Ingesta de Datos

Deborah Famadas Rodriguez

Universidad de la Habana

18 de noviembre de 2024

Ingesta de datos

- Posterior a la captura, se realiza la ingesta de datos; este es un proceso fundamental en cualquier pipeline de datos. Su propósito es recopilar e importar datos desde diversas fuentes, como sitios web, aplicaciones o plataformas de terceros, y luego cargarlos en otra capa para ser utilizados.
- Durante esta etapa, los datos se **limpian** y **almacenan**, lo que permite su posterior procesamiento y análisis.

Ingesta de datos

El proceso de ingestión de datos consta de dos elementos. Vamos a entenderlos uno por uno.

- **Fuente**: El objetivo de la ingestión de datos es recopilar y cargar datos. Por lo tanto, uno de los elementos básicos de la ingestión de datos es la fuente. Aquí, fuente se refiere a cualquier al origen de los datos capturados.

Fuentes de datos

- **Sistemas financieros**
- **Proveedores de datos de terceros**
- **Plataformas de redes sociales**
- **Dispositivos IoT**
- **Aplicaciones SaaS**
- **Aplicaciones empresariales locales**

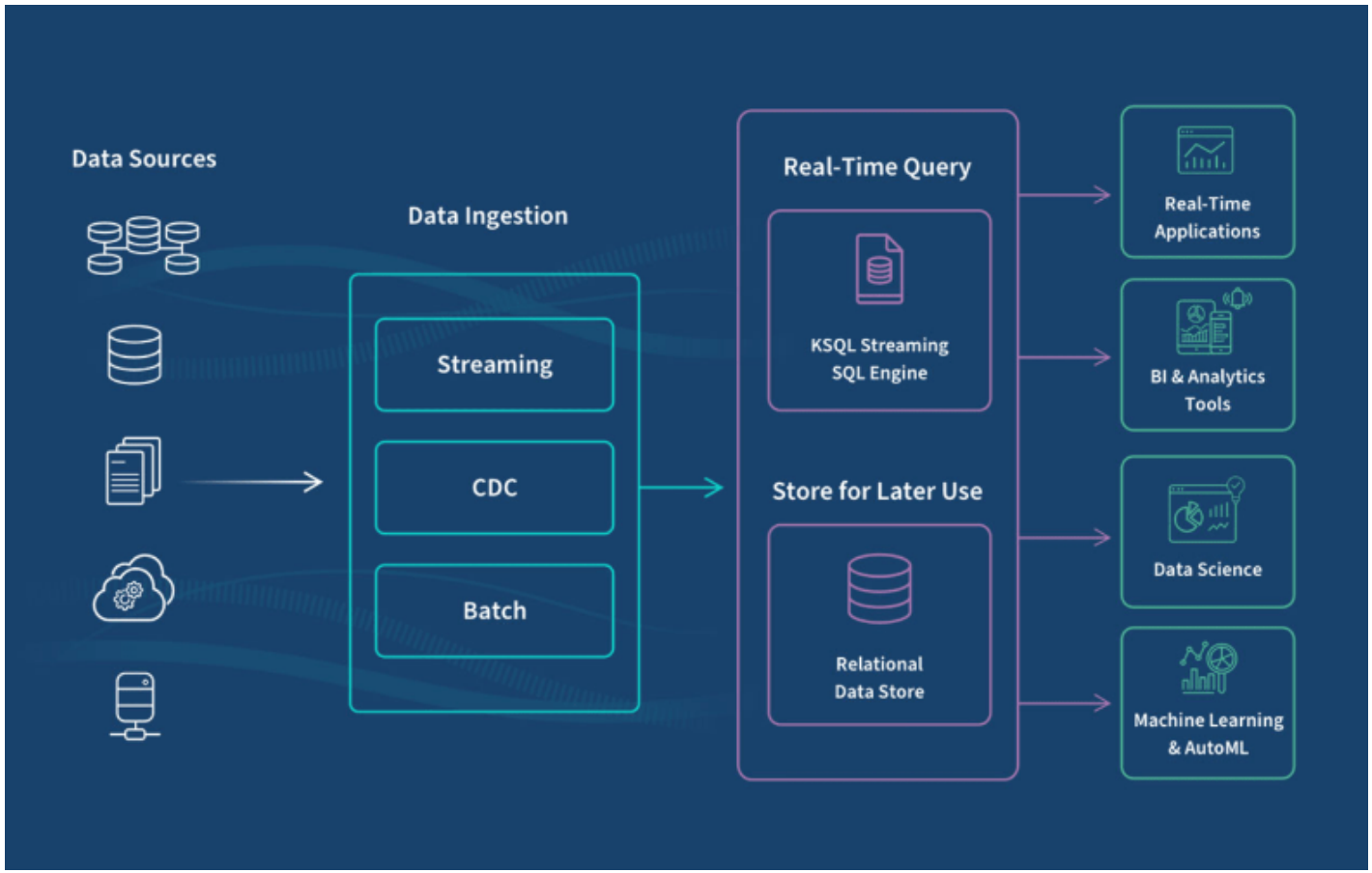
Ingesta de datos

El proceso de ingestión de datos consta de dos elementos. Vamos a entenderlos uno por uno.

- **Fuente**
- **Destino**: Dado que estas fuentes contienen datos estructurados y no estructurados una vez ingeridos, los datos pueden almacenarse en lagos de datos, almacenes de datos, data lakehouses, data marts, bases de datos relacionales y sistemas de almacenamiento de documentos.

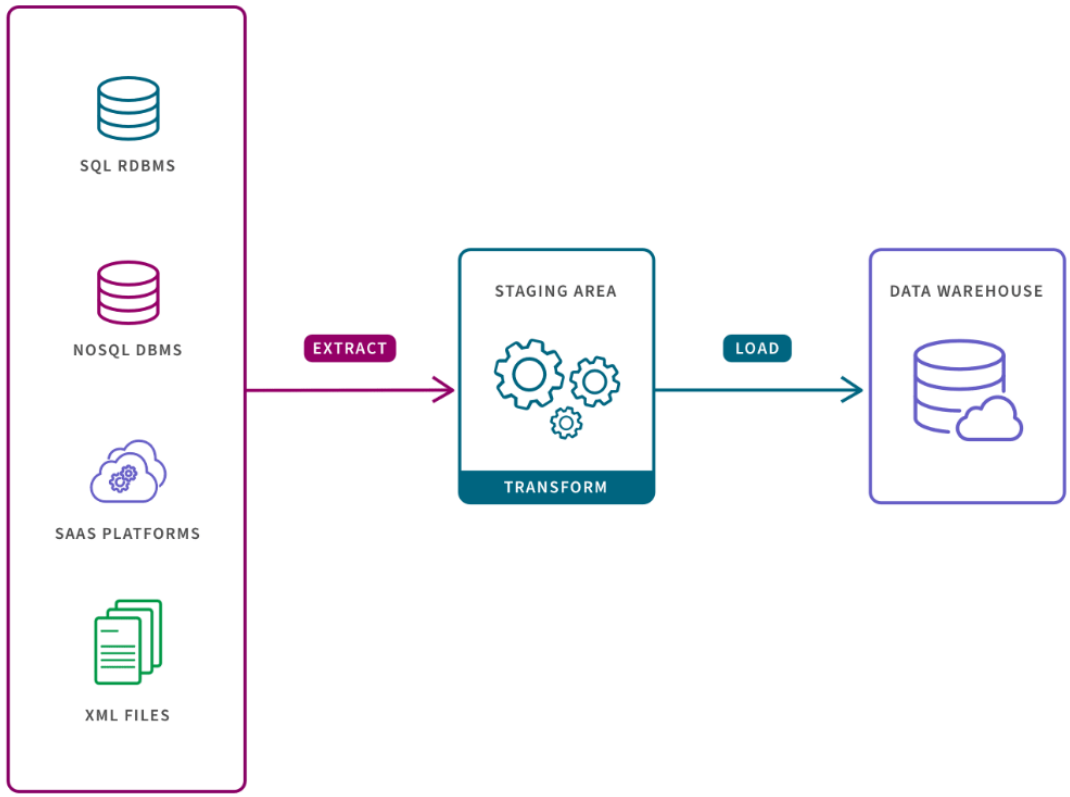
Tipos de ingesta de datos

La ingesta de datos puede realizarse de muchas maneras. Existen principalmente tres formas de realizar la ingesta de datos: en tiempo real, or lotes y lambda (combinación de ambos).



Preprocesamiento por lotes (Batch processing)

El procesamiento por lotes transfiere datos históricos a un sistema de destino a intervalos programados, activados automáticamente, ordenados lógicamente, iniciados por consultas o provocados por eventos de la aplicación.



Preprocesamiento por microlotes (microbatching)

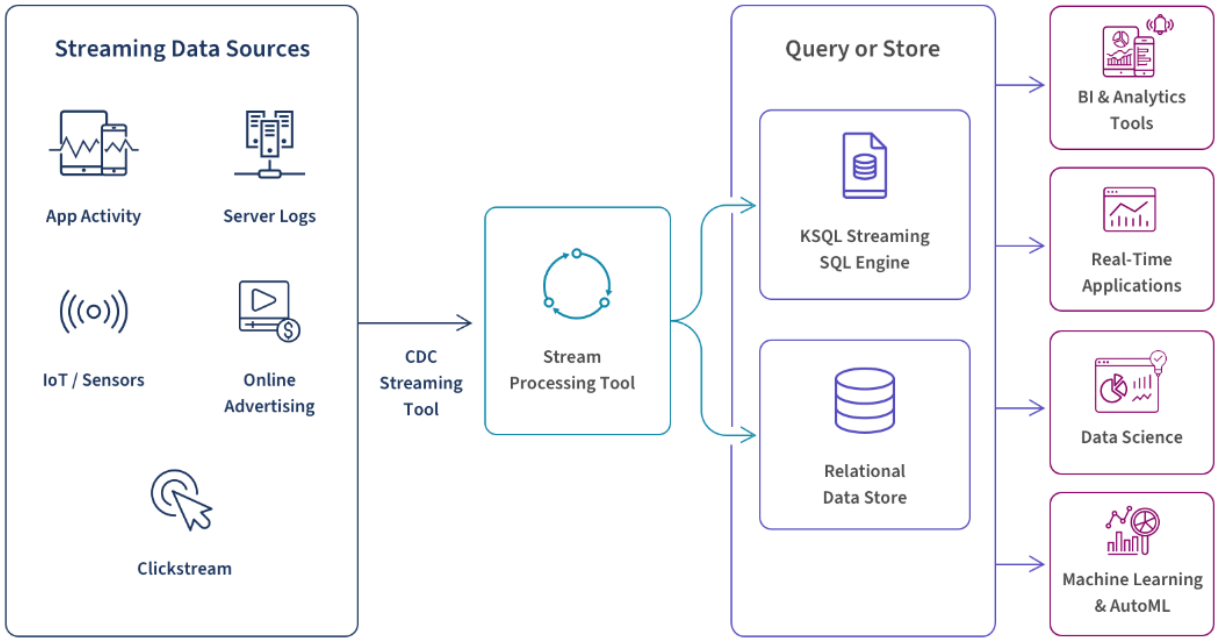
Microbatching

Es una variante del procesamiento por lotes, ofrece resultados similares a los datos en tiempo real, atendiendo a diversas necesidades analíticas.

Si necesita datos puntuales, casi en tiempo real, pero la arquitectura que se tiene es de integración de datos y le impide emplear el procesamiento de flujos, la microloteado es una buena opción a tener en cuenta. En el microbatching, los datos se dividen en grupos y se ingieren en pequeños incrementos, simulando el streaming en tiempo real.

Preprocesamiento en tiempo real

El procesamiento en tiempo real, también conocido como procesamiento de flujo, es un método que permite mover datos desde la fuente al destino inmediatamente después de ser reconocidos por la capa de ingesta en la canalización de datos.



Peprocesamiento en tiempo real

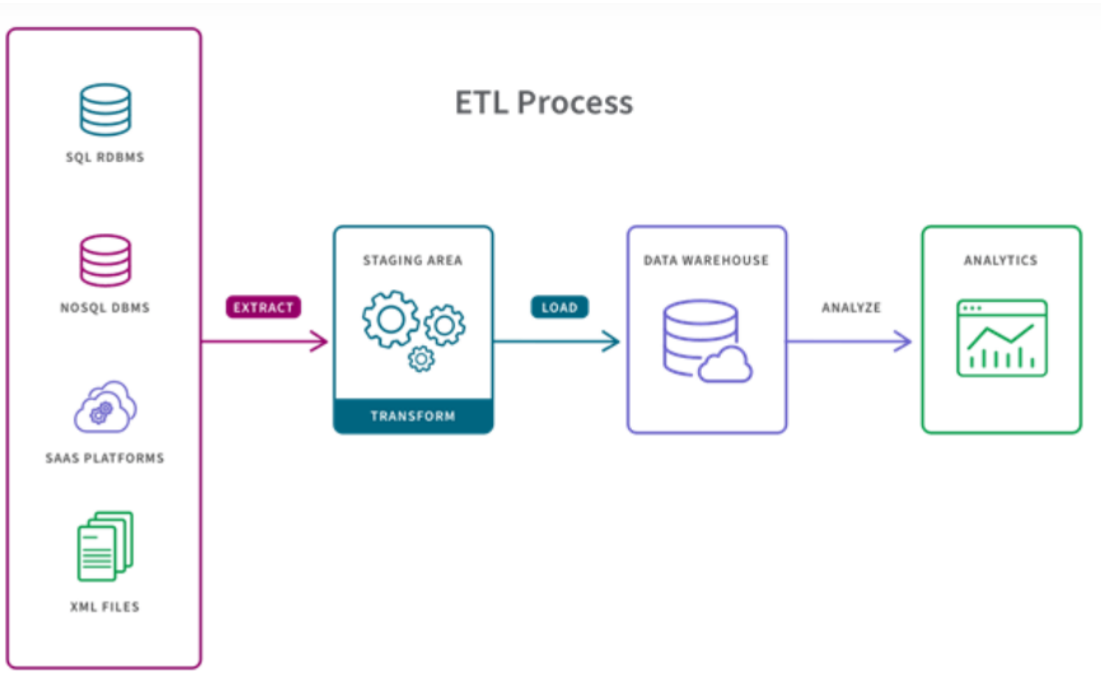
- Este método de flujo de datos suele utilizarse en situaciones en las que se requiere información actualizada al minuto, como el análisis en tiempo real, la detección de fraudes y la supervisión.
- Una forma de integración de datos en tiempo real, la captura de datos de cambios (CDC), aplica las actualizaciones realizadas en los datos de los sistemas de origen a los almacenes de datos y otros repositorios. A continuación, estos cambios pueden aplicarse a otro repositorio de datos o ponerse a disposición en un formato consumible por ETL, por ejemplo, u otros tipos de herramientas de integración de datos.

Tecnicas utilizadas en ingesta de datos

Un problema habitual al que se enfrentan las organizaciones es cómo recopilar datos de múltiples fuentes, en múltiples formatos. Luego hay que trasladarlos a uno o varios almacenes de datos. El destino puede no ser el mismo tipo de almacén de datos que la fuente. A menudo, el formato es diferente o es necesario dar forma o limpiar los datos antes de cargarlos en su destino final.

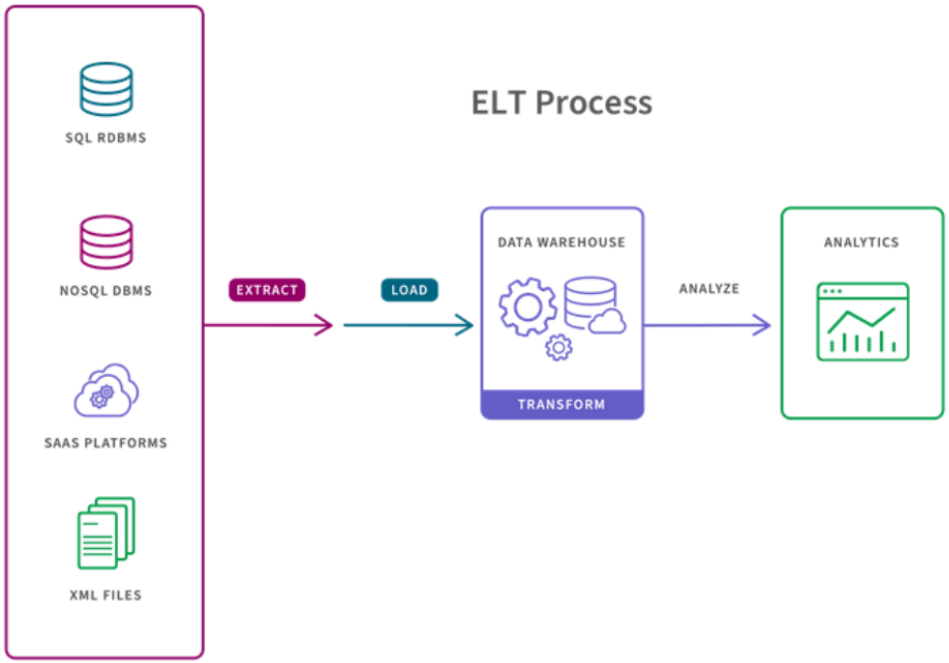
ETL vs ELT

Extraer, transformar, cargar (ETL) es una tecnica de data pipeline que se utiliza para recopilar datos de diversas fuentes. Luego se transforman los datos y los carga en un almacén de datos de destino.



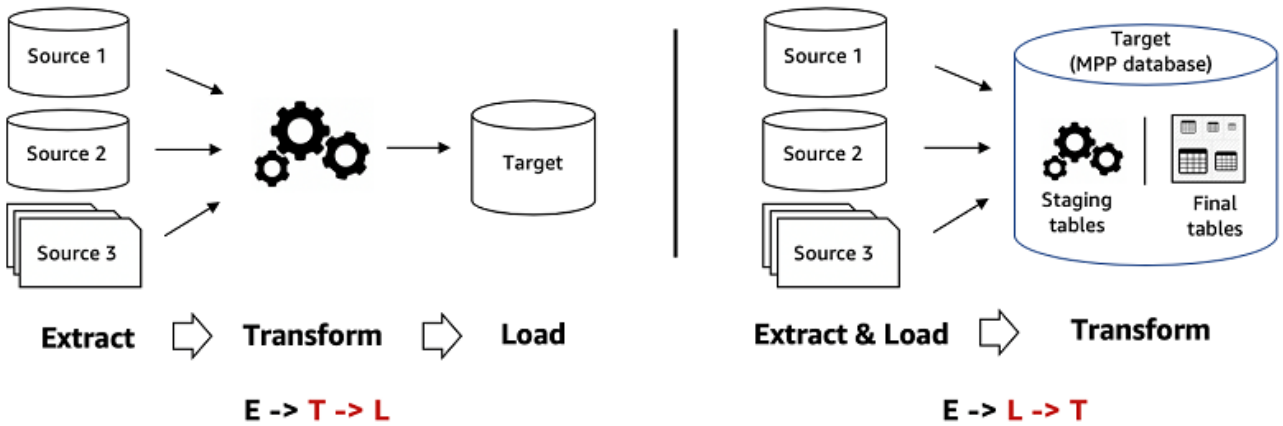
ETL vs ELT

Extraer, cargar, transformar (ELT) difiere de ETL únicamente en dónde tiene lugar la transformación. En el pipeline de ELT, la transformación se produce en el almacén de datos de destino. En lugar de utilizar un motor de transformación independiente, se utilizan las capacidades de procesamiento del almacén de datos de destino para transformar los datos. Esto simplifica la arquitectura al eliminar el motor de transformación del canal.



ETL vs ELT

En ETL, los datos se extraen primero de las fuentes de origen, luego se transforman para cumplir con el formato y la estructura del sistema de destino y, finalmente, se cargan en el almacenamiento (generalmente un data warehouse). Mientras que en ELT, los datos se extraen y luego se cargan directamente en el sistema de almacenamiento (generalmente un data lake o una base de datos en la nube), y la transformación ocurre después, dentro del mismo almacenamiento.



Data Ingestion Pipeline

- Paso 1 - Identificación de las fuentes de datos

Data Ingestion Pipeline

- Paso 1 - Identificación de las fuentes de datos
- Paso 2 - Identificación del sistema de destino

Data Ingestion Pipeline

- Paso 1 - Identificación de las fuentes de datos
- Paso 2 - Identificación del sistema de destino
- Paso 3 - Identificación del método de ingestión de datos

Data Ingestion Pipeline

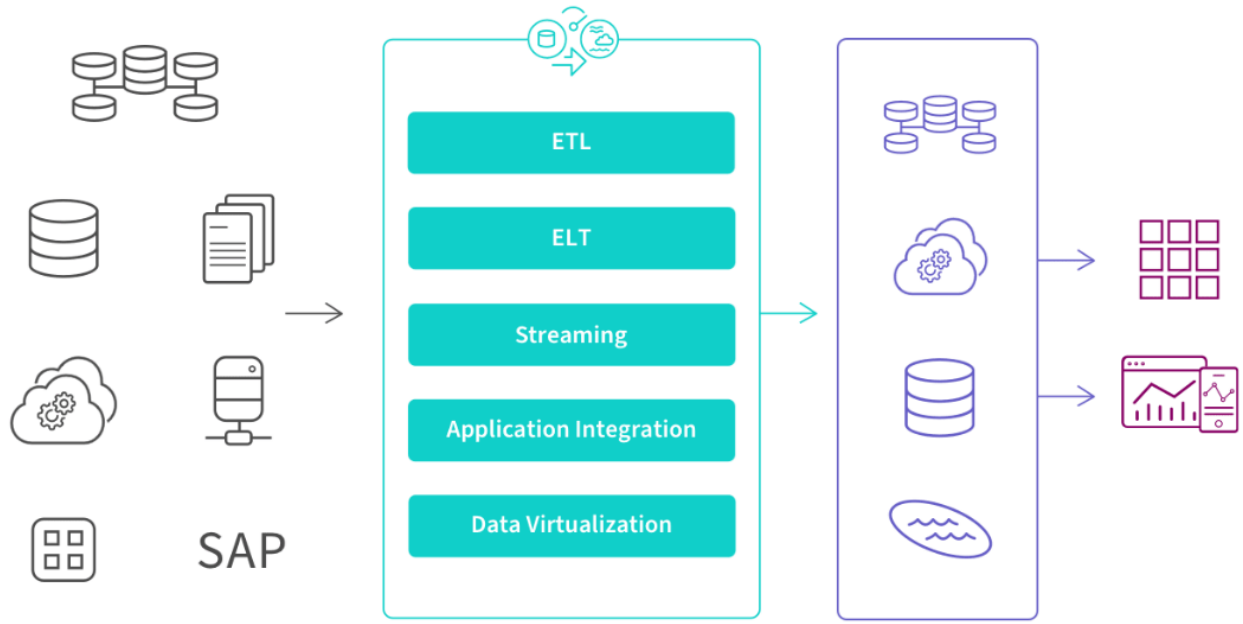
- Paso 1 - Identificación de las fuentes de datos
- Paso 2 - Identificación del sistema de destino
- Paso 3 - Identificación del método de ingestión de datos
- Paso 4 - Diseño del proceso de integración de datos

Data Ingestion Pipeline

- Paso 1 - Identificación de las fuentes de datos
- Paso 2 - Identificación del sistema de destino
- Paso 3 - Identificación del método de ingestión de datos
- Paso 4 - Diseño del proceso de integración de datos
- Paso 5 - Supervisión y optimización

Ingesta de datos vs Integración de datos

La integración de datos se refiere al proceso de reunir datos de múltiples fuentes en toda una organización para proporcionar un conjunto de datos completo, preciso y actualizado para BI, análisis de datos y otras aplicaciones y procesos empresariales. Incluye la replicación, la ingesta y la transformación de datos para combinar diferentes tipos de datos en formatos estandarizados que se almacenarán en un repositorio de destino, como un almacén de datos, un lago de datos o un data lakehouse.



Fin


WE SHOULD
TAKE THE DATA
AND MOVE IT
SOMEWHERE ELSE
imgflip.com