

Gestión de datos

Modelos de datos y sistemas de información

Lorena Etcheverry (lorenae@fing.edu.uy)
Instituto de Computación, FING, Udelar

¿de qué
hablamos
cuando
hablamos de
datos?

Datos, información y conocimiento

Datos

- Un parámetro o hecho, un número, una afirmación, una imagen
- Representan algo en el mundo real
- Son la materia prima para la producción de información

Información

- Datos con un significado en cierto contexto
- Datos relacionados
- Datos luego de su manipulación

Conocimiento

- Experiencia e información acumulada

**Pero,
¿de dónde
salen los
datos?**

*Se puede decir que los datos casi
siempre provienen de un Sistema de
Información*

Los sistemas de información (SI)

Conjunto de componentes que interactúan con el objetivo de **almacenar**, **recuperar** y **procesar datos** e información para crear nueva información.

Los componentes de un SI son software y hardware, pero es fundamental el rol de las personas.

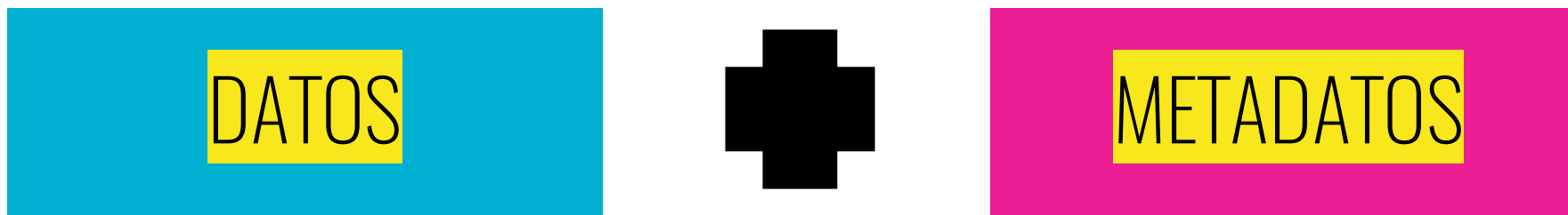
Un ejemplo de los primeros Sistemas de Información: Censos (de personas y/o bienes) babilonios año 3800 a.c. !!!!

Los Sistemas de Información Informáticos

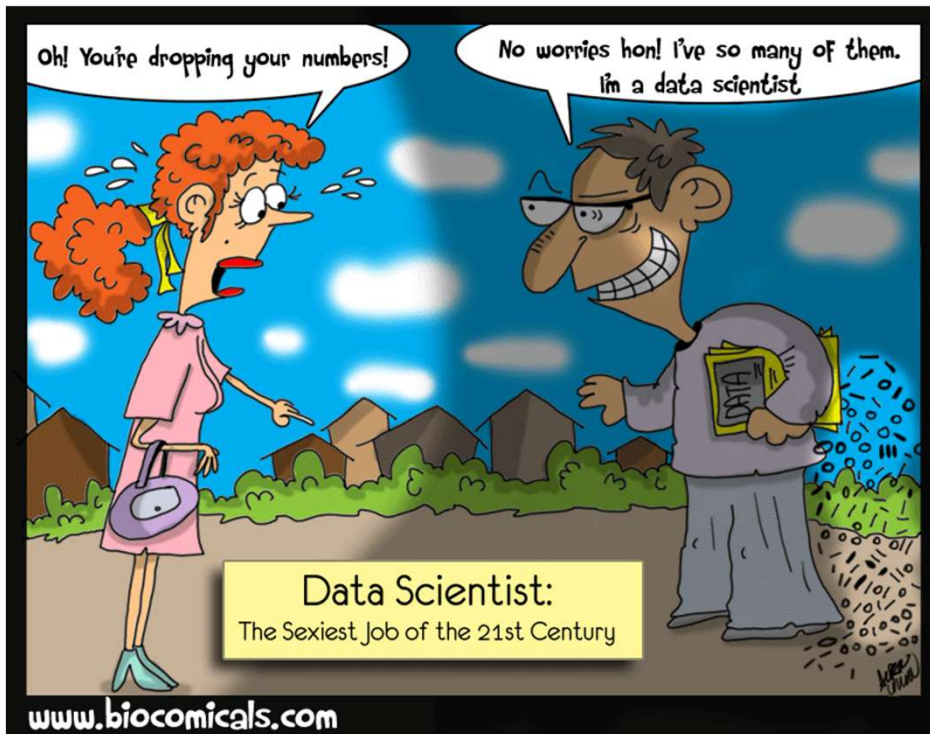
Utilizan tecnologías informáticas para realizar algunas de sus tareas.

Cumplen con 3 funciones principales:

- **Memoria:** mantienen una representación del estado de cierto **dominio**
- **Informativa:** proveen información acerca del estado de cierto **dominio**
- **Activa:** realizan acciones que cambian el estado de cierto **dominio**

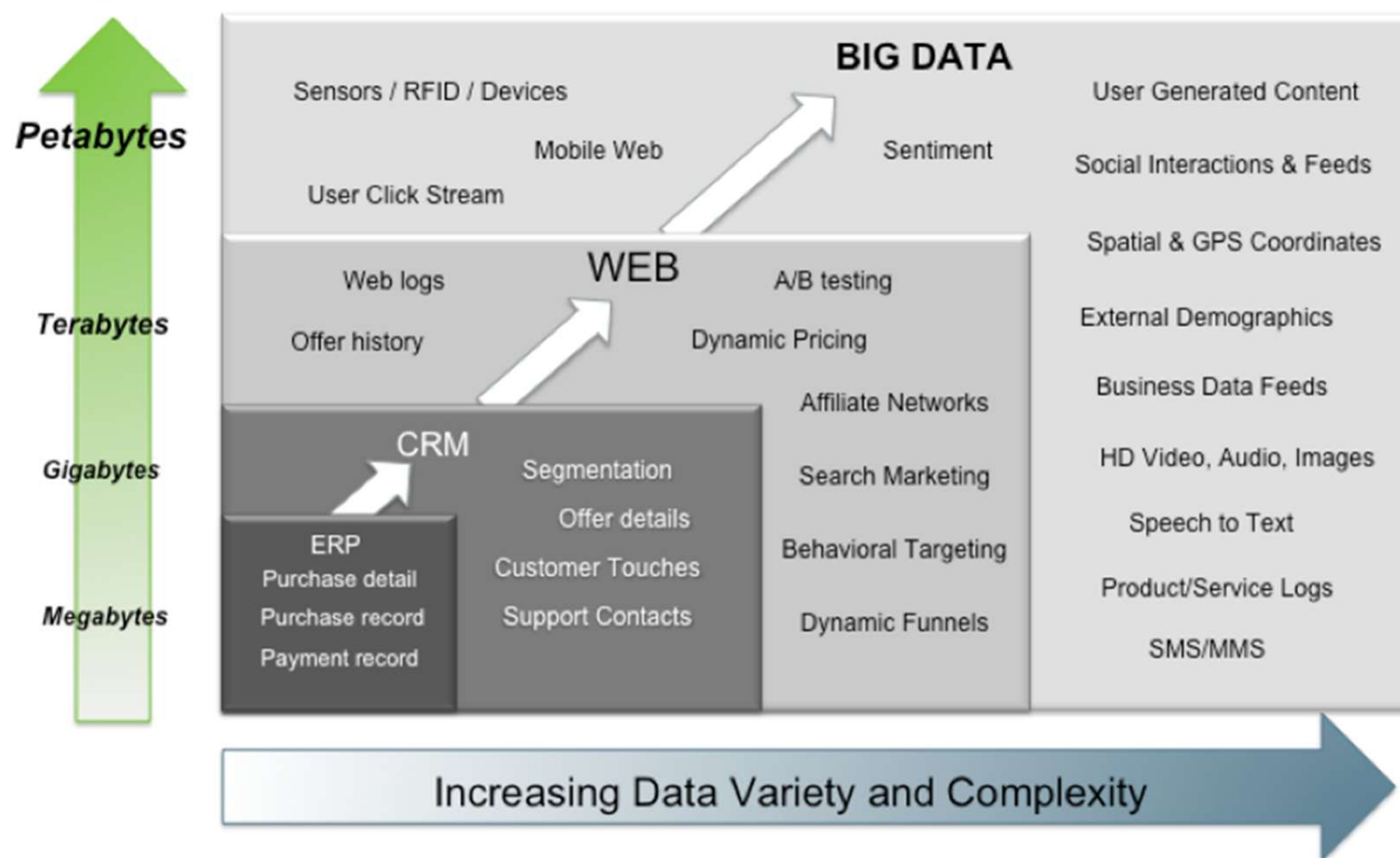


En los últimos 20 años el panorama es más complejo



La naturaleza de los datos y su volumen

Big Data = Transactions + Interactions + Observations



Source: Contents of above graphic created in partnership with Teradata, Inc.

¿qué son los
modelos de
datos?

Modelo de datos: definición

Los modelos de datos son **lenguajes** usados para especificar y manipular Bases de Datos

Un Modelo de Datos permite expresar:

- **Estructuras:** Elementos de los problemas.
- **Restricciones:** Reglas que deben cumplir los datos para que la base sea considerada válida.
- **Operaciones:** Insertar, borrar y **consultar** la BD.

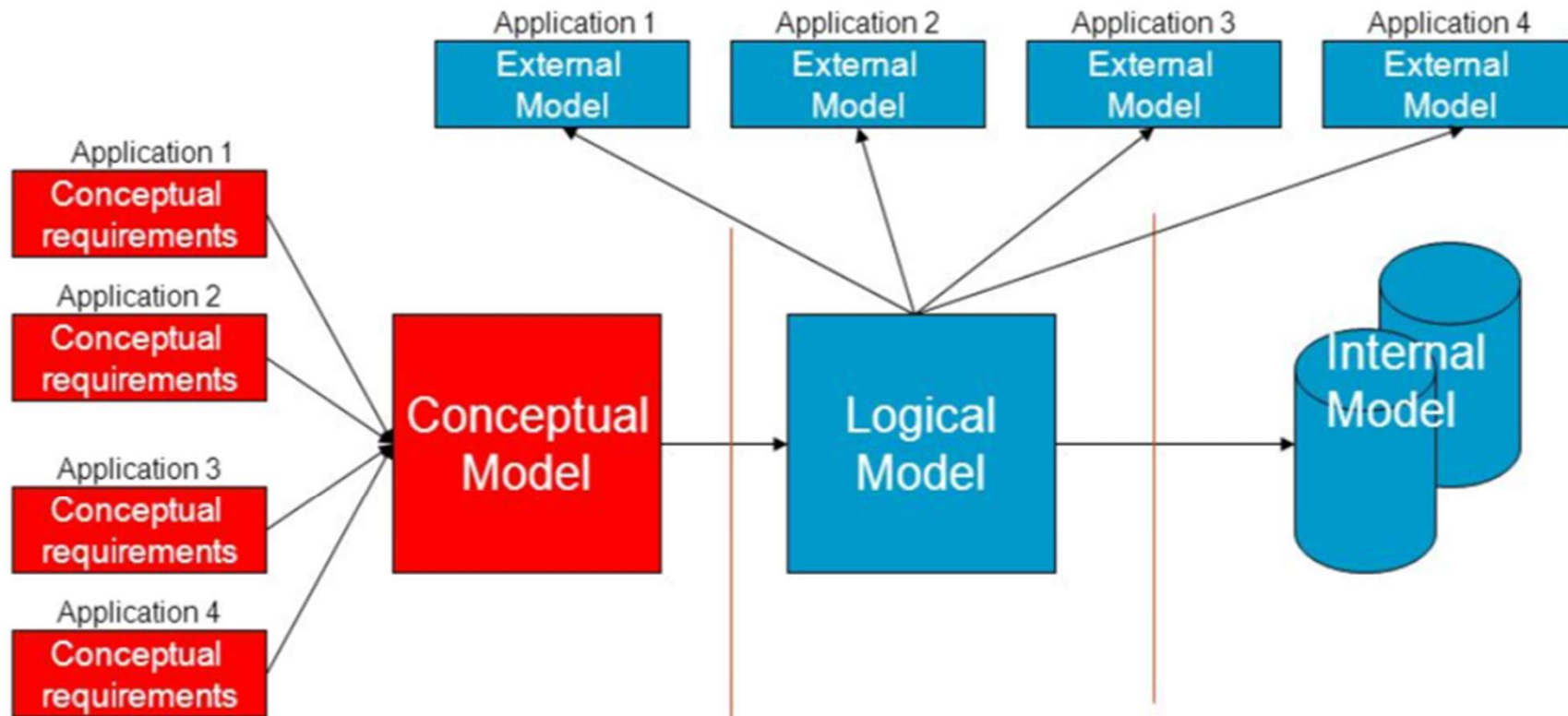
Clasificación de los modelos de datos: nivel de abstracción

Conceptuales: Representan la realidad independientemente de cualquier implementación de BD.

Lógicos: Son implementados en un manejador de bases de datos particular.

Físicos: Corresponden a cómo está implementado el manejador de bases de datos (estructuras de datos)

El proceso de diseño de una base de datos



Ejemplos

El **modelo Entidad-Relación** es un ejemplo de modelo Conceptual.

El **modelo Relacional** es un ejemplo famoso y exitoso de modelo lógico.

Otros modelos lógicos:

- el modelo de **property-graphs** para bases de datos de grafos
- el modelo **documental**

Modelos Conceptuales

Nos concentraremos en el modelo **Entidad-Relación**

Para profundizar/repasar referirse al material recomendado en el EVA:

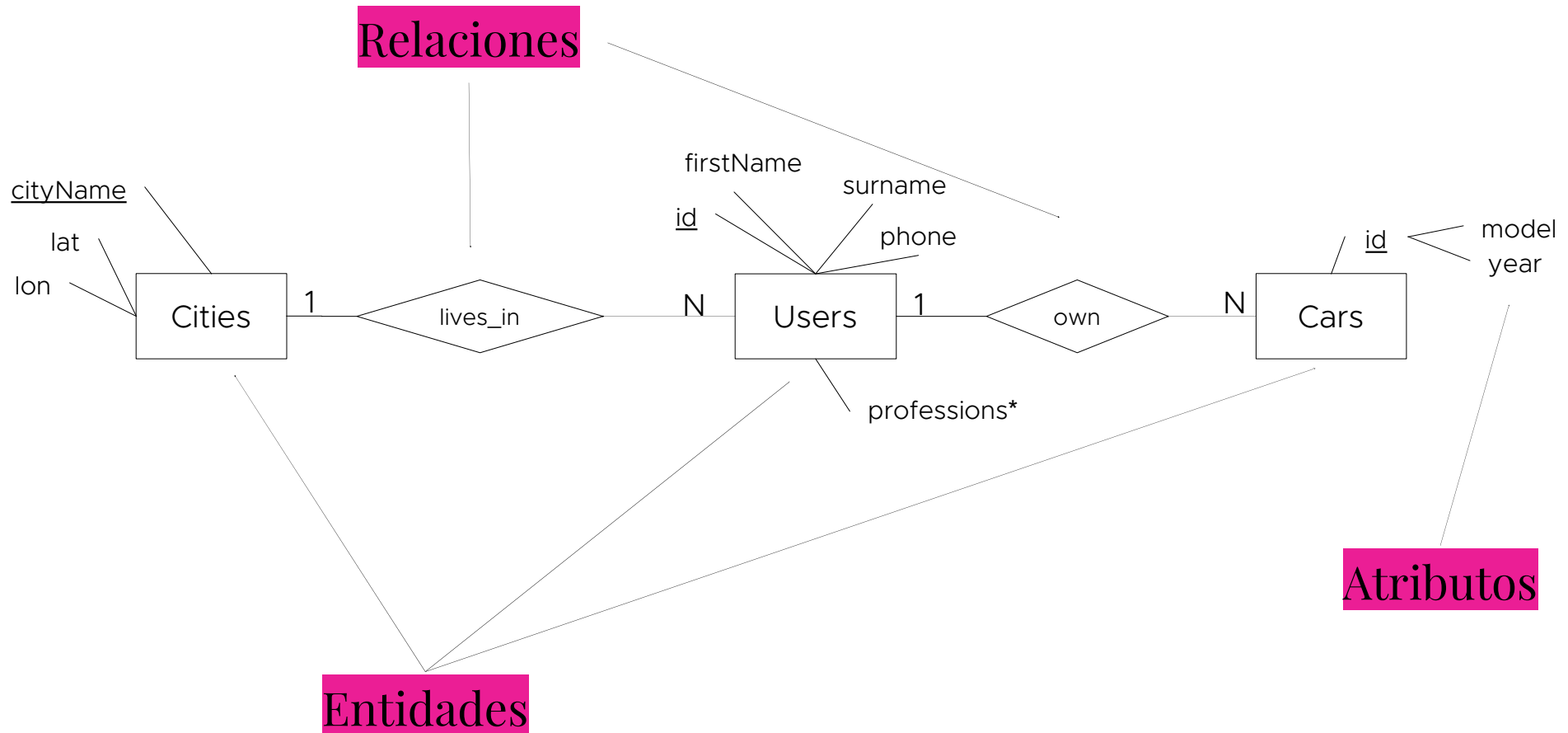
- los videos del curso Fundamentos de Bases de Datos (disponibles en OpenFING)
- Capítulo 7 ([Fundamentals of Database Systems, Elmasri and Navathe, 6th edition](#))

Proceso de elaboración de un modelo conceptual

Ejemplo: Queremos modelar una base de datos de propietarios de automóviles. De cada usuario se conoce su nombre y apellido, un número de teléfono y una o más profesiones. Cada usuario se identifica por un número de usuario, que es único. Se conoce además la ciudad en donde vive cada usuario, y para cada ciudad (identificada por su nombre) se conocen sus coordenadas (lat y long). Además interesa registrar los vehículos que posee cada usuario. Cada vehículo se identifica por el modelo y el año.

- 1- identificamos entidades (conceptos relevantes de la realidad)
- 2- identificamos atributos de las entidades
- 3- relaciones entre entidades y restricciones

Un posible diseño conceptual de la realidad anterior



Modelos lógicos

Son los modelos que implementan los motores de bases de datos.

Se construyen a partir de un diseño conceptual.

El **modelo Relacional** (Codd, 1970) es un ejemplo famoso y exitoso de modelo lógico.

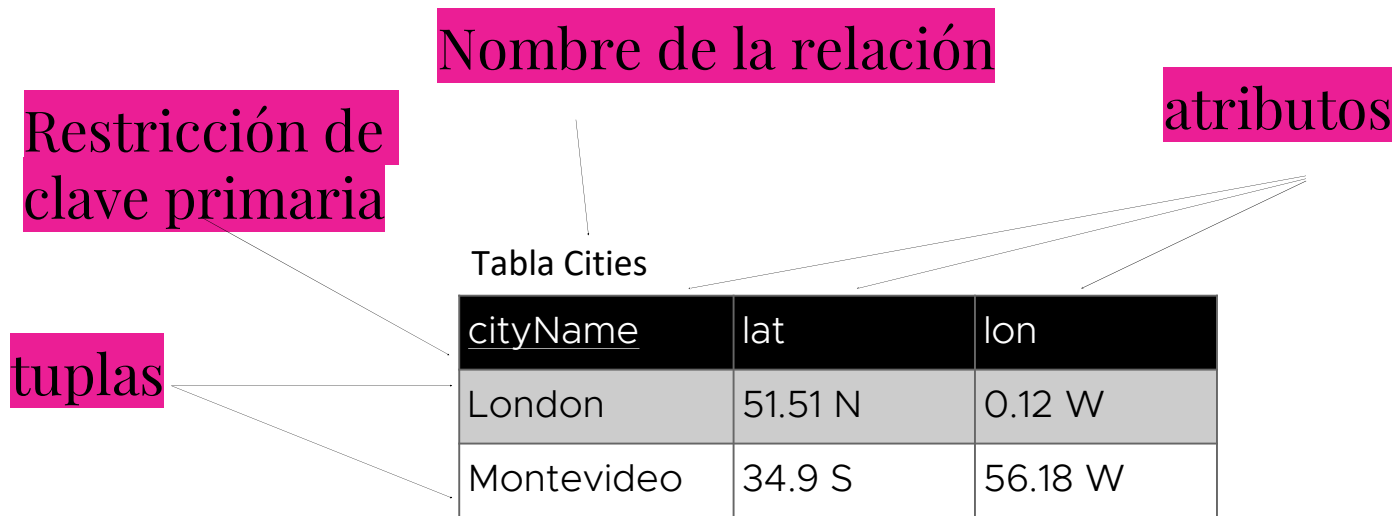
Modelos lógicos

Son los modelos que implementan los motores de bases de datos.

Se construyen a partir de un diseño conceptual.

El **modelo Relacional** (Codd, 1970) es un ejemplo famoso y exitoso de modelo lógico.

Modelo relacional



Los datos se representan como una colección de **relaciones** (tablas).

Cada relación tiene un nombre y un conjunto de atributos.

Las instancias de la relación (valores) son conjuntos de tuplas (sin repetidos, sin orden).

Además hay restricciones (ej: restricción de clave primaria, de clave foránea)

Un diseño relacional de nuestro ejemplo

Tabla Cities

<u>cityName</u>	lat	lon
London	51.51 N	0.12 W
Montevideo	34.9 S	56.18 W

Tabla Users

<u>id</u>	firstName	surname	phone	city
1	Paul	Miller	5111111	London
2	Laura	Rodriguez	3333333	Montevideo

Tabla User_professions

<u>userId</u>	<u>profession</u>
1	banking
1	finance
1	trader

Tabla Cars

<u>userId</u>	<u>model</u>	<u>year</u>
1	Bentley	1973
1	Rolls Royce	1965

Esquema vs Instancia

ESQUEMA: describe qué datos hay en la base, cómo se relacionan esos datos entre sí y qué restricciones de integridad deben cumplir

Estructuras + Restricciones

Bastante estático

INSTANCIA: Conjunto de datos almacenados en una base.

Es el valor de la base en un instante de tiempo.

Volátiles.

Si respetan todas las restricciones, se considera que la instancia es correcta.

Esquema de nuestro ejemplo

Cities (cityName, lat, lon)

Users (id, firstName, surname, phone, city)

User_professions (userId, profession)

Cars (userId, model, year)

Restricciones de clave primaria:

los atributos subrayados en cada tabla son determinantes

Restricciones de clave foránea:

- userId en User_professions tiene que ser un id en Users
- userId en Cars tiene que ser un id en Users

Diseño lógico de BDs relacionales

- Hay heurísticas que permiten obtener un diseño a partir del modelo conceptual, garantizando cierta calidad del diseño obtenido
 - Pasaje de MER a ER ([ver material](#))
- Esta calidad es independiente de las consultas a realizar.
- Las Formas Normales garantizan propiedades para todas las instancias

Lenguaje SQL

- Lenguaje declarativo estándar sobre bases de datos relacionales

Todos los usuarios que viven en Montevideo

```
SELECT firstName, lastname  
FROM users  
WHERE city = "Montevideo"
```

Dueños de autos de 1973

```
SELECT u.firstName, u.lastname  
FROM users u , cars c  
WHERE u.id= c.userId and c.year= "1973"
```

Tabla Users

id	firstName	surname	phone	city
1	Paul	Miller	5111111	London
2	Laura	Rodriguez	333333	Montevideo

Tabla Cars

userId	model	year
1	Bentley	1973
1	Rolls Royce	1965

Pero hay otros modelos lógicos

- Bases de datos **documentales** (colecciones de documentos tipo JSON)
- Bases de datos de **grafos**

JSON: Javascript Object Notation

objeto

valor

clave

arreglo (lista de valores)

```
{ "glossary": {  
  "title": "example glossary",  
  "GlossDiv": {  
    "title": "S",  
    "GlossList": {  
      "GlossEntry": {  
        "ID": "SGML",  
        "SortAs": "SGML",  
        "GlossTerm": "Standard Generalized Markup Language",  
        "Acronym": "SGML",  
        "Abbrev": "ISO 8879:1986",  
        "GlossDef": {  
          "para": "A meta-markup language, used to  
            create markup languages such as DocBook.",  
          "GlossSeeAlso": ["GML", "XML"]  
        },  
        "GlossSee": "markup"  
      }  
    }  
  }  
}
```

Las *claves* son
strings.

Los *valores*
pueden ser
cualquier
elemento de la
notación.

valor

string

number

object

array

true

false

null



BD relacional

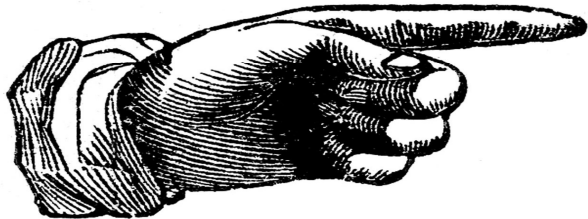


BD documental

Tabla	Colección
Fila (tupla)	Documento
Columna (atributo)	Campo (<i>field</i>)
SQL Join	Documentos embebidos y referencias
SQL Group-by (agregación)	<i>Aggregation pipeline</i>

- Los documentos dentro de una colección pueden tener campos diferentes (diferente esquema).
- Se *puede* exigir la validación de cierto esquema definido con [JSON Schema](#)
 - Ejemplos de [esquemas](#)

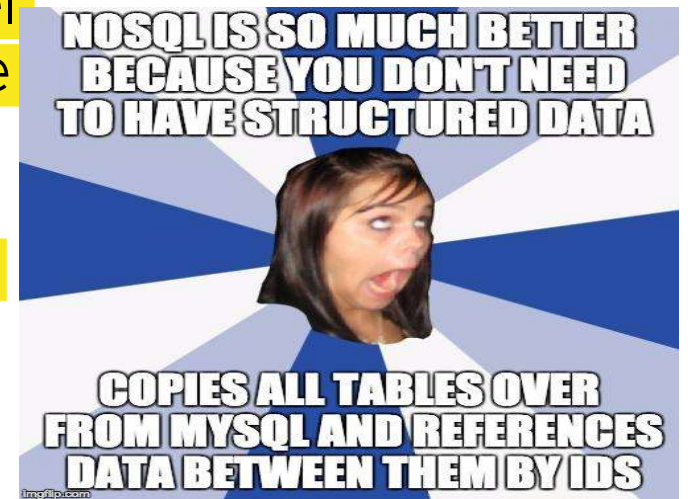
Please Notice This



La ausencia de un esquema fijo (*schema-less*) no quiere decir que no exista una etapa de diseño y modelado

Diseño de BDs documentales

- No hay heurísticas que permitan pasar desde conceptual a documental como en el caso relacional que garanticen la calidad de la solución
- El diseño depende del tipo de operaciones que voy a hacer además de la naturaleza de los datos
- Diseñar “a la relacional” no suele ser lo mejor



Diseño de BDs documentales

¿Cuáles son las variables que tenemos?

- Decidir cómo se van a representar los elementos de la realidad.
- Decidir cómo representar las relaciones entre elementos.

Estrategias para representar relaciones

```
{
  "imdbID":"tt0944947",
  "Type":"series",
  "Title":"Game of Thrones",
  "Genre":"Adventure, Drama, Fantasy",
  "Actors":"Peter Dinklage, Lena Headey, Emilia Clarke, Kit Harington",
  "imdbRating":"9.5",
  "imdbVotes":"1,031,056"
}
{
  "imdbID":"tt1877832",
  "Type":"movie"
  "Title":"X-Men: Days of Future Past",
  "Year":"2014",
  "Genre":"Action, Adventure, Fantasy",
  "Actors":"Hugh Jackman, Michael Fassbender, Jennifer Lawrence, Peter Dinklage",
  "imdbRating":"8.0",
  "imdbVotes":"514,203"
}
```



Estrategia 1: documentos embebidos

```
{
  "imdbID": "tt0944947",
  "Type": "series",
  "Title": "Game of Thrones",
  "Genre": "Adventure, Drama, Fantasy",
  "Actors": [
    { "imdbID": "nm0227759", "name": "Peter Dinklage" },
    { "imdbID": "nm0372176", "name": "Lena Headey" },
    { "imdbID": "nm3229685", "name": "Kit Harington" }
  ]
  "imdbRating": "9.5",
  "imdbVotes": "1,031,056"
}

{
  "imdbID": "tt1877832",
  "Type": "movie",
  "Title": "X-Men: Days of Future Past",
  "Year": "2014",
  "Genre": "Action, Adventure, Fantasy",
  "Actors": [
    { "imdbID": "nm0413168", "name": "Hugh Jackman" },
    { "imdbID": "nm1055413", "name": "Michael Fassbender" },
    { "imdbID": "nm2225369", "name": "Jennifer Lawrence" },
    { "imdbID": "nm0227759", "name": "Peter Dinklage" }
  ]
  "imdbRating": "8.0",
  "imdbVotes": "514,203"
}
```



Documentos embebidos

VENTAJAS

- Recupero toda la información relevante en una sola consulta.
 - Evita implementar joins en el código de la aplicación o utilizar \$lookup.
- Actualizo la información relacionada como una única operación atómica.
 - Por defecto, todas las operaciones CRUD sobre un mismo documento son compatibles con ACID.

LIMITACIONES

- Duplicación de datos
- Los documentos grandes suponen más sobrecarga si la mayoría de los campos no son relevantes.
 - Se puede aumentar el rendimiento de las consultas limitando el tamaño de los docs
- MongoDB tiene un límite de tamaño de documento de 16 MB.
 - Si estoy colocando demasiados datos dentro de un único documento, podría alcanzar este límite.

Estrategia 2: documentos referenciados

```
{
  "imdbID":"tt0944947",
  "Type":"series",
  "Title":"Game of Thrones",
  "Genre":"Adventure, Drama, Fantasy",
  "Actors":
    ["nm0227759", "nm0372176", "nm3229685" ]
  "imdbRating":"9.5",
  "imdbVotes":"1,031,056"
}
{
  "imdbID":"tt1877832",
  "Type":"movie"
  "Title":"X-Men: Days of Future Past",
  "Year":"2014",
  "Genre":"Action, Adventure, Fantasy",
  "Actors":
    [ "nm0413168", "nm1055413", "nm2225369", "nm0227759" ]
  "imdbRating":"8.0",
  "imdbVotes":"514,203"
}
...
{ "id":"nm0227759",
  "name":"Peter Dinklage",
  "description":"Actor, Battle of the Bastards"
}
```



Documentos referenciados

VENTAJAS

- Evito duplicación de datos
- Menor tamaño en los documentos
- La información a la que se accede con poca frecuencia no se retorna en cada consulta.

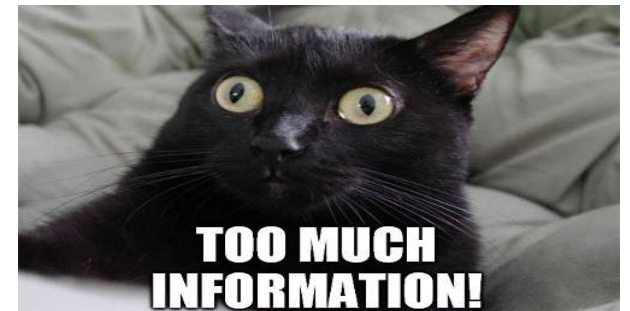
LIMITACIONES

- Para recuperar toda la información relevante debo utilizar \$lookup.
- La actualización puede impactar a más de un documento (transacción).
- Aunque el operador de transacción está disponible desde la versión 4.0, es un anti-patrón depender demasiado de su uso.

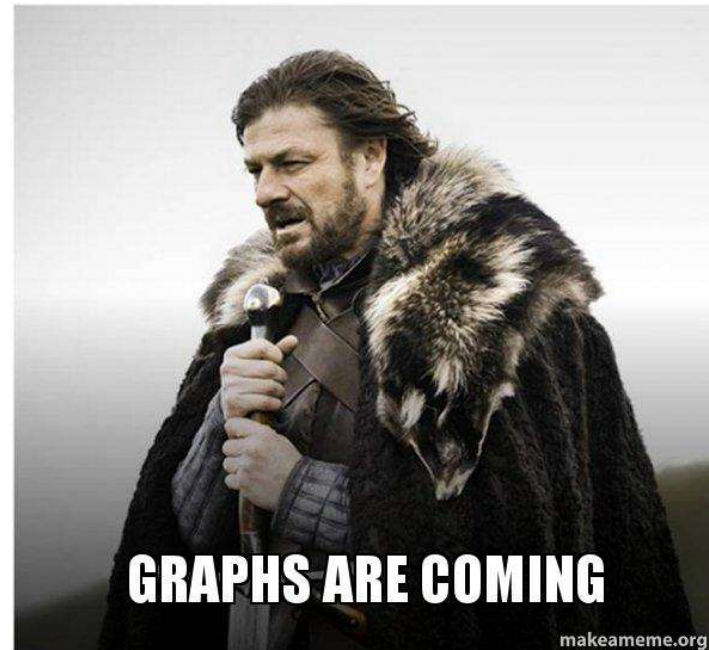
Aspectos a tener en cuenta en el diseño

- Atomicidad de las operaciones (ACID a nivel de documento)
- Restricción de tamaño máximo de cada documento.
- Restricciones de tamaño y cantidad de documentos en las colecciones
- Operaciones a realizar y su frecuencia
- Ciclo de vida del dato (permanente, volátil?)
- Necesidades de particionamiento (sharding)

Estos aspectos condicionan los diseños
posibles

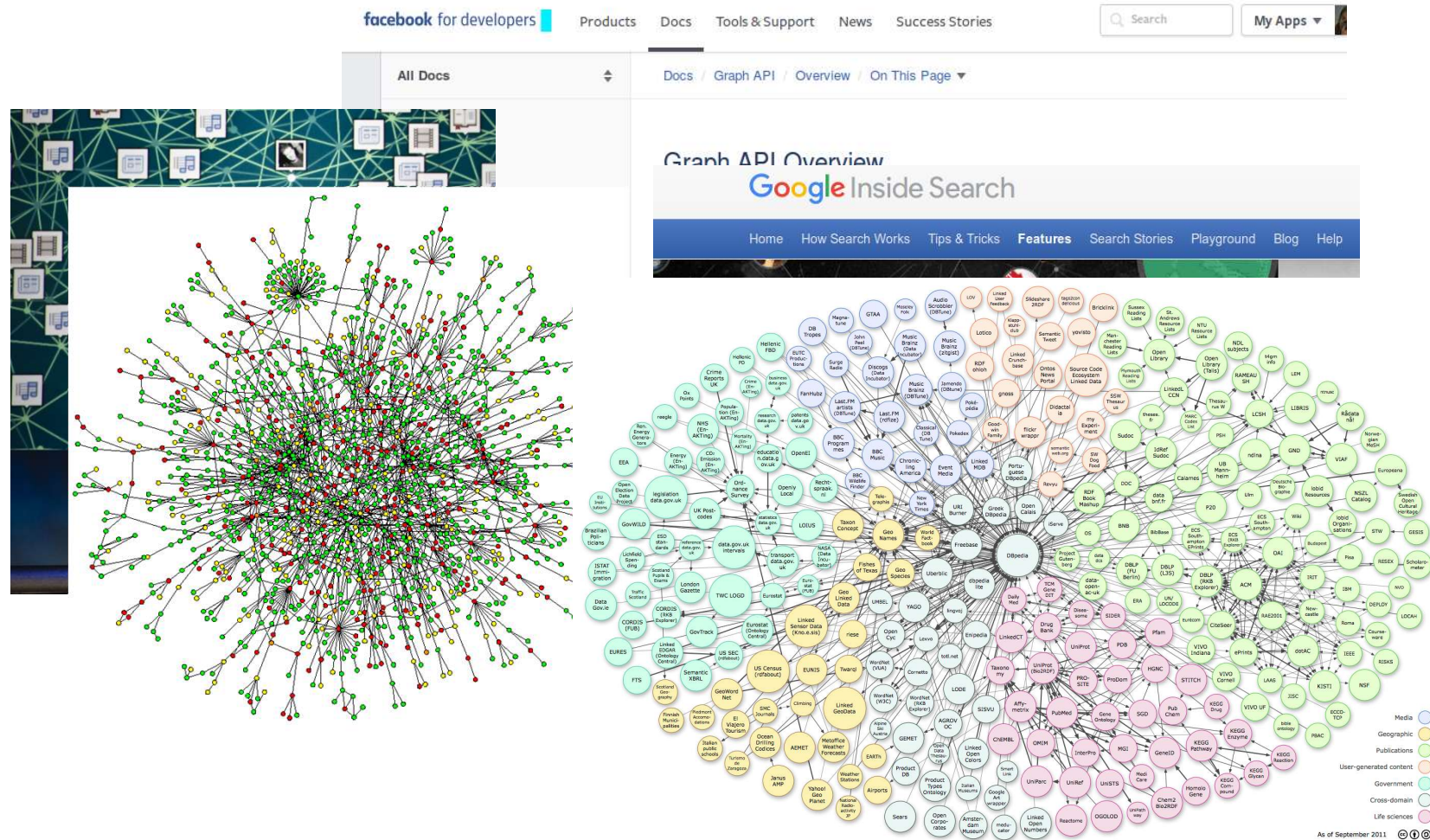


Bases de datos de grafos



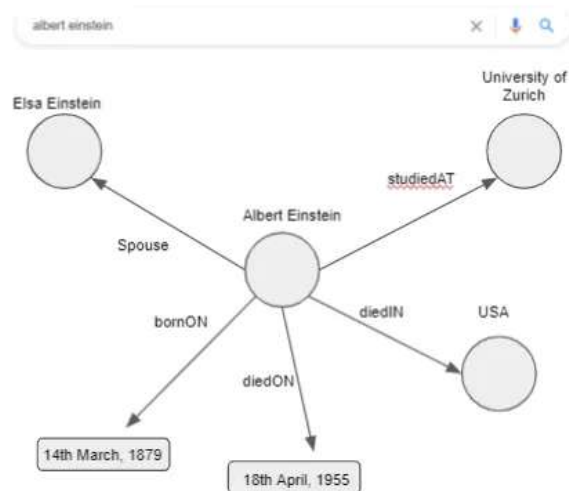
Instituto de Computación, FING, Udelar
CC-BY Lorena Etcheverry lorenae@fing.edu.uy

Hay casos en que interesa más modelar las relaciones entre cosas que las cosas!



Knowledge Graphs

Google Knowledge Panel



Albert Einstein
Theoretical physicist

Albert Einstein was a German-born theoretical physicist, widely acknowledged to be one of the greatest physicists of all time. Einstein is known widely for developing the theory of relativity, but he also made important contributions to the development of the theory of quantum mechanics. [Wikipedia](#)

Born: 14 March 1879, Ulm, Germany
Died: 18 April 1955, Penn Medicine Princeton Medical Center, New Jersey, United States
Spouse: [Elsa Einstein](#) (m. 1919–1936), [Mileva Marić](#) (m. 1903–1919)
Education: [University of Zurich](#) (1905), [ETH Zürich](#) (1896–1900), [MORE](#)

Books [View 35+ more](#)

- [Relativity : the special a...](#) 1916
- [The World As I see It](#) 1934
- [Out of My Later Years](#) 1950
- [The Evolution of Physics](#) 1938

Quotes [View 7+ more](#)

- Imagination is more important than knowledge.*
- If you can't explain it simply, you don't understand it well enough.*
- Life is like riding a bicycle. To keep your balance you must keep moving.*

People also search for [View 15+ more](#)

- [Eduard Einstein](#)
- [Isaac Newton](#)
- [Elsa Einstein](#)
- [Stephen Hawking](#)

Fuente: <https://towardsdatascience.com/a-guide-to-the-knowledge-graphs-bfb5c40272f1>

¿en qué escenarios son adecuadas las BD de grafos?

- Las entidades están muy conectadas a través de relaciones descriptivas.
- Existen relaciones cíclicas o entidades autorreferenciadas.
 - Esto suele ser un reto cuando se utilizan bases de datos relacionales o documentales.
- Las relaciones entre entidades evolucionan dinámicamente.
 - especialmente aplicable a datos jerárquicos o estructurados en árbol con muchos niveles.
- Existen relaciones de muchos a muchos entre las entidades.
- Existen requisitos de escritura y lectura tanto en las entidades como en las relaciones.

Agenda

- Modelos en bases de datos de grafos
 - El modelo *Property graph*
 - RDF y modelos en la web semántica
 - Diseño de bases de datos de grafos
- Motores de bases de datos de grafos
 - Estrategias de almacenamiento
 - Lenguajes de consulta
 - Modelos de procesamiento distribuido

Modelos en bases de datos de grafos

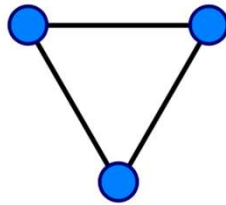
¿Qué es un grafo?

Un grafo G consiste en
un conjunto de nodos o vértices V ,
y un conjunto de aristas E .
Las aristas conectan nodos entre si.

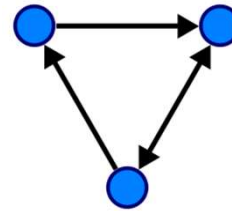
Las bases de datos de grafos
implementan diferentes variantes¹.

¹ *Survey of Graph Database Models*, Angles and
Gutierrez,
ACM Computing Surveys, 2008

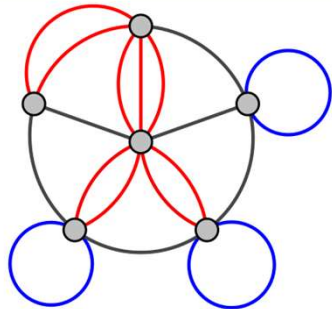
Grafos según su estructura



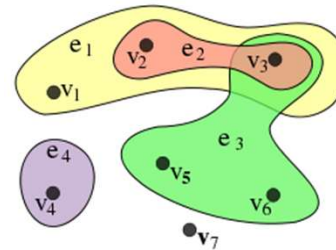
Grafo no
dirigido



Grafo dirigido

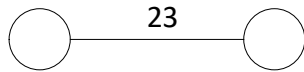


Multigrafos y
pseudografos

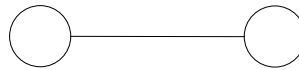


Hipergrafos

Grafos según los datos asociados



Etiquetas en las
aristas



Bob Alice

Etiquetas en los
nodos



[Bob,25]

[Alice,30]

Atributos en los
nodos

¿qué tipos de grafos se utilizan en las bases de datos de grafos?

Existen muchos modelos

(ver *Survey of Graph Database Models*, Angles and Gutierrez, ACM Computing Surveys, 2008)

Los sistemas más populares actualmente implementan el modelo *property graph* o RDF

Property Graph Model (PGM)

Pseudografos dirigidos.

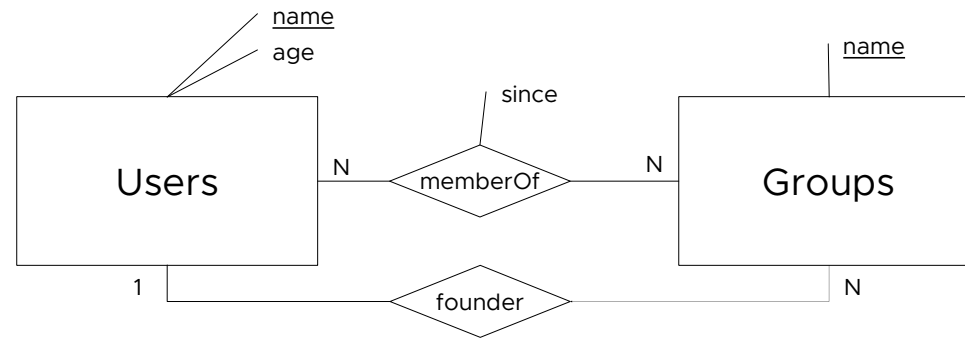
Parejas (clave,valor) llamadas *propiedades* asociadas a nodos y aristas.

Además puedo etiquetar nodos y aristas.

Cada nodo o arista puede tener más de una etiqueta

Neo4j y Titan son ejemplos de sistemas de bases de datos que soportan PGM.

CONCEPTUAL



LÓGICO

Users

<u>name</u>	age
Bob	25
Alice	null

Groups

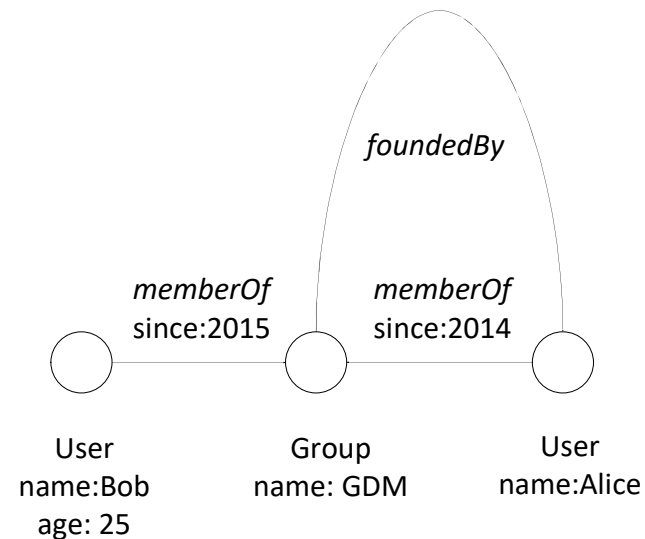
<u>name</u>	founder
GDM	Alice

GroupMembers

<u>userName</u>	<u>groupName</u>	since
Bob	GDM	2015
Alice	GDM	2014

Relacional

Property Graphs



Comparación entre Modelos: Cuándo Usar Cada Uno

- **Modelo Relacional:** Ideal cuando la estructura de datos es bien definida y se requiere integridad referencial. Por ejemplo, una base de datos de inventario con productos, clientes y órdenes.
- **Modelo Documental:** Más flexible para datos semiestructurados, como registros de clientes con preferencias personalizadas o logs de actividad en aplicaciones web.
- **Modelo de Grafos:** Ideal para analizar conexiones entre entidades, como redes de colaboración científica o relaciones entre palabras en un corpus de texto.



Brindarles definiciones y conceptos generales sobre gestión de datos y *data engineering*.

Desarrollar capacidades básicas, y proveer referencias para que uds. puedan profundizar.

Contenido



Por que no todo es una
planilla de cálculo



Data profiling

Conceptos de
Calidad de Datos



¿cómo se integran datos de
diferentes fuentes?

Algunas ideas y
herramientas

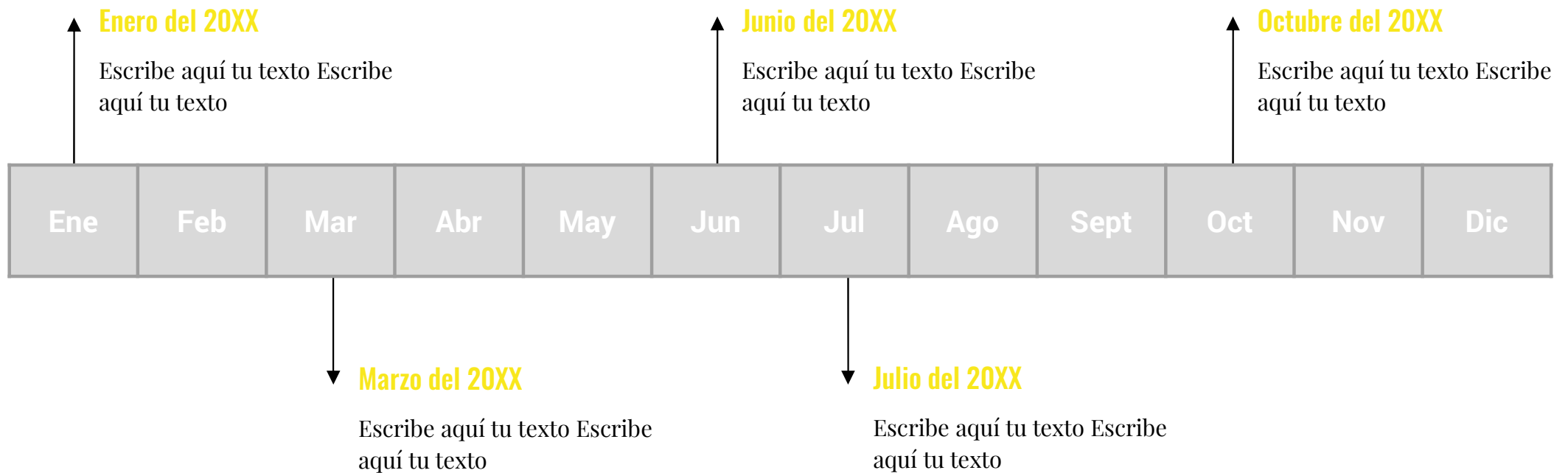


Privacidad y
anonimización

Modelos y sesgo

Hitos

Indica en qué fase del proyecto te encuentras y qué hitos faltan por completar



Apéndice

Muestra a la audiencia que te has anticipado a sus preguntas.

Deja un tiempo para que los asistentes hagan preguntas. Incluye en el Apéndice aquellas que crees que podrían surgir y respóndelas de forma precisa con datos que refuercen tus argumentos. Pon a prueba hasta qué punto ha entendido la audiencia tanto el problema como la solución que has propuesto. Las preguntas son una buena forma de iniciar un debate sobre el enfoque y, además, te permiten demostrar tu dominio de la materia.

Puesta en práctica

Paso 1

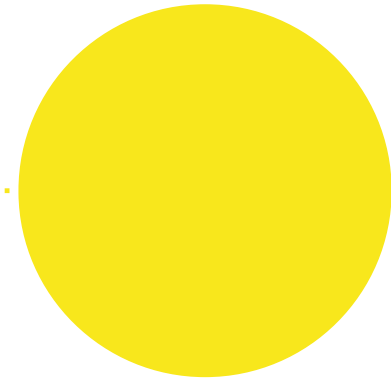
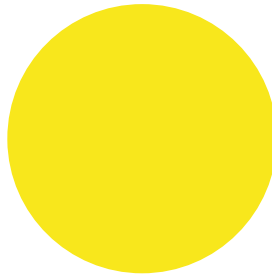
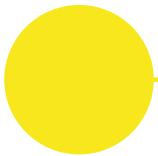
Escribe aquí tu texto Escribe
aquí tu texto

Paso 2

Escribe aquí tu texto Escribe
aquí tu texto

Paso 3

Escribe aquí tu texto Escribe
aquí tu texto





Tecnología: GPS + RFID

Modelo de ingresos

Escribe aquí tu texto Escribe
aquí tu texto Escribe aquí tu
texto Escribe aquí tu texto
Escribe aquí tu texto



¿Por qué ahora?

Escribe aquí tu texto

Escribe aquí tu texto

Escribe aquí tu texto

Escribe aquí tu texto

Escribe aquí tu texto

