Vector Databases: What's Really New and What's Next? (VLDB 2024 Panel)

Jianguo Wang
Purdue University
csjgwang@purdue.edu

Eric Hanson
SingleStore
hanson@singlestore.com

Guoliang Li
Tsinghua University
liguoliang@tsinghua.edu.cn

Yannis Papakonstantinou
Google Cloud
yannispap@google.com

Harsha Simhadri
Microsoft
harshasi@microsoft.com

Charles Xie
Zilliz
charles@zilliz.com

ABSTRACT

Vector databases have recently emerged as a hot topic in the field of databases, especially in industry. This is due to the widespread interest in Large Language Models (LLMs), where vector databases provide the relevant context for LLMs to produce more accurate responses. However, vector data is not new. It has been studied for more than two decades, leading to many efficient algorithms and indexes for vector similarity search. Thus, a natural question is: *What is really new and what is next for vector databases?* This panel will bring together several leading experts in vector databases to share their insights and experiences from various perspectives. The panel will also discuss the broader role of databases, beyond just vector databases, in the era of generative AI.

PVLDB Reference Format:
Jianguo Wang, Eric Hanson, Guoliang Li, Yannis Papakonstantinou, Harsha Simhadri, and Charles Xie. Vector Databases: What's Really New and What's Next? (VLDB 2024 Panel). PVLDB, 17(12): 4505-4506, 2024.
doi:10.14778/3685800.3685911

1 INTRODUCTION

1.1 Why Vector Databases?

Despite the significant success of Large Language Models (LLMs), they still face several inherent limitations, such as hallucination, a lack of domain knowledge (e.g., enterprise data), and the inability to incorporate up-to-date information. Vector databases effectively address these issues by storing the vector embeddings of domain-specific and real-time data, serving as an external knowledge base for LLMs. When a user poses a question (e.g., to ChatGPT), it is first sent to vector databases to retrieve relevant data via vector similarity search. The results are then augmented with the original question to form a prompt that provides a more comprehensive context, enabling LLMs to generate more accurate answers. This process is known as Retrieval-Augmented Generation (RAG).

As a result, vector databases have become a critical component in LLM ecosystems, which leads to an explosion of recent efforts in developing vector databases. Examples include Milvus, Pinecone, Chroma, Qdrant, Weaviate, and Vespa. Moreover, existing relational databases such as Oracle, many Google Cloud database systems (e.g., AlloyDB and Spanner), Azure SQL, Alibaba AnalyticDB, SingleStore, PostgreSQL, and MySQL now also support vector similarity search to better support LLM applications.

Proceedings of the VLDB Endowment, Vol. 17, No. 12 ISSN 2150-8097.
doi:10.14778/3685800.3685911

1.2 Questions Discussed in the Panel

The panel will discuss the following questions:

- **Q1** What is really new about vector databases (considering that vector search has been around for two decades)?
- **Q2** Does it need to be a specialized vector database, or just a vector search feature within a relational database? We may compare it with the example of XML.
- **Q3** What are the fundamental challenges in vector databases that are not yet addressed?
- **Q4** What are the unique challenges and opportunities of implementing vector search inside relational databases?
- **Q5** Will vector databases still be relevant when large language models are powerful enough?
- **Q6** What are the challenges of RAGs that can or cannot be addressed by vector databases?
- **Q7** Beyond vector databases: What is the role of databases in general in generative AI? How to build future database systems to better support generative AI?

2 BIOGRAPHIES

Our panel comprises six members with extensive experience in vector databases, representing both industry and academia. The industry members come from major tech companies (actively working on vector databases) as well as a leading vector database startup. They have worked on both specialized vector databases (e.g., Milvus) and extended vector databases (e.g., Google AlloyDB, SingleStore, and Azure SQL). Interestingly, some members have prior experience with XML, yet another non-relational data model that was very popular 20 years ago. They can compare and contrast the vector and XML data models. Overall, their diverse backgrounds uniquely position them to address the questions in the panel.

Moderator

Jianguo Wang is an Assistant Professor of Computer Science at Purdue University. He obtained his Ph.D. degree from the University of California, San Diego. He has worked or interned at Zilliz, Amazon AWS, Microsoft Research, Oracle, and Samsung on various

database systems. His current research interests include Database Systems for the Cloud and Large Language Models, including Disaggregated Databases and Vector Databases. He co-authored Milvus (SIGMOD'21) and SingleStore-V (VLDB'24), two leading vector databases in the field, where Milvus is a specialized vector database and SingleStore-V is an integrated vector database within the relational database SingleStore. He also co-authored other vector database papers published in SIGMOD'24 and ICDE'24. He is a recipient of the NSF CAREER Award.

Panelists

Eric Hanson is a Director of Product Management at SingleStore, responsible for query processing, storage and extensibility feature areas. He joined the SingleStore product management team in 2016. He is a PhD computer science graduate of UC Berkeley, was an Air Force officer, an associate professor of computer science at the University of Florida, and a principal program manager, lead program manager, and developer for SQL Server at Microsoft. At Microsoft he was a founder of the columnstore project for SQL Server. Eric was an Apache Hive committer for contributions to Stinger. He has extensive background on database triggering and alerting, data warehousing, column stores, columnar query execution, and vector database search.

Guoliang Li is a Full Professor of Department of Computer Science at Tsinghua university. His research interests include database systems, machine learning for databases, and large-scale data cleaning and integration. He got VLDB 2017 Early Research Contribution Award, TCDE 2014 Early Career Award, SIGMOD 2023 Best Papers, VLDB 2023 Best Industry Paper Runner-up, VLDB 2020 Best Papers, CIKM 2017 Best Paper Award, KDD 2018 Best Papers, ICDE 2018 Best Papers, DASFAA 2023 Best Paper Award, DASFAA 2014 Best Paper Runnerup, APWeb 2014 Best Paper Award, EDBT 2013 Similarity Join and Search Champion. He was SIGMOD 2021 general chair, VLDB 2021 Demo chair, and ICDE 2022 industry chair. He regularly served as PC Member of SIGMOD, VLDB, ICDE, KDD, WWW. He was serving as associate editor for IEEE TKDE and VLDB Journal.

Yannis Papakonstantinou is a Distinguished Engineer, working on Query Processing and GenAI, at Google Cloud. He is also an Adjunct Professor of Computer Science and Engineering at the University of California, San Diego, following many years of having been a UCSD regular faculty member. Previously, he was an architect in query processing & ETL at Databricks. Earlier, he was a Senior Principal Scientist at Amazon Web Services from 2018-2021 and was a consultant for AWS since 2016. He was the CEO and Chief Scientist of Enosys Software, which built and commercialized an early Enterprise Information Integration platform for structured and semi-structured data. The Enosys Software was OEMed and sold under the BEA Liquid Data and BEA AquaLogic brand names, eventually acquired in 2003 by BEA Systems. His R&D work has been mostly on query processing with a focus on querying semi-structured data. He has published over 120 research articles that have received over 20,000 citations. Yannis holds a Diploma of Electrical Engineering from the National Technical University of Athens, MS, and PhD in Computer Science from Stanford University (1997).

Harsha Simhadri is a Senior Principal Researcher at Microsoft, and leads the development of the DiskANN library which supports at-scale vector search in various Microsoft products such as web search, advertisement, Microsoft 365 and Windows copilots. He is also the tech lead for this topic in Azure Databases, which recently announced CosmosDB vector search. He co-leads a cross-industry and academia effort called big-ann-benchmarks.com to model real-world use cases, datasets and benchmarks. He holds a B.Tech. from IIT Madras and a Ph.D. from Carnegie Mellon University.

Charles Xie is the founder and CEO of Zilliz, a company dedicated to developing a cutting-edge unstructured data platform for AI applications. He is the creator of Milvus, a leading open-source vector database that is used by over 5,000 enterprises worldwide. In addition to his role at Zilliz, Charles has been actively involved with the LF AI & Data Foundation, serving as a board member and as the chairperson from 2020 to 2021. Before founding Zilliz, he was one of the founding engineers behind Oracle's 12c cloud database project. Charles holds a Master's degree in Computer Science from the University of Wisconsin-Madison.