

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?

Comprehensive

- ☐ Coverage (usecase, components)
- ☐ Reliable Proxy implementation
- ☐ Target usage

Usable

- ☐ Easy to use Kit
- ☐ Simplified Metric
- ☐ Support and Maintenance

Based on Open Standards

- ☐ Peer Reviewed Spec and Code
- ☐ Public Availability
- ☐ Industry Acceptance

Flexible

- ☐ Adaptive
- ☐ Modularized
- ☐ Reusable

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



William Vambenepe @vambenepe

24 Oct

MapR sets new world record for Hadoop **TeraSort**. On Google Compute Engine. insights.wired.com/video/mapr-goo...

[Expand](#)



Rohit Valia @rohitvalia

24 Oct

IBM demonstrates breakthrough [#hadoop](#) 100TB **Terasort** results on private cloud. More details at my session [@strataconf](#) tinyurl.com/HiPerfHadoop

[Expand](#)



Nyrra Jvggranhre @nyrrajvggranhre

24 Oct

Stop using terasort as a benchmark! It is almost always very far from reality or measures different things than you think it does!

[Expand](#)



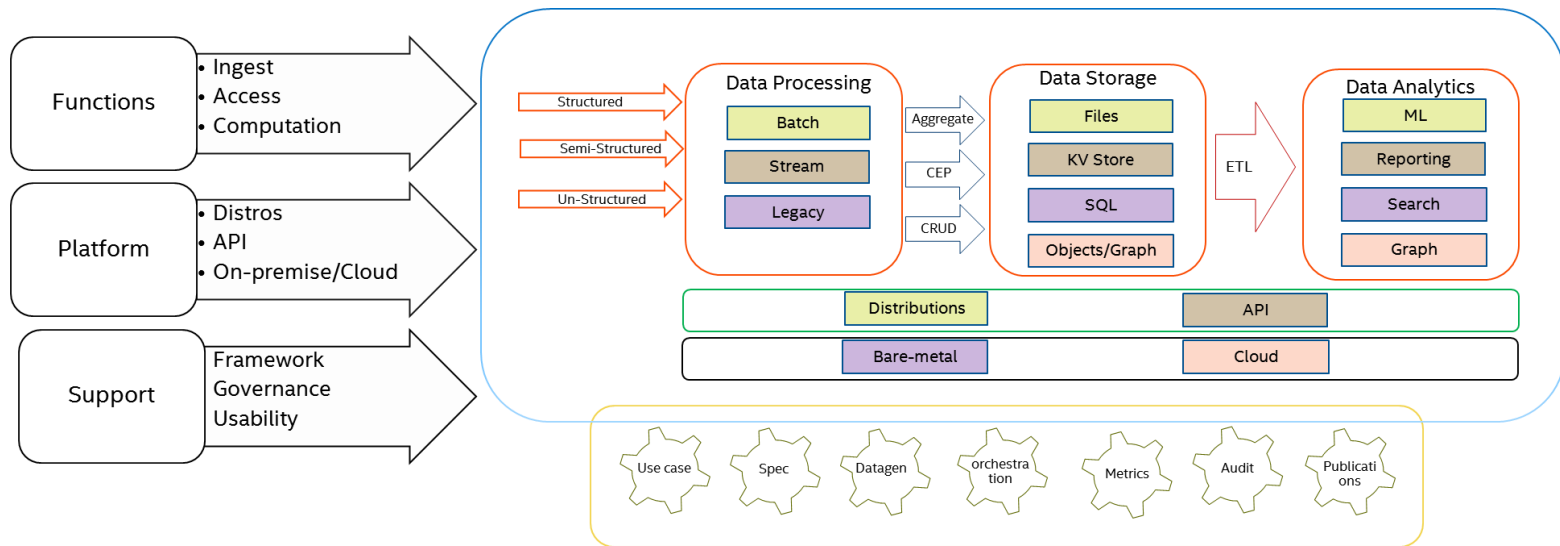
Andrew Purtell @akpurtell

24 Oct

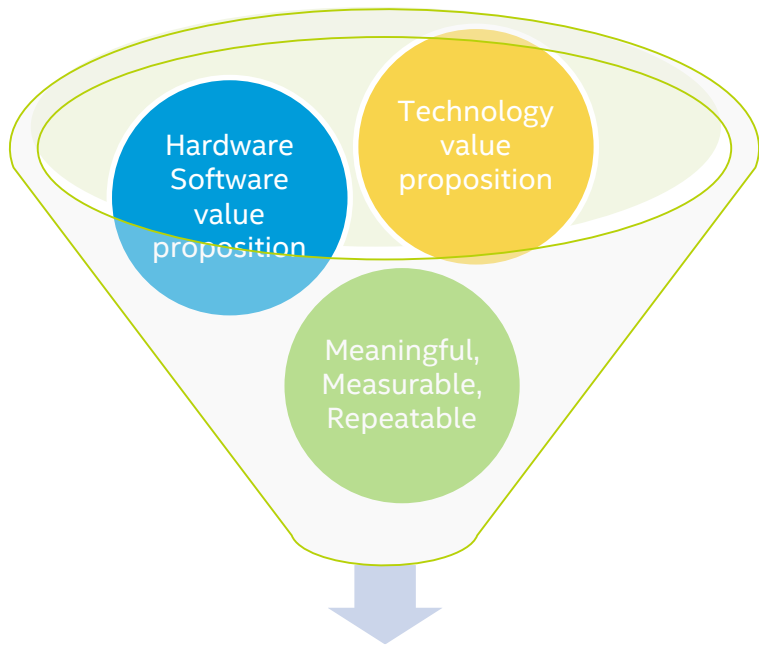
Prevalence of **Terasort** benchmark as yardstick in various talks means we desperately need an industry benchmark suite for Hadoop. [#strataconf](#)

[Expand](#)

Big Data Benchmark Pipeline



The Case for Standards and Specifications



Benchmark usage

- Performance Analysis
- Reference Architectures, Collaterals
- Influence Roadmaps and Features
- Aid Customer Deployments
- Illustrative $\neq$ Informative

Industry Standards

TPC

TPC

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&prelayout=flat>

² http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.aspx

³ http://www.odbs.org/wp-content/uploads/2014/12/bigbench-workload-tpctc_final.pdf

TPCx-BB v1.0 Implementation

Implementation

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

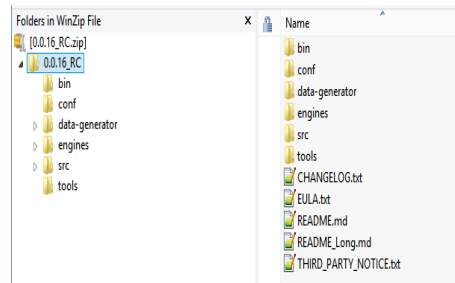
Kit Features

- Easy setup
- Size and Concurrency Scaling
- Versatile Driver
- Modular to support New APIs'

Availability

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

Kit Structure



Supported Hadoop Engines



Metrics and Reporting

$$BBQpm@SF = \frac{SF * 60 * M}{T_{LD} + \frac{2}{\sqrt{T_{PT}}} * T_{TT}}$$

$$\$ / BBQpm@SF = \frac{C}{BBQpm@SF}$$

Test	Stream Number	Query Number	Start Time	End Time	Elapsed Time in Seconds
POWER_TEST	0	1	12/07/2015 01:02:30 AM	12/07/2015 01:10:05 AM	469
POWER_TEST	0	2	12/07/2015 02:10:58 AM	12/07/2015 03:58:07 AM	6429
POWER_TEST	0	3	12/07/2015 03:58:07 AM	12/07/2015 04:40:06 AM	2369
POWER_TEST	0	4	12/07/2015 04:40:06 AM	12/07/2015 06:21:17 AM	6002
POWER_TEST	0	5	12/07/2015 06:21:17 AM	12/07/2015 07:51:34 AM	5037
POWER_TEST	0	6	12/07/2015 07:51:34 AM	12/07/2015 08:17:00 AM	9431
POWER_TEST	0	7	12/07/2015 08:17:00 AM	12/07/2015 08:40:55 AM	1430
POWER_TEST	0	8	12/07/2015 08:40:55 AM	12/07/2015 09:09:36 AM	1721
POWER_TEST	0	9	12/07/2015 09:09:36 AM	12/07/2015 09:29:31 AM	1197
POWER_TEST	0	10	12/07/2015 09:29:31 AM	12/07/2015 09:38:42 AM	349
POWER_TEST	0	11	12/07/2015 09:38:42 AM	12/07/2015 09:43:11 AM	275
POWER_TEST	0	12	12/07/2015 09:43:11 AM	12/07/2015 09:59:42 AM	967
POWER_TEST	0	13	12/07/2015 09:59:42 AM	12/07/2015 10:12:30 AM	768
POWER_TEST	0	14	12/07/2015 10:12:30 AM	12/07/2015 10:16:06 AM	203
POWER_TEST	0	15	12/07/2015 10:16:06 AM	12/07/2015 10:19:36 AM	210

Modular



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

