Introducción a SQL

El valor de los datos

El Data Scientist es el responsable de obtener valor de los datos

¿Qué son los datos?

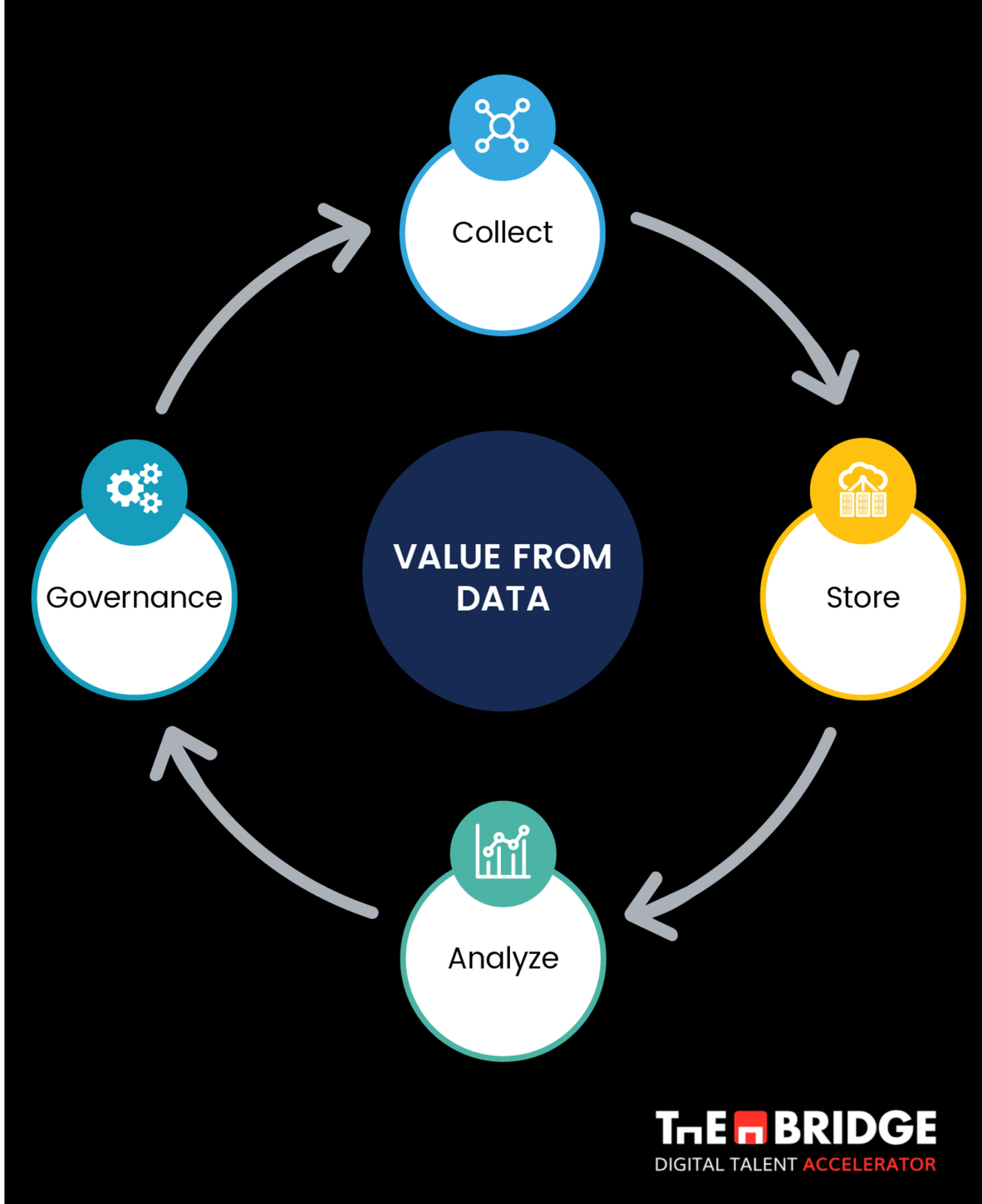
¿Dónde están los datos?

¿Cómo adquirir datos?

¿Cómo almacenar datos?

¿Cómo analizar datos?

¿Cómo visualizar datos?



¿Qué es un dato?

WIKIPEDIA
La enciclopedia libre

Portada
Portal de la comunidad
Actualidad
Cambios recientes
Páginas nuevas
Página aleatoria
Ayuda
Donaciones
Notificar un error

Dato

Para una antigua ciudad griega de Tracia, véase Dato (Tracia).

Véase también: Archivo informático

Un **dato** es una representación simbólica (numérica, alfabética, algorítmica, espacial, etc.) de un atributo o variable cuantitativa o cualitativa. Los datos describen hechos empíricos, sucesos y entidades. Es un valor o referente que recibe el computador por diferentes medios, los datos representan la información que el programador manipula en la construcción de una solución o en el desarrollo de un algoritmo.

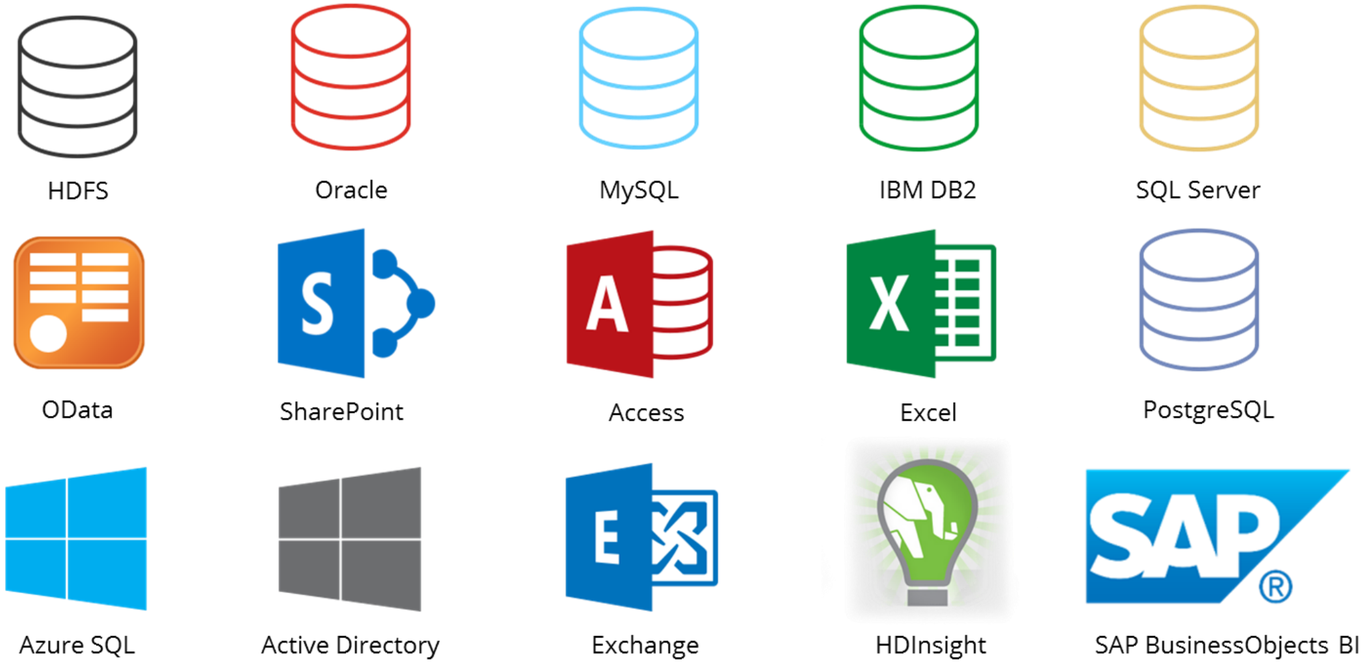
Hoy en día, un dato es cualquier elemento que se pueda digitalizar

Procesos ETL

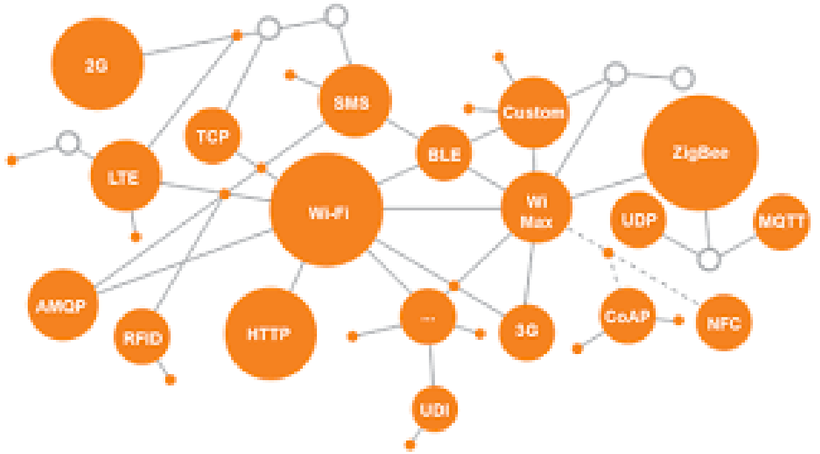
- **Extraer**: Obtener datos de diferentes fuentes
- **Transformar**: Filtrar, limpiar, homogeneizar y agrupar la información.
- **Cargar**: Organizar y actualizar los datos en bases de datos

Extracción


AVI
JPG
TXT
MP3
PDF
XLS
PNG
MOV
WAV
DOC
EPS
PPT
ZIP
CSS
RAR
HTML
EML
RAW
CSV
TTF
Aa


HDFS
Oracle
MySQL
IBM DB2
SQL Server
OData
SharePoint
Access
Excel
PostgreSQL
Azure SQL
Active Directory
Exchange
HDInsight
SAP BusinessObjects BI


You
Tube


3G
SMS
TCP
Custom
LTE
BLE
ZigBee
Wi-Fi
Wi Max
UDP
MQTT
AMQP
RFID
HTTP
...
3G
CoAP
NFC
UDI

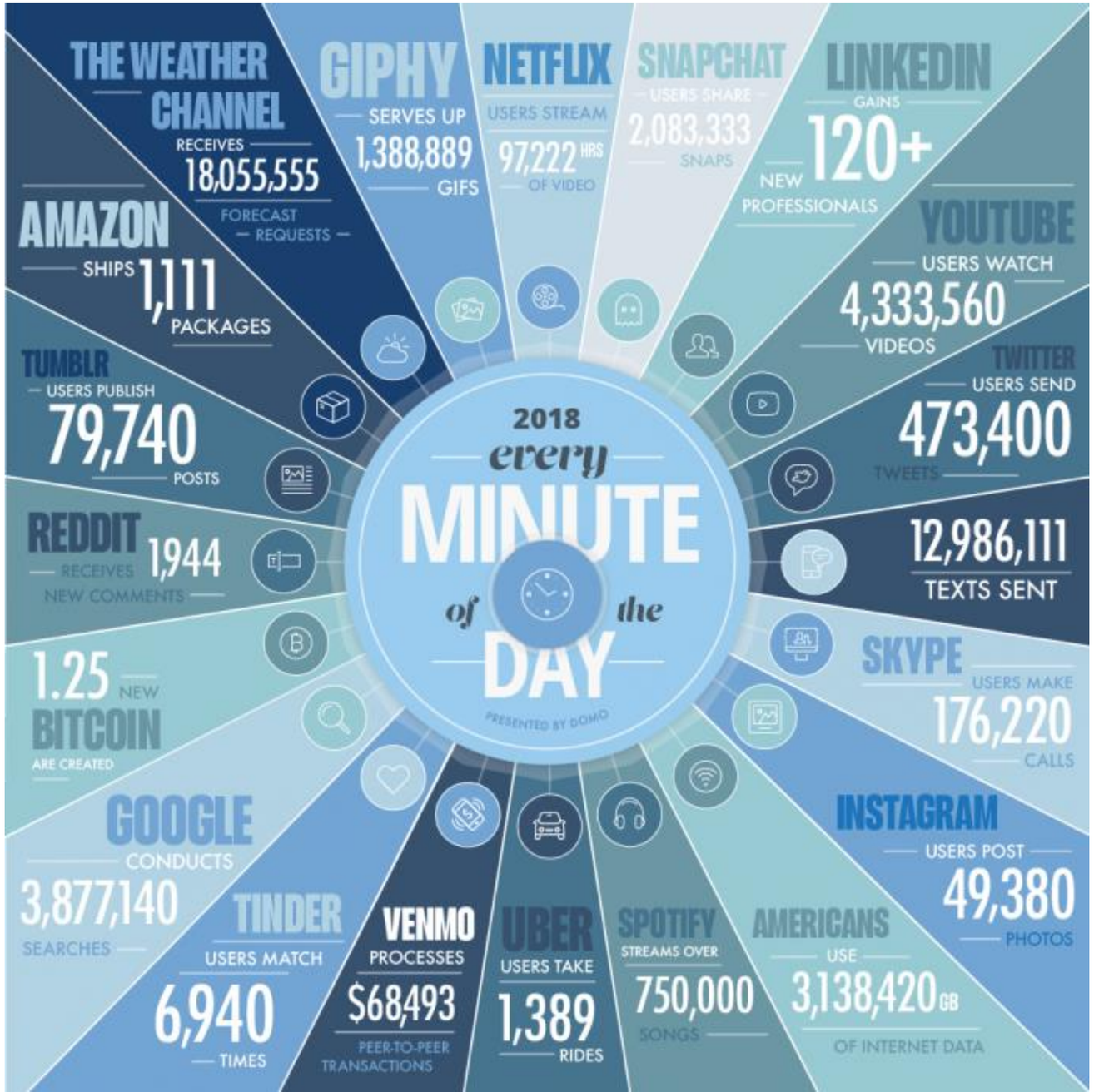
Extracción

Datos web

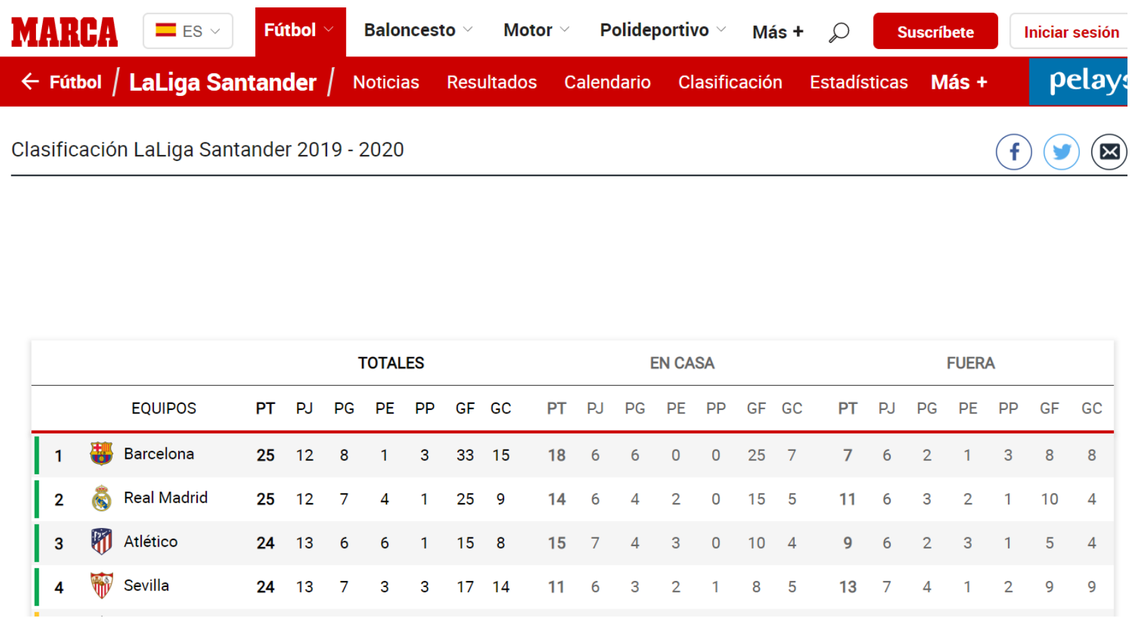
- Podemos acceder a datos web a través de:

- **Web Scraping** es el proceso de obtener datos de una web utilizando programas automáticos

- **APIs** (Application Programming Interfaces) son un conjunto de procedimientos que permiten el acceso a los datos de otra aplicación

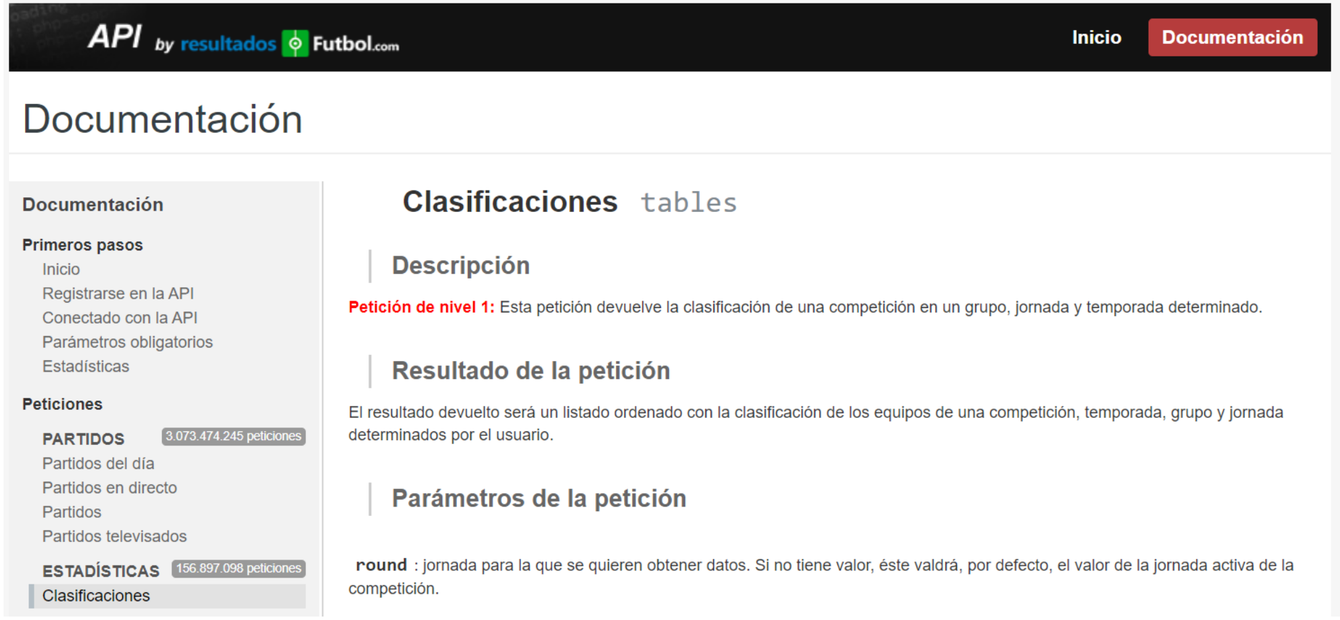


Extracción



		TOTALES							EN CASA							FUERA						
	EQUIPOS	PT	PJ	PG	PE	PP	GF	GC	PT	PJ	PG	PE	PP	GF	GC	PT	PJ	PG	PE	PP	GF	GC
1	Barcelona	25	12	8	1	3	33	15	18	6	6	0	0	25	7	7	6	2	1	3	8	8
2	Real Madrid	25	12	7	4	1	25	9	14	6	4	2	0	15	5	11	6	3	2	1	10	4
3	Atlético	24	13	6	6	1	15	8	15	7	4	3	0	10	4	9	6	2	3	1	5	4
4	Sevilla	24	13	7	3	3	17	14	11	6	3	2	1	8	5	13	7	4	1	2	9	9

Web Scraping

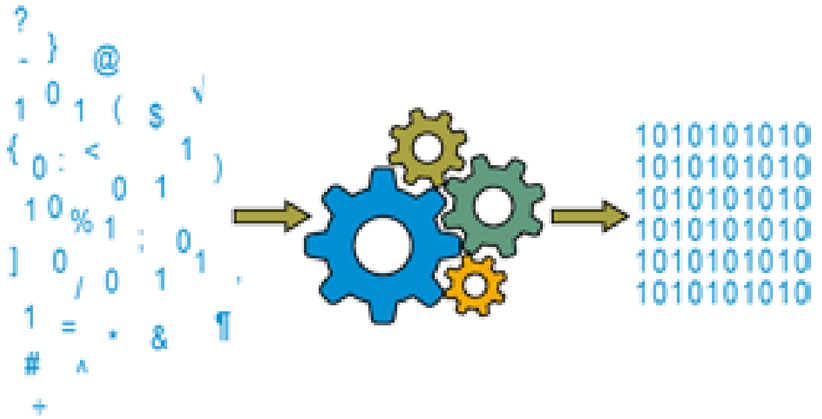


API

- Si una web comparte sus datos a través de APIs, es preferible usarlas antes que realizar web scraping

Transformación

- Estructurar datos
- Aplicar funciones o reglas de negocio
- Seleccionar o filtrar (filas, columnas)
- Traducir códigos: {“Alto”, “Medio”, “Bajo”} = {3, 2, 1}
- Anonimizaciones
- Nuevos valores: TotalVenta = Precio x Cantidad
- Unir datos de múltiples fuentes
- Agrupar filas
- Etc.



Carga

¿Qué es una base de datos?

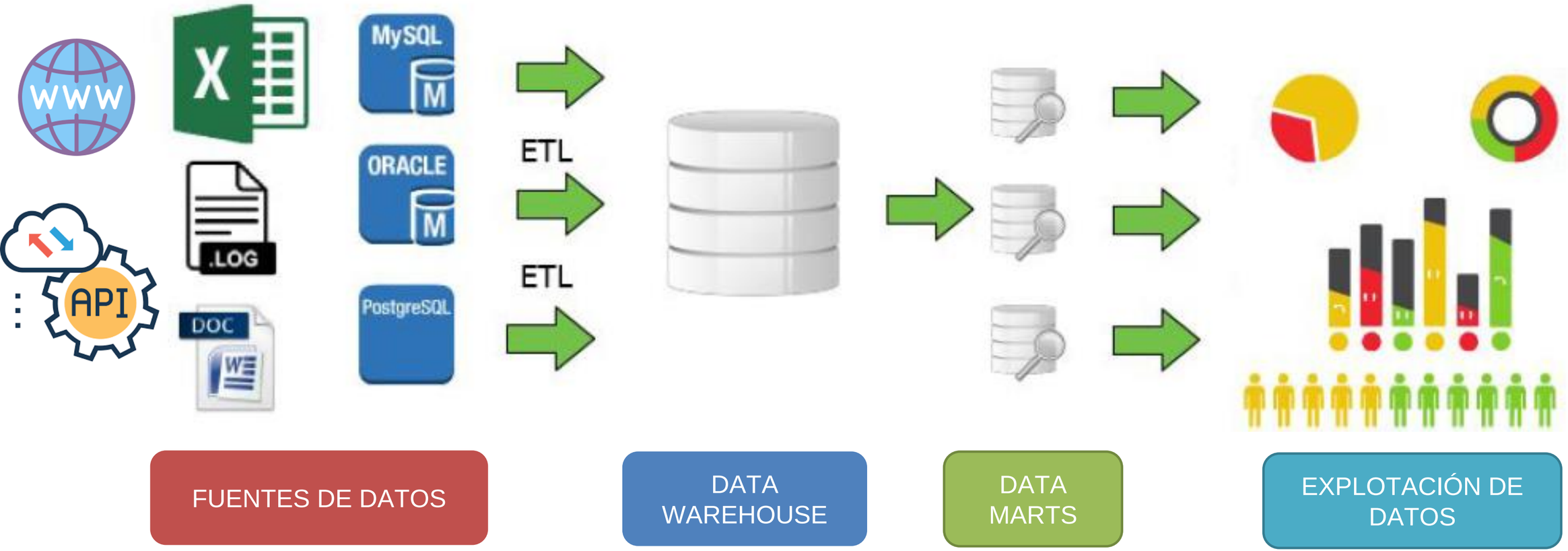
- Almacenes que nos permiten guardar datos de forma organizada
- Los sistemas de gestión de bases de datos (SGBD) son programas desarrollados explícitamente para gestionar bases de datos (MySQL, Oracle, SQL Server…)
- Para poder acceder a la información de una base de datos se emplean lenguajes de consulta (SQL es el más utilizado)
- Tipos de bases de datos:
  - Relacionales
  - Multidimensionales
  - NoSQL

Carga

¿Qué no es una base de datos?

- Excel **NO** es una base de datos, es una herramienta de hoja de cálculo

Arquitectura Data Warehouse



- El **data warehouse** es un almacén de información que integra los datos de toda las fuentes de información
- Un **data mart** es una base de datos departamental que se nutre del data warehouse

Modelo de datos

Modelo Relacional

- Las **bases de datos relacionales** son las más tradicionales y comunes
- Ordena los datos en **tablas**, las cuales se relacionan a partir de uno o más campos (columnas).
- El lenguaje para consultar estas bases de datos es **SQL** (Structured Query Language)
- SGBDs de SQL más conocidos:

Microsoft SQL Server

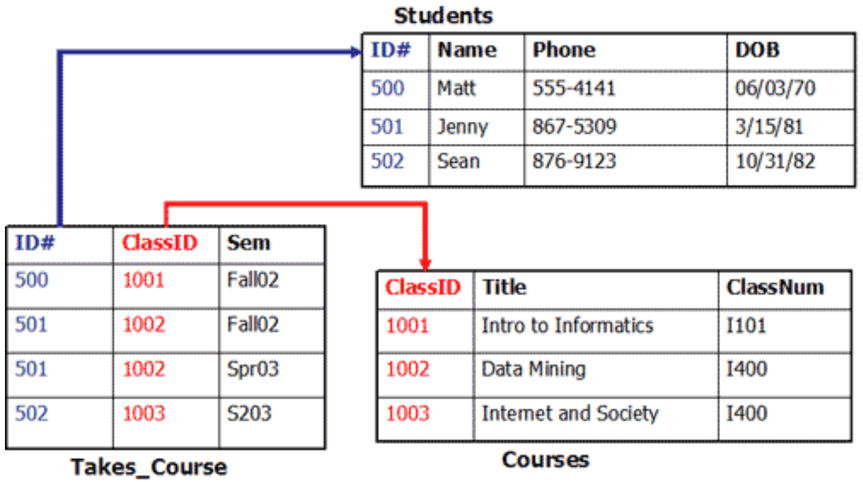
MySQL

ORACLE DATABASE

PostgreSQL

teradata.

IBM DB2



Students

ID#	Name	Phone	DOB
500	Matt	555-4141	06/03/70
501	Jenny	867-5309	3/15/81
502	Sean	876-9123	10/31/82

Takes_Course

ID#	ClassID	Sem
500	1001	Fall02
501	1002	Fall02
501	1002	Spr03
502	1003	S203

Courses

ClassID	Title	ClassNum
1001	Intro to Informatics	I101
1002	Data Mining	I400
1003	Internet and Society	I400

Modelo de datos

SQL

Tabla ALUMNOS

NOMBRE	APELLIDO	EDAD
Rafael	López	50
Laura	Sanz	33
Fernando	Sánchez	22
Ana	Martín	21
Luisa	Lorenzo	55
Carlos	Fernández	20

query→

```
SELECT NOMBRE, EDAD
FROM ALUMNOS
WHERE EDAD>30
```

NOMBRE	EDAD
Rafael	50
Laura	33
Luisa	55

Modelo de datos

Ejemplo de modelo relacional

- Queremos diseñar una base de datos para una plataforma online de cursos de idiomas

- Primer diseño:

Cliente	Idioma	Nivel	Suscripción	Precio	Descuento	Precio Final
Pedro	Inglés	Intermedio	Mensual	70	0	70
Pedro	Chino	Principiante	Mensual	90	0	90
Ana	Francés	Avanzado	Anual	80	25	60
Luis	Inglés	Intermedio	Trimestral	70	10	63

- ¿Problemas?

TABLAS DE DIMENSIÓN

CLIENTE

cliente_id	cliente
1	Pedro
2	Ana
3	Luis

PROGRAMA

programa_id	idioma	nivel	precio
1	Alemán	Principiante	70
2	Chino	Principiante	90
3	Francés	Avanzado	80
4	Inglés	intermedio	70

SUSCRIPCION

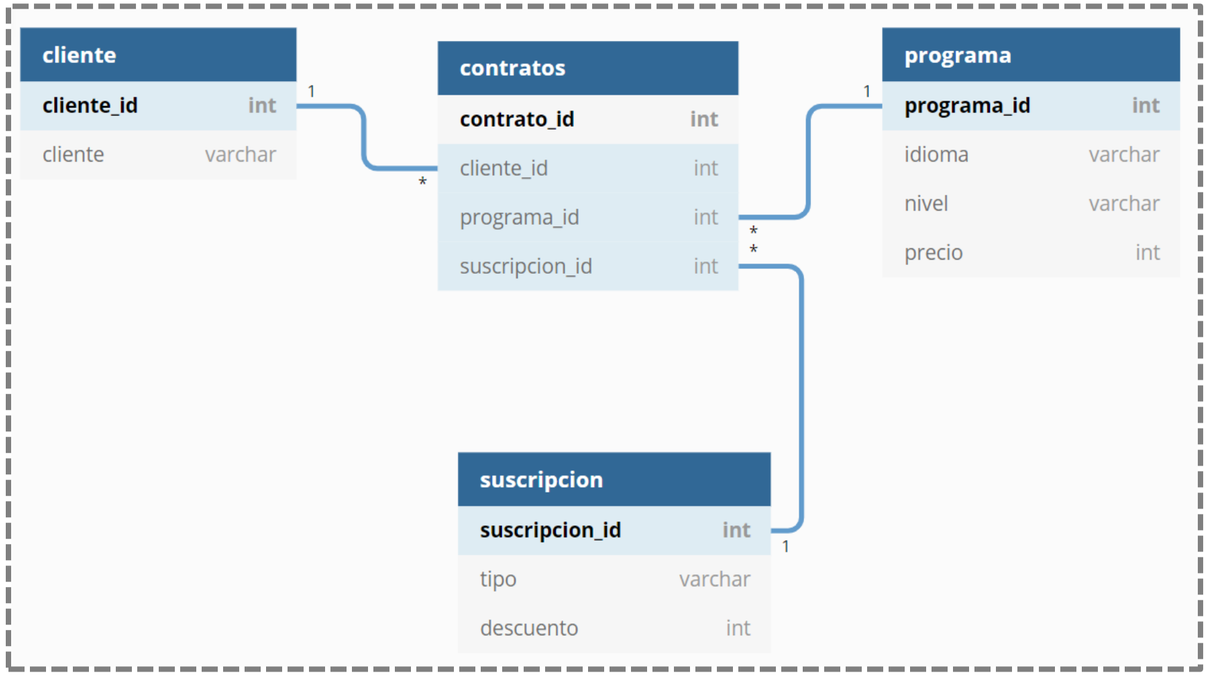
suscripcion_id	tipo	descuento
1	Mensual	0
2	Trimestral	10
3	Anual	25

TABLAS DE HECHOS

CONTRATOS

Contrato_id	cliente_id	programa_id	suscripción_id
1	1	4	1
2	1	2	1
3	2	3	3
4	3	4	2

Cliente	Idioma	Nivel	Suscripción	Precio	Descuento	Precio Final
Pedro	Inglés	Intermedio	Mensual	70	0	70
Pedro	Chino	Principiante	Mensual	90	0	90
Ana	Francés	Avanzado	Anual	80	25	60
Luis	Inglés	Intermedio	Trimestral	70	10	63



TABLAS DE DIMENSIÓN

CLIENTE

cliente_id	cliente
1	Pedro
2	Ana
3	Luis

PROGRAMA

programa_id	idioma	nivel	precio
1	Alemán	Principiante	70
2	Chino	Principiante	90
3	Francés	Avanzado	80
4	Inglés	intermedio	70

SUSCRIPCION

suscripcion_id	tipo	descuento
1	Mensual	0
2	Trimestral	10
3	Anual	25

TABLAS DE HECHOS

CONTRATOS

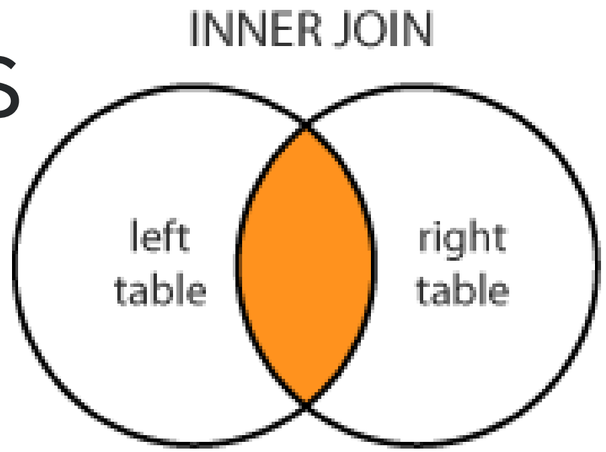
Contrato_id	cliente_id	programa_id	suscripción_id
1	1	4	1
2	1	2	1
3	2	3	3
4	3	4	2

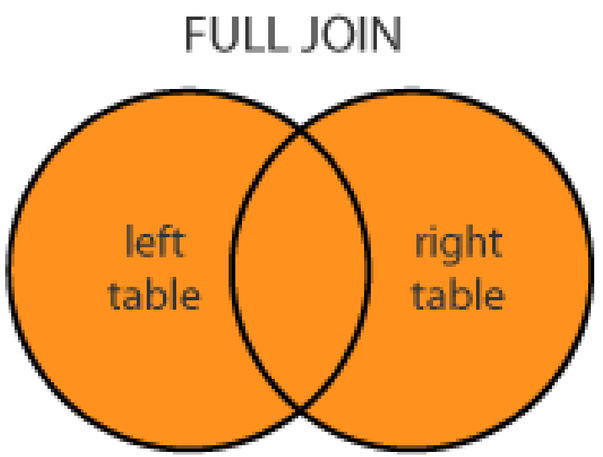
Cliente	Idioma	Nivel	Suscripción	Precio	Descuento	Precio Final
Pedro	Inglés	Intermedio	Mensual	70	0	70
Pedro	Chino	Principiante	Mensual	90	0	90
Ana	Francés	Avanzado	Anual	80	25	60
Luis	Inglés	Intermedio	Trimestral	70	10	63

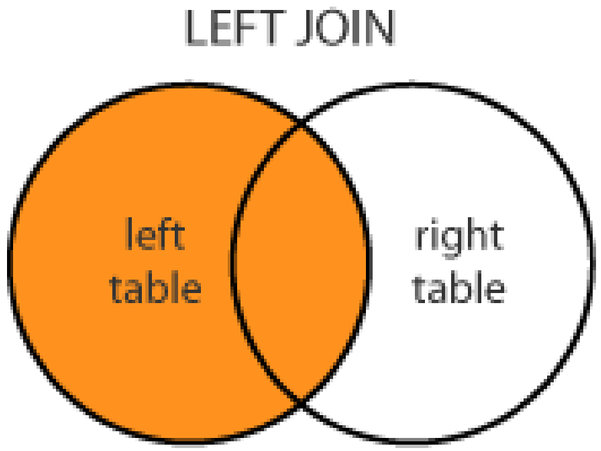
```
SELECT CLI.CLIENTE,
       PRO.IDIOMA,
       PRO.NIVEL,
       SUS.TIPO AS SUSCRIPCION,
       PRO.PRECIO,
       SUS.DESCUENTO,
       PRO.PRECIO * (1-SUS.DESCUENTO/100)
            AS PRECIO_FINAL

FROM CONTRATOS CON

JOIN CLIENTE CLI

ON CON.CLIENTE_ID = CLI.CLIENTE_ID

JOIN PROGRAMA PRO
ON CON.PROGRAMA_ID = PRO.PROGRAMA_ID
JOIN SUSCRIPCION SUS
ON CON.SUSCRIPCION_ID = SUS.SUSCRIPCION_ID
```

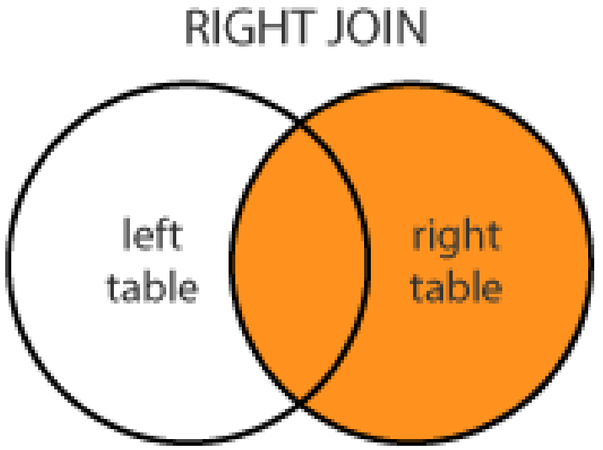
Modelo de datos

SQL JOINS



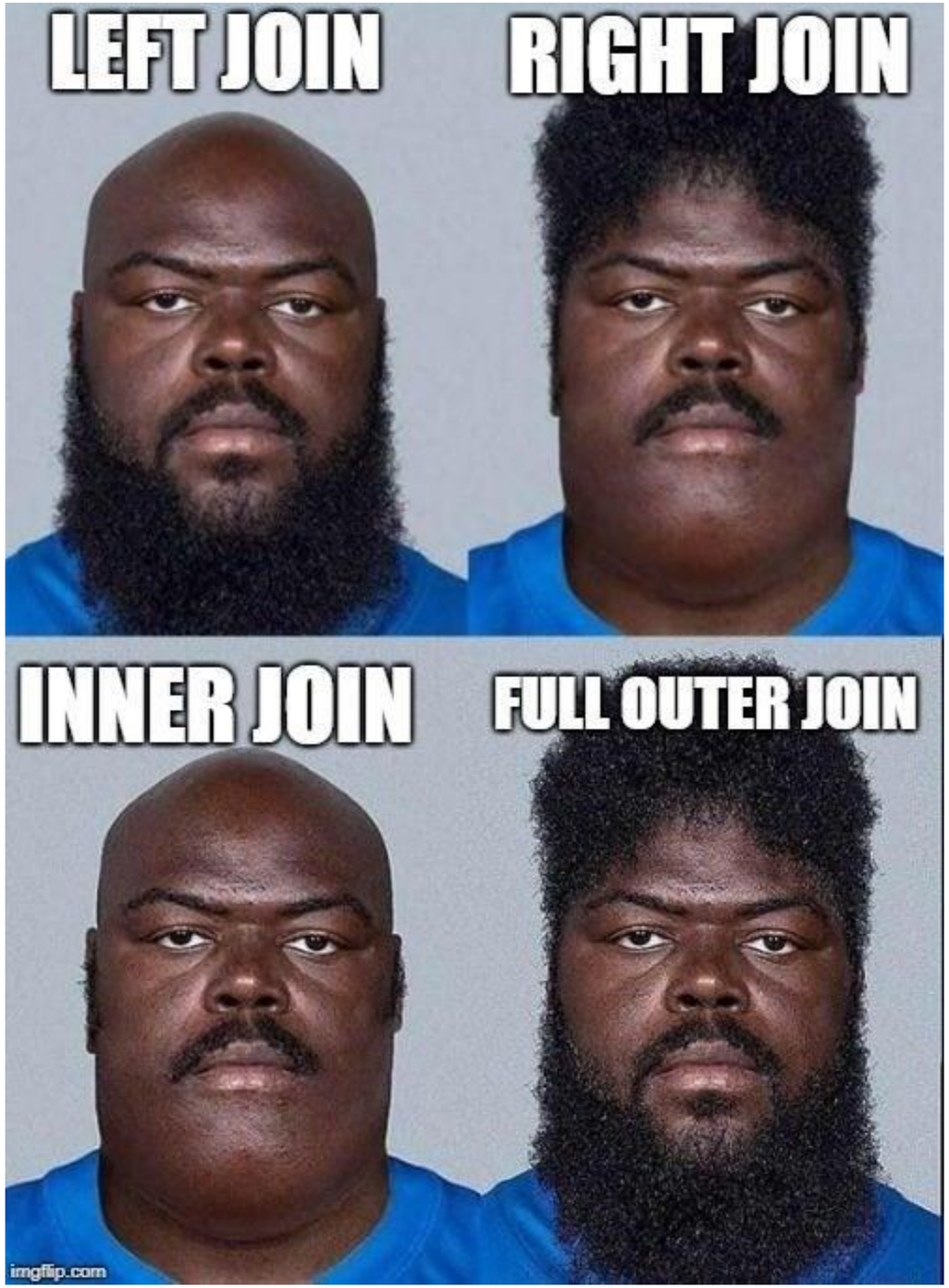






Modelo de datos

SQL JOINS



Modelo de datos

Tabla EDADES

Jugador	Edad
Ramos	33
Benzema	30
Messi	32
Aspas	31

Tabla SALARIOS

Jugador	Salario
Suárez	150
Ramos	120
Piqué	100
Messi	500

```
SELECT EDA.Jugador, EDA.Edad, SAL.Salario
FROM EDADES EDA
JOIN SALARIOS SAL
ON EDA.JUGADOR = SAL.JUGADOR
```

Jugador	Edad	Salario
Ramos	33	120
Messi	32	500

Modelo de datos

Tabla EDADES

Jugador	Edad
Ramos	33
Benzema	30
Messi	32
Aspas	31

Tabla SALARIOS

Jugador	Salario
Suárez	150
Ramos	120
Piqué	100
Messi	500

```
SELECT EDA.Jugador, EDA.Edad, SAL.Salario
FROM EDADES EDA
LEFT JOIN SALARIOS SAL
ON EDA.JUGADOR = SAL.JUGADOR
```

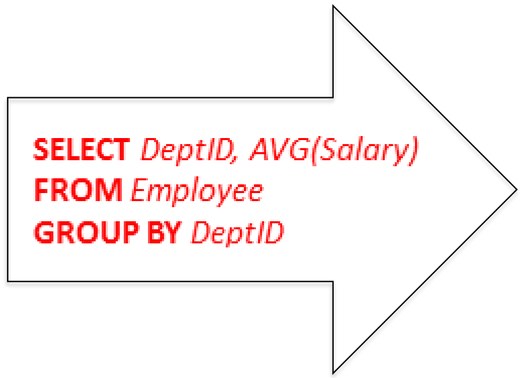
Jugador	Edad	Salario
Ramos	33	150
Benzema	30	NULL
Messi	32	500
Aspas	31	NULL

Modelo de datos

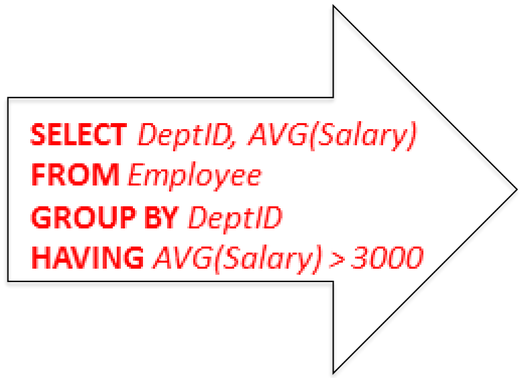
Agrupaciones SQL

Employee

EmployeeID	Ename	DeptID	Salary
1001	John	2	4000
1002	Anna	1	3500
1003	James	1	2500
1004	David	2	5000
1005	Mark	2	3000
1006	Steve	3	4500
1007	Alice	3	3500



DeptID	AVG(Salary)
1	3000.00
2	4000.00
3	4250.00



DeptID	AVG(Salary)
2	4000.00
3	4250.00

Funciones de agregación:

`SUM(), COUNT(), AVG(), MIN(), MAX()`

Modelo de datos

SUBQUERIES

Employee

EmployeeID	Ename	DeptID	Salary
1001	John	2	4000
1002	Anna	1	3500
1003	James	1	2500
1004	David	2	5000
1005	Mark	2	3000
1006	Steve	3	4500
1007	Alice	3	3500

Position

IdEmployee	Pos
1004	Director
1029	President
1022	Director
1033	Vicepresident
1001	Director

Quiero filtrar aquellos empleados con la posición de director

```
SELECT * FROM Employee
WHERE EmployeeID IN (SELECT IdEmployee
                     FROM Position
                     WHERE Pos = “Director”)
```

(1004,1022,1001)

Gracias

Rafa Zambrano

rafael@thebridgeschool.es

THE BRIDGE