Microsoft Research

Interactive and Collaborative Data Management in the Cloud

Magdalena Balazinska
Assistant Professor
University of Washington

Introduction

Microsoft Research

- Assistant Professor - University of Washington
- PhD from MIT in February 2006
  - Advisors: Hari Balakrishnan and Mike Stonebraker
  - Topic: distributed stream processing engines
- Research area: databases and distributed systems
- Microsoft Research Faculty Fellow 2007

Vision - 2006

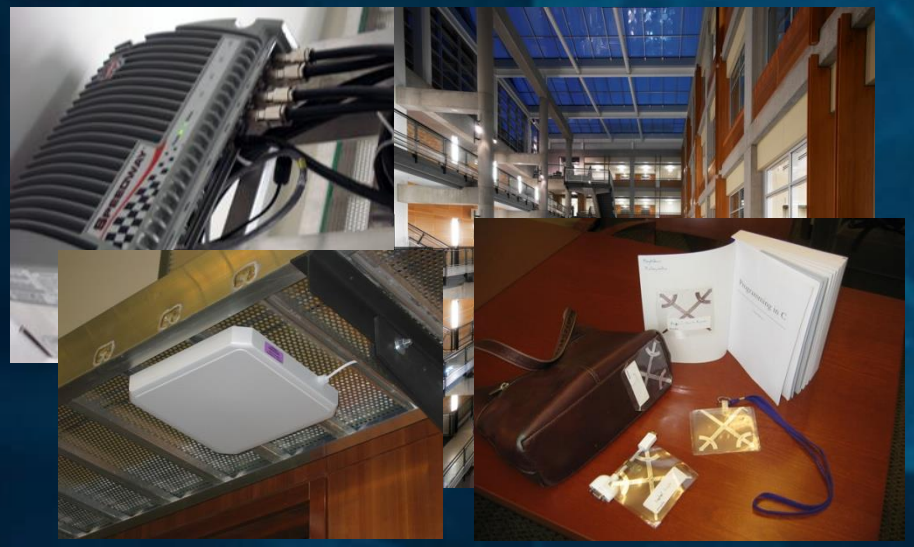
Microsoft Research

- New world
  - Millions of heterogeneous sensors
  - Petabytes of data
  - Real-time, streaming, distributed data

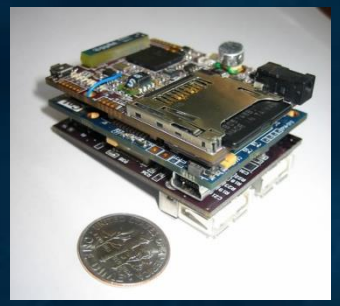
Sensor Deployments

Microsoft Research

Pervasive computing applications

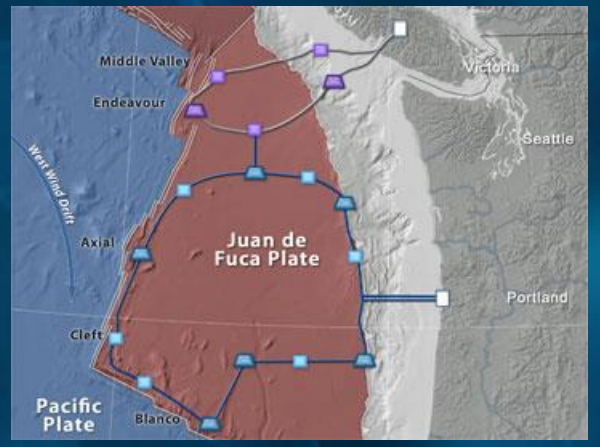


RFID-based tracking @ UW
http://rfid.cs.washington.edu



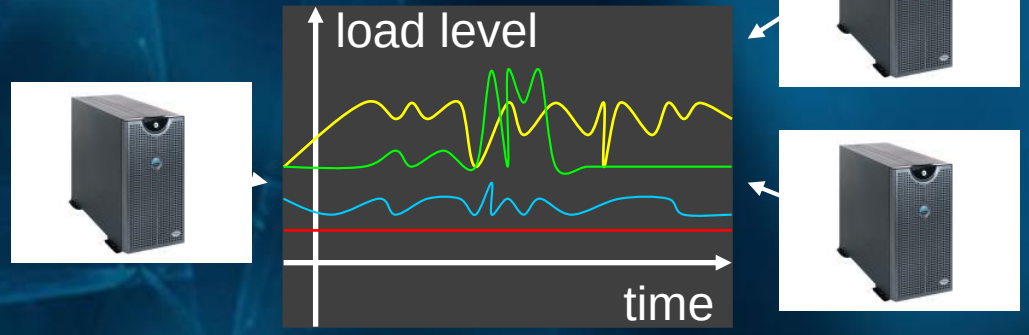
Mobile sensor network @ UW
http://www.cs.washington.edu/homes/yanokwa/

Scientific applications



Neptune project @ UW
http://www.neptune.washington.edu

Computer systems and network monitoring



Research Questions

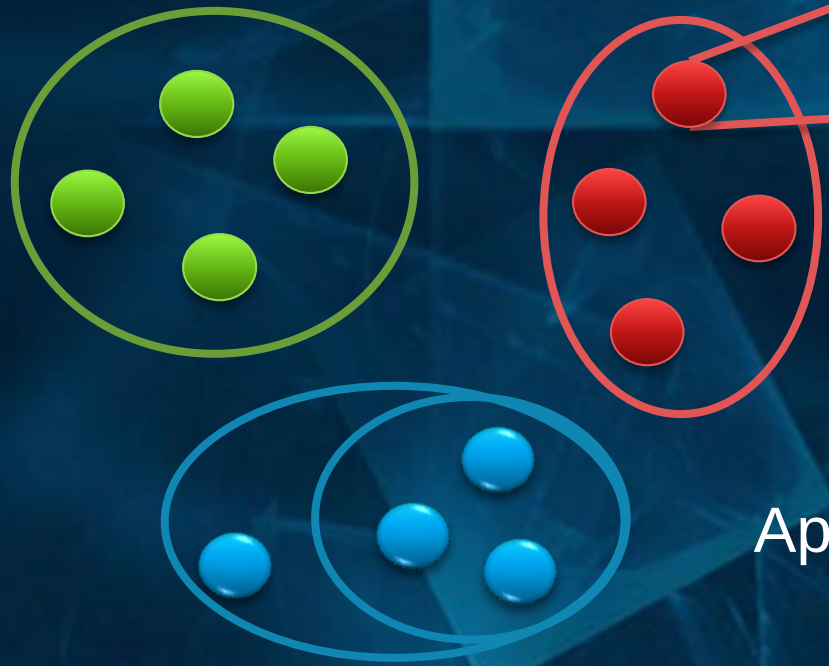
Microsoft Research

- Combining live and archived stream data processing
  - Processing and querying data streams in near real-time
  - Archiving and accessing historical data streams
  - Querying both simultaneously
- Managing noisy, sensor data
  - Managing data errors and ambiguity

Live and Archived Stream Processing

Microsoft Research

- Moirae system, CDM framework, and MCStream algorithm
  - Complex event clustering in streaming fashion



Event

Event ID	Time	Server ID
110345	1:30pm	456

Event context

Apps …
…

Users …
…

Software …
…

Resources …
…

Applications: Any type of monitoring applications

- Ex1: Computer network monitoring
- Ex2: Sensor-based monitoring

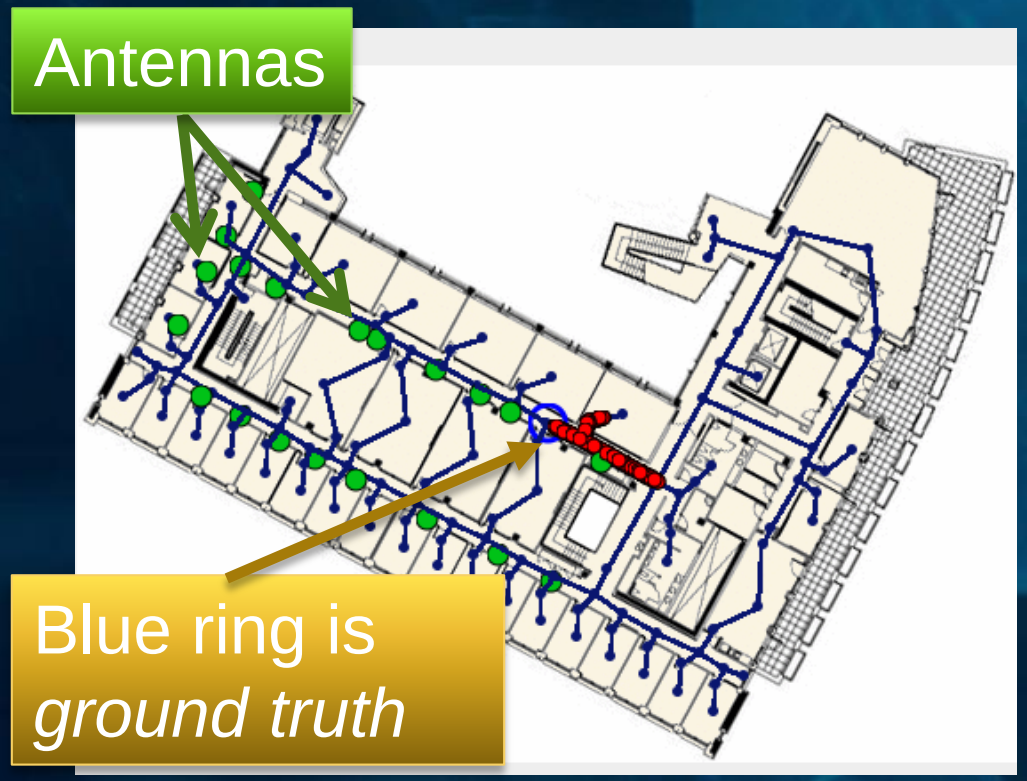
More Information

Microsoft Research

- Project website: http://db.cs.washington.edu/moirae/
- Selected publications
  - Y. Kwon, W. Y. Lee, M. Balazinska, and G. Xu.
    Clustering Events on Streams using Complex Context Information
    MCD 2008
  - Y. Kwon, M. Balazinska, and A. Greenberg:
    Fault-tolerant Stream Processing using a Distributed, Replicated File System
    VLDB 2008
  - M. Balazinska, Y. Kwon, N. Kuchta, and D. Lee.
    Moirae: History-Enhanced Monitoring
    CIDR 2007

Managing Noisy Sensor Data

Microsoft Research



Approach: build a probabilistic model

At 10am, the doctor was either in room 525 (25%) or the hall (75%)

At 10:01am, she was in room 525 (30%) or the hall (70%)

Enable sophisticated queries over the model

If the doctor goes from her office to her patients' rooms and back to her office, generate a "doctor round" event

Managing Noisy Sensor Data

Microsoft Research

- Lahar: Complex event processing over Markovian streams
  - Streams with probabilities and correlations
  - Indexing techniques for Markovian streams
  - Approximation and compression of Makovian streams
- RFID Ecosystem: building-scale RFID deployment
  - 8,000 square meters, 7 floors, and 160 distinct locations
  - 67 participants with more than 300 RFID tags
- Cascadia: Event specification, extraction, and management

More Information

Microsoft Research

- Project websites
  - Lahar: http://lahar.cs.washington.edu
  - RFID Ecosystem: http://rfid.cs.washington.edu

- Selected publications
  - J. Letchner, C. Ré, M. Balazinska, and M. Philipose. Access Methods for Markovian Streams. ICDE 2009
  - C. Ré, J. Letchner, M. Balazinska, and D. Suciu. Event Queries on Correlated Probabilistic Streams. SIGMOD 2008
  - E. Welbourne, K. Koscher, E. Soroush, M. Balazinska, G. Borriello. Longitudinal Study of a Building-wide RFID Ecosystem. Mobisys 2009

How Did My Fellowship Help?

Microsoft Research

- Funding is a chicken-egg problem
  - In order to get $$$, must have preliminary results
  - In order to get preliminary results, must have $$$
- This can be rather stressful!
- Fellowship enables risk-taking
  - I believe X is an important problem!
  - But I don't yet have money for it…. And what if the reviewers don't like it? Can I take this chance?
  - Yes! Can use fellowship until other funding becomes available!

Vision - 2009

Microsoft Research

- Sciences are increasingly data rich
- Scientists need effective tools to manage data
  - Storage, Analysis, Organization, Sharing
- Existing database systems do not meet scientists needs

Use-Case: Astronomy Simulation

Microsoft Research

- Studying evolution of structure in the universe is difficult
- Astronomers rely on large-scale simulations (TB of data)
  - Universe is modeled as a set of particles (gas, stars, dark matter)
  - Particles interact through gravity and hydrodynamics
  - Simulator outputs a snapshot of the universe every few timesteps
  - Astronomers analyze these results

Research Questions

Microsoft Research

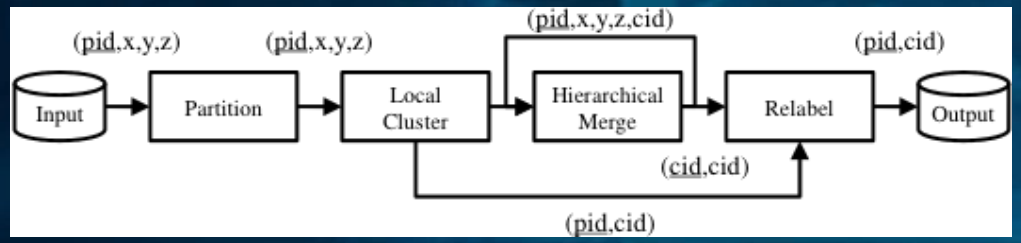
- Long term: Data management as a service for scientists
- Short term: Runtime Query Management
  - How can we facilitate large-scale data analysis?
  - Can we enable users to manage their queries during execution?
- Short term: Offline Query Management
  - Can we develop tools to ease query composition?
  - Can we promote query sharing and reuse?

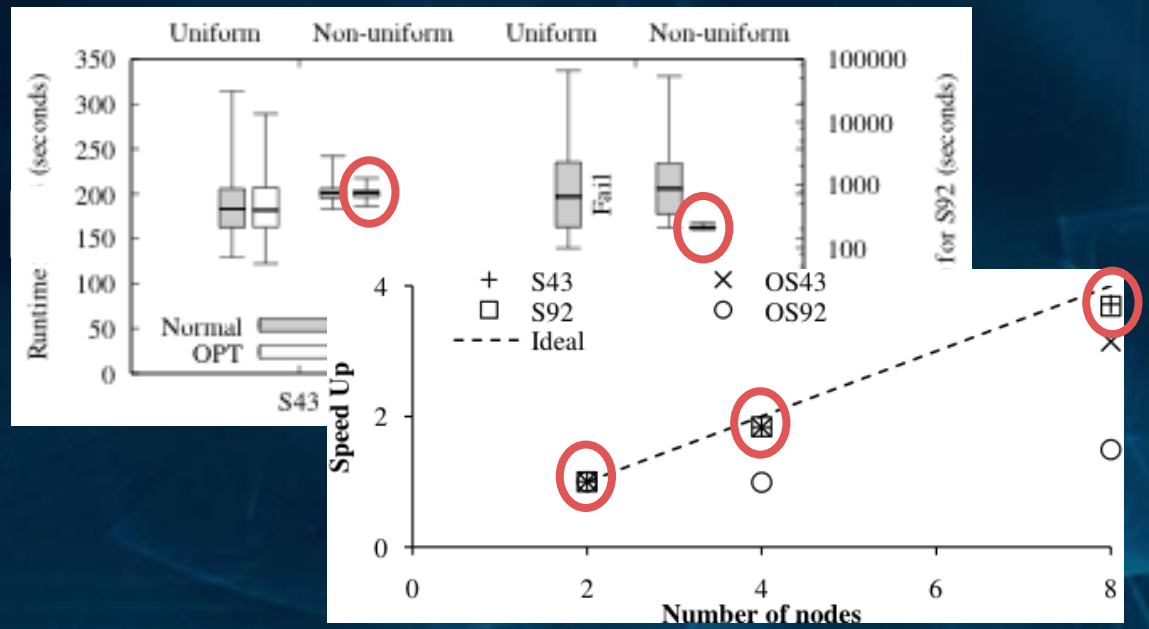
Runtime Query Management

Microsoft Research

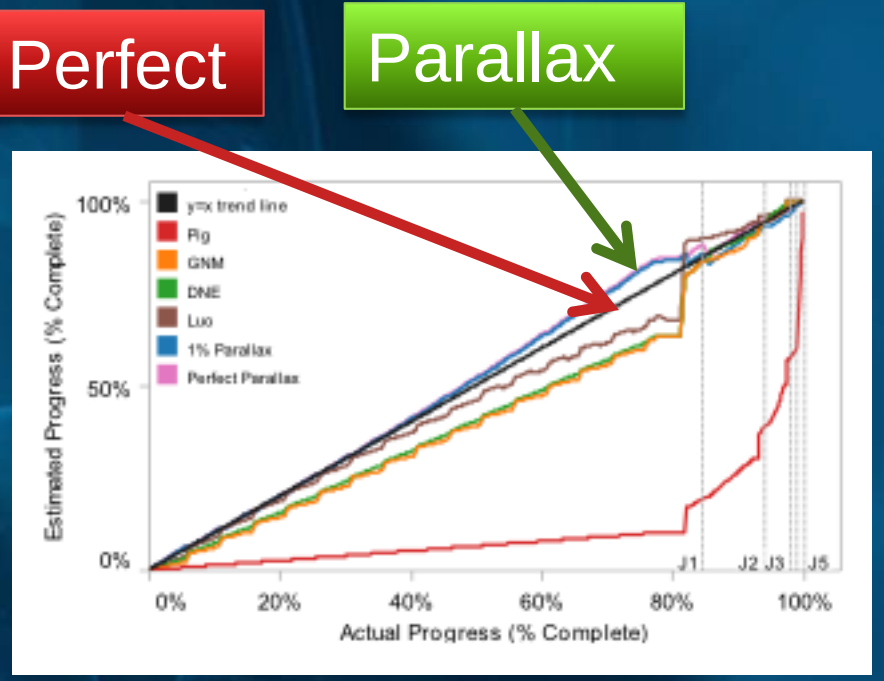
Project name: Nuage

Efficient clustering algorithm for Dryad (written in DryadLINQ)





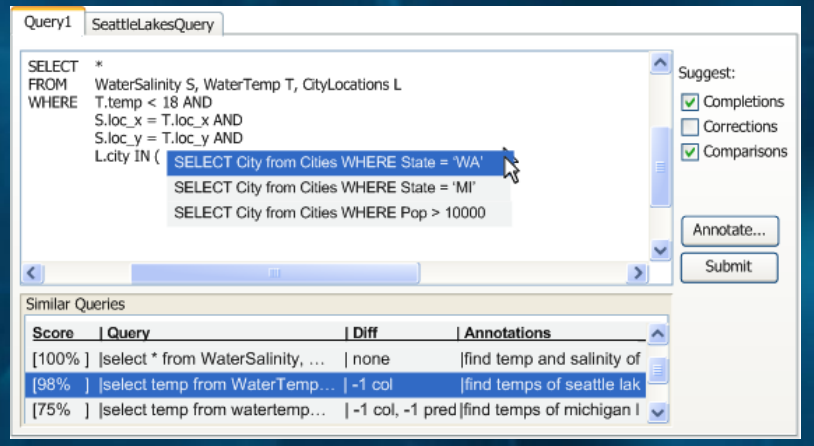
Parallax: Accurate progress indicator for Pig/Hadoop



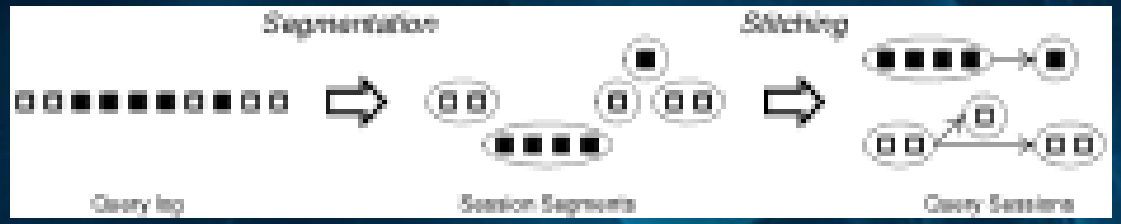
Offline Query Management

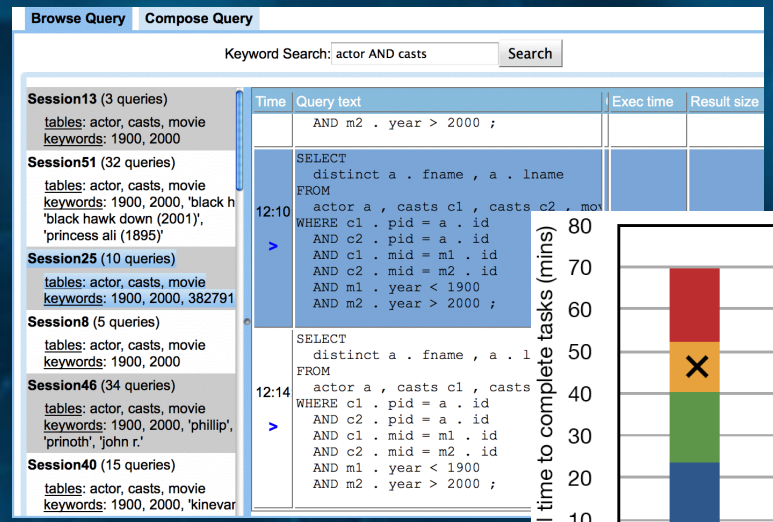
Microsoft Research

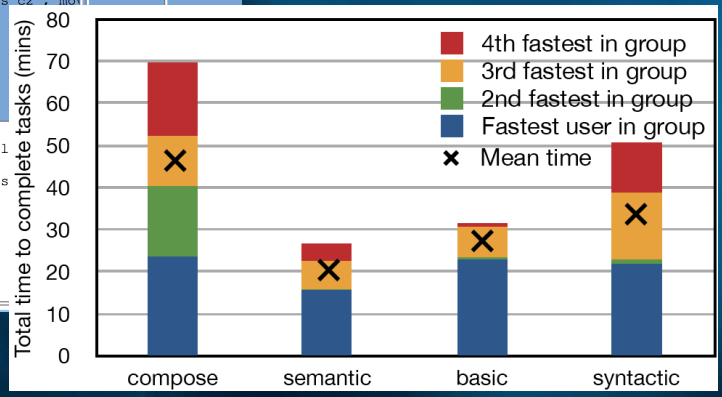
Collaborative Query Management System (CQMS)



Smart Query Browser Improves query reuse







More Information

Microsoft Research

- Project websites
  - Nuage: http://nuage.cs.washington.edu
  - CQMS: http://cqms.cs.washington.edu

- Selected publications
  - Y. Kwon, D. Nunley, J. Gardner, M. Balazinska, B. Howe, and S. Loebman. Scalable clustering algorithm for N-body simulations in a shared-nothing cluster. Tech report. 2009
  - K. Morton, A. Friesen, M. Balazinska, D. Grossman. Toward A Progress Indicator for Parallel Queries. Tech report 2009
  - N. Khoussainova, M. Balazinska, W. Gatterbauer, Y. Kwon, and D. Suciu. A Case for a Collaborative Query Management System. CIDR 2009 (Persp.)

Acknowledgments

Microsoft Research

- Many great students: Nodira Khoussainova, YongChul Kwon, Julie Letchner, Kristi Morton, Emad Soroush, Prasang Upadhyaya, Evan Welbourne, and many undergraduate students

- Many great collaborators: Gaetano Borriello, Jeff Gardner, Wolfgang Gatterbauer, Albert Greenberg, Dan Grossman, Bill Howe, Matthai Philipose, Christopher Ré, Dan Suciu, the SciDB team, and others

Acknowledgments

- This research was partially supported by NSF CAREER award IIS-0845397, NSF grant IIS-0713123, NSF CRI grants CNS-0454425 and CNS-0454394; by gifts from Cisco Systems Inc., Intel Research, and Microsoft Research including a gift under the SensorMap RFP; by an HP Labs Innovation Research Award, by a Mitre contract; and by Magdalena Balazinska's Microsoft Research New Faculty Fellowship

Summary

Microsoft Research

- The greatest impact of this fellowship
- Has been to give me courage to pursue my ideas
- And worry about funding later

- This model seems to be working great so far!

Thank you Microsoft Research!