Exploring DuckDB potential

Josef Machytka
2 min read · 5 days ago

This article serves as an introduction to my series on the DuckDB analytical database. In this series, I have so far published the following articles (listed from newest to oldest):

- PostgreSQL and DuckDB: Supercharging Ad-Hoc Data Analysis and ETL (with recorded online talk)
- DuckDB Performance Problems with Inappropriate Pivoting Queries on Very Large Datasets
- DuckDB Database File as a New Standard for Sharing Data?
- Data Exports to Parquet Format Using DuckDB
- Statistics and Histograms with DuckDB
- Pivot Tables with DuckDB
- Extending DuckDB ETL Capabilities with Python
- DuckDB as a Rudimentary Data Migration Tool
- Cross-Database Selects with DuckDB
- Using DuckDB as an Intelligent ETL tool for PostgreSQL
- How DuckDB handles data not fitting into memory?

How I Met DuckDB

My interest in DuckDB started during the PostgreSQL conference Swiss PG Day 2024. There, a big poster compared PostgreSQL and DuckDB, presenting this new tool as superior to my long-time favorite PostgreSQL database. Naturally, I was intrigued and decided to dig deeper, conducting some performance tests to see what DuckDB was truly capable of.

At first, my search led me to articles that felt overly hyped and disappointingly shallow. Many of them were so generic that replacing the word “DuckDB” with the name of another tool would leave the content practically unchanged. However, my perspective shifted when I discovered the e-book *DuckDB in Action* on the website motherduck.com. After diving into its main chapters, I became captivated by DuckDB’s potential.

DuckDB is not just another enthusiastic open-source project destined to fade away in six months because its creator loses interest. It was created at the National Research Institute for Mathematics and Computer Science in Amsterdam, which is also credited as a birthplace of the Python programming language.

Being familiar with PostgreSQL, I find it fascinating that DuckDB takes significant inspiration from PostgreSQL. It closely follows PostgreSQL SQL syntax, achieving it thanks to its use of a repackaged PostgreSQL SQL parser. This means that anyone familiar with PostgreSQL commands can immediately feel at home using DuckDB. Beyond this compatibility, DuckDB also introduces some really interesting innovative features, which I explore in details across my articles.

Importantly, DuckDB is not positioned as a competitor to traditional large-scale databases like PostgreSQL or MySQL. Instead, it occupies a unique niche as an embedded analytical database optimized for heavy parallel analytical workloads, leveraging its vectorized query engine. Remarkably, it can efficiently process very large datasets — even those not fitting into memory. I believe DuckDB holds significant promise as a complementary tool to established databases, as well as to modern data infrastructures like Data Lakes and Lakehouses. Let’s explore its potential deeper.

Image created by the author using DeepDreamGenerator

Postgresql Duckdb Data Analysis Etl Tool

Edit profile

Written by Josef Machytka

58 Followers · 15 Following

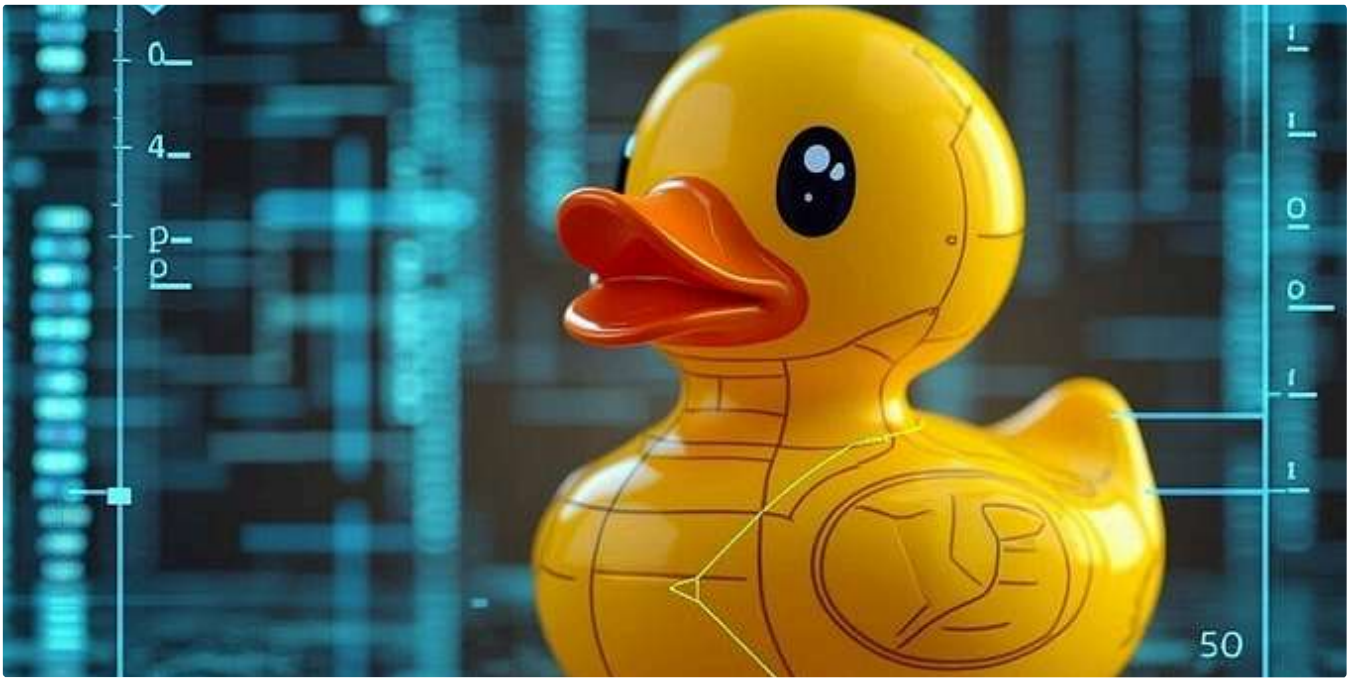
I work as PostgreSQL specialist & database reliability engineer at NetApp Deutschland, Open Source Services division.

No responses yet

What are your thoughts?

Respond

More from Josef Machytka



Josef Machytka

DuckDB Database File as a New Standard for Sharing Data?

This is not my original idea; I came across it in an excellent article titled “DuckDB Beyond the Hype” by Alireza Sadeghi. However, it...

Dec 30, 2024 17 2

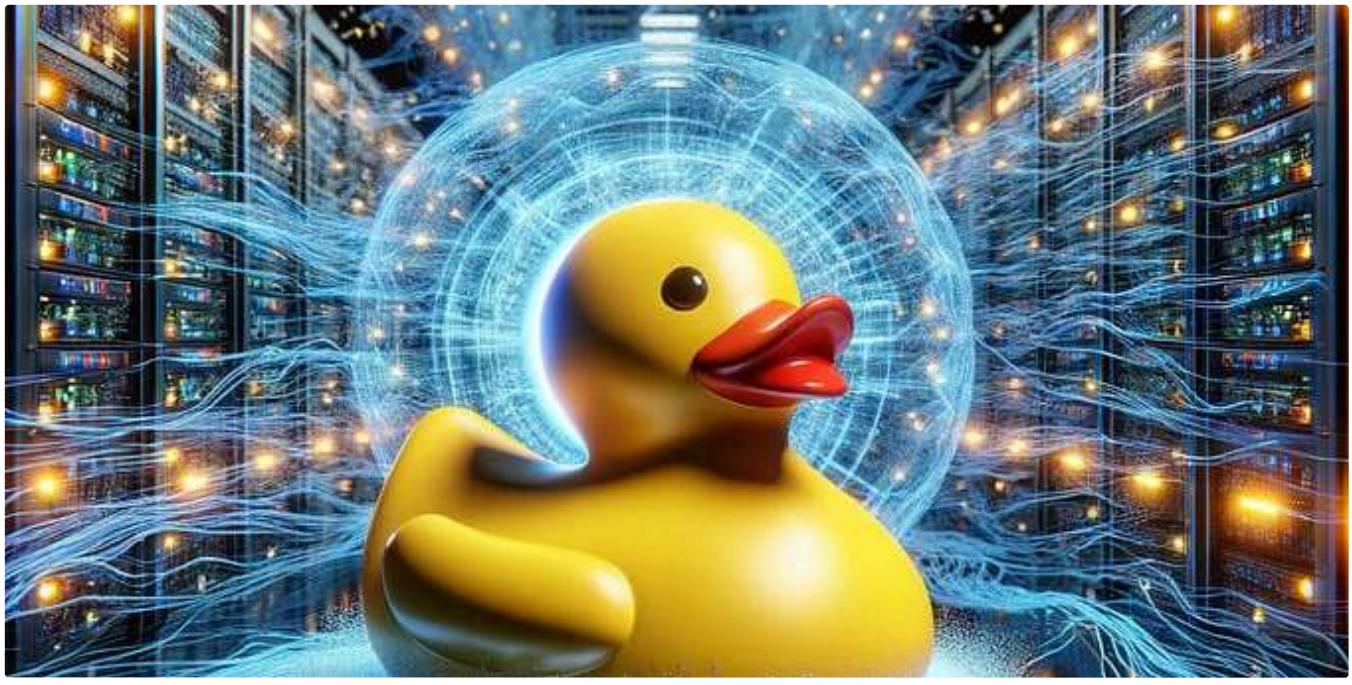
probe_id int64	2024-12-18_avg_tem… double	2024-12-19_avg_tem… double	2024-12-20_avg_tem… double	2024-12-21_avg_tem… double	…	2024-12-27_avg_tem… double	2024-12-28_avg_tem… double	2024-12-29_avg_tem… double	2024-12-30_avg_tem… double
1	-0.09333333333333331	0.008709499040937112	-0.09530979615314230	0.010285515229008652	…	0.024082900616364426	-0.06485498315668406	0.012230606453306028	2.89
2	-1.8681481481481474	-0.011747858418378…	-0.19133368175211093	-0.06350527768649375	…	-0.04171747826158102	-0.02843359476698743	-0.025157060000087…	-31.82
3	0.5514285714285712	0.03776647082685746	0.27300376356349165	0.14048107877274826	…	0.011376852696737617	0.0502324666659877	0.08683195299201234	-3.990000000000001
4	0.15520325203252047	-0.008692187717719168	0.08583476862546609	-0.061396022144761 16	…	0.19196071702861273	-0.08186630396370788	0.08936674686468202	19.36
5	0.5163888888888885	-0.045700490736644817	0.01953642221439854	-0.015551540087672…	…	-0.01359002275741 25	-0.07238289422942779	-0.18353784234809068	
6	1.0165486725663717	0.029071114059955196	0.08107486316960297	0.017795231646764362	…	0.11904802583941525	0.04725992074913865	0.08915156107911504	-27.380000000000003
7	0.07568807339449547	0.14391917325689826	0.05512613147262228	-0.09983962752198643	…	0.16227462051478873	-0.15890324113132445	-0.05464776857339331	
8	2.034600000000001	0.11482386883886646	0.09522098292944779	0.062771995286724 14	…	0.10036929933070622	0.13754822380761314	-0.05208855536214023	17.05
9	1.5455856179775286	0.09855607778079753	-0.03765412237862671	0.04697267741617703	…	-0.06521116188527977	-0.0032089106887395…	0.06777951553103001	-19.71
10	-0.6838000000000001	-0.28492896786129136	-0.08407264992227768	-0.05535888615888 6…	…	0.061434627070469434	0.08977096362267186	0.07613119220133417	-13.08
11	-3.7764347826086957	-0.10416217933853829	0.07367137134759187	0.18342240982852617	…	0.04691991417919546	0.0842402263416545	0.07204996928117939	-19.3
12	3.520606060606606	0.2100927168344198	0.18093850718864263	-0.055756935746621 55	…	0.06861218940928169	-0.0077779324699063…	0.03938419688864 1266	45.96
13	-0.44719626168224432	-0.04059024910542247	0.05239493131063208	-0.021405589594085 01	…	-0.01592269770570314	0.07835466119365161	-0.03261512269386271	4.923333333333334
14	-2.001339285714285 8	0.07316068336385598	-0.05005799448684 42…	-0.022370491099405826	…	0.03994954500485442 6	0.07726930817610073	-0.03379195482260548	-1.7650000000000006
15	-2.209239313043478 27	-0.00514496682415 5…	-0.05662676811376 70…	0.08211320367634989 29	…	-0.0502924418318254	0.05335375404928863	0.06826877209785106	5.020000000000005
16	-5.104608695652175	-0.18033826233814361	0.01292838487297233 4	-0.0595786342858545…	…	0.09010858405407832	0.03176688704763877	-0.00866681223794503	20.44
17	4.781382978723406	0.06598824767680733	0.16312028213207262	-0.07717668773888778	…	0.03365020383722565	0.0891822301972314	0.09128568507993609	-13.65
18	0.09867256637168122	-0.00259735235966 6…	0.03803206675119623 3	0.05646905735349664 5	…	-0.0561575135985831…	0.04528215540037215	-0.07462724619622213	
19	2.0930851063829783	-0.01148322710644 8…	-0.02934949084170585	-0.12917833341526482	…	0.09808158720020972	-0.06237591444411659	-0.14364666680662121	
20	0.6865853658536589	-0.10475527495408 83	-0.02643816057244126	0.05332479169556990 4	…	0.03488868009314119	0.08472395769024974 1	-0.09954026202597237	
.	.	.	.	.		.	.	.	.
.	.	.	.	.		.	.	.	.
.	.	.	.	.		.	.	.	.
981	6.274818181818182	-0.05899642646971017	-0.15701241468368753	-0.01428184948352 1…	…	0.005729976703744055	0.07811990462155151	0.08303317574310037	35.695
982	4.146236559139785	-0.00621086647346 6…	-0.12709462668479013	0.07320853030092017	…	0.1600774813251191	0.0847571846019098	0.06076915535876202 5	45.92
983	-2.2424489795918374	0.1292289729655479	-0.03578956372968345	-0.11036742167256543	…	0.09850752882287772	0.08381679949749958	0.1301149031511466	26.12
984	-4.861891891891899	-0.01495983417621 3…	-0.06395120195650242	0.10667740909053837	…	0.04613073286371114	0.18371431376077715	-0.10319072996226 13	9.035
985	-1.8353846153846156	0.003296385458702291	0.09753330664989428	0.14356711686992685	…	0.1461250838808932	0.04264527854992811	0.04525775774073553 54	45.31
986	-4.5237272727272274	0.021925843797540274	0.01824524486002695 3	-0.02690612438702 8…	…	0.007001466575995972	0.05016651993445028	0.0948050401255 1043	8.267499999999998
987	0.9406930693069931	-0.17302650062481123	-0.12358152747873527	-0.04427029467290464	…	0.003127898452236 7…	-0.0578733656777165…	0.05902411002970675	7.306666666666666
988	-0.9697247706422023	-0.02859980578515256	-0.03257230155877223	0.02456938782219897 4	…	-0.04804684167358322	0.08498042434439989	-0.08818101494486 37	-17.515
989	3.5681886792452837	0.1939678855708561 7	0.09489573405199327	-0.1713985185185183	…	0.08176643332483276	0.0672998507794916	-0.09539878056797271	6.27
990	-4.490602409638554	0.012684438773310816	0.13265936665410001	-0.16057133716685 68	…	0.03820050655322511	-0.01058317870572 6…	-0.02406956821144 2…	2.5549999999999997
991	-2.1278494623655992	-0.06413519889512372	0.03634113253189271	0.05266669924101 76	…	0.11720178216846 89	0.00435620669368 53…	-0.00249060365395 3…	
992	-0.5800917431192666	-0.02636662790196839	0.06159776805763601	0.07686053872739868	…	-0.11079811859732074	0.07561719025622922	-0.00243283753235 2…	-46.48
993	3.4088999025688072	-0.03593746508409 87…	-0.03272355366839524	-0.05592654793368 5…	…	-0.00509854754816 9…	-0.03095633874322 4…	-0.10192112685241857	
994	1.0146872180451113	0.06102599214551422	-0.02018151491365789	0.09304260496105 29	…	0.09056344801167159	0.10671040530775777	-0.09901044093960255	
995	-2.299230769230777	0.28774707867194425	0.04415908120009207	-0.05555994590366952	…	-0.0670810119571237	-0.05968745899488238	0.01184777014386458 7	
996	1.4535245901639349	0.02414991936423966	0.04559683800928681	-0.13891605347874647	…	-0.08022091745526695	-0.0321399521138 77…	0.04877205139074573	40.25
997	-2.591138211382114	-0.08115191341321969	-0.06748481691747602	-0.12257004855989407	…	-0.07479163742613288	0.09856108471698426	0.02814124843597 0503	

Josef Machytka

DuckDB Performance Problems with Inappropriate Pivoting Queries on Very Large Datasets

The DuckDB documentation clearly states that this tool is designed for handling datasets fitting into memory. I fully understand that I'm…

Jan 3 6



Josef Machytka

Quick and Easy Data Exports to Parquet Format Using DuckDB

The Parquet format has become almost an industry standard for Data Lakes and Data Lakehouses, thanks to its efficiency and compact storage...

Dec 5, 2024 17

```
ate the table
 TABLE special_data_types (
 INT AUTO_INCREMENT PRIMARY KEY,
me VARCHAR(50) NOT NULL,
atus ENUM('active', 'inactive', 'pending') NOT NULL,
rmissions SET('read', 'write', 'execute') NOT NULL,
all_number TINYINT NOT NULL,
dium_number MEDIUMINT NOT NULL,
scription TEXT,
ta BLOB,
eated_at DATE NOT NULL


ert 10 rows of data
 INTO special_data_types (name, status, permissions, small_number, medium_number, description, data, created_at)
e', 'active', 'read,write', 5, 1000, 'Alice description', 'Alice data', '2023-01-01'),
, 'inactive', 'read', 10, 2000, 'Bob description', 'Bob data', '2023-02-01'),
lie', 'pending', 'write,execute', 15, 3000, 'Charlie description', 'Charlie data', '2023-03-01'),
d', 'active', 'read,write,execute', 20, 4000, 'David description', 'David data', '2023-04-01'),
, 'inactive', 'execute', 25, 5000, 'Eve description', 'Eve data', '2023-05-01'),
k', 'pending', 'read,write', 30, 6000, 'Frank description', 'Frank data', '2023-06-01'),
e', 'active', 'read', 35, 7000, 'Grace description', 'Grace data', '2023-07-01'),
', 'inactive', 'write,execute', 40, 8000, 'Hank description', 'Hank data', '2023-08-01'),
, 'pending', 'read,write,execute', 45, 9000, 'Ivy description', 'Ivy data', '2023-09-01'),
'. 'active', 'execute', 50, 10000, 'Jack description', 'Jack data', '2023-10-01');
```

 Josef Machytka

DuckDB as a Rudimentary Data Migration Tool

After exploring how to use DuckDB as an intelligent ETL tool for PostgreSQL, and how to extend its ETL capabilities with simple Python...

Nov 30, 2024 26 2

See all from Josef Machytka

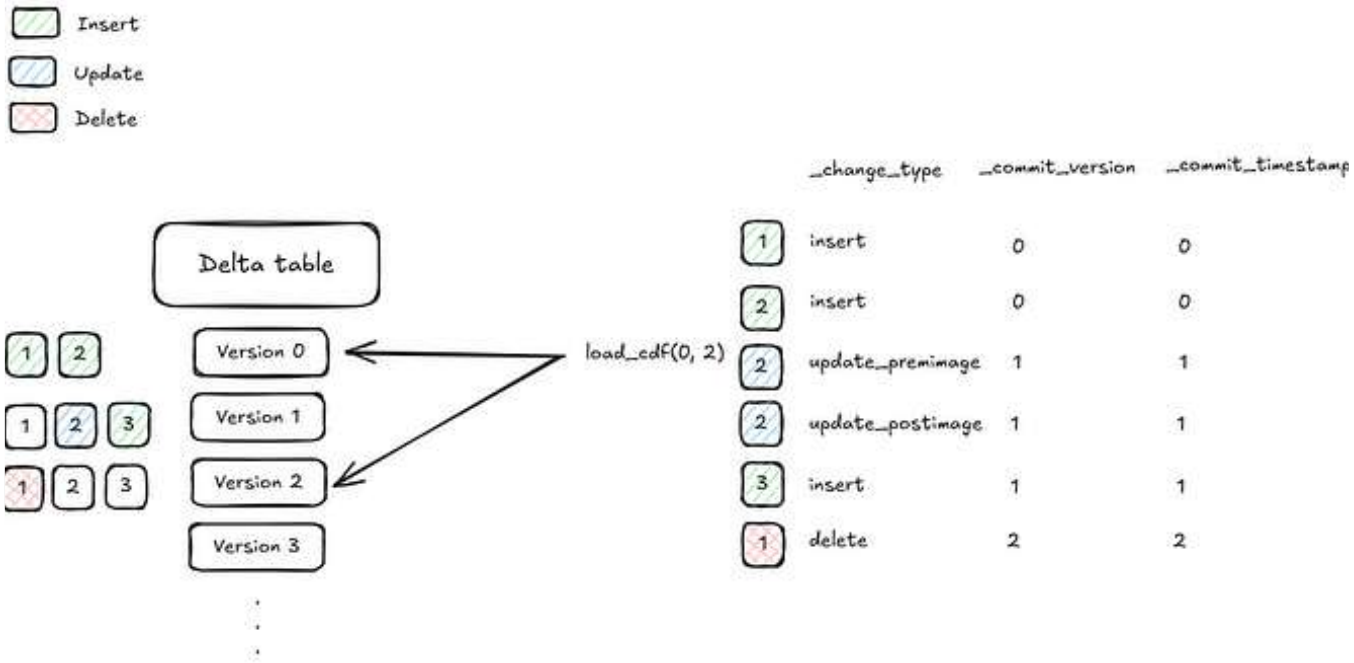
Recommended from Medium

In Python in Plain English by Kiran Maan

How I Speed Up My Python Scripts by 300%

You won't regret learning this.

6d ago 783 6



Pablo Cabeza

Incremental pipelines with Delta and Polars

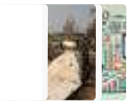
How to apply watermark pipelines at a small scale with Change Data Feed

Nov 29, 2024 2

Lists

Practical Guides to Machine Learning
10 stories · 2155 saves

ChatGPT prompts
51 stories · 2472 saves



Staff picks
800 stories · 1569 saves

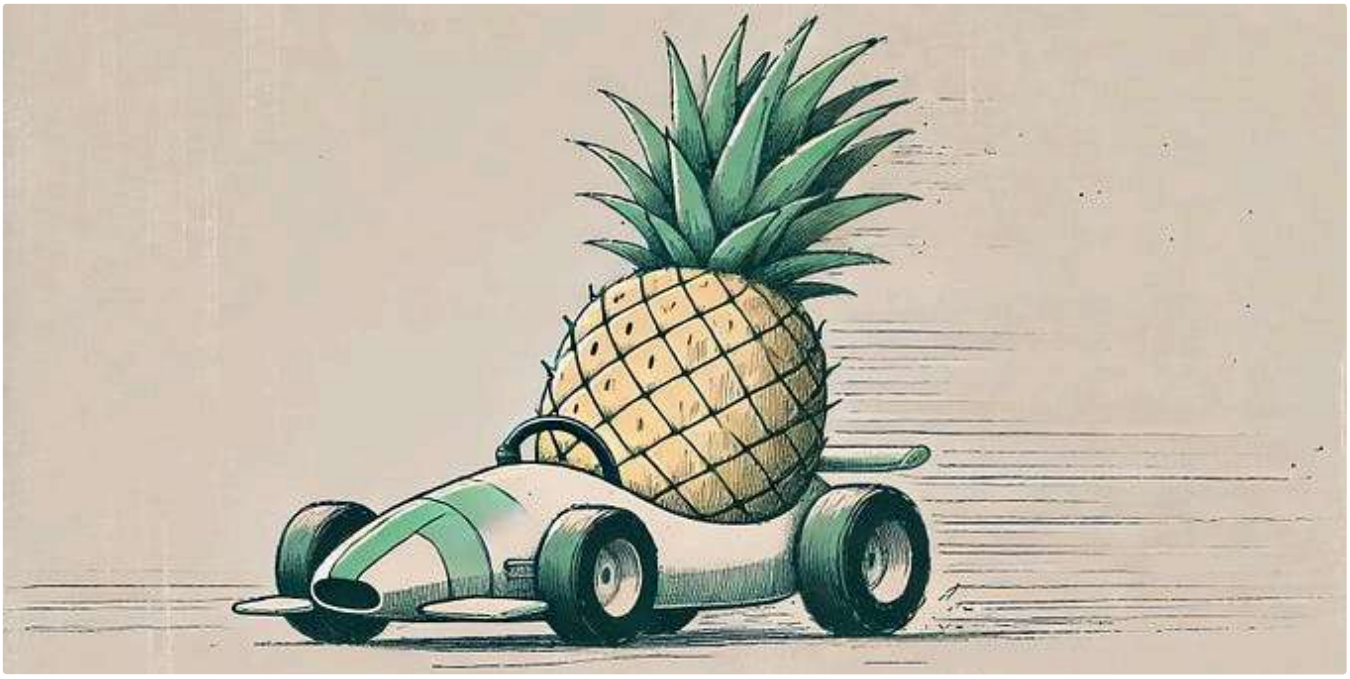
Natural Language Processing
1889 stories · 1544 saves

Varun Singh

Python 3.14 Released — Top 5 Features You Must Know

Faster Annotations & Mind-Blowing Updates You NEED to Know!

Dec 31, 2024 1K 5

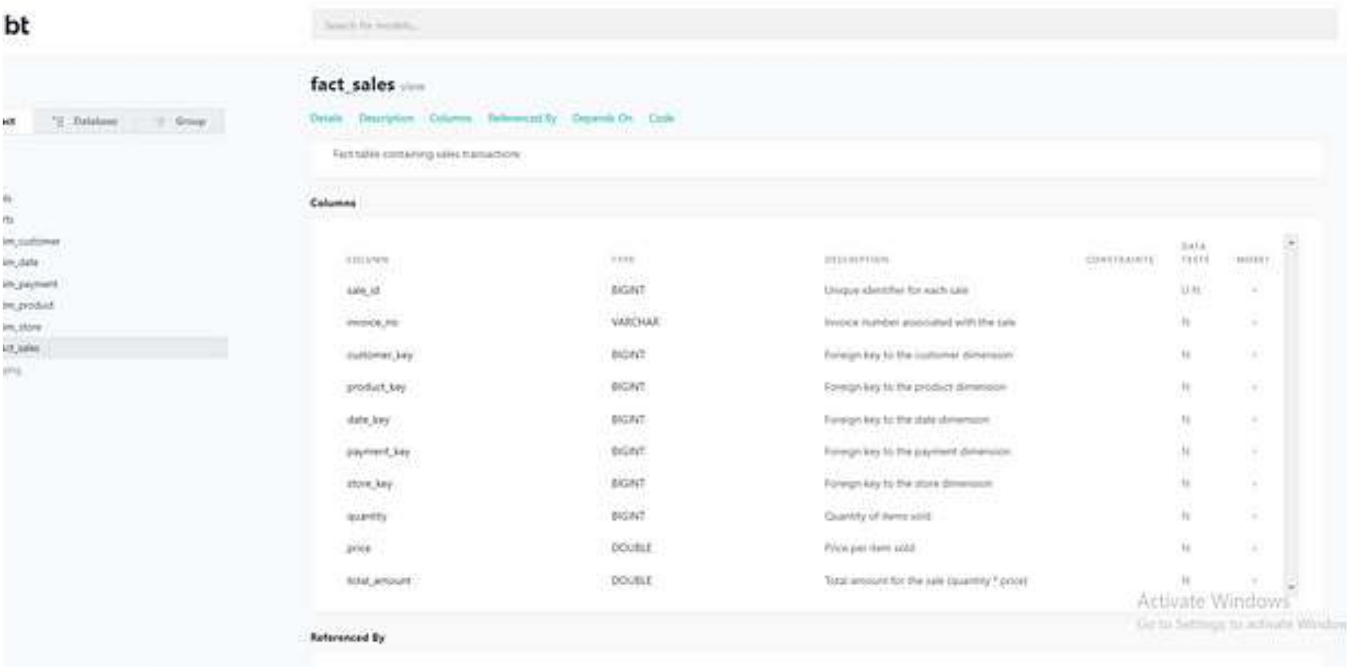


Dor Yardeni - Tech Pineapple

Should We Stop Using Spark DataFrame API?

Benchmarking Spark SQL vs. DataFrame API: What the Docs Won't Tell You

Jan 6 77 10



In Towards Dev by Wissem Hammoudi

Building an Efficient Data Pipeline: Setting Up dbt Core with DuckDB

A Comprehensive Guide to Setting Up Your Data Warehouse Project Using dbt Core and DuckDB

Nov 28, 2024 97

In Towards Data Science by Jiayan Yin

Advanced SQL Techniques for Unstructured Data Handling

Everything you need to know to get started with text mining