Gaurav Tarlok Kakkar

4^{th} Year Ph.D. Student
College Of Computing
Georgia Institute of Technology
gkakkar7@gatech.edu

Research Interests

My research lies at the intersection of databases and machine learning, with a specific focus on improving resource efficiency, enhancing usability, and expanding query capabilities. I have worked extensively on developing novel query optimization and execution algorithms to accelerate workloads in video analytics and large language model (LLM) applications. I am actively involved in leading the development of EvaDB, a novel database system designed to accelerate emerging AI applications. EvaDB has gained significant recognition, amassing approximately **2.7K** GitHub stars and receiving acknowledgment on platforms such as HackerNews (Discussion 1), HackerNews (Discussion 2), and Decibel.

More recently, my research has expanded to focus on natural language to SQL (NL2SQL) systems. In this work, I explore the trade-offs between accuracy, latency, and cost in NL2SQL pipelines, aiming to enhance the efficiency and usability of database interactions through natural language interfaces. This transition builds upon my prior work on optimizing query processing for multimedia and ML workloads, extending these insights to structured data with a focus on improving accessibility and efficiency.

Education

- **Georgia Institute of Technology** — Atlanta, GA
  Ph.D. in Computer Science — *Aug. 2021 – ongoing*
- **Georgia Institute of Technology** — Atlanta, GA
  Master's in Computer Science; **GPA: 4.0/4.0** — *Aug. 2019 – May. 2021*
- **Indian Institute of Technology** — Kanpur, India
  Bachelor of Engineering in Computer Science; **GPA: 9.5/10.0** — *Aug. 2013 – July. 2017*

Publications

A. Published Conference or Journal Articles

- Yeounoh Chung, **Gaurav Tarlok Kakkar**, Yu Gan, Brenton Milne, Fatma Ozcan
  Is Long Context All You Need? Leveraging LLM's Extended Context for NL2SQL,
  To be presented at VLDB, 2025
- **Gaurav Tarlok Kakkar***, Jiashen Cao*, Aubhro Sengupta, Joy Arulraj, and Hyesoon Kim.
  Aero: Adaptive Query Processing of ML Queries,
  To be presented at SIGMOD, 2025
- Mohammadreza Pourreza, Hailong Li, Ruoxi Sun, Yeounoh Chung, Shayan Talaei, **Gaurav Tarlok Kakkar**, Yu Gan, Amin Saberi, Fatma Ozcan, Sercan O Arik

Chase-sql: Multi-path reasoning and preference optimized candidate selection in text-to-sql,
ICLR, 2025

- **Gaurav Tarlok Kakkar***, Jaeho Bang*, Pramod Chunduri, Subrata Mitra, and Joy Arulraj.
  Seiden: Revisiting Query Processing in Video Database Systems,
  VLDB23: 48th Intl Conf on Very Large Data Bases, Vancouver, Canada, 2023.
  Proceedings of the VLDB Endowment, Vol. 16, No. 11.

- **Gaurav Tarlok Kakkar***, Zhuangdi Xu*, Joy Arulraj, and Umakishore Ramachandran.
  EVA: A Symbolic Approach to Accelerating Exploratory Video Analytics with Materialized Views,
  SIGMOD22: 49th ACM SIGMOD Intl Conf. on the Management of Data, Philadelphia, PA, 2022.

B. Workshop Publications

- **Gaurav Tarlok Kakkar** et al.
  EVA: An End-to-End Exploratory Video Analytics System
  DEEM @ SIGMOD23: 7th Workshop on Data Management for End-to-End Machine Learning, Seattle, WA, 2023.

C. Demonstrations

- **Gaurav Tarlok Kakkar**, Aryan Rajoria, Myna Prasanna Kalluraya, Ashmita Raju, Jiashen Cao, Kexin Rong, and Joy Arulraj
  Interactive Demonstration of EVA
  Proceedings of the VLDB Endowment, 2023

Patents

- **GT Kakkar**, M Singh Text Wrap Detection - US Patent 11,151,370 (link)
- M. Rastogi, P. Mehrotra, S. Sinha, **G.Kakkar**. Mapping annotations to ranges of text across documents - US Patent 11,151,307 (link)
- M. Rastogi, P. Mehrotra, S. Sinha, **G.Kakkar**. Digital Annotation And Digital Content Linking Techniques - US Patent 11,048,864 (link)

Experience

A. Research Experience

- **Graduate Research Assistant** GaTech
  Advised by Prof. Arulraj Joy *Aug 2021 - ongoing*

EVA - A new database management system (DBMS) tailored to efficiently and accurately enable ML queries at scale. This project focuses on the following problem:

- Materialized Views: EVA accelerates exploratory video analytics by 4 with minimal storage overhead (1.001) by automatically materializing and reusing expensive UDF results. Unlike traditional DBMS reuse techniques, EVA (1) targets UDFs instead of sub-plans, (2) uses symbolic predicate analysis to detect overlap, and (3) incorporates reuse into cost-based optimization decisions. Presented at SIGMOD 2022.
- Adaptive Query Processing: Traditional static query optimization does not work well for ML queries, as UDFs often dominate the queries, and it is non-trivial to collect accurate estimates of UDF statistics, such as selectivity and cost. To address this limitation, in EVA, we propose an extensive and generalizable adaptive query processing framework to adjust the query plan at the execution stage dynamically. We achieves up to 6.4× speedup compared to a state-of-the-art ML-centric DBMS with no impact on accuracy.
- Distributed Execution: EVA's execution engine leverages heterogeneous computational units (CPUs, GPUs). To support distributed query execution, EVA leverages Ray, a distributed framework. The modular and extensible nature of EVA enables users to write custom UDFs using deep learning frameworks like PyTorch, Tensorflow, etc.
- Revisiting Query Processing in VDBMS: We observe a diminishing gap in inference time between oracle and proxy models in VDBMSs, thereby challenging the assumptions made in SoTA VDBMSs. In this work, we leverage the oracle model and temporal video continuity to surpass SoTA query processing approaches. It incorporates a novel multi-arm bandit-based sampling algorithm for accelerated retrieval and aggregate queries, balancing exploration of unexplored regions while exploiting high-rewarding video segments. Empirical evaluations showcase a $6.6\times$ speed-up.

B. Industrial Experience

- **Google** Sunnyvale, USA
  Research Intern in **Systems Research @ Google** *May 2024 - Dec 2024*
  - Led a research project on In-Context Learning for NL2SQL using Google's Gemini 1.5 Pro.
  - Investigated accuracy-latency trade-offs of long-context LLMs for NL2SQL generation.
  - Evaluated the impact of rich contextual signals (e.g., schema, example values, hints, online synthetic examples) on model performance without finetuning.
- **Snowflake** Sunnyvale, USA
  Software Development Intern in **SQL Optimization Team** *Summer'21*
  - Led the research project to drive new query optimizations by analyzing production workloads based on query inter arrival patterns, query types, and execution statistics.
  - Designed architecture to collect and report back the query runtime statistics to the optimization engine to optimize future queries.
- **Google** Sunnyvale, USA
  Software Development Intern in **Cloud SQL** *Summer'20*
  - Led the project to accelerate OLAP (Online Analytical Processing) queries by automatically building columnar cache indexes

    - Improved the query statistics collection engine and building ML driven columnar cache index advisor.
    - 5x improvement in query execution time with no manual cost overhead and worked towards US patent.

- **Adobe Systems** Noida, India
  Member of Technical Staff *Jul 2017 - Aug 2019*
    - **Regenerate Layout from PDFs**: Research project
        - Built a multitude of ML algorithms processing together along with Software level heuristics to provide a single click layout generator from an inspiration pdf
        - Implemented a deep learning model, modified Faster RCNN to detect shape agnostic text wrap in a given pdf
        - Tackled challenges viz. detecting white space cover, creating master pages, organizing raw text runs into well defined text frames and intelligently figuring out object styles.

      Key Achievements: Filed patent in US on shape agnostic text wrap detection, implemented document analysis techniques and researched on core text properties.
    - **Import PDF Comments**: Keynote feature shipped with InDesign Max 2019
        - Implemented a novel approach to import and easily track the feedback made on a pdf version of document, solving the most in demand feature request of our million designers
        - Mastered the existing PDF library, to tackle the challenges of associating text or graphics with the annotation.

      Key Achievements: Filed patent in US and mastered PDF Library APIs

- **Adobe Systems** Bangalore, India
  Research Intern in Big Data Experience Lab *May 2016 - July 2016*
    - Generating personalized bundles of products for customers of e-Commerce website that are needs-driven.
    - Formulated a novel approach of incorporated common sense knowledge, Concept net along with data driven insights.
    - Formed candidate set using hierarchical and minimum spanning tree-based clustering algorithm over customer-centric data enriched by semantic analysis. (slides)

Recognition and Awards

2017	Dr. Elizabeth & Varkey Cherian Award(Best UG project with an impact on campus community)
2017	Academic Excellence Award, IIT Kanpur (awarded to top 7% students in the institute)
2014	Academic Excellence Award, IIT Kanpur (awarded to top 7% students in the institute)
2013	All India Rank 236, IIT-JEE Advanced (among 150,000 candidates).

Relevant Courses

- **Systems**: High Performance Parallel Computing, Advanced Operating Systems, Database Technologies, Computer Architecture, Compiler Design, Computer Networks, Computer Security, Advanced Data Structure and Algorithms
- **ML/Data Science**: Data analysis using Deep Learning, Recent Advances in Computer Vision, Machine Learning Tools and Techniques, Natural Language Processing, Data Visualization and Analysis

Other Projects

- **[ML Systems] Fast Array of Wimpy GPUs (FAWG)** GaTech
  Research Project with Prof. Alexey Tumanov *Sep. 2020 – Jan. 2021*
  - Serve memory-hungry models using model parallelism on cheap wimpy GPUs while meeting the latency SLOs
  - Proactive planner regresses over the batching parameters, model partitions, operator replicas, hardware types, and operator placement to search for a cost-effective model serving plan
  - The reactive planner behaves as a high-frequency tuner to auto-scale to meet tail latency goals in response to changes in the query arrival process.
- **[Databases] Cafeteria Automation System** IITK, India
  Under-Graduate Project with Prof. Sumit Ganguly *Jan. 2016 – Dec 2016*
  - Designed a desktop app in C# incorporating mess menu creation, consumption Records, items BOM management, worker Attendance and salary management.
  - Won **Dr. Elizabeth & Varkey Cherian Award** - **Best UG project** with an impact on campus community.
  - As of May 2017, managed over 2,00,000 transactions of worth greater than INR 3.4 million. (slides) (code)

Technical Skills

Languages	Python, C++, C, SQL, OpenMP, MPI, Js, Go, Bash, Assembly, HTML, CSS
ML	Tensorflow, Keras, scikit-learn, OpenCV, PyTorch