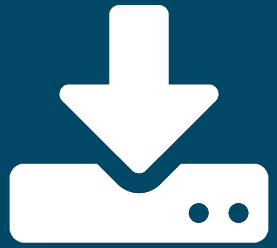


Almacenamiento y captura de datos

Claudio Aracena
GobLab - Universidad Adolfo Ibáñez
Chatbot Chile

Base de datos no relacionales (NoSQL)

NoSQL es el término genérico usado para referirse a almacenamiento de datos que no sigue el modelo tradicional de base de datos relacionales.
Específicamente, la data no sigue el modelo entidad-relación y no utiliza SQL como lenguaje de consulta.

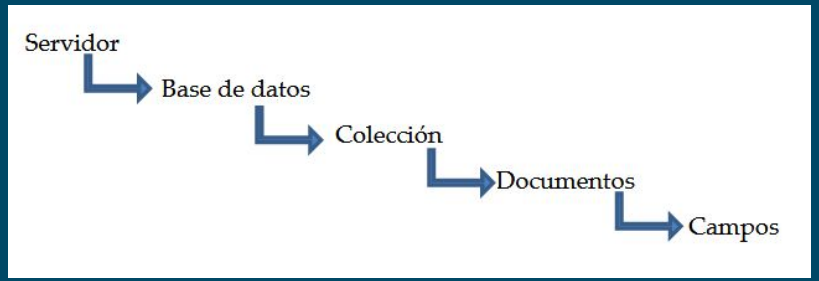
Ejemplos de estas bases de datos son

- MongoDB (document-oriented)
- Cassandra, Hbase (column-oriented)
- Redis (key-value)
- Neo4j (graph-oriented)

https://towardsdatascience.com/datastore-choices-sql-vs-nosql-database-ebec24d56106

Base de datos documentales

- Una base de datos documental está constituida por un conjunto de programas que almacenan, recuperan y gestionan datos de documentos o datos de algún modo estructurados.
- A diferencia de las bases de datos relacionales, estas bases de datos están diseñadas alrededor de una noción abstracta de "Documento".
- En MongoDB un documento es un conjunto de datos almacenado en formato JSON



mongoDB

https://www.mongodb.com/online

Base de datos basadas en grafos

- Representa la información como nodos de un grafo y sus relaciones con las aristas del mismo
- **Ventajas**
  - Consultas realmente rápidas cuando busca relaciones entre nodos
  - Realmente rápido para recorrer nodos
  - Puede representar múltiples dimensiones
- **Desventajas**
  - Inapropiado para información transaccional, como registros contables donde las relaciones entre registros son más simples
  - Es difícil hacer consultas agregadas de manera eficiente

https://sandbox.neo4j.com/

Base de datos llave-valor

- Guardan tuplas que contienen una clave y su valor.
- Cuándo se quiere recuperar un dato, simplemente se busca por su clave y se recupera el valor.
- En general es para tipos de datos simples, o cuando queremos buscar un dato en particular en forma rápida
- **Desventajas**:
  - No es muy útil para almacenar relaciones.
  - Es difícil mantener llaves únicas cuando los datos aumentan

https://try.redis.io/

Base de datos columnares

- Una base de datos en columnas está optimizada para lograr una recuperación rápida de columnas de datos
- Normalmente son usadas en aplicaciones analíticas.
  - Para realizar operaciones de agregación (min, max, mean) sobre datos en particular
  - Esto es eficiente pues recupera toda la columna en forma rápida

Contenidos

- Captura de datos desde archivos
- Base de datos
- **Captura y almacenamiento de datos en BD**
- Captura de datos de la Web (Web scraping)
- Captura de datos de API (ej: Twitter)
- Captura y almacenamiento en arquitecturas Big data

Códigos y clase en:
https://github.com/caracena/almacenamiento-captura-datos

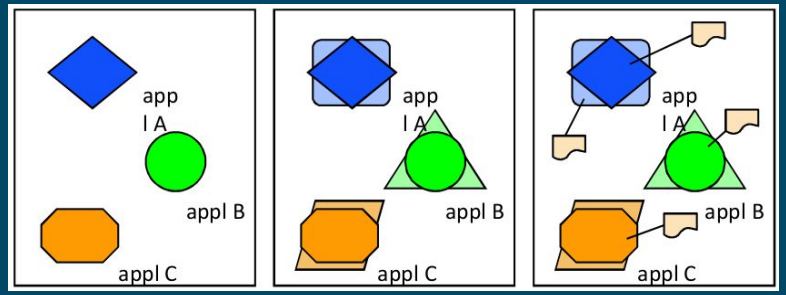
Clase de hoy

Captura y almacenamiento de datos en BD

- Data Warehouse
- Extract-Transform-Load

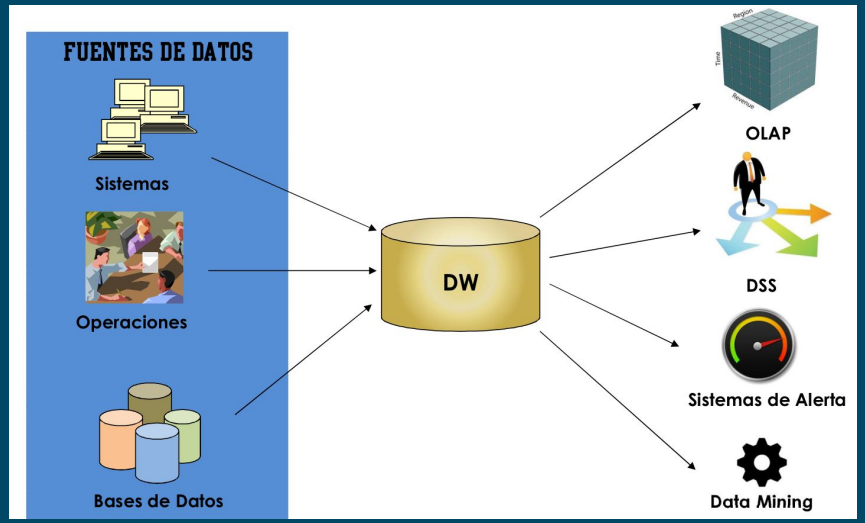
Data Warehouse

- El problema de acceso a los datos
  - Los sistemas operacionales (transaccionales) no están pensados para ser sistemas de análisis.
  - Las sistemas se complejizan a lo largo del tiempo
  - Para el análisis de datos usualmente se requiere integración entre diversas fuentes de datos
  - El equipo de TI no da abasto con las solicitudes de información y datos



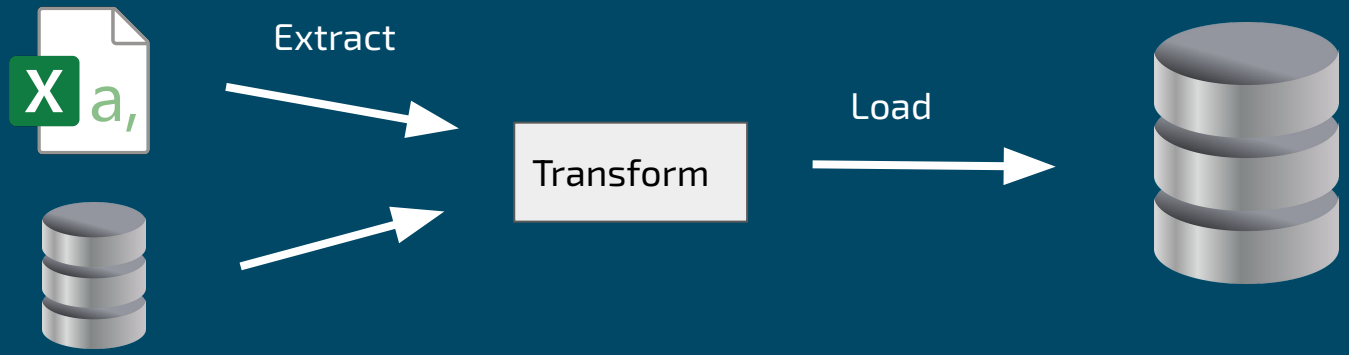
Data Warehouse

- Data Warehouse
  - Un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte de la toma de decisiones de la gerencia



Extract-Transform-Load (ETL)

Se refiere al proceso de extraer datos desde una o diversas fuentes, transformarlos, formatearlos y limpiarlos, y finalmente cargarlos en otro lugar, ya sea una base de datos o sistema de archivos.



Herramientas disponibles para ETL

- Desarrollo a medida
- Librerías
  - petl (https://github.com/petl-developers/petl)
  - bonobo (https://www.bonobo-project.org/)
  - mara (https://github.com/mara)
- Frameworks
  - Apache Nifi (https://nifi.apache.org/)
  - Apache Airflow (https://airflow.apache.org/)
  - Luigi (https://github.com/spotify/luigi)

Ejemplo

Queremos aplicar el siguiente proceso de ETL:

- Extraer las ventas y los clientes que realizan esas ventas
- Transformar el monto de las ventas a pesos chilenos (USD = 820 CLP)
- Cargar los datos transformados en una nueva base de datos llamada ventas