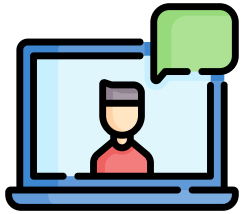
Dara Warehouses

Definição da tecnologia

- Armazém de dados
- Coleção de dados orientada a assunto
- Características:
  - Base SQL
  - Pré-processamento
  - Visualização
  - ML/AI
- Otimizada
- Diversas tecnologias
- Data Mart
- Data Lake

Introdução

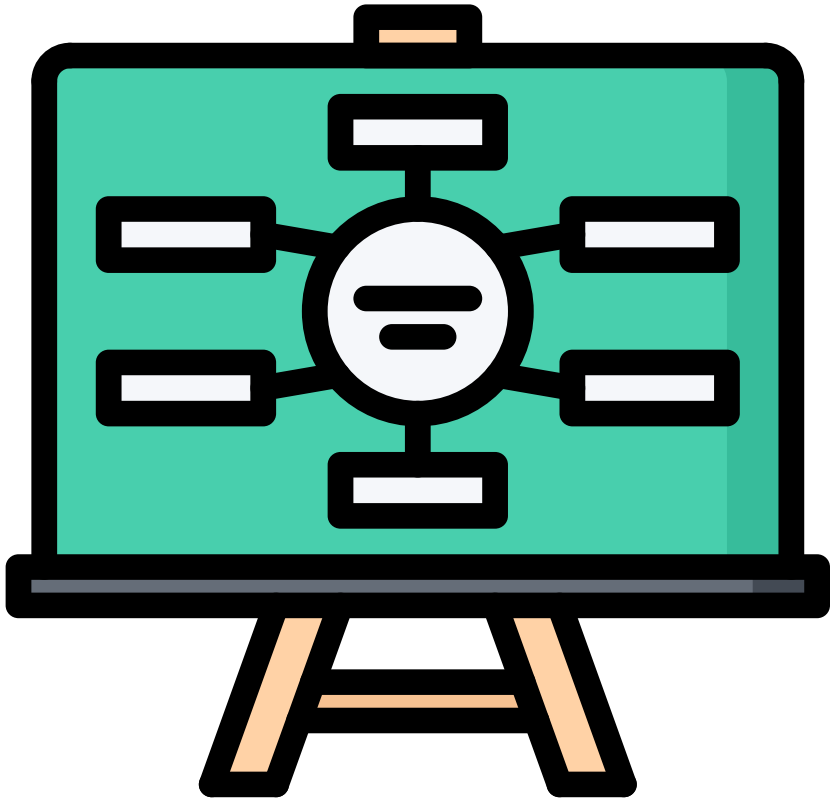


- Base de dados foi encontrada no Kaggle
- Base de dados sobre tweets da eleição dos EUA.
- Base de Dados no formato de arquivo csv
- O porquê essa base de dados foi escolhida

Schema de um data warehouse

Conceitos

- Fact tables que são tabela principal que reúne os atributos das tabelas secundárias mais os seus próprios atributos principais (exemplo: valor total da venda);
- Dimension tables, tabela secundária de características sobre a tabela principal (exemplo: horário que a venda foi realizada).



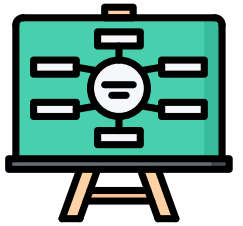
Modelando o Banco

- Análise do CSV
- Modelagem SQL
- Script Python para tratamento dos dados
  - ETL(Extract Transform Load)
  - Pandas
  - SQLAlchemy

pandas

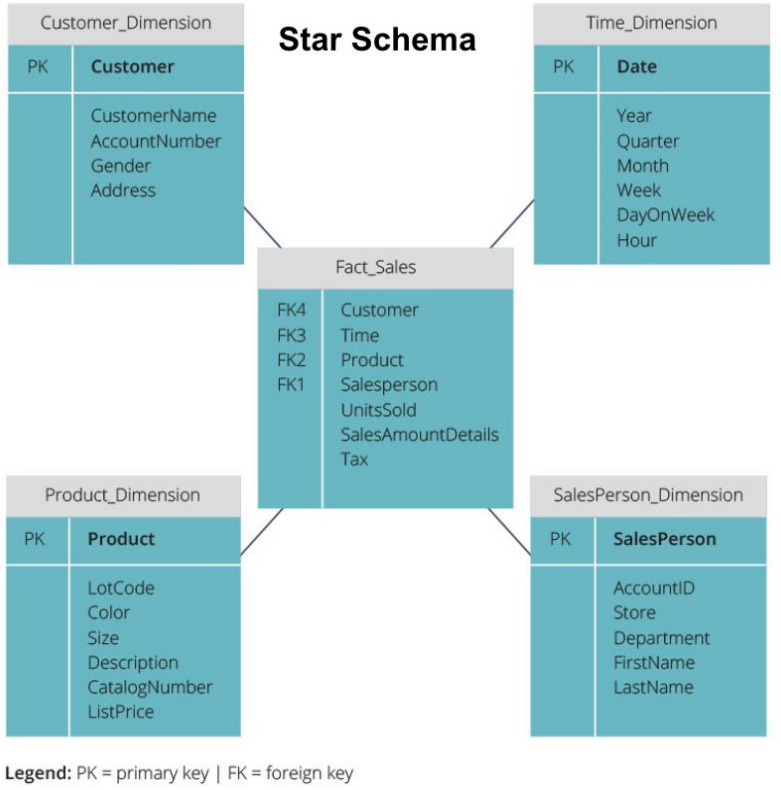
SQLAlchemy

Escolha de um Schema



Star Schema

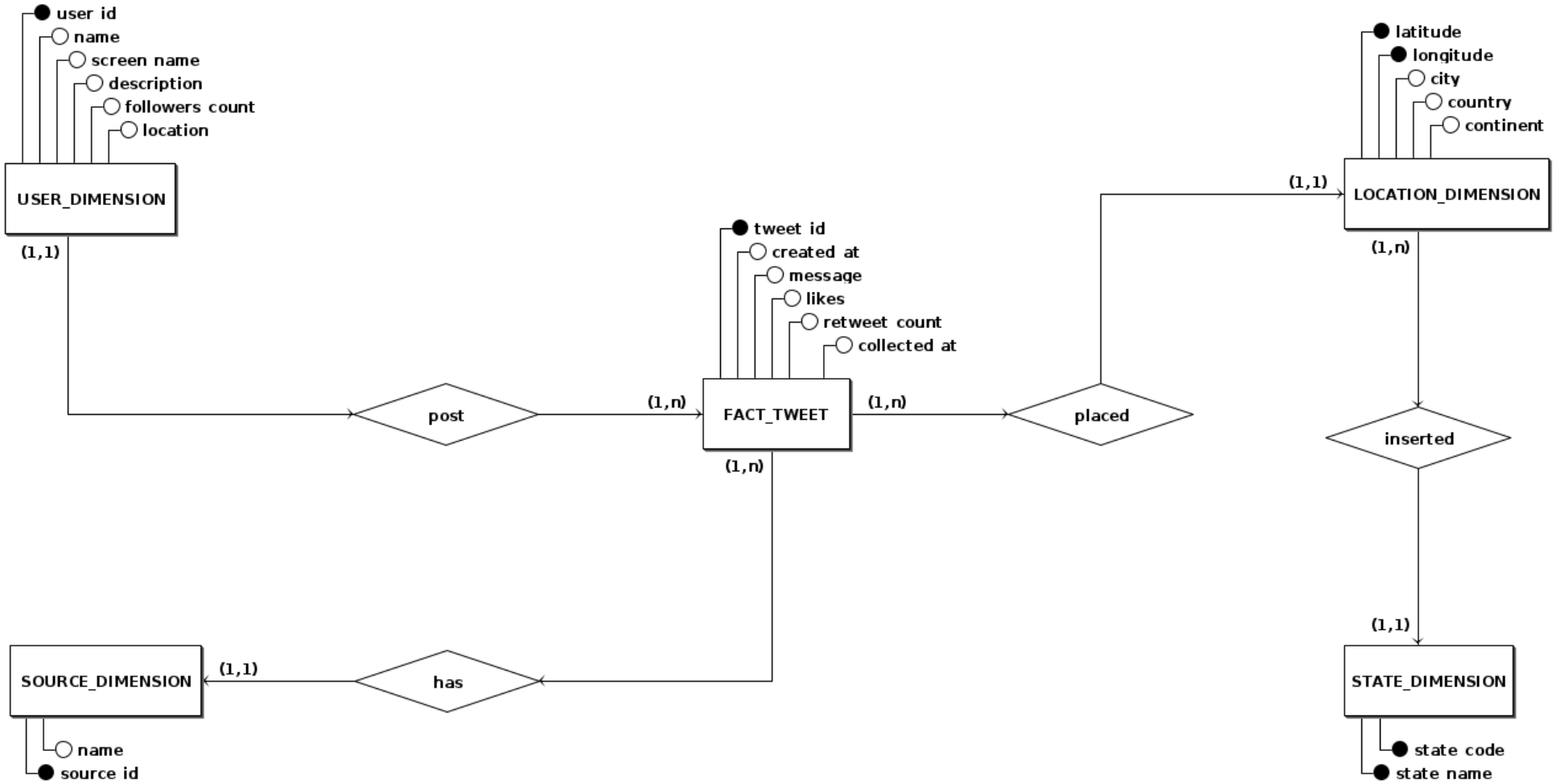
- Dividir fact table em dimension tables desnormalizadas.
- Maior velocidade de consultas por realizar apenas uma junção das dimension tables com a fact table.
- Escolha mais adequada para o contexto



Documentação da Base de Dados

Kaggle

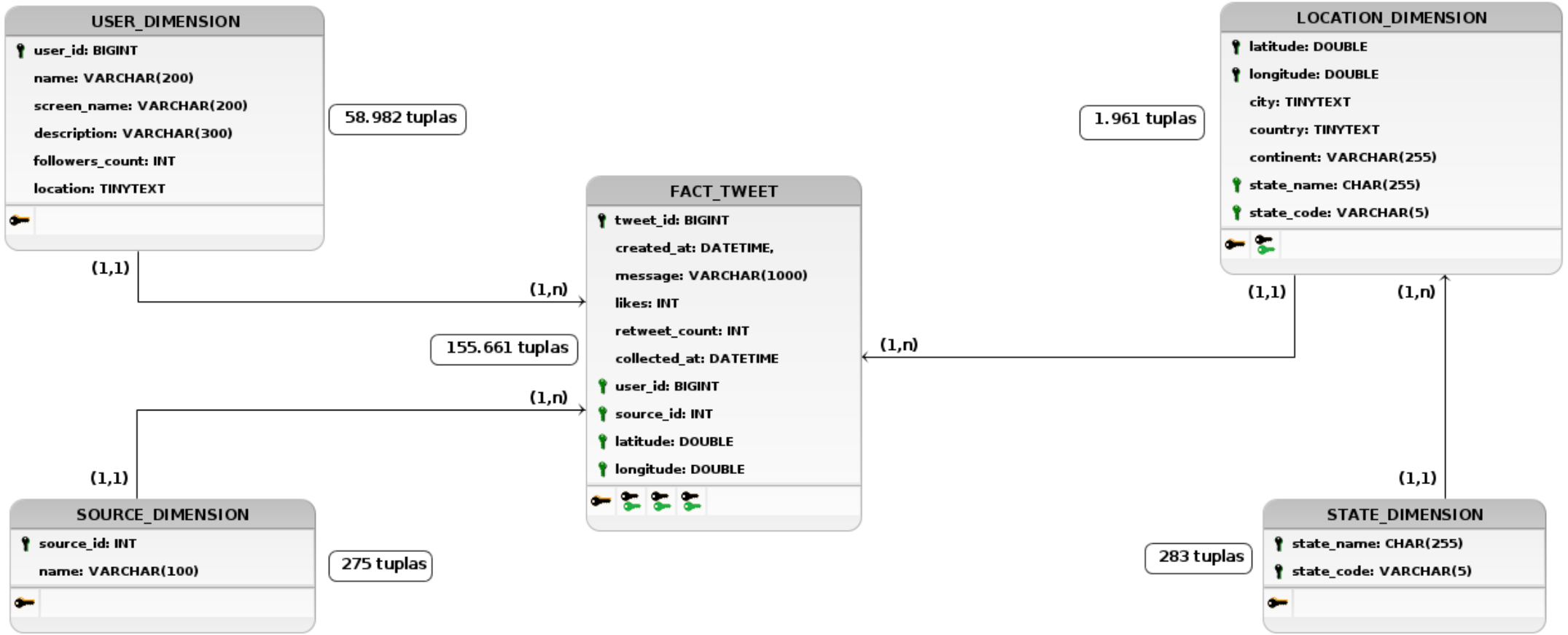
Dump



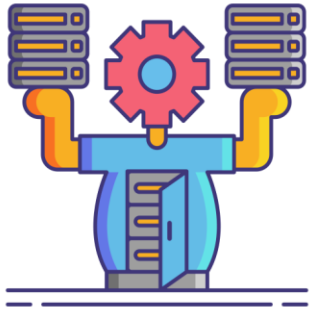
Documentação da Base de Dados

Kaggle

Dump

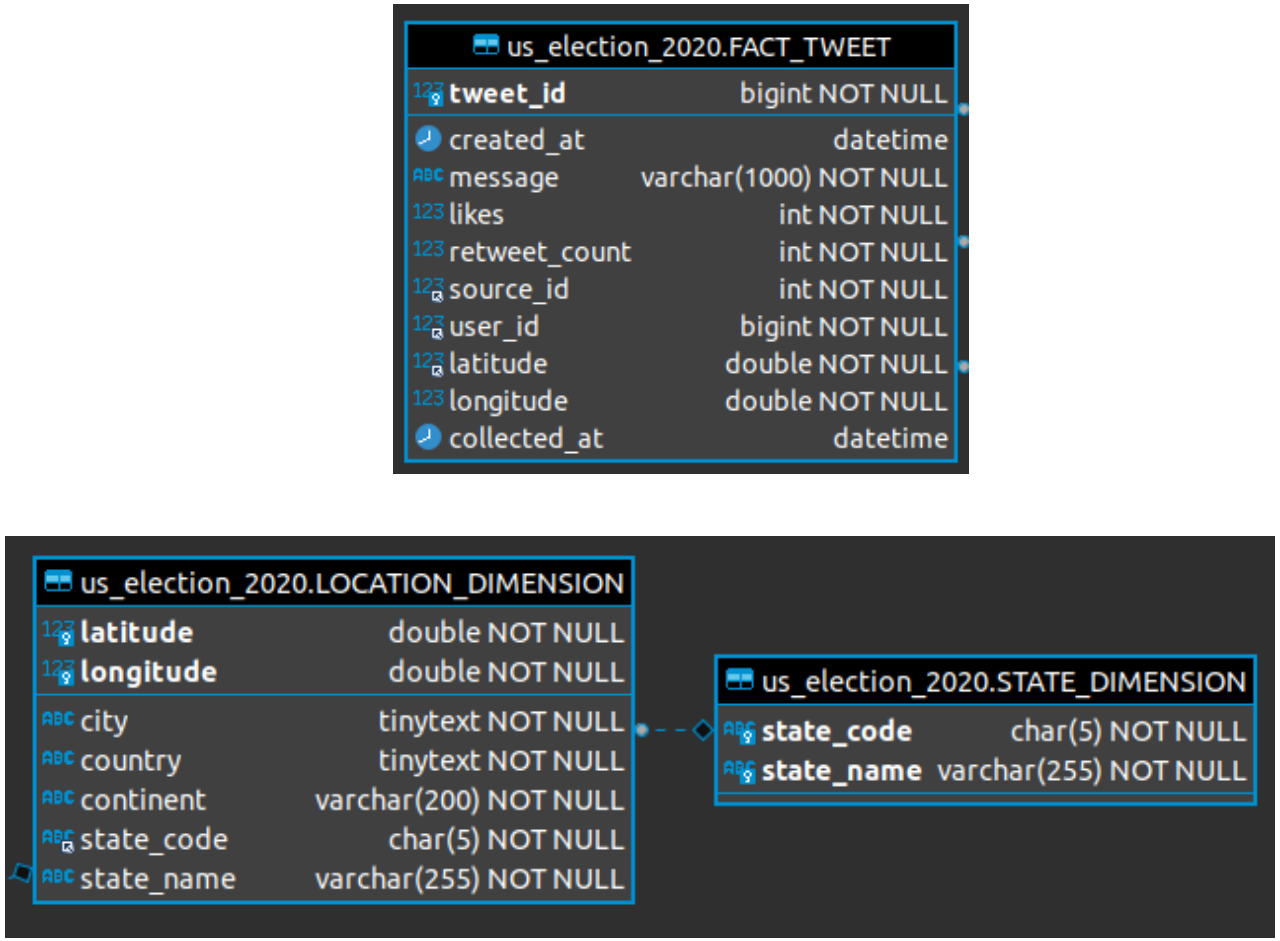


Trabalhando com Big Data



- Dump gerado
- Problemas com exportação de dados (Enconding)
- Inserts do Dump
- Formatação via Regex

Consultas


us_election_2020.FACT_TWEET
tweet_id bigint NOT NULL
created_at datetime
message varchar(1000) NOT NULL
likes int NOT NULL
retweet_count int NOT NULL
source_id int NOT NULL
user_id bigint NOT NULL
latitude double NOT NULL
longitude double NOT NULL
collected_at datetime
us_election_2020.LOCATION_DIMENSION
latitude double NOT NULL
longitude double NOT NULL
city tinytext NOT NULL
country tinytext NOT NULL
continent varchar(200) NOT NULL
state_code char(5) NOT NULL
state_name varchar(255) NOT NULL
us_election_2020.STATE_DIMENSION
state_code char(5) NOT NULL
state_name varchar(255) NOT NULL
us_election_2020.TWEETS_BY_STATE
state char(5) NOT NULL
name varchar(255) NOT NULL
quantity bigint NOT NULL

Consultas

```
SELECT * FROM TWEETS_BY_STATE LIMIT 10
```

	state	name	quantity
1	NY	New york	18,112
2	CA	California	15,935
3	ENG	England	11,786
4	DC	District of columbia	6,703
5	TX	Texas	6,691
6	IDF	Ile-de-france	6,669
7	FL	Florida	5,405
8	IL	Illinois	3,478
9	DL	Delhi	3,477
10	ON	Ontario	3,280

Consultas

us_election_2020.FACT_TWEET	
tweet_id	bigint NOT NULL
created_at	datetime
message	varchar(1000) NOT NULL
likes	int NOT NULL
retweet_count	int NOT NULL
source_id	int NOT NULL
user_id	bigint NOT NULL
latitude	double NOT NULL
longitude	double NOT NULL
collected_at	datetime

us_election_2020.USER_DIMENSION	
user_id	bigint NOT NULL
name	varchar(200) NOT NULL
screen_name	varchar(200) NOT NULL
description	varchar(300) NOT NULL
followers_count	int NOT NULL
location	tinytext NOT NULL

us_election_2020.USERS_ANALYTICS	
id	bigint NOT NULL
username	varchar(200) NOT NULL
followers	int NOT NULL
qtd_tweets	bigint NOT NULL
likes_avg	decimal(14,4)
retweet_avg	decimal(14,4)

Consultas

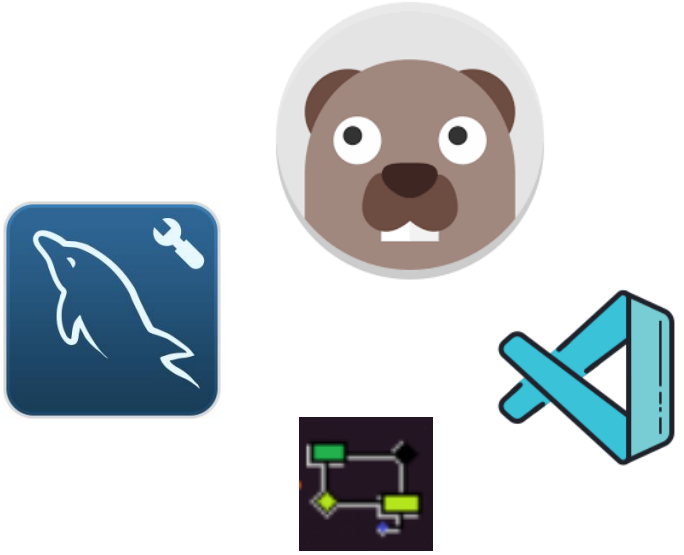
```
SELECT * FROM USERS_ANALYTICS WHERE likes_avg BI
```

Enter a SQL expression to filter results (use Ctrl+Space)

	id	username	followers	qtd_tweets	likes_avg	retweet_avg
1	105,165,371	andrebercoff	88,139	5	2,243.8	797.4
2	550,377,428	AshokShrivasta6	101,742	3	2,407.6667	373.3333
3	19,409,508	bobbyberk	444,512	2	2,854.5	102
4	2,329,492,448	OliviaTroye	118,266	2	2,813	608.5
5	18,208,368	kurtbardella	68,810	2	2,757.5	847
6	509,981,951	PorcherThomas	96,998	2	2,608	595
7	196,401,315	johndelancie	71,438	2	2,127	407
8	394,250,799	Br4mm3n	335,858	2	2,063	14.5
9	1,591,400,760	TaraSetmayer	105,668	2	2,011	178
10	56,039,856	abhijitmajumder	547,842	1	2,797	956
11	87,643,561	Rajput_Ramesh	91,070	1	2,749	231
12	1,066,505,848,791,474,176	FamilyKabs	38,636	1	2,605	352
13	14,085,099	KDKA	193,015	1	2,525	683
14	1,664,075,966	HeatherTDay	38,102	1	2,348	92
15	412,591,867	masonsixtencox	39,062	1	2,299	99
16	393,421,515	BernardineEvari	52,353	1	2,268	99
17	74,169,715	MzKatieCassidy	671,142	1	2,248	139
18	122,104,353	c_lindner	414,994	1	2,199	115
19	14,957,147	TheYoungTurks	429,163	1	2,123	469
20	1,249,274,011,109,949,440	TorreyHarris901	1,508	1	2,035	204

Ferramentas

- DBeaver
- MySQL Workbench
- Vscode
- BrModelo



Obrigado!