

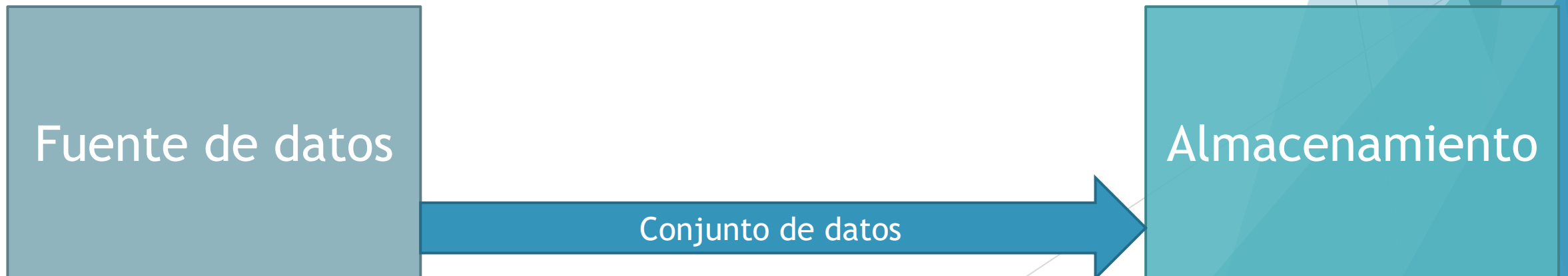
# Las fuentes de datos

Big Data Aplicado

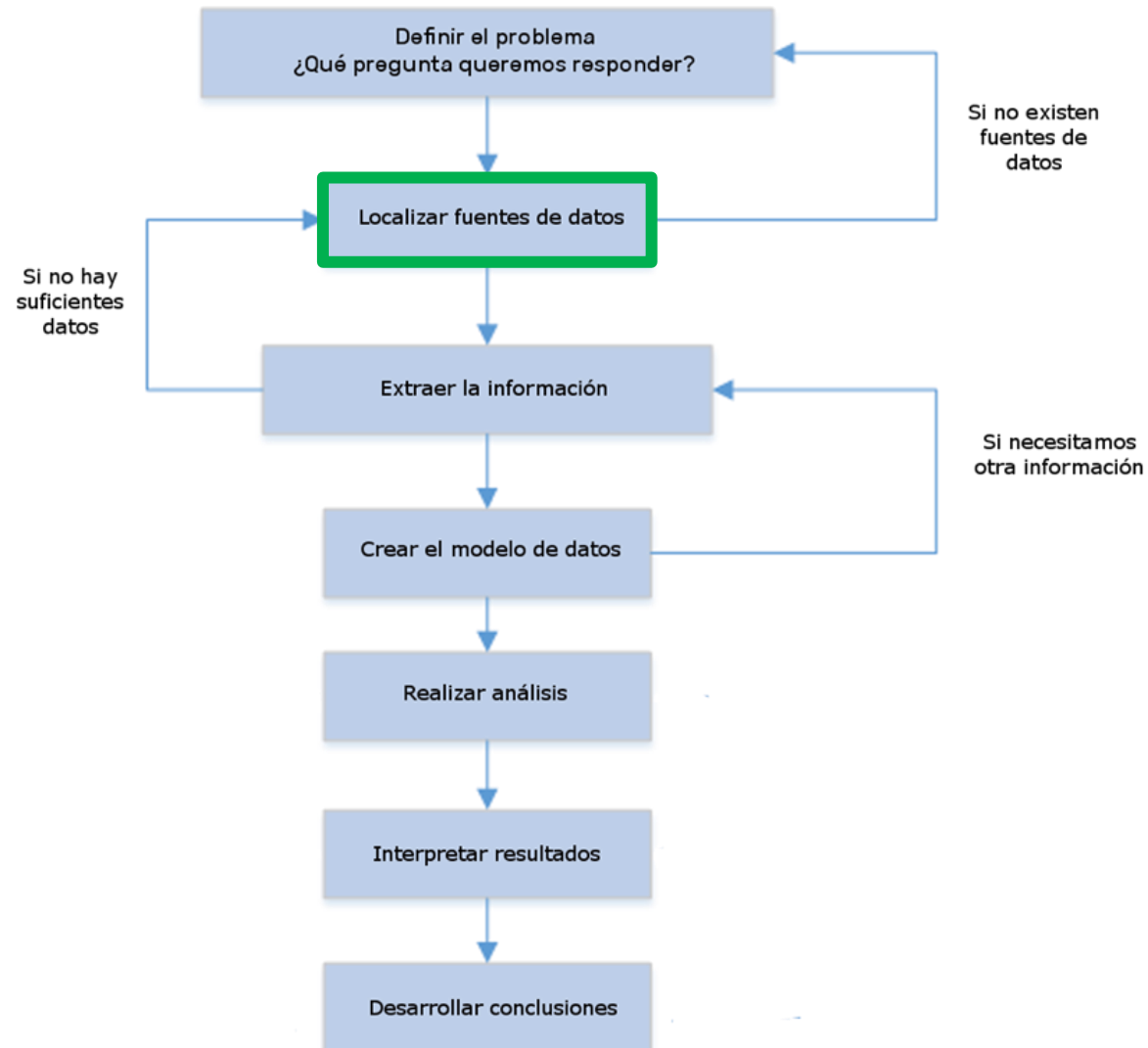
Dr. Francisco E. Cabrera

# Fuentes de datos

- ▶ Las fuentes de datos son aquellos lugares de los cuales obtenemos información **potencialmente relevante** para nuestros objetivos de análisis.
  - ▶ Los datos pueden provenir de orígenes muy variados según los análisis que pretendamos realizar.
  - ▶ Dependiendo de las fuentes escogidas, los datos pueden venir en distintos formatos.



# Localizar fuentes de datos



# Identificación y extracción de los datos

Aspectos a tener en cuenta

- ▶ Tipo de fuente.
- ▶ Tipo de contenido.
- ▶ Origen.
- ▶ Tiempo.
- ▶ Estructura.
- ▶ Derechos sobre los datos.



# Localizar fuentes de datos

**Problema:** Detección de fraude en transacciones financieras.

- ▶ Identificar transacciones sospechosas en tiempo real para prevenir fraudes en una plataforma de pago.
- ▶ Fuentes de datos a contemplar:



# Localizar fuentes de datos

**Problema:** Detección de fraude en transacciones financieras.

- ▶ Identificar transacciones sospechosas en tiempo real para prevenir fraudes en una plataforma de pago.
- ▶ Fuentes de datos a contemplar:
  - ▶ Historial de transacciones de los clientes.
  - ▶ Datos de geolocalización (ubicación de la transacción).
  - ▶ Dispositivos y direcciones IP utilizadas.
  - ▶ Cantidad, frecuencia y patrón de gasto de cada usuario.
  - ▶ Registro de intentos de inicio de sesión fallidos.
  - ▶ Datos externos sobre actividades fraudulentas previas.



# Localizar fuentes de datos

**Problema:** Mantenimiento predictivo en la industria.

- ▶ Predecir fallos en máquinas y equipos para reducir tiempos de inactividad en una fábrica.
- ▶ Fuentes de datos a contemplar:



# Localizar fuentes de datos

**Problema:** Mantenimiento predictivo en la industria.

- ▶ Predecir fallos en máquinas y equipos para reducir tiempos de inactividad en una fábrica.
- ▶ Fuentes de datos a contemplar:
  - ▶ Datos de sensores IoT en las máquinas (temperatura, vibración, presión, etc.).
  - ▶ Historial de mantenimiento y reparaciones previas.
  - ▶ Condiciones ambientales (humedad, temperatura externa).
  - ▶ Uso y carga de trabajo de la máquina.
  - ▶ Registro de errores o fallos pasados.
  - ▶ Información técnica del fabricante.





# Localizar fuentes de datos

**Problema:** Análisis de sentimiento en redes sociales para empresas.

- ▶ Conocer la percepción pública sobre una marca o producto en redes sociales.
- ▶ Fuentes de datos a contemplar:



# Localizar fuentes de datos

**Problema:** Análisis de sentimiento en redes sociales para empresas.

- ▶ Conocer la percepción pública sobre una marca o producto en redes sociales.
- ▶ Fuentes de datos a contemplar:
  - ▶ Publicaciones en redes sociales (Reddit, X, Facebook, Instagram, etc.).
  - ▶ Comentarios y reseñas de clientes en plataformas como Amazon, Google Reviews o Trustpilot.
  - ▶ Sentimiento expresado en el lenguaje (positivo, negativo, neutro).
  - ▶ Volumen y frecuencia de menciones de la marca.
  - ▶ Datos de la competencia para comparar tendencias.



# Localizar fuentes de datos

**Problema:** Análisis de percepción política en elecciones.

- ▶ Evaluar el sentimiento y las opiniones de la población sobre candidatos, partidos políticos o propuestas antes y durante una campaña electoral.
- ▶ Fuentes de datos a contemplar:



# Localizar fuentes de datos

**Problema:** Análisis de percepción política en elecciones.

- ▶ Evaluar el sentimiento y las opiniones de la población sobre candidatos, partidos políticos o propuestas antes y durante una campaña electoral.
- ▶ Fuentes de datos a contemplar:
  - ▶ **Redes sociales:** Tweets, publicaciones en Facebook, Instagram y TikTok sobre los candidatos o temas políticos.
  - ▶ **Análisis de sentimiento:** Clasificación de comentarios en positivos, negativos o neutros.
  - ▶ **Tendencias y hashtags:** Qué temas políticos son más mencionados y en qué contexto.
  - ▶ **Foros y blogs:** Opiniones en Reddit, Quora y otras plataformas de discusión política
  - ▶ **Encuestas y datos históricos:** Comparación con elecciones anteriores.
  - ▶ **Cobertura mediática:** Noticias y artículos de prensa sobre los candidatos y su impacto en la opinión pública.
  - ▶ **Geolocalización:** Identificar zonas donde un candidato tiene más apoyo o rechazo.



# La importancia de cada fuente

La importancia de las fuentes de datos puede depender del **enfoque concreto** en el que nos queramos centrar.

Medir el Impacto de un debate en tiempo real



Identificar los temas que más preocupan a la población



# La importancia de cada fuente

La importancia de las fuentes de datos puede depender del enfoque concreto en el que nos queramos centrar.

Fuentes a contemplar:

**Medir el impacto de un debate en tiempo real**

- ▶ Redes sociales.
- ▶ Análisis de sentimiento en comentarios y publicaciones.
- ▶ Cobertura mediática (noticias, sitios de opinión, etc.).
- ▶ Foros y blogs políticos.

**Identificar los temas que más preocupan a la población**

- ▶ Encuestas y datos históricos.
- ▶ Redes sociales.
- ▶ Análisis de noticias.
- ▶ Foros y otros sitios especializados.



# Morfología de los datos

# Morfología de los datos

No todos los datos son iguales

- ▶ Según su origen.
- ▶ Según su periodicidad.
- ▶ Datos según su estructura.
- ▶ Según su nivel de agregación





# Según su origen

## Información Interna

- ▶ Es información generada por la propia organización.
  - ▶ Transacciones comerciales, inventarios, datos de consumo, etc.
- ▶ La información suele estar estructurada.

## Información Externa

- ▶ Información proporcionada por otras organizaciones a través de internet.
  - ▶ Información acerca de la apreciación del público, la competencia, los proveedores, etc.
- ▶ La información no suele estar estructurada

# Según su periodicidad

Datos en Tiempo Real

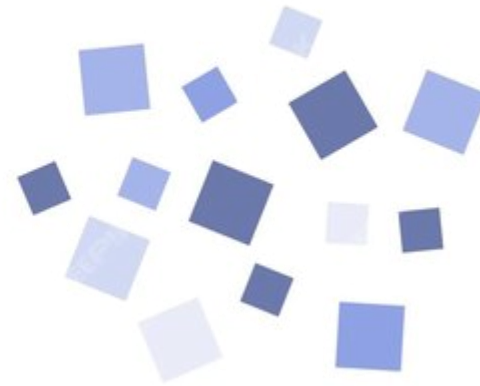
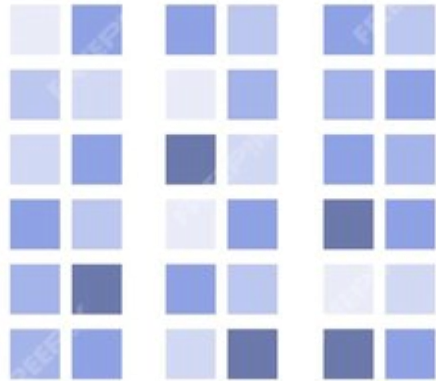
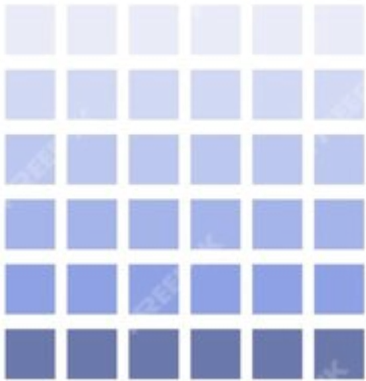


Datos en Lotes



# Según su estructura

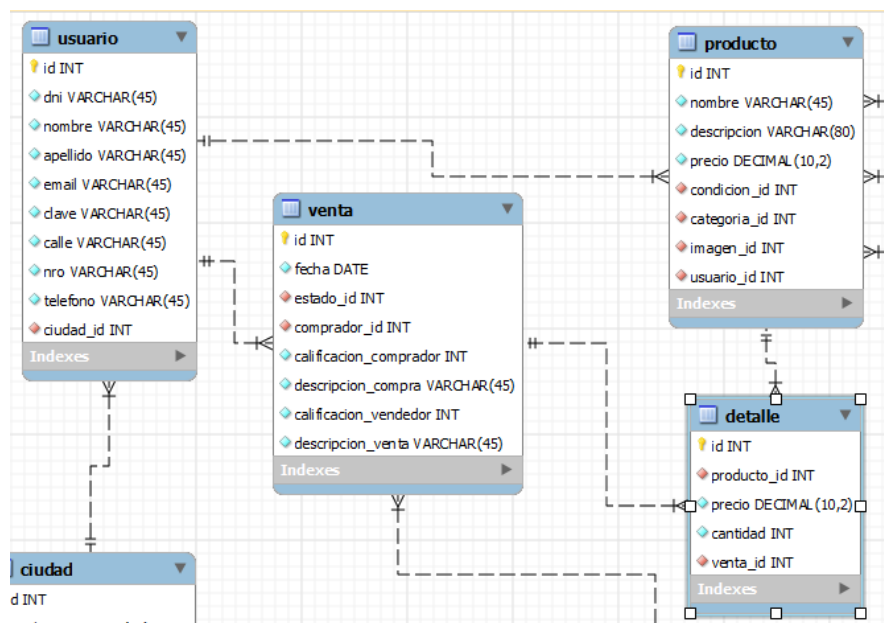
- ▶ Datos estructurados.
- ▶ Datos semiestructurados.
- ▶ Datos no estructurados.



# Datos estructurados

- ▶ Son datos organizados en un **formato fijo y predefinido**
- ▶ Se pueden buscar y analizar fácilmente con consultas estructuradas.
- ▶ **Características:**
  - ▶ Organizados en filas y columnas.
  - ▶ Fáciles de almacenar y procesar.
  - ▶ Uso de esquemas predefinidos (tablas con tipos de datos específicos).
- ▶ **Ejemplos:**
  - ▶ Bases de datos de clientes, registros de transacciones bancarias inventario de productos, etc.

# Datos estructurados



	A	B	C	D	E
1	Last Name	Sales	Country	Quarter	
2	Smith	\$16,753.00	UK	Qtr 3	
3	Johnson	\$14,808.00	USA	Qtr 4	
4	Williams	\$10,644.00	UK	Qtr 2	
5	Jones	\$1,390.00	USA	Qtr 3	
6	Brown	\$4,865.00	USA	Qtr 4	
7	Williams	\$12,438.00	UK	Qtr 1	
8	Johnson	\$9,339.00	UK	Qtr 2	
9	Smith	\$18,919.00	USA	Qtr 3	
10	Jones	\$9,213.00	USA	Qtr 4	
11	Jones	\$7,433.00	UK	Qtr 1	
12	Brown	\$3,255.00	USA	Qtr 2	
13	Williams	\$14,867.00	USA	Qtr 3	
14	Williams	\$19,302.00	UK	Qtr 4	
15	Smith	\$9,698.00	USA	Qtr 1	
16					



# Datos semiestructurados

- ▶ **No tienen una estructura rígida**, pero contienen etiquetas o marcadores que los organizan parcialmente.
- ▶ No se almacenan fácilmente en bases de datos relacionales tradicionales.
- ▶ **Características:**
  - ▶ No siguen un esquema fijo, pero tienen cierta organización.
  - ▶ Pueden contener metadatos o etiquetas para estructurar la información.
  - ▶ Más flexibles que los datos estructurados, pero requieren procesamiento adicional.
- ▶ **Ejemplos:**
  - ▶ JSON, XML, YAML, Logs de servidores.
  - ▶ Correos electrónicos (Asunto, fecha, remitente y cuerpo del mensaje).

# Datos semiestructurados

```
1 {  
2     "count": 7,  
3     "items": ["socks", "pants", "shirts", "hats"],  
4     "manufacturer": {  
5         "name": "Molly's Seamstress Shop",  
6         "id": 39233,  
7         "location": {  
8             "address": "123 Pickleton Dr.",  
9             "city": "Tucson",  
10            "state": "AZ",  
11            "zip": 85705  
12        }  
13    },  
14    "total_price": "$393.23",  
15    "purchase_date": "2022-05-30",  
16    "country": "USA"  
17 }
```



```
error.log x  
error.log  
1002  
1003 Traceback (most recent call last):  
1004   File "C:\Python312\Lib\site-packages\django\template\base.py", line 906, in  
    _resolve_lookup  
1005     raise VariableDoesNotExist(  
1006 django.template.base.VariableDoesNotExist: Failed lookup for key [name] in  
    <URLResolver <URLPattern list> (admin:admin) 'admin/'>  
1007 Not Found: /  
1008 "GET / HTTP/1.1" 404 3458  
1009 File C:\Python312\Lib\site-packages\django\contrib\messages\storage\cookie.py  
    first seen with mtime 1725448709.5825415  
1010 File C:\Python312\Lib\site-packages\django\contrib\messages\storage\session.py  
    first seen with mtime 1725448709.5825415  
1011 File C:\Python312\Lib\site-packages\django\contrib\messages\storage\fallback.py  
    first seen with mtime 1725448709.5825415  
1012 File C:\Python312\Lib\site-packages\django\contrib\sessions\serializers.py first  
    seen with mtime 1725448709.8349962  
1013 "GET /debug/ HTTP/1.1" 200 27  
1014 "GET /warning/ HTTP/1.1" 200 29  
1015 "GET /debug/ HTTP/1.1" 200 27  
1016 "GET /critical/ HTTP/1.1" 200 30  
1017 "GET /info/ HTTP/1.1" 200 26  
1018 |
```



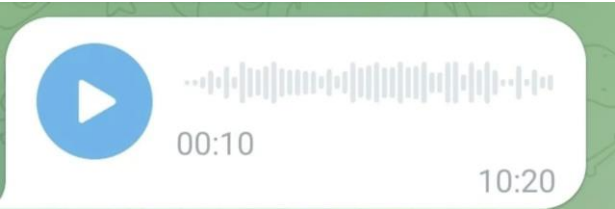


# Datos no estructurados


- ▶ **No tienen una estructura predefinida.**
- ▶ Necesitan procesamiento previo para almacenarse en bases de datos relacionales.
- ▶ Representan la mayor parte de los datos en el mundo.
- ▶ **Características:**
  - ▶ Sin estructura definida ni formato estandarizado.
  - ▶ Difíciles de analizar directamente sin herramientas especializadas.
  - ▶ Generalmente requieren técnicas de procesamiento de datos computacionalmente costosas como NLP (procesamiento de lenguaje natural) o visión por computadora.
- ▶ **Ejemplos:**
  - ▶ Imágenes, vídeos, audios, textos largos, publicaciones en redes sociales, documentos destinados a su lectura...
  - ▶ Formatos como MP4, MP3, JPG, PNG, TXT, PDF.



# Datos no estructurados



541 people found this review helpful  
2,870 people found this review funny 44

 Recommended  
16,241.0 hrs on record

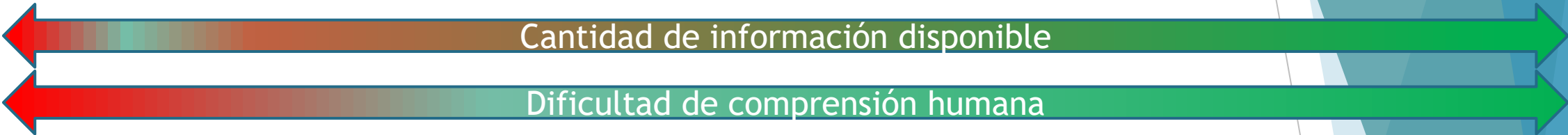
Posted: December 13, 2018

EARLY ACCESS REVIEW

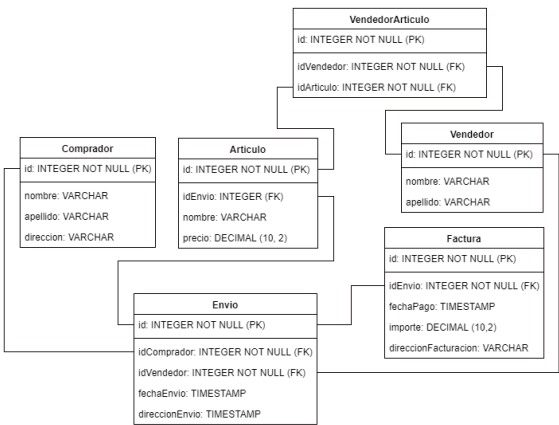
The game is ok, but only good for a couple of hours.

EDIT: I played a couple more hours and I think since the game's full release it's still pretty good.

# Estructura de los datos



## Fuentes estructuradas



## Fuentes semiestructuradas

Object	(1) ObjectId("5d1b521b58949341...")	{ 10 fields }	Object
Object	ObjectId("5d1b521b5894934157b49a37")	ObjectId	ObjectId
String	name	Tourism Criteria: Basic	String
Object	locale	{ 10 fields }	Object
Date	date_created	2019-07-02 13:46:01.754Z	Date
String	description	Basic criteria that valuate the tourism exp...	String
Array	criteria	[ 3 elements ]	Array
ObjectId	verifier	ObjectId("5d1b52185894934157b49a35")	ObjectId
ObjectId	expert	ObjectId("5d1b52175894934157b49a34")	ObjectId
Array	tags	[ 2 elements ]	Array
String	[0]	tourism	String
String	[1]	spain	String
Boolean	from_criterion	false	Boolean
Object	(2) ObjectId("5d1b534ddc244133...")	{ 9 fields }	Object
Object	(3) ObjectId("5d1b538edc244133...")	{ 9 fields }	Object
Object	(4) ObjectId("5d1b5506dc244133...")	{ 9 fields }	Object
Object	(5) ObjectId("5d1b5533dc244133...")	{ 9 fields }	Object

## Fuentes no estructuradas

El Mundo Ciencia retweeted

**EL MUNDO** @elmundoes · 20 jul.

Este sábado Aldrin y Collins han sido recibidos por Donald Trump en la Casa Blanca

**Buzz Aldrin bromea con ayudar a la tripulación de un avión a desp...**

Pese a su avanzada edad, Buzz Aldrin no parece dispuesto a dejar de volar, aunque ahora tenga que conformarse con viajes dentro de la Tierr...

elmundo.es



# Nivel de agregación

## Granularidad de los datos

- ▶ Datos agregados.
  - ▶ Promedio de ventas del día.
- ▶ Datos detallados (grano fino).
  - ▶ Registro de temperaturas por minuto de un sensor.
- ▶ Datos resumidos.
  - ▶ Informe mensual en comparación con los registros en tiempo real.

# Consideraciones previas

- ▶ Antes de extraer los datos debería hacerme estas preguntas:
  - ▶ ¿Tengo derecho legal para acceder a estos datos?
  - ▶ ¿Cumpló con regulaciones como GDPR o CCPA?
  - ▶ ¿Los datos están bien estructurados y listos para procesar?
  - ▶ ¿Mi sistema puede manejar la carga de extracción y almacenamiento?
  - ▶ ¿Estoy protegiendo la privacidad y seguridad de los datos?
  - ▶ ¿El uso que daré a los datos es ético y responsable?

Si alguna de estas respuestas es “No” debería descartar la fuente o volver al primer paso y plantearme la definición del problema.

# ¿Dónde conseguir los datos?

- ▶ Datos internos.
- ▶ Fuentes de datos abiertos.
- ▶ APIs de diferentes servicios.
- ▶ Datos de investigaciones científicas.
- ▶ Web Scraping.
- ▶ Compra de datos.



# Algunas fuentes de datos abiertos

- ▶ Instituto Nacional de Estadística
  - ▶ <https://ine.es>
- ▶ CIS
  - ▶ <https://www.cis.es/catalogo-estudios/resultados-definidos/buscador-estudios>
- ▶ Portal de datos del gobierno
  - ▶ <https://datos.gob.es/es/>
- ▶ Portal de datos abiertos Junta de Andalucía
  - ▶ <https://www.juntadeandalucia.es/datosabiertos/>
- ▶ NASA
  - ▶ <https://data.nasa.gov>
- ▶ WorldBank
  - ▶ <https://data.worldbank.org>

# Algunas fuentes de datos abiertos

- ▶ Kaggle
  - ▶ <https://www.kaggle.com>
- ▶ DataHub
  - ▶ <https://datahub.io>
- ▶ Google Dataset Search
  - ▶ <https://datasetsearch.research.google.com>
- ▶ GitHub
  - ▶ <https://github.com/datasets>

# Ejercicio: Buscar fuentes de datos

Encontrar 4 fuentes de datos y para cada una:

- ▶ Identifique el **dominio** de la información provista por la fuente de datos.
- ▶ Identifique el **proveedor** de los datos.
- ▶ Identifique la **frecuencia** de actualización de los datos de la fuente.
- ▶ Identifique la **morfología** de la fuente de datos.
  - ▶ En que formato se obtienen los datos.
  - ▶ Que tipo de estructura tienen los datos.

¿En que casos podría ser útil la fuente de datos?



# El proceso ETL



# El proceso ETL

Extracción, transformación y carga



ETL



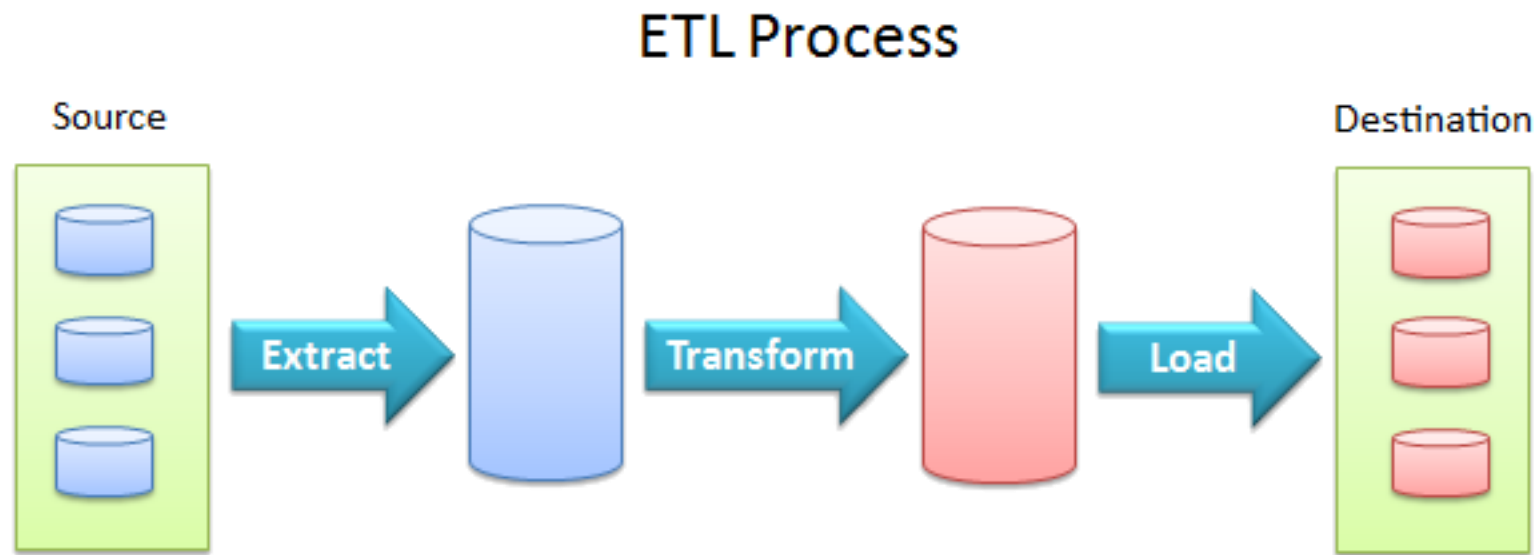
# ¿Qué hace el proceso ETL?

El proceso ETL (Extract, Transform, Load) es esencial en **Data Engineering** y **Data Science**.

Objetivo:

- ▶ Pasar de tener datos de diversas fuentes a tener datos cargados en un almacenamiento final, habiendo transformado esos datos según los requerimientos del negocio.

# El proceso ETL



# Extracción

Es la fase donde se recopilan los datos desde diversas fuentes.

- ▶ **Aspectos a tener en cuenta:**

- ▶ **Formato de los datos:** Cada fuente puede tener un formato distinto.
- ▶ **Velocidad y volumen:** Algunas fuentes generan datos en tiempo real (ej. sensores IoT).
- ▶ **Calidad de los datos:** Puede haber información incompleta, duplicada o incorrecta.
- ▶ **Relevancia con respecto al problema:** Hay que considerar qué datos pueden aportar las fuentes para resolver nuestro problema.

# Transformación

Es la fase donde los datos extraídos se limpian, estructuran y convierten en información útil.

## ► Tareas comunes en esta fase:

- **Limpieza de datos:** Eliminar registros duplicados, corregir valores nulos o inconsistentes.
- **Conversión de formatos:** Convertir fechas, cambiar unidades de medida, normalizar texto.
- **Integración de datos:** Unificar información de diferentes fuentes en un solo formato.
- **Cálculos y agregaciones:** Calcular promedios, sumar ventas, identificar tendencias.
- **Enriquecimiento:** Añadir datos externos (ej. agregar información meteorológica a ventas).

# Carga

En esta fase, los datos transformados se almacenan en un destino final para ser analizados y usados en reportes o modelos de machine learning.

- ▶ **Aspectos a tener en cuenta:**

- ▶ **Tipos de carga:** Completa, Incremental o en Tiempo Real (ETL Streaming)
- ▶ Rendimiento y escalabilidad.
- ▶ Integridad de los datos.
- ▶ Seguridad y cumplimiento normativo.
- ▶ Monitorización y manejo de errores.
- ▶ Formato y estructura de los datos en destino.



# ¿Cuándo debería aplicar el ETL?

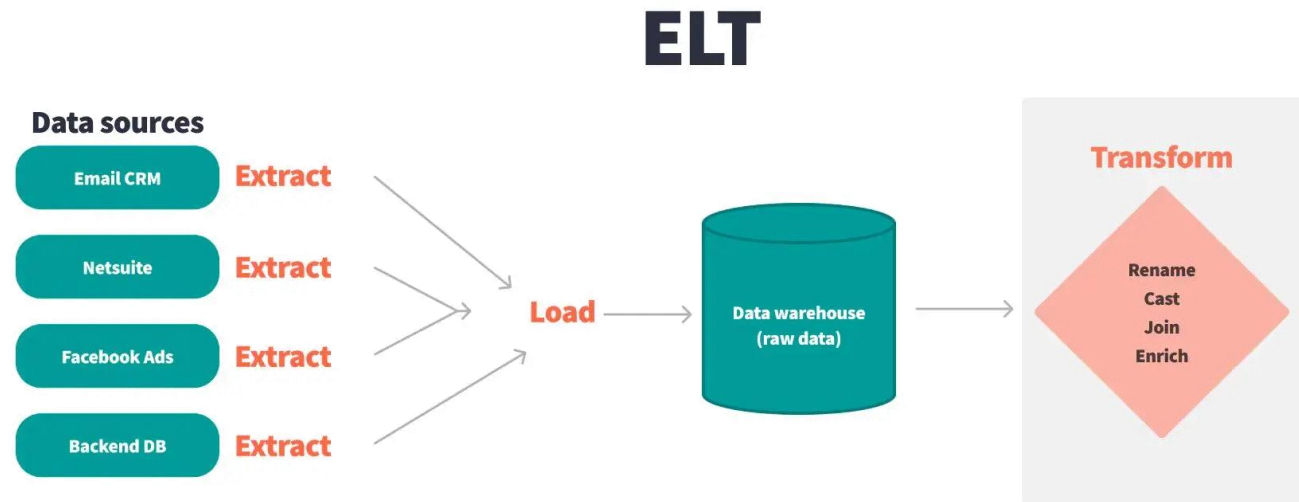
Plantear un ETL cuando hay que:

- ▶ Integrar datos desde múltiples fuentes
  - ▶ APIs, bases de datos, archivos, etc.
- ▶ Tengo datos desestructurados o en formatos diferentes
  - ▶ Los datos necesitan ser limpiados y normalizados.
- ▶ El análisis o modelado requiere estructuras limpias y consistentes.
- ▶ Hacer actualizaciones de datos en un Data Warehouse.
- ▶ Hace falta mejorar el rendimiento de consultas,
  - ▶ Se pueden optimizar bastante los datos antes de cargarlos en sistemas de BI.
- ▶ Necesito asegurar consistencia y calidad antes de almacenarlos.
- ▶ La cantidad de datos me obliga a mejorar la eficiencia en los pipelines de datos.



# Otra opción es el ELT

Extracción, Carga y Transformación.



- Cargar los datos sin procesar y transformarlos luego.

# Utilidades del ELT

¿Cuándo me puede convenir planter un ELT?

- ▶ Al trabajar con muchos datos no estructurados.
- ▶ Necesito flexibilidad según el caso de uso.
- ▶ Tengo un Data Lake con tecnología que me lo permite.
  - ▶ Google BigQuery, AmazonS3, Azure DataLake
- ▶ No dispongo de suficiente procesamiento durante la etapa de captura.
  - ▶ Pero luego voy a disponer de más potencia en el Data Warehouse.
- ▶ Datos en streaming.
  - ▶ La capa de transformación de ETL puede ser un cuello de botella.
- ▶ Costos en la nube.

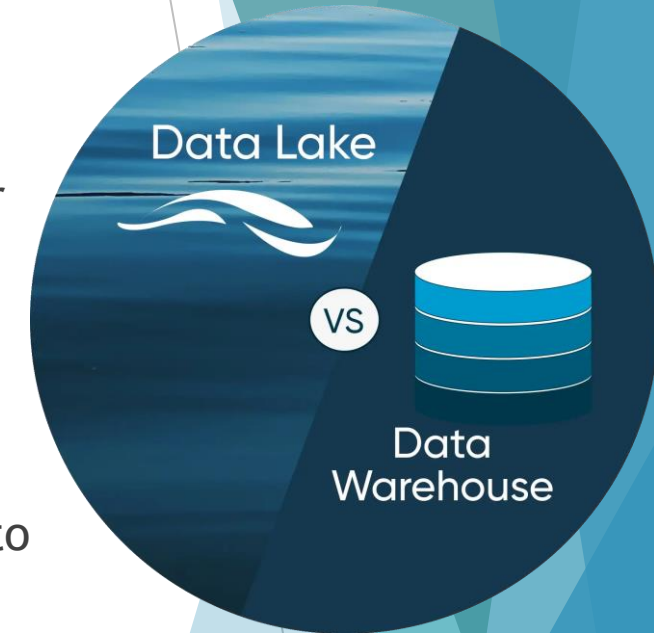
# Data Warehouse y Data Lake

## Data Warehouse

- ▶ Almacén de datos estructurados para análisis y Business Intelligence
- ▶ Datos estructurados o semiestructurados.
- ▶ Proceso ETL.
- ▶ Procesamiento SQL muy rápido en las consultas.
- ▶ Difícil escalabilidad horizontal.

## Data Lake

- ▶ Repositorio de datos sin procesar
- ▶ Todos los formatos.
- ▶ Proceso ELT.
- ▶ Batch y tiempo real.
- ▶ Suele requerir más procesamiento en las consultas.
- ▶ Fácil escalabilidad.



# ¿Y por qué no los dos?

El enfoque híbrido.

- ▶ Cargar rápidamente los datos en un Data Lake sin transformarlos (ELT).
- ▶ Transformar solo los datos estructurados y de alta prioridad para almacenarlos en un Data Warehouse.
- ▶ Mantener el Data Lake como un respaldo de datos crudos, permitiendo transformaciones adicionales en análisis futuros.