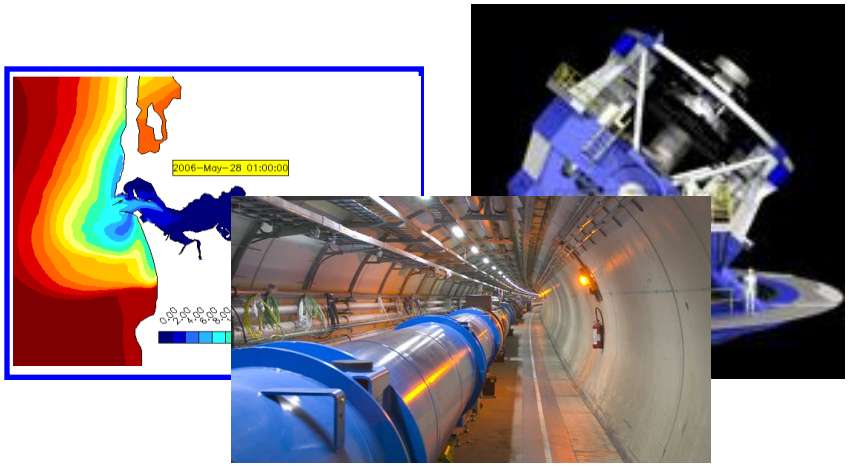
Big-Data Analytics Usability

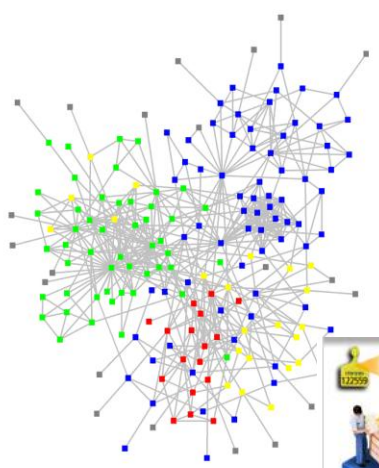
Magdalena Balazinska

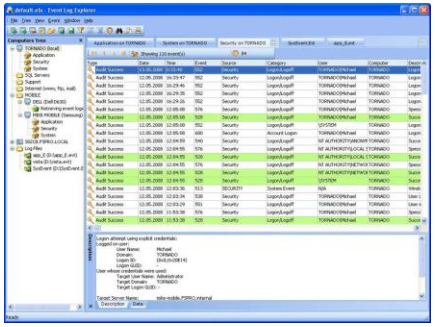
UNIVERSITY OF WASHINGTON

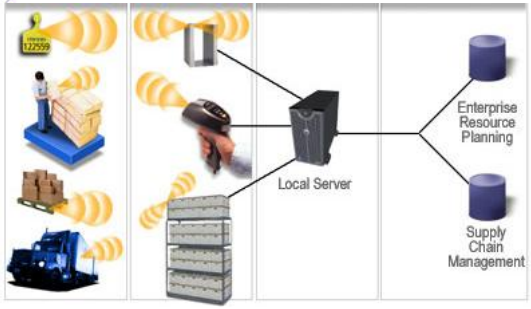
http://www.cs.washington.edu/people/faculty/magda

Big Data Challenges Go Beyond High-Performance Data Processing



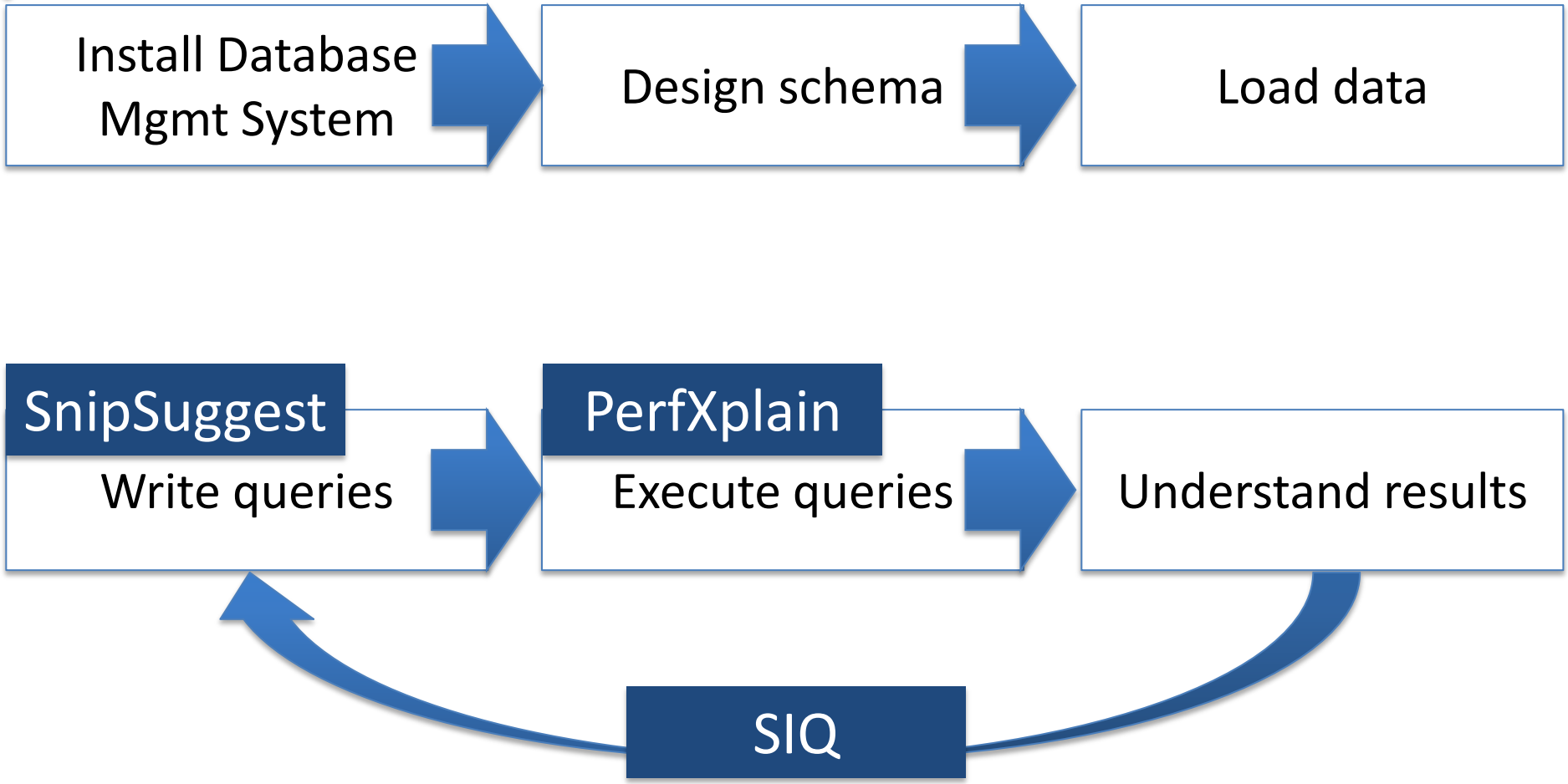




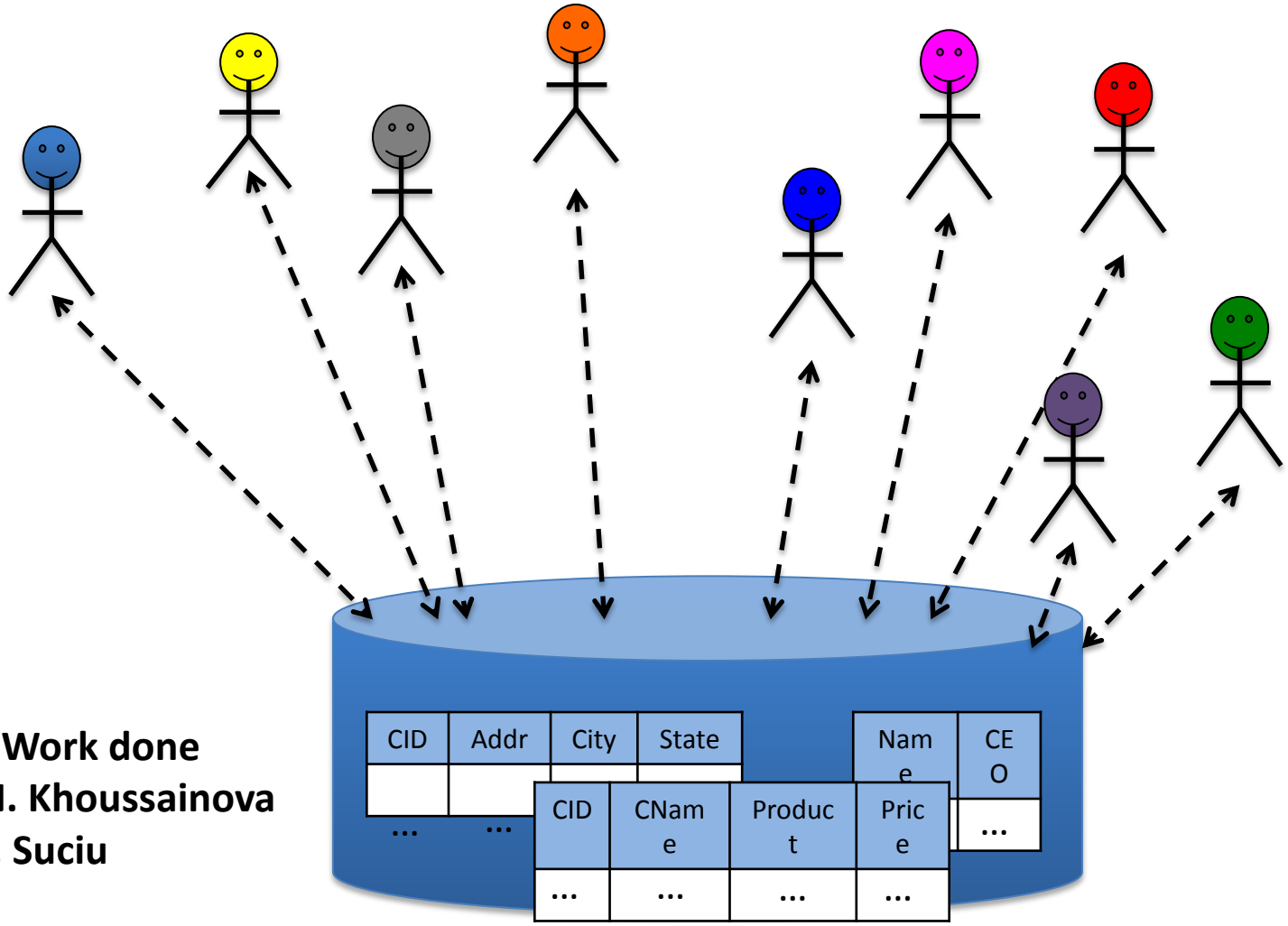


- Big data management requires
  - Tools that ***can be used by data scientists***
  - Experts in their data and their domain
  - But not database or systems experts

Data Management Systems are Known to be Hard to Use



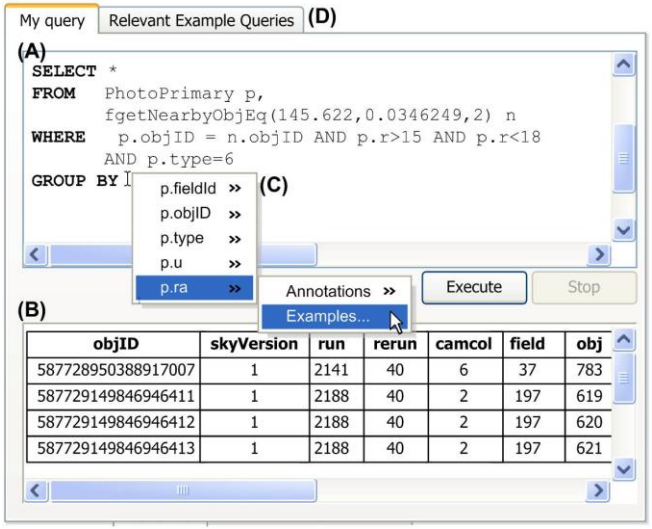
Can Leverage **Collaborative** Nature of Big Data Processing to Ease some Pains



ACKS: Work done with N. Khoussainova and D. Suciu

[KBGKS CIDR 2009, KKBS VLDB 2011]

SnipSuggest



SQL Auto-complete system

- Recommends **snippets** of SQL
- **On-the-go** as a user types query
  - Recommendations are **context-aware**

- **Mines a query log** and curates its set of recommendations from **similar past queries**

- Evaluated on 10,000 queries over Sloan Digital Sky Survey (10-fold validation)
  - Achieves **93.7% average precision**, and
  - a **mean response time of 14 ms** per query

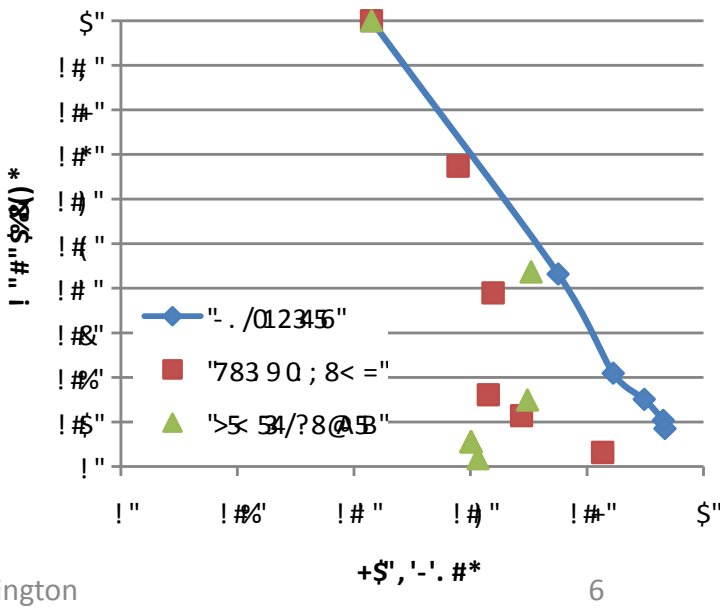
PerfXplain

Performance explanations for MapReduce systems

- Enables users to ask performance questions
- Explains **problems with configuration parameters, cluster conditions, the data, and the query**
- **Mines the log** of past query executions to find answer

Why was the second job as slow as the first job? I expected it to be much faster!

Because your DFS block size was large.



Other Tools

Other tools being developed in the UW DB group:

- SIQ: Tool for generating sample databases for query debugging (resistant to query edits)

- Tool for a unified, visual analytics process comprising
  - Data integration
  - Data cleaning
  - Visual data exploration

- Rethinking what it means to offer data processing and analysis as a service