



Especialización en Inteligencia  
de Datos Aplicada



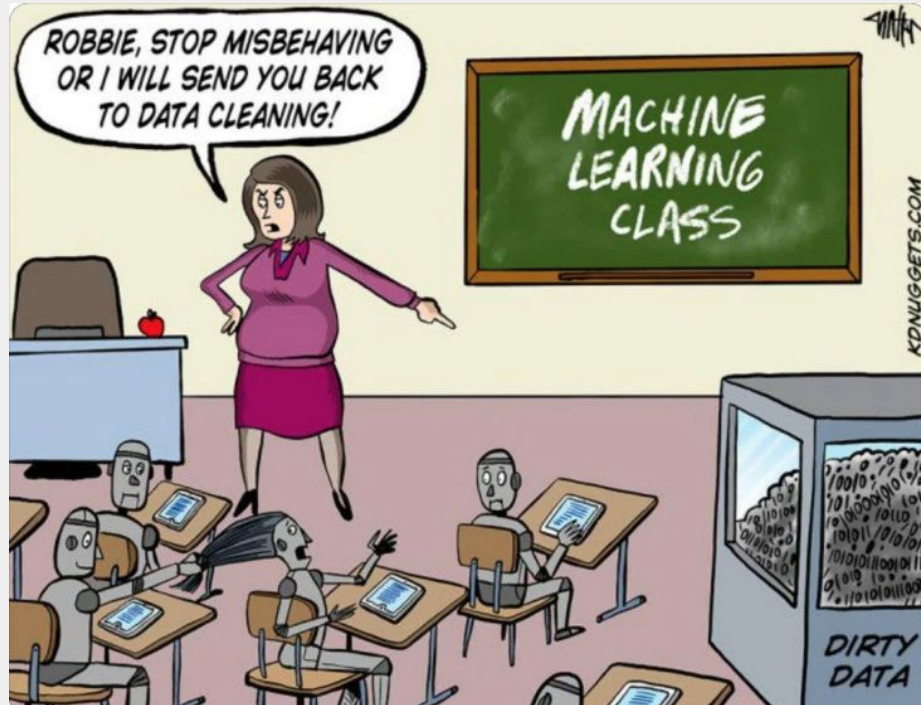
Facultad de Informática  
Universidad Nacional del Comahue

# EXTRACCIÓN, PREPARACIÓN Y ALMACENAMIENTO DE LOS DATOS

Segundo Cuatrimestre  
Unidad III

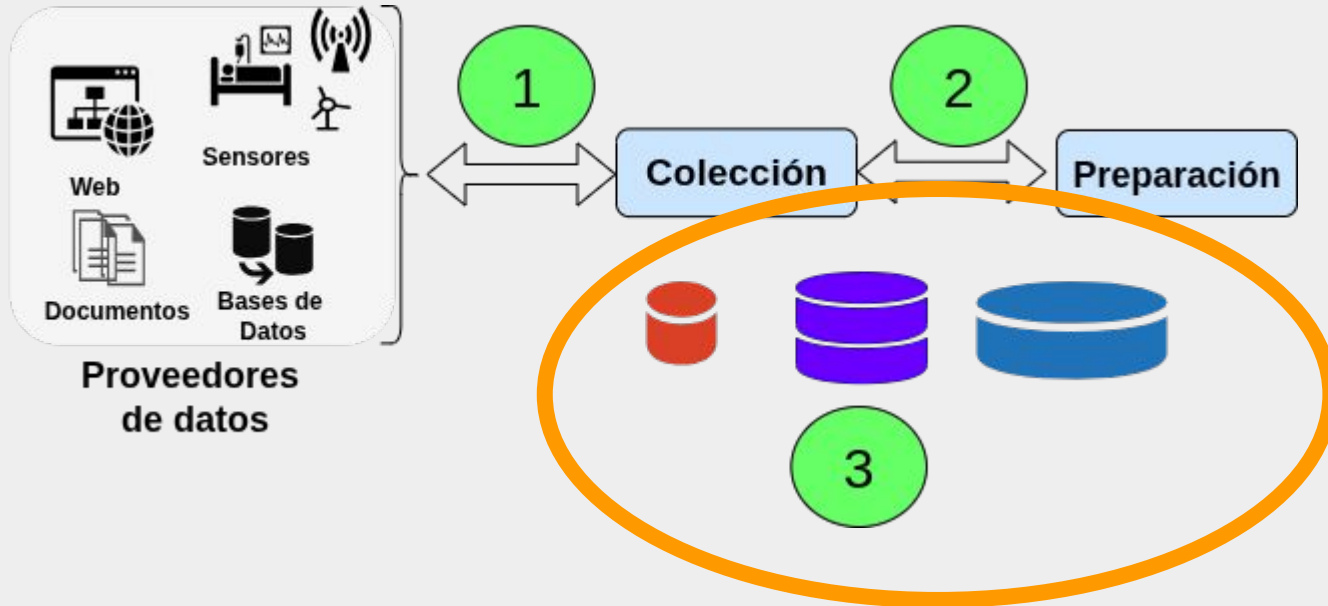
# PROCESO DE ANÁLISIS DE DATOS

## Preparación de los Datos, es así?



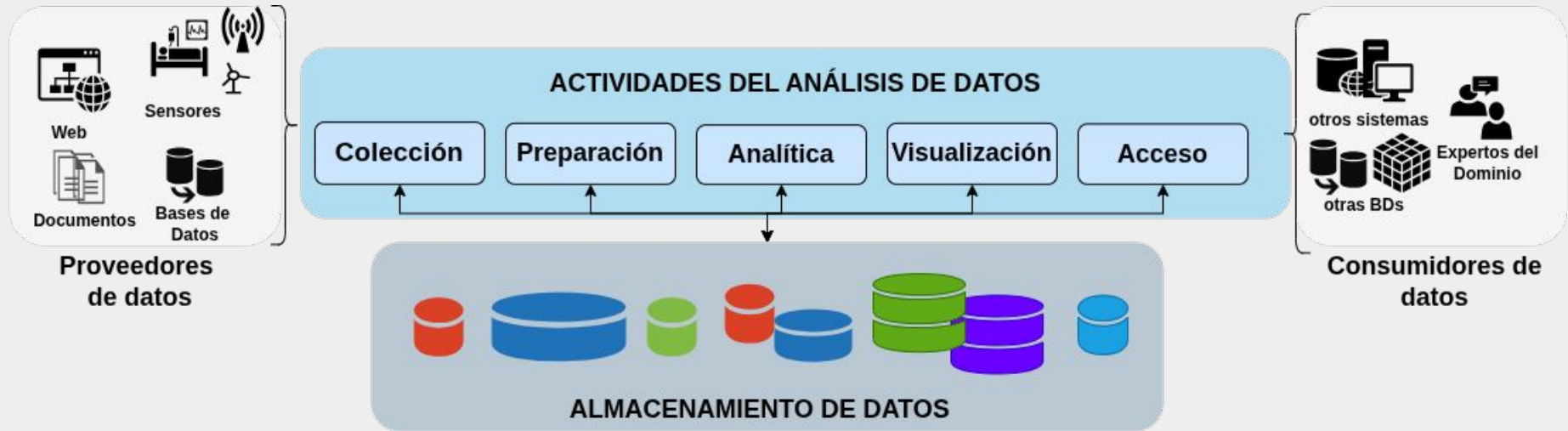
# PROCESO DE ANÁLISIS DE DATOS

## Con qué seguimos?



# PROCESO DE ANÁLISIS DE DATOS

## Actividades



# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento de Datos

- Lamentablemente y afortunadamente hay **muuuuchos** paradigmas de almacenamiento que funcionan bajo diferentes reglas y entornos.



# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento de Datos

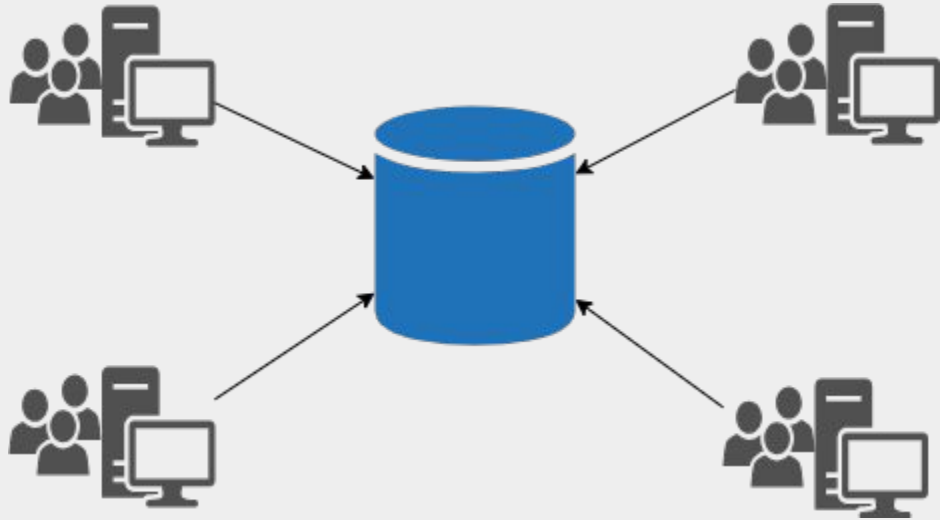
---

- Pero... veremos cada uno detalladamente durante la Carrera.
- Aquí empezamos sólo con algunos



# PROCESO DE ANÁLISIS DE DATOS

## Entornos centralizados vs distribuidos

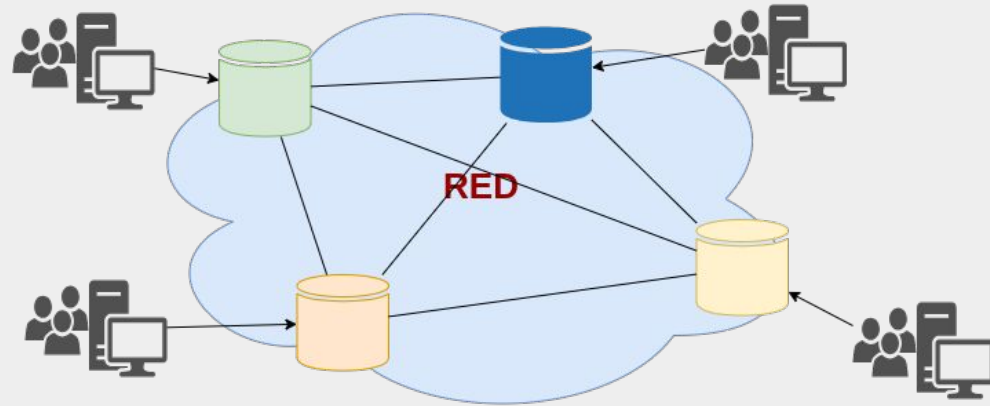


**Centralizado:** los datos se almacenan y mantienen en una única ubicación, al cual acceden los usuarios o aplicaciones a través de una red.

# PROCESO DE ANÁLISIS DE DATOS

## Entornos centralizados vs distribuidos

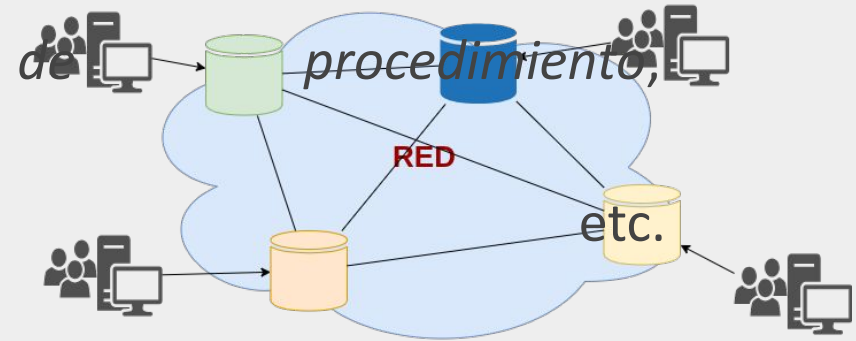
**Distribuido:** es una colección de sitios (o nodos) conectados mediante alguna red de comunicación, donde cada uno representa un sistema en sí mismo, y los sitios están de acuerdo en trabajar juntos (si es necesario).



# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento distribuido

- **Ventajas:** *Eficiencia  
accesibilidad, disponibilidad,  
rendimiento,*

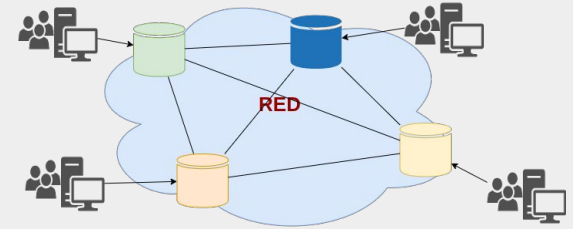


- **Regla básica:** *Para un usuario un Sistema Distribuido debería verse exactamente igual que un Sistema no-distribuido*

# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento distribuido

- **Características más importantes:** Fragmentación, Replicación, Transparencia, Procesamiento Distribuido, Escalabilidad horizontal



# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento de Datos - Reglas

¿Qué es una OPERACIÓN?

SELECT

DELETE

UPDATE

INSERT

¿Qué es una TRANSACCIÓN?

Un conjunto de operaciones que responden a una lógica del negocio

**A**

T  
O  
M  
I  
C  
I  
D  
A  
D

**C**

O  
N  
S  
I  
S  
T  
E  
N  
C  
I  
A

**I**

A  
I  
S  
L  
A  
C  
I  
Ó  
N

**D**

U  
R  
A  
B  
I  
L  
I  
D  
A  
D

...pensamos en transacciones que deben ser ACID



# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento de Datos - Reglas

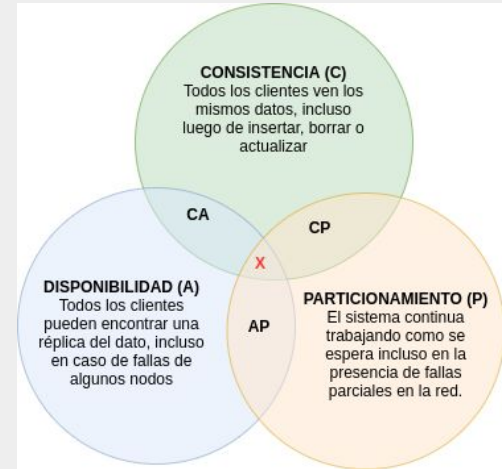
- **Teorema CAP:** se aplica a entornos distribuidos
  - **Consistencia (C):** Todos los nodos deben poseer la misma información. Para cumplir con esta propiedad los nodos deben estar siempre comunicados.
  - **Disponibilidad (A):** el sistema debe estar siempre funcionando, incluso cuando un nodo cae, otro puede ocupar su lugar para responder a una petición de lectura o escritura. El sistema siempre debe responder a un cliente ya sea con éxito o con falla.
  - **Tolerancia a Particiones de Red (P):** el sistema debe poder funcionar aún cuando existan particiones en la red que dividan los nodos en una o más particiones.

# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento de Datos - Reglas

### ● Teorema CAP

- **C+P**: los nodos no pueden mantenerse disponibles (A) mientras se quiera lograr un estado de consistencia (C), debido a las particiones (P).
- **A+P**: la consistencia (C) no es posible debido al particionamiento (P) y mantener la disponibilidad (A).
- La elección es entre **C y A** solo cuando ocurre una partición de red o falla



# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento de Datos - Reglas

---

- **Principio BASE** - basado en el Teorema CAP
  - *Básicamente disponible*: la BD siempre reconocerá un requerimiento de un cliente y responderá con éxito o con falla.
  - *Estado suave*: la BD puede caer en un estado inconsistente cuando un dato es leído, es decir que los resultados de una consulta a un dato puede cambiar cuando el mismo dato es requerido.

# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento de Datos - Reglas

- **Principio BASE** - basado en el Teorema CAP
  - *Eventualmente consistente*: el sistema puede encontrarse en un estado inconsistente en algunos momentos, pero con el paso del tiempo se volverá consistente.

Cuando un SGBD soporta BASE favorece la disponibilidad contra la consistencia, es decir cumple A+P del teorema CAP.

# PROCESO DE ANÁLISIS DE DATOS

## Clasificación de Bases de Datos

| SQL          | vs                  | NoSQL                                |
|--------------|---------------------|--------------------------------------|
| RIGIDO       | ESQUEMA             | FLEXIBLE                             |
| VERTICAL     | ESCALABILIDAD       | HORIZONTAL                           |
| ACID         | INTEGRIDAD DE DATOS | BASE/CAP                             |
| CENTRALIZADO | ENTORNO             | DISTRIBUIDO                          |
| CONSISTENCIA | PRIORIDADES         | DISPONIBILIDAD Y TOLERANCIA A FALLAS |

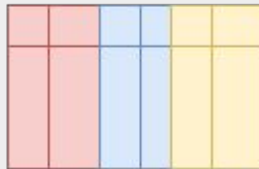
# PROCESO DE ANÁLISIS DE DATOS

## Clasificación de Bases de Datos

SQL

Lenguaje principal SQL

BD  
Relacionales



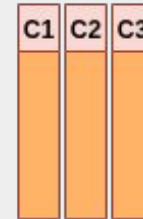
NoSQL

No solo SQL

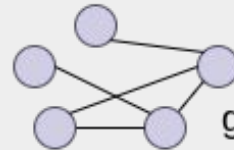
basada en  
documentos



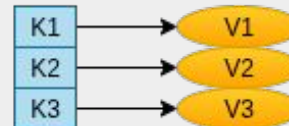
orientado  
a columnas



grafos



clave-valor



# PROCESO DE ANÁLISIS DE DATOS

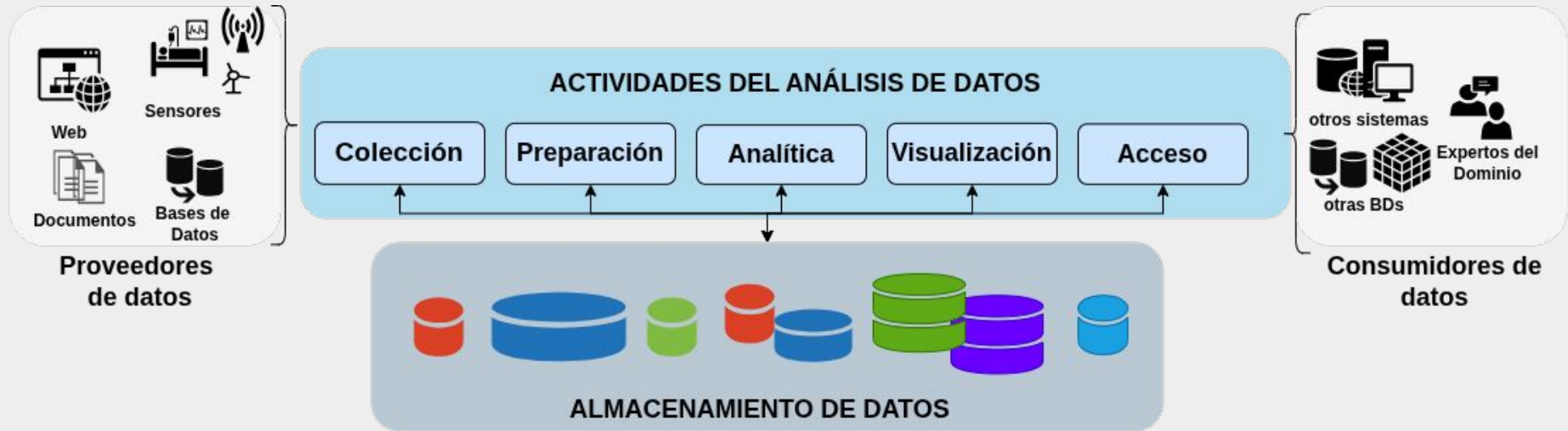
## Clasificación de Bases de Datos



# PROCESO DE ANÁLISIS DE DATOS

## Utilidad de BDs

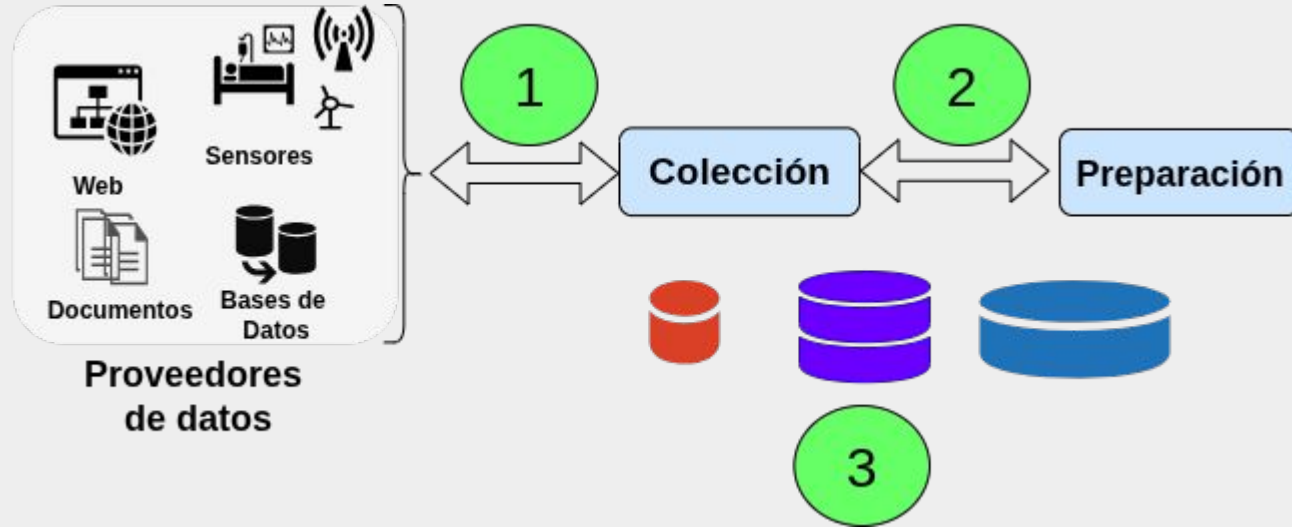
- Volvemos otra vez al proceso



# PROCESO DE ANÁLISIS DE DATOS

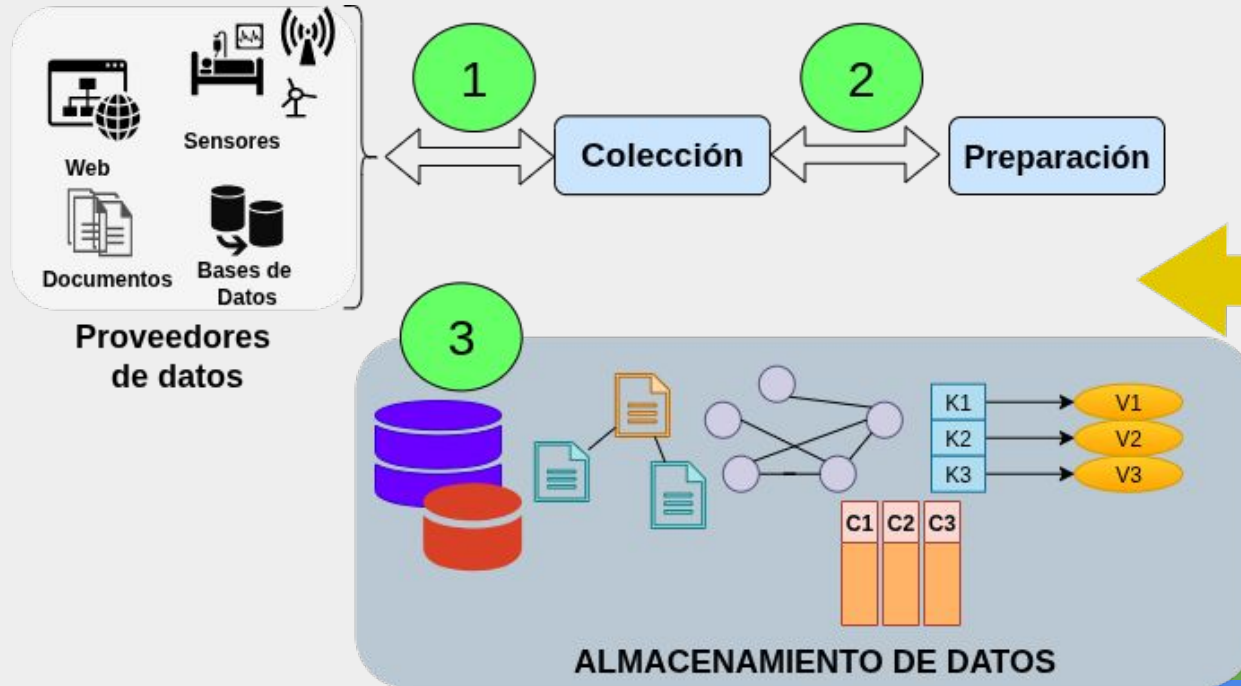
## Utilidad de BDs

- Volvemos otra vez al proceso - estabamos aca



# PROCESO DE ANÁLISIS DE DATOS

## Utilidad de BDs

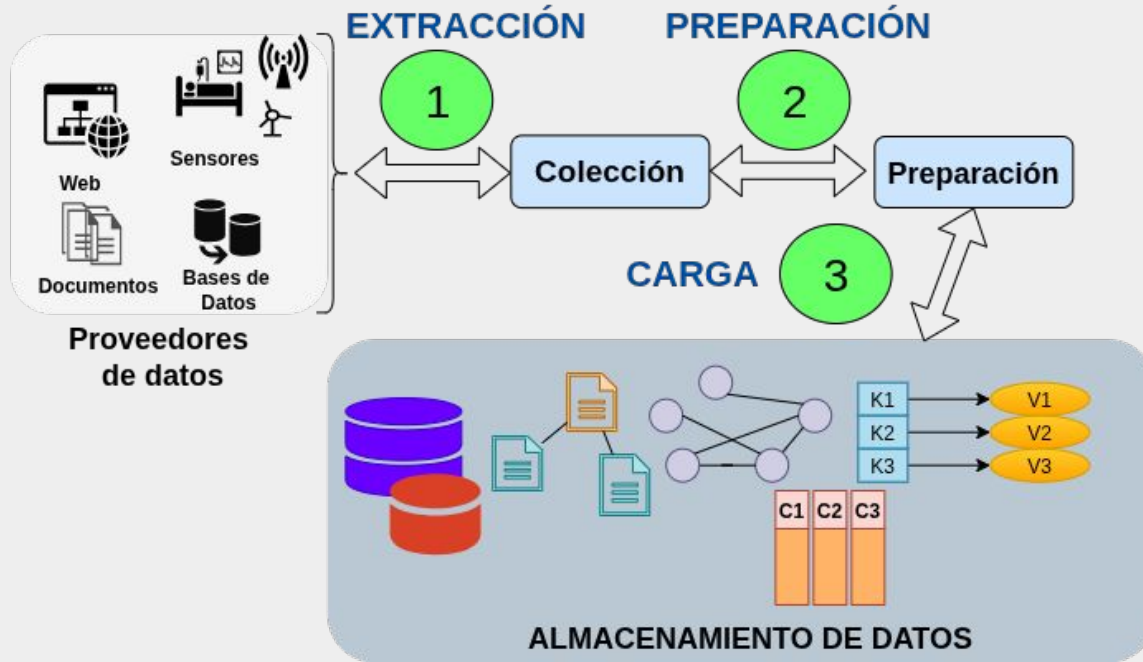


¿Cómo interactúa la **Colección** y **Preparación** con el **Almacenamiento**?

# PROCESO DE ANÁLISIS DE DATOS

## Proceso ETL

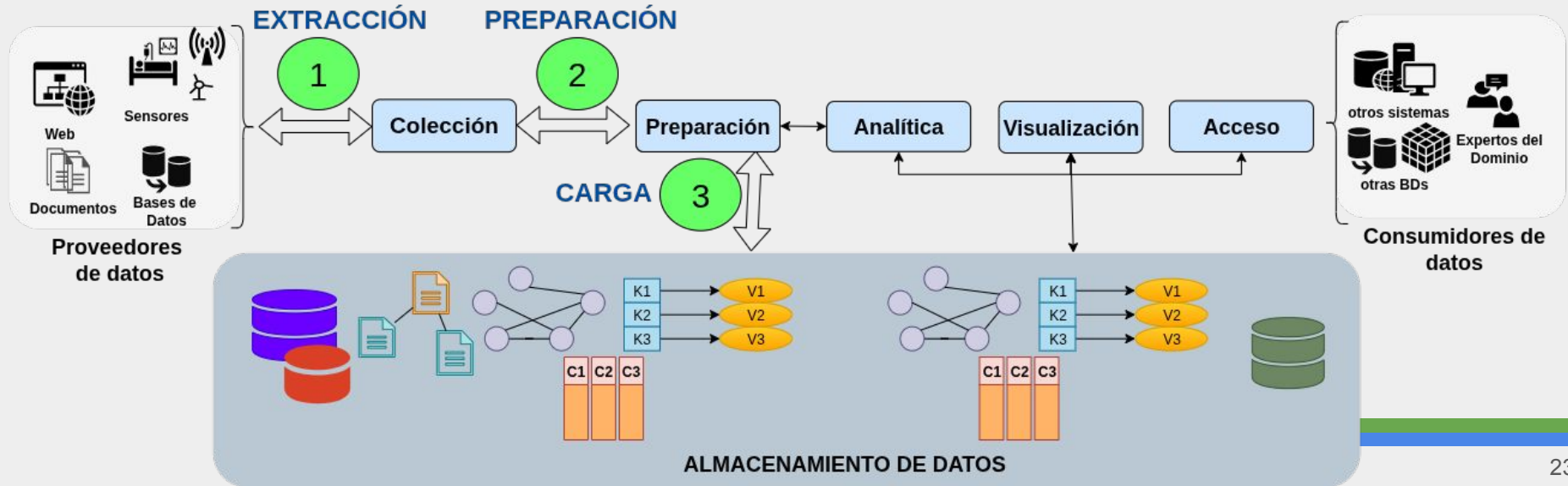
- Proceso **ETL**: *extracción-transformación-carga*



# PROCESO DE ANÁLISIS DE DATOS

## Proceso ETL Completo

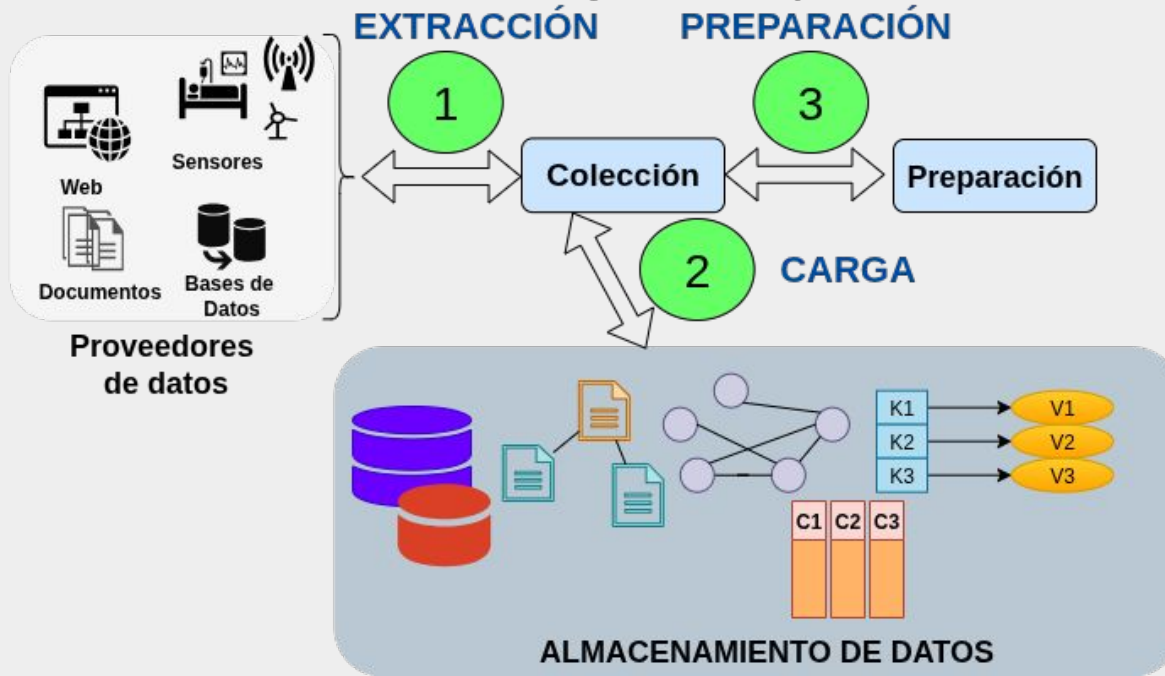
Requerimiento  
Definido



# PROCESO DE ANÁLISIS DE DATOS

## Proceso ELT

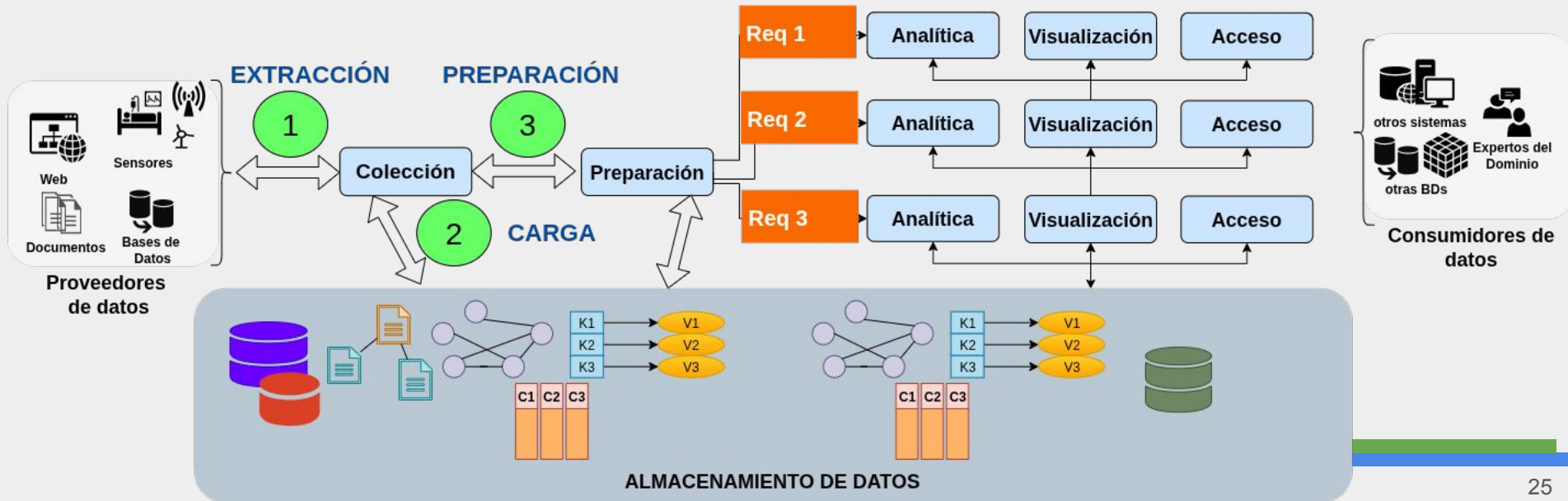
- Proceso **ELT**: *extracción-carga-transformación*



# PROCESO DE ANÁLISIS DE DATOS

## Proceso ELT

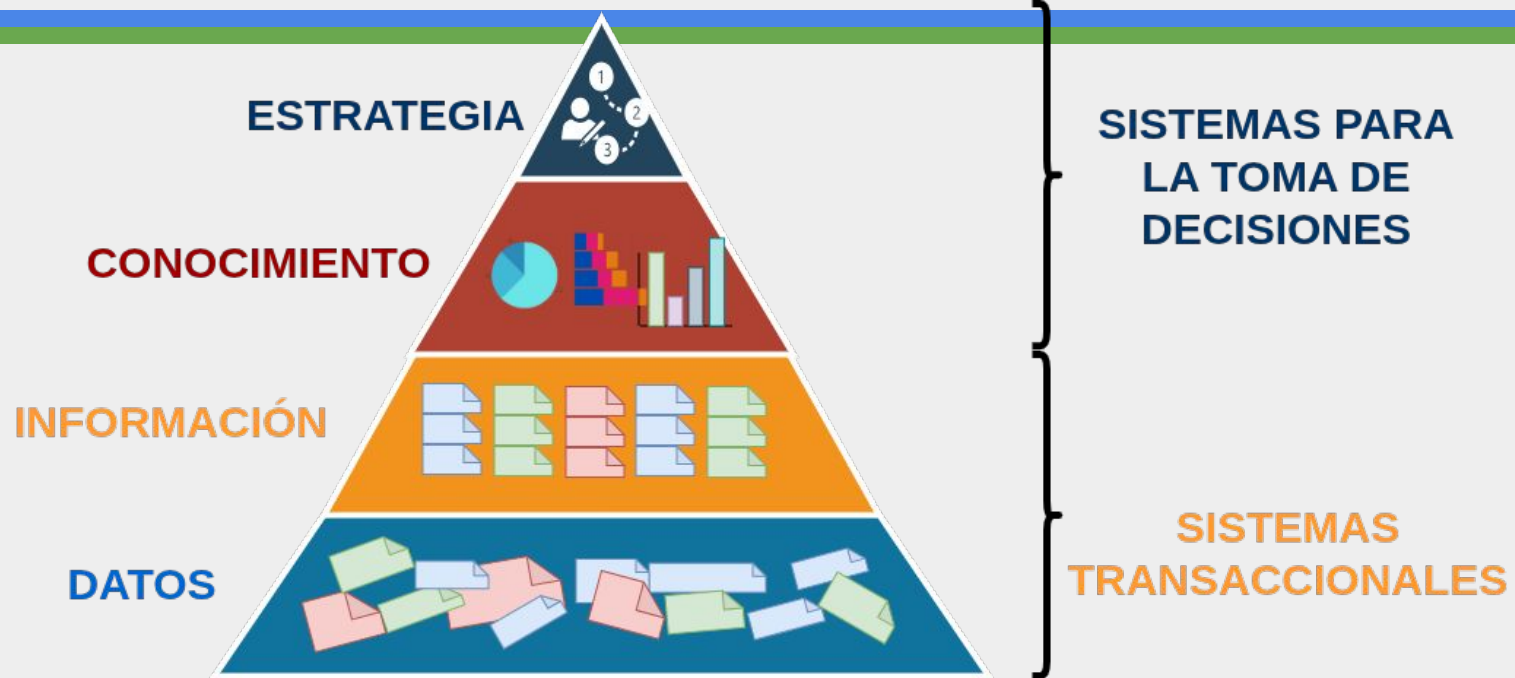
- Proceso **ELT**: *extracción-carga-transformación*



# PROCESO DE ANÁLISIS DE DATOS

| ETL                    | vs            | ELT                      |
|------------------------|---------------|--------------------------|
| RIGIDO Y DETALLADO     | ESQUEMA       | RAPIDO Y FLEXIBLE        |
| ALTO                   | COSTO         | BAJO                     |
| POCO/MEDIO VOLUMEN     | DATOS         | GRANDES VOLUMNEES        |
| ALTA                   | SEGURIDAD     | MENOS CONTROLADA         |
| ESCALABILIDAD VERTICAL | ESCALABILIDAD | ESCALABILIDAD HORIZONTAL |

# Tipos de Sistemas



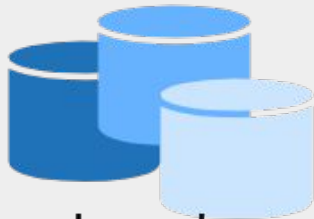
# Tipos de Sistemas

---

- Podemos distinguir entre dos tipos fundamentales:
  - **Sistemas Transaccionales:** gestión de las transacciones diarias, automatización de los procesos empresariales.
  - **Sistemas para la Toma de Decisiones:** proporcionar una visión amplia, integral y profunda de los datos de una organización, de forma de tomar decisiones informadas.

# Soporte según tipos de sistemas

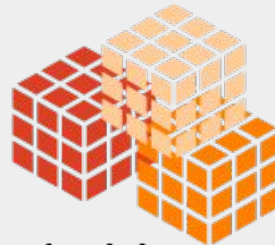
**OLTP (OnLine Transaction Processing):** Procesamiento en línea para el manejo transaccional



bases de  
datos SQL

bancos, aerolíneas,  
universidades,  
seguros, etc.

**OLAP (OnLine Analytical Processing):** Procesamiento en línea para la toma de decisiones



depósitos  
de datos

información agregada,  
gerencial, toma de decisiones

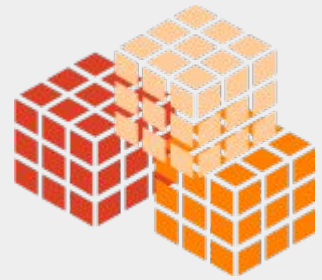


lagos de  
datos

# PROCESO DE ANÁLISIS DE DATOS

## Depósitos de Datos - ETL

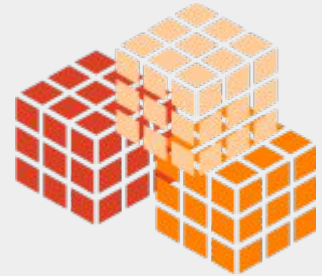
- Los **Depósitos de Datos o Data Warehouses** son OTRO tipo de almacenamiento, con las características de:
  - Una base de datos diseñada para tareas analíticas
  - Con datos de múltiples aplicaciones
  - Uso de datos especialmente para la lectura
  - Interacción directa con el sistema sin asistencia IT
  - Contenido modificado periódicamente y estable



# PROCESO DE ANÁLISIS DE DATOS

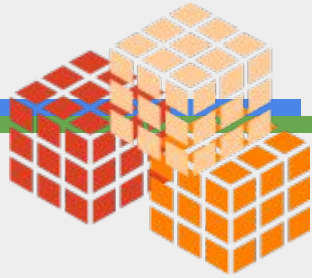
## Depósitos de Datos - ETL

- Los **Depósitos de Datos** tienen las características de:
  - Contenido que incluya datos actuales e históricos
  - Habilidad de los usuarios para ejecutar consultas y obtener resultados en línea
  - Habilidad de los usuarios de realizar reportes



# PROCESO DE ANÁLISIS DE DATOS

## Depósitos de Datos - ETL



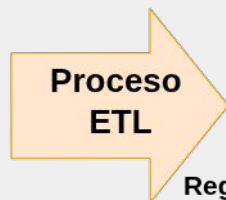
- Un **Depósito de Datos** es un entorno que:
  - Provee una vista integrada y total de la empresa
  - Hace que la información actual e histórica esté fácilmente disponible para la toma de decisiones estratégicas
  - Hace posibles transacciones de soporte a las decisiones sin afectar los sistemas operacionales
  - Es una fuente de información estratégica y consistente

## Datos transaccionales

| paquetes |         |       |
|----------|---------|-------|
| P1       | perfume | ..... |
| P2       | remera  | ....  |
| P3       | vino    | ....  |
| ....     | ....    | ....  |

| envios |                 |            |              |              |
|--------|-----------------|------------|--------------|--------------|
| P1     | Jet747          | 12/12/2025 | Buenos Aires | Buenos Aires |
| P2     | Tren4           | 13/12/2025 | New York     | New York     |
| P3     | Royal Caribbean | 13/12/2025 | Paris        | Paris        |
| ....   | ....            | ....       | ....         | ....         |

| ciudades     |         |       |
|--------------|---------|-------|
| Buenos Aires | ARG     | ..... |
| New York     | USA     | ....  |
| Paris        | Francia | ....  |
| ....         | ....    | ....  |



Región

Emisferio Oriental

Emisferio Occidental

Asia

Africa

Australia

Francia

América Sur

América Norte

Ruta

Agua

mar

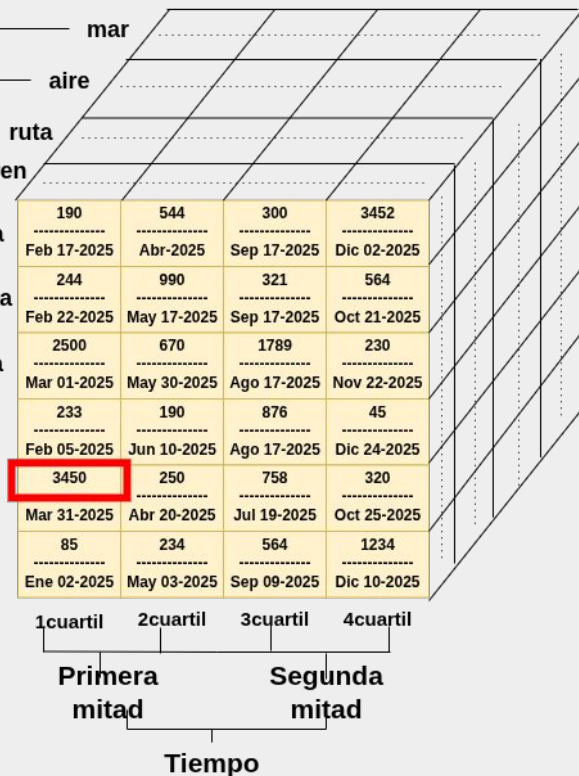
aire

Tierra

ruta

tren

## Datos agregados y combinados



La cantidad de envios (medida) al final del 1er trimestre (dimensión 1) en América del Sur (dimensión 2) por tren (dimensión 3) es **3450**

# PROCESO DE ANÁLISIS DE DATOS

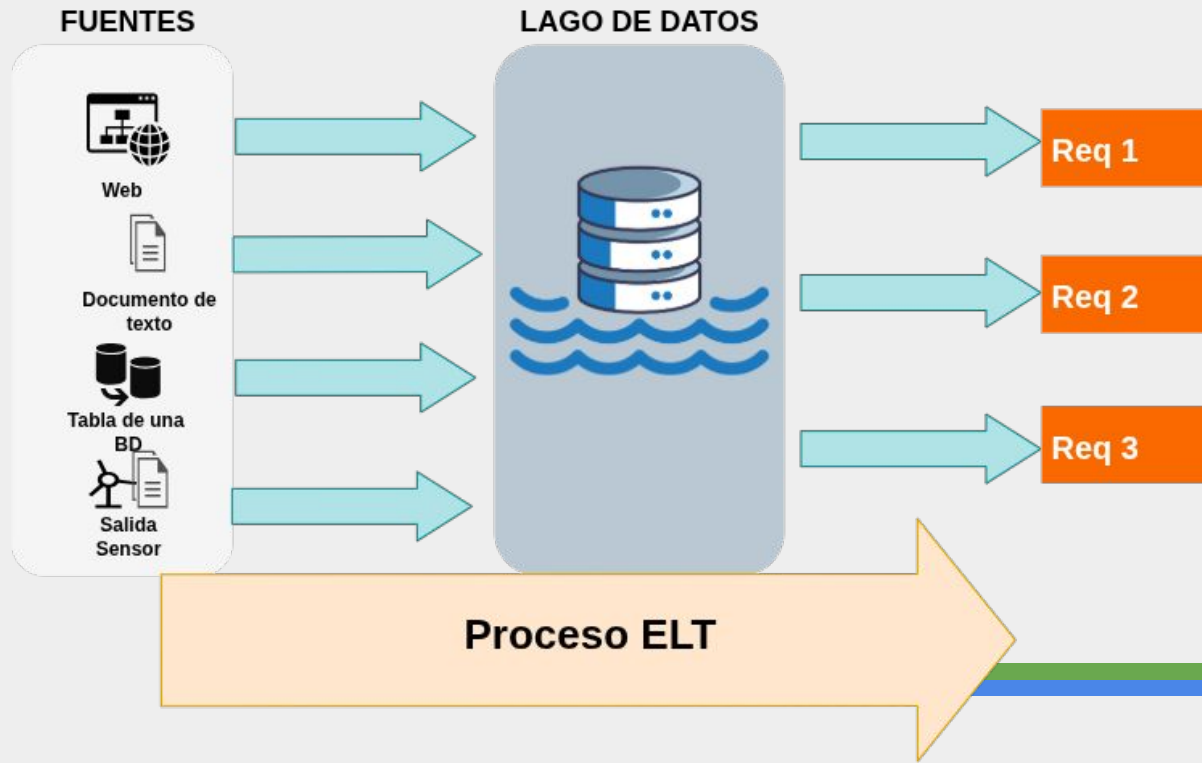
## Lago de Datos - ELT

Un **lago de datos** es un repositorio (para almacenamiento) que contiene **grandes cantidades de datos de varias fuentes en un formato granular y sin procesar**. Puede guardar datos estructurados, semiestructurados o no estructurados, lo que significa que los datos pueden conservarse en un formato más flexible para usarlos en un futuro. Al guardar datos, un lago de datos los asocia con identificadores y etiquetas de metadatos para poder extraerlos más rápidamente



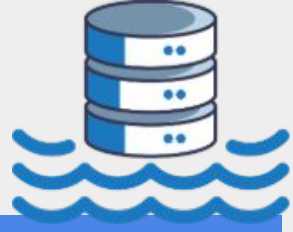
# PROCESO DE ANÁLISIS DE DATOS

## Lago de Datos - ELT



# PROCESO DE ANÁLISIS DE DATOS

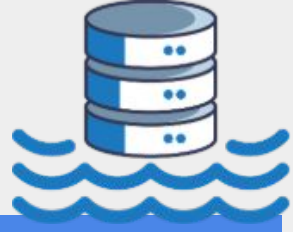
## Lago de Datos - ELT



- Un **Lago de Datos** es un repositorio de datos muy diferente:
  - Los datos se almacenan en su formato original, no se transforman hasta que las aplicaciones los llaman
  - Se evita la transformación costosa previa de datos.
  - Las operaciones de transformación solo se realizarán cuando los datos se lean desde el lago de datos.
  - El enfoque de los lagos de datos se denomina enfoque de **"Esquema de lectura" (Schema on read)**, es decir *la estructura (esquema) se define en la lectura*

# PROCESO DE ANÁLISIS DE DATOS

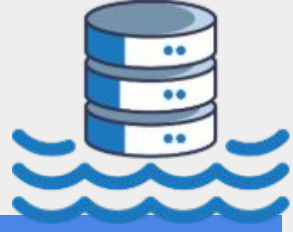
## Lago de Datos - ELT



- **Ventajas:**
  - Permiten que los usuarios tengan acceso inmediato a toda la información almacenada en el lago.
  - Permite diferentes tipos de fuentes de información, no limitadas solo a datos relacionales o transaccionales.
  - Acelera la entrega permitiendo que las unidades de negocio alimenten las aplicaciones rápidamente.

# PROCESO DE ANÁLISIS DE DATOS

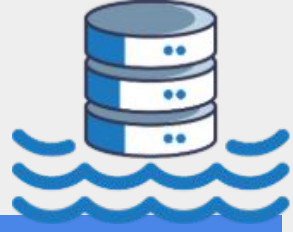
## Lago de Datos - ELT



- **Ventajas:**
  - Los datos se preparan *"según sea necesario"*, lo que reduce los costos de preparación sobre el procesamiento inicial.
  - Mayor flexibilidad al no necesitar todas las respuestas por adelantado.
  - Posibilidad de almacenar datos en bruto que pueden refinarse en el tiempo para que su comprensión mejore.

# PROCESO DE ANÁLISIS DE DATOS

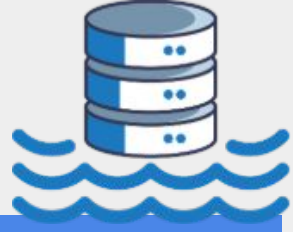
## Lago de Datos - ELT



- **Desventajas:**
  - Diferentes interpretaciones de usuario o aplicación de los datos que pueden entrar en conflicto.
  - Inexactitud en los metadatos haciendo difícil encontrar datos específicos.
  - Se debe considerar tiempo extra para garantizar seguridad. Las fuentes de datos del lago pueden contener información confidencial que requerirá la implementación de medidas de seguridad apropiadas en la organización.

# PROCESO DE ANÁLISIS DE DATOS

## Lago de Datos - ELT

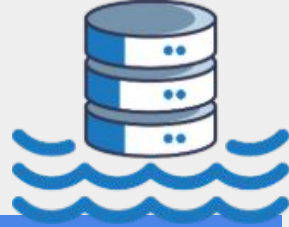
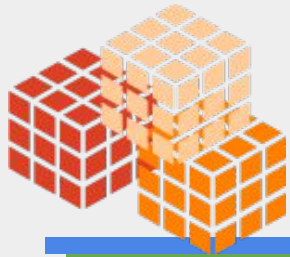


- **Desventajas:**

- Los datos se pueden volver obsoletos debido a un indiscriminado acaparamiento.
- Los datos pueden estar corruptos debido a la falta de comprobaciones iniciales.
- Los datos obsoletos y corruptos pueden convertir un lago de datos en un **"pantano de datos"**. La curación adecuada de los datos puede evitar que esto suceda, aunque esto significa un esfuerzo adicional.

# PROCESO DE ANÁLISIS DE DATOS

## DW vs Lago de Datos



DW

vs



Lago de  
Datos

Estructurados,  
datos procesados

DATOS

Sin estructura, sin  
procesar

Esquema de  
escritura

PROCESAMIENTO

Esquema de  
lectura

Caro y  
confiable

ALMACENAMIENTO

Barato y  
dudoso

Configuración  
robusta y fija

FACILIDAD  
USO

Configuración  
ligera y flexible

Madura

SEGURIDAD

Madurando

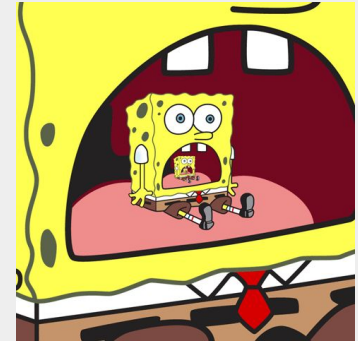
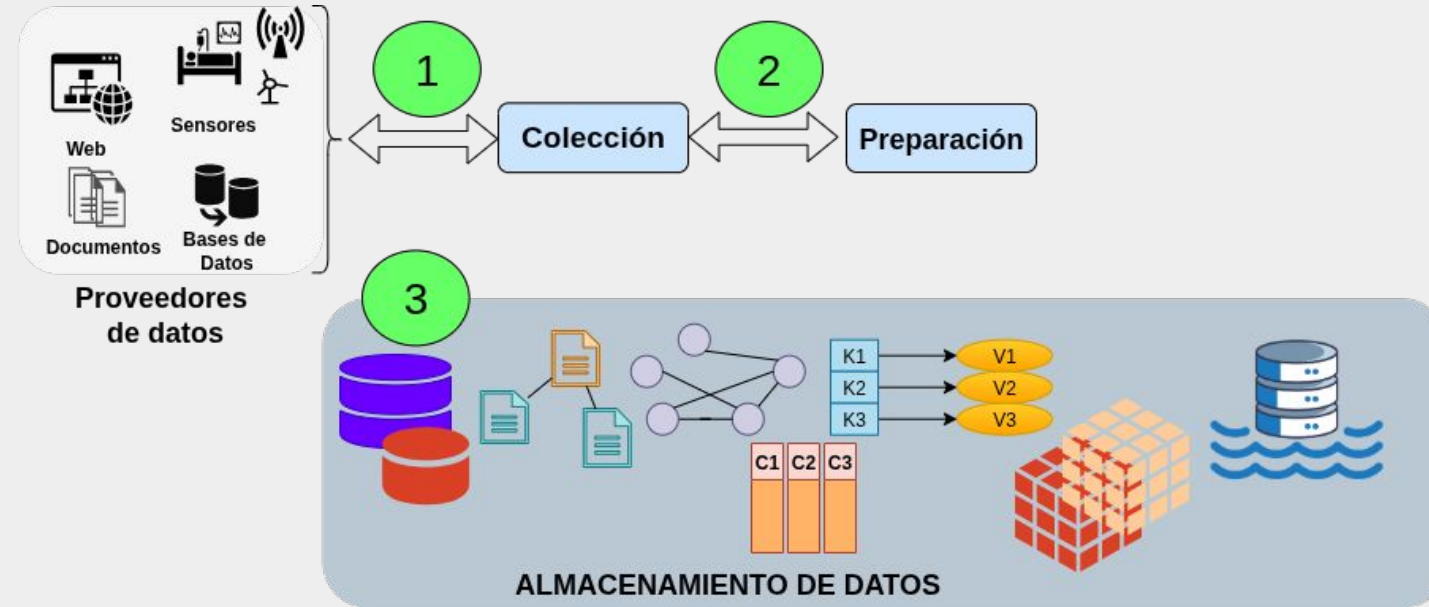
Profesionales,  
empresarios,  
CEOs

USUARIOS

Científicos de datos  
familiarizados con el  
dominio

# PROCESO DE ANÁLISIS DE DATOS

## Entonces? dónde almacenamos?



# PROCESO DE ANÁLISIS DE DATOS

## Almacenamiento de Datos

🍺 *Entran a un bar...*

🔑 **Una base de datos Clave-Valor,**  
📄 **una Orientada a Documentos,**  
📊 **una Columnar,**  
🔗 **una Relacional,**  
🏠 **un Data Warehouse,**  
🌊 **y un Data Lake.**

🍺 El bartender pregunta:  
—¿Qué van a tomar?

🔑 **Clave-Valor:**

—Lo de siempre, clave “cerveza”, valor “IPA”.

📄 **Orientada a Documentos:**

—Te paso un JSON con mi orden completa, toppings incluidos.

📊 **Columnar:**

—Solo quiero la columna “alcohol”, ignora el resto.

🔗 **Relacional:**

—Primero pásame la tabla de cervezas, después la de precios, y hago un *JOIN*.

🏠 **Data Warehouse:**

—¿Tienes algo preagregado por tipo, día y marca?

🌊 **Data Lake:**

—Tráeme todo lo que tengas... ya veré qué hago con eso mañana.

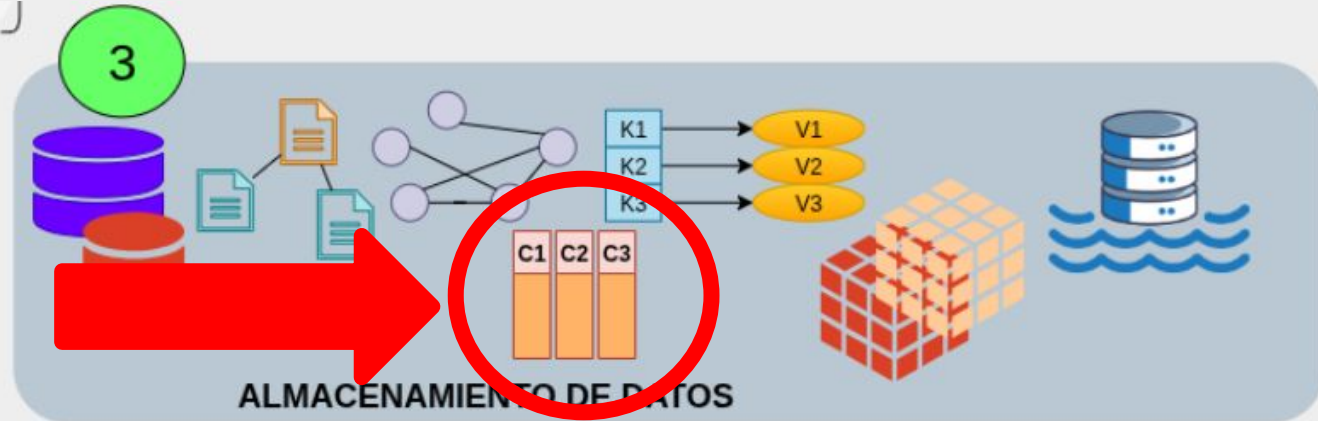
📁 **Mientras tanto, un CSV borracho grita desde la esquina:**

—¡Nadie me entiende si no me abren con Excel!

# PROCESO DE ANÁLISIS DE DATOS

## BDs Orientadas a columnas

- Una **BD orientada a columnas** es un repositorio de gestión datos que almacena su contenido por columnas, en lugar de por filas.
- La columna es el elemento base y está compuesto por un nombre y un valor.



# PROCESO DE ANÁLISIS DE DATOS

## BDs Orientadas a columnas

| LIBRO       |                |               |                |
|-------------|----------------|---------------|----------------|
| <u>ISBN</u> | titulo         | cantidadHojas | editorial      |
| 9-983-33    | Compiladores   | 936           | McGraw-Hill    |
| 5-889-22    | Redes          | 1053          | Oxford         |
| 4-432-11    | Bases de Datos | 867           | Addison-Wesley |

Aunque las columnas se almacenan **por separado**, los valores están **ordenados y alineados por el número de fila original** (índice posicional).

### BD Orientadas a Columnas

| <u>ISBN</u> | titulo         | cantidadHojas | editorial      |
|-------------|----------------|---------------|----------------|
| 9-983-33    | Compiladores   | 936           | McGraw-Hill    |
| 5-889-22    | Redes          | 1053          | Oxford         |
| 4-432-11    | Bases de Datos | 867           | Addison-Wesley |

# PROCESO DE ANÁLISIS DE DATOS

## BDs Orientadas a columnas

| LIBRO       |                |               |                |
|-------------|----------------|---------------|----------------|
| <u>ISBN</u> | titulo         | cantidadHojas | editorial      |
| 9-983-33    | Compiladores   | 936           | McGraw-Hill    |
| 5-889-22    | Redes          | 1053          | Oxford         |
| 4-432-11    | Bases de Datos | 867           | Addison-Wesley |

Por Filas

|          |                |      |                |
|----------|----------------|------|----------------|
| 9-983-33 | Compiladores   | 936  | McGraw-Hill    |
| 5-889-22 | Redes          | 1053 | Oxford         |
| 4-432-11 | Bases de Datos | 867  | Addison-Wesley |

Por Columnas

|              |          |                |
|--------------|----------|----------------|
| 9-983-33     | 5-889-22 | 4-432-11       |
| Compiladores | Redes    | Bases de Datos |
| 936          | 1053     | 867            |
| McGraw-Hill  | Oxford   | Addison-Wesley |

# PROCESO DE ANÁLISIS DE DATOS

| Relacionales                          | vs              | Columnas                               |
|---------------------------------------|-----------------|----------------------------------------|
| Rápido para transacciones             | QUERY           | Rápido para agregaciones               |
| Baja debido a heterogeneidad de filas | COMPRESIÓN      | Alta debido a homogeneidad de columnas |
| Rápida                                | ACTUALIZACIONES | Lenta, modifica muchas columnas        |
| Rapido                                | JOINS           | Requiere índices                       |
| ACID                                  | TRANSACCIONES   | No hay transacciones                   |

# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

- **Maria DB Column Store** es un repositorio de datos orientado a columnas:
  - Es otro ENGINE, en el curso de BD vieron InnoDB
  - Ofrece una compresión del 90%, lo que reduce la E/S a disco y el tamaño de los datos en disco.
  - Los datos se pueden distribuir en múltiples instancias y consultar en paralelo para aumentar el rendimiento de las consultas y dar soporte a grandes conjuntos de datos.

# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

### Row-oriented vs. Column-oriented format

SELECT Fname FROM Table 1 WHERE State = 'NY'

| ID | Fname    | Lname | State | Zip   | Phone          | Age | Sex |
|----|----------|-------|-------|-------|----------------|-----|-----|
| 1  | Bugs     | Bunny | NY    | 11217 | (718) 938-3235 | 34  | M   |
| 2  | Yosemite | Sam   | CA    | 95389 | (209) 375-6572 | 52  | M   |
| 3  | Daffy    | Duck  | NY    | 10013 | (212) 227-1810 | 35  | M   |
| 4  | Elmer    | Fudd  | ME    | 04578 | (207) 882-7323 | 43  | M   |
| 5  | Witch    | Hazel | MA    | 01970 | (978) 744-0991 | 57  | F   |

| ID | Fname    | Lname | State | Zip   | Phone          | Age | Sex |
|----|----------|-------|-------|-------|----------------|-----|-----|
| 1  | Bugs     | Bunny | NY    | 11217 | (718) 938-3235 | 34  | M   |
| 2  | Yosemite | Sam   | CA    | 95389 | (209) 375-6572 | 52  | M   |
| 3  | Daffy    | Duck  | NY    | 10013 | (212) 227-1810 | 35  | M   |
| 4  | Elmer    | Fudd  | ME    | 04578 | (207) 882-7323 | 43  | M   |
| 5  | Witch    | Hazel | MA    | 01970 | (978) 744-0991 | 57  | F   |

- **Row oriented**
  - Rows stored sequentially in a file
  - Scans through every record row by row
- **Column oriented**
  - Each column is stored in a separate file
  - Scans the only relevant column



# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

### Single-Row Operations - Insert

| Key | Fname    | Lname   | State | Zip   | Phone          | Age | Sex |
|-----|----------|---------|-------|-------|----------------|-----|-----|
| 1   | Bugs     | Bunny   | NY    | 11217 | (718) 938-3235 | 34  | M   |
| 2   | Yosemite | Sam     | CA    | 95389 | (209) 375-6572 | 52  | M   |
| 3   | Daffy    | Duck    | NY    | 10013 | (212) 227-1810 | 35  | M   |
| 4   | Elmer    | Fudd    | ME    | 04578 | (207) 882-7323 | 43  | M   |
| 5   | Witch    | Hazel   | MA    | 01970 | (978) 744-0991 | 57  | F   |
| 6   | Marvin   | Martian | CA    | 91602 | (818) 761-9964 | 26  | M   |

| Key | Fname    | Lname   | State | Zip   | Phone          | Age | Sex |
|-----|----------|---------|-------|-------|----------------|-----|-----|
| 1   | Bugs     | Bunny   | NY    | 11217 | (718) 938-3235 | 34  | M   |
| 2   | Yosemite | Sam     | CA    | 95389 | (209) 375-6572 | 52  | M   |
| 3   | Daffy    | Duck    | NY    | 10013 | (212) 227-1810 | 35  | M   |
| 4   | Elmer    | Fudd    | ME    | 04578 | (207) 882-7323 | 43  | M   |
| 5   | Witch    | Hazel   | MA    | 01970 | (978) 744-0991 | 57  | F   |
| 6   | Marvin   | Martian | CA    | 91602 | (818) 761-9964 | 26  | M   |

#### Row oriented:

new rows appended to the end.

#### Column oriented:

new value added to each file.



# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

### Single-Row Operations - Update

| Key | Fname    | Lname | State | Zip   | Phone          | Age | Sex |
|-----|----------|-------|-------|-------|----------------|-----|-----|
| 1   | Bugs     | Bunny | NY    | 11217 | (718) 938-3235 | 34  | M   |
| 2   | Yosemite | Sam   | CA    | 95389 | (209) 375-6572 | 52  | M   |
| 3   | Daffy    | Duck  | NY    | 10013 | (212) 227-1810 | 35  | M   |
| 4   | Elmer    | Fudd  | ME    | 04578 | (207) 882-7323 | 43  | M   |
| 5   | Witch    | Hazel | MA    | 01970 | (978) 744-0991 | 57  | F   |

| Key | Fname    | Lname | State | Zip   | Phone          | Age | Sex |
|-----|----------|-------|-------|-------|----------------|-----|-----|
| 1   | Bugs     | Bunny | NY    | 11217 | (718) 938-3235 | 34  | M   |
| 2   | Yosemite | Sam   | CA    | 95389 | (209) 375-6572 | 52  | M   |
| 3   | Daffy    | Duck  | NY    | 10013 | (212) 227-1810 | 35  | M   |
| 4   | Elmer    | Fudd  | ME    | 04578 | (207) 882-7323 | 43  | M   |
| 5   | Witch    | Hazel | MA    | 01970 | (978) 744-0991 | 57  | F   |

#### Row oriented:

Update 100% of rows  
means change 100%  
of blocks on disk.

#### Column oriented:

Just update the blocks  
needed to be updated



# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

### Single-Row Operations - Delete

| Key | Fname    | Lname   | State | Zip   | Phone          | Age | Sex |
|-----|----------|---------|-------|-------|----------------|-----|-----|
| 1   | Bugs     | Bunny   | NY    | 11217 | (718) 938-3235 | 34  | M   |
| 2   | Yosemite | Sam     | CA    | 95389 | (209) 375-6572 | 52  | M   |
| 3   | Daffy    | Duck    | NY    | 10013 | (212) 227-1810 | 35  | M   |
| 4   | Elmer    | Fudd    | ME    | 04578 | (207) 882-7323 | 43  | M   |
| 5   | Witch    | Hazel   | MA    | 01970 | (978) 744-0991 | 57  | F   |
| 6   | Marvin   | Martian | CA    | 91602 | (818) 761-9964 | 26  | M   |

**Row oriented:**  
Entire rows deleted

| Key | Fname    | Lname   | State | Zip   | Phone          | Age | Sex |
|-----|----------|---------|-------|-------|----------------|-----|-----|
| 1   | Bugs     | Bunny   | NY    | 11217 | (718) 938-3235 | 34  | M   |
| 2   | Yosemite | Sam     | CA    | 95389 | (209) 375-6572 | 52  | M   |
| 3   | Daffy    | Duck    | NY    | 10013 | (212) 227-1810 | 35  | M   |
| 4   | Elmer    | Fudd    | ME    | 04578 | (207) 882-7323 | 43  | M   |
| 5   | Witch    | Hazel   | MA    | 01970 | (978) 744-0991 | 57  | F   |
| 6   | Marvin   | Martian | CA    | 91602 | (818) 761-9964 | 26  | M   |

**Column oriented:**  
Value deleted from  
each file



# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

### Changing the table structure

| Key | Fname    | Lname | State | Zip   | Phone          | Age | Sex | Active |
|-----|----------|-------|-------|-------|----------------|-----|-----|--------|
| 1   | Bugs     | Bunny | NY    | 11217 | (718) 938-3235 | 34  | M   | Y      |
| 2   | Yosemite | Sam   | CA    | 95389 | (209) 375-6572 | 52  | M   | N      |
| 3   | Daffy    | Duck  | NY    | 10013 | (212) 227-1810 | 35  | M   | N      |
| 4   | Elmer    | Fudd  | ME    | 04578 | (207) 882-7323 | 43  | M   | Y      |
| 5   | Witch    | Hazel | MA    | 01970 | (978) 744-0991 | 57  | F   | N      |

| Key | Fname    | Lname | State | Zip   | Phone          | Age | Sex | Active |
|-----|----------|-------|-------|-------|----------------|-----|-----|--------|
| 1   | Bugs     | Bunny | NY    | 11217 | (718) 938-3235 | 34  | M   | Y      |
| 2   | Yosemite | Sam   | CA    | 95389 | (209) 375-6572 | 52  | M   | N      |
| 3   | Daffy    | Duck  | NY    | 10013 | (212) 227-1810 | 35  | M   | N      |
| 4   | Elmer    | Fudd  | ME    | 04578 | (207) 882-7323 | 43  | M   | Y      |
| 5   | Witch    | Hazel | MA    | 01970 | (978) 744-0991 | 57  | F   | N      |

#### Row oriented:

Requires rebuilding of the whole table

#### Column oriented:

Create new file for the new column



# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

- Cuando y porque usar **Maria DB Column Store**:
  - Conjuntos de datos muy grandes
    - Muchas columnas
    - Muchos millones de filas
  - Agregaciones sobre gran cantidad de datos
  - Rápida inserción de gran cantidad de datos
    - Cuanto mayor sea el lote, mejor

# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

- Cuando y porque usar **Maria DB Column Store**:
  - *Escalabilidad Horizontal*: Entorno distribuido
  - Solo se procesan y devuelven las columnas requeridas, lo que optimiza la ejecución de la consulta
  - Son (casi) ACID ya que no soporta un modelo de transacciones tradicional (no permite múltiples operaciones en una transacción)

# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

- Características de **Maria DB Column Store**:
  - Usa SQL, pero hay que indicarle `ENGINE=ColumnStore;` al crear tablas
  - No soporta claves primarias, ni claves foráneas.
  - Las inserciones se pueden hacer como un `INSERT INTO` tradicional, pero se recomienda un volcado masivo con `LOAD DATA INFILE`
  - Los joins entre tablas diferentes se dividen en la forma en la que se efectúan: si se hacen en disco o en memoria. Dependiendo si se necesita disco para datos intermedios

# PROCESO DE ANÁLISIS DE DATOS

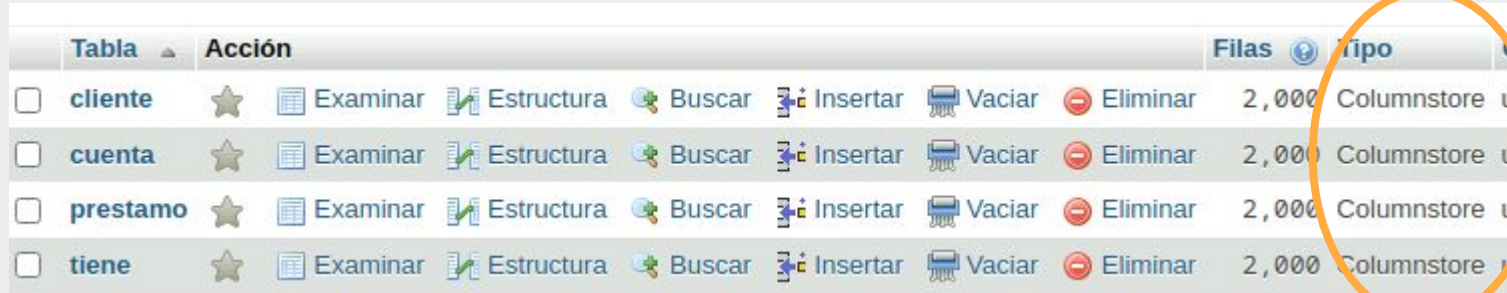
## Maria DB ColumnStore

- Características de **Maria DB Column Store**:
  - Como las operaciones DML (INSERTAR, ACTUALIZAR y ELIMINAR) permiten realizar cambios a nivel de fila, son más lentas en ColumnStore que en InnoDB.
  - ColumnStore está optimizado para modificaciones masivas, por lo que estas operaciones son más lentas.
  - Documentación: <https://mariadb.com/docs/columnstore/reference>

# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

- Características de **Maria DB Column Store**:
  - No hay diferencia en la visualización de las tablas ya sean InnoDB o ColumnStore. Se ven todas igual.
  - Entonces, cómo hago para saber qué motor está usando?



| Tabla                             | Acción                                                | Filas | Tipo          |
|-----------------------------------|-------------------------------------------------------|-------|---------------|
| <input type="checkbox"/> cliente  | ★ Examinar Estructura Buscar Insertar Vaciar Eliminar | 2,000 | Columnstore u |
| <input type="checkbox"/> cuenta   | ★ Examinar Estructura Buscar Insertar Vaciar Eliminar | 2,000 | Columnstore u |
| <input type="checkbox"/> prestamo | ★ Examinar Estructura Buscar Insertar Vaciar Eliminar | 2,000 | Columnstore u |
| <input type="checkbox"/> tiene    | ★ Examinar Estructura Buscar Insertar Vaciar Eliminar | 2,000 | Columnstore u |

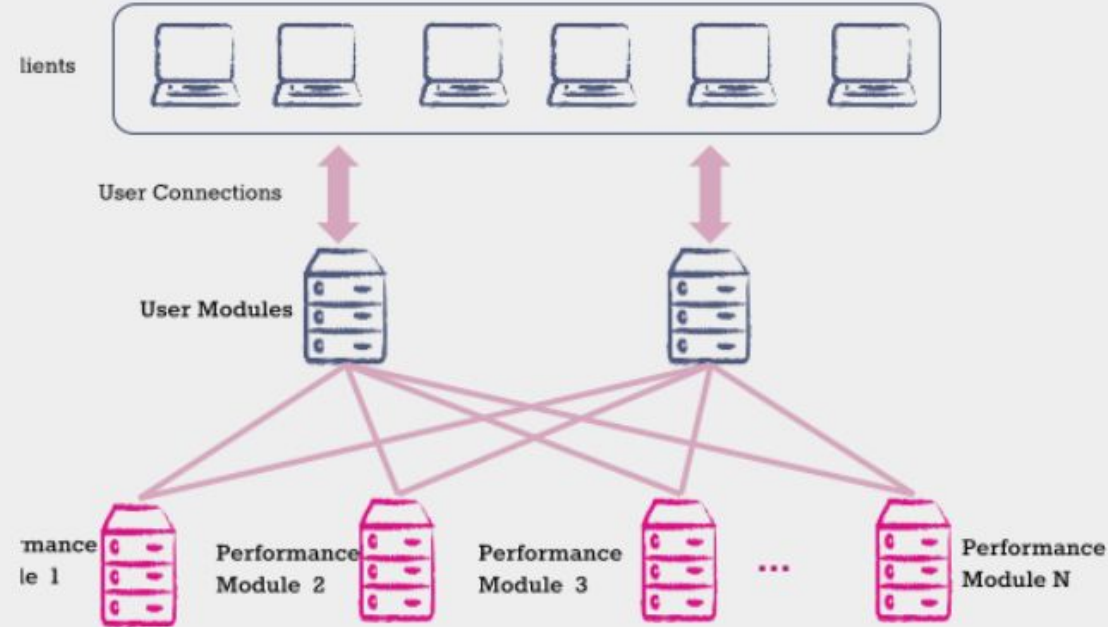
# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore

- Características de **Maria DB Column Store**:
  - Entonces la diferencia va a estar sobre todo en **las operaciones que vamos a realizar** dependiendo el motor utilizado, ya que debemos conocer qué operaciones son mejores para qué motor.

# PROCESO DE ANÁLISIS DE DATOS

## Maria DB ColumnStore



- \* Cuando un cliente hace una consulta, el **MODULO DE USUARIO (UM)** la interpreta y divide el trabajo.
- \* Luego, los **MÓDULOS DE PERFORMANCE (PMs)** ejecutan los fragmentos de la consulta en paralelo sobre los datos que tienen.
- \* Finalmente, **UM** recolecta los resultados y los devuelve al cliente

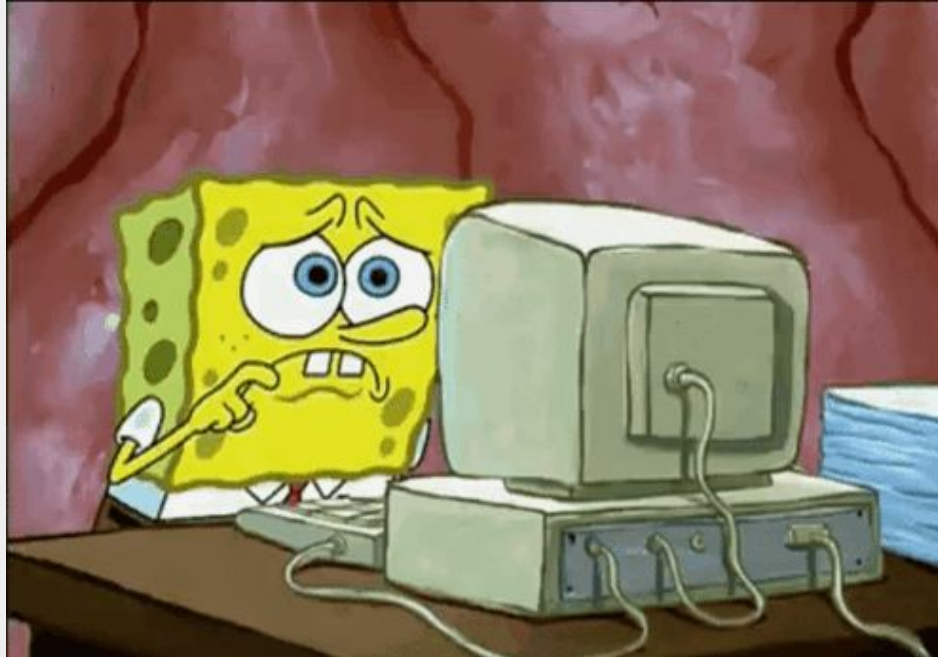
# PROCESO DE ANÁLISIS DE DATOS

## Maria DB: InnoDB vs ColumnStore

---

- Qué diferencias podemos encontrar?
  - vemos el resumen [aca](#)

# PROCESO DE ANÁLISIS DE DATOS A TRABAJAR!!



# PROCESO DE ANÁLISIS DE DATOS y ahora?

