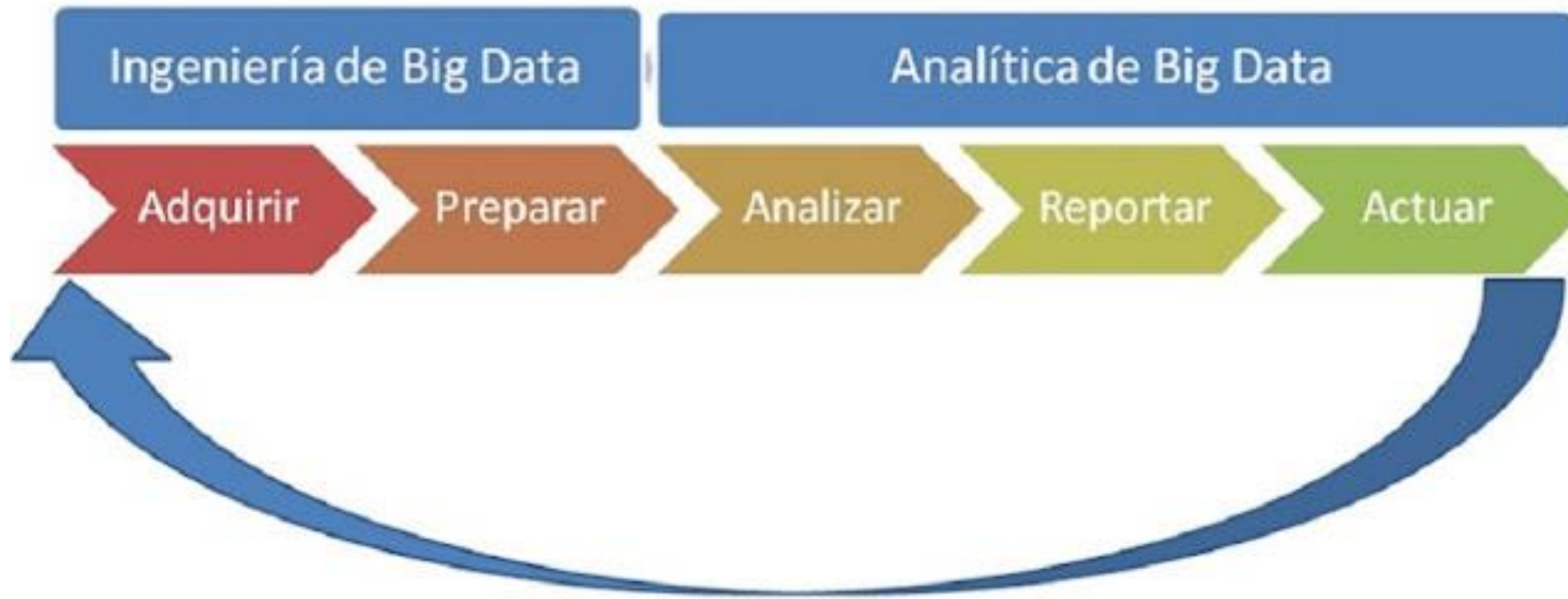


Introduccción al BigData

Ing. Ricardo Velasteguí (M.B.A.)

Proceso - BigData



► Pasos de un Proceso Big Data

- Adquirir datos
- Preparar/Limpiar datos
- Analizar los datos
- Reportar los datos (visualización)
- Actuar sobre los datos (Tomar decisiones)

► Arquitectura BigData

- Consultar arquitectura genérica para BigData.

Adquisición de Datos - ETL

- ▶ El proceso ETL está compuesto por las fases de extracción, transformación y carga.



▶ Seleccionar Fuentes

- ▶ Determinar los datos disponibles.
- ▶ Procedencia múltiple fuentes.
- ▶ Heterogéneos.
- ▶ Datos Estructurados de organizaciones.

▶ Fuentes convencionales y no convencionales Big Data

- ▶ Pueden provenir de ficheros (texto y hojas de cálculo)
- ▶ Datos de sitios Web (Técnica Webscraping)
- ▶ Servicios web (API/ curl / websockets)
- ▶ Datos públicos (open data)

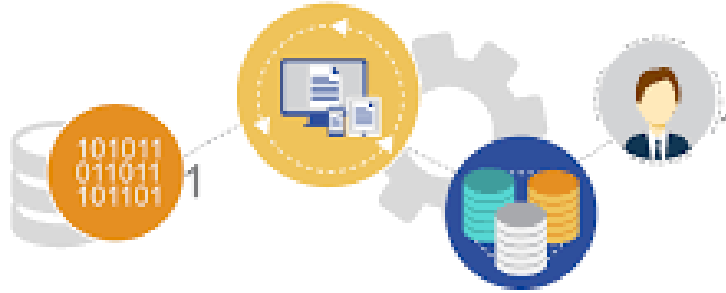
Adquisición de Datos - ETL



► Extracción

- El objetivo de esta fase es la extracción de los datos de cada una de las fuentes identificadas, realizando una copia exacta (sin procesamiento) de los datos de origen.
- En esta fase se realiza el acceso a los datos de origen de una forma eficiente, periódica y automatizada, definiendo la conexión/integración con las fuentes de datos internas o externas que se hayan identificado.

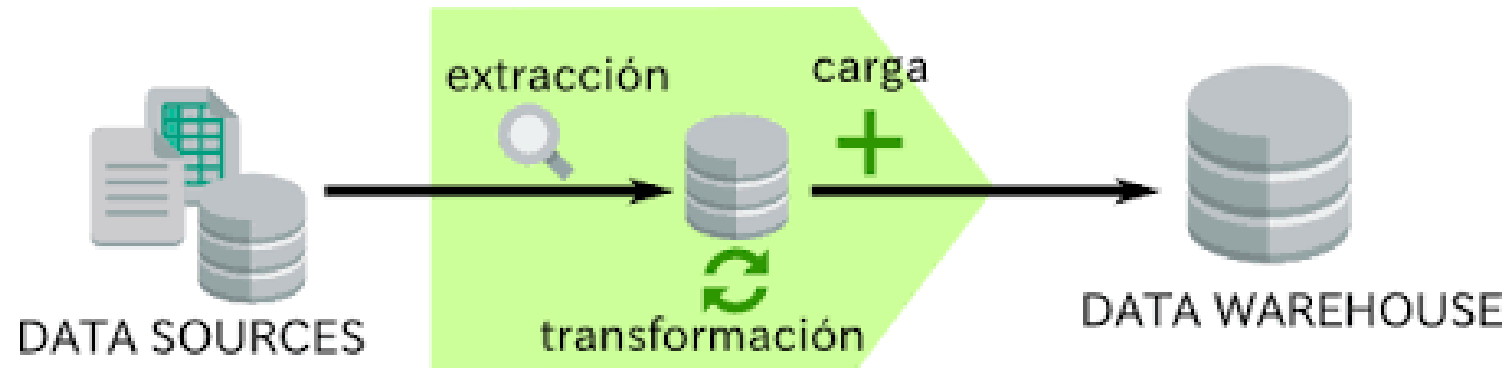
Adquisición de Datos - ETL



► Transformación

- Es el segundo de los procesos y consiste, principalmente, en limpiar y conformar la información extraída de todas las fuentes. Este paso es el más laborioso y en el que ETL agrega más valor.
- En esta fase, también se realiza el conformado y unificación de fuentes y datos, lo que aporta unicidad a la información, veracidad y genera maestros.
- En esta fase se da la propia transformación de la información, en la que se aplican las reglas de negocio, comprobaciones, datos referenciales, etc.

Adquisición de Datos - ETL

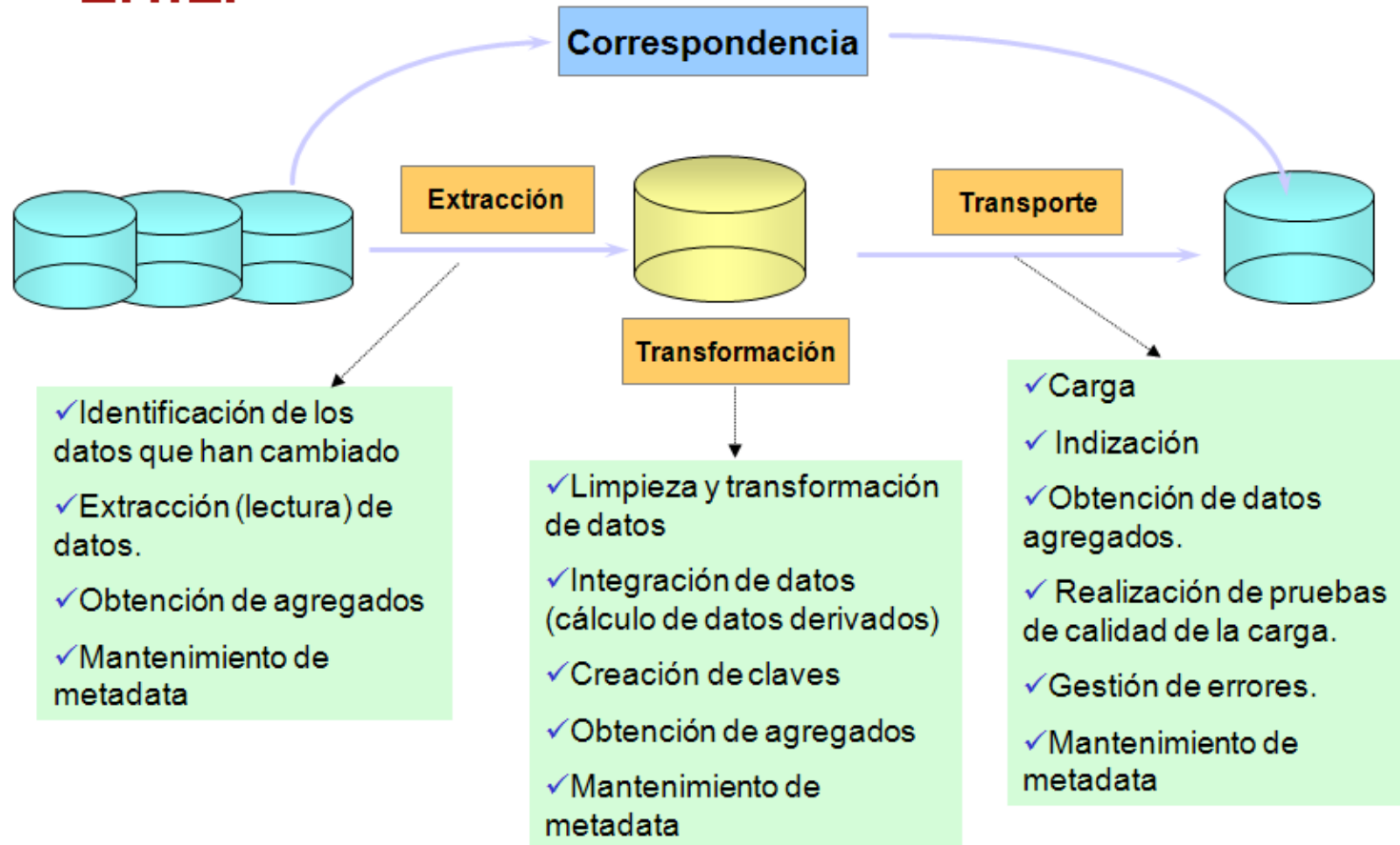


► Carga

- Es el paso final del proceso y su objetivo es realizar de forma eficiente, la carga de datos en el almacén de datos y, más concretamente, en cada uno de los modelos multidimensionales definidos. Por tanto, su diseño debe prestar especial atención a minimizar los tiempos de carga y tener en cuenta que se trata de procesos intensivos de inserción —y actualización, si fuera necesario— de registros en tablas de elevada volumetría.

Adquisición de Datos - ETL

E.T.L.



Fuentes de datos - Webscraping

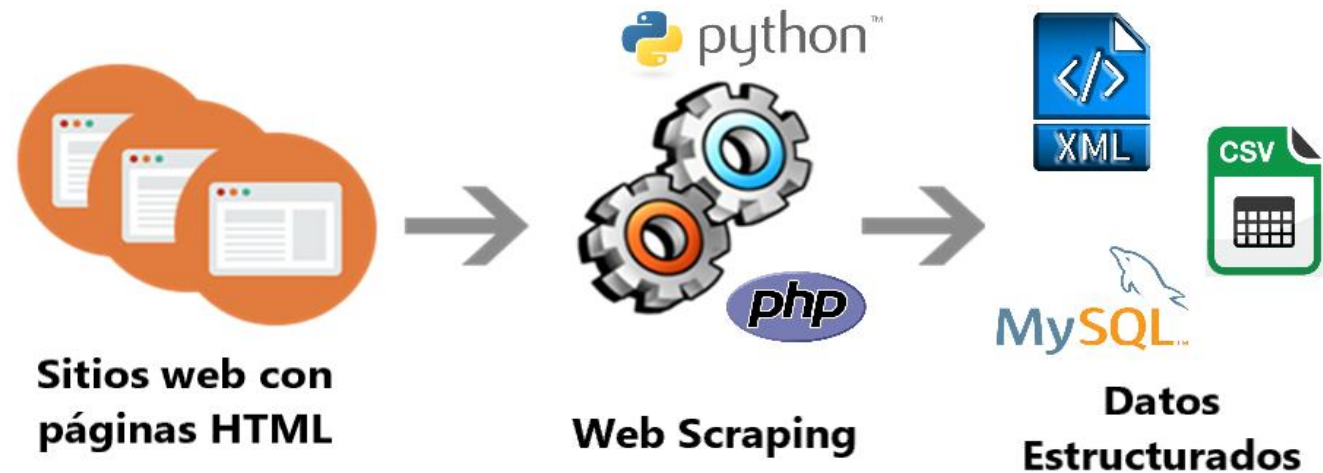


Es una técnica que permite extraer información específica de una página web. el ejemplo más común es cuando se realiza la acción de "copiar y pegar" información de varias páginas web y utilizarla de manera diferente en un documento nuevo.



Los buscadores como Google, Yahoo, Bing están programados para realizar web scraping ya que se encargan de rastrear sitios por toda la red, analizan sus contenidos y finalmente lo clasifican mostrando un listado con las coincidencias encontradas.

Webscraping con datos estructurados



► Ejemplos Herramientas Básicas de Webscraping

- ImportHTML de Google Sheets(Hojas de cálculo).
- Plugin Chrome - Table Capture.
- Tabula (Captura datos de PDF).
- Hunter.io (Captura de direcciones de e-mail).
- Octoparse (Herramienta profesional de Webscraping)

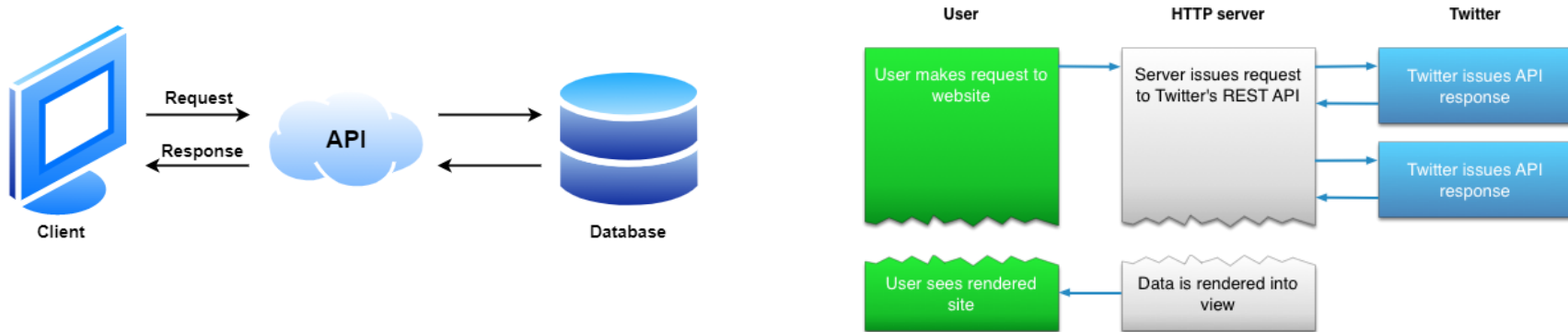
► Otras Herramientas Webscraping

- Consultar otras herramientas de webscraping para BigData.

Fuentes de Datos - API con datos semiestructurados

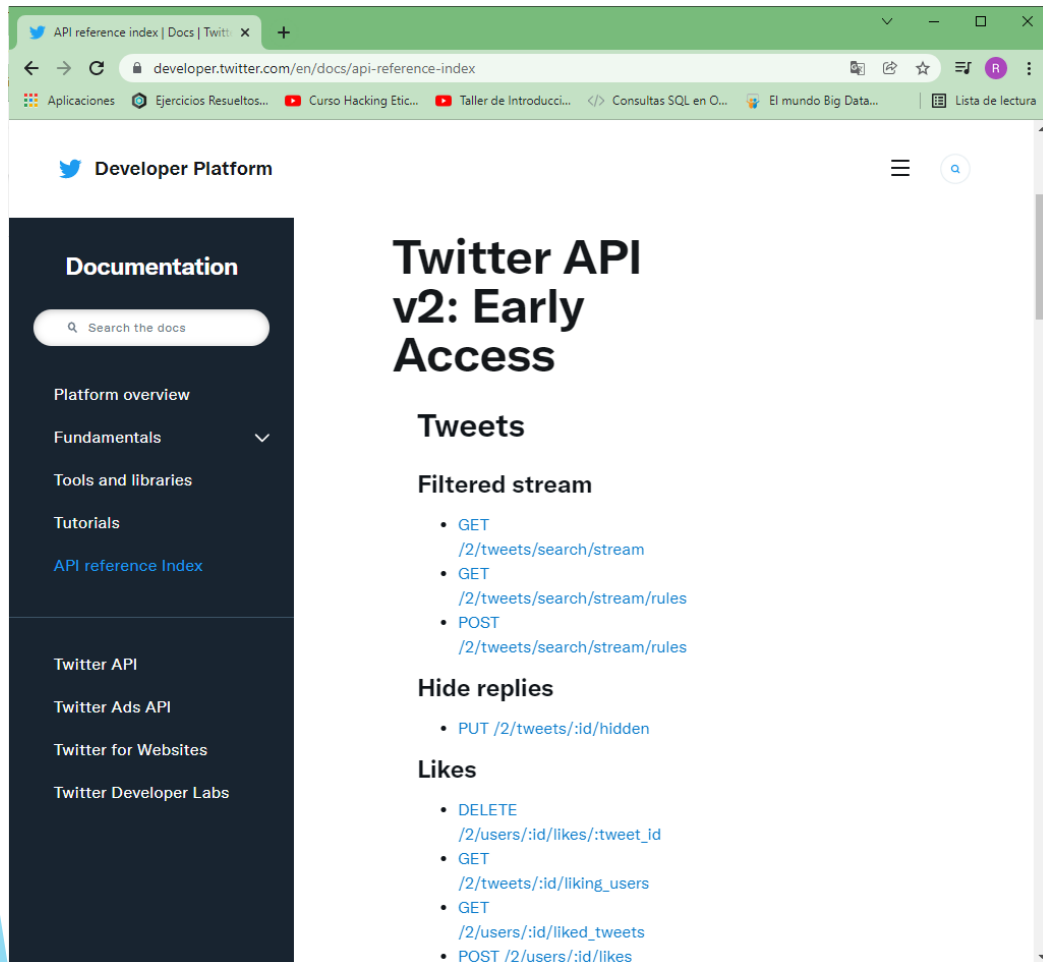
► API

- El término API es una abreviatura de Application Programming Interfaces, que en español significa interfaz de programación de aplicaciones. Se trata de un conjunto de definiciones y protocolos que se utiliza para desarrollar e integrar el software de las aplicaciones, permitiendo la comunicación entre dos aplicaciones de software a través de un conjunto de reglas.



API	Petición	Max. datos por petición (página)	Cada 15 minutos
REST	<u>GET statuses/user_timeline</u>	200 tuits	$900 * 200 = 180.000$ tuits
REST	<u>GET users/show</u>	1 perfil	$1 * 900 = 900$ perfiles
REST	<u>GET followers/list</u>	200 perfiles	$200 * 15 = 3.000$ perfiles
REST	<u>GET followers/ids</u>	5.000 ids de usuario	$5.000 * 15 = 75.000$ ids
Search	<u>GET search/tweets</u>	100 tuits	$180 * 100 = 18.000$ tuits
Streaming	<u>POST statuses/filter</u>	—	Máximo de 45.000 tuits

Fuentes de Datos - API con datos semiestructurados



API reference index | Docs | Twitter

Developer Platform

Documentation

Search the docs

Platform overview

Fundamentals

Tools and libraries

Tutorials

API reference Index

Twitter API

Twitter Ads API

Twitter for Websites

Twitter Developer Labs

Twitter API v2: Early Access

Tweets

Filtered stream

- GET </2/tweets/search/stream>
- GET </2/tweets/search/stream/rules>
- POST </2/tweets/search/stream/rules>

Hide replies

- PUT </2/tweets/:id/hidden>

Likes

- DELETE /2/users/:id/likes/:tweet_id
- GET /2/tweets/:id/liking_users
- GET /2/users/:id/liked_tweets
- POST </2/users/:id/likes>

Error response structures

Twitter API v2:

The error structure from the Twitter API v2 will always include the JSON object and elements below. The type includes a direct link to the error description on this index page.

```
1 {
2   "errors": [
3     {
4       "parameters": {
5         "end_time": [
6           "2026-10-31T23:59Z"
7         ]
8       },
9       "message": "Invalid 'end_time': '2026-10-31T23:59Z'. 'end_time' must be a minimum"
10    }
11  ],
12  "title": "Invalid Request",
13  "detail": "One or more parameters to your request was invalid.",
14  "type": "https://api.twitter.com/2/problems/invalid-request"
15 }
```

Twitter API HTTP status codes

Code	Text	Version	Description	Troubleshooting tips
200	OK	V1.1 V2	The request was successful!	

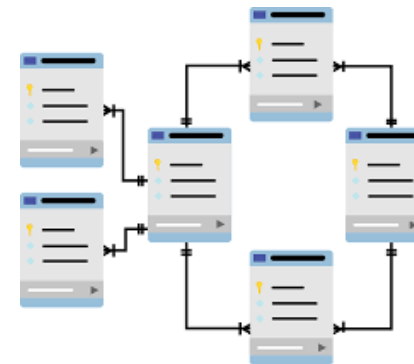
Tipos de Datos - BigData



Datos estructurados (Structured data)

- ▶ Los datos estructurados tienen perfectamente definido la longitud, el formato y el tamaño de sus datos.
 - ▶ Se almacenan en formato tabla, hojas de cálculo o en bases de datos relacionales.

	nombre	color	edad	altura	peso	puntuacion
1:	Paco	Rojo	24	182	74.8	83
2:	Juan	Green	30	170	70.1	500
3:	Andres	Amarillo	41	169	60.0	20
4:	Natalia	Green	22	183	75.0	865
5:	Vanesa	Verde	31	178	83.9	221
6:	Miriam	Rojo	35	172	76.2	413
7:	Juan	Amarillo	22	164	68.0	902



- ▶ Los podríamos ver como si fuese un archivador perfectamente organizado donde todo está identificado, etiquetado y es de fácil acceso.
- ▶ Para este propósito se almacena la información en estructuras denominadas tablas, estas tablas pueden estar conectadas entre sí por claves comunes, el modelo no resulta sencillo de consultar por el usuario ya que puede requerir una compleja combinación de tablas.
- ▶ Los datos estructurados son cualquier tipo de dato que se encuentre en un campo fijo dentro archivo o registro, regularmente utilizados con sistemas gestores de base de datos relacionales (SGBD).

SGDB - Sistema Gestor de Base de Datos

- ▶ Un SGDB (en inglés DBMS: DataBase Management System) es un sistema de software que permite la definición de bases de datos; así como la elección de las estructuras de datos necesarios para el almacenamiento y búsqueda de los datos, ya sea de forma interactiva o a través de un lenguaje de programación SQL.
- ▶ Un SGDB relacional es un modelo de datos que facilita a los usuarios describir los datos que serán almacenados en la base de datos junto con un grupo de operaciones para manejar los datos a través de las propiedades ACID (en inglés de Atomicity, Consistency, Isolation and Durability: Atomicidad, Consistencia, Aislamiento y Durabilidad, en español.).



- ▶ Las características que se contemplan en un SGDB relacional son:
 - ▶ Una base de datos se compone de varias tablas, interdependientes entre sí denominadas relaciones.
 - ▶ No pueden existir dos tablas con el mismo nombre ni registro.
 - ▶ Cada tabla es a su vez un conjunto de campos (columnas) y registros (filas).
 - ▶ La relación entre una tabla padre y un hijo se lleva a cabo por medio de las llaves primarias y llaves foráneas (o ajenas).
 - ▶ Las llaves primarias son la clave principal de un registro dentro de una tabla y estas deben cumplir con la integridad de datos.
 - ▶ Las llaves foráneas se colocan en la tabla hija, contienen el mismo valor que la llave primaria del registro padre; por medio de estas se hacen las formas relacionales.

SGDB - Sistema Gestor de Base de Datos

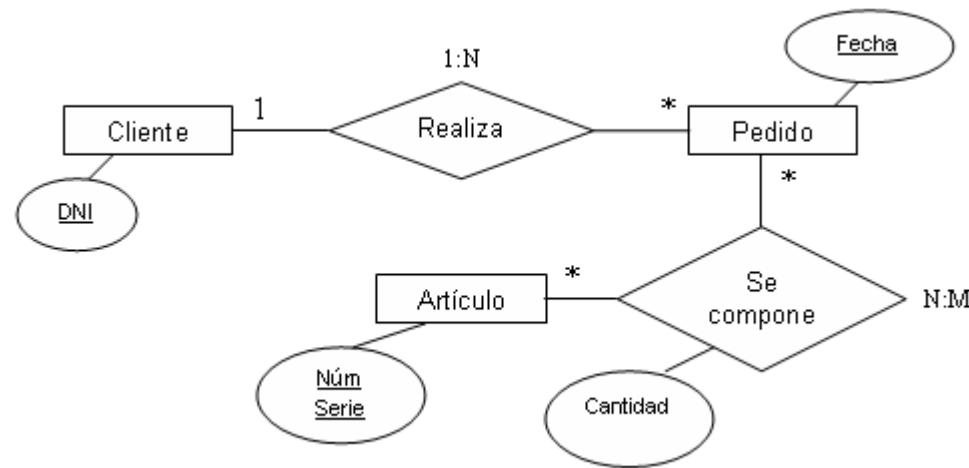
► Algunos SGDB populares son:

- [Oracle](#).- SGDB empresarial, el más completo y robusto, destacando soporte de transacciones, estabilidad, escalabilidad, multiplataforma.
- [Mysql](#).- Es multihilo y multiusuario utilizado en las páginas web actuales. El más usado con software libre, es código abierto con licencia comercial disponible.
- [Microsoft SQL Server](#).- Basado en Transact-SQL, es un sistema cliente/servidor, proporciona integridad de datos, optimización de consultas, control de concurrencia, backup y recuperación.
- [MariaDB](#).- Es de código abierto, reemplaza a MySQL. Es rápida, escalable y robusta, con un rico ecosistema de motores de almacenamiento, complementos y muchas otras herramientas.
- [PostgreSQL](#).- Está orientado a objetos y es libre, publicado bajo la licencia BSD. PostgreSQL es manejado por una comunidad de desarrolladores, denominada el PGDG (PostgreSQL Global Development Group).
- [SQLite](#).- De dominio público, con autocontenido, fiabilidad e integrable con muchas aplicaciones, creado por D. Richard Hipp, es una pequeña librería, totalmente libre, muy utilizada en el desarrollo de dispositivos móviles.



Representación de Datos estructurados

- ▶ Los datos estructurados pueden ser representados gráficamente a través de un diagrama relacional.
- ▶ Modelo Entidad-Relación.- Un diagrama entidad-relación, también conocido como modelo entidad relación o ERD, es un tipo de diagrama de flujo que ilustra cómo las "entidades", como personas, objetos o conceptos, se relacionan entre sí dentro de un sistema.

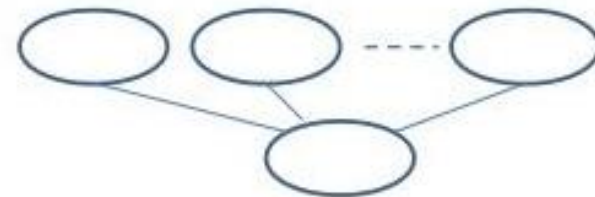
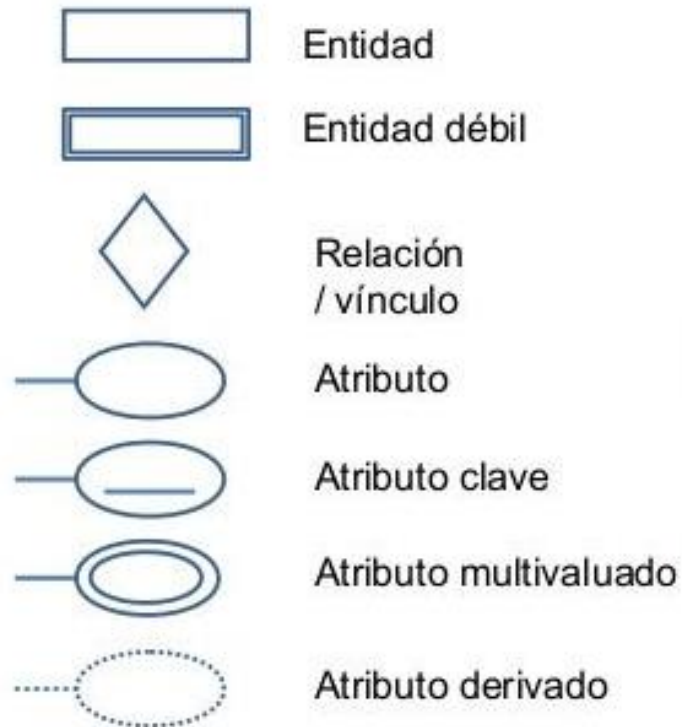


- ▶ Emplean un conjunto definido de símbolos, tales como rectángulos, diamantes, óvalos y líneas de conexión para representar la interconexión de entidades, relaciones y sus atributos. Son un reflejo de la estructura gramatical y emplean entidades como sustantivos y relaciones como verbos.

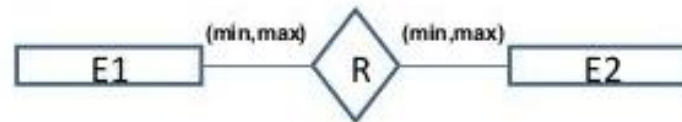
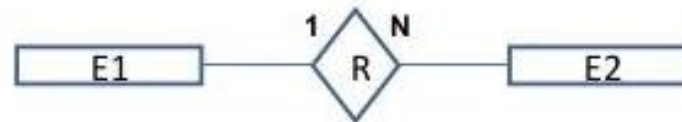
Representación de Datos estructurados

- Los datos estructurados pueden ser representados gráficamente a través de un diagrama relacional.

Simbología



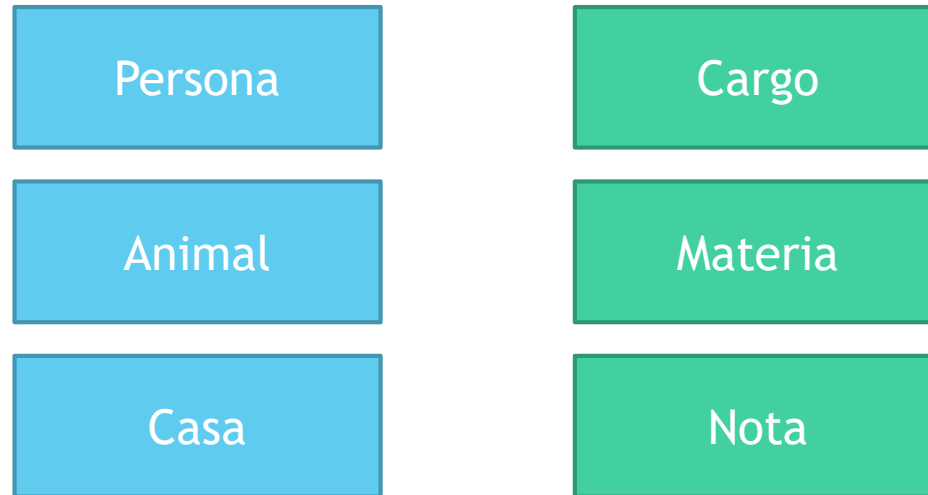
Atributo compuesto



Representación de Datos estructurados

► Partes de un diagrama Entidad-Relación (representación lógica).

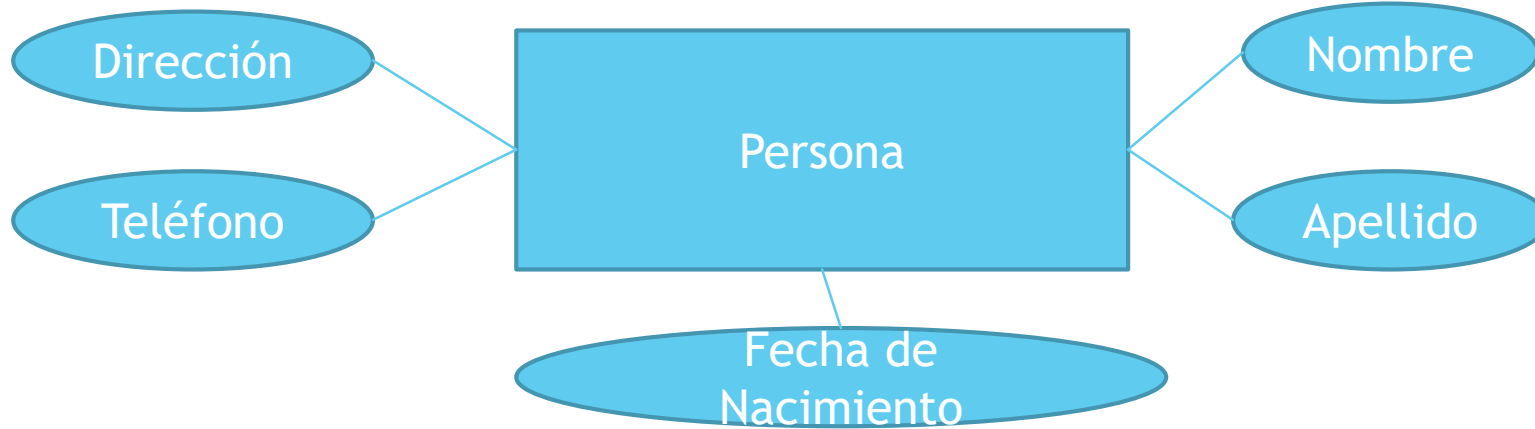
- Los diagramas ER se componen de entidades, relaciones y atributos.
- Entidad.- Las entidades representan cosas u objetos (ya sean reales o abstractos), que se diferencian claramente entre sí a través de atributos. Por ejemplo: un cliente, estudiante, auto o producto. Por lo general se muestran como un rectángulo.



- Una entidad puede ser un objeto con existencia física como: una persona, un animal, una casa, etc. (entidad concreta); o un objeto con existencia conceptual como: un puesto de trabajo, una asignatura de clases, notas de alumno, etc. (entidad abstracta).
- Las entidades son el fundamento del modelo entidad relación. Podemos adoptar como definición de entidad cualquier cosa o parte del mundo que es distinguible del resto. Por ejemplo, en un sistema bancario, las personas y las cuentas bancarias se podrían interpretar como entidades.

Representación de Datos estructurados

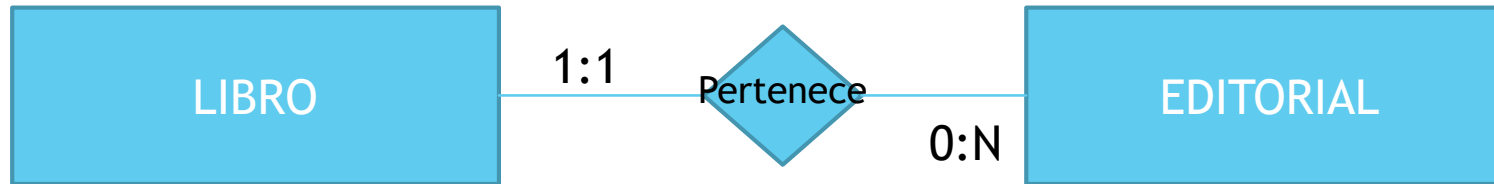
- ▶ Partes de un diagrama Entidad-Relación (representación lógica).
 - ▶ Atributos.- Los atributos son las características propias que poseen las Entidades, como bien podría ser por ejemplo en el caso de una Persona: Nombre, Apellido, Edad, Dirección, Teléfono, Fecha de Nacimiento, etc.



- ▶ Existen también diferentes tipos de atributos.
 - ▶ **Atributos Simples:** Los atributos simples son aquellos que tienen un solo valor para la entidad.
 - ▶ **Atributos Compuestos:** Son aquellos que pueden subdividirse en "sub-Atributos" por ejemplo el atributo Teléfono puede subdividirse a su vez en Celular, o Fijo.
 - ▶ **Atributos Derivados:** Son aquellos que derivan o se calculan a partir de otro/s atributos, por ejemplo "Edad" normalmente se puede calcular a partir de la fecha de nacimiento de la persona.

Representación de Datos estructurados

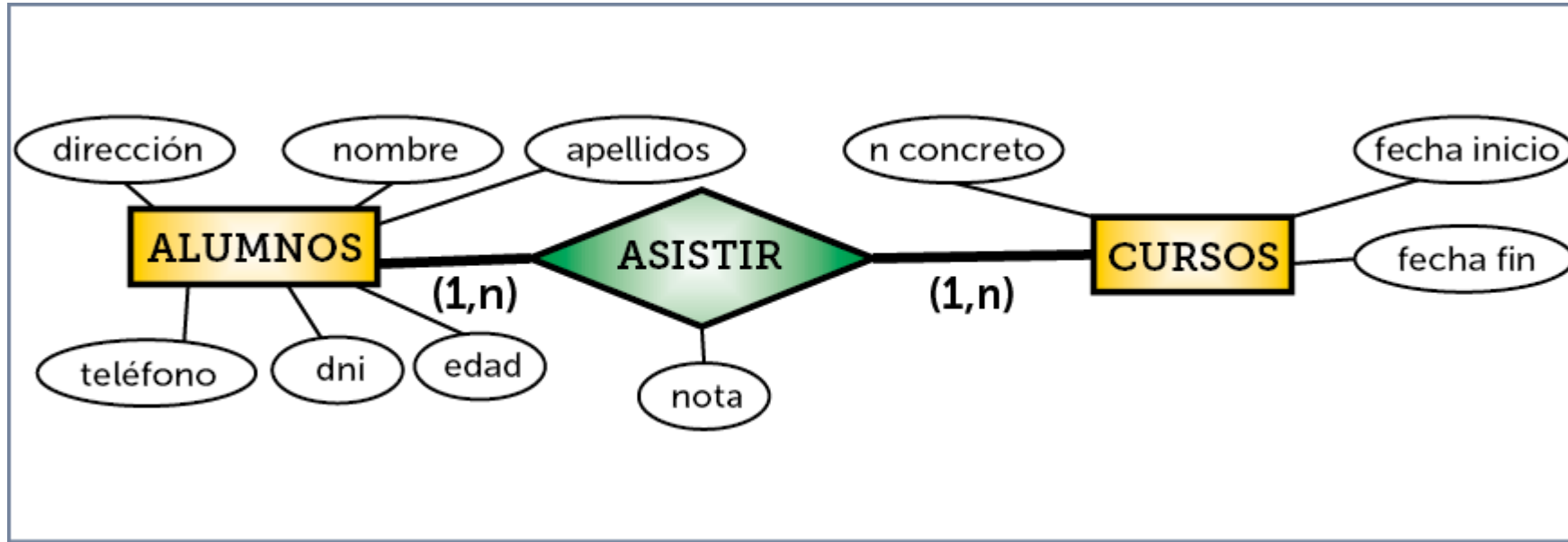
- ▶ Partes de un diagrama Entidad-Relación (representación lógica).
 - ▶ Relacion.- Describe cierta dependencia o asociación entre las entidades.



- ▶ Existe una cardinalidad en el conjunto de relaciones:
 - ▶ Relación de cardinalidad “uno a uno”.
 - ▶ Relación de cardinalidad “uno a varios”
 - ▶ Relación de cardinalidad “varios a uno”
 - ▶ Relación de cardinalidad “varios a varios”

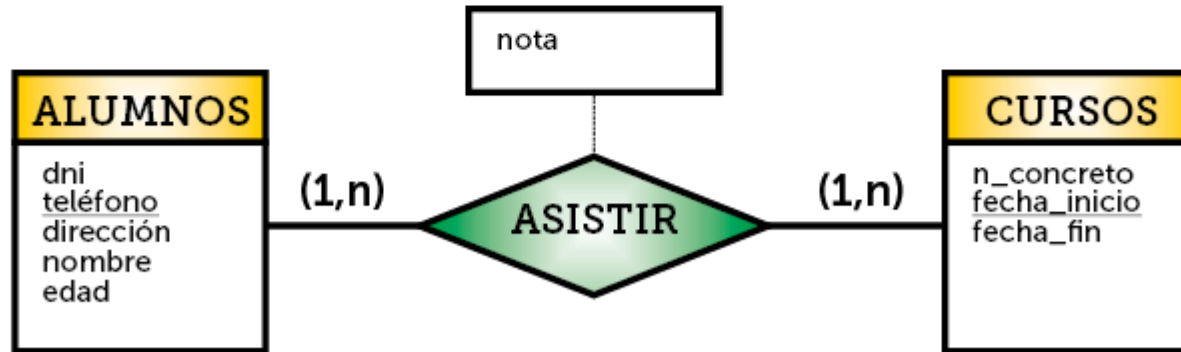
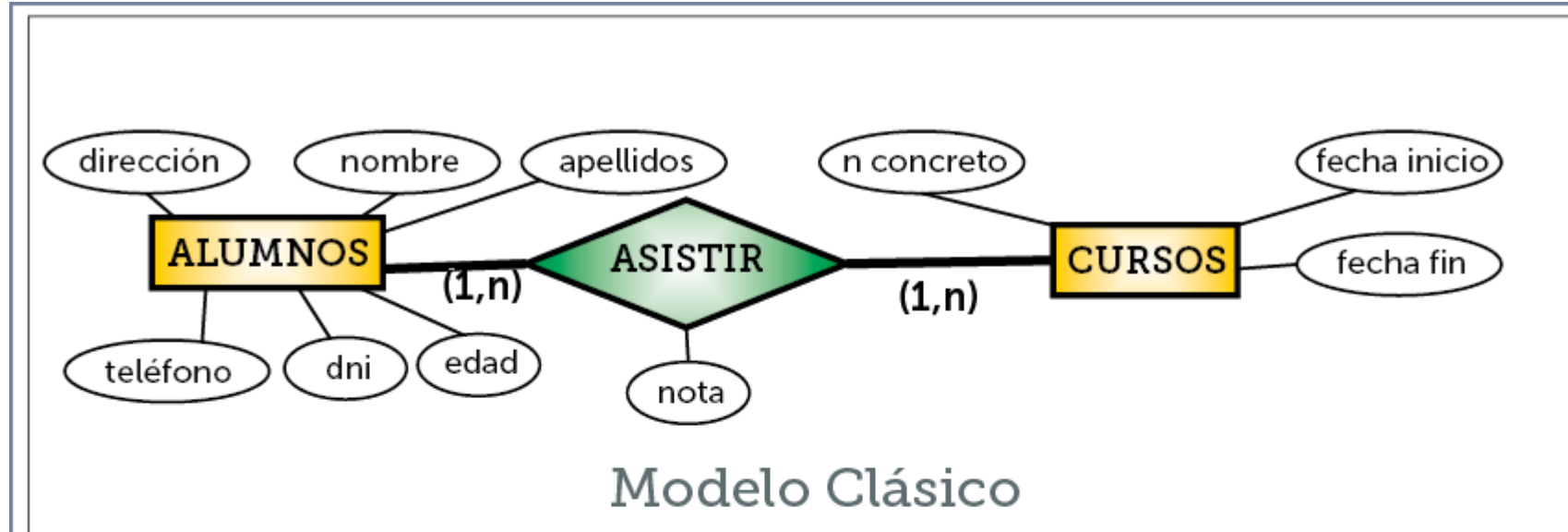
Representación de Datos estructurados

- Partes de un diagrama Entidad-Relación (representación lógica).



Representación de Datos estructurados

- Partes de un diagrama Entidad-Relación (representación lógica).



Representación de Datos estructurados

- ▶ Partes de un diagrama Entidad-Relación (representación lógica).
 - ▶ Claves.- Una clave es aquel atributo que nos permite diferenciar a una entidad de otra inequívocamente dentro de un conjunto de entidades. Normalmente los atributos que forman una clave se suelen encontrar subrayados.



PERSONAS	
CÉDULA	PK LLAVE PRIMARIA
NOMBRES	
APELLIDO	
GÉNERO	

- ▶ Existen diferentes tipos de claves:



- ▶ **Clave primaria:** definida por el diseñador de la base de datos permite identificar inequívocamente a la entidad del conjunto.
- ▶ **Clave foránea:** Es aquella que permite identificar inequívocamente una relación entre una o más de las entidades.

Representación de Datos estructurados

- Partes de un diagrama Entidad-Relación (representación lógica).

Claves Primarias y Claves Foráneas

Cada entidad tiene una **clave primaria** o **campo llave** que **identifica unívocamente** al conjunto de datos.

Cuando en una entidad figura la clave primaria de otra entidad, ésta se denomina **clave foránea**.

Las entidades se relacionan entre sí a través de las **claves foráneas**.

