Módulo 1: Procesos ETL

Rafael Zambrano

ETL

- ¿Qué son los procesos ETL?

- Extracción (**Extract**): obtención de datos de distintas fuentes.
- Transformación (**Transform**): filtrado, limpieza, homogeneización y agrupación de la información.
- Carga (**Load**): organización y actualización de los datos en la base de datos.

Extracción

- Proceso de extracción

- Hacer llegar los datos desde sus orígenes hasta un lugar propietario dentro de la Organización, llamado área de *staging*
- Esta área puede ser desde una base de datos a un sistema de ficheros
- Los datos se extraen en bruto (*raw data)* sin ninguna intervención intermedia
- Modos de extracción:
    - Full Extract
    - Incremental
    - Update notification

Extracción

- ¿Qué es un dato?

WIKIPEDIA
La enciclopedia libre

Portada
Portal de la comunidad
Actualidad
Cambios recientes
Páginas nuevas
Página aleatoria
Ayuda
Donaciones
Notificar un error

Dato

Para una antigua ciudad griega de Tracia, véase Dato (Tracia).

Véase también: Archivo informático

Un **dato** es una representación simbólica (numérica, alfabética, algorítmica, espacial, etc.) de un atributo o variable cuantitativa o cualitativa. Los datos describen hechos empíricos, sucesos y entidades. Es un valor o referente que recibe el computador por diferentes medios, los datos representan la información que el programador manipula en la construcción de una solución o en el desarrollo de un algoritmo.

Hoy en día, un dato es cualquier elemento que se pueda digitalizar

Extracción


AVI
JPG
TXT
MP3
PDF
XLS
PNG
MOV
WAV
DOC
EPS
PPT
ZIP
CSS
RAR
HTML
EML
RAW
CSV
TTF
Aa

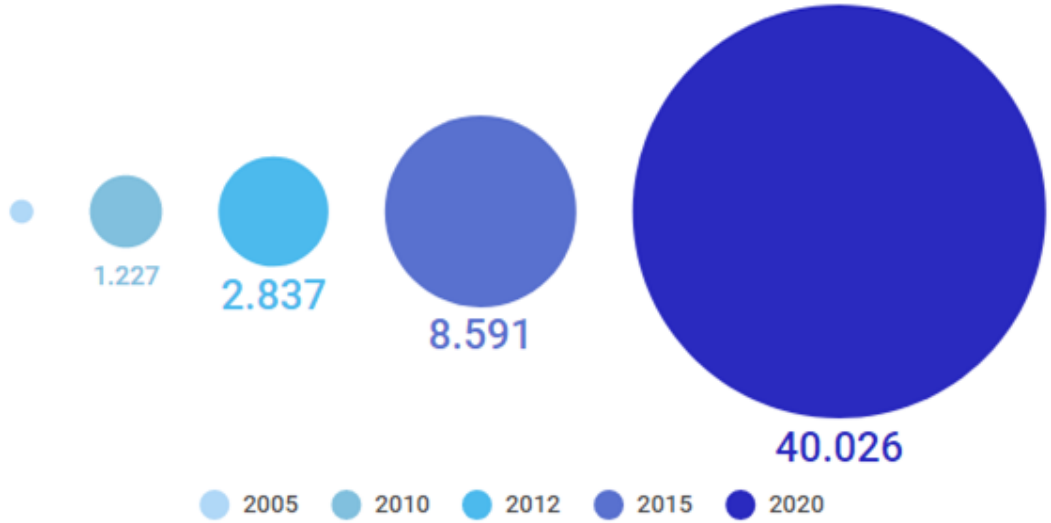

HDFS
Oracle
MySQL
IBM DB2
SQL Server
OData
SharePoint
Access
Excel
PostgreSQL
Azure SQL
Active Directory
Exchange
HDInsight
SAP BusinessObjects BI


You
Tube


2G
SMS
TCP
Custom
BLE
ZigBee
LTE
Wi-Fi
Wi
Max
UDP
MQTT
AMQP
RFID
HTTP
3G
CoAP
NFC
UDI

Extracción

- Volumen mundial de datos



Cifras expresadas en exabytes (1 exabyte=1.000 millones de GB). Fuente: IDC.

Extracción

- Tipos de datos

Extracción

- Archivos

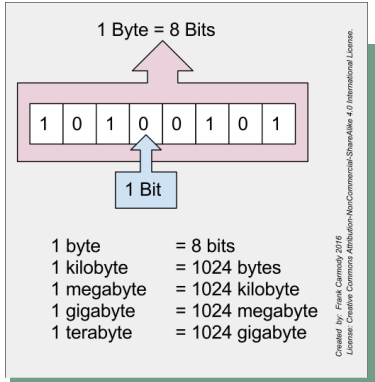
- Toda la información asociada a un dato se almacena en archivos

- Según su **funcionalidad**:
  - Ejecutables
  - Archivos de datos

- Según el **tipo de información**:
  - Compresión
  - Audio
  - Imágenes
  - Vídeo
  - Texto
  - Bases de datos
  - Etc.

Extracción

- Tamaño de archivos



Extracción

- Archivos de texto

- **Texto enriquecido**: docx, pdf, xlsx…

- **Texto plano**:
  - Ficheros con información: txt, log, bat, sh, cfg, csv, xml, json…
  - Ficheros con código de programación: java, r, py, php, jsp…

Transformación

- Estructurar datos
- Aplicar funciones o reglas de negocio
- Seleccionar o filtrar (filas, columnas)
- Traducir códigos: {“Alto”, “Medio”, “Bajo”} = {3, 2, 1}
- Anonimizaciones
- Nuevos valores: TotalVenta = Precio x Cantidad
- Unir datos de múltiples fuentes
- Agrupar filas
- Etc.

Carga

- ¿Qué es una base de datos?

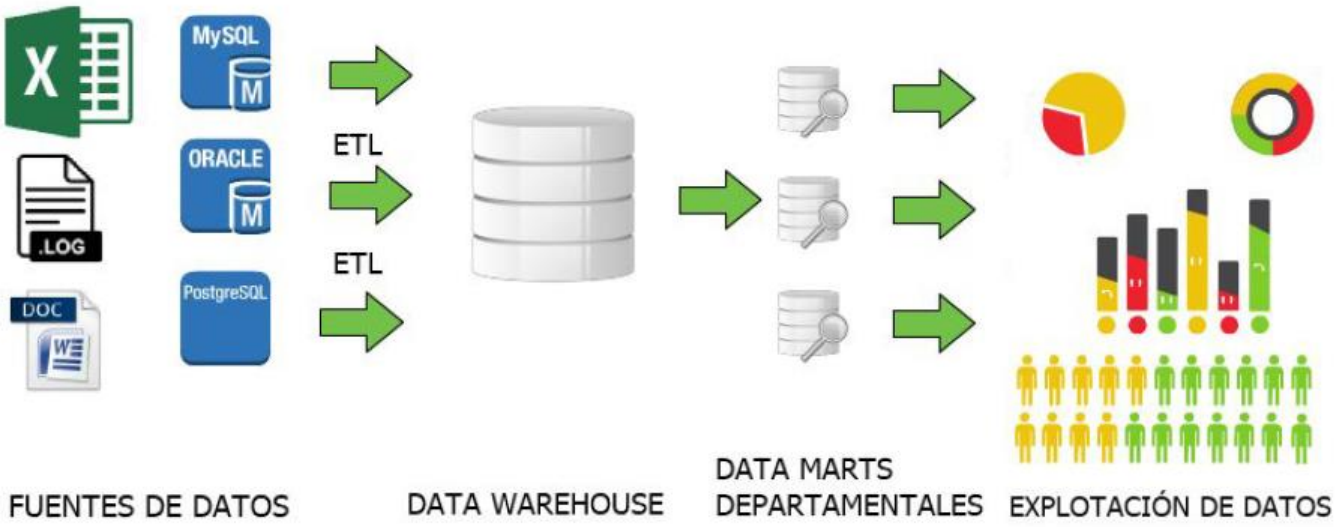
- **Almacenes** que nos permiten guardar datos de forma organizada
- Los sistemas de gestión de bases de datos (SGBD) son programas desarrollados explícitamente para gestionar bases de datos (MySQL, Oracle, SQL Server…)
- Para poder acceder a la información de una base de datos se emplean lenguajes de consulta (SQL es el más utilizado)
- Tipos de bases de datos:
  - Relacionales
  - Multidimensionales
  - NoSQL

Carga

- ¿Qué no es una base de datos?

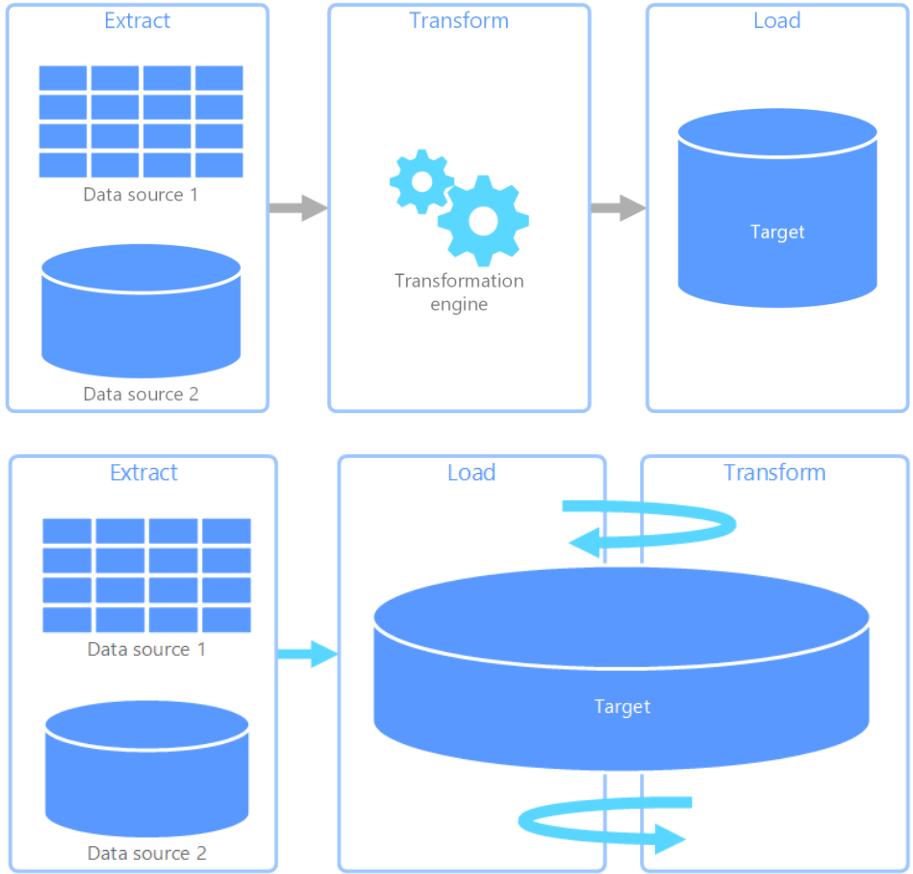
- Excel **NO** es una base de datos, es una herramienta de hoja de cálculo

Arquitectura Data Warehouse (DWH)



- El **data warehouse** es un almacén de información que integra los datos de toda las fuentes de información
- Un **data mart** es una base de datos departamental que se nutre del data warehouse

ETL vs ELT


Extract
Data source 1
Data source 2
Transform
Transformation engine
Load
Target
Extract
Data source 1
Data source 2
Load
Target
Transform

ETL vs ELT

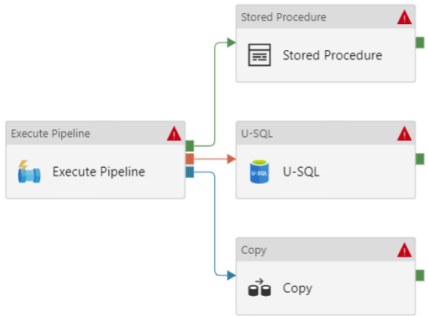
- En los procesos ELT (Extraer, cargar y transformar) los datos extraídos se cargan primero en un **data lake**
- Utilizado con cantidades grandes de datos en infraestructuras *cloud*
- Todos los datos están siempre disponibles y el acceso es más rápido

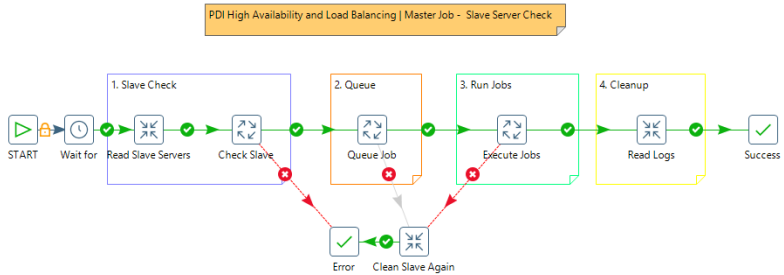
DATA WAREHOUSE
(ETL)

DATA LAKE
(ELT)

Herramientas ETL

- **Software empresarial**: Informática PowerCenter, IBM Infosphere DataStage, Oracle Data Integrator, Microsoft SQL Server Integration Services (SSIS), SAP Data Services
- **Open Source**: Talend Open Studio, Pentaho Data Integration
- **ETL personalizadas**: **Python**, Java, Spark
- **Servicios Cloud**: AWS EMR, Azure Data Factory, Google Cloud Dataflow





Lecturas recomendadas

- **Por qué guardar las fechas en UTC en la base de datos:** https://picodotdev.github.io/blog-bitix/2016/08/por-que-guardar-las-fechas-en-utc-en-la-base-de-datos/

- **ETL vs ELT: 5 Critical Differences** https://www.xplenty.com/blog/etl-vs-elt/

¡Gracias!