Talk Title:

Beyond Indexes for Lakehouse Systems: Bridging the Semantic Gap in RAG via Schema-Graphs and Optimized Physical Layouts

Abstract:

While vector indexes are often evaluated on their ability to reach 99% recall, this metric frequently fails to translate into actual Question Answering (QA) accuracy for complex semantic queries. This talk examines 'two sides of the retrieval coin' within the IBM Watsonx.data ecosystem to address this gap. First, we evaluate unstructured retrieval-investigating how embedding models and indexing strategies (such as ParadeDB DiskANN) satisfy semantic questions on BEIR-style benchmarks. Second, we address relational data through Schema-GraphRAG (ICDE 2026 Demo), which utilizes graph-based traversals to resolve join-paths that standard vector-only approaches often ignore.

To support these applications, we move beyond traditional indexing at the foundational layers. At the physical layer, we discuss Predictive Table Layout Optimization (PTO) (SIGMOD ‘26), where data is reorganized to satisfy strict latency SLAs without the overhead of maintaining traditional relational indexes. Finally, we address the logical layer through ALFA (VLDBJ ‘24), using active learning for Ontology Graph Alignment. This offline alignment creates a consistent semantic layer across diverse sources, providing the necessary logical foundations for the complex, multi-source retrieval required in production Lakehouse environments.

Bio:

Dr. Vamsi Meduri is a Staff Research Scientist at IBM Research, which he joined in 2022. His work focuses on the intersection of database systems and applied machine learning, where he currently develops hybrid retrieval and RAG components for IBM Watsonx.data. He received an IBM Outstanding Technical Achievement Award (OTA) in 2025 for his contributions to the Db2 optimizer-as-a-service for the Presto lakehouse ecosystem. His earlier research focused on Active Learning for data integration and Predictive Analytics for query optimization. Specifically, his past projects span across data preparation and application layers, entity matching, knowledge graphs, predictive modeling, and systems optimization. His research outcomes have been published in top-tier data management venues, such as SIGMOD, VLDB, ICDE, EDBT, VLDB Journal, and ACM TODS. He holds a Ph.D. from Arizona State University and an M.S. from the National University of Singapore. He serves on the Program Committees for major database conferences, including SIGMOD, VLDB, ICDE, and EDBT.