Especialización en Inteligencia de Datos Aplicada

Facultad de Informática
Universidad Nacional del Comahue

EXTRACCIÓN, PREPARACIÓN Y ALMACENAMIENTO DE LOS DATOS

Segundo Cuatrimestre
Unidad I

EXTRACCIÓN, PREPARACIÓN Y ALMACENAMIENTO DE LOS DATOS

CONTENIDOS MÍNIMOS

Captura de datos. Fuente de datos. Captura de datos en la Web. Almacenamiento en bases de datos relacionales y no relacionales. Repositorios NoSQL. Depósitos de datos. Limpieza de datos. Lagos de datos.

UNIDAD I

IDENTIFICACIÓN Y RECOLECCIÓN DE DATOS

Ingesta de datos. Tipos de datos fuente. Extracción de datos de la Web (Web Scraping, Web Crawling). Patrones de ingesta de datos en el contexto de Big Data.

UNIDAD II

PREPARACIÓN DE DATOS

Calidad de Datos. Limpieza de Datos. Ruido y detección de anomalías en los datos. El proceso ETL y ELT. Integración de los datos.

UNIDAD III

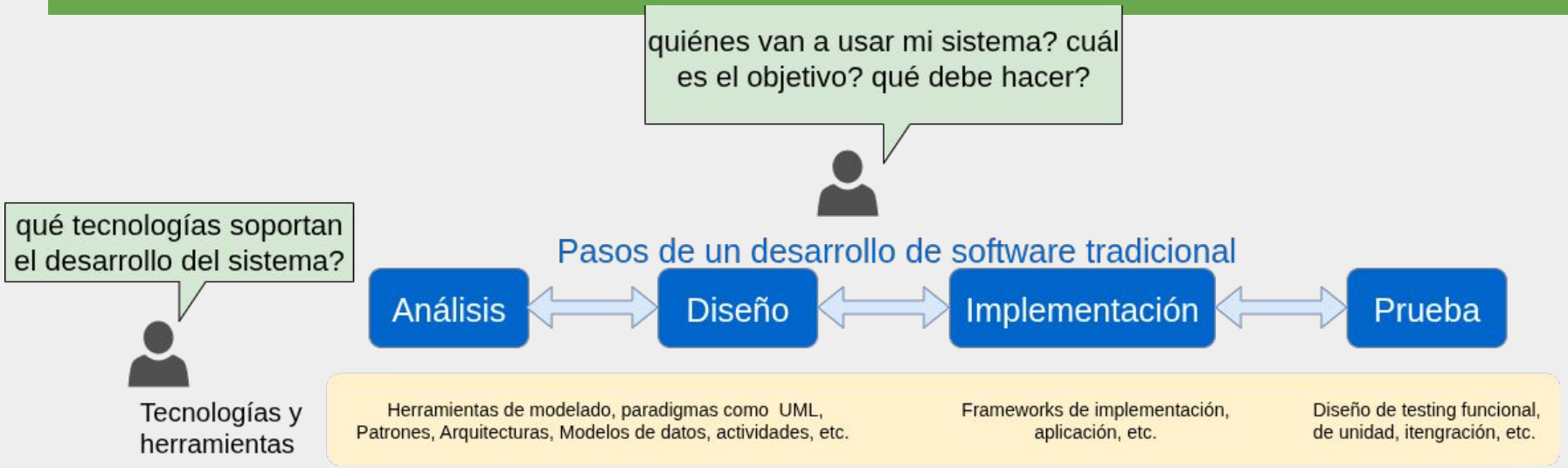
ALMACENAMIENTO DE DATOS

Tipos de almacenamiento. ACID-CAP-BASE. NoSQL, Distribuidas. Depósitos de Datos vs Lago de Datos. Conceptos del almacenamiento de grandes volúmenes de datos.

pero,

cómo comenzamos?

Cómo se construye un sistema?



- **Qué tipo de sistema voy a construir?**

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.**

Tipos de Sistemas

SISTEMAS PARA LA TOMA DE DECISIONES

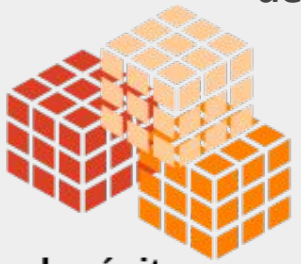
SISTEMAS TRANSACCIONALES

Soporte según tipos de sistemas

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

bases de datos SQL

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



depósitos de datos

lagos de datos

bancos, aerolineas, universidades, seguros, etc.

información agregada, gerencial, toma de desiciones

Ahora sí.....


FIELD
Cube

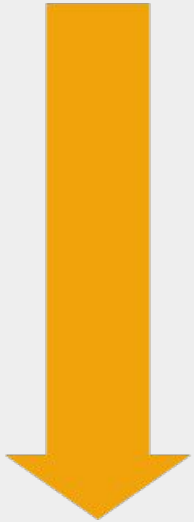

REITERAMOS
Crómica
16:21
26·5
HOY,
TE CONVERTIS
EN HEROE

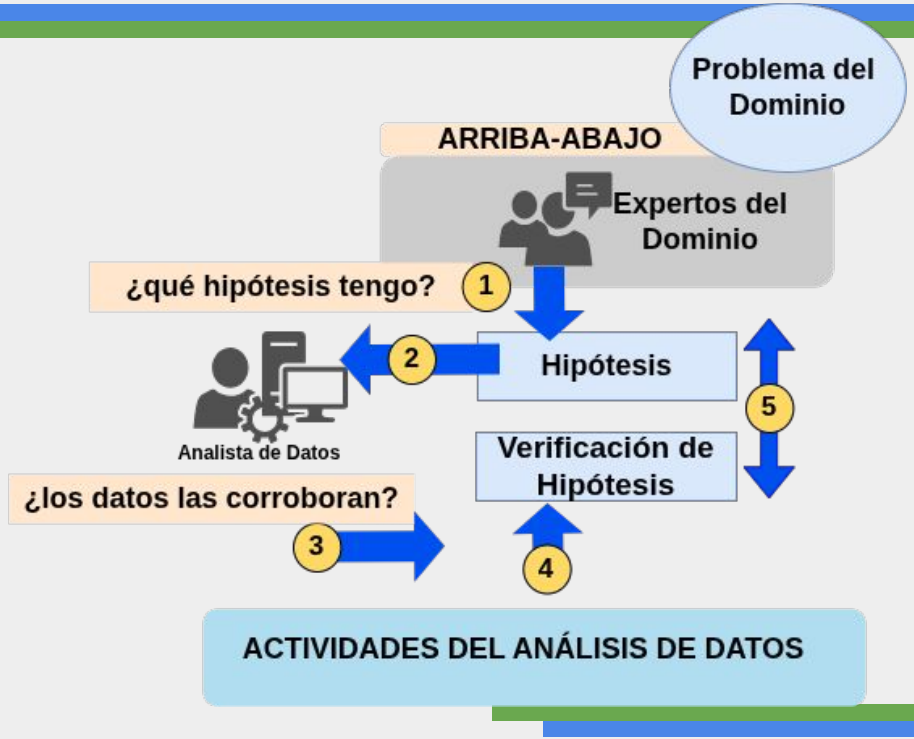
Sistemas para la Toma de Decisiones

- Se deben seguir una serie de pasos para su construcción
- Al conjunto de estos pasos lo podemos llamar

PROCESO DE ANÁLISIS DE DATOS
de arriba a abajo

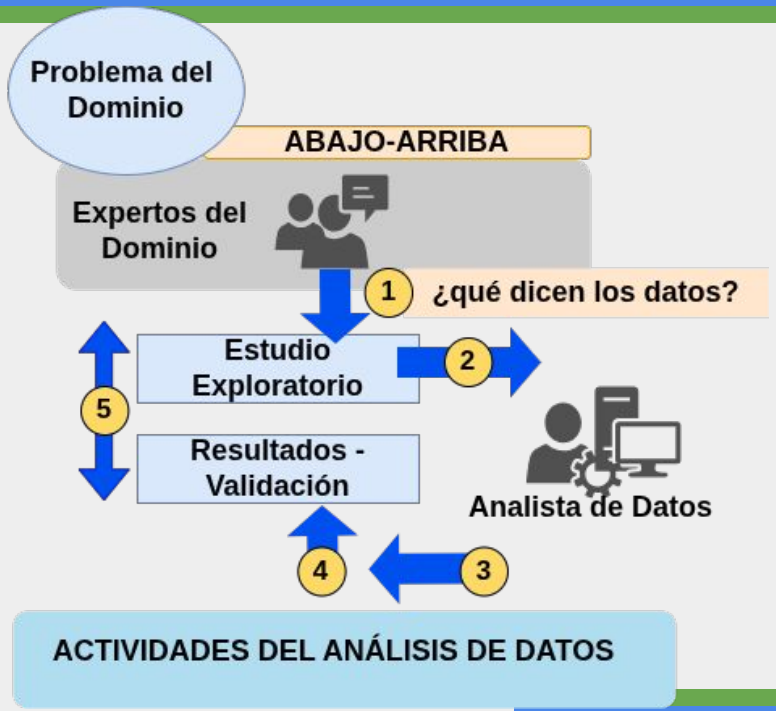
Análisis basado en Hipótesis (o inferencial)



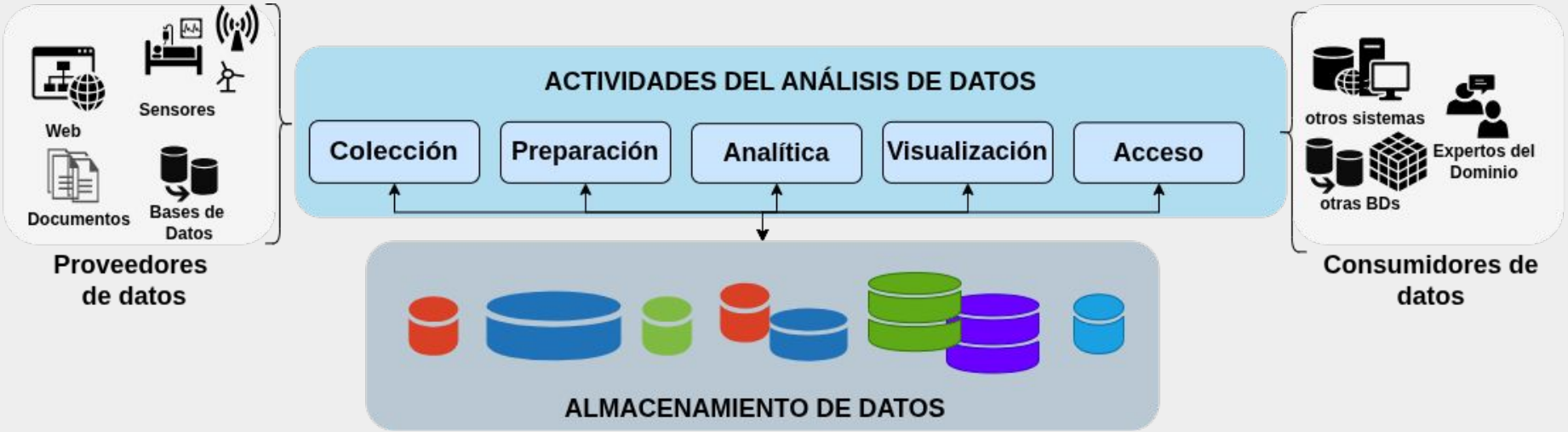


PROCESO DE ANÁLISIS DE DATOS
de abajo a arriba

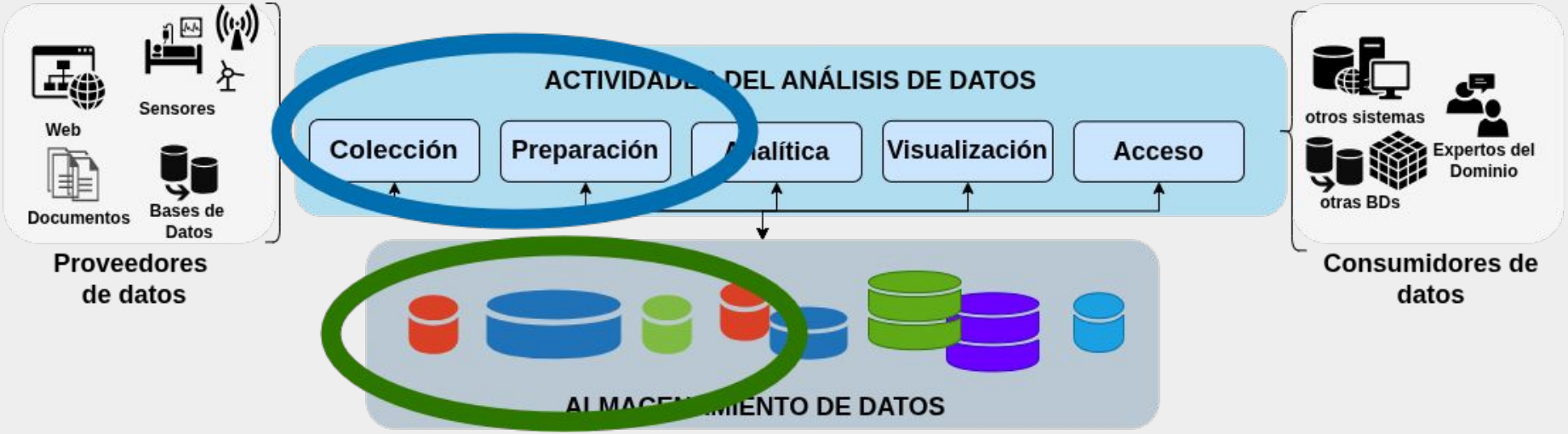
Análisis Exploratorio de Datos (EDA)



PROCESO DE ANÁLISIS DE DATOS
Actividades

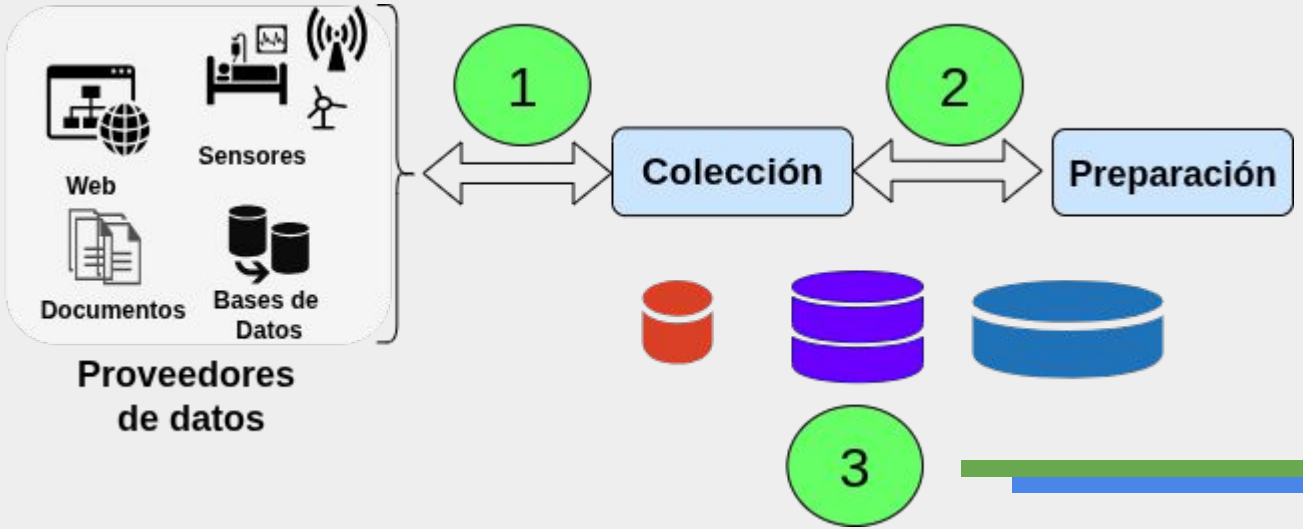


PROCESO DE ANÁLISIS DE DATOS
Actividades

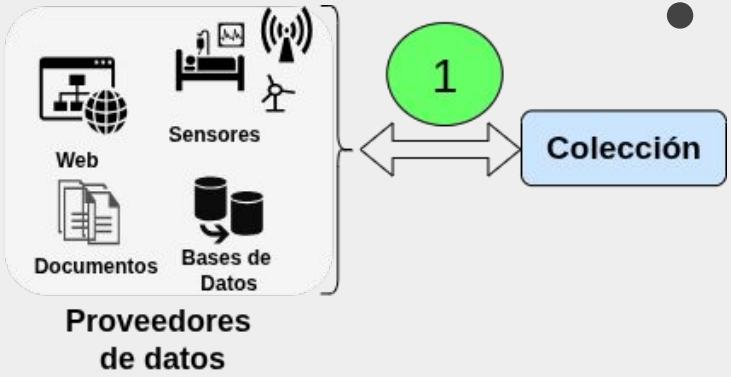


PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos

Es el proceso de coleccionar datos de diferentes fuentes y colocarlos en un repositorio destino



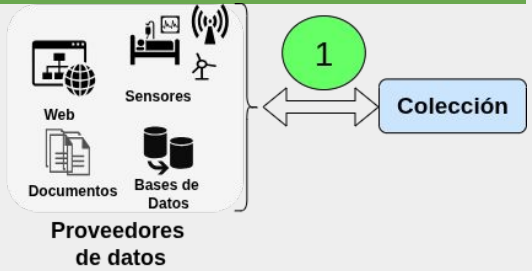
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos



- **Cómo deben ser identificados y recolectados los datos?**
  - Muchas fuentes de datos involucradas
  - Diferentes formatos de datos
  - y además….

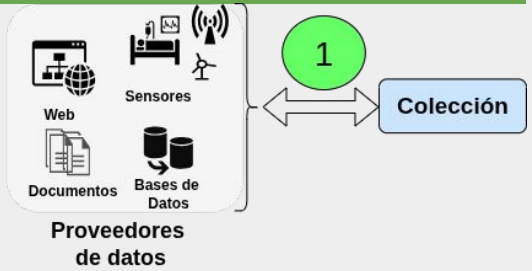
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos

- **Qué aspectos debemos considerar?**
  - ***Evolución de las fuentes***
    - *Las fuentes de datos pueden cambiar, por ejemplo las Web, estructuras de las BDs, nuevas fuentes a agregar.*



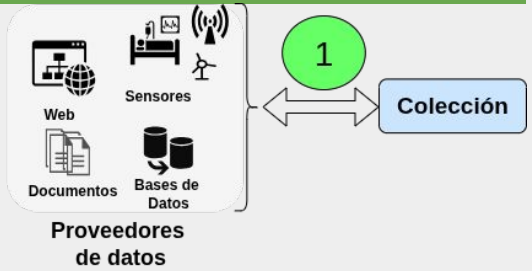
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos

- **Qué aspectos debemos considerar?**
  - ***Seguridad y Privacidad***
    - *De dónde obtengo los datos? pueden los usuarios conocer esa información?*



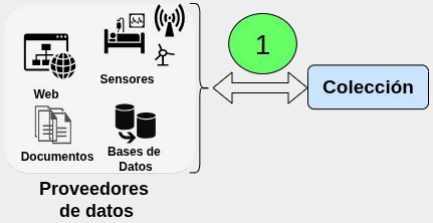
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos

- **Qué aspectos debemos considerar?**
  - ***Periodicidad de recolección***
    - *Con qué frecuencia las fuentes deben ser recuperadas/coleccionadas? De qué depende?*



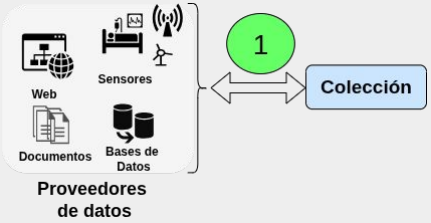
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos - Periodicidad

- La **periodicidad** en que se necesitan los datos se enmarca en tres modos específicos:
  - Por lotes (batch)
  - Tiempo real (stream)
  - Micro-lotes (micro-batch)



PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos - Lotes

- Se coleccionan grandes volúmenes de datos cada cierta cantidad de tiempo como horas, días, meses, etc.
- No requieren una acción inmediata.
- Los lotes pueden estar definidos por tamaño o por tiempo.
- ***Usos en Análisis*:** Análisis de datos históricos, predicciones, optimización, clasificación.

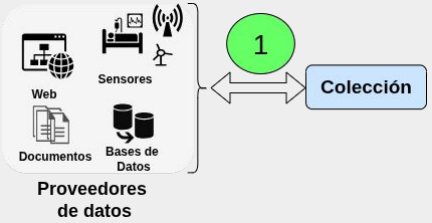


PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos - Tiempo real

- Se coleccionan los datos ni bien son generados (o casi) por las fuentes.
- Requieren una acción inmediata.
- Complejidad en cuanto al envío de la información que se genera (fallas de red, consistencia, envíos duplicados).
- ***Usos en Análisis***: detección de fraudes, logística (transporte), alertas, etc.

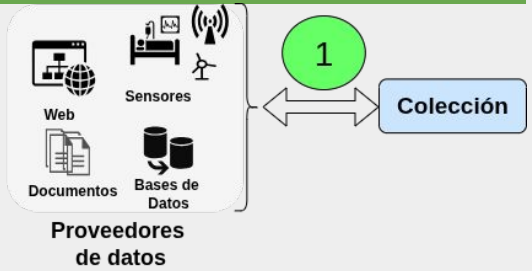
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos - Micro-Lotes

- Se coleccionan lotes más chicos y con más frecuencia.
- ***Usos en Análisis***: comportamientos de usuarios, minería de la web, etc.



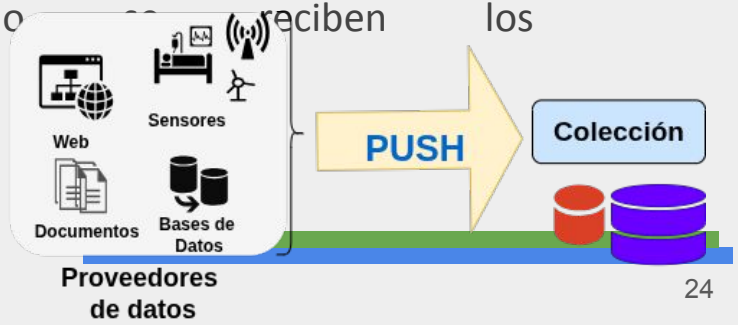
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos

- Otros tipos de patrones para la colección de los datos son:
  - **Pull** (extracción): el destino lee datos directamente de las fuentes
  - **Push** (inserción): las fuentes envían datos al destino.



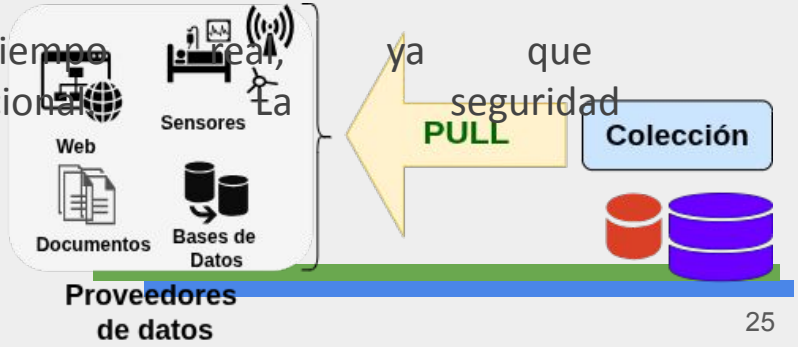
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos - Push

- Los datos se envían desde el origen (fuentes) al destino (sistema) en cuanto están disponibles.
- El origen puede generar muchos/pocos datos en corto/largo tiempo.
- **Ventajas**: *Tiempo real* ya que nuevos datos se envían al destino, *eficiente* cuando las fuentes producen datos constantemente y *seguridad* ya que se hace en destino.
- **Desventajas:** Es difícil controlar cómo reciben los datos en cuanto a su frecuencia, tamaño, etc.



PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos - Pull

- El destino envía una solicitud al origen y obtiene una respuesta con o sin datos.
- Permite al consumidor obtener datos de forma controlada. El consumidor puede retrasarse y recuperarse cuando lo desee.
- **Ventajas:** *escalabilidad* ya que se puede modificar según se desee y a preferencia del consumidor.
- **Desventajas:** No propicio para tiempo real, ya que la comunicación es bidireccional. La seguridad está en las fuentes.



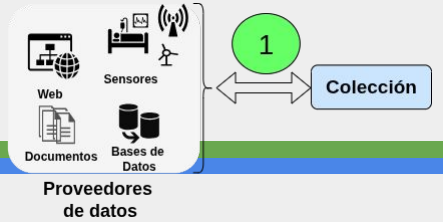
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos -Tipos de Datos

Proveedores de datos

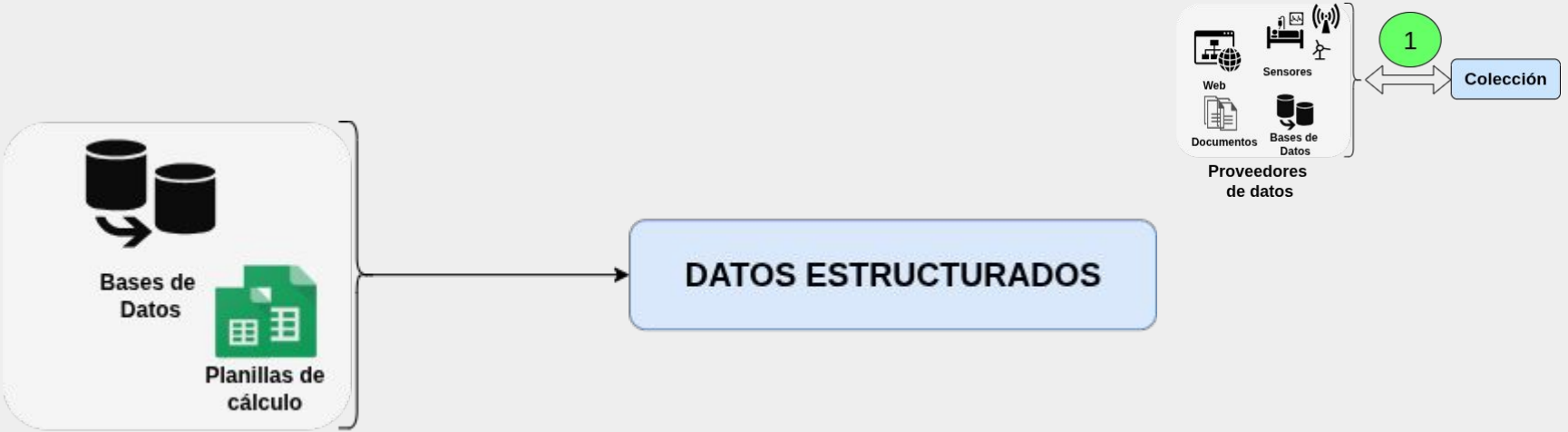
DATOS ESTRUCTURADOS — planillas de cálculo, datos de BDs

DATOS NO ESTRUCTURADOS — texto, imágenes, video, audio

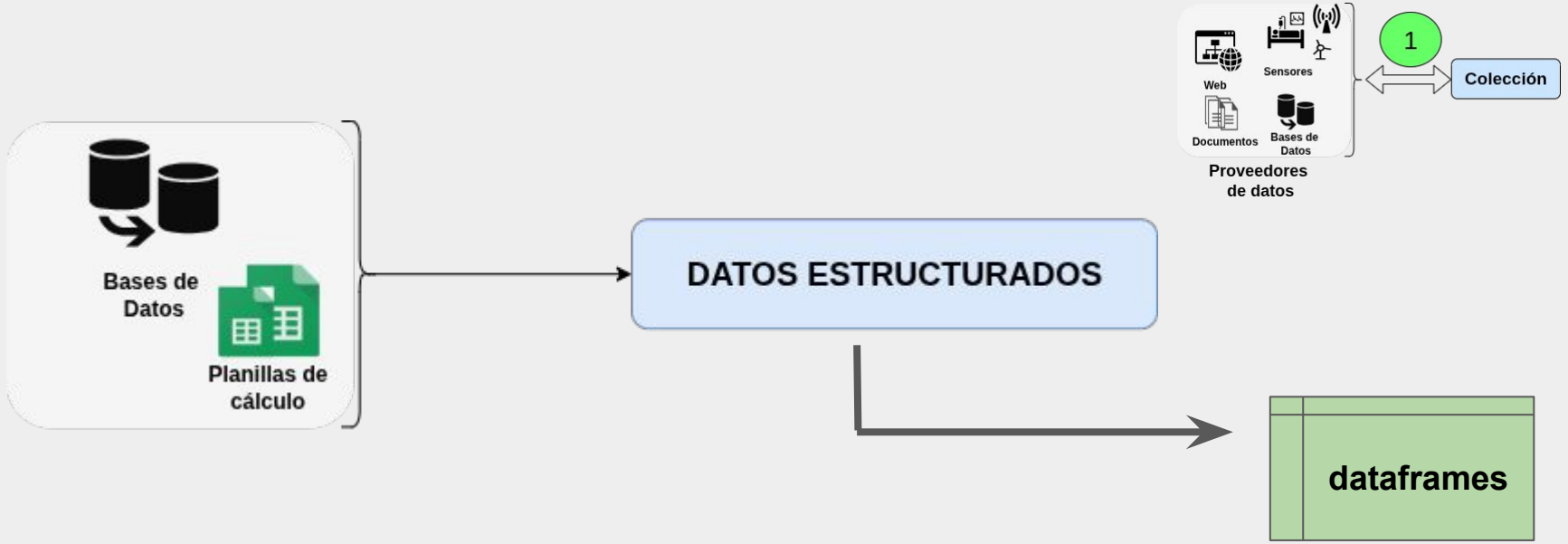
DATOS SEMI-ESTRUCTURADOS — mails, páginas web, logss



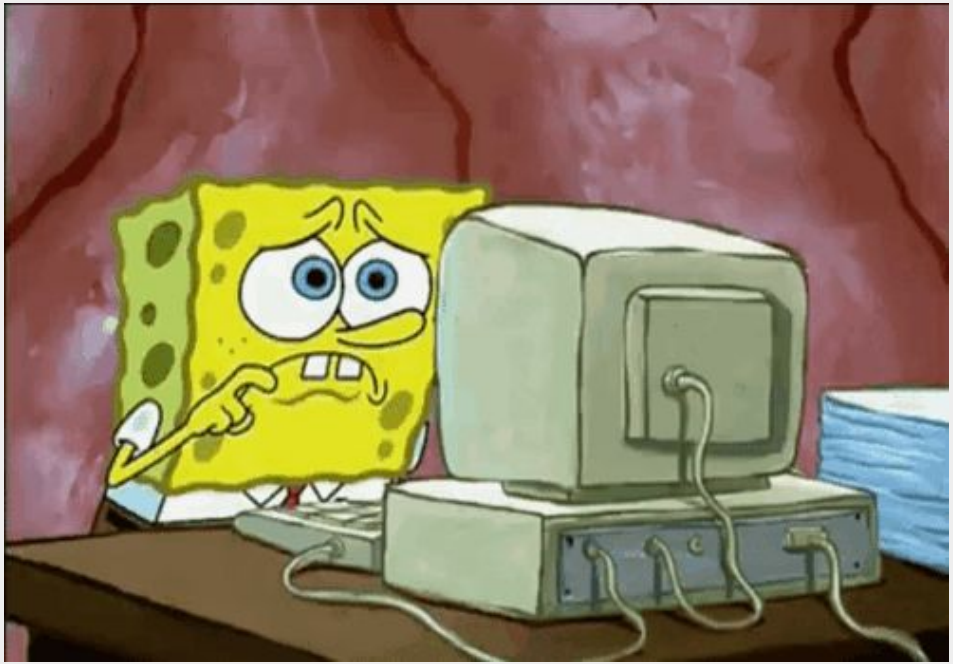
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos - Estructurados



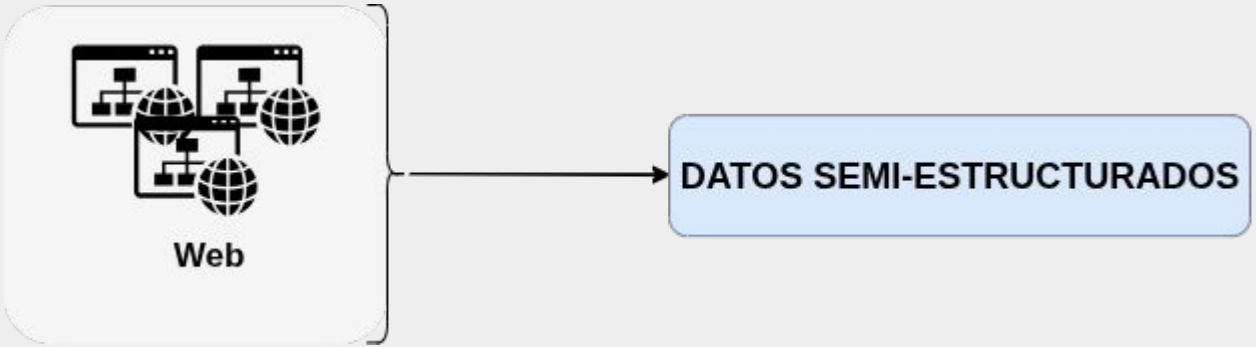
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos - Estructurados

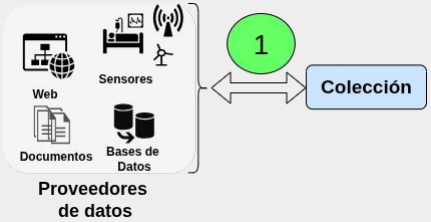


PROCESO DE ANÁLISIS DE DATOS
A TRABAJAR!!



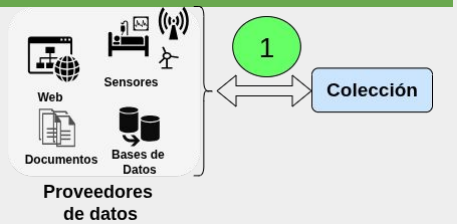
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos -SemiEstructurados





PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos - Búsqueda Web

Colección de datos desde la Web

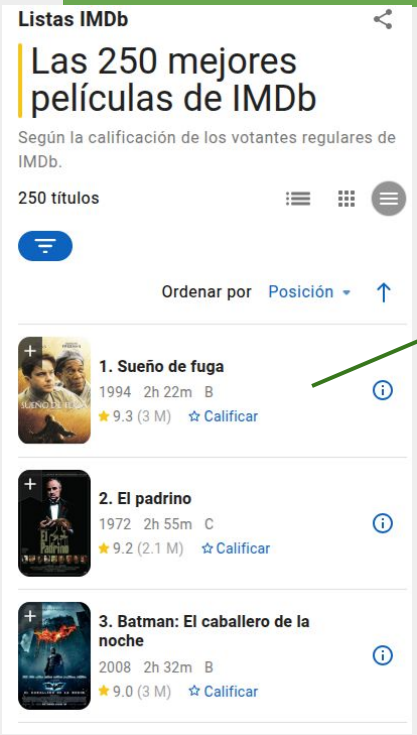


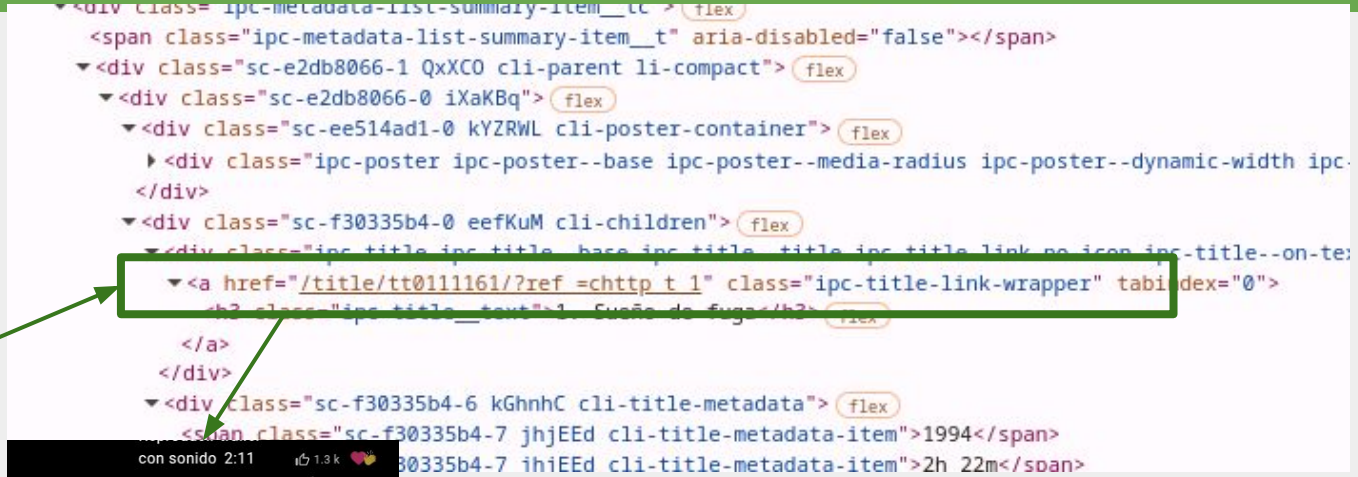
- basados en tags de HTML, XML y contenido
- Recursos:
  - **web crawling**
  - **web scraping**

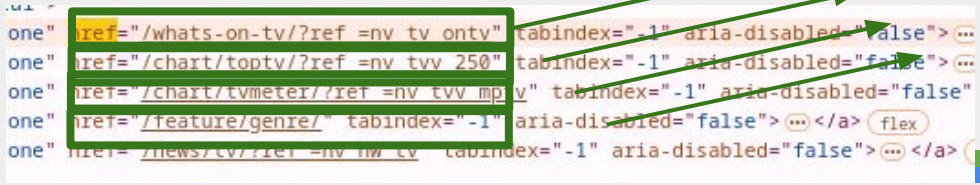
PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos -Búsqueda Web

- **Web crawling (o rastreadores Web)** es el proceso de explorar la Web (mediante programas automáticos) para recuperar datos.
  - Se navega las páginas usando los enlaces (hyperlinks) y recuperando las páginas siguientes.
  - Los datos coleccionados se utilizan para indexación (de los motores de búsqueda) o procesamiento posterior.
- **Web scraping (o raspado Web)**

PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos -Búsqueda Web

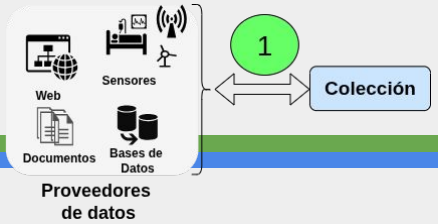




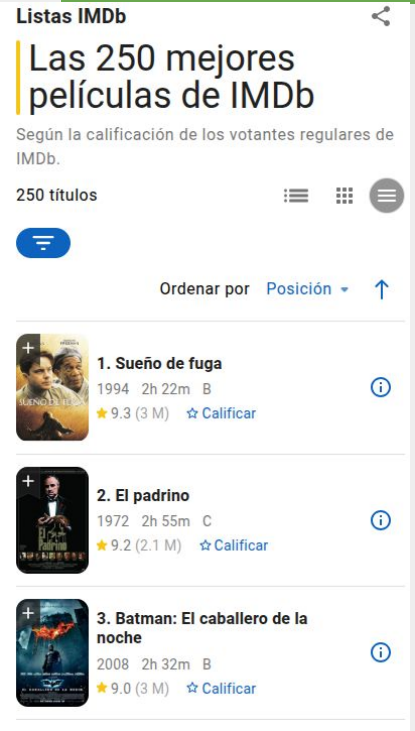


PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos -Búsqueda Web

- Web crawling (o rastreadores Web)
- **Web scraping (o raspado Web)** se centra en extraer datos específicos de los sitios y exportarlos para un almacenamiento en formato XML, Excel, JSON o SQL. Su automatización permite recopilar datos para la minería.



PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos -Búsqueda Web



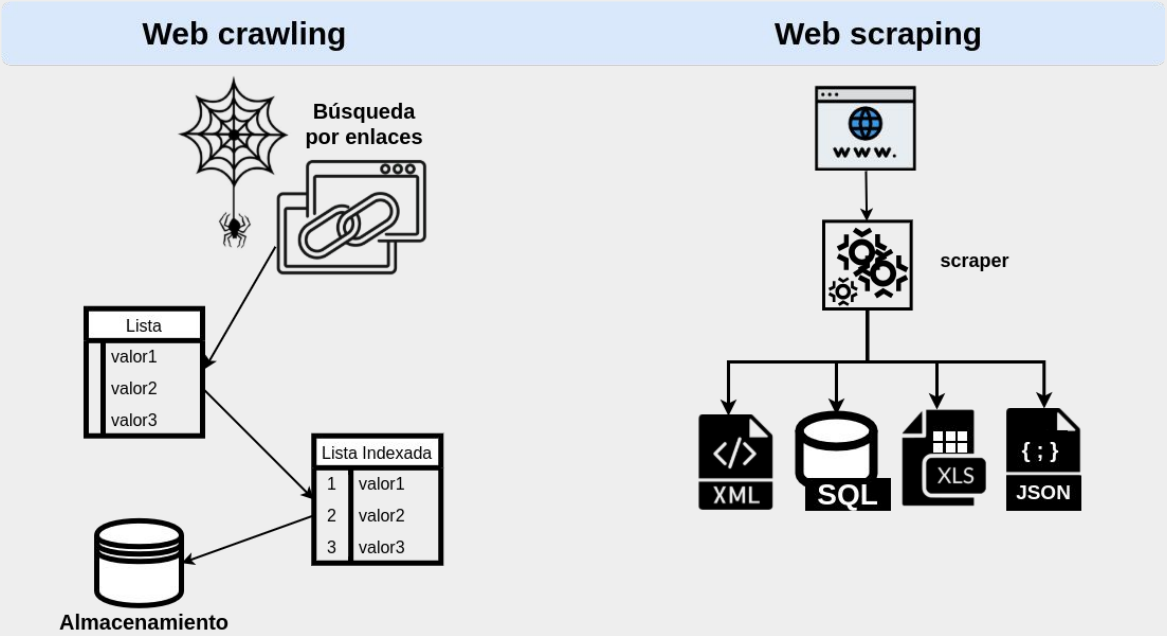
scraping

titulo de pelicula	año	duración	categoría	calificación
Sueño de Fuga	1994	2hs	B	9.3
El padrino	1972	3hs	C	9.2
Batman	2008	2.30hs	B	9

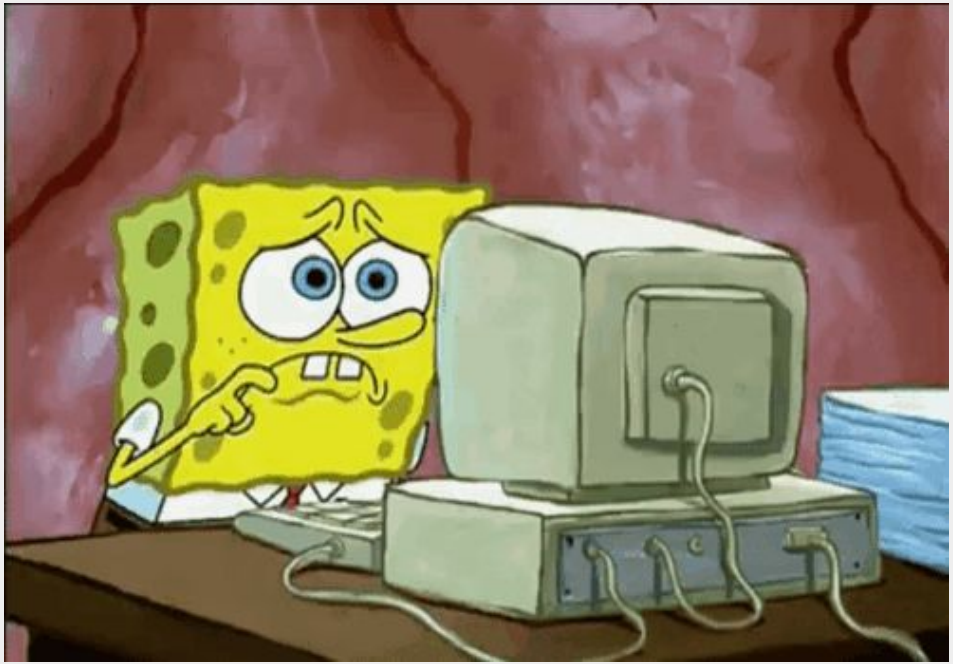
.CSV

```
titulo de pelicula,año,duración,categoría,calificación
Sueño de Fuga,1994,2hs,B,9.3
El padrino,1972,3hs,C,9.2
Batman,2008,2.30hs,B,9
```

PROCESO DE ANÁLISIS DE DATOS
Ingestión de Datos -Búsqueda Web



PROCESO DE ANÁLISIS DE DATOS
A TRABAJAR!!



PROCESO DE ANÁLISIS DE DATOS
con qué seguimos?