# PROCESADO DE DATOS (ETL)

* Extract
* Transform
* Load

## Extract

* Hacer llegar los datos desde sus orígenes hasta un lugar propietario dentro de la organización, llamado área de staging
* Puede ser un conjunto de ficheros o un gestor de bases de datos
* Los datos llegan en bruto, crudo o raw, no hay un procesamiento intermedio (T)

### Archivos

Se almacenan según su:

* Funcionalidad: ejecutables o archivos de datos
* El formato del archivo y por tanto la forma de acceso depende del tipo de dato que almacene:
  + Texto plano: txt, csv, xml, json
  + Texto enriquecido (pdf, word...)
  + Hojas de calculo
  + Bases de datos

## Transform

* Unir fuentes de datos
* Analizar y tratar duplicados
* Limpiar datos y unificar formatos
* Generar nuevas columnas/datos a raíz de los existentes
* Simplificar datos existentes
* Tratar los datos faltantes

## Load

* Sistema de archivos local o en la nube
* Base de datos
* Almacenes: datos de forma organizada
* Gestor de bases de datos (SGBD)
* Oracle, MySQL, PostgreSQL, SQL Server
* Para acceder: lenguajes especializados (SQL)