TPCx-AI Subcommittee

Transaction Processing Performance Council's

TPCx-AI Express Benchmark

Presented by:
Hamesh Patel, Chair of the TPCx-AI Committee

www.tpc.org

TPC Transaction Processing Performance Council

TPCx-AI Subcommittee

Agenda

- Benchmarking, standards & the TPC
- Challenges in ML\AI benchmarking
- TPC benchmark development goals & target usage
- TPCx-AI Key Features & Overview
- Summary

What Makes a Good Benchmark?

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

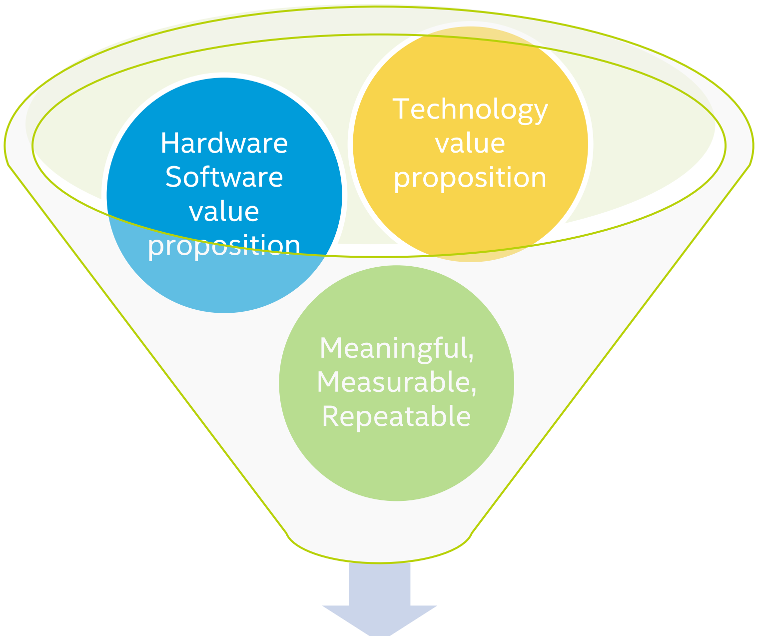
Usable
- ❑ Easy to use Kit
- ❑ Simplified Metric
- ❑ Support and Maintenance

Based on industry standards
- ❑ Peer Reviewed Specification and Code
- ❑ Public Availability
- ❑ Industry Acceptance

Flexible
- ❑ Adaptive
- ❑ Modularized
- ❑ Reusable

TPCx-AI Subcommittee

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

TPC

Transaction Performance Council(TPC)

- Mission

The Transaction Processing Performance Council (TPC) is a non-profit corporation whose mission is to develop data-centric benchmarks and to disseminate objective, verifiable TPC performance data to the industry.

- Background
  - The Transaction Processing Performance Council (TPC) was established in August 1988.
  - Distributes vendor-neutral performance data to the industry.
  - Vendors use TPC benchmarks to illustrate performance competitiveness for their existing products, and to improve and monitor the performance of their products under development.
  - Current Benchmark standards:
    - TPC-C, TPC-E, TPC-H, TPCx-BB, TPCx-IOT, TPC-DS TPCx-HCI, TPCx-HS, TPCx-V
  - Specifications consistent across all benchmark standards
    - TPC-Energy Specification: Augments existing TPC benchmarks with energy metrics
    - TPC-Pricing Specification: Single pricing specification ensures prices used in published results are verifiable

Challenges in ML\AI Benchmarking

Lack of diversity across several dimensions:

- Data types (textual, numerical, audio, image)
- Problem class (supervised, unsupervised)
- Method (classification, clustering, regression, Deep Learning vs ML)
- Complexity (simple, complex tasks)
- Scale

No representative benchmark emulating end-to-end datascience pipelines

Lack of AI production ready, commerically available solutions

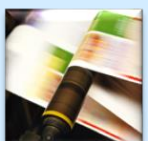
Lack of comparability across platforms & solutions

- No Unified Primary Metric, Price Performance metric and Energy metric

AI innovation continues at rapid pace

TPC Benchmark Development Goals

Build an industry standard that represents key AI use case(s) relevant in today's Datacenter and Cloud



Allow publication and audit of TPC results based on commercially available AI solutions.

Keep current with industry trends in AI and update the standard specification and\or implementations accordingly.

TPCx-AI benchmark standard usage

The standard addresses the needs of:

- Technology providers
  - Hardware innovators
  - Software solutions and services
- Companies adopting AI to:
  - Compare Performance, price performance, TCO
  - Improve existing deployments
- Academia and Research
- Vendors wanting to benchmark on large realistic datasets

TPCx-AI Subcommittee

Contributors

CISCO

IBM

DELL

Microsoft

Hewlett Packard Enterprise

intel

Red Hat

TTA

ORACLE

vmware

TPCx-AI: Key Features

- First Industry Standard benchmark emulating representative end-to-end Data science pipelines
- 7 Machine Learning & 3 Deep Learning use cases (Version 1.0.0) – Retail Datacenter
- Framework Agnostic
- 2 implementations in the kit (commercially available Scikit-Learn and Spark). Use either one
- Unified Primary Metric
- Ability to Scale to large datasets
- Priced configurations for TCO

Benchmark Standard Overview

python™

Keras

pandas

TensorFlow

Spark

HOROVOD

PySpark

[URL: TPCx-AI Homepage]

Implementation

- Self Contained Kits
- Scikit-Learn implementation
- Spark implementation
- Use either implementation to publish
- 10 Machine learning & Deep learning use cases

Kit Features

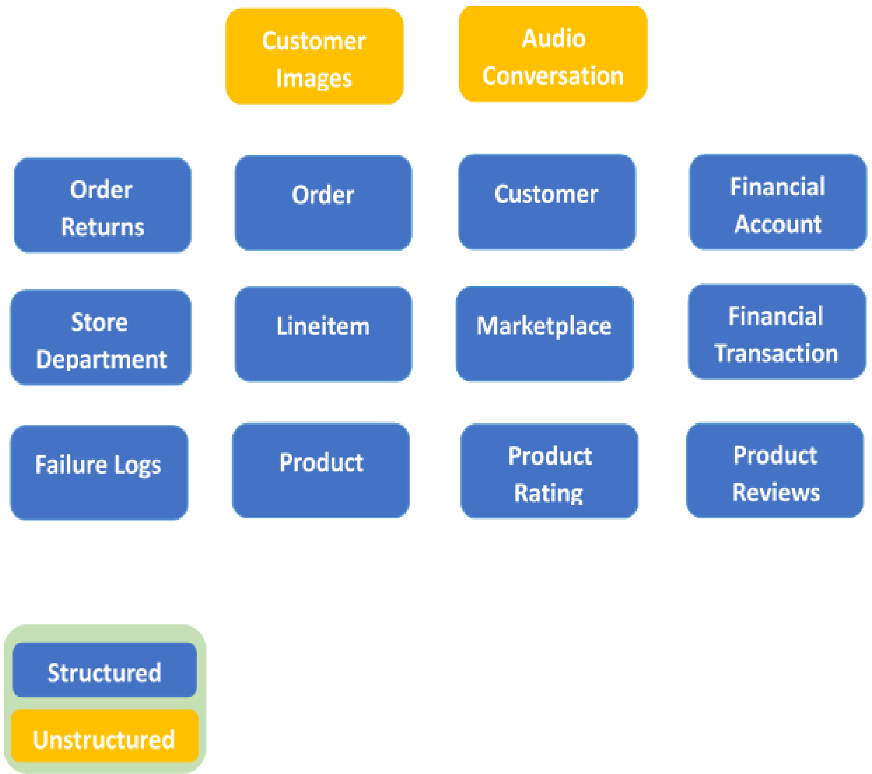
- Easy setup
- Dataset size and Concurrency Scaling
- Versatile
- Modular Driver to support Future APIs

Availability

- TPCx-AI Version 1.0.0
- Download from the TPC website (www.tpc.org)
- Open to contributions

TPCx-AI Subcommittee

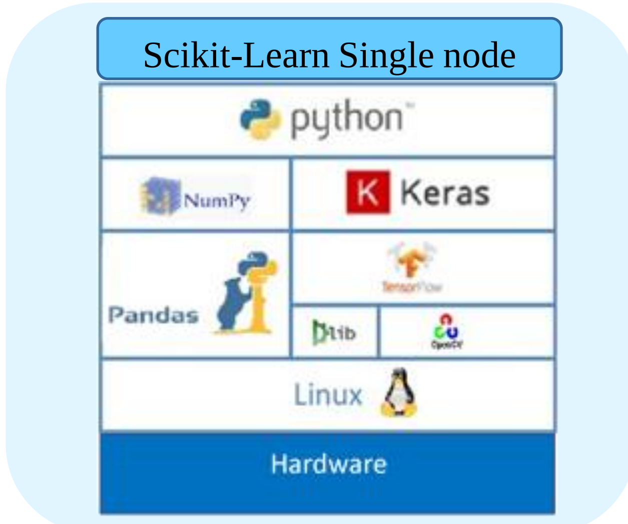
TPCx-AI: Data Model

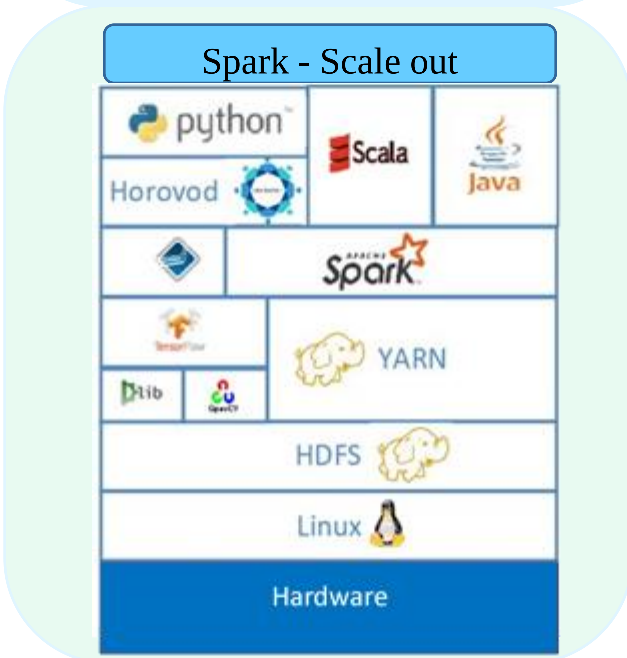


Unified Schema

- Inspired by real-world dataset schemas
- Data types
  - Text –Logs, Structured
  - Audio Conversations
  - Images
- Scaling and non-scaling Tables
- Parallel Data Generation

TPCx-AI: Kit Implementation





Scikit-Learn implementation

- For smaller datasets
- For smaller hardware configurations
- Uses Sci-kit Learn, Python, Pandas, Keras, Tensorflow, etc.

Spark implementation

- For very large datasets
- clustered configurations
- Uses Spark, pyspark, Horovod, Tensorflow, etc.

Use case # mapping

- Machine learning - 1, 3, 4, 6, 7, 8, 10
- Deep learning – 2, 5, 9

Flexible & Modular

- Ability to tune parameters
- Flexibility for optimizing data load, training & inference
- Can use other commercially supported SUT software implementations

TPCx-AI: Benchmark Workflow

Cluster Setup → Define Parameters → Run Benchmark → Analyze Data → Results

Cluster Setup

- ✓ Min 1 Node – Python
- ✓ Min 3 nodes - Spark
- ✓ Concurrent Streams 2-n

Define Parameters

- ✓ Scale Factor up to 10000GB
- ✓ Tuning Parameters for Training and Serving Tests

Run Benchmark

- ✓ Data Generation
- ✓ Load Test,
- ✓ Power Training Test,
- ✓ Power Serving Tests,
- ✓ Scoring Test,
- ✓ Throughput Test

Analyze Data

- ✓ Reports
- ✓ Metrics
- ✓ Utilization

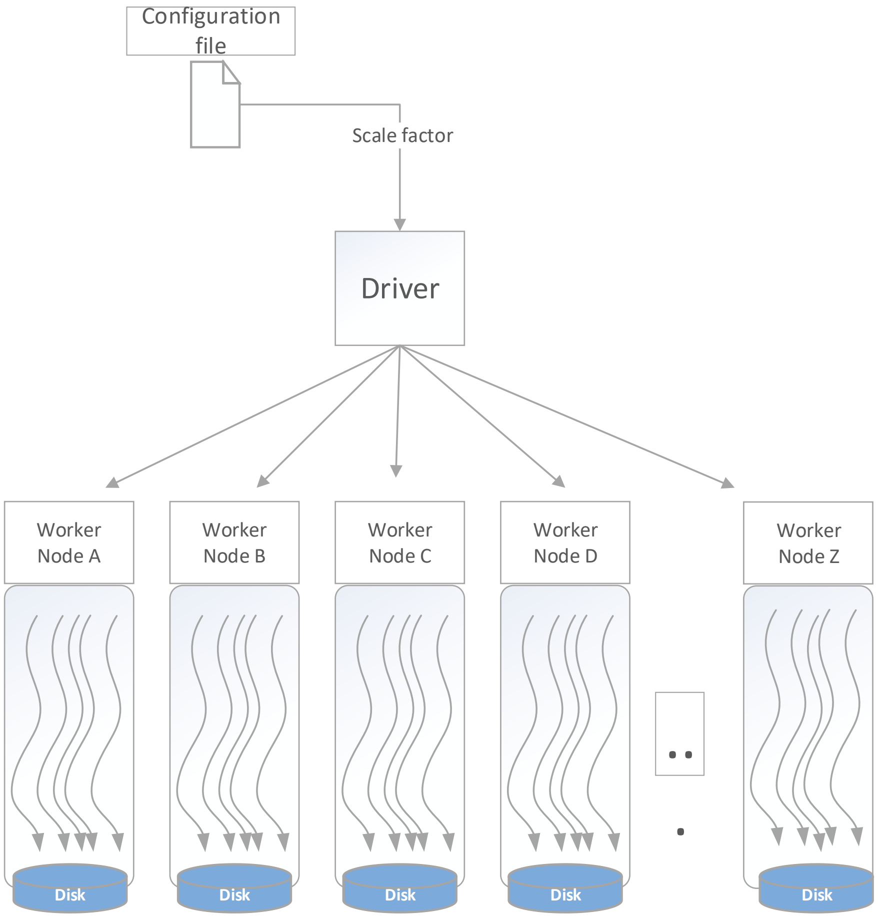
Results

- ✓ Publication
- ✓ Collateral
- ✓ POC

TPCx-AI Subcommittee

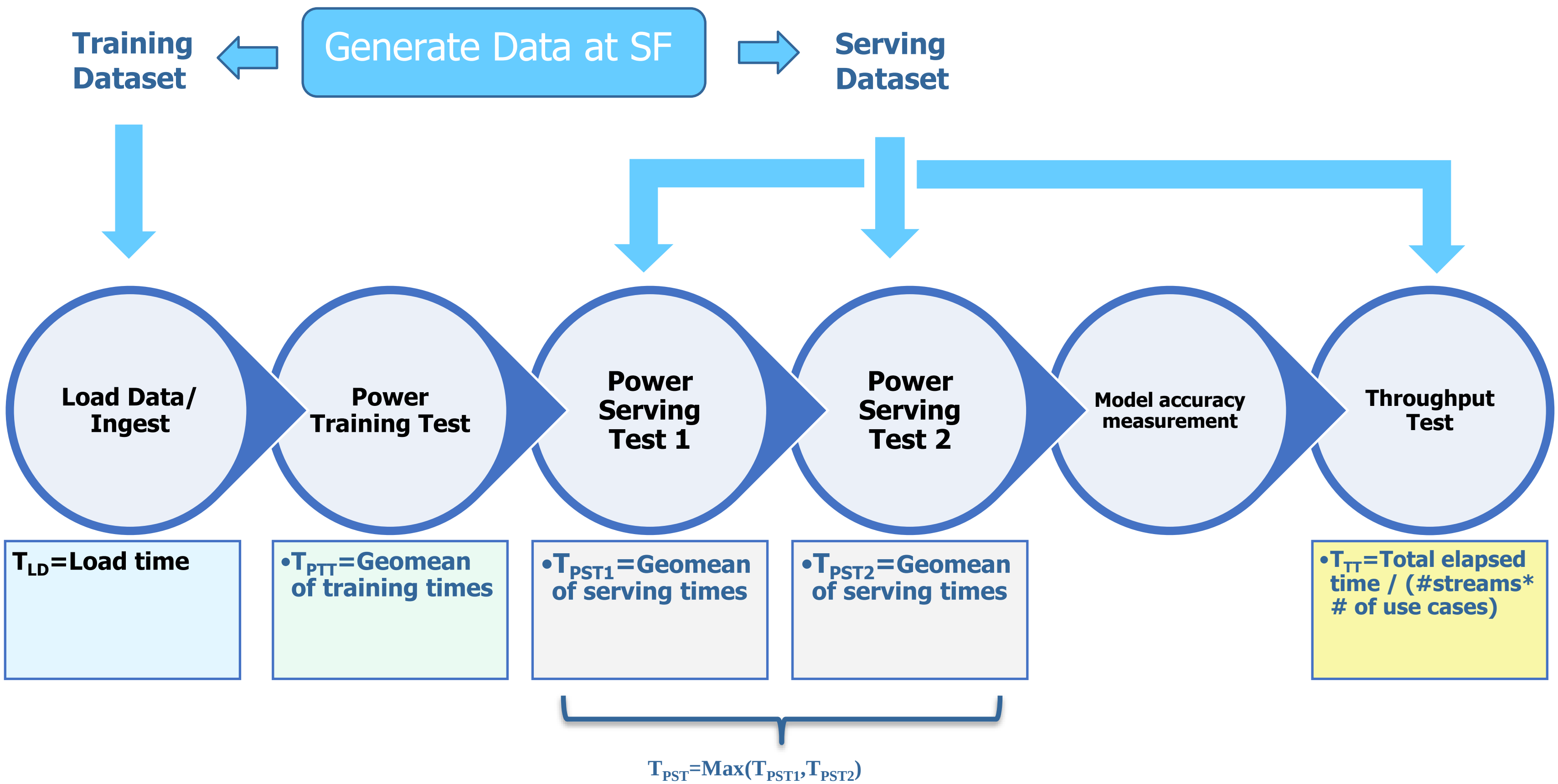
Benchmark Driver and Data Generator

- The benchmark includes a parallel synthetic data generator.
- The Driver reads multiple settings and user input parameters from configuration files or the command line.
- The scale factor (SF) is a configuration parameter specified by the user. It sets the target input dataset size in GB. E.g SF=100 equals 100GB.
- The driver spawns multiple data generation threads across all worker nodes in the cluster to quickly generate the amount of data specified by the SF.
- The synthetic data generator is capable of generating datasets of varied sizes (From Gigabytes to Terabytes) while maintaining the main characteristics of the dataset as a whole.



TPC Transaction Processing Performance Council

Benchmark Test Run

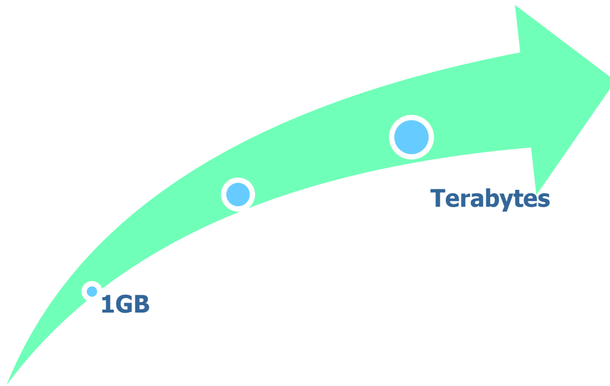


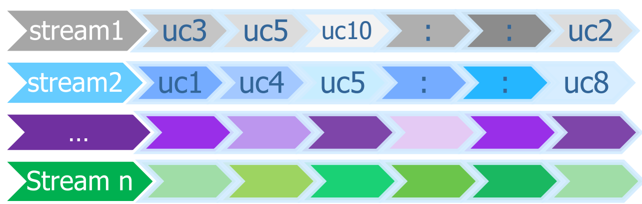
Performance Metric: $AIUCpm@SF = \dfrac{SF * N^{*} * 60}{\sqrt[4]{T_{PTT} * T_{PST} * T_{TT} * T_{LD}}}$

*N=Number of use cases

TPCx-AI Subcommittee

ScaleFactor, Streams & Metrics





AI Use cases per min @ SF

$ / AI Use Cases per min @ SF

Data & Scale Factors

- Data scales to Terabytes
- Diverse Dataset (Audio, Images, Text)
- Real world representative dataset

Streams

- Easy setup
- Dataset size and Concurrency Scaling
- Flexibility to show performance leadership
- Demonstrates TCO

Metrics

- Primary Metric
  - Includes time to load, manage data, train & serve
- Price Performance Metric
  - $/Primary Metric

TPCx-AI: Future developments

- AI innovation continues at a rapid pace
  - TPC will continue to keep pace
- Modularity in benchmarking is key
  - TPC plans to include new use cases in future versions
  - continue to modularize key stages
- Inclusion of additional stages
  - TPC will continue to develop additional stages for future versions

Summary -TPCx-AI: A Big Step…

- Representative key production ML\DL use cases
- Benchmarks end-to-end data science pipelines
  - essential in today's ML\AI environments
- Benchmark standard can scale a diverse and representative dataset
- Provides Price performance & overall AI solution TCO for published results

www.tpc.org