Benchmarking Linked Open Data technology

SRbench: A Benchmark for Streaming RDF Storage Engines

Ying Zhang, Peter Boncz (CWI, Amsterdam)
What is Database Benchmarking?

Standard test to measure and understand how technology performs

- Dataset definition
  - at various scales (100GB, 300GB, 1TB, 3TB, etc)
  - mimicks a recognizable relevant usage scenario

- Database Queries
  - often between 10-100 queries, with parameters
  - + rules/programs that specify how these queries are posed

- Result Metrics
  - a number to understand the result
  - tps = “transactions/second”
  - $/QphH@size = “price per query per hour”

- Audit Rules
  - allow results to be checked by independent auditors
  - prevent/limit cheating
Why Benchmarking?

- make competing products comparable
- accelerate progress, make technology viable
Benchmarking LOD Technology

LOD = Linked Open Data

- web addressable data $\rightarrow$ RDF data format (W3C)
- lots of useful data on the web ("LOD cloud")
Benchmarcking LOD Technology

LOD = Linked Open Data
- web addressable data ➔ RDF data format (W3C)
- lots of useful data on the web ("LOD cloud")

LOD technology (SPARQL) benchmarks:
- BSBM, DBpedia Benchmark, SIB
- SRbench ← topic of this talk
- New industry cooperation:
vendor cooperation to establish accepted RDF/Graph database benchmarks and benchmark results

* tentative/expected project
LDBC Goals

1. Create the LDBC Foundation of graph and RDF DB vendors

2. Equip de LDBC Foundation with a good initial set of benchmarks, and benchmark results

spin-off
Benchmarking
Linked Open Data technology

**SRbench: A Benchmark for Streaming RDF Storage Engines**

Ying Zhang, *Peter Boncz* (CWI, Amsterdam)
SRbench: **Streaming** RDF Benchmark

Traditional Database System vs.

**Stream** Database System

- Persistent Data
- Persistent Queries
  - "continuous queries"
  - "push" based query answering

- Data stream
- Stream of queries
- "pull" based query answering
Data Streams (1/4): Stock Market

Finance

26.90  -0.82  (-2.96%)

Pre-market: 26.71  -0.19  (-0.71%)

NASDAQ real-time data - Disclaimer

Company name | Price  | Change  | Chg %  | M Cap  | Valuation
---|---|---|---|---|---
FB  Facebook Inc  | 26.90  | -0.92  | -3.96%  | 57.51B  | 9.56
LVWD  Livewire, Inc  | 0.210  | 0.000  | 0.00%  | 6.96M  | 0.24
IMIN  Immersive Tech(NDA)  | 1.50  | 0.00  | 0.00%  | 23.80M  | 0.11
THWM  TheWrap Inc  | 0.0430  | 0.0000  | 0.00%  | 2.44M  | 0.01
CCLG  CycleLogic, Inc  | 0.0001  | 0.0000  | 0.00%  | 10.00  | 0.00
LNKD  LinkedIn Corporation  | 91.69  | -0.42  | -0.46%  | 9.41B  | 9.47
TTGT  TechTarget, Inc  | 4.92  | -0.13  | -2.57%  | 193.47M  | 4.07
VLOG  MergeLogic Corporation  | 0.0032  | 0.0000  | 0.00%  | 18.500  | 0.01
MRGN  MergeLogic Corporation  | 0.0035  | 0.0000  | 0.00%  | 1.58M  | 0.01
VLOG  VillageEDEOS, Inc  | 0.360  | 0.0000  | 0.00%  | 139.455  | 0.01
YVNO  Yahoo, Inc  | 26.90  | -0.82  | -3.96%  | 57.51B  | 9.56

Volatility: High

30-day average volatility: 23.80%

Key stats and ratios

<table>
<thead>
<tr>
<th>Stat</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net profit margin</td>
<td>19.99%</td>
</tr>
<tr>
<td>Operating margin</td>
<td>36.64%</td>
</tr>
<tr>
<td>EBITDA margin</td>
<td>55.02%</td>
</tr>
<tr>
<td>Return on assets</td>
<td>12.47%</td>
</tr>
<tr>
<td>Return on equity</td>
<td>12.22%</td>
</tr>
<tr>
<td>Employees</td>
<td>3,539</td>
</tr>
<tr>
<td>Carbon Disclosure Rating</td>
<td>-</td>
</tr>
</tbody>
</table>

Address
Data Streams (2/4): Social Chatter

- Detect breaking news
- Analyze Marketing campaigns
Data Streams (3/4): Car Traffic

- monitor positions and speeds of cars ➔ detect accidents, traffic jams
- Applications: better safety, improved logistics
Monitor health of elderly in their homes

**Why?**
- Difficult to reach locations
- Make health care more affordable

**How?**

Who are the users?

---

*Figure 1. Basic telehealth system.*
Streaming RDF data benefits:

- apply Linked Open Data (LOD) principles to streaming data
  - Link streaming data to data on the web (enrichment)
  - Publish data streams on the web
- support (simple) reasoning semantics in stream queries

⇒ Richer semantics than relational streaming database systems
SRbench: Streaming **RDF** Benchmark

Streaming RDF data **challenges:**

- Proper benchmark dataset
  - use real-world datasets from LOD

- No standard query language
  - natural language query definition +
    three implementations (SPARQLStream, CQELS, C-SPARQL)

- Limited systems support
  - evaluate on the strRS system (UPM)
SRbench: used Datasets

Use case: weather information application

LinkedSensorData

LinkedObservationData

Observation

om-owl:procedure

ResultData

Instant

om-owl:samplingTime

om-owl:result

MeasureData

TruthData

System

om-owl:hasLocatedNearRel

Point

om-owl:processLocation

LocatedNearRel

om-owl:hasLocation

DBpedia

Airport

owl:sameAs

GeoNames

Feature

om-owl:hasLocation

Ying Zhang, Peter Boncz – Benchmarking Linked Open Data Technology

June 7, 2012 @EDF Copenhagen
<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
<th>Q11</th>
<th>Q12</th>
<th>Q13</th>
<th>Q14</th>
<th>Q15</th>
<th>Q16</th>
<th>Q17</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 patter matching</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A,F</td>
<td>A,F,U</td>
<td>A,F</td>
<td>A,F</td>
<td>A,F</td>
<td>A,F</td>
<td>A,F</td>
<td>A,F</td>
</tr>
<tr>
<td>3 query form</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 reasoning</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 CQL feature</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>7 data access</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 2.** Addressed features per query. Operators are abbreviated in per row unique capital letters, defined as: 1. **And**, **Filter**, **Union**, **Optional**; 2. **Projection**, **Distinct**, **Limit**; 3. **Select**, **Construct**, **Ask**; 4. **Aggregate**, **Subquery**, **Negation**, **Expr in select**, assign**M**ent, **Functions&operators**, **Property path**; 5. **subClassOf**, **subPropertyOf**, **owl:sameAs**; 6. **Time-based window**, tu**P**le-based window, **I**stream, **D**stream, **R**stream; 7. **LinkedObservationData**, **LinkedSensorMetadata**, **GeoNames**, **Dbpedia**.
Summary

- the importance of
  - Database System Benchmarking
  - RDF Database System Benchmarking
  - Streaming RDF Database System Benchmarking

- SRbench
  - Developed in PlanetData (CWI, UPM)
  - First dedicated streaming RDF/SPARQL benchmark

- SRbench future work:
  - performance evaluation
  - results verification (not easy!)
Thank You!

Questions?

- Ying Zhang (zhang@cwi.nl)
- Peter Boncz (boncz@cwi.nl)