Josef Machytka: Speaker Portfolio

Josef Machytka
8 min read · 7 hours ago

Expert Talks on PostgreSQL, Databases, Data Ingestion and Data Analysis

Introduction

With over 30 years of experience in database technologies, I have cultivated a deep understanding of diverse database systems, including extensive work with PostgreSQL. I worked with multiple heterogeneou sdata ingestion and data processing pipelines, focusing on delivery of valuable business results. The need to solve often very complicated technical problems equipped me with the expertise necessary for insightful and impactful talks. I am passionate about sharing my knowledge, empowering others in the field of databases and data processing and making complex topics accessible.

My talks are living and evolving, I steadily improve them with new content reflecting progress in new versions of the corresponding software. My speaking style combines practical insights with important technical details, ensuring that audiences walk away with really valuable knowledge they can

apply. Slides contain a lot of additional information, making them valuable resources even outside the context of my presentations.

Thank you for considering me as a speaker for your next event. I look forward to the opportunity to share my knowledge and help your audience gain deeper insights into PostgreSQL and database topics.

About Me

Name: Josef Machytka
My current job: Professional Service Consultant, PostgreSQL specialist at NetApp Open Source Systems

Experience:
* 30+ years of production experience with different databases: PostgreSQL 12 years, BigQuery 7y, Oracle 15y, MySQL 12y, Elasticsearch 5y, MS SQL 5y, Sybase ASE, FoxPro
* 10+ years of experience with high volume and velocity data ingestion pipelines, Data Analysis, Data Warehouse and Data Lakehouse
* 1.5+ years of practical experience with different LLMs including their architecture and principles

Accounts: LinkedIn, ResearchGate.net, Academia.edu
Mastodon: @JosefMachytka@me.dm

Current Talk Portfolio

- GIN, BTREE_GIN, GIST, BTREE_GIST, HASH and BTREE indexes on JSONB data
- PostgreSQL and DuckDB: Analytics Performance and Use Cases

- Partitioning and Clustering: An Overview of Solutions with a deep dive into PostgreSQL implementation
- Blending AI and Human Expertise in PostgreSQL: Lessons from Two Years of AI-Assisted Troubleshooting and Development
- Building a Data Lakehouse with PostgreSQL: Dive into Formats, Tools, Techniques, and Strategies
- Understanding Statistics in PostgreSQL: Beyond Obsession with Indexes
- PostgreSQL AI Muscles: Comparison of Vector Databases for RAG-Powered Applications

GIN, BTREE_GIN, GIST, BTREE_GIST, HASH and BTREE indexes on JSONB data

Duration: 45 minutes
Target Audience: Application developers, data analysts

Overview: Talk summarizes several months long and still ongoing internal project testing usage and performance of GIN and BTREE_GIN with different operator classes, GIST and BTREE_GIST indexes for GeoJSON data and also standard HASH and BTREE indexes specifically on JSONB data. Tested on several real life datasets with a total size of dozens of GBs. Also, the influence of TOAST compression algorithms, parallelism, memory settings, table statistics on processing JSONB data was tested. Objective of this project was to gather relevant experience to be able to help our customers with their problems, because the majority of articles on the web about JSONB data in PostgreSQL show only trivial examples without any reasonable value for developers solving multiple performance issues related to JSONB data. The talk also discusses practical limitations developers would face if they try to fully decompose JSONB data into relational tables.

Key Takeaways:

- Understanding of use cases and performance of different types of indexes for JSONB data
- The impact of system settings like TOAST compression, parallelism, and memory on performance and usage of indexes
- Limitations and considerations for decomposing JSONB data into relational structures

Slides: available on academia.edu

Presented at:

- Prague PostgreSQL Developer Day 2024 (article on NetApp blog)
- Swiss PG day 2024 (article on NetApp blog)
- Berlin PostgreSQL MeetUp October 2024 (MeetUp entry)

PostgreSQL and DuckDB: Analytics Performance and Use Cases

Duration: 45 minutes
Target Audience: Data Analysts, Data Scientists, App developers

Overview: This talk explores the capabilities and features of PostgreSQL and DuckDB in the context of Data Analytics. It highlights key differences, focusing on the strengths and weaknesses of each database for analytical workloads. The presentation also includes benchmark results from some typical business use cases, providing real-world performance comparisons.

We discuss how to combine PostgreSQL and DuckDB effectively in data analytics pipelines and ETL processes to leverage the best of both worlds.

Key Takeaways:

- Small Data Manifesto as the new trend in Data Analysis
- Understanding of of the main features of DuckDB, its strengths and weaknesses
- Usage of DuckDB as an very efficient ETL tool for PostgreSQL
- Optimization of analytical workload by combining PostgreSQL and DuckDB

Slides: older version available on academia.edu

Presented at:

- Prague PostgreSQL MeetUp October 2024

Partitioning and Clustering: An Overview of Solutions with a deep dive into PostgreSQL implementation

Duration: 45 minutes
Target Audience: App developers, system architects

Overview: In this presentation, we will examine the implementation of partitioning and clustering in several database systems, such as BigQuery, Snowflake, Oracle and MySQL. Following that, we will discuss a detailed analysis of PostgreSQL's approach to inheritance, partitioning, and clustering. We will check database parameters that affect performance of

these solutions, compare the results of performance tests between a single large table and partitioned tables on different datasets, look at efficiency of indexes, and discuss the application of multi-level partitioning for different use cases. Additionally, we will delve into details of memory usage, statistics and query optimization.

Key Takeaways:

- Understanding of partitioning & clustering across different database systems
- In-depth understanding of PostgreSQL implementation
- Practical tips and tricks for performance optimization

Slides: not available online yet

Presented at: presented so far only internally

Blending AI and Human Expertise in PostgreSQL: Lessons from Two Years of AI-Assisted Troubleshooting and Development

Duration: 45 minutes
Target Audience: App developers, database administrators

Overview: In this talk, we will share our real-world experiences from nearly two years of using various AI models to troubleshoot issues and develop solutions in PostgreSQL. We will delve into the phenomenon of AI hallucinations, explaining their causes and discussing prompting techniques to minimize their occurrence. The presentation will highlight both the strengths and limitations of using AI for PostgreSQL-related problem-solving. The speaker has deeper knowledge of LLM architecture and

underlying principles, he delivered already several talks on this topic. He works with a variety of commercial, internal, and open-source AI models.

Key Takeaways:

- Strength and limitations of AI for PostgreSQL troubleshooting
- Improving veracity and quality of AI answers by different prompt engineering techniques
- Typical problems encountered in everyday work with different LLMs

Slides: not available online yet

Presented at: presented so far only internally

Building a Data Lakehouse with PostgreSQL: Dive into Formats, Tools, Techniques, and Strategies

Duration: 45 minutes
Target Audience: App developers, database administrators, data analysts

Overview: This talk shares practical insights from building a Data Lakehouse architecture using PostgreSQL, BigQuery, and Google Cloud Storage. Provides explanations of various data formats such as Parquet or HDF5, and frameworks like Apache Iceberg, Delta Lake, or Apache Hudi. Explains how to effectively combine relational and non-relational data in PostgreSQL and discusses optimization techniques for scaling PostgreSQL to handle large datasets. It dives deeper into the problem of maintaining robust Data Governance, ensuring compliance with privacy and security standards, and implementing proper data quality checks, data cleansing and data

transformation processes. Do not allow your Data Lake to turn into a dark swamp full of digital monsters everyone would fear!

Key Takeaways:

- Comprehensive Overview of Data Lakehouse Architecture
- Understanding of Data Lakehouse Formats and Frameworks
- Practical experience with implementing robust Data Governance

Slides: not available online yet

Presented at: presented so far only internally

Understanding Statistics in PostgreSQL: Beyond Obsession with Indexes

Duration: 45 minutes
Target Audience: App developers, database administrators, system architects

Overview: Statistics in PostgreSQL are fundamental to query optimization and performance tuning. They guide the planner in deciding whether to utilise indexes or perform a sequential table scan. However, some developers are fixated on forcing PostgreSQL to use indexes, even when it's inefficient. This talk will explore practical examples using publicly available datasets to demonstrate how statistics influence execution plan selection. We will dive into the settings that affect the precision of gathered statistics and how they impact the planner's decisions. Goal of this session is to give attendees a comprehensive understanding of PostgreSQL statistics, and help them to understand the reasons behind the planner's chosen execution plans

Key Takeaways:

- Understanding of values collected in statistics
- Insight into conditions under which planner uses indexes
- Influence of different settings on the precision of statistics and chosen execution plan

Slides: not available online yet

Presented at: presented so far only internally

PostgreSQL AI Muscles: Comparison of Vector Databases for RAG-Powered Applications

Duration: 45 minutes
Target Audience: App developers, database administrators, system architects

Overview: The talk focuses on the growing importance of vector databases in AI applications, specifically for Retrieval-Augmented Generation (RAG). The integration of vector-based search, where AI models can retrieve contextually relevant information, has become crucial for advanced tasks like question-answering, recommendation engines, and other large language model (LLM) powered applications. PostgreSQL has entered this space through the pgvector extension, offering vector similarity search capabilities. This presentation will compare PostgreSQL pgvector with specialised vector databases, highlighting key features, performance metrics, scalability, and ease of use for AI workloads.

Key Takeaways:

- Understanding the Role of Vector Databases in AI
- Comparative Insights into PostgreSQL pgvector Extension and Specialized Vector Databases

Slides: not available online yet

Presented at: presented so far only internally

My latest speaking engagements

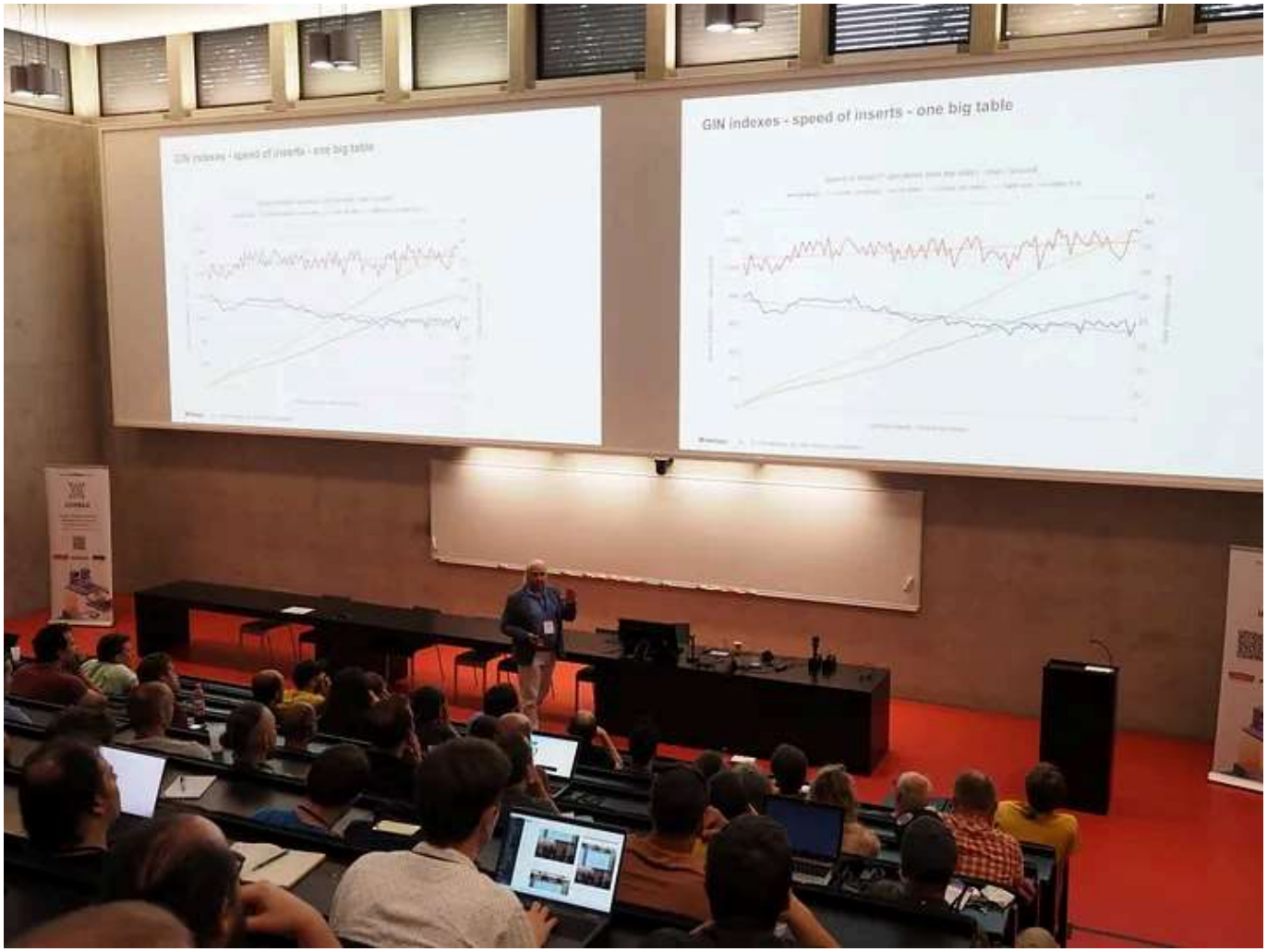
1. **2024.10.29 Prague PostgreSQL MeetUp October 2024**
   "PostgreSQL and DuckDB"
2. **2024.10.16 NetApp internal workshop** — "AI Workshop: Exploring Artificial Intelligence" — 3 talks:
   - "The AI Dilemmas: The Philosophical, Ethical, and Legal Implications of AI"
   - "Understanding the Magic of LLMs: A Deep Dive into the Internal Structure of LLMs"
   - "Many facets of AI hallucinations: factual errors, deep fakes, and creativity"
3. **2024.10.09 PostgreSQL MeetUp Berlin October 2024**
   - "GIN, BTREE_GIN, GIST, BTREE_GIST, HASH and BTREE indexes on JSONB data"
4. **2024.09.27 NetApp internal talk**
   - "Beyond the Buzzwords: A pragmatic explanation of LLM terminology"
5. **2024.10.13 NetApp internal talk**
   - "The Art and Science of AI Prompt Engineering"

6. **2024.07.16 PostgreSQL MeetUp Berlin July 2024**
   - Lightning talk: “Can PostgreSQL have a more prominent role in the AI boom?”

7. **2024.06.27 Swiss PG day 2024**
   - “GIN, BTREE_GIN, GIST, BTREE_GIST, HASH and BTREE indexes on JSONB data”

8. **2024.06.05 Prague PostgreSQL Developer Day 2024**
   - “GIN, BTREE_GIN, GIST, BTREE_GIST, HASH and BTREE indexes on JSONB data”

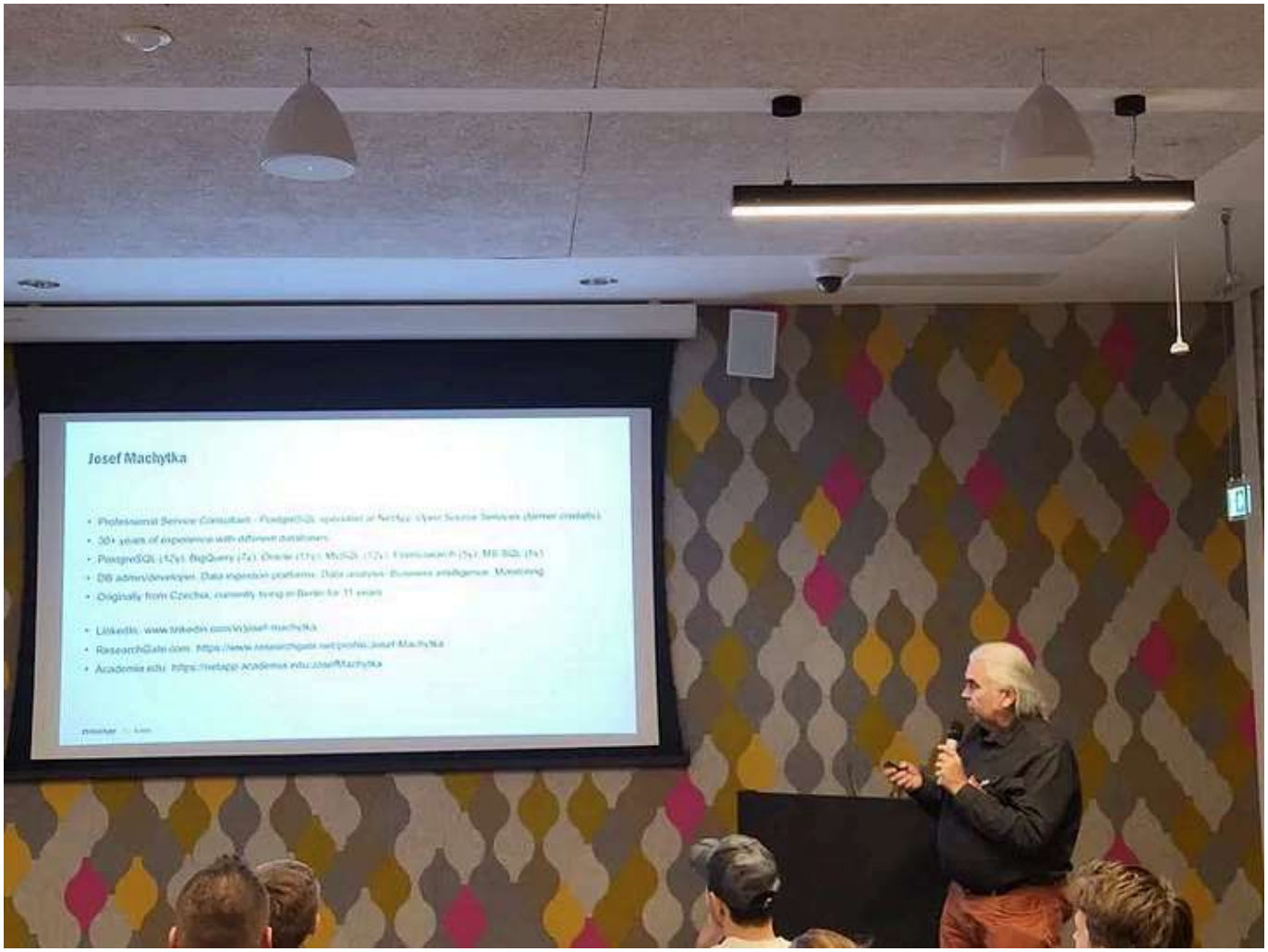
Photos from my talks



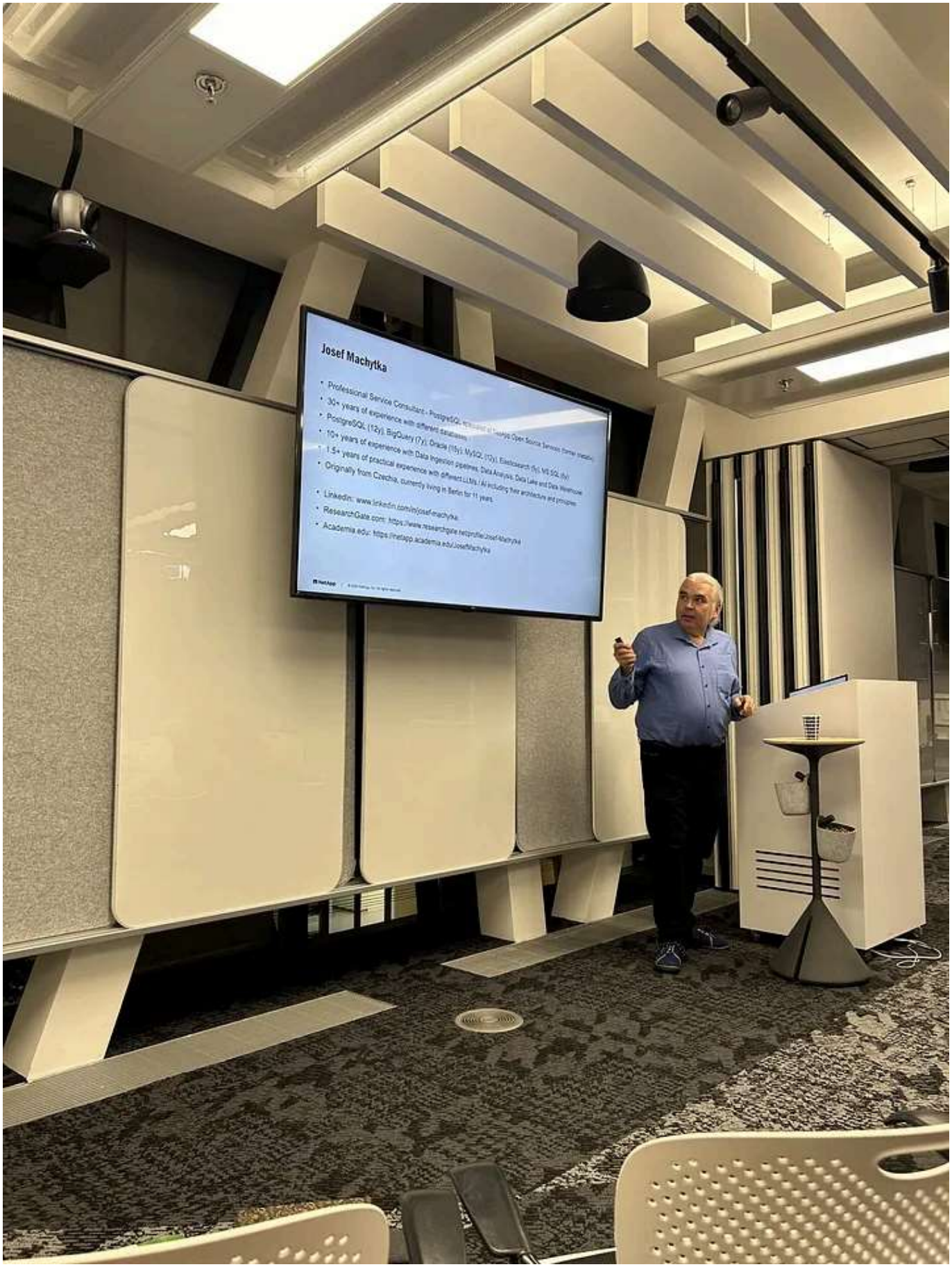
© Tomas Vondra P2D2 — Prague PostgreSQL Developer Day 2024 — talk about indexes on JSONB data

© organizers of Swiss PG day — Swiss PG day 2024 — talk about indexes on JSONB data

© Andreas Scherbaum — PostgreSQL Meetup Berlin July 2024 — lightning talk about PostgreSQL and AI



© Andreas Scherbaum — PostgreSQL Meetup Berlin Octobec 2024 — talk about indexes on JSONB data



© Igor Gavrilov (LinkedIn post) — Prague PostgreSQL Meetup October 2024 — talk about DuckDB

Postgresql Data Analysis Data Ingestion Duckdb

Written by Josef Machytka

Edit profile

0 Followers

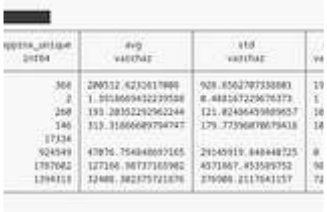
I work as Professional Service Consultant - PostgreSQL specialist in NetApp Deutschland GmbH, Open Source Services division.

More from Josef Machytka

Josef Machytka

Using DuckDB as an Intelligent ETL tool for PostgreSQL

There is a lot of hype around DuckDB these days. At one PostgreSQL conference, I even saw a large poster comparing...



1d ago

See all from Josef Machytka

Recommended from Medium

Abdur Rahman in Stackademic

Python is No More The King of Data Science

5 Reasons Why Python is Losing Its Crown

Oct 23 2K 16

F. Perry Wilson, MD MSCE

How Old Is Your Body? Stand On One Leg and Find Out

According to new research, the time you can stand on one leg is the best marker of...

Oct 23 7.2K 167

Lists

Practical Guides to Machine Learning

10 stories · 1995 saves

ChatGPT prompts

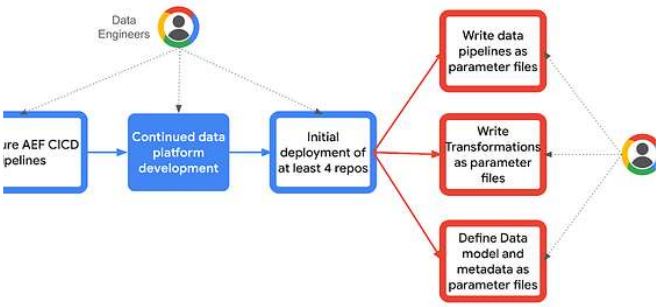
50 stories · 2169 saves

Staff Picks

755 stories · 1417 saves

Natural Language Processing

1789 stories · 1394 saves

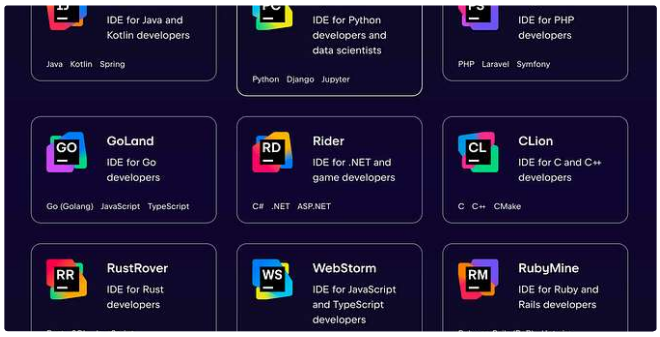


Oscar Pulido in Google Cloud - Community

Stop Thinking in Data Pipelines, Think in Data Platforms:...

Imagine a world where you could deploy your entire enterprise-ready data platform in...

5d ago 214 2



Saeed Zarinfam in ITNEXT

Reasons behind the recent changes in JetBrains products...

VS Code is getting popular and powerful, and Fleet is getting late!

6d ago 382 9

Salvatore Raieli in Towards Data Science

The Savant Syndrome: Is Pattern Recognition Equivalent to...

Exploring the limits of artificial intelligence: why mastering patterns may not equal...

2d ago 1.2K 28

Ignacio de Gregorio

Apple Speaks the Truth About AI. It's Not Good.

Are We Being Lied To?

Oct 23 4.8K 148

See more recommendations