Extracción, transformación y carga de datos (ETL)

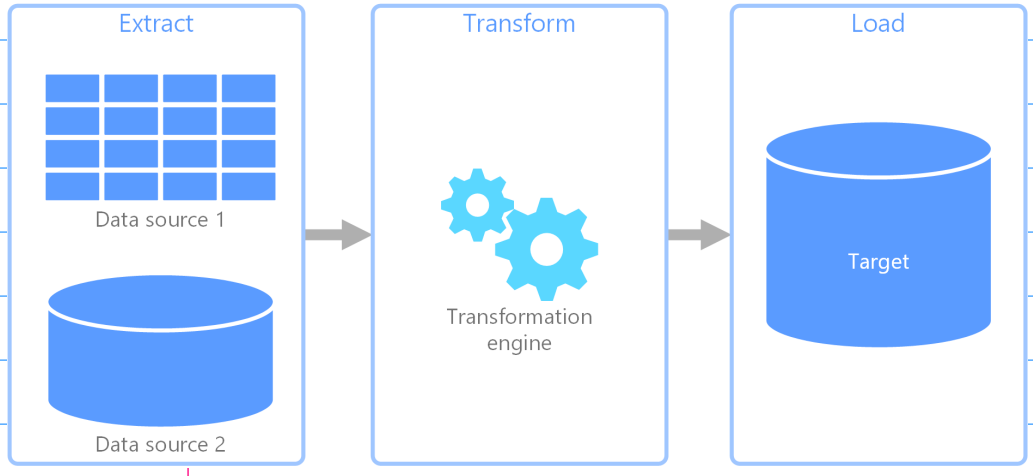
Un problema habitual al que se enfrentan las organizaciones es cómo recopilar datos de varios orígenes, en varios formatos. A continuación, tendrá que moverlos a uno o varios almacenes de datos. Es posible que el destino no sea el mismo tipo de almacén de datos que el origen. A menudo el formato es diferente, o bien es necesario dar forma a los datos o limpiarlos antes de cargarlos en el destino final.

Con los años se han desarrollado varias herramientas, servicios y procesos para ayudarle a afrontar estos desafíos. Independientemente del proceso que se utilice, hay una necesidad común de coordinar el trabajo y aplicar cierto nivel de transformación de datos en la canalización de datos. Las secciones siguientes resaltan los métodos más habituales utilizados para realizar estas tareas.Un problema habitual al que se enfrentan las organizaciones es cómo recopilar datos de varios orígenes, en varios formatos. A continuación, tendrá que moverlos a uno o varios almacenes de datos. Es posible que el destino no sea el mismo tipo de almacén de datos que el origen. A menudo el formato es diferente, o bien es necesario dar forma a los datos o limpiarlos antes de cargarlos en el destino final.

Con los años se han desarrollado varias herramientas, servicios y procesos para ayudarle a afrontar estos desafíos. Independientemente del proceso que se utilice, hay una necesidad común de coordinar el trabajo y aplicar cierto nivel de transformación de datos en la canalización de datos. Las secciones siguientes resaltan los métodos más habituales utilizados para realizar estas tareas.

Proceso de extracción, transformación y carga (ETL)

Extracción, transformación y carga (ETL) es una canalización de datos que se usa para recopilar datos de varios orígenes. A continuación, transforma los datos según las reglas de negocio y los carga en un almacén de datos de destino. El trabajo de transformación en ETL tiene lugar en un motor especializado y, a menudo, implica el uso de tablas de almacenamiento provisional para conservar los datos temporalmente a medida que estos se transforman y, finalmente, se cargan en su destino.



La transformación de datos que tiene lugar a menudo conlleva varias operaciones como filtrado, ordenación, agregación, combinación de datos, limpieza de datos, desduplicación y validación de datos.

Frecuentemente, las tres fases del proceso ETL se ejecutan en paralelo para ahorrar tiempo. Por ejemplo, mientras se extraen datos, puede que esté funcionando un proceso de transformación sobre los datos ya recibidos y de preparación para la carga, y puede que empiece a funcionar un proceso de carga sobre los datos preparados, en lugar de tener que esperar a que termine todo el proceso de extracción.

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. Se utiliza a menudo para construir un almacén de datos. Durante este proceso, los datos se toman (extraen) de un sistema de origen, se convierten (transforman) en un formato que se puede almacenar y se almacenan (cargan) en un data warehouse u otro sistema. Extraer, cargar, transformar (ELT) es un enfoque alterno pero relacionado diseñado para canalizar el procesamiento a la base de datos para mejorar el desempeño.

Las empresas han confiado en el proceso ETL por muchos años para obtener una vista consolidada de los datos que que dé lugar a mejores decisiones de negocios. Hoy día, este método de integración de datos de múltiples sistemas y fuentes sigue siendo un componente central de la caja de herramientas de integración de datos de una organización.

->Software de integración de datos de SAS
El software de integración de datos de SAS distribuye tareas de integración en cualquier plataforma y se conecta virtualmente a cualquier almacén de datos de origen o destino.

Los procesos ETL (extract, transform y load) son aquellos mediante los cuales se extrae información de uno o varios orígenes de datos. Ésta se transforma para adaptarla a las necesidades del negocio y posteriormente se carga en un sitio compartido para su consulta por todas las partes interesadas. Una de las características más importantes del sistema ETL es que permite integrar en un entorno homogéneo aquellos datos de orígenes heterogéneos. Sin embargo, la configuración del flujo de trabajo tiene un cierto nivel de complejidad y un proceso ETL mal diseñado puede causar problemas operacionales de un coste muy elevado.
-> Fases
1. Extracción (extract)
2. Transformación (transform)
3. Carga (load)