DIGITALHOUSE

Etapas de ETL

Certified Tech Developer
The Ultimate Degree

01 Etapa de extracción

Extracción

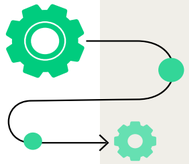
Características de esta etapa:

- Tiene que ver con la recolección de los sistemas de origen, conocida también como Data Sources o Sistemas Operacionales.
- Puede ser realizada por el propio sistema de origen, fuente de los datos, o por cualquier herramienta ETL.
- En la propia extracción pueden existir reglas de negocio aplicadas. Por ejemplo, extraer del sistema de registro de clientes aquellos datos que indiquen más de 60 años de edad y/o que posean un ingreso por encima de un valor x.

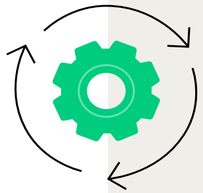
Observemos los tipos más frecuentes de extracción de datos:

Extracción por lote/total – La herramienta de ETL siempre extrae todos los datos que se ajustan a las reglas de extracción.

Actualización incremental – La herramienta ETL puede comparar lo que ya ha extraído con la fuente de datos. Por lo tanto, la misma puede identificar y extraer sólo lo nuevo.



Actualización automática (Push / Webhook / Callback / HTTP Request) – El sistema de origen puede notificar cuando se han modificado los registros. Esto genera que la herramienta extraiga únicamente información nueva, es decir, la que ha sido modificada.



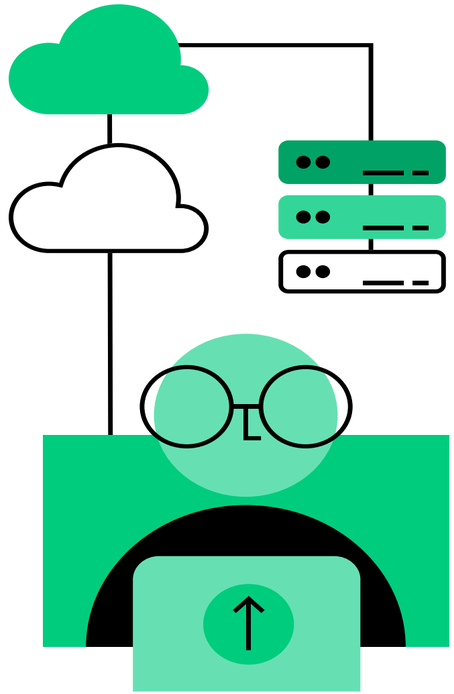
Extracción continua (Streaming de dados) – La fuente de datos envía constantemente nuevos datos. La herramienta de extracción debe soportar la captura de este volumen continuo.

02

Etapa de transformación

Transformación

- El proceso de transformación de datos está compuesto por varias etapas: estandarización, limpieza, calidad, consolidación e integridad.
- El mayor peso del proceso de ETL está en la etapa de transformación, debido al gran número de fases requeridas.
- Dependiendo del lenguaje o de la herramienta utilizada, el proceso de transformación puede ser muy rápido o extremadamente lento.



Transformaciones básicas

- Limpieza datos
- Estandarización de formatos
- Estandarización de domínios
- Remoción de duplicados
- Filtrado de datos

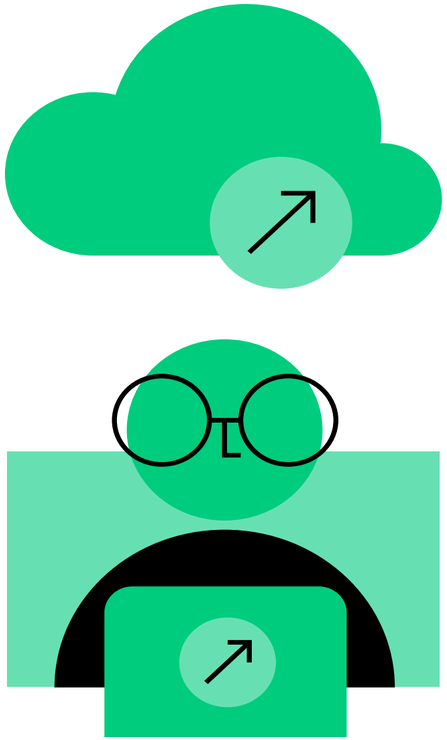
Transformaciones avanzadas

- Unir archivos/tablas
- División de archivos/tablas
- Campos de cálculo
- Enriquecimento datos/campos
- División o consolidación de campos
- Unidades de conversión
- Transformación de llaves
- Modificación de encoding

03 Etapa de carga

Carga

- El proceso de carga es lo más simple del flujo. Se encuentra condicionado y debe orientarse al tipo de archivos o datos de destino que puede trabajar.
- Es extremadamente estratégico. Esto se debe a que convertir un archivo de origen a un nuevo modelo, más pequeño y con mayor tasa de compresión, ayuda mucho a mantener el entorno Big Data.
- Cuando trabajamos con grandes volúmenes de datos, estos son almacenados en un **datawarehouse, o en un data lake**.
- En cambio si trabajamos sobre conjuntos de datos más pequeños, podemos “cargar” el resultado del ETL en el repositorio de la herramienta de Data Analytics con la que estemos trabajando en la visualización de datos, por ejemplo Power BI.



04 Conclusiones

Conclusiones

En suma, los sistemas ETL:

- Se conectan a diversas fuentes de información.
- Son robustos.
- Disponen de sistemas de programación de actualización de datos.
- Validan la consistencia de las operaciones de carga de datos.
- Entregan la información adquirida y procesada a un almacén de datos.
- Son muy importantes para poder consolidar y unificar datos que vienen de muchas fuentes de información que tienen estructuras diferentes entre sí.