



Almacenamiento y captura de datos

Claudio Aracena



Preguntas de la clase anterior

- Revisión ejercicio 5 - Trabajo con excel
- Lectura de archivos grandes
 - <https://www.kaggle.com/rohanrao/tutorial-on-reading-large-datasets>
 - <https://www.kaggle.com/c/riiid-test-answer-prediction/data?select=train.csv>



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



Base de datos

- Sistema Gestor de Base de datos
- Modelo entidad-relación
- SQL (Structured Query Language)





Base de datos

Una Base de Datos es una colección de información perteneciente a un mismo contexto (o problema), que está almacenada de forma organizada en archivos.

Una base de datos está organizada mediante **tablas** que almacenan información concerniente a algún objeto o suceso. Estas tablas se relacionan formando vínculos o **relaciones** entre ellas, que ayudan a mantener la información de los diversos objetos de forma ordenada y coherente.

Cada una de estas tablas es una estructura que se parece a las hojas de cálculo, pues está dispuesta mediante filas y columnas (**registros** y **campos**).



Sistema Gestor de Base de datos

Se define un Sistema Gestor de Base de Datos (SGBD o DBMS en inglés) como el conjunto de herramientas que facilitan la consulta, uso y actualización de una base de datos.

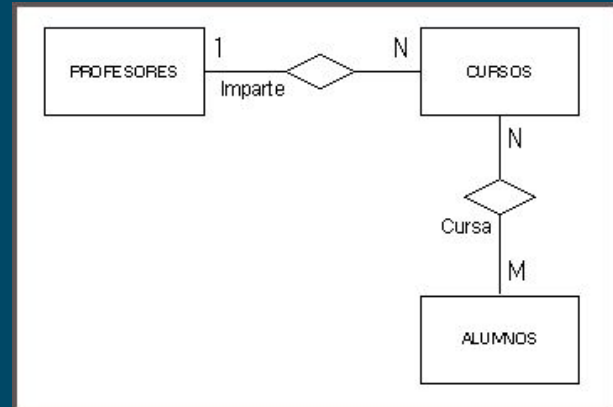




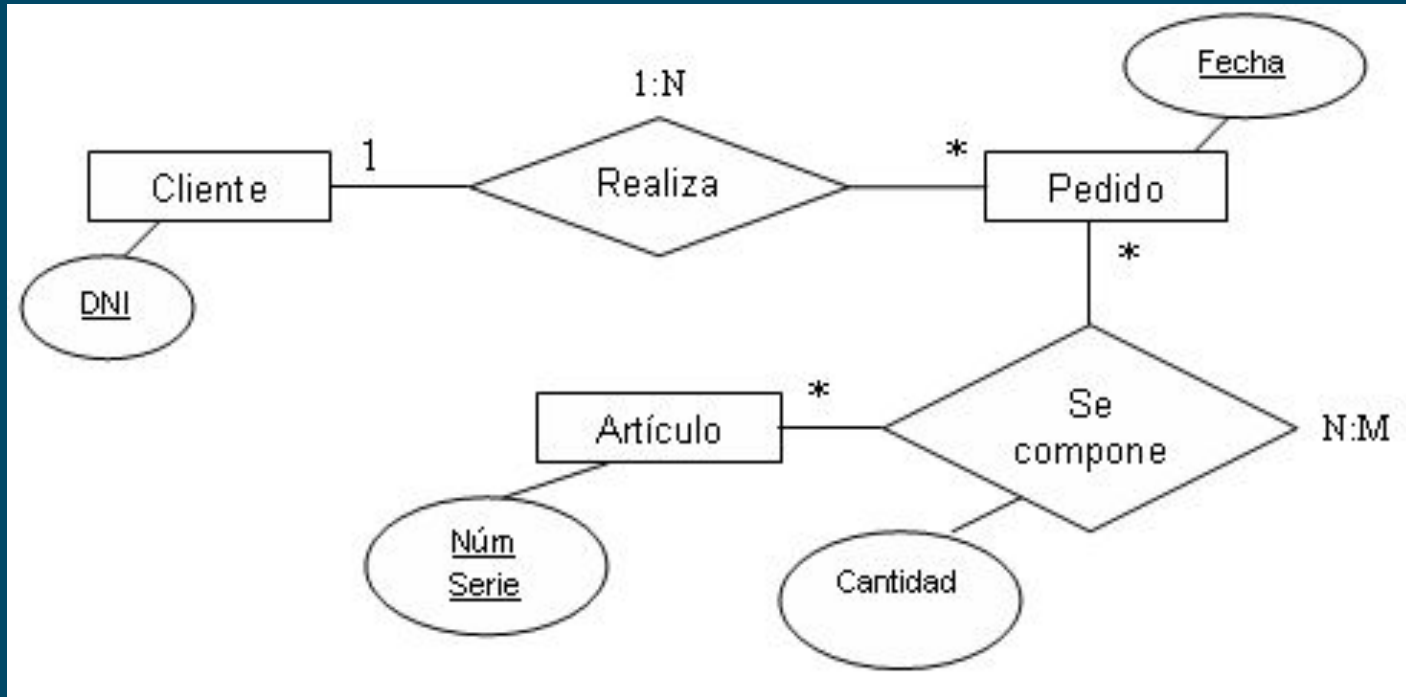
Modelo entidad-relación

Este modelo consiste en plasmar el resultado del análisis del problema mediante diagramas entidad-relación.

- Entidad: Cualquier tipo de objeto o concepto sobre el que se recoge información; cosa, persona, concepto abstracto o suceso
- Relación: Una relación (o interrelación), es una correspondencia o asociación entre dos o más entidades.



Modelo entidad-relación



Ejercicio entidad-relación

Le contrataron para hacer una BD que permita apoyar la gestión de una PYME. La empresa necesita tener registro de sus proveedores, productos, ventas y clientes.

- Un proveedor tiene un ID, nombre, dirección, teléfono y web.
- Un cliente también tiene ID, nombre, dirección y teléfono.
- Un producto tiene un ID, nombre, precio y stock actual, y es vendido por un proveedor. Además los productos se organizan por categorías donde, un producto puede tener más de una categoría. A su vez la categoría tiene un ID, nombre y descripción.
- Las ventas se registran con un ID, fecha, cliente y monto final. Además se debe guardar el precio al momento de la venta, la cantidad vendida y el monto total por cada producto.



SQL (Structured Query Language)

La principal herramienta de un gestor de base de datos es la interfaz de programación con el usuario. Este interfaz consiste en un lenguaje muy sencillo llamado SQL, Structured Query Language. SQL está estandarizado por la ISO, es decir, todas las bases de datos que soporten SQL deben tener la misma sintaxis a la hora de aplicar el lenguaje.

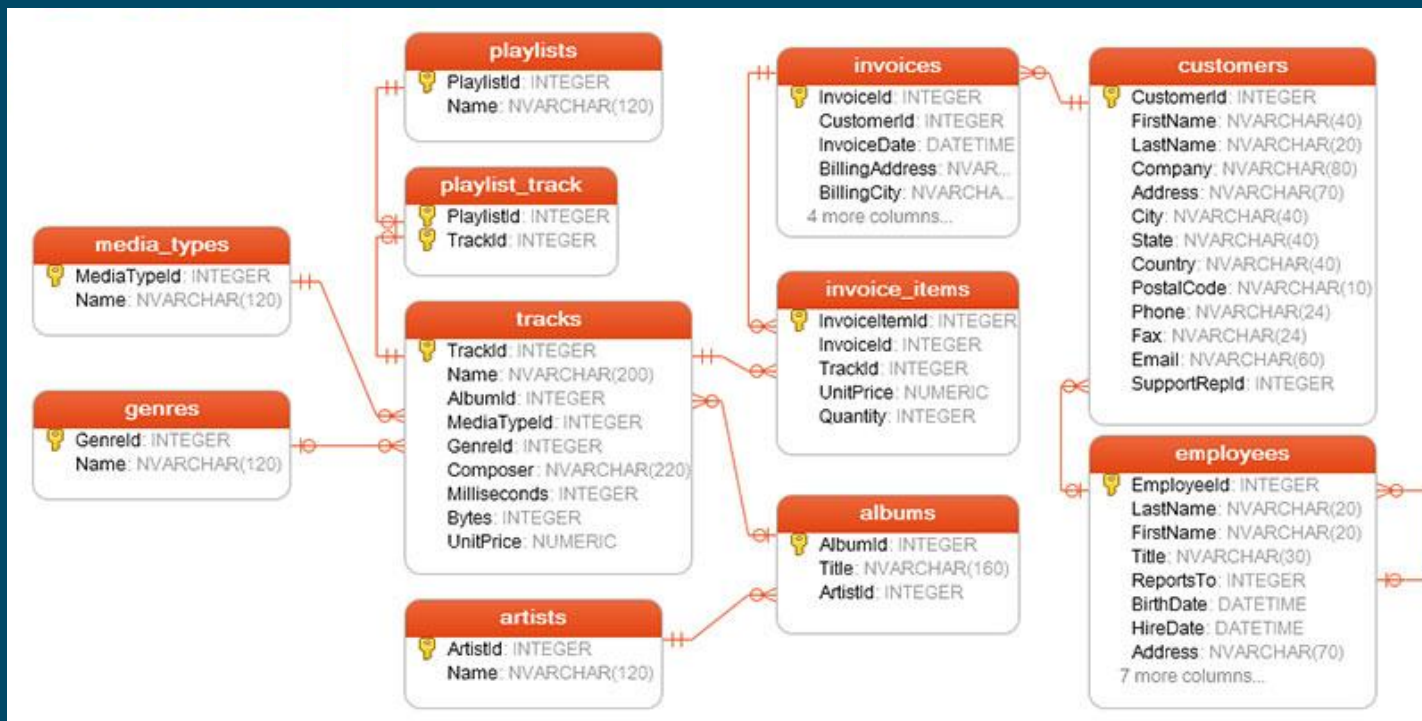
Comandos más comunes:

- DML (Data Manipulation Language): SELECT, UPDATE, INSERT, DELETE,
- DDL (Data Definition Language): CREATE, ALTER, DROP

SQL (Structured Query Language)



- Abrir <https://sqliteonline.com/> e importar chinook.sql





Ejemplos SQL

- Primer nombre de los empleados:
 - `SELECT FirstName FROM Employees;`
- Empleados que sean IT Staff
 - `SELECT * FROM Employees WHERE Title = 'IT Staff';`
- Álbumes que comiencen con la letra a
 - `SELECT * FROM Albums WHERE title LIKE 'A%';`
- Suma de los totales de los invoices para Alemania
 - `SELECT sum(Total) FROM Invoices WHERE BillingCountry = 'Germany';`
- Suma de los totales de los invoices agrupado por país
 - `SELECT BillingCountry, sum(Total) AS monto FROM Invoices GROUP BY BillingCountry;`
- Artistas y sus álbumes
 - `SELECT * FROM artists JOIN albums ON artists.ArtistId = albums.ArtistId;`



Google Cloud Platform

Cloud SQL

- Base de datos alojada en la nube de Google
- 3 opciones de motores (SGBD) de base de datos
 - MySQL
 - PostgreSQL
 - SQL Server
- Se paga por cpu, ram, espacio, tipo de espacio, ubicación de los servidores, entre otros