Data Agents: Rethinking Data Systems in the AI Agent Era

Guoliang Li
Tsinghua University
Beijing, China
liguoliang@tsinghua.edu.cn

Yuyu Luo
The Hong Kong University of Science and Technology
Guangzhou, China
yuyuluo@hkust-gz.edu.cn

ABSTRACT

The database community is at a pivotal moment. Large Language Models (LLMs) and AI agents are rapidly changing how users interact with data systems. The traditional model—where human experts write SQL queries or navigate complex BI tools—is being disrupted by a new vision: data agents capable of understanding natural language, reasoning about data semantics, autonomously executing multi-step analytical workflows, and collaborating with humans to derive insights. This shift introduces fundamental questions that span database systems, human-computer interaction, AI/ML infrastructure, and programming languages.

This panel will convene leading researchers and practitioners to discuss the future of data agents. A "*data agent*" is defined as an autonomous system capable of perceiving data in various forms (structured, semi-structured, and unstructured), planning and executing complex data manipulation and analysis tasks, interacting with humans through natural language or other intuitive interfaces, and learning from feedback to improve over time. The panel will examine whether this vision is achievable, the technical obstacles that need to be addressed, and how the database community should adapt to meet this emerging challenge.

PVLDB Reference Format:
Guoliang Li and Yuyu Luo. Data Agents: Rethinking Data Systems in the AI Agent Era. PVLDB, 19(12): 4946 - 4949, 2026.
doi:10.14778/3827998.3838708

1 WHY IS THIS TOPIC OF INTEREST TO THE COMMUNITY?

Data agents represent both an existential challenge and a tremendous opportunity for the database community [14, 17, 18, 23, 29, 34, 35]. On one hand, the rise of agentic AI threatens to bypass traditional database interfaces, potentially commoditizing decades of research on query optimization, indexing, and transaction processing [8, 13, 15, 19, 21, 22, 26, 33]. On the other hand, it opens vast new research frontiers: How should database systems be architected to serve as reliable substrates for agent workloads? What new abstractions are needed to support agents that operate across structured and unstructured data? How do we ensure correctness, efficiency, and security when autonomous agents access critical data?

Proceedings of the VLDB Endowment, Vol. 19, No. 12 ISSN 2150-8097.
doi:10.14778/3827998.3838708

The topic is timely because we are witnessing a convergence of enablers: LLMs with sufficient reasoning capability, enterprise data platforms with unified access patterns, and growing user demand for self-service analytics. Recent industrial developments, including major data platform vendors integrating AI agents into their products, signal that this is not a distant research fantasy but an imminent reality. The database research community must actively shape this transition rather than passively adapt to it.

2 HOW WILL THIS PANEL CHANGE THE COMMUNITY'S THINKING?

This panel aims to challenge a central assumption in database research: that databases remain the primary locus of data interaction and optimization. Data agents may shift the center of gravity from query execution to task execution, from SQL interfaces to mixed natural-language/programmatic interaction, and from passive data systems to agent-native environments with verification, memory, and governance [7]. The discussion will force the community to debate whether data agents are merely a new interface layer, or whether they require a fundamental rethinking of database architecture, benchmarks, and correctness guarantees.

3 OPENING QUESTIONS AND DISCUSSION PROMPTS

Each panelist will be asked to prepare a 3-minute opening statement addressing the following question:

> *"What is the single most important technical or conceptual barrier that must be overcome for data agents to become a production reality, and what is the database community's role in overcoming it?"*

Following the opening statements, the moderators will use the following questions to guide the discussion:

(1) **Agent-Native Architecture:** Do we need new database engines designed specifically for agent workloads, or can agents be effectively built on existing systems? What would an agent-native database look like?
(2) **Query Interfaces:** Is natural language sufficient for serious data analysis, or do we need new formal languages that blend the accessibility of NL with the precision of SQL? What can we learn from decades of HCI research on data visualization and interaction?
(3) **Correctness and Trust:** How can users trust data agents with mission-critical decisions? What verification, explanation, and human-override mechanisms are needed? How do we benchmark correctness in open-ended agent tasks?
(4) **Performance and Cost:** LLM calls are expensive and slow. What optimization techniques, caching strategies,

and model selection approaches can make data agents economically viable at scale?

(5) **Multi-Modal Data:** Real-world multi-modal data is messy, heterogeneous, and poorly documented. How do agents discover, understand, and integrate data from diverse sources? What metadata and documentation infrastructure is needed?

(6) **Industry vs. Academia:** What problems are industry players actually facing in deploying data agents, and how can academic research best contribute? Are we researching the right problems?

4 AUDIENCE ENGAGEMENT PLAN

To ensure the panel is entertaining and maximally engaging, we plan the following interactive elements:

- **Live Polling:** We will use a polling platform to survey the audience on key questions (e.g., "Will NL replace SQL for 50% of analytics tasks by 2030?") before and during the panel, displaying results in real-time to spark debate.
- **Audience Q&A:** At least 30 minutes will be reserved for audience questions. The moderators will actively seek questions that challenge panelist positions.
- **Position Statements:** Before the panel, each panelist will submit a one-paragraph position statement that will be shared with attendees, allowing the audience to come prepared with targeted questions.
- **Demo Challenge:** We will solicit short failure cases before the panel; if no audience submissions are available, moderators will prepare two representative failure cases to seed the discussion.

5 CONFIRMED PANELISTS

The following invited panelists have confirmed that they are willing to participate in person if the panel is accepted. They are listed in alphabetical order by surname.

Carsten Binnig *Professor, TU Darmstadt*

Prof. Dr. Carsten Binnig is Professor of Data Management at TU Darmstadt and Head of the Department of Data and AI Systems. His research focuses on data management systems, scalable data systems on modern hardware, and machine learning for scalable data systems. He received his Ph.D. from the University of Heidelberg in 2008, worked as a postdoctoral researcher at ETH Zurich, and later worked at SAP on in-memory databases, where he played a key role in the development of SAP HANA. Before joining TU Darmstadt in 2017, he joined Brown University in 2014. He is a founding faculty of hessian.AI, leads the Systemic AI for Decision Support research area at DFKI Darmstadt, and was elected to the VLDB Endowment Board of Trustees in 2024.

Rationale: Carsten brings a strong systems perspective on how data agents can be supported by scalable, high-performance, and production-oriented data infrastructures. His experience spanning academic database research, modern hardware-aware systems, and industrial in-memory database development makes him well positioned to discuss whether data agents require fundamentally new system architectures or can be built on top of existing database engines [7, 8].

Samuel Madden *College of Computing Distinguished Professor, MIT*

Samuel Madden is a Distinguished Professor at MIT, where he heads the Data Systems Group and the Data Science and AI Lab (DSAIL). He is a pioneer of column-oriented database systems (C-Store, Vertica), sensor networks (TinyDB), and learned database systems. His recent work on Palimpzest and the "Databases Unbound" vision explores declarative systems for AI workloads over unstructured data. He is the recipient of the SIGMOD Edgar F. Codd Innovations Award, an ACM Fellow, and has received multiple test-of-time awards. He is also co-founder and Chief Scientist at Cambridge Mobile Telematics.

Rationale: Sam brings both historical perspective (having shaped multiple generations of database technology) and cutting-edge vision (with his declarative AI systems work) [5, 30, 32]. His "Databases Unbound" vision directly addresses the panel topic, arguing for extending database principles to all of the world's data through AI [25].

Fatma Ozcan *Principal Software Engineer, Google*

Fatma Ozcan is a Principal Software Engineer at Google, with over 20 years of industrial research experience. Prior to Google, she was a Distinguished Research Staff Member at IBM Almaden Research Center. Her research focuses on large-scale data analysis platforms, knowledge graphs, natural language querying, and query processing and optimization. She is an elected member of the SIGMOD Executive Committee, serves on the board of trustees for the VLDB Endowment, and is an ACM Distinguished Member.

Rationale: Fatma brings essential industrial perspective on building and deploying data systems at scale [6, 11, 27]. Her expertise in natural language interfaces to data and her leadership roles in the database community add significant weight to the panel. She represents the voice of industrial research that must balance innovation with production constraints.

Aditya Parameswaran *Associate Professor, UC Berkeley*

Aditya Parameswaran is an Associate Professor at UC Berkeley, where he co-directs the EPIC Data Lab focused on low/no-code data tooling powered by LLMs. He co-founded Ponder (acquired by Snowflake in 2023) and has developed widely-used tools, including DocETL, DocWrangler, DataScout, Modin, and Lux, with tens of millions of downloads. His research leverages data management, HCI, and AI techniques to develop next-generation data tooling.

Rationale: Aditya brings unique expertise in building and deploying LLM-powered data tools that practitioners actually use [20, 24]. His group's work on data agent benchmarks is directly relevant to the panel topic, and his HCI perspective ensures the discussion remains grounded in user needs [1, 12].

Nan Tang *Associate Professor, HKUST(GZ)*

Nan Tang is an Associate Professor at The Hong Kong University of Science and Technology (Guangzhou), with an affiliated position at HKUST. He is also an Associate Dean of the Information Hub at HKUST(GZ). His research interests lie at the intersection of data management and artificial intelligence, including data preparation, data-centric AI, multimodal data management, and natural language interfaces for data. He has published over 130 papers in major database venues and received numerous awards, including the VLDB Best Paper Award, ACM SIGMOD Reproducibility

Award, SIGMOD Best Paper Award, and SIGMOD Research Highlight Award.

Rationale: Nan brings a unique perspective on how modern data management systems should evolve in the era of AI agents. His research on data-centric AI, data preparation, and natural language interfaces directly addresses the foundations needed for reliable and effective data agents [3, 16, 30, 37]. His experience bridging database systems and emerging AI techniques will provide valuable insights into how data agents can transform data discovery, analysis, and management workflows.

Immanuel Trummer *Associate Professor, Cornell University*

Immanuel Trummer is an Associate Professor at Cornell, where his research makes data analysis more efficient and user-friendly through AI. He has developed systems for automated database tuning (DB-BERT, UDO), LLM-powered SQL generation (SQLBarber, CodexDB), automated claim verification (CEDAR, Scrutinizer), and LLM-driven data analysis (NaturalMiner). His work spans formal methods, optimization, and machine learning for databases.

Rationale: Immanuel represents the rigorous automation perspective, bringing expertise in combining LLMs with principled database techniques. His work on verification and optimization of LLM-generated database operations addresses key correctness challenges for data agents [9, 10, 31, 31].

Eugene Wu *Associate Professor, Columbia University*

Eugene Wu is an Associate Professor at Columbia University and Co-director of the Data, Agents, and Processes Lab (DAPLab). He is broadly interested in the foundations of computing infrastructure needed for AI agents to safely, reliably, and efficiently automate complex work. His research spans the computing stack from visualization and HCI to core data systems. He received his PhD from MIT and was a postdoc in the AMPLab at Berkeley. He has received the VLDB 10-year test of time award, NSF CAREER award, and faculty awards from Google, Adobe, and Amazon.

Rationale: Eugene's lab name (Data, Agents, and Processes) reflects his direct focus on the panel topic. His broad perspective spanning HCI, visualization, and data systems, combined with his recent work on agentic data environments and benchmarks, makes him an ideal panelist to bridge different communities and identify cross-cutting challenges [2, 4, 28, 36].

6 PANEL MODERATORS

Guoliang Li (https://dbgroup.cs.tsinghua.edu.cn/ligl/) is a full professor in the Department of Computer Science, Tsinghua University. His research interests mainly include learning-based databases and data-centric AI. He is an ACM Fellow and an IEEE Fellow. He got the VLDB 2017 Early Research Contribution Award, TCDE 2014 Early Career Award, ICDE 2025 Best Paper Runner-up, VLDB 2023 Industry Best Paper Runner-up, Best of SIGMOD 2023, SIGMOD 2023 research highlight award, DASFAA 2023 Best Paper Award, and CIKM 2017 Best Paper Award. He has extensive experience organizing major database venues, having served as SIGMOD 2021 General Co-Chair, ICDE 2027 PC Co-Chairs, VLDB 2022 Demo Co-Chair, ICDE 2022 Industrial Co-Chair, SIGMOD 2022 Tutorial Co-Chair, and SIGMOD 2024 Program Contest Co-Chair.

Yuyu Luo (https://luoyuyu.vip/) is an Assistant Professor at HKUST (Guangzhou) and HKUST. He received his Ph.D. in Computer Science from Tsinghua University in 2023. His research interests lie at the intersection of data and AI, with a focus on data agents and data-centric AI. His work has been published in top-tier database venues (SIGMOD, VLDB, TODS) and AI/ML venues (ICML, NeurIPS, ICLR), with over 70 papers to date. He is a recipient of the Best of SIGMOD 2023 Paper award. He organized the Second Workshop on LLM+Vector Data at ICDE 2026 and the first Agentic Data System Workshop @ VLDB 2026, and has presented tutorials at SIGMOD 2026, VLDB 2025, and SIGMOD 2023. He also organized the KDD Cup 2026 *(Data Agents for Complex Data Analysis Track).*

ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China (2023YFB4503600), NSF of China (62525202, 62232009, 62402409), Shenzhen Project (CJGJZD20230724093403007), China Railway Science Research Institute Group Co., Ltd, Zhongguancun Lab, Huawei, Tsinghua-Tencent Joint Laboratory and Beijing National Research Center for Information Science and Technology (BNRist). Guoliang Li and Yuyu Luo are the corresponding authors.

REFERENCES

[1] Shubham Agarwal, Asim Biswal, Sepanta Zeighami, Alvin Cheung, Joseph Gonzalez, and Aditya G. Parameswaran. 2026. Arming Data Agents with Tribal Knowledge. *CoRR* abs/2602.13521 (2026).

[2] Elaine Ang, Chenxi Huang, Georgios Liargkovas, Jerry Liu, Jinhui Liu, Nikos Pagonas, Charlie Summers, Haonan Wang, Jiakai Xu, Tianle Zhou, Yusen Zhang, Zhou Yu, Zhuo Zhang, Tianyi Peng, Kostis Kaffes, and Eugene Wu. 2026. Agentic Data Environments. *IEEE Data Eng. Bull.* 50, 1 (2026), 5–21.

[3] Chengliang Chai, Nan Tang, Ju Fan, and Yuyu Luo. [n.d.]. Demystifying Artificial Intelligence for Data Preparation. In *Companion of the 2023 International Conference on Management of Data.*

[4] Yiru Chen, Xupeng Li, Jeffrey Tao, Lana Ramjit, Subrata Mitra, Javad Ghaderi, Ravi Netravali, Aditya G. Parameswaran, Dan Rubenstein, and Eugene Wu. 2025. Physical Visualization Design: Decoupling Interface and System Design. *Proc. ACM Manag. Data* 3, 3 (2025), 197:1–197:27.

[5] Zui Chen, Lei Cao, Sam Madden, Tim Kraska, Zeyuan Shang, Ju Fan, Nan Tang, Zihui Gu, Chunwei Liu, and Michael Cafarella. 2023. SEED: Domain-specific data curation with large language models. *arXiv preprint arXiv:2310.00749* (2023).

[6] Yeounoh Chung, Gaurav Tarlok Kakkar, Yu Gan, Brenton Milne, and Fatma Ozcan. 2025. Is Long Context All You Need? Leveraging LLM's Extended Context for NL2SQL. *Proc. VLDB Endow.* 18, 8 (2025), 2735–2747.

[7] Timo Eckmann and Carsten Binnig. 2026. A Vision for Autonomous Data Agent Collaboration: From Query-by-Integration to Query-by-Collaboration. In *CIDR.*

[8] Timo Eckmann, Matthias Jasny, Johannes Wehrstein, and Carsten Binnig. 2026. The Future Is Bespoke: Synthesizing One-Size-Fits-One DBMSs with LLM Coding Agents. *IEEE Data Eng. Bull.* 50, 1 (2026), 88–103.

[9] Victor Giannakouris and Immanuel Trummer. 2025. λ-Tune: Harnessing Large Language Models for Automated Database System Tuning. *Proc. ACM Manag. Data* 3, 1 (2025), 2:1–2:26.

[10] Jiale Lao and Immanuel Trummer. 2026. GenDB: The Next Generation of Query Processing - Synthesized, Not Engineered. *CoRR* abs/2603.02081 (2026).

[11] Jiale Lao, Andreas Zimmerer, Olga Ovcharenko, Tianji Cong, Matthew Russo, Gerardo Vitagliano, Michael Cochez, Fatma Özcan, Gautam Gupta, Thibaud Hottelier, H. V. Jagadish, Kris Kissel, Sebastian Schelter, Andreas Kipf, and Immanuel Trummer. 2026. SemBench: A Benchmark for Semantic Query Processing Engines. *Proc. VLDB Endow.* 19, 8 (2026), 1754–1767.

[12] Quentin Romero Lauro, Shreya Shankar, Sepanta Zeighami, and Aditya G. Parameswaran. 2026. RAG Without the Lag: Enabling "What-If" Analysis for Retrieval-Augmented Generation Pipelines. In *CHI.* ACM, 368:1–368:16.

[13] Boyan Li, Chong Chen, Zhujun Xue, Yinan Mei, and Yuyu Luo. 2025. DeepEye-SQL: A Software-Engineering-Inspired Text-to-SQL Framework. *arXiv preprint arXiv:2510.17586* (2025).

[14] Boyan Li, Yuyu Luo, Chengliang Chai, Guoliang Li, and Nan Tang. 2024. The Dawn of Natural Language to SQL: Are We Fully Ready? *VLDB* (2024).

[15] Boyan Li, Jiayi Zhang, Ju Fan, Yanwei Xu, Chong Chen, Nan Tang, and Yuyu Luo. 2025. Alpha-SQL: Zero-Shot Text-to-SQL using Monte Carlo Tree Search. In *Forty-second International Conference on Machine Learning.*

[16] Changlun Li, Chenyu Yang, Yuyu Luo, Ju Fan, and Nan Tang. 2025. Weak-to-Strong Prompts with Lightweight-to-Powerful LLMs for High-Accuracy, Low-Cost, and Explainable Data Transformation. *Proceedings of the VLDB Endowment* (2025).

[17] Guoliang Li, Jiayi Wang, Chenyang Zhang, and Jiannan Wang. 2025. Data+AI: LLM4Data and Data4LLM. In *SIGMOD*.

[18] Guoliang Li, Xuanhe Zhou, and Xinyang Zhao. 2024. LLM for data management. *Proceedings of the VLDB Endowment* (2024).

[19] Chunwei Liu, Matthew Russo, Michael J. Cafarella, Lei Cao, Peter Baile Chen, Zui Chen, Michael J. Franklin, Tim Kraska, Samuel Madden, Rana Shahout, and Gerardo Vitagliano. 2025. Palimpzest: Optimizing AI-Powered Analytics with Declarative Query Processing. In *CIDR*. www.cidrdb.org.

[20] Shu Liu, Soujanya Ponnapalli, Shreya Shankar, Sepanta Zeighami, Alan Zhu, Shubham Agarwal, Ruiqi Chen, Samion Suwito, Shuo Yuan, Ion Stoica, Matei Zaharia, Alvin Cheung, Natacha Crooks, Joseph Gonzalez, and Aditya G. Parameswaran. 2026. Supporting Our AI Overlords: Redesigning Data Systems to be Agent-First. In *CIDR*. www.cidrdb.org.

[21] Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. 2025. A Survey of Text-to-SQL in the Era of LLMs: Where are we, and where are we going? *IEEE Transactions on Knowledge and Data Engineering* (2025).

[22] Xinyu Liu, Shuyu Shen, Boyan Li, Nan Tang, and Yuyu Luo. 2025. Nl2sql-bugs: A benchmark for detecting semantic errors in nl2sql translation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*.

[23] Yuyu Luo, Guoliang Li, Ju Fan, and Nan Tang. 2026. Data Agents: Levels, State of the Art, and Open Problems.

[24] Ruiying Ma, Shreya Shankar, Ruiqi Chen, Yiming Lin, Sepanta Zeighami, Rajoshi Ghosh, Abhinav Gupta, Anushrut Gupta, Tanmai Gopal, and Aditya G. Parameswaran. 2026. Can AI Agents Answer Your Data Questions? A Benchmark for Data Agents. *CoRR* abs/2603.20576 (2026).

[25] Samuel Madden, Michael J. Cafarella, Michael J. Franklin, and Tim Kraska. 2024. Databases Unbound: Querying All of the World's Bytes with AI. *Proc. VLDB Endow.* 17, 12 (2024), 4546–4554.

[26] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).

[27] Fatma Ozcan, Yeounoh Chung, Yannis Chronis, Yu Gan, Yawen Wang, Carsten Binnig, Johannes Wehrstein, Gaurav Tarlok Kakkar, and Sami Abu-el-haija. 2025. LLMs and Databases: A Synergistic Approach to Data Utilization. *IEEE Data Eng. Bull.* 49, 1 (2025), 32–44.

[28] Shreya Shankar, Tristan Chambers, Tarak Shah, Aditya G. Parameswaran, and Eugene Wu. 2025. DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing. *Proc. VLDB Endow.* 18, 9 (2025), 3035–3048.

[29] Zhaoyan Sun, Jiayi Wang, Xinyang Zhao, Jiachi Wang, and Guoliang Li. 2025. Data Agent: A Holistic Architecture for Orchestrating Data+AI Ecosystems. *IEEE TCDE Data Bulletin* (2025).

[30] Nan Tang, Chenyu Yang, Zhengxuan Zhang, Yuyu Luo, Ju Fan, Lei Cao, Sam Madden, and Alon Y. Halevy. 2024. Symphony: Towards Trustworthy Question Answering and Verification using RAG over Multimodal Data Lakes. *IEEE Data Eng. Bull.* 48, 4 (2024), 135–146.

[31] Immanuel Trummer. 2025. Customizing Operator Implementations for SQL Processing via Large Language Models. *IEEE Data Eng. Bull.* 49, 1 (2025), 45–56.

[32] Gerardo Vitagliano, Jun Chen, Peter Baile Chen, Ferdi Kossmann, Eugenie Lai, Chunwei Liu, Matthew Russo, Sivaprasad Sudhir, Anna Zeng, Ziyu Zhang, Michael J. Cafarella, Tim Kraska, and Sam Madden. 2025. Towards AI-Enabled Data-to-Insights Systems. *IEEE Data Eng. Bull.* 49, 3 (2025), 48–64.

[33] Jiayi Wang and Guoliang Li. 2025. AOP: Automated and interactive llm pipeline orchestration for answering complex queries. CIDR.

[34] Jiayi Wang, Guoliang Li, and Jianhua Feng. 2025. iDataLake: An LLM-Powered Analytics System on Data Lakes. *IEEE Data Eng. Bull.* 49, 1 (2025), 57–69.

[35] Kuncan Wang, Ziting Wang, Peizhuo Lv, Haoyang Li, Guoliang Li, Gao Cong, and Wei Dong. 2026. Data Agents Under Attack: Vulnerabilities in LLM-Driven Analytical Systems. *CoRR* abs/2606.08661 (2026).

[36] Eugene Wu. 2025. Database Theory in Action: Database Visualization. In *ICDT (LIPIcs)*, Vol. 328. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 35:1–35:6.

[37] Chenyu Yang, Yuyu Luo, Chuanxuan Cui, Ju Fan, Chengliang Chai, and Nan Tang. 2025. Data Imputation with Limited Data Redundancy Using Data Lakes. *Proc. VLDB Endow.* (2025). https://doi.org/10.14778/3748191.3748200