Modeling Mixed Workloads for Large-Scale Database Benchmarks

Dina Bitton, Ph.D., Boris Gelman, Ph.D., Hualin Wang, Ph.D.
SAP Labs LLC, Palo Alto, CA
Why New Benchmarks are Needed?

- “Real-world“ data to be represented in benchmarks
- Data sizes are growing dramatically
- Categories of “real world” applications have to be represented in benchmarks:
  1. Data refresh functionality
  2. Transaction management
  3. Real-time analytics
  4. Advanced application functions / processes
- New technologies are evolving:
  1. SQL / non-SQL / Search-based data processing
  2. In-memory databases
  3. Cloud computing
- Existing benchmarks performance metrics have to be re-evaluated to make them more relevant
Categories of Enterprise Applications

• TPC-C, TPC-DS, TPC-E, TPC-H, … were the earlier efforts to represent categories of enterprise applications

• Enterprise applications are evolving, and new categorization becomes necessary

• Dimensions of application categorization:
  1. Functional Categorization:
     ➢ Sales and Distribution
     ➢ CRM
     ➢ Financial, etc
  2. Structural Categorization:
     ➢ Insert only
     ➢ Insert and update
     ➢ Real-time analytics
     ➢ Large-scale analytics, etc
  3. Categorization of Data:
     ➢ Simple data structures
     ➢ Rich / complex data structures
     ➢ Unstructured data
Modeling Enterprise Data

• **Richer benchmark data needed:**

  1. Wide tables:
     - ERP tables have up to 300 columns
  2. Realistic data types:
     - Densely and sparsely populated columns
     - Variety of realistic data types
  3. Skews and distributions – challenges for conventional optimizers:
     - Non-uniform data distributions
     - Typical cases to be represented
  4. Realistic Cases of Big Data:
     - Functional cases
     - Industry-specific cases:
       a) Healthcare
       b) Telecommunication
       c) Stock Trading
       d) Retail, etc.
Need for Data Generators

• **Data Generators:**
  1. Data Generators were originally introduced with TPC-C, TPC-DS, … benchmarks

• **New Benchmarks require new Data Generators**
  1. Data generation rules to be based on profiling of “real world” data
  2. Scaling factors to have better business meaning

• **Examples:**

<table>
<thead>
<tr>
<th>Table Name</th>
<th>No of Columns in TPC-H Benchmark</th>
<th>No of Columns in ERP Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orders</td>
<td>9</td>
<td>149</td>
</tr>
<tr>
<td>Lineitem</td>
<td>16</td>
<td>250</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column Name</th>
<th>ERP Database Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales Organization</td>
<td>• Large SO – 62 % of rows</td>
</tr>
<tr>
<td></td>
<td>• Medium SO – 10 % of rows</td>
</tr>
<tr>
<td></td>
<td>• Small SO – 1 % of rows</td>
</tr>
<tr>
<td></td>
<td>• BA1 – 3% of rows</td>
</tr>
<tr>
<td></td>
<td>• BA2 – 49% of rows</td>
</tr>
<tr>
<td></td>
<td>• BA3 – 2% of rows</td>
</tr>
<tr>
<td>Business Area</td>
<td>• ...</td>
</tr>
<tr>
<td></td>
<td>• ...</td>
</tr>
</tbody>
</table>
Data Processing Patterns

• **Data refresh patterns:**
  1. **“Old” ERP style:** Data inserts and related data updates
  2. **“New” Big Data style:** Streams of data inserts representing real-time events

• **Analytics and Reporting:**
  1. **“Old” style:**
     - Reports are based on Materialized Views, or Special Tables
     - Real-Time has limited definition
     - Metrics are not truly dynamic
  2. **“New” style:**
     - Real-time is for real
     - No pre-materialization of results
     - Any metric can be computed dynamically, even so called non-materializable metrics

• **Special Processes**
  1. Advanced Analytics
  2. Simulations, such as Financial Simulations
Prototyping Future Benchmarks – Performance Tests

• **Designing and Implementing Performance Tests:**
  1. Performance test is the first step towards open future benchmarks
  2. Recently, several performance tests were implemented at SAP:
     - SAP-H – modified TPC-H benchmark based on enriched TPC-H Data Model
     - Large-scale performance test for real-time, dynamic analytics based on sales and distribution data
  3. Several performance tests are currently under development at SAP

• **Performance Tests as Benchmark Candidates:**
  1. Performance tests can be presented to the Benchmark Council as benchmark candidates for specific application categories
  2. Performance tests will go through discussion and approval process to become open performance benchmarks
  3. We would like to encourage technology companies to become more active in designing and implementing performance tests, and presenting them to the community
Thank You!

Contact information:

Boris Gelman, Ph.D.
VP Development
SAP Labs, LLC
3412 Hillview Ave, Palo Alto, CA 94304
boris.gelman@sap.com