Carnegie Mellon University
Database Group
Industry Affiliates

CMU Database Group

OVERVIEW

Fall 2024

https://db.cs.cmu.edu

CMU DATABASE GROUP

Research

Teaching

Seminars

RESEARCH AGENDA (2013-2023)

Our focus in the last decade has been on the use of AI/ML to automatically configure, optimize, and operate database management systems.

→ Includes both creating the ML methods for tuning systems and designing new system architectures to support automated control.

Existing Systems: OtterTune *(dead)*

Self-Driving Systems: NoisePage *(defunct)*, DB Gym

OBSERVATION

Lots of innovation in the last 10 years on modern DBMS architectures in both academia and industry.

→ Many of those players are in this room right now.
→ Most academics are building off of **DuckDB** these days.

But the most efficient system and robust ML methods are wasted if the DBMS chooses bad query plans and cannot adapt at runtime…

RESEARCH AGENDA (2024-???)

New cost-based query optimization service (**optd**) designed for modern data-intensive systems.

Optimization is **not** a one-shot operation. Service will generate a plan, follow its execution behavior to learn whether the chosen was correct or not, and make incremental changes.

Current prototype relies on DataFusion front-end:
→ Input: SQL
→ Output: DataFusion plan (switching to Substrait)
→ Stretch Goal: Emit SQL with plan hints.

OPTD: DESIGN GOALS

Goal #1: Parallel Search

→ Multiple asynchronous threads simultaneously exploring solution space for the same query.

→ Enable the service to explore wider solution space for each query more efficiently.

Goal #2: Workload/Pipeline Optimization

→ Support holistic optimization for multiple queries for ETL/ELT pipelines (i.e., dbt).

→ Identify redundancies and opportunities to combine queries to reduce execution time/cost.

OPTD: DESIGN GOALS

Goal #3: Explainable Decision Making

→ Maintain meta-data about a query's optimization search progress and why it makes certain decisions.

→ Makes it easier for human and other tools to decipher the reasoning why query plans look the way they do.

Goal #4: Incremental / Restartable Searches

→ Service can use same debug meta-data to pause and restart optimization for individual queries.

→ Automatic invalidation of cached state based on new information derived from runtime observations.

ADDITIONAL RESEARCH PROJECTS

➡ **User-bypass Database Architectures (eBPF)**
→ *Embedding DBMS logic inside OS kernel.*

Database Gyms (PostgreSQL)
→ *Batteries-included ML/AI training platform for databases.*

UDF Compilation Magic (PL/SQL, Python)
→ *Automatic optimization of UDFs via inlining/outlining.*

Future File Formats
→ *Next generation open-source columnar file format.*

CMU DATABASE COURSES

Intro to Database Systems (15-445/645)
- → Fundamentals of disk-oriented DBMS architectures.
- → Enrollment: ~130 students per semester (BS/MS)

Advanced Database Systems (15-721)
- → Latest research on modern data-intensive systems.
- → Enrollment: 20-40 students per semester (BS/MS/PhD)

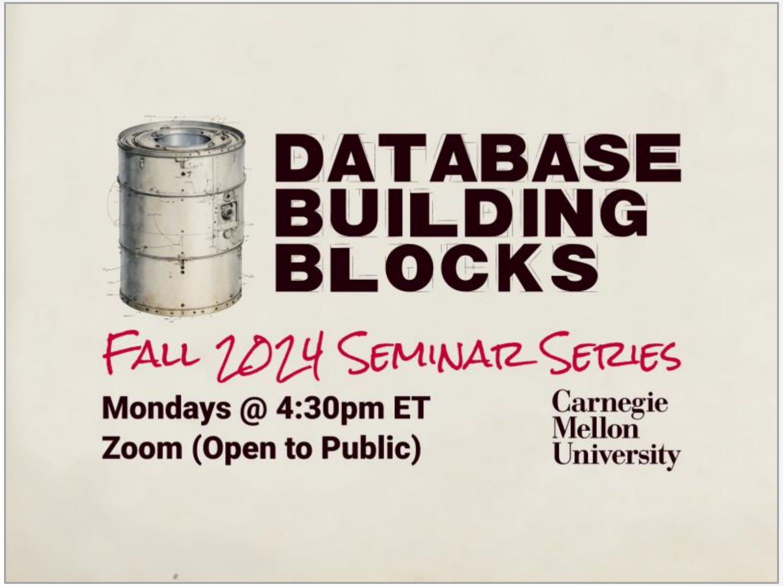
Special Topics in Databases (15-799)
- → Spring 2022: Self-Driving Databases
- → Spring 2025: Query Optimizers
- → Enrollment: 20 students (BS/MS/PhD)

Database Building Blocks

Mondays @ 4:30pm (starting on 9/23)

Live on Zoom. Published to YouTube afterwards. Open to the public.

https://db.cs.cmu.edu/seminar2024



DATABASE BUILDING BLOCKS

FALL 2024 SEMINAR SERIES

Mondays @ 4:30pm ET
Zoom (Open to Public)

Carnegie Mellon University

DATA FUSION

APACHE DATAFUSION COMET™

ParadeDB

THESEUS

Synnada

influxdb

GreptimeDB

APACHE OpenDAL™

Exon

GlareDB

Thank You to Our Inaugural IAP Members

ClickHouse

DATASTAX

FIREBOLT

RelationalAI

TiDB

CONFLUENT

dbt

NEON

startree

Weaviate