TPCx-BB (BIGBENCH)

Big Data Benchmark
The end-user Perspective of Big Data benchmarks

- Analyze Volume of data in Defined SLA
- Select Right Framework At Optimum TCO
What Makes a Good Benchmark?

<table>
<thead>
<tr>
<th>Comprehensive</th>
<th>Usable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage (usecase, components)</td>
<td>Easy to use Kit</td>
</tr>
<tr>
<td>Reliable Proxy implementation</td>
<td>Simplified Metric</td>
</tr>
<tr>
<td>Target usage</td>
<td>Support and Maintenance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Based on Open Standards</th>
<th>Flexible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peer Reviewed Spec and Code</td>
<td>Adaptive</td>
</tr>
<tr>
<td>Public Availability</td>
<td>Modularized</td>
</tr>
<tr>
<td>Industry Acceptance</td>
<td>Reusable</td>
</tr>
</tbody>
</table>
Benchmark Types, use and limitation

**Micro-Benchmarks**
- Basic insights
- SPEC CPU, DFSIO
- Reduced complexity

**Functional Benchmarks**
- Specific High-level function
- Sort, Ranking
- Limited Applicability

**Application Benchmarks**
- Relevant real-world usage
- TPC-H, TPC-E
- Complex implementation
Big Data Benchmark Pipeline

- Functions
  - Ingest
  - Access
  - Computation
- Platform
  - Distros
  - API
  - On-premise/Cloud
- Support
  - Framework
  - Governance
  - Usability

Data Processing
- Structured
- Semi-Structured
- Un-Structured

Data Storage
- Files
- KV Store
- SQL
- Objects/Graph

Data Analytics
- ML
- Reporting
- Search
- Graph

Use case
- Spec
- Datagen
- orchestration
- Metrics
- Audit
- Publications
The Case for Standards and Specifications

Industry Standard Benchmark

Benchmark usage
- Performance Analysis
- Reference Architectures, Collaterals
- Influence Roadmaps and Features
- Aid Customer Deployments
- Illustrative ≠ Informative

Industry Standards

Hardware Software value proposition
Technology value proposition
Meaningful, Measurable, Repeatable
Introducing TPCx-BB (BigBench)

Origin
• Presented in 2013 Sigmod paper¹
• Adapted from McKinsey Business cases²
• TPC standardization 2014

Features
• 30 Batch Analytic use-cases
• Framework Agnostic
• Structured, Semi-Structured & Unstructured

Collaboration
• Papers and Standardization
• Industry and Academic Presentations
• Specification and Publications

Partners and Contributors

¹ http://dl.acm.org/citation.cfm?id=2463712&preflayout=flat
TPCx-BB v1.0 Implementation

Implementation
- Self Contained Kit
- SQL on Hadoop API’s
- Spark MLLIB Machine Learning
- Natural Language Processing

Kit Features
- Easy setup
- Size and Concurrency Scaling
- Versatile Driver
- Modular to support New APIs’

Availability
- TPCx-BB v1.0
- Download from TPC
- Open to Contribution

Metrics and Reporting

\[
BBQpm@SF = \frac{SF \times 60 + M}{T_{LD} + \frac{1}{T_{PY}} + T_{TT}}
\]

\[
\frac{S}{BBQpm@SF} = \frac{c}{BBQpm@SF}
\]

Supported Hadoop Engines
- TEZ
- HIVE
- Spark
- SQL
- APACHE DRILL

Modular

Kit Structure

Supported Hadoop Engines

Availability
- TPCx-BB v1.0
- Download from TPC
- Open to Contribution

Metrics and Reporting

\[
BBQpm@SF = \frac{SF \times 60 + M}{T_{LD} + \frac{1}{T_{PY}} + T_{TT}}
\]

\[
\frac{S}{BBQpm@SF} = \frac{c}{BBQpm@SF}
\]
Test Phases

- **Load Test**
  - Data Aggregated from various sources
  - Permute and Loads Aggregated Data
  - Test's Storage, Compute, Codecs, Formats etc

- **Power Test**
  - Each Use Case runs once
  - Helps Identify Optimization areas
  - Varied utilization Pattern

- **Throughput Test**
  - Multiple Jobs running in parallel
  - Realistic Hadoop Cluster usage
  - Optimize for Cluster Efficiency

Simulate real-world usage

Use case driven utilization pattern

Coherently test Hardware and Software
Benchmark Workflow

**Cluster Setup**
- Min 3 Nodes
- CDH 5.5, HDP 2.3
- Concurrent Streams 2-n

**Define Parameters**
- Scale Factor 1TB,3TB
- HoMR, HoS,HoT
- Tuning Parameters

**Run Benchmark**
- Data Generation
- Load, Power, Throughput Tests

**Analyze Data**
- Reports
- Metrics
- Utilization

**Results**
- Publication
- Collateral
- POC