Zezhou (Zachary) Huang

Email : zh2408@columbia.edu

http://www.columbia.edu/~zh2408/

1272 Amsterdam Ave, New York, NY 10027-5047 United States

GitHub: https://github.com/zachary62

Education

Columbia University — New York City, NY

Ph.D. in Computer Science; GPA: 4.00; Advisor: Prof. Eugene Wu — *Sep. 2019 – May. 2024 (Expected)*

M.S. in Computer Science; GPA: 4.00 — *Sep. 2019 – May. 2021*

University of Wisconsin-Madison — Madison, WI

B.S. in Computer Science; GPA: 3.89 — *May. 2019*

Industry Experience

- **Microsoft** — Redmond, WA

  Research Intern — *May. 2023 – Aug. 2023*

  Developed a prototype for database engines using novel hardware. When tested with production workloads, our system performs over an order of magnitude faster and more cost-efficient than SQL Server and PowerBI.

- **Databricks** — San Francisco, CA

  Software Engineer Intern — *May. 2022 – Aug. 2022*

  Implemented data structures for query optimization and view coverage. Delivered to **Databricks Runtime 11.1**. Experimented IVM over join using delta table with dynamic pruning, low shuffle merge, and deletion vectors. Implemented MV strategies in **Enzyme**. (advised by Prof. Yannis Papakonstantino)

- **Tusimple** — San Diego, CA

  Software Engineer Intern — *May. 2021 – Aug. 2021*

  Built the back-end of Trip Data Collection Service that performs ETL over three data sources.

Research Experience

- **Columbia University** — New York City, NY

  Graduate Research Assistant — *Sep. 2019 – Present*

  - **Scalable, Interactive and Private Wide-Table Data Analytics**:
    I am developing systems that facilitate Wide-Table Data Analytics. The core insight is to model selection-projection-join-aggregation queries as a message-passing procedure while extensively reusing messages for collaborative efforts. My projects enable analytics to scale across thousands of tables, provide interactive data exploration within 100 milliseconds, and incorporate differential privacy.

- **University of Wisconsin-Madison** — Madison, WI

  Undergraduate Research Assistant — *Aug. 2018 – May. 2019*

  - **Managed Storage Hierarchy in WiscKey [Github]**:
    Assessed the read and write performance of **WiscKey** and **LevelDB** on solid-state drives in **C++** and **Go**. Exploited the inner data structure of **LSM tree** to balance the read and write performance. Evaluated the system on a 100-GB database. Improved performance by 17.3% under 4-KB values. Added a layer between **LSM tree** and APIs to balance the range query and random lookup performance.

Publications

1. **From Ambiguity to Clarity: How Documenting Data and Queries Improves GPT's Text-to-SQL.**
   Zezhou Huang, Pavan Kalyan Damalapati, and Eugene Wu.
   In Review
2. **The Fast and the Private: Task-based Dataset Search.**
   Zezhou Huang, Jiaxiang Liu, Haonan Wang, Eugene Wu.
   In Review
3. **Lightweight Materialization for Fast Dashboards Over Joins.**
   Zezhou Huang, and Eugene Wu.
   SIGMOD 2024

4. **Saibot: A Differentially Private Data Search Platform.**
   Zezhou Huang, Jiaxiang Liu, Daniel Gbenga Alabi, Raul Castro Fernandez, and Eugene Wu.
   VLDB 2023
5. **Kitana: Efficient Data Augmentation Search for AutoML.**
   Zezhou Huang, Pranav Subramaniam, Raul Castro Fernandez, and Eugene Wu.
   Coming soon
6. **Random Forests over normalized data in CPU-GPU DBMSes.**
   Zezhou Huang, Pavan Kalyan Damalapati, Rathijit Sen, and Eugene Wu.
   DaMoN@SIGMOD 2023
7. **JoinBoost: Grow Trees Over Normalized Data Using Only SQL.**
   Zezhou Huang, Rathijit Sen, Jiaxiang Liu, and Eugene Wu.
   VLDB 2023, Video
8. **Aggregation Consistency Errors in Semantic Layers and How to Avoid Them.**
   Zezhou Huang, Pavan Kalyan Damalapati, and Eugene Wu.
   HILDA@SIGMOD 2023
9. **Reptile: Aggregation-level Explanations for Hierarchical Data.**
   Zezhou Huang, and Eugene Wu.
   SIGMOD 2022, Video, News, Interview
10. **Calibration: A Simple Trick for Wide-table Delta Analytics.**
    Zezhou Huang, and Eugene Wu.
    Arxiv