

[Open in app](#)

Medium

 Search

Building a Data Lakehouse with PostgreSQL: Dive into Formats, Tools, Techniques, and Strategies (Josef Machytka: Speaker portfolio)



Josef Machytka

2 min read · Feb 9, 2025



Listen



Share



More

(See my bio, other talks in my portfolio and my speaker experience [in the covering article](#).)

Duration: 45 minutes

Target Audience: Application Developers, Database Administrators, Data Analysts

Overview:

The evolution of Data Warehouses, Data Lakes, and Data Lakehouses has been marked by many buzzwords, fluctuating trends, and tools that often over-promised but under-delivered. While there are numerous materials on these topics, most of them provide mostly introductory overviews and focus narrowly on a single technology. And there are even many different opinions about what exactly is Data Lakehouse.

This talk discusses different ways how to understand this topic. It explores data formats and frameworks like Parquet, Apache Iceberg, Delta Lake, Apache Hudi. Discusses different architectures of Data Lakehouse solutions. Also key challenges will be addressed, such as effective Data Governance, compliance with privacy and security standards, and comprehensive data quality checks.

Last part of the talk address current AI hype with its many promises and proposes realistic overview of real capabilities of current Large Language Models and their use cases in Data Lakehouses.

PostgreSQL is extremely well equipped to play a major role in the current Data Lakehouse and AI boom.

Key Takeaways:

- A comprehensive overview of Data Lakehouse architecture
- Insights into key data formats and frameworks in modern Data Lakehouses
- Practical ideas for implementing Data Governance practices
- Realistic view of real capabilities of current LLMs in scope of Data Lakehouses

Slides: [on my GitHub](#)

Presented at:

- [Prague PostgreSQL Developer Day 2025](#)
- NetApp Internal Talk 2025.01.23



© organizers of Prague PostgreSQL Developer Day 2025



© organizers of Prague PostgreSQL Developer Day 2025

Postgresql

Data Lake

Data Lakehouse

Data Analysis

[Edit profile](#)

Written by Josef Machytka

68 Followers · 25 Following

I work as PostgreSQL specialist & database reliability engineer at credativ GmbH.

No responses yet



Josef Machytka he/him

What are your thoughts?

More from Josef Machytka

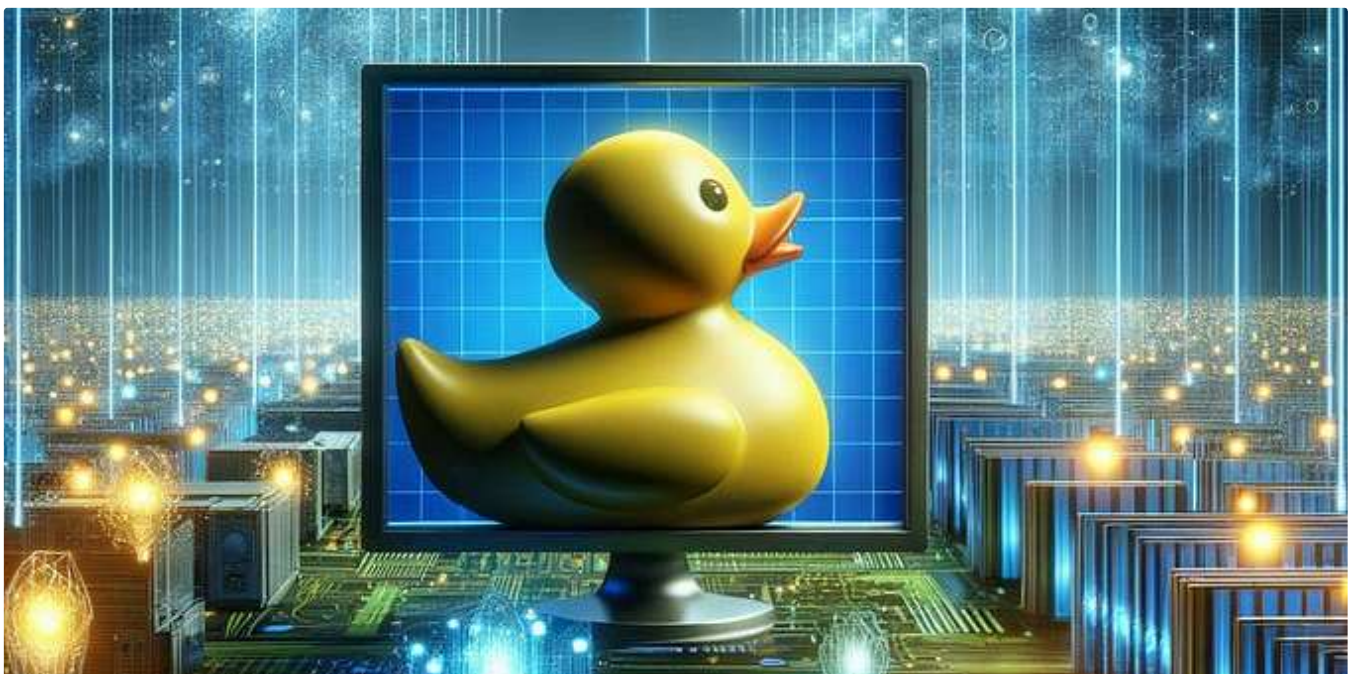


 Josef Machytka

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  24  3



 Josef Machytka

Quick and Easy Statistics and Histograms with DuckDB

DuckDB is an exceptional tool that demonstrates how tasks requiring sometimes considerable manual effort in other tools can be accomplished...



 Josef Machytka

PostgreSQL JSONB Operator Classes of GIN Indexes and Their Usage

Throughout 2024, I worked on an internal project exploring the use of JSONB data in PostgreSQL and its various indexing options. During...

Bob	2100.0	600.0				
Charlie	2300.0	1500.0	1100.0			

D pivot pg.sales on (product,year) using sum(sales_amount) group by salesperson order by salesperson;

salesperson varchar	(Laptop, 2022) double	(Laptop, 2023) double	(Phone, 2022) double	(Phone, 2023) double	(Tablet, 2022) double	(Tablet, 2023) double
Alice	1200.0	1400.0	800.0	900.0	300.0	400.0
Bob	1000.0	1100.0	600.0			
Charlie	1100.0	1200.0	700.0	800.0	500.0	600.0

D pivot pg.sales on (year,product) using sum(sales_amount) group by salesperson order by salesperson;

salesperson varchar	(2022, Laptop) double	(2022, Phone) double	(2022, Tablet) double	(2023, Laptop) double	(2023, Phone) double	(2023, Tablet) double
Alice	1200.0	800.0	300.0	1400.0	900.0	400.0
Bob	1000.0	600.0		1100.0		
Charlie	1100.0	700.0	500.0	1200.0	800.0	600.0

D pivot pg.sales on (year) using sum(sales_amount) group by salesperson order by salesperson;

--	--	--	--	--	--	--

 Josef Machytka

Easy and Intelligent Pivot Tables with DuckDB

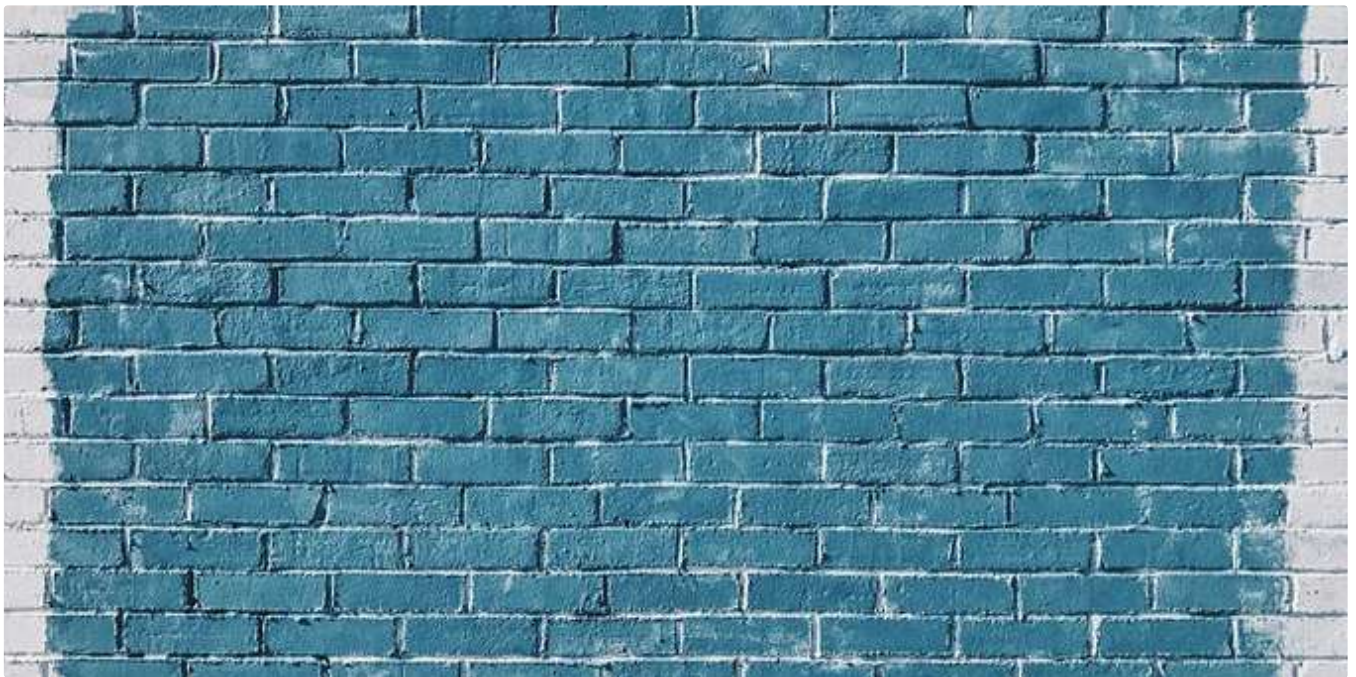
After exploring the various capabilities of DuckDB in my earlier articles, I want to focus more on its powerful data analytical...

Dec 4, 2024 🖱 7 💬 1



See all from Josef Machytka

Recommended from Medium



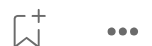
Vijay Gadhave

Delta Lake 4.0: Next-Level Big Data Management

Note: If you're not a medium member, [CLICK HERE](#)



Feb 21 🖱 16 💬 1





In Timescale by Team Timescale

Data Visualization in PostgreSQL With Apache Superset

Looking for data visualization tools for Postgres? We discuss a few options & provide a step-by-step guide on PostgreSQL and Apache...

Jan 24 🖱️ 23

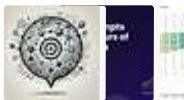


Lists



Practical Guides to Machine Learning

10 stories · 2215 saves



ChatGPT prompts

51 stories · 2605 saves



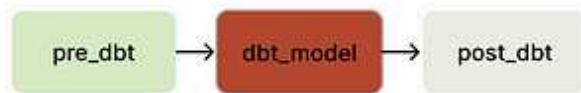
Staff picks

819 stories · 1637 saves

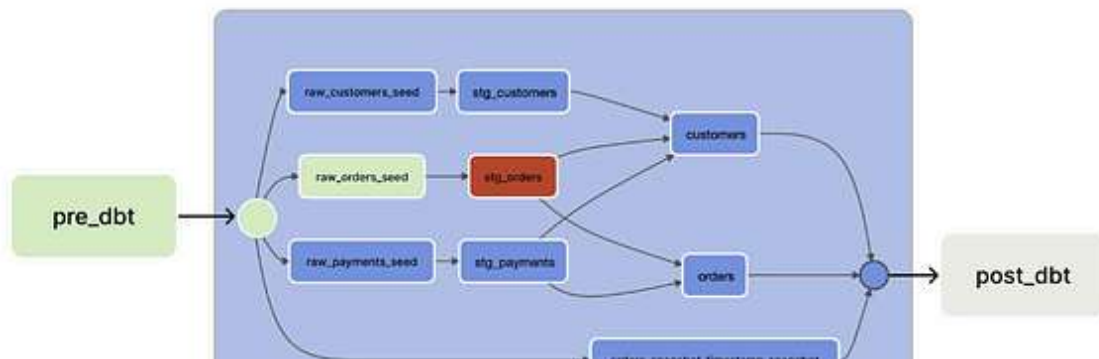



Natural Language Processing

1958 stories · 1605 saves



With Cosmos



 In Apache Airflow by Arjun Anandkumar

Running dbt models on Airflow (MWAA)

Premise

Feb 21  5

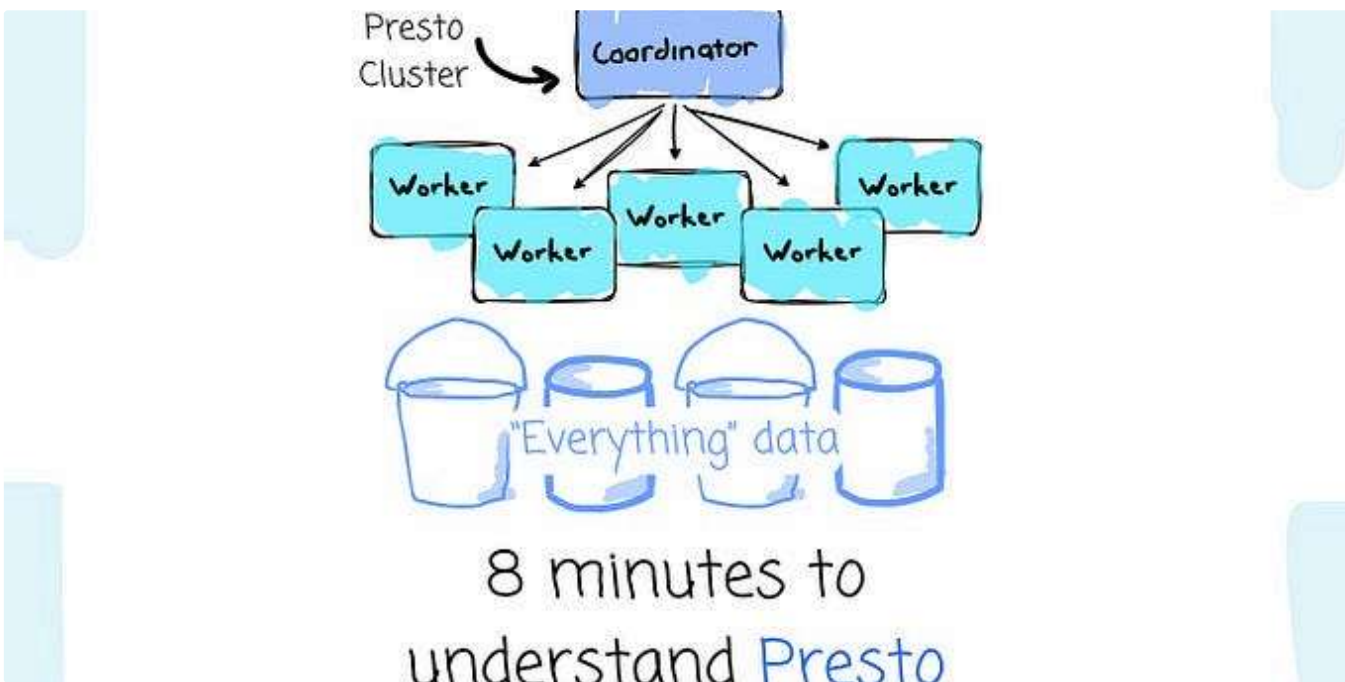


 In Dev Genius by Shashank Mayya

Python, Pandas, and Iceberg: A Seamless Data Engineering Workflow with Trino

The modern data landscape is increasingly demanding. Data engineers need tools that offer flexibility, performance, and robust data...

Feb 19 149

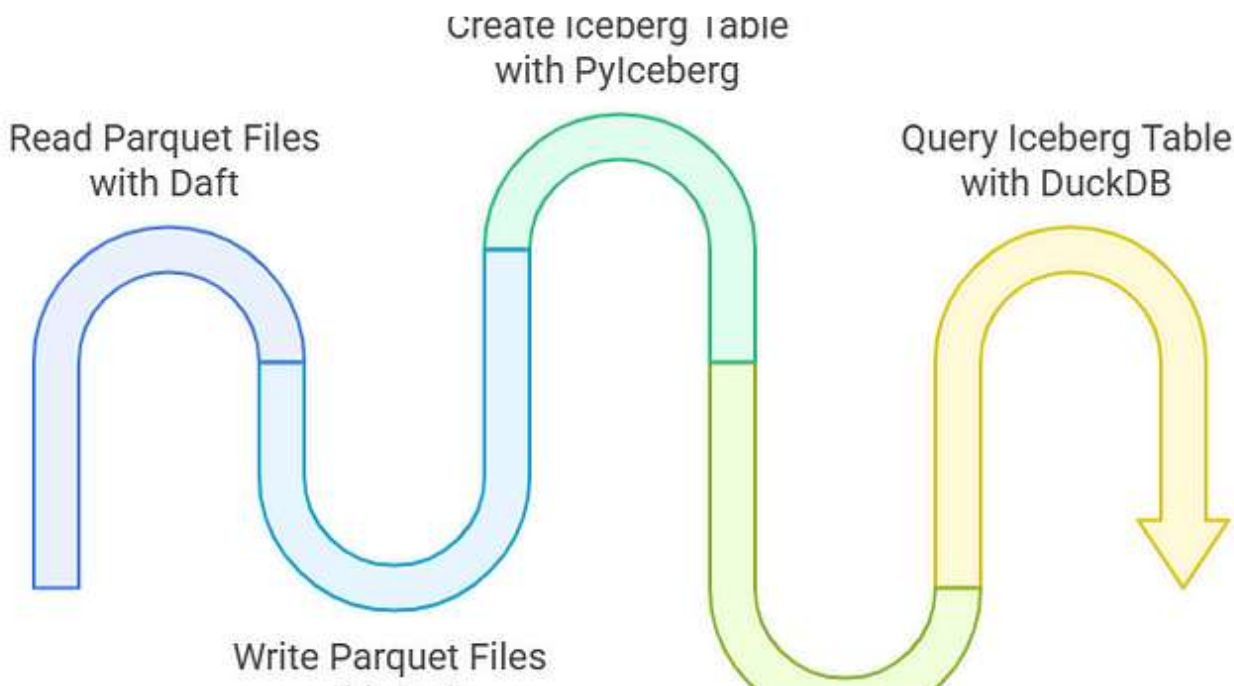


In Data Engineer Things by Vu Trinh

8 minutes to understand Presto

Uber, Netflix, Airbnb, and LinkedIn use this query engine.

Feb 20 199 3



In Towards Dev by Ashkan Golehpour

Building an Apache Iceberg Table from Parquet with Daft/Polars, Pylceberg (SQLite Catalog), and...

This article will demonstrate a fully Python-based workflow to convert a pure Parquet dataset into an Apache Iceberg table. We leverage...

Feb 16  3



See more recommendations