Scientific Data Management in the Cloud: Overview of the Nuage Project

Magdalena Balazinska
University of Washington
http://www.cs.washington.edu/homes/magda
• Science is becoming a data management problem
• Existing database management systems are insufficient
  – Wrong data model, wrong features, insufficient scalability

• Nuage project goals (http://nuage.cs.washington.edu/)
  – Focus on scientific applications
  – Massive-scale parallel query processing
  – Cloud computing: DBMS as a service for science

• Current application domains

Oceanography
Bill Howe, UW eScience

Astronomy
Jeff Gardner
Andrew Connolly
People

• **Students**: Nodira Khoussainova, YongChul Kwon, Sarah Loebman, Kristi Morton, Emad Soroush, and Prasang Upadhyaya,

• **Collaborators**: Andrew Connolly, Jeff Gardner, Dan Grossman, Bill Howe, and Dan Suciu
Astronomy Simulation Use Case

- Evolution of large scale structure in the universe
  - Universe is a set of particles (gas, dark matter, stars)
  - Particles interact through gravity and hydrodynamics
  - Output snapshot every few simulation timesteps

- Analysis needs:
  - Select-project-join (SPJ) queries over snapshot data
  - Data clustering within snapshot
  - SPJ and recursive queries over clustered data

<table>
<thead>
<tr>
<th>Simulation</th>
<th>No. Particles</th>
<th>Snapshot Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbtest128g</td>
<td>4.2 million</td>
<td>169MB</td>
</tr>
<tr>
<td>cosmo50</td>
<td>33.6 million</td>
<td>1.4 GB</td>
</tr>
<tr>
<td>cosmo25</td>
<td>916.8 million</td>
<td>36 GB</td>
</tr>
</tbody>
</table>

Few dozen to few hundred snapshots per run
Astronomy Simulation Use Case

- Implemented SPJ queries over raw data
  - Relational DBMS (single site and distributed)
  - Pig/Hadoop
  - IDL: State-of-the-art in astronomy

- Designed, implemented, and evaluated algorithm for parallel data clustering
Current Research Focus

• Given magnitude of data and queries
• Need more than efficient query processing

• Users need tools for managing queries at runtime:
  – Accurate, time-based progress indicators
  – The ability to see representative partial results
  – The ability to suspend and resume queries
  – Intra-query fault-tolerance
  – Agile query scheduling, resource management, skew handling

• All this without too much runtime overhead
Nuage Project

- **Nuage project goals** (http://nuage.cs.washington.edu/)
  - Focus on scientific applications
  - Massive-scale parallel query processing
  - Cloud computing: DBMS as a service for science
DBMS As a Service for Science

• **SciFlex**: A Cross-scale Cross-domain Scientific Data Management Service
  – Schema recommendation & data upload utilities
  – Query, archive, and visualization services
  – Data intensive computing!
  – Data, schema, and tool sharing + tool recommendation
  – Annotations, tagging, disagreement, discussions
  – Security: need to share safely
  – SLAs for science

• Interesting systems issues involved in building SciFlex

• In collaboration with Microsoft Research
Conclusion

• Sciences are increasingly data rich

• Need efficient, large-scale query processing

• Need other data management services too

• Nuage/SciFlex project strives to address these needs
Acknowledgments

• This research is partially supported by
  – NSF CAREER award IIS-0845397
  – NSF Cluster Exploratory Award IIS-0844572
  – NSF CRI grant CNS-0454425
  – An HP Labs Innovation Research Award
  – Gifts from Yahoo!
  – Gifts from Microsoft Research
  – Balazinska's Microsoft Research Faculty Fellowship
  – University of Washington's eScience institute.