

Research topics

Topic #1: Semantic Duality Views: A Unified Framework for Consistent Cross-Modal Semantic Search in Oracle Database

Development contact: Avneesh Pant

Abstract:

Modern enterprises increasingly store related information across multiple modalities—relational tables, JSON documents, Redis-style key-value structures, Kafka streams, and Iceberg objects. Oracle’s Duality Views elegantly unify these representations at the **structural** level, but they lack a corresponding **semantic** unification layer that enables AI systems to understand, index, and retrieve meaning across modalities in a consistent way. This internship explores the concept of **Semantic Duality Views**, a research-driven extension to the duality model that derives canonical “semantic documents” from existing duality projections and maps them into a unified semantic index stored within Oracle 23ai.

Topic #2: Data Gathering Framework for Learning-Based DB Autonomy

Development contact: Vikramraj Sitpal

Abstract

Field: ML for Systems

Databases have a growing opportunity to better support the collection and use of high-quality data for autonomous database research. Today, work on self-driving databases - such as anomaly detection, and hang analysis often relies on fragmented tooling and bespoke data pipelines, which can slow experimentation. This project proposes a six-month exploratory study to evaluate an integrated experimentation environment that streamlines training data acquisition for Oracle DBMS research. The proposed platform would offer a unified API with pluggable components for data collection, labeling, transformation, synthetic data generation, quality validation, and benchmarking. By standardizing and automating the path from raw system signals to ML-ready datasets, the environment aims to enable more reproducible experiments, faster model iteration, and scalable progress toward end-to-end autonomous DBMS capabilities. The study will assess the feasibility, impact, and long-term value of this architecture as a foundational research asset for future initiatives. (Inspired by Database Gyms paper out of CMU)

Topic #3: A Unified Observability Framework Enhanced by eBPF

Development contact: Vikramraj Sitpal

Abstract

Field: Observability and Diagnostics

Modern database systems often lack sufficiently rich, correlated observability data for diagnosing performance and correctness issues. Engineers rely on logs, traces, and limited metrics that provide only partial visibility, forcing issue resolution to depend on problem reproduction or expert-driven hypothesis of building through manual inspection. To remove data insufficiency as a root obstacle, we propose a unified observability framework that aggregates correlatable signals from multiple sources into a consistent, developer-friendly format with minimal overhead. The framework leverages eBPF as a low-overhead mechanism to enrich existing observability tools. Building on recent eBPF instrumentations like passive stack-sampling techniques, it captures system-level signals - such as CPU scheduling effects, kernel and user stack contexts, lock contention, and I/O behavior - and correlates them with application logs, traces, and runtime metadata via shared timestamps, thread IDs, and request identifiers. By standardizing and correlating these signals, the framework enables a holistic view of application behavior, resource usage, and kernel interactions during triage. This correlation-first approach strengthens analysis, reduces reliance on expert intuition, shortens debugging cycles, and eliminates the need to reproduce complex or rare failures - making data insufficiency effectively a non-factor while preserving production performance.

Topic #4: A Framework for CXL Memory Offloading for Oracle AI Database

Development Contact: Pei Li

Abstract:

Modern database systems face growing memory demands due to large-scale workloads and diverse query patterns, especially in vector databases with high-dimensional embeddings. These workloads are bursty, requiring dynamic memory management. Disaggregated memory via CXL (Compute Express Link) helps by pooling memory across servers and decoupling it from compute nodes, providing flexible, low-latency expansion. However, managing data between DRAM and CXL remains challenging. The system consists of two key components: a characterization study and profiling that captures real-time access patterns and semantics of DB workloads, and a CXL management layer inside the database that uses this data to identify and manage hot pages. By dynamically promoting or demoting hot memory through a combination of hybrid migration techniques or fine-grained hot-spot tracking, the database itself controls the movement, optimizing memory placement based on real-time needs to ensure better prediction and performance, whether through application-level hints or autonomous memory scaling.

Topic #5: Memory Safety Through AI

Development contact: Ranjit Noronha

Abstract

Field: AI based corruption detection and analyses

Memory safety in legacy code written in memory-unsafe languages like C/C++ is a critical challenge. Issues such as buffer overflows, stale reads, and overwrites are difficult to detect during testing due to the exponential number of execution paths in large codebases, leaving many vulnerabilities undiscovered until production. Tools like ASAN can detect these bugs but incur significant memory and performance overhead, limiting their use to internal testing. Hardware support such as Memory Tagging Extension (MTE) enables validation of memory accesses in production, making continuous memory safety checks feasible. Recent AI models like GPT-5 provide strong reasoning capabilities to statically analyze code, hypothesize and explore potential corruption paths, and generate targeted test cases using fuzzing. Combined with MTE, this approach can uncover new memory corruption issues. This project focuses on developing a tool to harden legacy codebases.