

DuckDB

O Patinho no Lago de Dados

Danilo Santos



DuckDB

O Patinho **IA** no Lago de Dados

Danilo Santos

PLATINA



UniAcademia

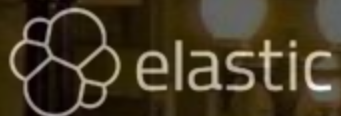


Rubeus

m3 ÓCULOS

« CodeXperience »
Zona da Mata
2024

PARCEIROS



DevOps Days
Juiz de Fora



Tech Hub
Juiz de Fora



« 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.





Agenda?

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



O que é DuckDB?

- 🐥 CWI Amsterdam, 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



Simple de usar



Simple de instalar



Sem Servidor



Arquivo Único



In-Memory



Vetorização



Paralelo



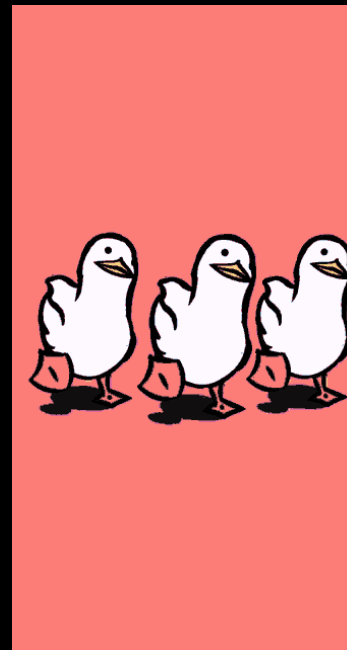
Nunca Falha *



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

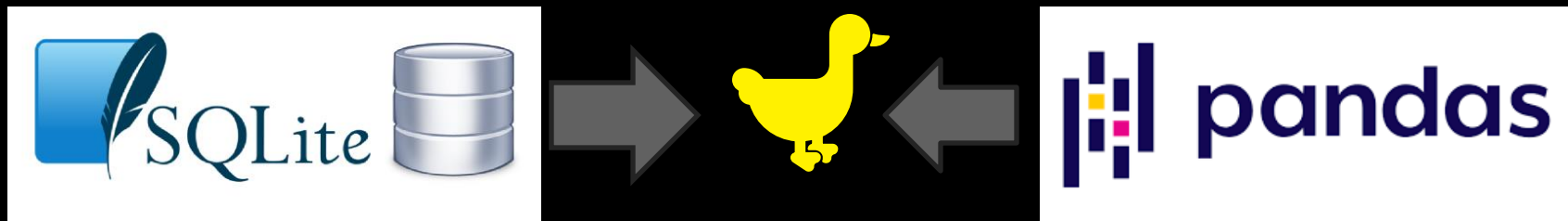
Sem Servidor

Armazena único arquivo

Otimizado cache CPU



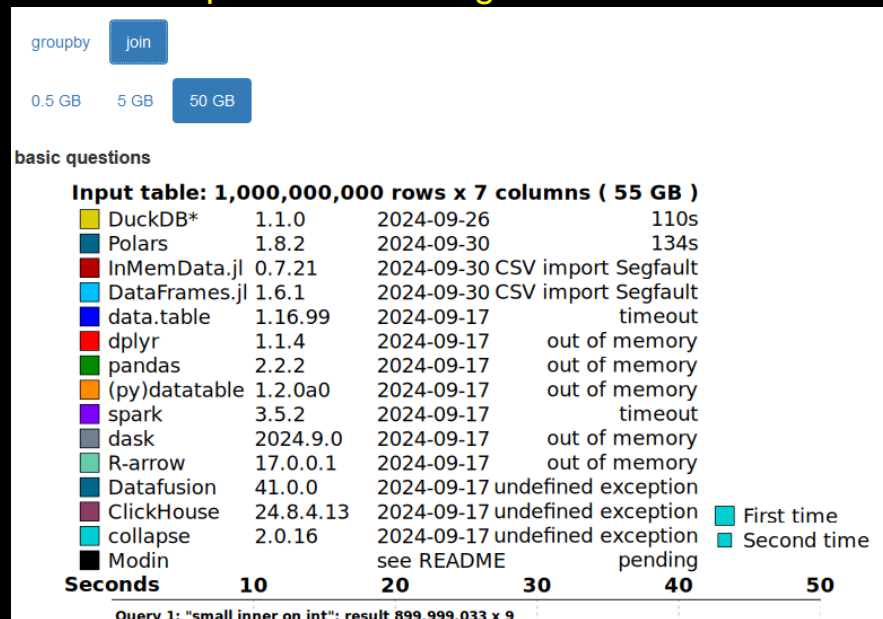
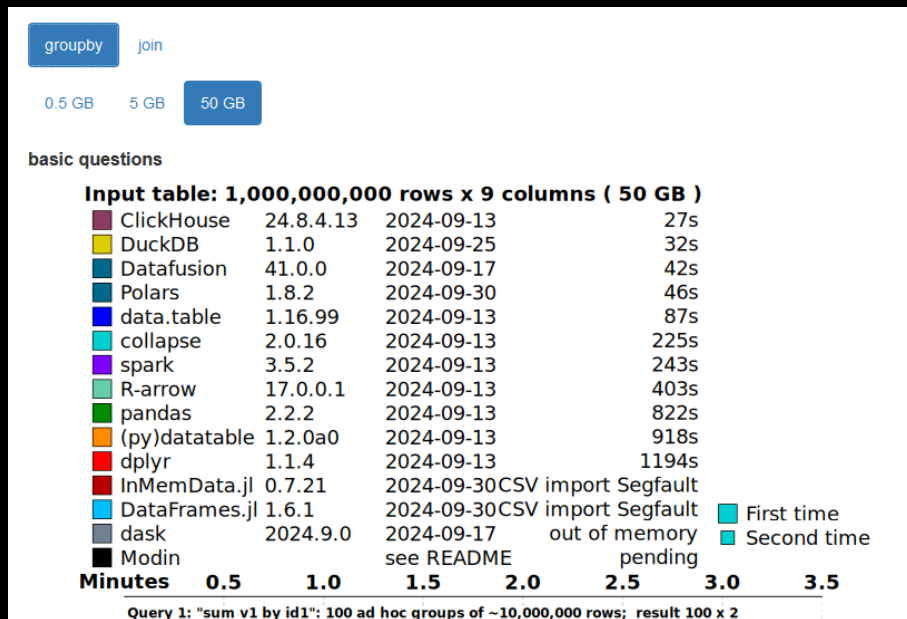
Principais Características



Principais Características



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



Principais Características

Python DB API

 Engine PostgreSQL

API Relacional

 **NumPy**

 **PyArrow**

 **Pandas**

 **Polars**

 **SQLAlchemy**

**** zero cópia**



Principais Características



Principais Características



Principais Características



Principais Características



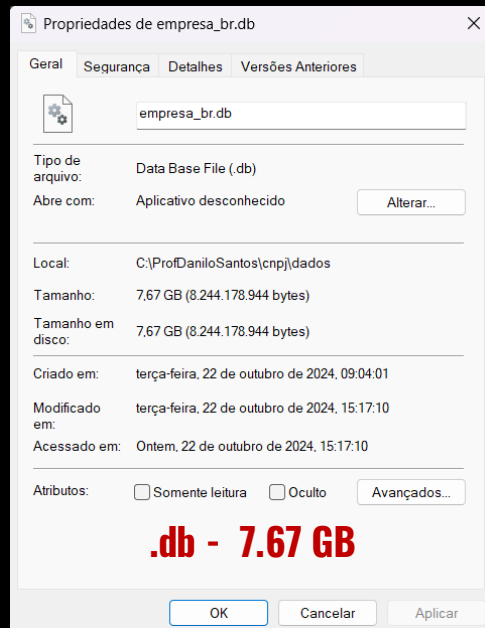
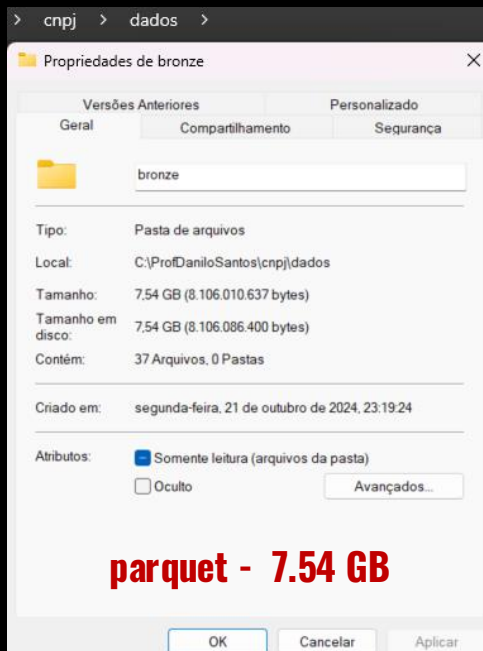
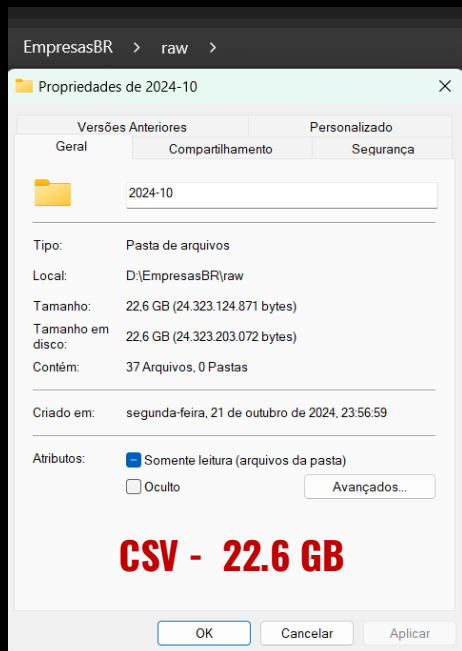
Data Saturday
Vitória 2024

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	<code>now()</code>
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





Documentação



🔍 Documentation ▾ Blog GitHub ★ 23.6k

🔍 Search ctrl+k

Installation

Documentation ▾

Getting Started

Connect ▾

Overview

Concurrency

Data Import ▾

Overview

Data Sources

CSV Files >

JSON Files >

Multiple Files >

Parquet Files >

Partitioning >

Appender

INSERT Statements

Client APIs >

SQL >

Configuration >

Extensions >

Guides >

Operations Manual >

Development >

Internals >

Sitemap

What's DuckDB?

DuckDB Installation

🌓 Light Mode 1.1 (stable) 🌙

This page contains installation options for DuckDB. For production use, we recommend the stable release.

Binaries are available for major programming languages and [platforms](#). If there are no pre-packaged binaries available, consider [building DuckDB from source](#).

Version

Stable release (v1.12)

Nightly build (bleeding edge)

Environment

Command line

Python

R

Java

Node.js

Rust

Go

C/C++

ODBC

Platform

Windows

macOS

Linux

Download method

Package manager

Direct download

Architecture

x86_64

arm64

Installation

`winget install DuckDB.cli`



Note: Each DuckDB client is installed without relying on any other DuckDB clients. For example, the Python library can use a different version than the CLI client. Therefore, they need to be updated separately.

Usage example

`duckdb`

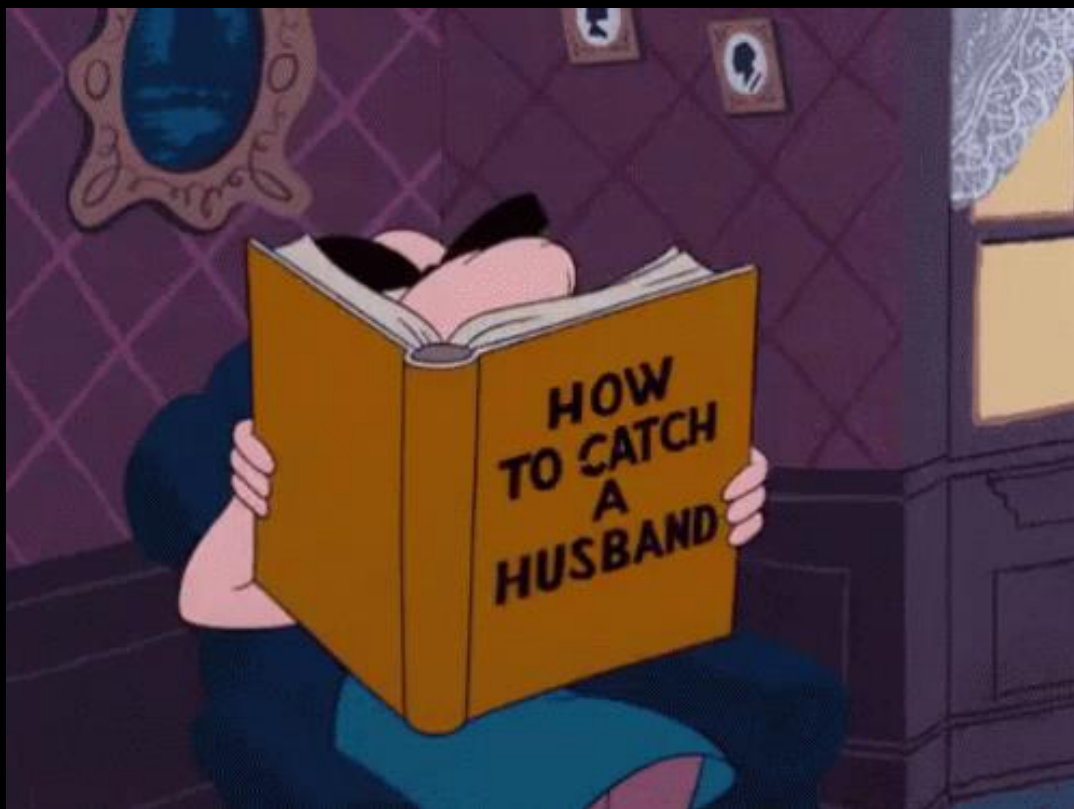


<https://duckdb.org/>

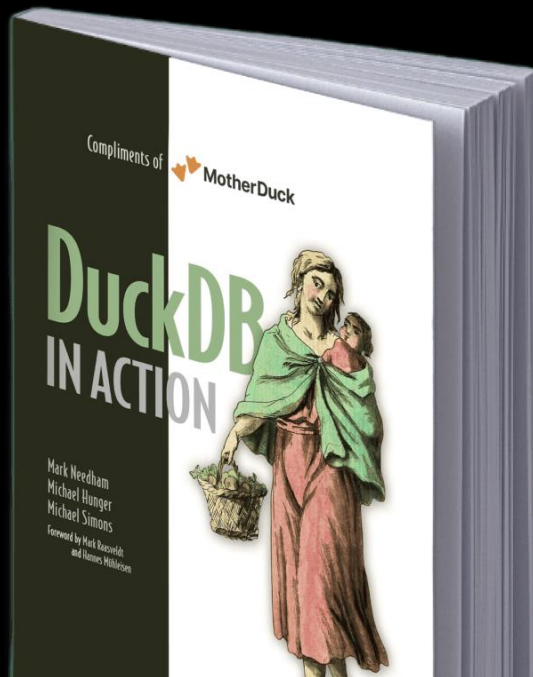


« CodeXperience »
Zona da Mata
2024





DuckDB In Action - Mother Duck



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



« CodeXperience »
Zona da Mata
2024



Falando?

DUCKDB: UM SGBD EMBARCADO PARA ANÁLISE/CIÊNCIA DE DADOS
Insight Lab • 1,3 mil visualizações • Transmitido há 3 anos

Omega Talks | Pedro Holanda • DuckDB: Um SGBD embarcado para análise/ciência de dados
Ômega Data Science • 713 visualizações • Transmitido há 3 anos

Processamento de Dados com DuckDB, AWS e Spark com Pedro Holanda
Iury Rosal • 700 visualizações • Transmitido há 5 meses

Instalando Configurando DuckDB DBEaver
Carreira Dados • 241 visualizações • há 5 meses

Supletivo DH - DuckDB: O Banco de Dados Analítico de Alta Velocidade
Rodrigo TEORIA • 4 mil visualizações • Transmitido há 1 ano

Como utilizar o DuckDB com Python: De Excel a Parquet com DuckDB!
DataWay BR • 3,3 mil visualizações • há 8 meses

DuckDB: Introdução (Gua Definitivo e Completo)
Iury Rosal • 1,7 mil visualizações • Transmitido há 7 meses

Demetrius Mata • Seguindo
Engenheiro de Dados SR | Arquiteto de Dados Azure | AWS | Databricks
5 m •

Sua dose sobre o mundo dos dados
#DuckDB #ProcessamentoDeDados #AnaliseDeDados #BigData #Mundo dos Dados

Luciano Vasconcelos Filho • 1º
Engenheiro de dados e professor da Jornada
Acesse meu site
9 m •

Você trabalha com SQL?
Se sim, tenho certeza que ta esc...

Iury R. (He/his) • Seguindo
Engenheiro de Dados | Mestrado em Informática pela PUC-RIO | Python, SQL...
5 m • Editado •

Rodrigo Santana • Seguindo
Data Engineer | Liderança em Dados e Inteligência Artificial
5 m •

Aqui na empresa, implementamos o **#DuckDB** em alguns de nossos clientes e resultados superaram nossas expectativas.

Na minha opinião, essas são algumas das vantagens do DuckDB:
- Rapidez e eficiência na execução

Téo Calvo • Seguindo
Eu transformo vidas por meio do ensino
2 m •

DUCKDB INTEGRADO NO SEU NOTEBOOK!!
Você já usou ou já ouviu falar do DuckDB? Vem comigo, que nesse post eu ...mais

Daniel S. • 2º
Data Engineer | Analytics Engineer | SQL | Python | Spark | Airf...
10 m •

Vale a pena dar uma olhada! DuckDB já é mega performático, com tu...
Só alegria!

Anderson Amaral
Building Client-Tested
4 d •

Ainda não viu ninguém usando DuckDB?
Empresa sem Agentes de IA

DuckDB
29.392 seguidores
10 m •

We wrote a performance guide for DuckDB users! This guide covers

Charles Lima • 1º
Analytics Engineer | Data Engineer | Data Vault 2.0 | Databricks | Power BI | P...
3 sem • Editado •

Quando eu falo de DuckDB as pessoas pensam que é só mais um banco de dados...
comum...
que de...

JOVIANO SILVEIRA • 1º
Microsoft MVP | Controller - Contador | POWER QUERY | Power BI | SQL | Py...
Acesse meu site
1 sem •

leia **INSERIR DADOS DE PLANILHAS NO BANCO DE DADOS**
Se você é daqueles que se pergunta "como armazenar meus dados sem toda aquela parafernália de servidor?", SQLite e DuckDB estão aqui para salvar ...mais

DuckDB: The New Way To Store and Process Your Data |



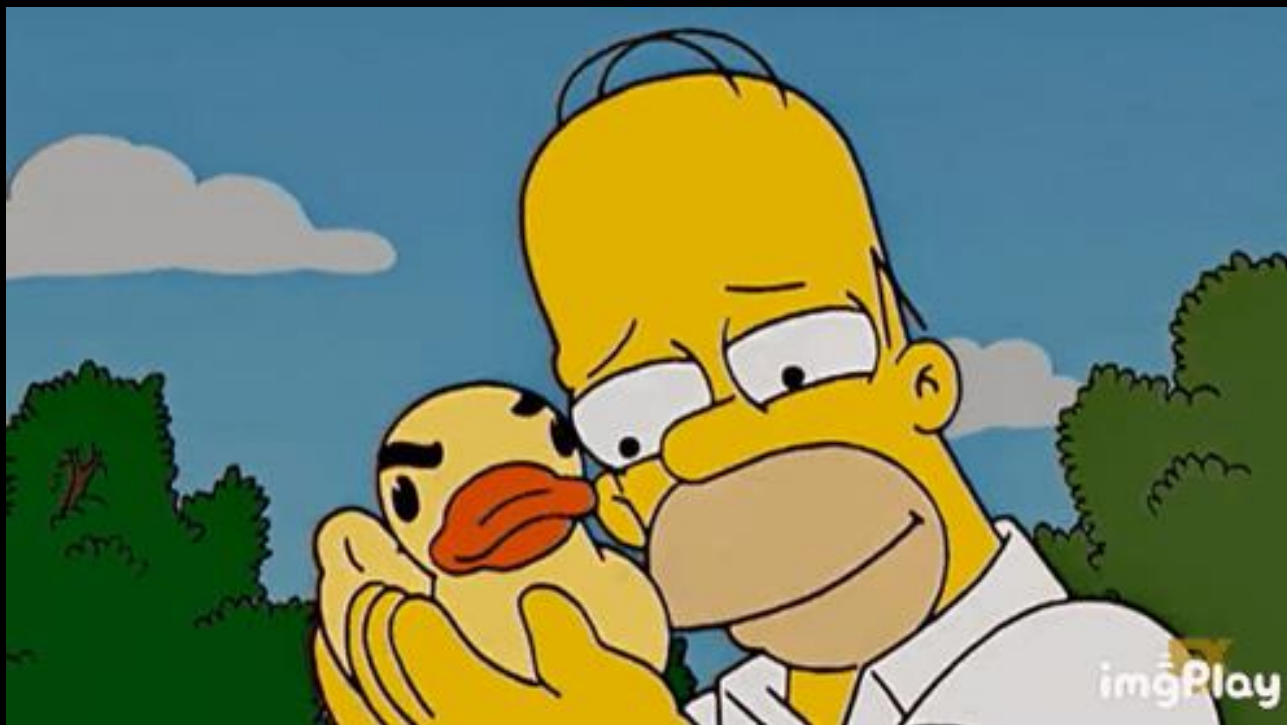
Falando?

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

423 systems in ranking, October 2024

Rank			DBMS	Database Model	Score		
Oct 2024	Sep 2024	Oct 2023			Oct 2024	Sep 2024	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





Vagas?





Caso de uso

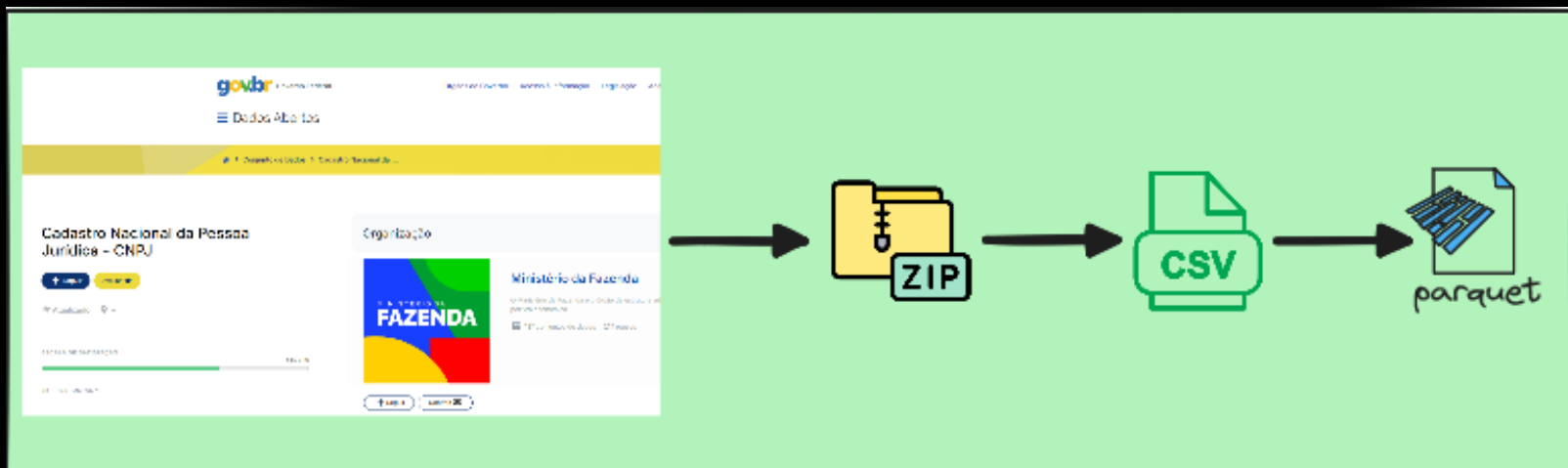
- 🐦 Análise de Cadastro Nacional de Pessoas Jurídicas, dados abertos do Gov.BR;
- 🐦 Apuração de concurso de rádio amador;
- 🐦 Integração de sistema entre APIs/Json e Mysql;
- 🐦 Documentação de pipelines Data Factory (poc);
- 🐦 Migrando o curso de BD para DuckDB;



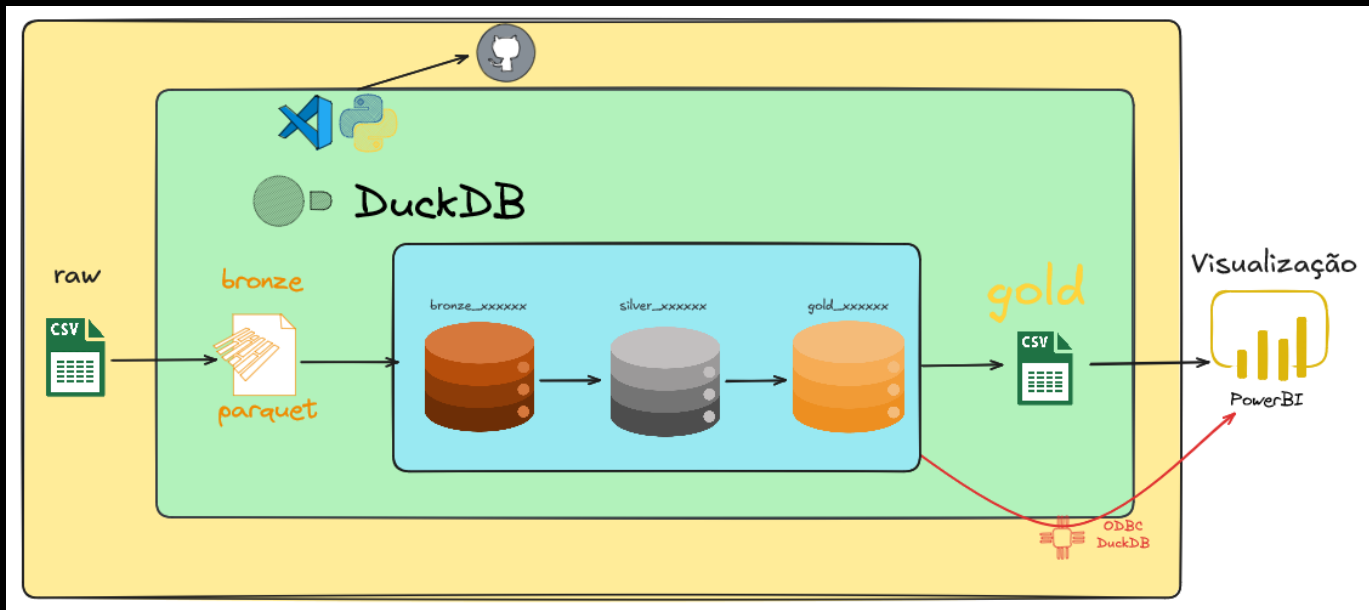


Hands on

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

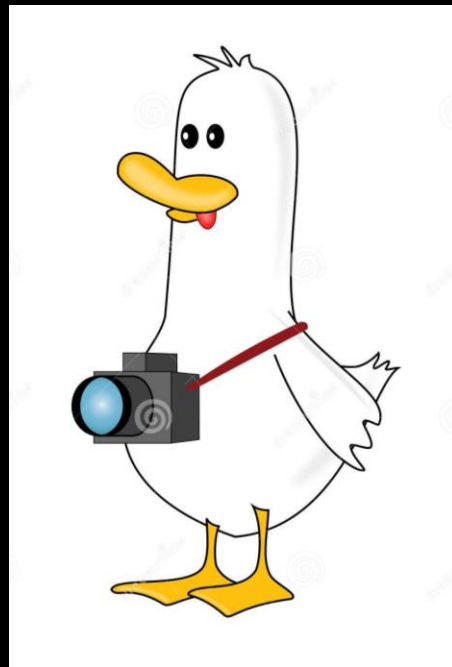


Hands on









Arquivos

https://github.com/profdanilosantos/duckdb_codexperience_2024_juizdefora



Danilo Santos

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

