Eugene Wu

Address
32 Vassar St Room G930
Cambridge, MA 02139
eugenewu@mit.edu
(510) 499-5876

Research Interests

Systems and user oriented database research, with a focus on high performance provenance.

Education

Ph.D., Electrical Engineering and Computer Science — Expected, 2014
Advisor: Samuel Madden
Thesis Topic: High Performance Provenance Systems and Use Cases
Massachusetts Institute of Technology, Cambridge, MA

Masters, Electrical Engineering and Computer Science — 2012
B.S.,, Electrical Engineering and Computer Science — 2006
University of California, Berkeley, Berkeley, CA.

Professional Experience

MIT CSAIL, Cambridge, MA — Spring 2007 - Present
Graduate Research Assistant, with Professors Samuel Madden and Michael Stonebraker.

Scorpion Project: Designed and implemented an analysis framework to explain outliers in the results aggregation queries by constructing predicates on the input data. I formalized the concept of predicate influence and identified several operator properties to enable higher performance on common statistical aggregates.

SubZero Project: Designed, prototyped, and evaluated a low overhead provenance system for large-scale scientific workflow applications that process gigabytes of data per second.

Qurk Project: This project pioneered the use of human computation platforms such as Mechanical Turk within a database query execution engine.

Index and Partitioning Techniques: I investigated the application of indexing and partitioning techniques for time-varying and skewed query workloads. Shinobi incrementally re-partitions and indexes database tables based on recent query access patterns. Our no-bits paper proposed the use of un-used space in B-tree indices as a cache for heavily accessed tuples.

Trajectory Optimized Storage: Implemented core storage system for TrajStore, a high performance data management system for storing and querying vehicle trajectory data by location and time. The system incrementally optimizes the storage layout as the query workload changes over time.

Google Inc., Mountain View, CA — Spring - Winter 2007
Intern, Data Management Research

Webtables Project: I worked in Alon Halevy's data management group on the WebTables project to mine

the Google web corpus for tabular data. I developed the table extraction pipeline and extracted more than 125 million tables. In addition, I built a table search engine that allows users to query over the structured data and automatically visualizes attributes in graphs or maps.

IBM Extreme Blue., Almaden, CA Spring 2005
Engineering Intern

Developed a new software patch service for DB2 for z/OS team that reduced patch application times from the order of months to a few minutes.

UC Berkeley Computer Science Department, Berkeley, CA Summer 2004 - Fall 2006
Undergraduate Researcher

High Performance Stream Processing: Designed and implemented one of the first high performance complex event processing systems for detecting high level events (e.g., shoplifting occured) from streams of raw sensor events (e.g., RFID tag XXX detected). Results were published at SIGMOD, the premier database conference.

The HiFi Project : Implemented the RFID reader interface for extracting raw events from early RFID readers and the interactive dashboard for the VLDB demonstration. HiFi is a research project around cascading stream architectures for large-scale receptor-based networks.

Teaching Experience

Instructor, Big Data Systems (MIT 6.885) Fall 2013
Co-developed and instructed MIT's first course focused on large scale data analysis tools and techniques. Topics ranged from data cleaning and integration, large-scale systems like Hadoop, to scalable visualization techniques. The course featured 8 new labs to give students hands-on experience with the systems covered in class.

Instructor, Data Analysis IAP Course Spring 2012
Co-developed and taught approximately 20 students introduction to data analysis course during MIT's Independent Activities Period in January. dataiap.github.io

Curriculum Head, MEET 2011 - 2012
Head of curriculum development for MEET. Help prepare incoming instructors during the summer. Successfully migrated the organization from a Java-based curriculum to a Python-oriented one. The rationale was both pedagogical (e.g, fast development iteration using the REPL), and practical (most MIT CS undergraduates are taught Python. Few are taught Java).

TA, Database Systems (MIT 6.830) Fall 2010
Assisted in writing and grading the assignments and projects.

Instructor, MEET Summer 2010
Mentored a group of 30 Israeli and Palestinian high schoo students through MIT's MEET program, a peace initiative in the Middle East.

Instructor, Introduction to Java Course Spring 2010, 2011
Instructed a class of 50 students in an introduction to the Java programming language.

TA, Database Systems (UCB CS186) Fall 2006
Taught approximately 30 students in weekly discussion sections. Assisted in writing and grading the assignments and projects.

Publications

[1] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. *VLDB*, 1(1):538–549, 2008.

[2] M. J. Cafarella, A. Y. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. Uncovering the relational web. *WebDB*, 2008.

[3] A. Cheung, L. Ravindranath, E. Wu, S. Madden, and H. Balakrishnan. Mobile applications need targeted micro-updates. In *APSys*, page 8. ACM, 2013.

[4] O. Cooper, A. Edakkunni, M. J. Franklin, W. Hong, S. R. Jeffery, S. Krishnamurthy, F. Reiss, S. Rizvi, and E. Wu. Hifi: A unified architecture for high fan-in systems. *VLDB*, pages 1357–1360, 2004.

[5] P. Cudre-Mauroux, E. Wu, and S. Madden. The case for rodentstore, an adaptive, declarative storage system. *CIDR*, 2009.

[6] P. Cudre-Mauroux, E. Wu, and S. Madden. Trajstore: An adaptive storage system for very large trajectory data sets. In *ICDE*, pages 109–120. IEEE, 2010.

[7] C. A. Curino, E. P. C. Jones, R. A. Popa, N. Malviya, E. Wu, S. R. Madden, H. Balakrishnan, N. Zeldovich, et al. Relational cloud: A database-as-a-service for the cloud. 2011.

[8] M. J. Franklin, S. R. Jeffery, S. Krishnamurthy, F. Reiss, S. Rizvi, E. Wu, O. Cooper, A. Edakkunni, and W. Hong. *Design considerations for high fan-in systems: The HiFi approach*, volume 5. CIDR, 2005.

[9] M. N. Garofalakis, K. P. Brown, M. J. Franklin, J. M. Hellerstein, D. Z. Wang, E. Michelakis, L. Tancau, E. Wu, S. R. Jeffery, and R. Aipperspach. Probabilistic data management for pervasive computing: The data furnace project. *IEEE Data Eng. Bull*, 29(1):57–63, 2006.

[10] S. Madden, E. Wu, S. Madden, Y. Zhang, E. Jones, C. Curino, et al. Relational cloud: The case for a database service. 2010.

[11] A. Marcus, E. Wu, D. Karger, S. Madden, and R. Miller. Human-powered sorts and joins. *VLDB*, 5(1):13–24, 2011.

[12] A. Marcus, E. Wu, D. Karger, S. Madden, and R. C. Miller. Demonstration of qurk: a query processor for human operators. In *SIGMOD*, pages 1315–1318. ACM, 2011.

[13] A. Marcus, E. Wu, D. R. Karger, S. R. Madden, R. C. Miller, et al. Crowdsourced databases: Query processing with people. 2011.

[14] A. Marcus, E. Wu, and S. Madden. Data in context: Aiding news consumers while taming dataspaces. *DBCrowd*, page 47, 2013.

[15] E. Wu. *Shinobi: Insert-aware partitioning and indexing techniques for skewed database workloads*. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2010.

[16] E. Wu, P. Cudre-Mauroux, and S. Madden. Demonstration of the trajstore system. *VLDB*, 2(2):1554–1557, 2009.

[17] E. Wu, C. A. Curino, S. R. Madden, et al. No bits left behind. *CIDR*, 2011.

[18] E. Wu, Y. Diao, and S. Rizvi. High-performance complex event processing over streams. In *SIGMOD*, pages 407–418. ACM, 2006.

[19] E. Wu and S. Madden. Partitioning techniques for fine-grained indexing. In *ICDE*, pages 1127–1138. IEEE, 2011.

[20] E. Wu and S. Madden. Scorpion: Explaining away outliers in aggregate queries. *VLDB*, 2013.

[21] E. Wu, S. Madden, and M. Stonebraker. A demonstration of dbwipes: clean as you query. *VLDB*, 5(12):1894–1897, 2012.

[22] E. Wu, S. Madden, and M. Stonebraker. Subzero: a fine-grained lineage system for scientific databases. *ICDE*, 2013.