Data Systems for AI and Quantum

Rihan Hai

Web Information System, TU Delft, Netherlands
r.hai@tudelft.nl

In this talk, I will explore a fundamental question in the era of generative AI and quantum computing:

What is data?

AI models as data. Traditionally, data refers to information such as numbers, text, images, and videos, which are typically stored in DBMSs or file systems. Modern AI processes such data and produces new artefacts such as model parameters, activations, and computation graphs.

Instead of bringing data to computation, we treat computation itself as data. We represent model weights, activation function, and computation graphs as database-native objects, e.g., tables, and operate over them using relational and tensor abstractions, integrating linear algebra with relational algebra to support querying and optimization [4, 12]. This shift in the understanding of data also enables the design of systems that compile transformer inference into executable SQL, supporting resource-aware LLM serving on constrained platforms such as edge devices and 6G infrastructure [11].

Moving from a single model to a repository of pre-trained models (i.e., model zoos) introduces a key challenge: selecting which model to fine-tune. We organize pre-trained models as datasets in a data lake, represented as graphs [5]. This allows us to build systems that model both datasets and models as a heterogeneous graph, enabling learning over model–data relationships to identify fine-tuning candidates and reduce the cost of transfer learning [7, 6].

In real-world settings such as edge devices, data is often insufficient, imbalanced, or subject to privacy constraints. We address this by extending the notion of data with *synthetic data*, and building systems that generate it while preserving statistical properties and satisfying high-level constraints such as metadata [9, 10].

From classical to quantum data. Quantum data refers to information encoded as quantum states. These may be classical data embedded into qubits or quantum outputs produced by quantum processors. We store and manage quantum states via classical representations (vectors, tensors, group generators), enabling efficient simulation [8] while exploiting structural properties such as sparsity [3, 2]. Building on this foundation, we envision quantum-native data systems that manage quantum states efficiently, which enable new privacy and security solutions grounded in the principles of quantum communication [1].

References

[1] (Poster) Giancarlo Gatti and Rihan Hai. Private quantum database. The 29th Annual Quantum Information Processing Conference, January 2026.

[2] Floris Geerts and Rihan Hai. QC meet CQ: Quantum conjunctive queries. In *Proceedings of the 2nd Workshop on Quantum Computing and Quantum-Inspired Technology for Data-Intensive Systems and Applications*, Q-Data '25, page 25, New York, NY, USA, 2025. Association for Computing Machinery.

[3] Rihan Hai, Shih-Han Hung, Tim Coopmans, Tim Littau, and Floris Geerts. Quantum data management in the NISQ era. *PVLDB*, 18(6):1720–1729, 2025.

[4] Rihan Hai, Christos Koutras, Andra Ionescu, Ziyu Li, Wenbo Sun, Jessie van Schijndel, Yan Kang, and Asterios Katsifodimos. Amalur: Data integration meets machine learning. In *ICDE*, pages 3729–3739, 2023.

[5] Rihan Hai, Christos Koutras, Christoph Quix, and Matthias Jarke. Data lakes: A survey of functions and systems. *TKDE*, 35(12):12571–12590, 2023.

[6] Ziyu Li, Wenbo Sun, Danning Zhan, Yan Kang, Lydia Chen, Alessandro Bozzon, and Rihan Hai. Amalur: The convergence of data integration and machine learning. *TKDE*, pages 1–14, 2024.

[7] Ziyu Li, Hilco Van Der Wilk, Danning Zhan, Megha Khosla, Alessandro Bozzon, and Rihan Hai. Model selection with model zoo via graph learning. In *ICDE*, pages 1296–1309, 2024.

[8] Tim Littau and Rihan Hai. Qymera: Simulating quantum circuits using RDBMS. In *SIGMOD*, pages 179–182, 2025.

[9] Aditya Shankar, Hans Brouwer, Rihan Hai, and Lydia Chen. Silofuse: Cross-silo synthetic data generation with latent tabular diffusion models. In *ICDE*, pages 110–123, 2024.

[10] Aditya Shankar, Lydia Chen, Arie van Deursen, and Rihan Hai. WaveStitch: Flexible and fast conditional time series generation with diffusion models. *SIGMOD*, 3(6):1–25, December 2025.

[11] Wenbo Sun, Qiming Guo, Wenlu Wang, and Rihan Hai. TranSQL+: Serving large language models with SQL on low-resource hardware. *SIGMOD*, 3(6):1–27, December 2025.

[12] Wenbo Sun and Rihan Hai. Ilargi: A gpu compatible factorized ml model training framework. In *Proceedings of the 26th International Conference on Web Information Systems Engineering (WISE 2025)*, 2025. to appear.