
DuckDB
O Patinho no Lago de Dados
Danilo Santos
CodeXperience
Zona da Mata
2024


DuckDB
O Patinho IA no Lago de Dados
Danilo Santos
CodeXperience
Zona da Mata
2024


PLATINA
UniAcademia
Rubeus
ÓCULOS
CodeXperience
Zona da Mata
2024


PARCEIROS
elastic
DevOps Days
Juiz de Fora
Tech Hub
Juiz de Fora
café das
MINAS
MEDEIROS GOURMET
BUFFET
CodeXperience
Zona da Mata
2024

Danilo Santos

Graduado em Sistemas de Informação, especialização em Engenharia de Dados.

Supervisor de Dados.

Professor de Banco de Dados e Python.

Projeto de consultoria, freelancer e treinamento.

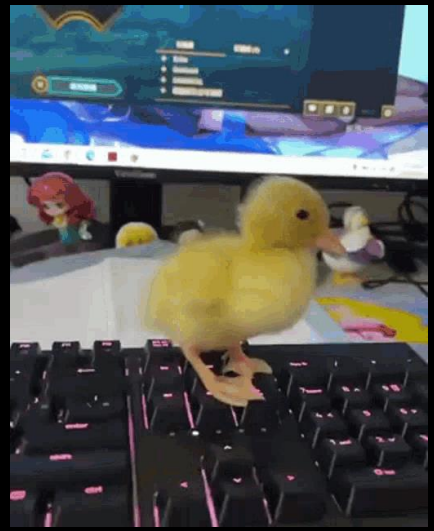
CodeXperience
Zona da Mata
2024

DJ KALANGO

CodeXperience
Zona da Mata
2024

Agenda?

- O que é DuckDB?
- Principais Características
- Documentação
- Livro MutherDuck
- Hands on



O que é DuckDB?

- CWI Amsterdan, Holanda, 2018
- É um banco de dados relacional
- Processamento Analítico - OLAP
- Embarcado
- Processamento em tempo de execução
- Código Aberto
- Gratuito

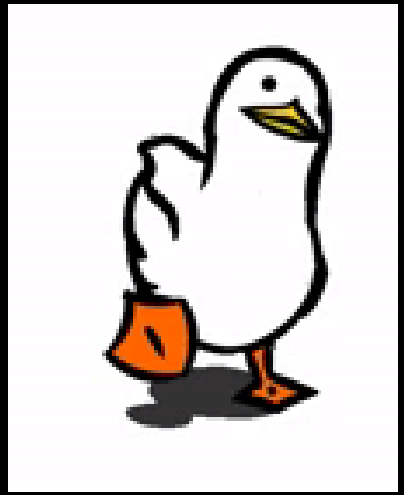
Porque DuckDB?

Hannes Mühleisen

Porque DuckDB?

Os patos são muito versáteis

- Voam
- Andam
- Nadam

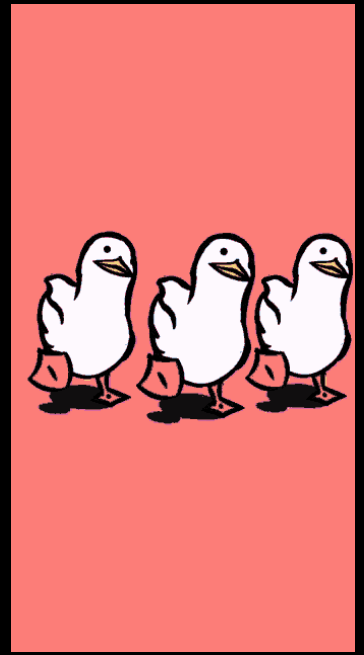


Principais Características

- Simples de usar
- Simples de instalar
- Sem Servidor
- Arquivo Único
- In-Memory
- Vetorização
- Paralelo
- Nunca Falha *

GBs

TBs



Principais Características

Otimização de Query

- Reescrita de execução
- Join, Agregação, Ordenação
- Achatamento de subquery
- Projeção e filtros

Principais Características

SGBDS

OLTP
Processamento em linha
SQL
Sem Data Frame
Servidor
Armazenamento complexo

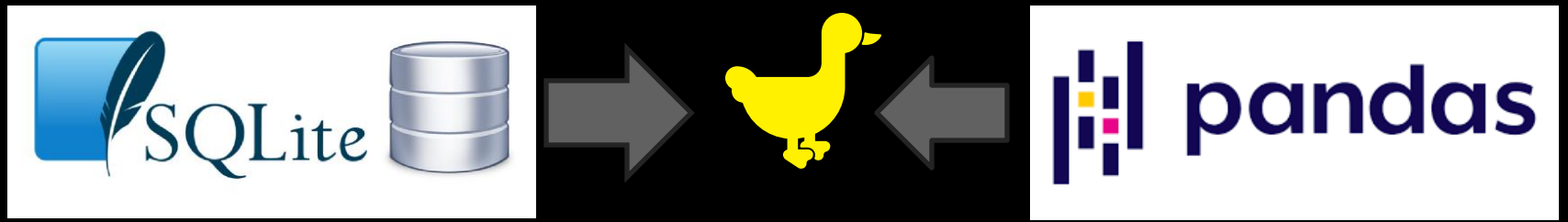
Pandas

OLAP
Processamento em Coluna
Sem SQL
Data Frame
Sem servidor
Sem armazenamento

DuckDB

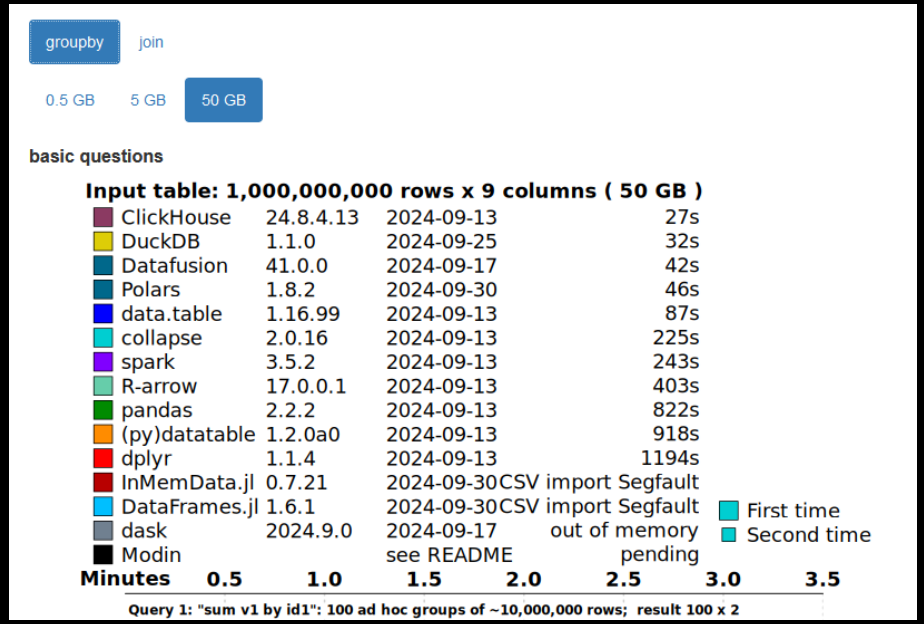
OLAP
Processamento Vetorizado
SQL
Data Frame
Sem Servidor
Armazena único arquivo
Otimizado cache CPU

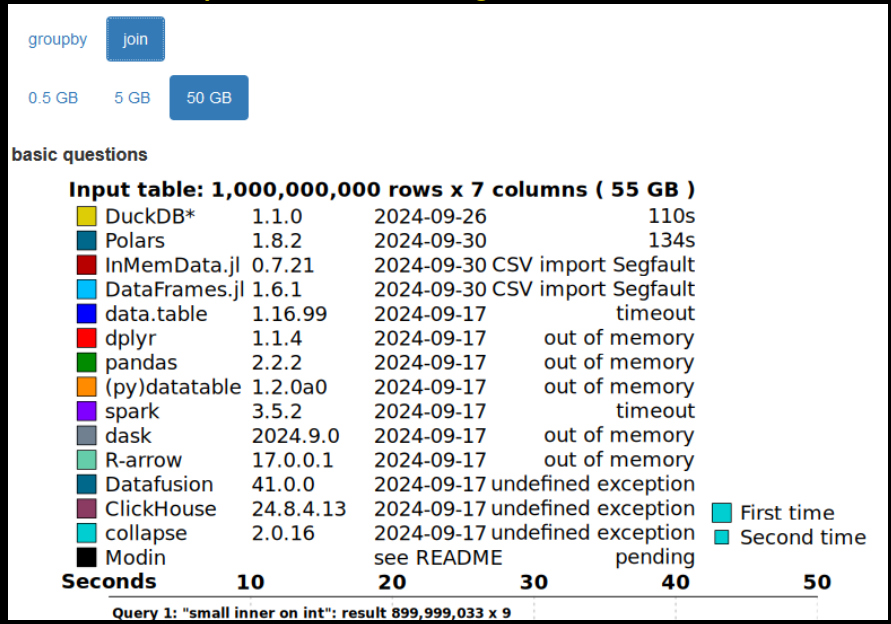
Principais Características



Principais Características

https://duckdblabs.github.io/db-benchmark/





Principais Características

Python DB API

- Engine PostgreSQL

API Relacional

- NumPy
- PyArrow
- Pandas
- Polars
- SQL Alchemy

** zero cópia

Principais Características








Principais Características

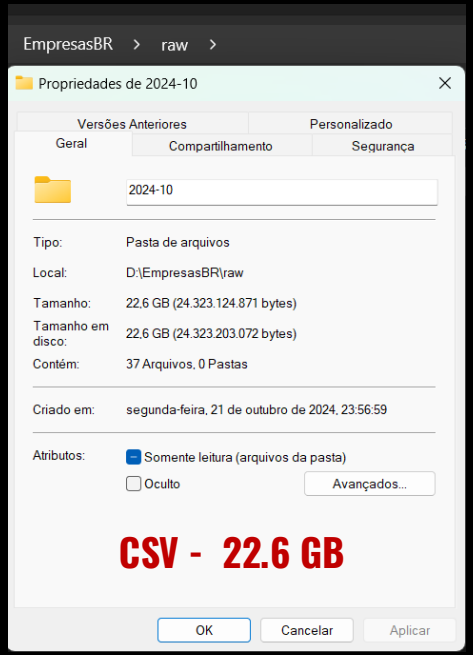
Principais Características

Principais Características

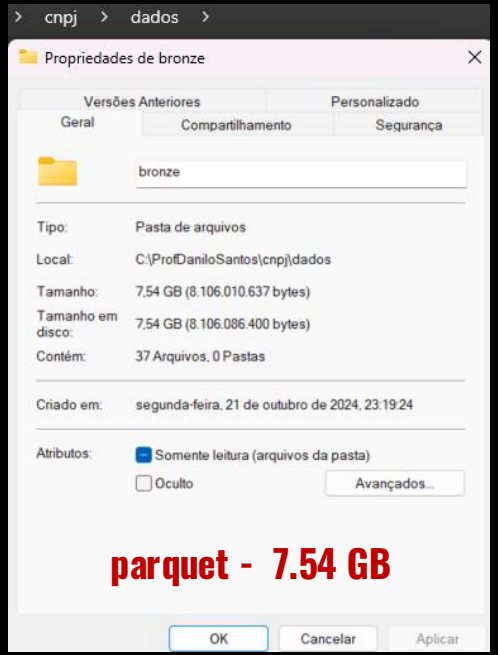
Compressão de Dados

Version	Taxi	On Time	Lineitem	Notes	Date
DuckDB v0.2.8	15.3GB	1.73GB	0.85GB	Uncompressed	July 2021
DuckDB v0.2.9	11.2GB	1.25GB	0.79GB	RLE + Constant	September 2021
DuckDB v0.3.2	10.8GB	0.98GB	0.56GB	Bitpacking	February 2022
DuckDB v0.3.3	6.9GB	0.23GB	0.32GB	Dictionary	April 2022
DuckDB v0.5.0	6.6GB	0.21GB	0.29GB	FOR	September 2022
DuckDB dev	4.8GB	0.21GB	0.17GB	FSST + Chimp	`now()`
CSV	17.0GB	1.11GB	0.72GB		
Parquet (Uncompressed)	4.5GB	0.12GB	0.31GB		
Parquet (Snappy)	3.2GB	0.11GB	0.18GB		
Parquet (ZSTD)	2.6GB	0.08GB	0.15GB		

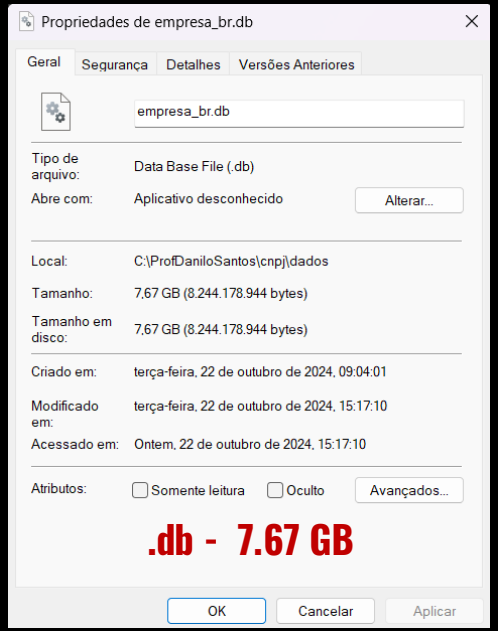
Compressão de Dados



CSV - 22.6 GB



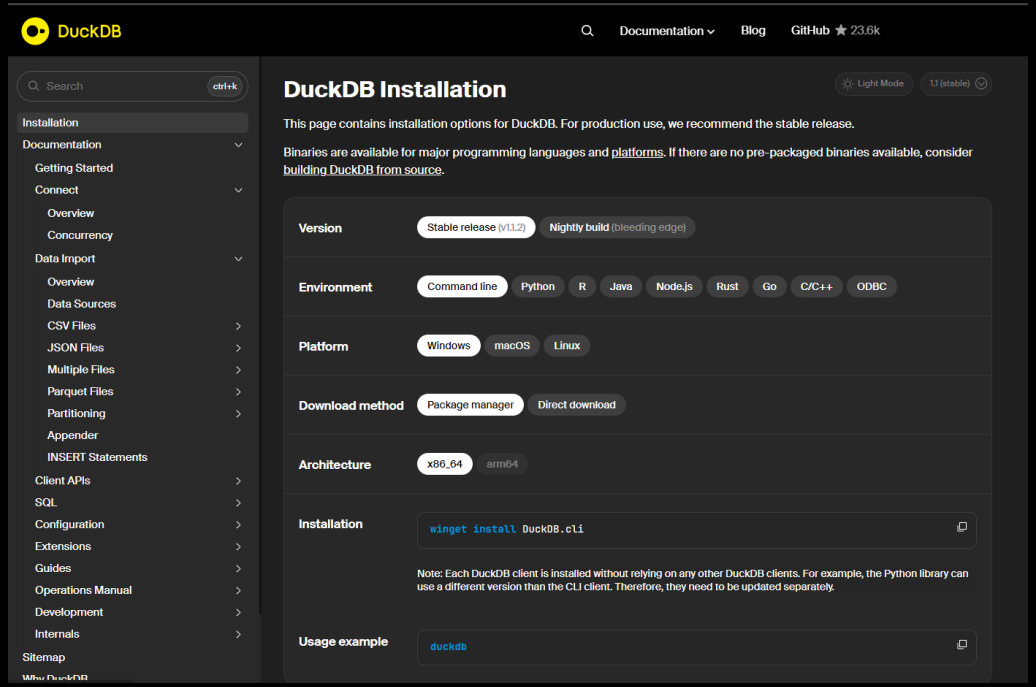
parquet - 7.54 GB



.db - 7.67 GB

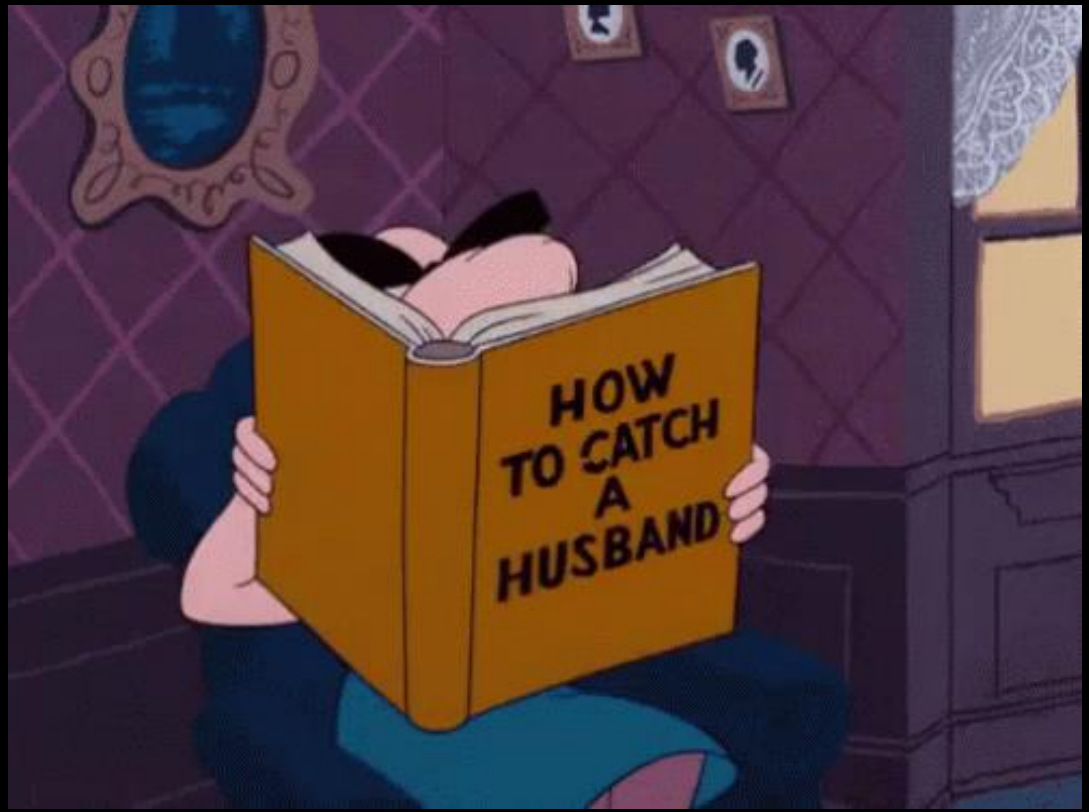

GIFAK.NET

Documentação

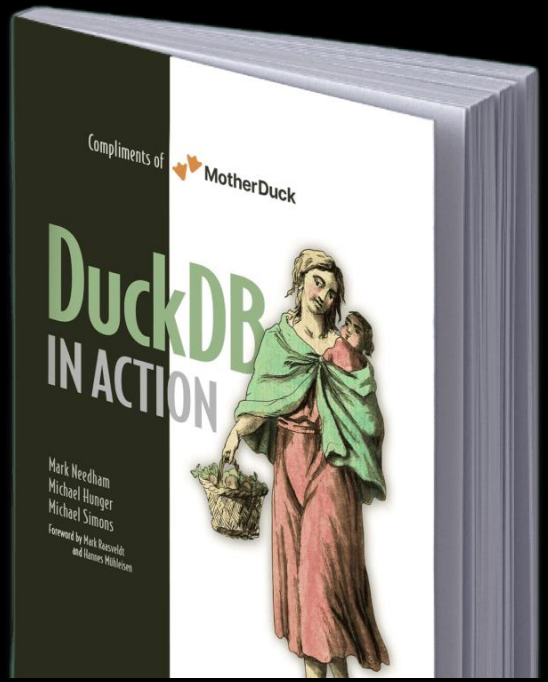


https://duckdb.org/

CodeXperience
Zona da Mata
2024


HOW
TO CATCH
A
HUSBAND

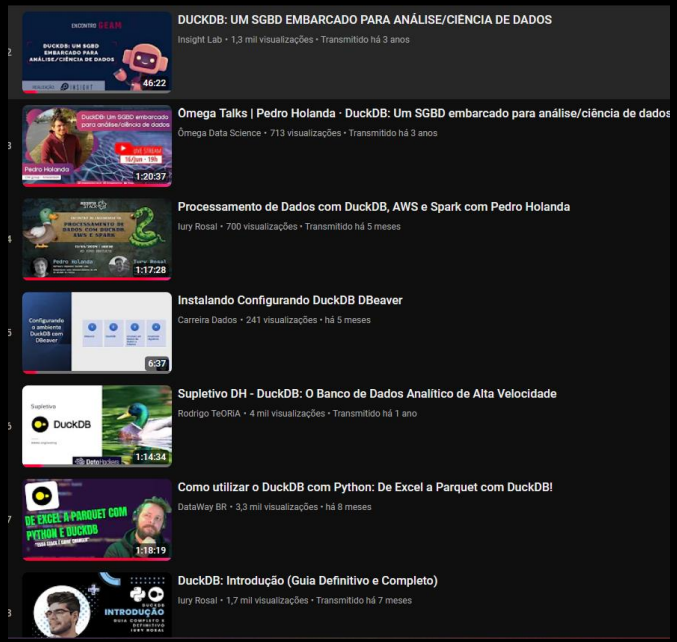
DuckDB In Action - Mother Duck

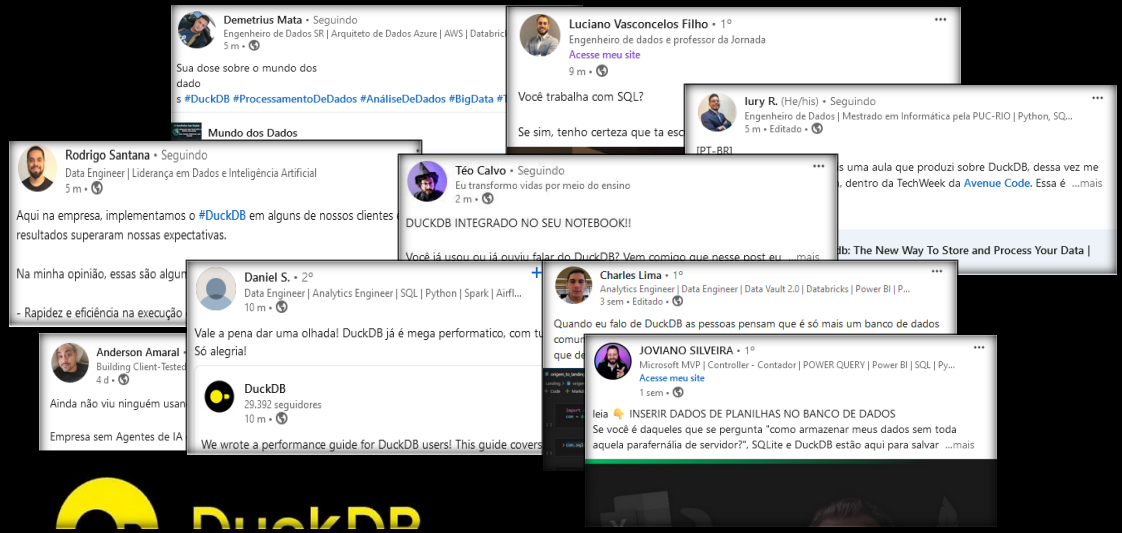


http://motherduck.com/duckdb-book-brief

CodeXperience
Zona da Mata
2024

Falando?





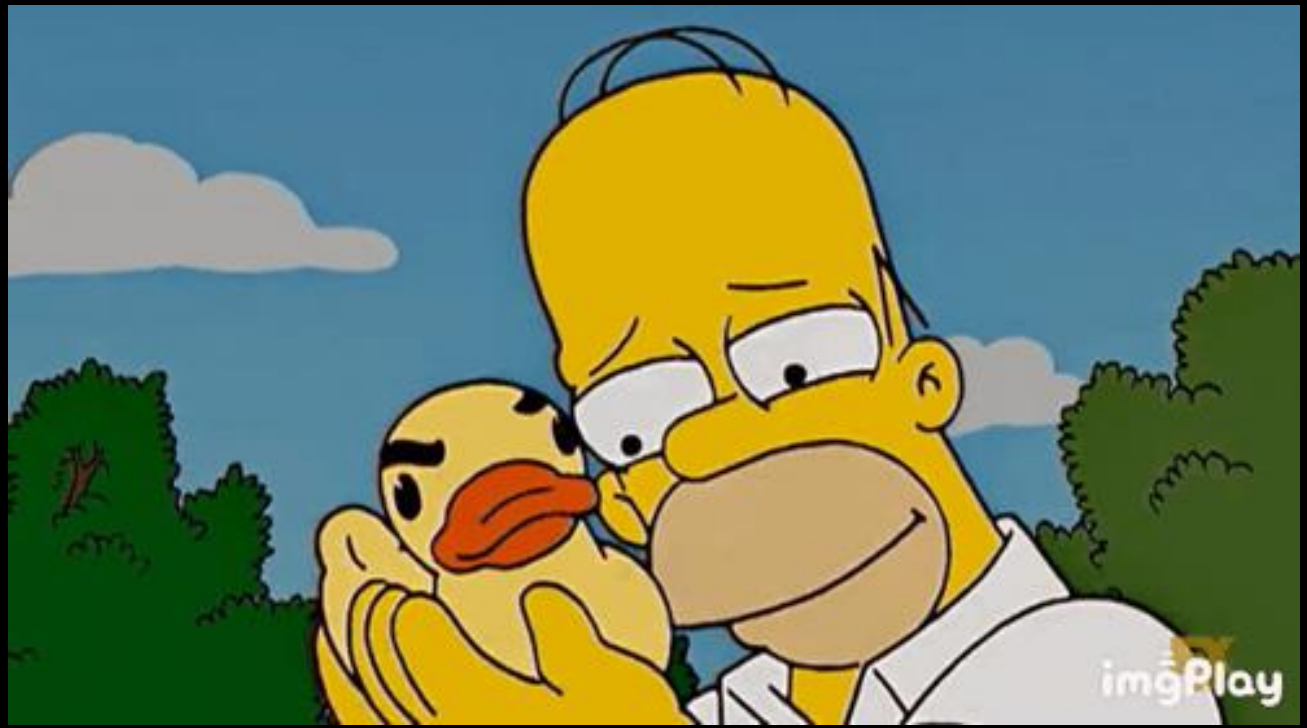
CodeXperience
Zona da Mata
2024

Falando?

https://db-engines.com/en/ranking

423 systems in ranking, October 2024

Rank Oct 2024	Rank Sep 2024	Rank Oct 2023	DBMS	Database Model	Score Oct 2024	Score Sep 2024	Score Oct 2023
1.	1.	1.	Oracle	Relational, Multi-model	1309.45	+22.85	+48.03
2.	2.	2.	MySQL	Relational, Multi-model	1022.76	-6.73	-110.56
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	802.09	-5.67	-94.79
4.	4.	4.	PostgreSQL	Relational, Multi-model	652.16	+7.80	+13.34
5.	5.	5.	MongoDB	Document, Multi-model	405.21	-5.02	-26.21
6.	6.	6.	Redis	Key-value, Multi-model	149.63	+0.20	-13.33
7.	7.	↑ 11.	Snowflake	Relational	140.60	+6.88	+17.36
8.	8.	↓ 7.	Elasticsearch	Multi-model	131.85	+3.06	-5.30
9.	9.	↓ 8.	IBM Db2	Relational, Multi-model	122.77	-0.28	-12.10
10.	10.	↓ 9.	SQLite	Relational	101.91	-1.43	-23.23
57.	↑ 60.	↑ 98.	DuckDB	Relational	5.98	+0.37	+2.34


imgPlay

Vagas?

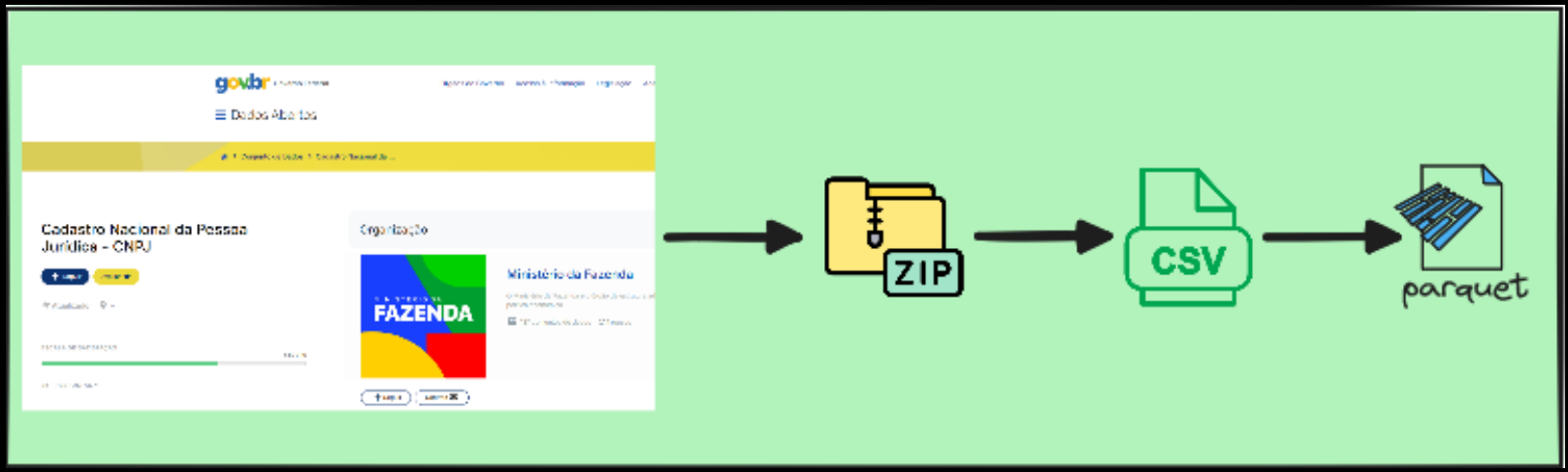
CodeXperience
Zona da Mata
2024

Caso de uso

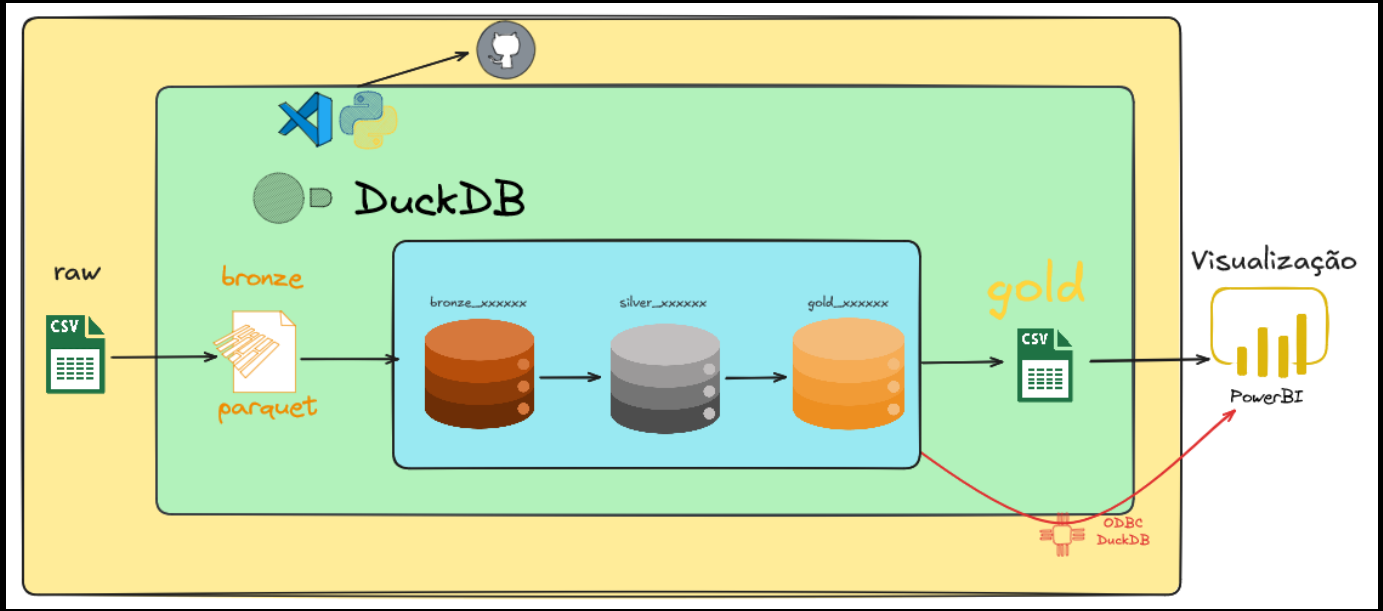
- Análise de Cadastro Nacional de Pessoas Jurídicas, dados abertos do Gov.BR;
- Apuração de concurso de rádio amador;
- Integração de sistema entre APIs/Json e Mysql;
- Documentação de pipelines Data Factory (poc);
- Migrando o curso de BD para DuckDB;

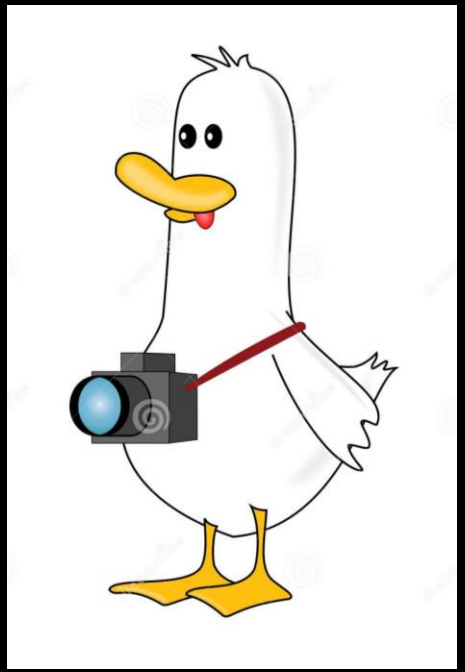
Hands on

https://dados.gov.br/dados/conjuntos-dados/cadastro-nacional-da-pessoa-juridica---cnpj



Hands on


DuckDB
raw
CSV
bronze
parquet
bronze_xxxxxx
silver_xxxxxx
gold_xxxxxx
gold
CSV
Visualização
PowerBI
ODBC
DuckDB



Arquivos

https://github.com/profdanilosantos/duckdb_codexperience_2024_juizdefora

CodeXperience
Zona da Mata
2024

Danilo Santos

- Email: danilo@danilosantos.dev.br
- linkedin.com/in/danilo-oliveira-santos
- @djkalango

CodeXperience
Zona da Mata
2024