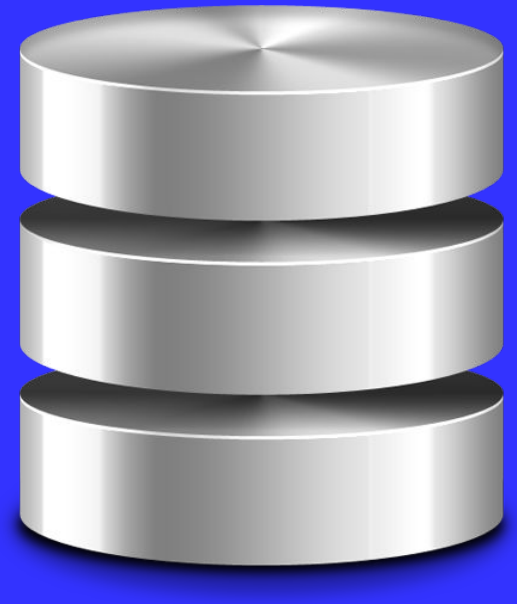
SQL para Data Science

O que é um banco de dados?

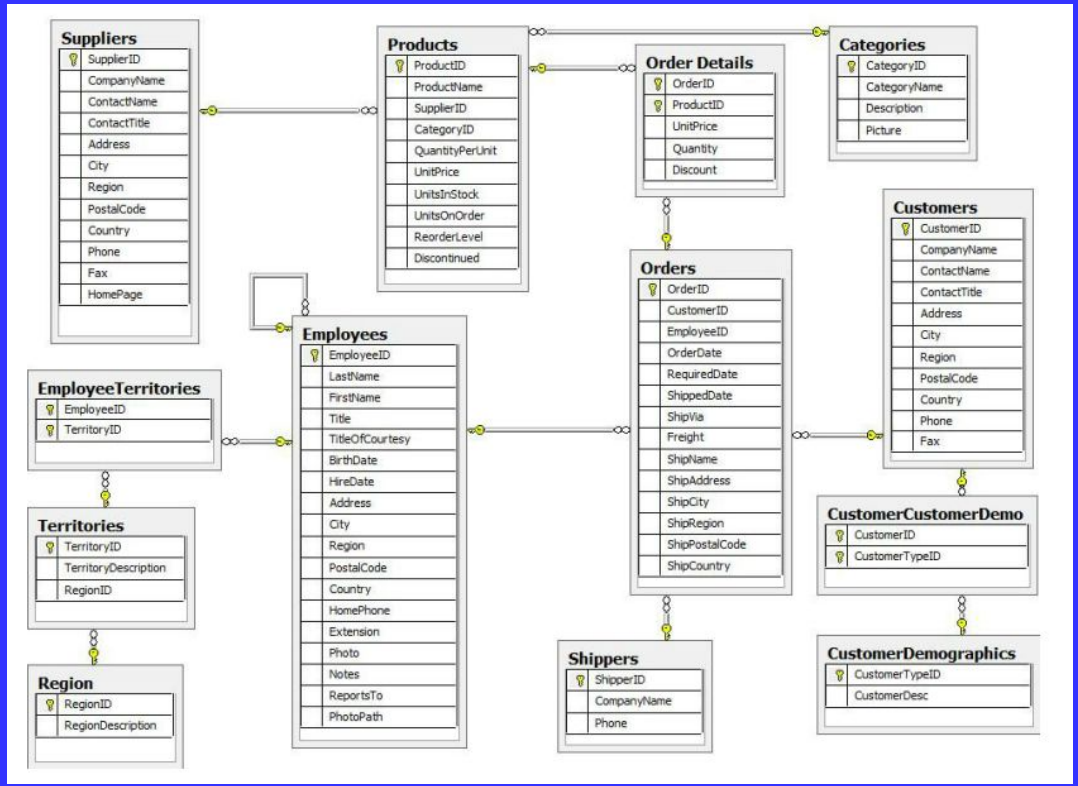
- Um conjunto de arquivos de dados usados para armazenar informações.
- Um banco de dados é gerenciado por um SGBD (Sistema Gerenciador de Banco de dados)
- Existem diversos SGBDs, como: Microsoft SQL Server, Oracle, MySQL, PostgreSQL etc.
- Para trabalhar com um banco de dados usamos uma linguagem chamada SQL.
    - Essa linguagem é usada para inserir, atualizar, deletar e recuperar dados em um banco de dados.



Arquitetura de Software

Esquema Relacional

- Os dados são armazenados em tabelas.
- As tabelas são relacionadas usando chaves.



Trabalhando com um banco de dados

- Linguagem SQL significa Structured Query language.
- Usada para recuperar e manipular dados no banco de dados.

Trabalhando com um banco de dados

- Linguagem de Manipulação de dados ou DML (Data Manipulation Language): É um subconjunto da linguagem SQL, utilizada para selecionar, inserir, atualizar e apagar dados.
- Linguagem de Definição de Dados ou DDL (Data Definition Language): A DDL permite ao usuário a manipulação de tabelas e elementos associados como chaves primárias, chaves estrangeiras, índices etc.
- Linguagem de Controle de Dados ou DCL (Data Control Language): A DCL controla os aspectos destinados a autorização de dados e licenças de usuários para manipulação de dados.
- Linguagem de Consultas de Dados ou DQL (Data Query Language): é a linguagem usada para recuperar dados em um banco de dados.

Funções de agregações

- **`SUM()`**: Retorna a soma total da coluna agrupada.
- **`COUNT()`**: Retorna o número de linhas do grupo.
- **`AVG()`**: Retorna o valor médio do grupo.
- **`MIN()`**: Retorna o valor mínimo do grupo.
- **`MAX()`**: Retorna o valor máximo do grupo.

Entendendo agregações

- Group By

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

```
SELECT DeptID, AVG(Salary)
FROM Employee
GROUP BY DeptID;
```

GROUP BY Employee Table using DeptID

DeptID	AVG(Salary)
1	3000.00
2	4000.00
3	4250.00

Cláusula Having

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

SELECT *DeptID, AVG(Salary)*
FROM *Employee*
GROUP BY *DeptID;*

GROUP BY Employee Table using DeptID

DeptID	AVG(Salary)
1	3000.00
2	4000.00
3	4250.00

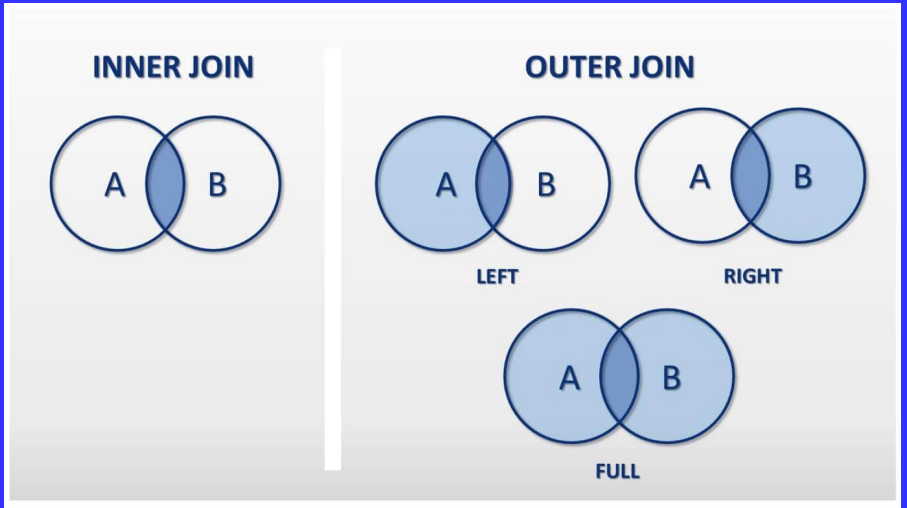
SELECT *DeptID, AVG(Salary)*
FROM *Employee*
GROUP BY *DeptID*
HAVING *AVG(Salary) > 3000;*

HAVING

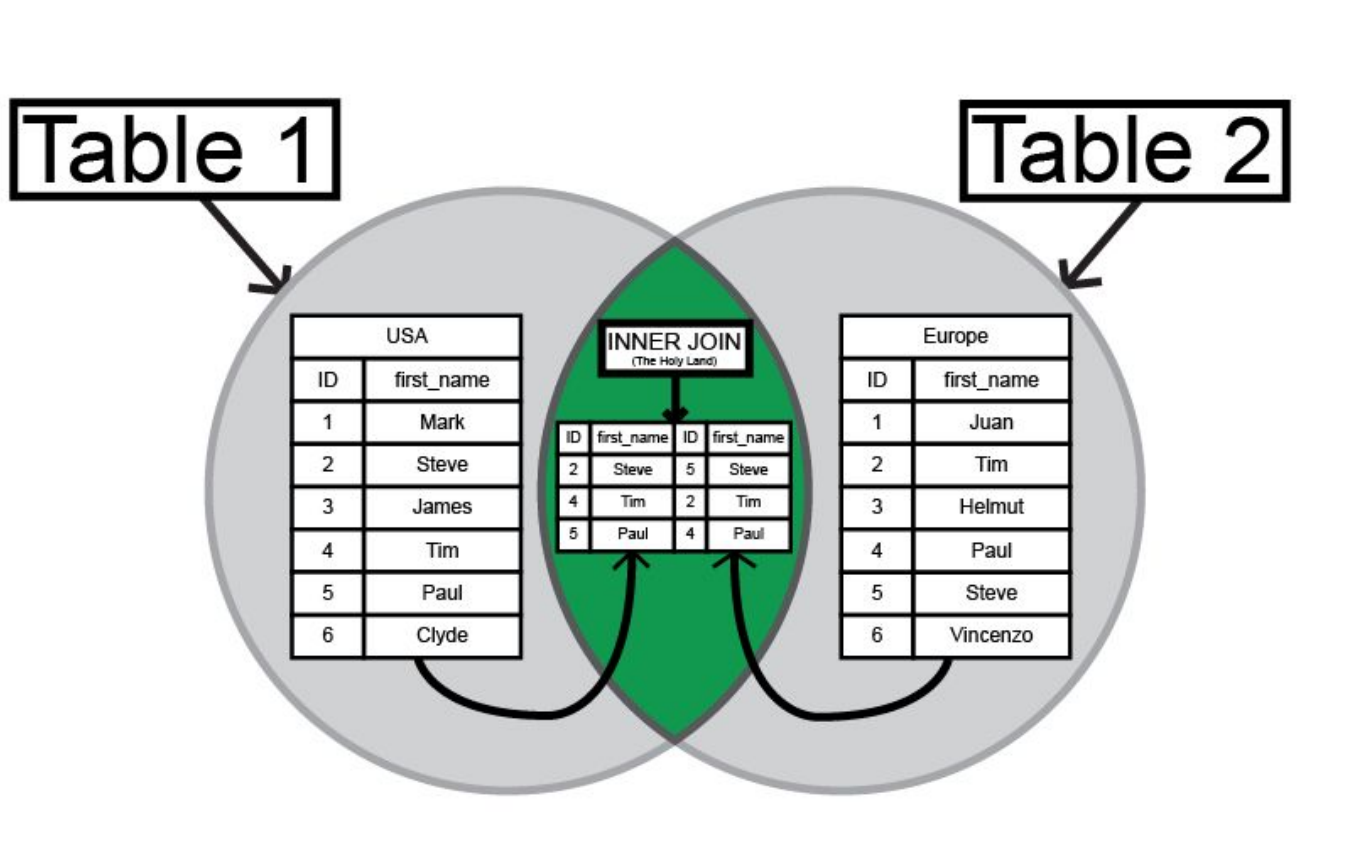
DeptID	AVG(Salary)
2	4000.00
3	4250.00

Trabalhando com Junções

- Os tipos de Joins são usados para recuperar dados entre várias tabelas
- Inner Join : Permite a recuperação de dados entre várias tabelas na qual todas as ligações estão presentes em todas as tabelas em questão.
- Left join: Permite a recuperação de dados onde existe os dados ao menos na tabela a esquerda.
- Right join: Permite a recuperação de dados onde existe os dados ao menos na tabela a direita.
- Full Join: Permite a recuperação dos dados de todas as tabelas.



Trabalhando com Junções



Trabalhando com Funções

- RANDOM()
  - Retorna um número pseudo-aleatório.
- MIN(), MAX(), SUM(), COUNT(), AVG()
  - Retornam o valor mínimo, máximo, a soma, a quantidade e a média respectivamente.
- PRAGMA table_info()
  - Retorna a estrutura física de uma tabela (DDL)
- CAST()
  - Retorna valores convertidos.
- UPPER()
  - Retorna caracteres em maiúsculo.
- LOWER()
  - Retorna caracteres em minúsculo.
- LENGTH()
  - Retorna o tamanho de uma variável (quantidade de caracteres)

SubQueries

Subconsultas são consultas aninhadas ou embutidas em outras consultas.

Uma subconsulta é usada para retornar dados que serão usados na consulta principal como uma condição para restringir ainda mais os dados a serem recuperados.

As subconsultas podem ser usadas com as instruções SELECT, INSERT, UPDATE e DELETE junto com os operadores como =, <,>,> =, <=, IN, BETWEEN, etc.

Trabalhando com Python

- Análises de dados mais ricas.
- Python + Banco de dados relacional.
- Integração do Pandas com SQL.