Transformación y carga

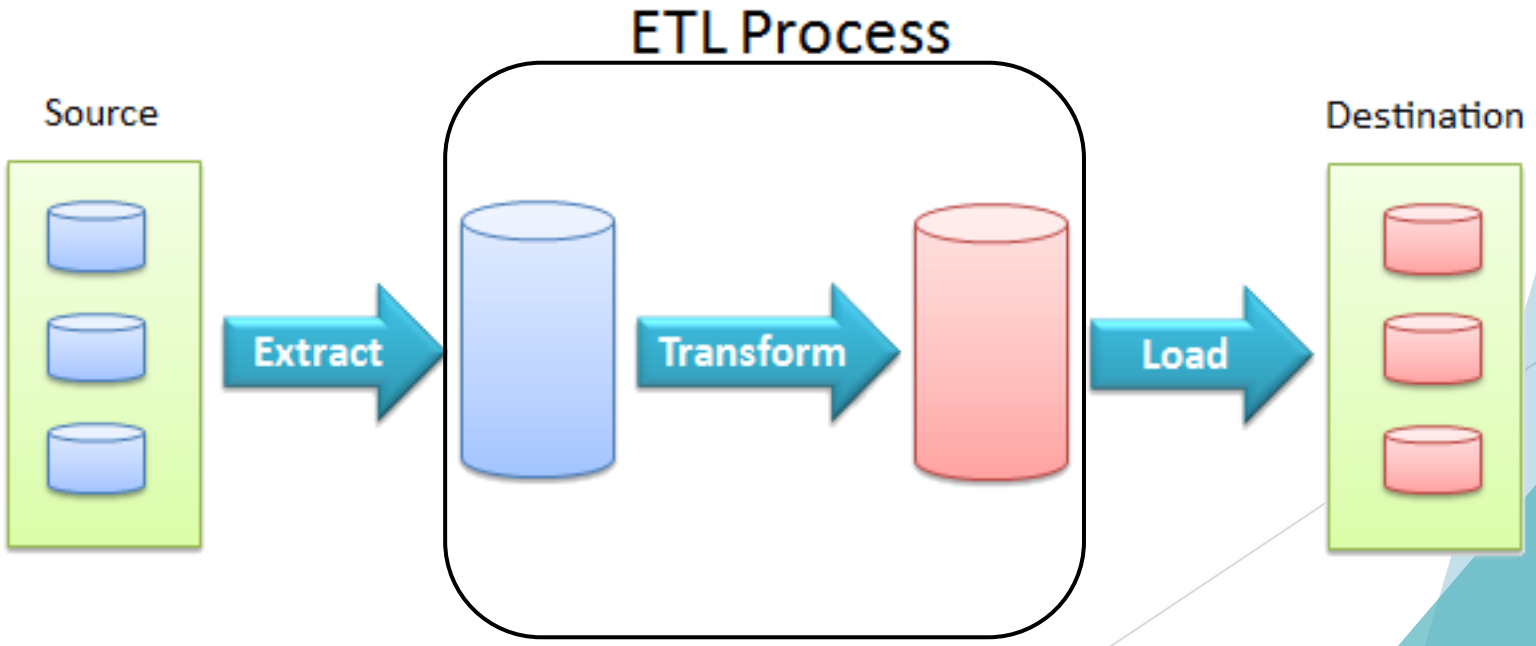
Big Data Aplicado

Dr. Francisco E. Cabrera

Transformación

Transformación

- Segunda fase de un proceso de ETL
- Esta fase consiste principalmente en aplicar funciones sobre los datos extraídos.



Transformación

- Las funciones aplicadas en el proceso de transformación dependen:
  - De las características de los datos obtenidos.
  - De los análisis que queramos realizar.
  - De la relación coste/beneficio de realizar estas transformaciones antes de empezar con el análisis o delegar esta tarea a la fase de análisis.

Transformación

El objetivo de esta fase es garantizar:

- Que los datos sean correctos.
- Que los datos no sean ambiguos.
- Que los datos sean consistentes.
- Que los datos sean completos.

En definitiva: La calidad de los datos.

Transformaciones Básicas

Transformaciones básicas

Se trata de transformaciones que no suelen requerir mucho procesamiento.

- Selección.
- Cambios de formato.
- Restructuración de claves.

Selección

Escoger los campos que puedan resultar relevantes para los análisis posteriores.

- Descartar campos que siempre están vacíos o tienen un valor constante.
- Descartar campos que no puedan aportar información para el dominio del problema o que no tengan significado útil para nosotros.
- Descartar los elementos duplicados.

Cambios de formato

Tener todos los datos en un mismo formato acelera el proceso de análisis posterior, por ejemplo:

- Formatos de fecha y hora:
  - Las fechas pueden venir como texto en forma de “MM/dd/yyyy HH:mm” y “dd/MM/yyyy HH:mm:ss” pero quedar almacenados como objeto DATETIME.
- Codificación de caracteres:
  - Pasar de ASCII a UTF-8 o viceversa.

Reestructuración de claves

Cuando se utilizan índices para realizar consultas, es conveniente la generación de identificadores criptográficos.

- Generar claves con valores compuestos para evitar la inserción de duplicados en la base de datos de destino.
- Ej: $cid = f(d['text'] + d['title'] + str(d['date']) + salt)$

  donde $f(x)$ es un hash criptográfico.

Transformaciones avanzadas

Transformaciones avanzadas

Se trata de operaciones que requieren un mayor procesamiento.

- Cálculo de valores derivados.
- Combinación de datos de diversas fuentes.
- Separación de datos.
- Agregación de datos.
- Validación de datos.
- Resumen de datos.
- Integración de datos.

Cálculo de valores derivados

Para agilizar el posterior proceso de análisis, se pueden obtener valores resultantes de aplicar una función a los datos originales.

Por ejemplo:

- Operaciones con números:
  - Calcular un valor en una escala que dependa de uno o varios campos.
- Operaciones para simplificar el análisis de textos:
  - Obtener una serie de tags a partir del texto en lenguaje natural.
  - Obtener una valoración de sentimiento de los textos.
  - Utilizar un LLM para detectar ciertos patrones en un texto.

Combinación de datos de diversas fuentes

La información necesaria para realizar el proceso de análisis puede provenir de varias fuentes, se puede unificar en esta fase. Por ejemplo:

- Se puede cotejar la información de un usuario con los datos ya existentes para recuperar información adicional.
- Se pueden realizar cruces entre información proveniente de varias llamadas a APIs diferentes.

Separación de datos

Se trata del cálculo de valores resultantes de dividir datos en distintos campos.

- La información no estructurada en ocasiones puede venir con un solo campo, por lo que es necesario separarla en campos para el análisis.
  - Ej: Puede llegar un log como texto con el siguiente formato:

    “WARNING: <FileNotFound> No se ha podido abrir el archivo, 02/07/2019.”

    - Que se puede separar en los campos: “log_level”, “error_type”, “”,“error_desc” y “date_reported”.

Este proceso es bastante habitual cuando nuestra fuente de datos se consulta por medio de scraping.

Agregaciones de datos

Se trata del cálculo de valores resultantes de combinar datos.

- Si todos los análisis se van a realizar cada cierto periodo de tiempo, se puede guardar la información agregada en el almacén de datos.
  - Ej: Calcular la emoción promedia por hora en lugar de almacenar la emoción individual para cada comunicación.

Validación de datos

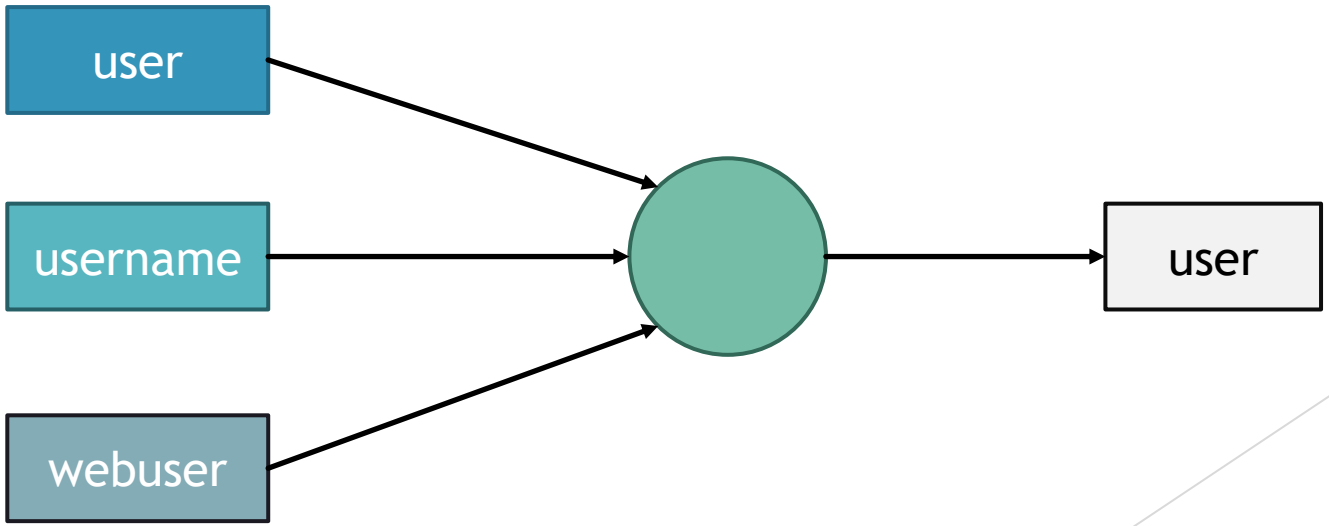
Asegurar el cumplimiento de una serie de reglas predefinidas. Las reglas de validación mas habituales suelen ser:

- Comprobar la presencia de datos obligatorios.
- Comprobar que no hay datos incongruentes.
- Comprobar los datos que deban ser claves únicas.
- Comprobar con los datos ya existentes en el almacén de datos.

Integración de los datos

Dotar a cada elemento un identificador y una descripción estándar.

- De esta manera, en el análisis los datos son recuperados con los mismos parámetros independientemente de la fuente de origen.



Resumen de datos

Es un caso especial de agregación de datos, se trata de almacenar la misma información en múltiples niveles.

De esta forma, cuando realizamos tareas de análisis, la búsqueda de la información es mas rápida.

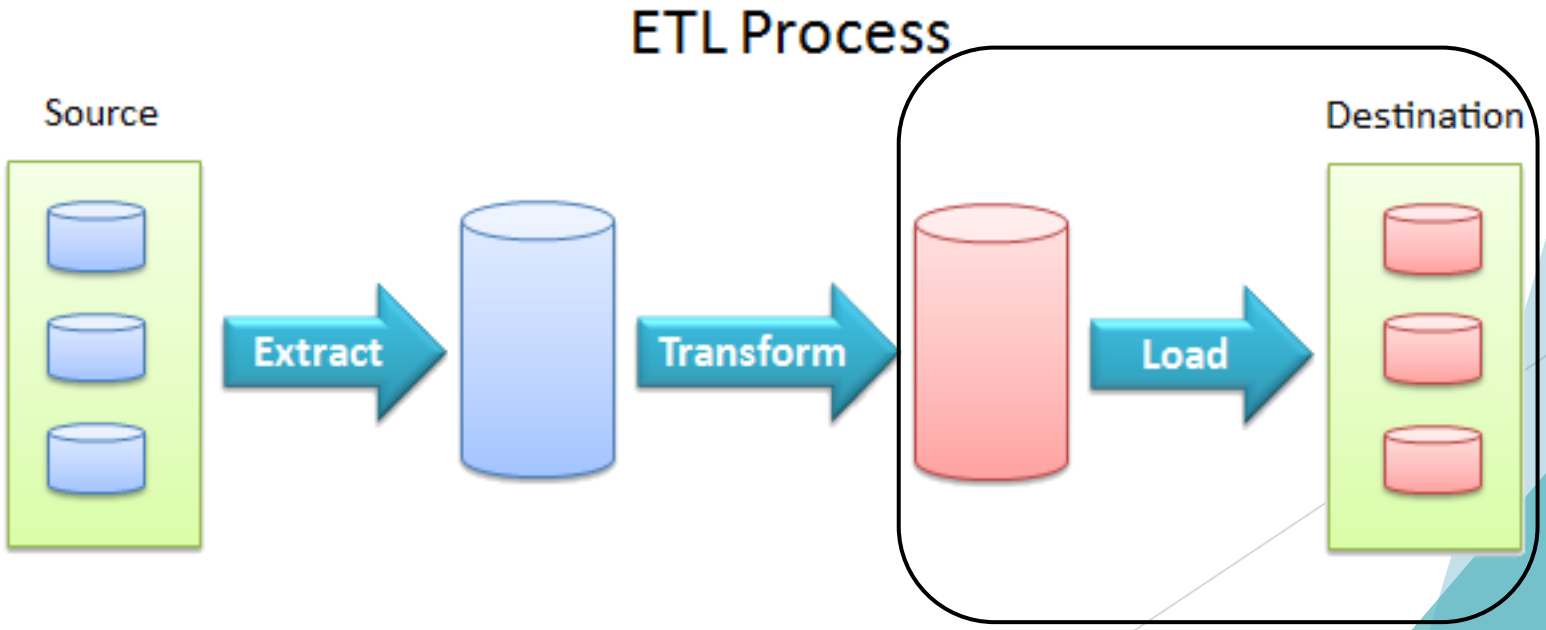
Por ejemplo:

- Almacenar datos agregados de la siguiente forma:
  - Por hora para la última semana.
  - Por día para los últimos 6 meses.
  - Por mes para datos anteriores a 6 meses.

Carga

Carga

- Tercera parte de un proceso ETL
  - Segunda fase en un proceso ELT
- En esta fase se guardan en su destino los datos transformados.



Carga

El lugar de destino será aquel desde donde se realizará posteriormente el proceso de análisis de los datos.

El destino de los datos puede depender de diversos factores:

- El tipo de análisis que se plantea realizar.
- La cantidad y tipo de datos.
- Las herramientas de análisis y visualización a utilizar.

Tipos de almacenamiento

El resultado de un proceso ETL se puede almacenar de varias formas:

- Archivos de datos.
- Bases de datos relacionales.
- Bases de datos no relacionales.

Debemos elegir el modelo de datos que mas se adecúe a nuestro caso de uso.