Expedited Rating of Data Stores Using Agile Data Loading Techniques

Sumita Barahmand and Shahram Ghandeharizadeh

USC Database Laboratory
Outline

- Performance Measurement
- Agile Data Loading Techniques
- Results
- Conclusion
Why Measure Database Performance?

• In recent years many different data stores have been developed [1].

• Is important for data store vendors to identify inefficiencies in their systems and try to improve it.

• Plays an important role in the application developer’s decision to utilize a data store.
  • Compare various data store with one another.
  • Identify tradeoffs.
  • Understand how suitable a solution is.

[1] Rick Cattel, Scalable SQL and NoSQL Data Store, SIGMOD Record, December 2010 (Vol. 39, No. 4).
**Motivation**

- Performance of a system depends on the workload issued against the system.
- The workload is characterized by the application.
- Realistic performance is measured with a realistic workload.

- **OLTP:**
  - Workload: Entering and delivering orders and recording payments.
  - Data: Customers and items.
- **Marketing applications:**
  - Workload: Providing recommendations and Sentiment analysis.
  - Data: users, user clicks and user reviews.
- **Social network:**
  - Workload: View profile, List friends and Add friends.
  - Data: Users, friendships, resources.
• Populating data store with realistic data is time consuming.

• Performance measurement is an iterative task.
  • Measure performance for different workloads in an application.
  • Fine tune and measure performance for a given workload.
  • Rating the data store: Find maximum capability for a given workload.
Agile Data Loading Techniques

• Use a high performance external tool to create the data in a native on disk format for a data store \[1\].

• Contributions:
  - *Database Image Loading (DBIL)*:
    • Relies on the capability of a data store to create a disk image of the database.
    • Uses this image repeatedly across different experiments.
    • Reduces 11 days to 30 minutes.
  - *RepairDB*:
    • Restores the database to its original state prior to the start of an experiment.
    • Reduces 11 days to 10 hours.
  - *Loadfree*: See paper for details.
    • Reduces 11 days to 0.

Database Image Loading (DBIL)

- This technique first populates the data store with the data and then creates a disk image of that for further use.

- Load time depends on how quickly the system copies the files pertaining to the database.
• At the end of each experiment it employs point-in-time recovery of a data store to restore the state of the data store.
• Use with workloads consisting of infrequent write actions.
• Depending on the data store and the workload issued, RepairDB may be faster/slower than actually reconstructing the database.
• For data stores with no point-in-time recovery mechanism, rating framework may implement RepairDB.
## Results

<table>
<thead>
<tr>
<th>Social Graph Size</th>
<th>Actual Construction</th>
<th>DBIL</th>
<th>RepairDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>100k</td>
<td>2.5 hrs</td>
<td>8 min</td>
<td>26 min</td>
</tr>
<tr>
<td>500k</td>
<td>6 hrs</td>
<td>10 min</td>
<td>2.5 hrs</td>
</tr>
<tr>
<td>1M</td>
<td>11 days</td>
<td>31 min</td>
<td>10 hrs</td>
</tr>
</tbody>
</table>

Loading time for BG Framework
Conclusions

• Understanding the performance behavior of data stores is important both for data store vendors and application developers.
• Performance evaluation/rating is an iterative task which may require populating the data store over and over.
• Populating a data store with large amount of data is time consuming.
• Without the use of agile loading techniques the rating process may not be feasible.
Questions?

THANK YOU