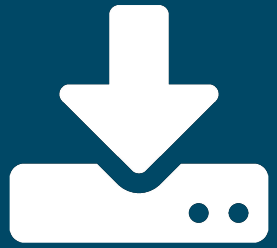


Almacenamiento y captura de datos

Claudio Aracena

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

- Captura de datos desde BD
- Almacenamiento en BD
- Procesamiento de datos en Python
- Data Warehouse
- Extract-Transform-Load

Conexión y consulta a base de datos

```
import sqlite3

conn = sqlite3.connect("data/chinook.db")

cur = conn.cursor()

cur.execute("SELECT * FROM albums")

rows = cur.fetchall()
```

Conexión y consulta a base de datos

```
import mysql.connector

conn = mysql.connector.connect(
    host="dirección servidor",
    user="usuario",
    password="constraseña",
    database="nombre_base"
)

cur = conn.cursor()

cur.execute("SELECT * FROM albums")

rows = cur.fetchall()
```

Consulta a base de datos con pandas

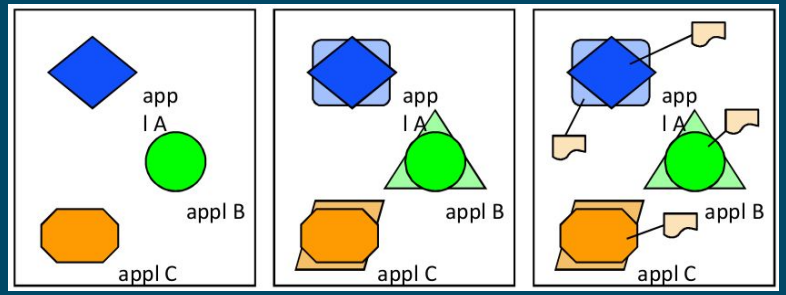
```
import pandas as pd

df = pd.read_sql("SELECT * FROM albums", conn)
```

Creación de base de datos

```
!pip install sqlalchemy

from sqlalchemy import create_engine

sql_engine = create_engine('sqlite:///data/test.db')

connection = sql_engine.raw_connection()

df.to_sql('artists_albums', connection, index=False)
```

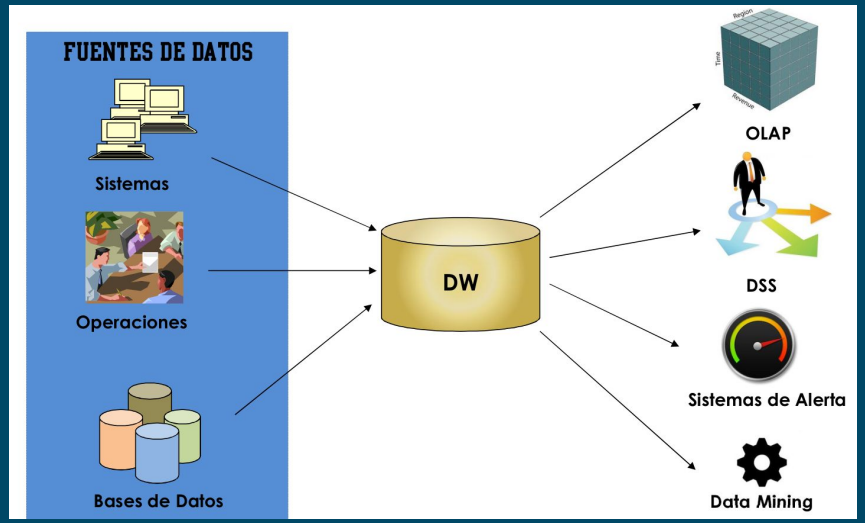
Data Warehouse

- 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



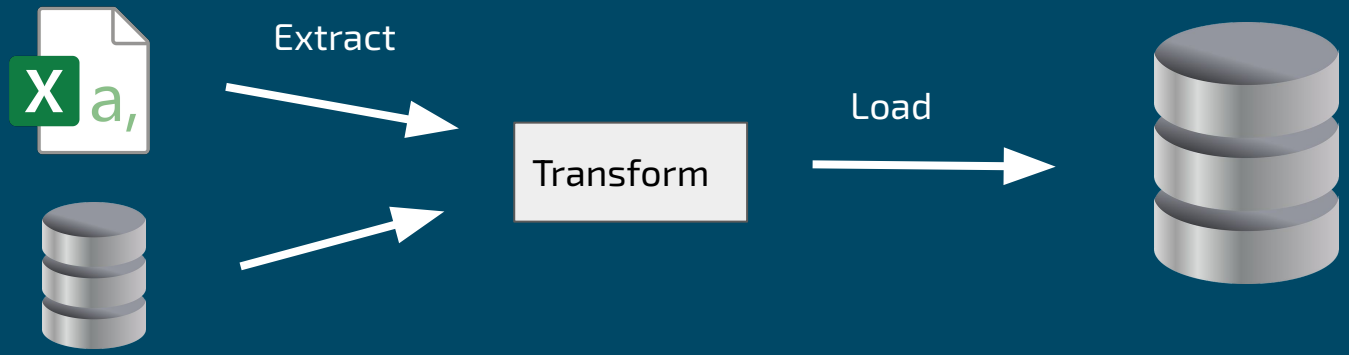
Data Warehouse

- 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 = 920 CLP)
- Cargar los datos transformados en una nueva base de datos llamada ventas