UW Database Group

Data management systems, cloud services, probabilistic databases, and data pricing in Computer Science & Engineering at the University of Washington.

TALKS & DEMOS

Intel Science & Technology Center Retreat

August 2016: The UW DB group participated in the Intel Science & Technology Center retreat in Hillsboro, Oregon, where Magda Balazinska and Bill Howe presented work done in the group on the Myria system and graduate students Parmita, Brandon, Helga, Jingjing, Dylan, and Jenny presented posters. Shrainik, Dylan, and Helga also contributed to this year's BigDAWG demonstration showing the BigDAWG stack for federated analytics, including the Myria system on a use-case from the oceanography domain.

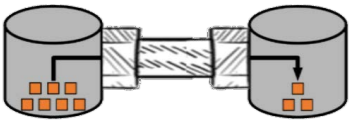
Dan Suciu: Uncertainty in Computation

October 4-7, 2016: Dan co-organized the Uncertainty in Computation Workshop at the Simons Institute for the Theory of Computing at UC Berkeley.
Uncertainty pervades our every interaction with the physical world and with all our attempts to organize our data and understanding. Probability is the best tool we have for managing uncertainty.

Definition A Probabilistic Database is (**W**, P), where **W** is an incomplete database, and P: **W** → [0,1] a probability distribution: $\sum_{i=1,n} P(W_i) = 1$

Brief Tutorial on Probabilistic Databases

In addition to co-organizing the Workshop, Dan conducted a tutorial on Probabilistic Databases. PDBs extend traditional relational databases by annotating each record with a weight, or a probability.



Brandon Haynes: PipeGen

In October, Brandon Haynes presented his paper titled "PipeGen: A Data Pipe Generator for Hybrid Analytics" at SOCC'16. His work was also recently featured on the Intel Science and Technology Center for Big Data (ISTC) blog. PipeGen is a tool used to automatically construct data pipes that efficiently move data between pairs of database systems. The resulting data pipes allow for inter-database transfer up to 3.8x faster than exporting and importing through the file system. More details -- including complete source code -- are available on the project website.

Magda Balazinska: Harvard CS Colloquium

October 13, 2016: Magda presented The Myria Big Data Management System and Cloud Service at the Harvard Computer Science Colloquium Series.

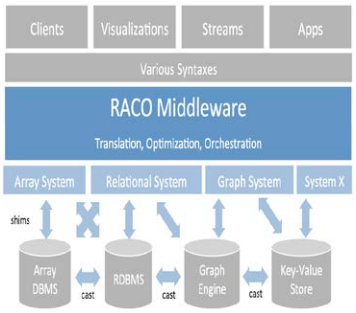
The Myria system and service is designed to meet the needs of modern data scientists, focusing on performance and productivity.

Bill Howe: Urban Analytics and Responsible Data Science

TechAlliance

November 9, 2016: Bill presented at SciTech Northwest.

The first decade of data science was characterized by what can be done: How can we extract actionable information from massive, noisy, heterogeneous, streaming datasets? The next decade will be characterized by what should be done: Identifying the right projects, ensuring accountability, and guarding personal privacy – while also avoiding algorithmic bias, and political inequities.



Polystore Team Demos @ Maryland Quarterly Review

Bill Howe, Shrainik Jain, and Dylan Hutchison traveled to Maryland to present their polystore work. The goal of the research is to expand the Raco relational optimizer to compile queries to execute on different data processing systems. Each part of the query should execute in the system on which it performs best.

POSTERS

Industry Affiliates

On October 19, 2016, the Database Lab presented posters at the CSE Industry Affiliates Meeting.

The main objective of the Industry Affiliates Program is to support the mutual needs of business, industry, and academia in computer research, development, and education. This is accomplished by providing appropriate mechanisms for technical exchange and collaboration and employment of students.

Poster Presentations

- **Casper: Using Verified Lifting for Spark**
  Maaz Ahmad, Alvin Cheung
- **Comparing systems for analyzing big neuroscience imaging data**
  Parmita Mehta, Sven Dorkenwald, Dongfang Zhao, Tomer Kaftan, Magda Balazinska, Alvin Cheung & Ariel Rokem
- **Cosette: An Automated SQL Solver**
  Shumo Chu, Chenglong Wang, Konstantin Weitz, Alvin Cheung, Dan Suciu
- **Deep Curation: Unsupervised Curation of Biological Repositories**
  Maxim Grechkin, Hoifung Poon, Bill Howe
- **Elastic Memory Management for Cloud Data Analytics**
  Jingjing Wang, Magdalena Balazinska
- **Gaussian Mixture Models Use-Case: In-Memory Analysis with Myria**
  Ryan Maas, Jeremy Hyrkas, Olivia Telford, Jake VanderPlas, Magdalena Balazinska, Andrew Connolly
- **Graphulo: native linear algebra in a NoSQL DB**
  Dylan Hutchison, Vijay Gadepally, Jeremy Kepner, Bill Howe
- **PipeGen: Data Pipe Generator for Hybrid Analytics**
  Brandon Haynes, Alvin Cheung, Magdalena Balazinska
- **PerfEnforce: Data Analytics with Performance Guarantees**
  Jennifer Ortiz, Brendan Lee, Magdalena Balazinska
- **Ocean genomic analysis with Myria**
  Dylan Hutchison, Shrainik Jain, Bill Howe, David Maier
- **Viska: Enabling Interactive Analysis of Performance Measurements**
  Helga Gudmundsdottir, Babak Salimi, Magdalena Balazinska, Dan R. K. Ports, Dan Suciu

---

Systems for Analyzing Big Neuroscience Imaging Data @ NeuroFutures 2016

Allen Institute for Brain Science

Presented by Parmita Mehta.

A. Rokem, P. Mehta, S. Dorkenwald,Z. Dongfang, M. Balazinska, A. Cheung.

The size, diversity and complexity of digital brain imaging data has thrust neuroscience researchers into the era of big data.

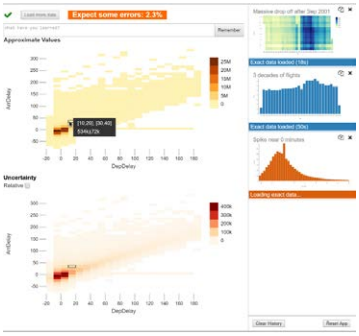
Viska: Enabling Interactive Analysis of Performance Measurements @ OSDI 2016

Presented by Helga Gudmundsdottir.

Helga Gudmundsdottir, Babak Salimi, Magdalena Balazinska, Dan R. K. Ports, and Dan Suciu

Viska is a toolkit for analyzing performance of database systems so that potential causes of anomalous or degraded behavior can be discovered and isolated efficiently.

NEWS



Dominik Moritz: Internship at MSR

This summer Dominik Moritz worked with Danyel Fisher at Microsoft Research on the user experience of approximate query processing (AQP). AQP has the potential to speed up exploratory data analysis but since approximations bear uncertainty analysts often don't trust the results. In our work we addressed these issues and designed a user interface that allows analysts to confirm critical observations.

Laurel Orr: Internship at MSR

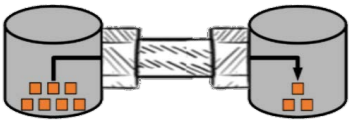
This summer Laurel Orr worked with Srikanth Kandula at Microsoft Research on using on-the-fly sampling to improve performance for approximate query processing on Microsoft's internal big-data system, Cosmos. On-the-fly samplers are where the query processing system samples the data while the query is being executed, and these samplers are beneficial for queries that make multiple passes over the data. The summer work was to use block-level sampling rather than the traditional row-level sampling while still maintaining a high level of accuracy.

Moore / Sloan Data Science Environments Data Science Summit

Magdalena Balazinska and Bill Howe participated in the Moore/Sloan Data Science Environments Data Science Summit as representatives of the University of Washington eScience Institute executive committee. Magda lead two breakout groups on Data Science Education as the Education Working Group lead for the eScience Institute. She also presented a lightning talk on image analytics in Myria. As Associate Director of the eScience Institute, Bill represented UW at the public portion of the event through an overview talk (a session that attracted 100+ national leaders in data science across universities, industry, and funding agencies) and participated in the overall organization of the event. He also organized a Software Working Group breakout session.

AWARDS

Madrona Prize**: Runner Up**



PipeGen: Data Pipe Generator for Hybrid Analytics

Brandon Haynes, Alvin Cheung, Magda Balazinska

VLDB Women in Database Research Award

Magdalena Balazinska for her inspirational research record on scalable distributed data systems.

COMING SOON

UW DB Industry Affiliates Meeting

December 2, 2016

Get an insider's look at current research, meet the students, and have an impact on future directions in the field.

Twitter Website

Computer Science & Engineering at the University of Washington is consistently ranked among the top programs in the nation. We educate tomorrow's innovators, conduct cutting-edge research in the principal areas of the field, lead a broad range of multi-disciplinary initiatives that demonstrate the transformative power of

computer science and computer engineering, and are widely recognized for our success in promoting diversity.

The Paul Allen Center for Computer Science & Engineering

Our home page is:
http://db.cs.washington.edu/

Our people are:
http://db.cs.washington.edu/people.html