

Danilo Santos





O Patinho IA no Lago de Dados

Danilo Santos



Patrocinadores















Patrocinadores



















Danilo Santos

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

Vero Internet: Supervisor de Dados.

Faculdade Ensin-e: Professor de Banco de Dados e Python.

Menara Dados: meu projeto de consultoria, freelancer e treinamento.



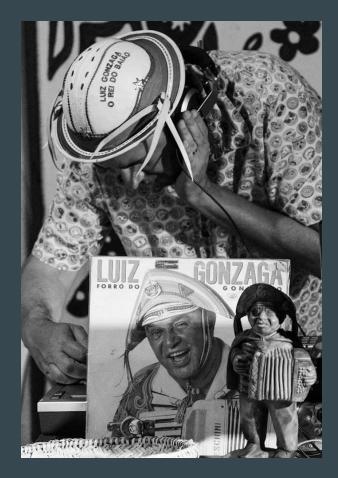
















O que vamos falar?

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





O que é DuckDB?

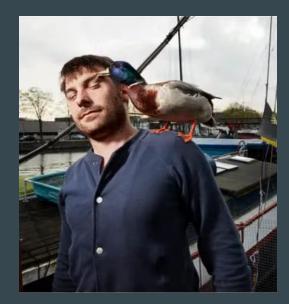
- CWI Amsterdan, Holanda, 2018
- É um banco de dados relacional
- Processamento Analítico OLAP
- Embarcado
- Processamento em tempo de execução
- Código Aberto
- Gratuito







Porque se chama DuckDB?



Hannes Mühleisen





Porque se chama DuckDB?

Os patos são muito versáteis



Voam



Andam



Nadam







Simples de usar



Simples de instalar



Sem Servidor



Arquivo Único



In-Memory



Vetorização



Paralelo



Nunca Falha *





Otimização de Query



Reescrita de execução



Join, Agregação, Ordenação



Achatamento de subquery



Projeção e filtros





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











API SQL

* PostgreSQL

API Relacional

* NumPy

* PyArrow

* Pandas

* Polares

* SQL Alchemy

** zero cópia



























































Compressão de Dados

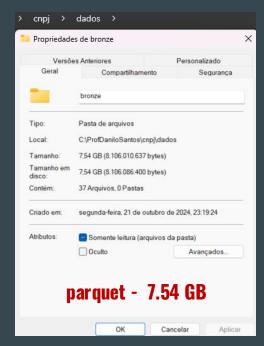
Version	Taxi	On Time	Lineitem	Notes	Date
DuckDB v0.2.8	15.3GB	1.73GB	0.85GB	Uncompressed	July 2021
DuckDB v0.2.9	11.2GB	1.25GB	0.79GB	RLE + Constant	September 2021
DuckDB v0.3.2	10.8GB	0.98GB	0.56GB	Bitpacking	February 2022
DuckDB v0.3.3	6.9GB	0.23GB	0.32GB	Dictionary	April 2022
DuckDB v0.5.0	6.6GB	0.21GB	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



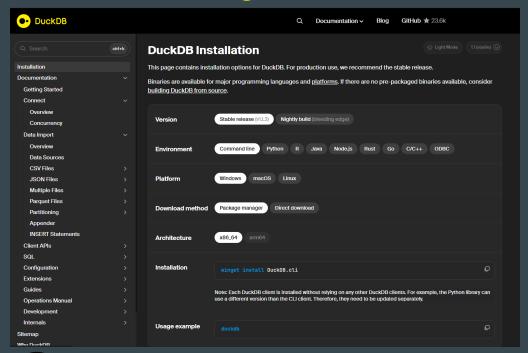


Propriedades de empresa_br.db								×	
Geral	Segura	nça	Detalhes	Versões Anteriores		iores			
***		empresa_l			r.db				
Tipo de arquivo		Dat	Data Base File (.db)						
Abre co	Abre com:		Aplicativo desconhecido Alterar						
Local:		C:\ProfDaniloSantos\cnpj\dados					-		
Taman	ho:	7,67 GB (8.244.178.944 bytes)							
Taman disco:	ho em	7,67 GB (8.244.178.944 bytes)							
Criado	em:	terça-feira, 22 de outubro de 2024, 09:04:01							
Modific em:	ado	terça-feira, 22 de outubro de 2024, 15:17:10							
Acessa	ado em:	Ontem, 22 de outubro de 2024, 15:17:10							
Atributo	os:		Somente leit	ura	Oculto		Avançados.	5	
		•	db -	7.	.67 G	B			
			ОК		Can	celar	Aplic	ar	





Documentação



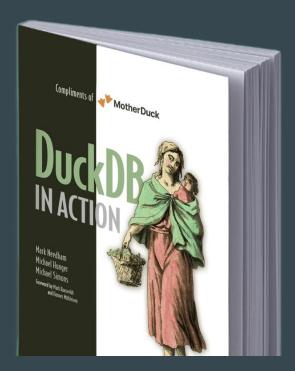
https://duckdb.org/







DuckDB In Action - Mother Duck



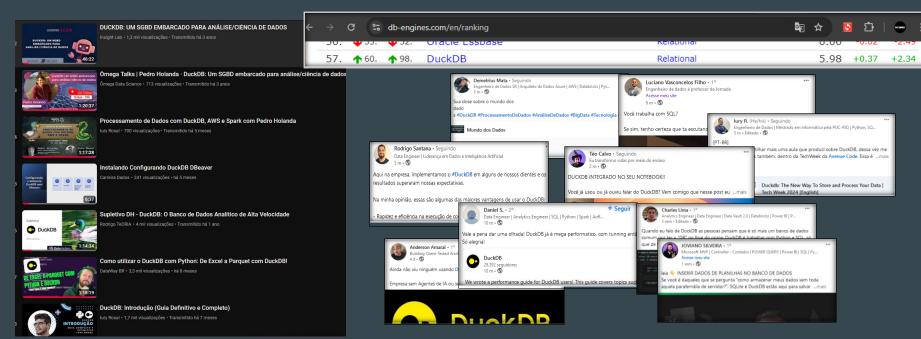
http://motherduck.com/duckdb-book-brief







O que estão falando?







O que estão falando?

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

	423 systems in ranking, October 2024							
	Rank				Score			
Oct 2024	Sep 2024	Oct 2023	DBMS	Database Model	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 [1]	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.	1 11.	Snowflake 🕂	Relational	140.60	+6.88	+17.36	
8.	8.	4 7.	Elasticsearch	Multi-model 👔	131.85	+3.06	-5.30	
9.	9.	4 8.	IBM Db2	Relational, Multi-model 👔	122.77	-0.28	-12.10	
10.	10.	4 9.	SQLite	Relational	101.91	-1.43	-23.23	
57.	1 60.	1 98.	DuckDB	Relational	5.98	+0.37	+2.34	







Vagas?





Caso de uso

- Analise 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

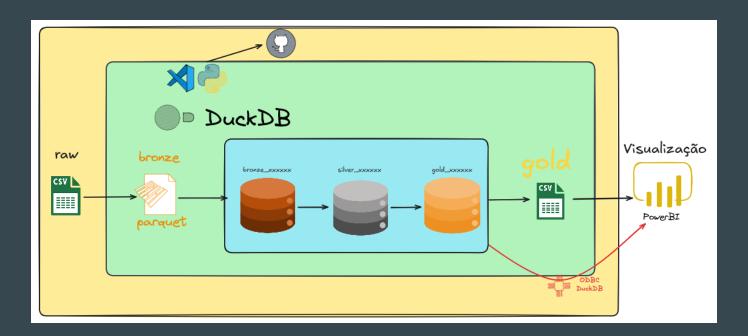
https://dados.gov.br/dados/conjuntos-dados/cadastro-nacional-da-pessoa-juridica---cnpj







Hands on







Meus Contatos

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







Arquivos

https://github.com/profdanilosantos/duckdb_data_saturday_2024_vitoria





