ETL facilita a los usuarios de negocios analizar y generar reportes sobre datos relevantes para sus iniciativas.
- ETL puede mejorar la productividad de los profesionales de los datos porque codifica y reutiliza procesos que mueven datos sin requerir habilidades técnicas para escribir código o scripts.

¿Por qué es importante un proceso ETL?

- ETL ha evolucionado para satisfacer requisitos de integración emergentes para cosas como los datos transmitidos por streaming.
- Las organizaciones necesitan ETL y ELT para conjuntar datos, mantener la precisión y proporcionar el recurso de auditoría que suele requerirse en los almacenes, reportes y análisis de datos.
- Permitir extraer y consolidar datos de múltiples fuentes.

Funcionamiento del ETL

¿Cómo funciona la extracción, transformación y carga de datos?

Extracción

El objetivo de un proceso ETL es producir datos limpios y accesibles que puedan utilizarse para analíticas u operaciones comerciales. Los datos en bruto deben extraerse de una variedad de fuentes, por ejemplo:

- Bases de datos existentes
- Registros de actividad como el tráfico de red, informes de errores, etc.
- Rendimiento y anomalías de aplicaciones
- Incidencias de seguridad
- Otras actividades transaccionales que deben comunicarse para dar cumplimiento normativo

Los datos extraídos en ocasiones se transfieren a otro destino como por ejemplo un data lake o un almacén de datos.

Proceso de extracción

- Extraer los datos desde los sistemas de origen.

- Analizar los datos extraídos obteniendo un chequeo.

- Interpretar este chequeo para verificar que los datos extraídos cumplen la pauta o estructura que se esperaba. Si no fuese así, los datos deberían ser rechazados.

- Convertir los datos a un formato preparado para iniciar el proceso de transformación

Transformación

La fase de transformación de los procesos de ETL aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Para potenciar su pragmatismo y eficacia, hay que asegurarse de que sean:

- Declarativas.
- Independientes.
- Claras.
- Inteligibles.
- Con una finalidad útil para el negocio.

Proceso de transformación

La transformación se efectúa mediante una serie de normas y reglamentos que se esbozan. Estos son algunos de los estándares que garantizan la calidad de datos y su accesibilidad durante esta fase:

- Normalización: definir qué datos entrarán en juego, cómo se formatearán y almacenarán, y otras consideraciones básicas que definirán las etapas sucesivas.

- Eliminación de duplicados: notificar los duplicados a los administradores de datos; excluyendo y/o eliminando los datos redundantes.

Proceso de transformación

- Verificación: ejecutar comprobaciones automatizadas para cotejar información similar, como tiempos de transacción o registros de acceso.

- Clasificación: maximizar la eficiencia de los almacenes de datos agrupando y clasificando elementos como los datos en bruto, audios, archivos multimedia y otros objetos en categorías.

- Las demás tareas las define usted y las configura para que se ejecuten automáticamente.

Carga

La última fase de un proceso de ETL típico es la carga de esos datos extraídos y transformados a su nuevo destino. Existen dos vías habituales de cargar los datos a un almacén de datos: la carga completa y la carga incremental.

- Carga completa: esta manera de cargar los datos consiste en realizar un resumen de todas las transacciones y transportar el resultado como una única transacción hacia el data warehouse, almacenando un valor calculado que consistirá típicamente en un sumatorio o un promedio de la magnitud considerada. Es la forma más sencilla y común de llevar a cabo el proceso de carga.

Carga

La última fase de un proceso de ETL típico es la carga de esos datos extraídos y transformados a su nuevo destino. Existen dos vías habituales de cargar los datos a un almacén de datos: la carga completa y la carga incremental.

- Carga incremental: Para ello se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo o diferentes niveles jerárquicos en alguna o varias de las dimensiones de la magnitud almacenada (por ejemplo, totales diarios, totales semanales, totales mensuales, etc.). Este proceso sería el más recomendable en los casos en que se busque mantener varios niveles de granularidad.

Aplicando un ETL en Python

Primeros pasos

Definir:

- ¿De qué fuente se va a extraer la información?¿Motor de base de datos, API?
- ¿Son datos estructurados, no estructurados?
- ¿Existe diccionario de datos o hay metadatos de la información?
- ¿Qué herramienta se usará para realizar la extracción y el cargue de la información?

Obtención de los datos

La información es extraída desde todas nuestras fuentes de datos, sean estas bases de datos relacionales, XML, o ficheros no estructurados.

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

SQL

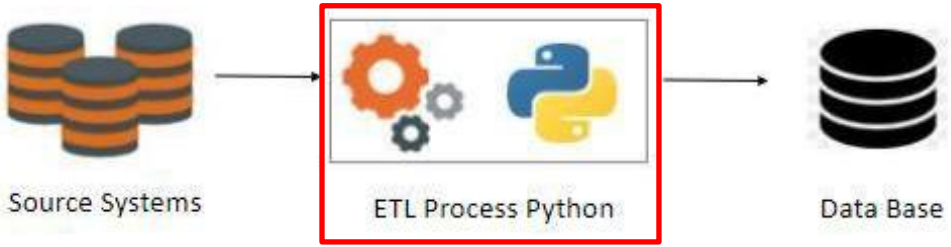
El lenguaje de consulta estructurado es el método más común para acceder a y transformar datos en una base de datos.

NoSQL

El lenguaje de consulta no estructurado surge de la necesidad de procesar información que no puede ser representada únicamente en tablas.

Transformaciones, reglas de negocios y adaptadores

Después de extraer datos, ETL utiliza reglas de negocios para transformar los datos en nuevos formatos. Los datos transformados se cargan después en el destino.



Transformaciones, reglas de negocios y adaptadores

Frameworks y librerías

- Una librería es una colección de funciones y objetos de ayuda para el proceso de desarrollo de software.
- Un framework puede consistir en una o más librerías, todas orientadas a un propósito común.

https://www.xplenty.com/blog/top-etl-python-frameworks/

Scripts

ETL es un método de automatización de los scripts (conjunto de instrucciones) que se ejecutan detrás de escena para mover y transformar datos.

Antes de que se integren los datos, a menudo se crea un área de preparación donde se pueden depurar datos, se pueden estandarizar valores de datos.

Gestión de datos maestros

Es el proceso de conjuntar datos para crear una vista única de los datos a través de múltiples fuentes.

Incluye capacidades ETL y de integración de datos para combinar los datos y crear un “registro de oro” o “el mejor registro”.

Bibliografía

- https://www.talend.com/es/resources/what-is-etl/
- https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/qu-son-los-procesos-etl
- https://www.sas.com/es_co/insights/data-management/what-is-etl.html
- https://blog.mdcloud.es/que-es-etl-extraccion-transformacion-y-carga/
- https://www.xplenty.com/blog/top-etl-python-frameworks/

Gracias!