Formatos
de los datos

Gabriel Moyà – gabriel.moya@uib.es

Contenidos

- Organización de la información
- Diferentes tipos de formatos
  - Text
  - CSV
  - XML
  - JSON
  - HDF
  - SQL

Organización de la información

Los formatos de archivos son la manera estándar en que la información está codificada para el almacenamiento en un medio informático.

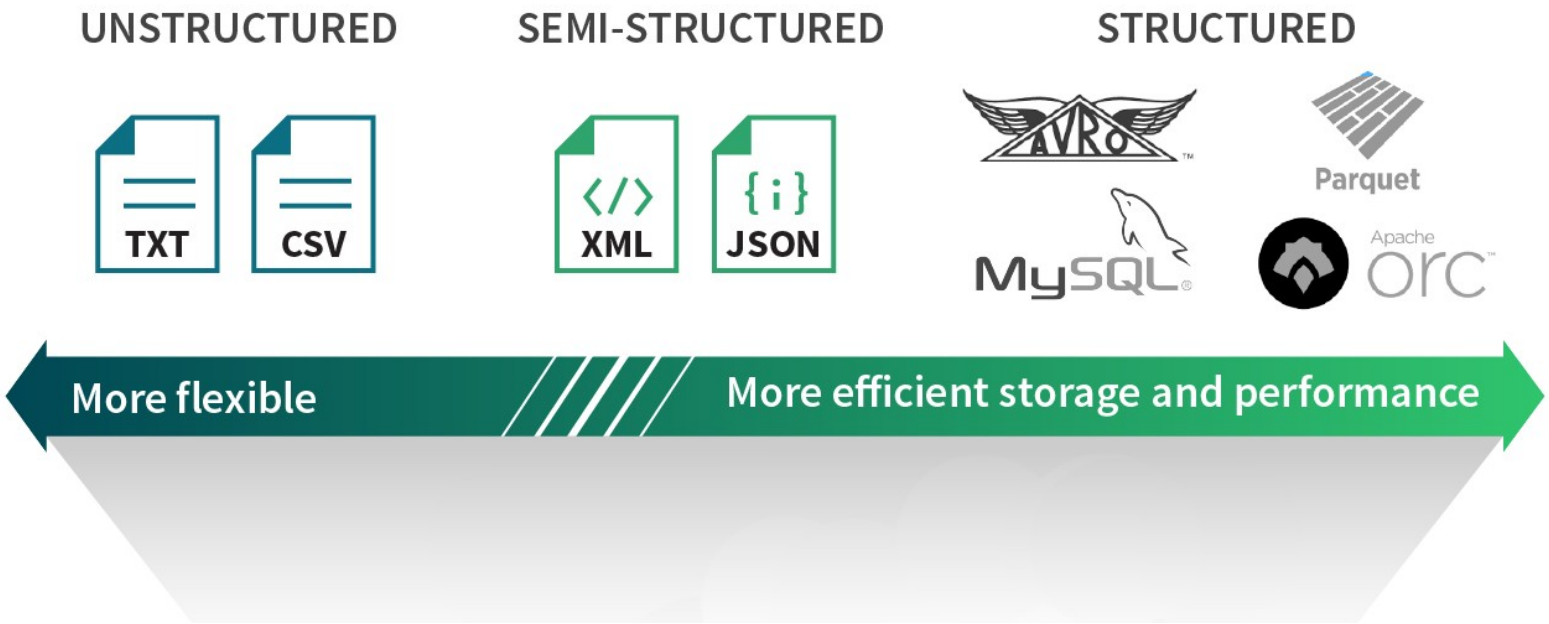
La información que podemos encontrar está organizada de diferentes formas. Para poderla procesar y almacenar es necesario conocer la existencia de los diferentes formatos de datos y como podemos operar con ellos.

Podemos clasificar estos formatos de datos en tres categorías: **datos estructurados, semi-estructurados y no estructurados**.

Organización de la información

Datos no estructurados

Las fuentes de datos no estructurados generalmente son los archivos de texto u objetos binarios que no contienen etiquetas ni metadatos (por ejemplo CSV) para definir la organización de datos.



Organización de la información

Datos semi estructurados

Las fuentes de datos semi-estructurados, presentan cierta estructura, pero no están organizadas en un modelo racional, como una tabla o un grafo.

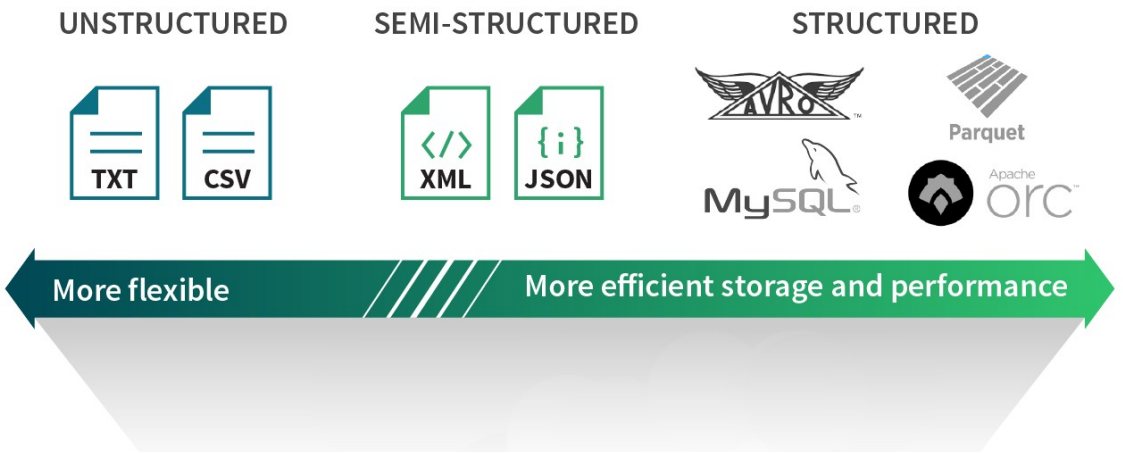
Como resultado, cada registro de datos tiene asociada cierta información que incrementa el conocimiento que podemos obtener de ella y la dota de cierto esquema. JSON y XML son los ejemplos más populares.

Las ventajas de este formato es que proporciona flexibilidad a la hora de expresar los datos, ya que cada registro se describe a sí mismo.

Organización de la información

Datos estructurados

Los datos estructurados tienen un alto nivel de organización que lo hace predecible, fácil de organizar y al que se puede acceder fácilmente. Los datos se introducen en campos específicos que contienen datos textuales o numéricas. Estos campos menudo tienen definido un tamaño máximo. Además de la estructura sólida de información, los datos estructurados tienen unas reglas muy específicas sobre cómo acceder a ellos.



Diferentes tipos de formato

Documentos de texto

Los documentos de texto (normalmente con extensión .txt) están diseñados para facilitar su lectura por parte de un humano. Típicamente, no incluyen metadatos estructurales, lo que significa que los desarrolladores de software necesitan crear un programa de análisis que pueda interpretar cada documento tal como aparece.

Al intercambiar archivos de texto plano entre sistemas operativos pueden aparecer algunos problemas. Por ejemplo: MS Windows, Mac OS X y otras variantes de Unix tienen su propia forma de decirle al ordenador que se ha llegado al final de la línea.

CSV

Los archivos CSV (separados por comas) son compactos y por tanto adecuados para transferir grandes conjuntos de datos con la misma estructura.

El formato es rudimentario, y esto implica que los datos son frecuentemente inservibles si no tenemos acceso a la documentación, ya que puede ser casi imposible saber el significado de las diferentes columnas. Por lo tanto, es importante que la documentación de los campos individuales sea precisa.

Es esencial que la estructura del archivo sea respetada, ya que la omisión de un único campo puede perturbar la lectura de todos los datos restantes del archivo.

CSV

Típicamente este tipo de ficheros son adecuados para almacenar la información que podemos representar en forma de tabla.

Los archivos del formato CSV se pueden importar y exportar desde programas que almacenan datos en tablas, como Microsoft Excel u OpenOffice Calc.

XML

XML (E**x**tensible **M**arkup **L**anguage) es un formato ampliamente utilizado para el intercambio de datos ya que ofrece una manera de organizar los datos de forma parcialmente estructurada.

Un lenguaje de marcado (como es XML) es un conjunto de símbolos que se pueden colocar en el texto de un documento para delimitar, etiquetar y relacionar sus diferentes partes.

Las etiquetas XML identifican los datos y se utilizan para almacenar y organizar los datos, estas etiquetas no están fijadas (como puede ocurrir con html) de ahí que se hable de un lenguaje extensible.

XML

XML es un estándar público, fue desarrollado por una organización llamada World Wide Web Consortium (W3C) y está disponible como estándar abierto.

Debido a la utilización de etiquetas de inicio y de final de informacion (entre '<' y '>') es un formato considerado pesado para la transferencia de información.

Ejemplo de XML

```
<note type = “card”>
   <to>Harry/to>
   <from>Hagrid</from>
   <heading>Recordatorio</heading>
    <body>¿Os apetece tomar el té conmigo esta tarde, a eso de las seis? </body>
</note>
```

JSON

JSON (**J**ava**S**cript **O**bject **N**otation) es un formato de archivo estándar y abierto que utiliza texto legible por humanos para transmitir conjuntos (u objetos) de datos que consisten en pares de clave-valor y arrays (o cualquier otro valor Serializable).

Una de las ventajas que tiene JSON es que puede ser leído por cualquier lenguaje de programación. Por lo tanto, puede ser usado para el intercambio de información entre diferentes tecnologías.

JSON

Actualmente es muy usado y en muchos casos sustituye XML como formato de intercambio de datos en la red debido a su **ligereza**, ya que al no tener marcas (<>) un fichero JSON con la misma estructura que un XML ocupa menos espacio en memoria.

Ejemplo de JSON

```
{
    “id”: 1,
    “nom”: “J.K Rowling”
    “aficions”= [“llegir”, “escriure”]
}
```

HDF

El formato HDF (Hierarchical Data Format)

- Permite obtener información acerca de los datos de un archivo desde dentro de ese archivo, sin necesidad de recurrir a fuentes externas.
- Permite almacenar datos de distinta naturaleza en un mismo archivo y relacionarlos entre ellos.
- Estandariza los formatos y las descripciones de los tipos de datos más comúnmente empleados.
- Se trata de un formato abierto, con sus especificaciones publicadas, lo que permite su implementación en diversas aplicaciones informáticas, facilitando la portabilidad, así como permitiendo al usuario desarrollar sus propias aplicaciones específicas.
- Es flexible y puede ser adaptado para almacenar cualquier tipo de dato.

Bases de datos

Las bases de datos, son un formato para ordenar e intercambiar información. SQL es el lenguaje estándar para almacenar, manipular y recuperar información a las bases de datos.

Existen dos grandes tipos de bases de datos:

Relacionales

No relacionales Las bases de datos no relacionales están orientadas a los documentos y permiten almacenar y recuperar datos en formatos que no sean tablas. Las bases de datos NoSQL son más flexibles y escalables.
Al trabajar con una base de datos NoSQL, se pueden agregar datos nuevos, sin tener que definirlos previamente en el esquema de la base de datos, lo que le permite procesar rápidamente grandes volúmenes de datos sin estructura o semiestructurados.

Bases de datos relacionales

Qué es una BDD relacional?

- Es una recopilación de elementos de datos con relaciones predefinidas entre ellos.
- Los elementos se organizan como un conjunto de tablas con columnas y filas
- Las tablas se utilizan para guardar información sobre los objetos que se van a representar en la base de datos.

Tablas

- Cada columna de una tabla guarda un determinado tipo de datos
- Las filas de la tabla representan una recopilación de valores relacionados de un objeto o entidad.
- Cada fila de una tabla podría marcarse con un identificador único denominado clave principal (en inglés, primary key o PK), mientras que filas de varias tablas pueden relacionarse con claves extranjeras (en inglés, foreign key, o FK).

Bases de datos relacionales

Primary Key and Foreign Key

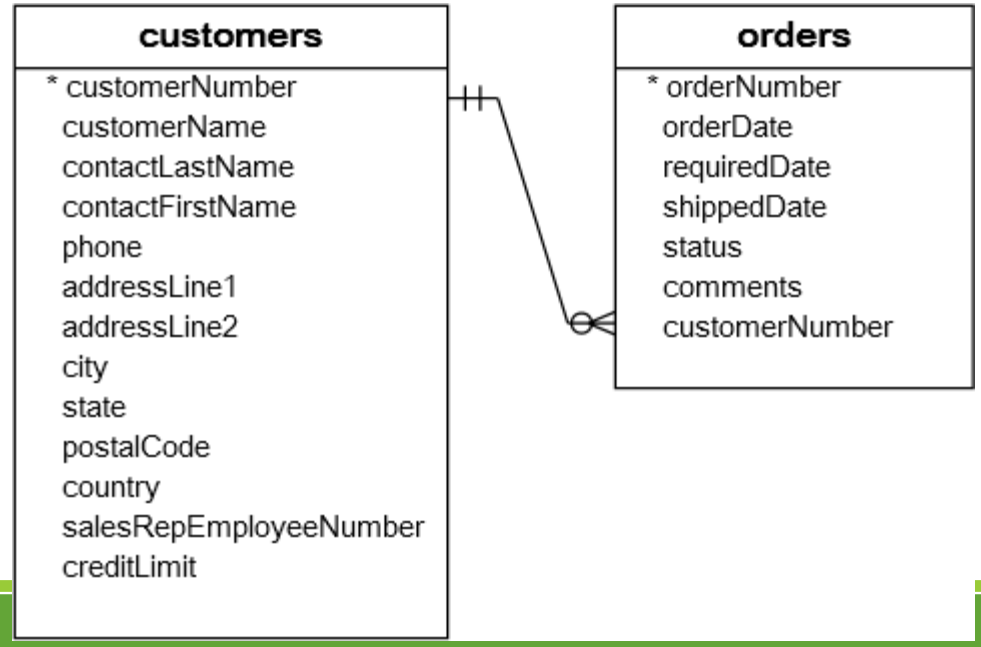
Primary key: Consiste en los valores de una o más columnas que se usán para identificar de forma univoca cada una de las filas contenidas en la tabla.

Foreign key: Es un conjunto de una o más columnas de una tabla a las que hace referencia la PK de otra tabla.

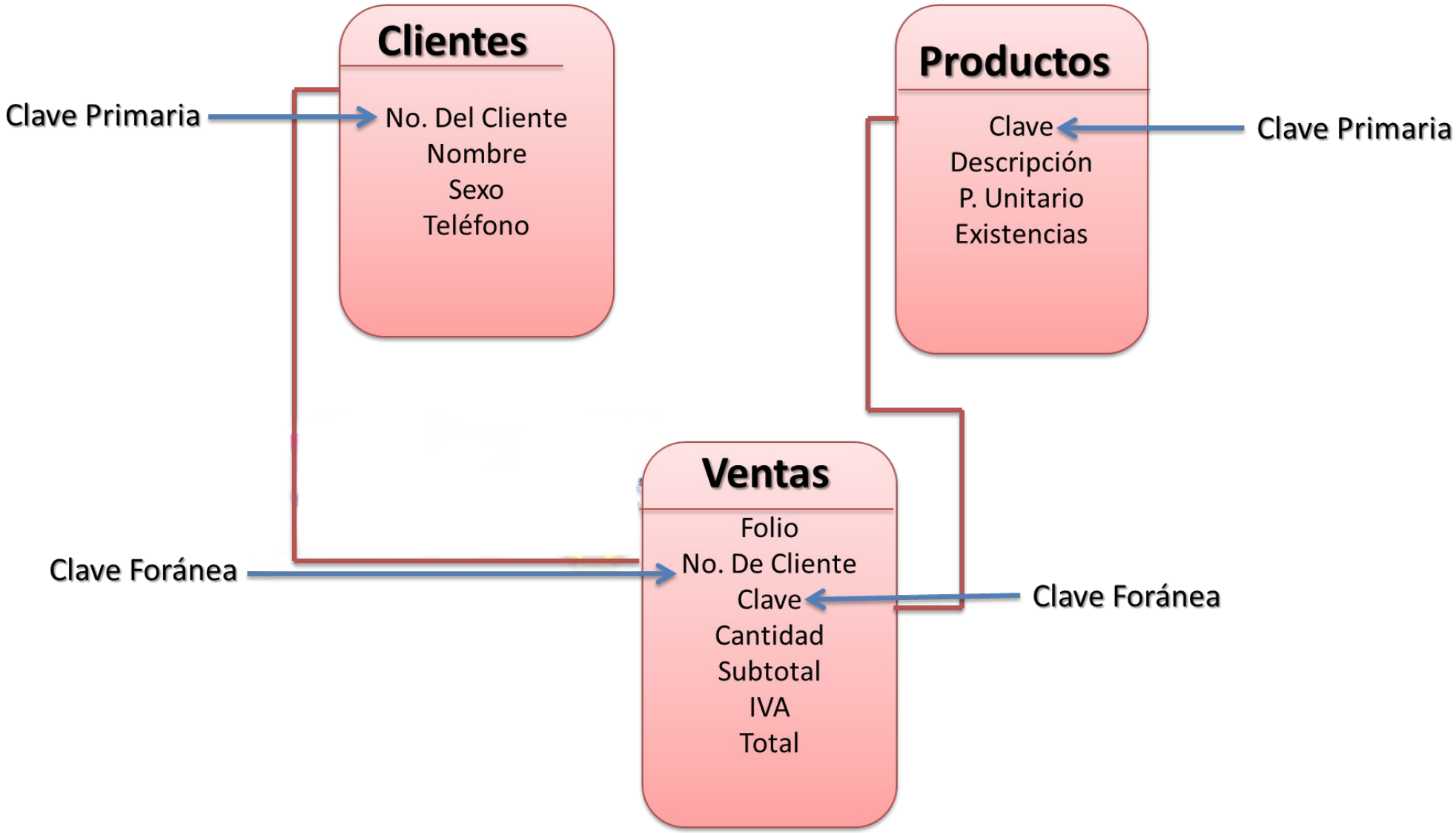
Diagrama de BDD

El campo primario de cada tabla se indica con un asterisco *

Se puede especificar o no el FK.



Modelo Relacional




HIVE
APACHE HBASE
mahout
FLUME
HUE
Microsoft SQL Server
ORACLE
ODBC
SQLite
PostgreSQL
MySQL
Zookeeper
hadoop
Pig
AVRO
Apache Solr
OOZIE
Impala

SQL

SQL es un lenguaje formal para comunicarse con una base de datos.

Permite articular de forma precisa qué información queremos obtener de una colección de tablas de una base de datos.

Con pequeñas diferencias, sirve como interfaz para diferentes proveedores de bases de datos.

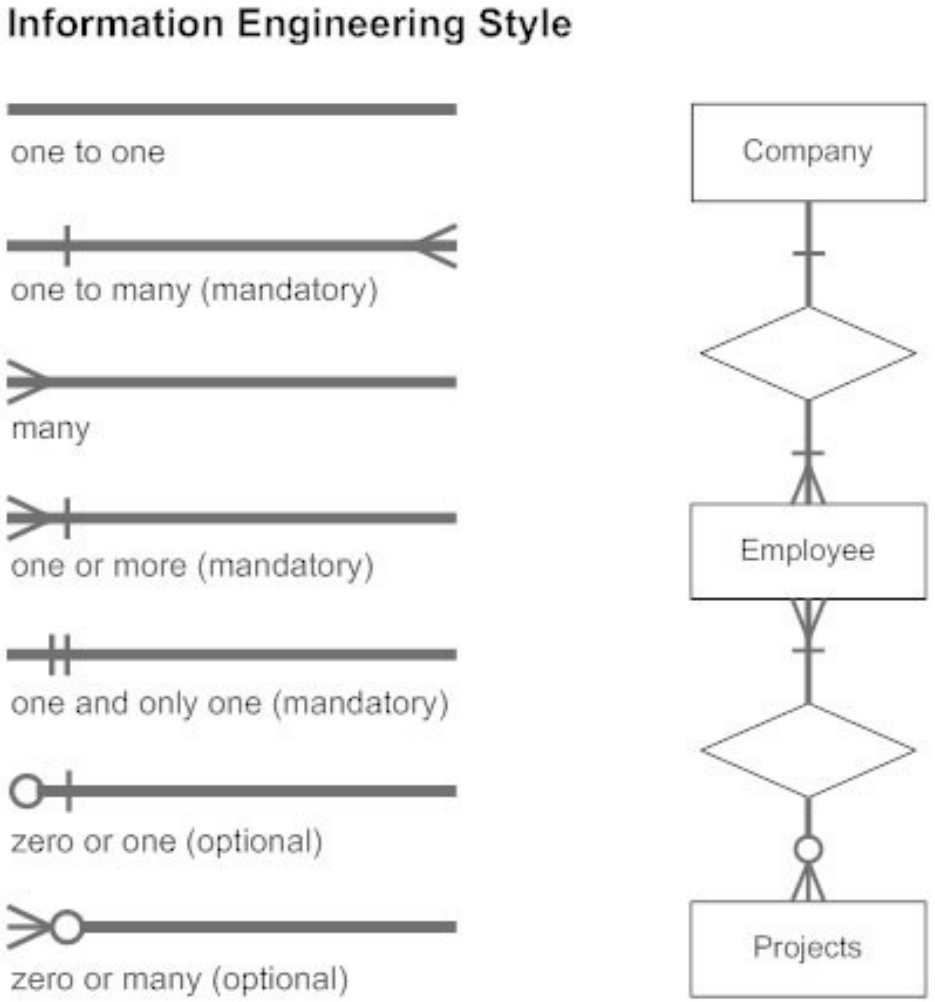
Bases de datos relacionales

Diagrama de Relación de Entidades

Se pinta la estructura de la base de datos como un diagrama que permite representar las entidades relevantes de un sistema de información así como sus interrelaciones y propiedades.

Las entidades se dibujan como puntos, poligonos, círculos u ovalos. Son el equivalente gramatical de nombres, tipo empleados, clientes, departamentos...Una entidad se define por sus atributos

Las relaciones entre las distintas entidades son el equivalente a los verbos que pueden realizar éstas, tipo comprar, ser miembro de un departamento, ... Las relaciones se definen en función del número de entidades que se asocian con ellas.



Más recursos

Sería recomendable que todos aquellos que no teneis conocimientos específicos en bases de datos, realizaseis (leyeseis) este tutorial con el objetivo de familiarizarse con los conceptos de este tema.

- Introducción a SQL 1
- Introducción a SQL 2