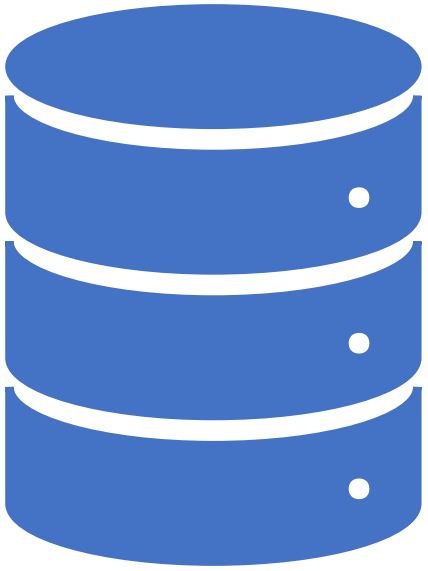


Economía y Ciencia de los Datos:

Bases de datos estructuradas - *Tópicos adicionales*

Carlos Alvarado & Pablo González

Algunas formas de almacenar los datos

CSV

Archivo .txt
(separar por pipes)

Parquet

¿Y bases de datos?

SQLite

Base de datos embedida más popular en el mundo

Auto-contenida

(Archivo → schema + data)

No require de un servidor ni mayor configuración

Soporta bases de hasta **140TB**

Es una base de datos

(Permite consultas, indices, full-text search, etc.)

Forma general (Consultas SQL)

SELECT v1, v2, v3, … vn

FROM tabla

(INNER/LEFT/RIGHT/FULL OUTER, etc) JOIN tabla2 ON …

WHERE predicado

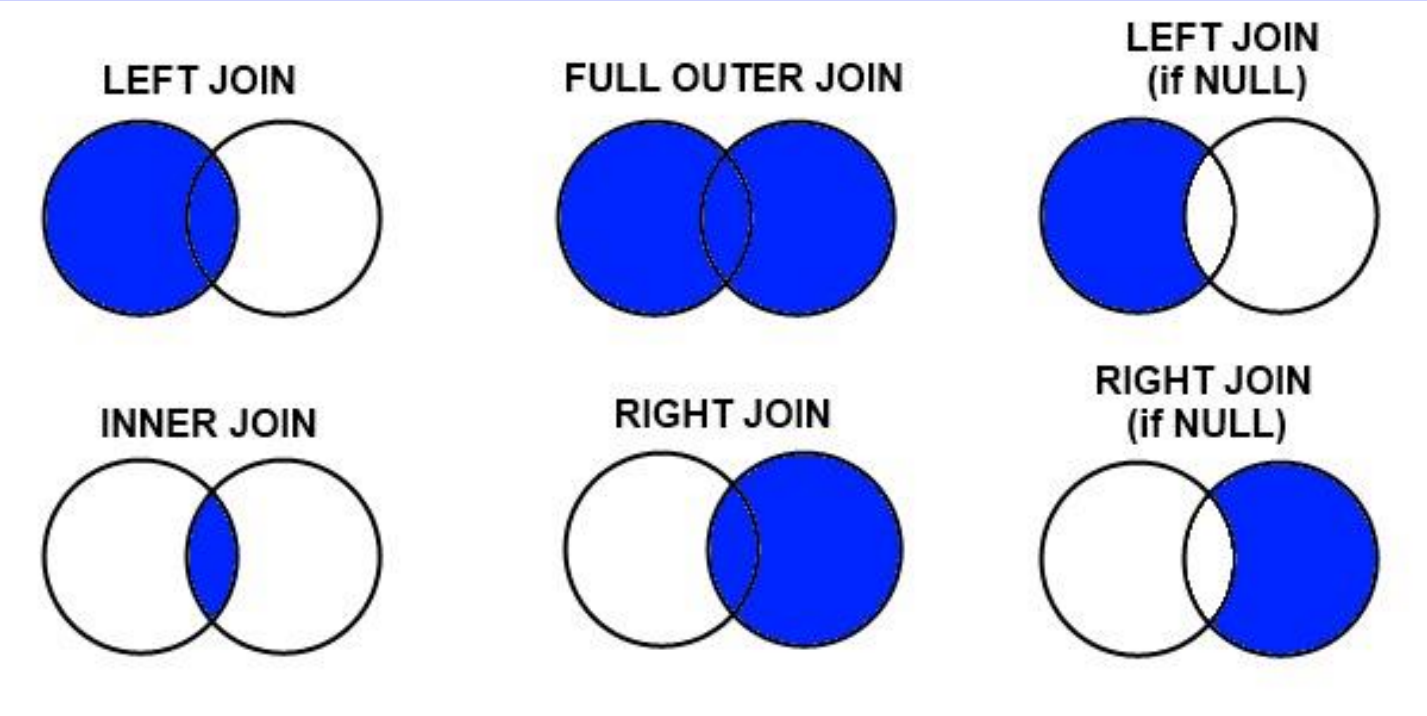
[ORDER BY …]

[GROUP BY …]

[HAVING …]

- *Nótese que transformaciones intermedias también se pueden realizar vía tablas temporales.*

Diagramas de Venn para JOINs



Conocimientos complementarios

- Iteración
- Ifnull/isnull
- Pivotes

Nótese:

Estos son los tipos de Join más utilizados. Sin embargo, también existen otros tales como el “CROSS JOIN”

Pivots

Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31

Long Format

Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

¿Cuándo sería mejor qué formato?

¿Cómo podemos pasar de uno al otro?

¿A qué se podría deber la lentitud de mi consulta?

Revisa que la base tenga índices.

Estructura de datos de B-tree

→ Velocidad de O(log n) para encontrar/añadir/eliminar un item

“Create INDEX nombre_index ON Tabla(llave)”

Diferencias entre “variantes” SQL

No todas las variants de SQL son iguales y ocaasionalmente trabajarán con otras versions, tales como TSQL. Algunas diferencias generals presents entre versions podrían ser:

DIFERENCIA DE FUNCIONES Y CLAUSULAS

CREACIÓN DE VARIABLES LOCALES

POSIBILIDAD/FORMA DE ITERAR

Iterando en TSQL