



Inspur Cloud Information Technology Co., Ltd.

TPC Express Benchmark™ AI Full Disclosure Report

InspurCloud Data-Cloud

with 20x Data-Cloud Server

using

InspurCloud Data Cloud Platform 5.2.0

running on

InLinux 23.12 (LTS-SP1)

TPCx-AI Version
Report Edition
Report Submitted

2.0.0
First
May 25, 2025

First Edition - May 2025

Inspur Cloud Information Technology Co., Ltd. (Inspur Cloud), the Sponsor of this benchmark test, believes that the information in this document is accurate as of the publication date. The information in this document is subject to change without notice. The Sponsor assumes no responsibility for any errors that may appear in this document.

The pricing information in this document is believed to accurately reflect the current prices as of the publication date. However, the Sponsor provides no warranty of the pricing information in this document.

Benchmark results are highly dependent upon workload, specific application requirements, and system design and implementation. Relative system performance will vary because of these and other factors. Therefore, TPC Express Benchmark™ AI should not be used as a substitute for a specific customer application benchmark when critical capacity planning and/or product evaluation decisions are contemplated.

All performance data contained in this report was obtained in a rigorously controlled environment. Results obtained in other operating environments may vary significantly. No warranty of system performance or price/performance is expressed or implied in this report.

Inspur Cloud and the Inspur Cloud Logo are trademarks of Inspur Cloud Information Technology Co., Ltd. and/or its affiliates in China and other countries. Third party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Inspur Cloud and any other company.

TPC Express Benchmark™ AI, TPCx-AI, and AIUCpm@3000, are registered certification marks of the Transaction Processing Performance Council.

The Inspur Cloud products, services or features identified in this document may not yet be available or may not be available in all areas and may be subject to change without notice. Consult your local Inspur Cloud business contact for information on the products or services available in your area. You can find additional information via Inspur Cloud's web site at <https://cloud.inspur.com>. Actual performance and environmental costs of Inspur Cloud products will vary depending on individual customer configurations and conditions.

Copyright© 2025 Inspur Cloud Information Technology Co., Ltd.

All rights reserved. Permission is hereby granted to reproduce this document in whole or in part provided the copyright notice printed above is set forth in full text or on the title page of each item reproduced.

Abstract

Inspur Cloud conducted the TPC Express Benchmark™ AI (TPCx-AI) on the InspurCloud Data-Cloud. The software used included InspurCloud Data Cloud Platform 5.2.0. This report provides full disclosure of the results. All testing was conducted in conformance with the requirements of the TPCx-AI Standard Specification, Revision 2.0.0.

Configuration Overview

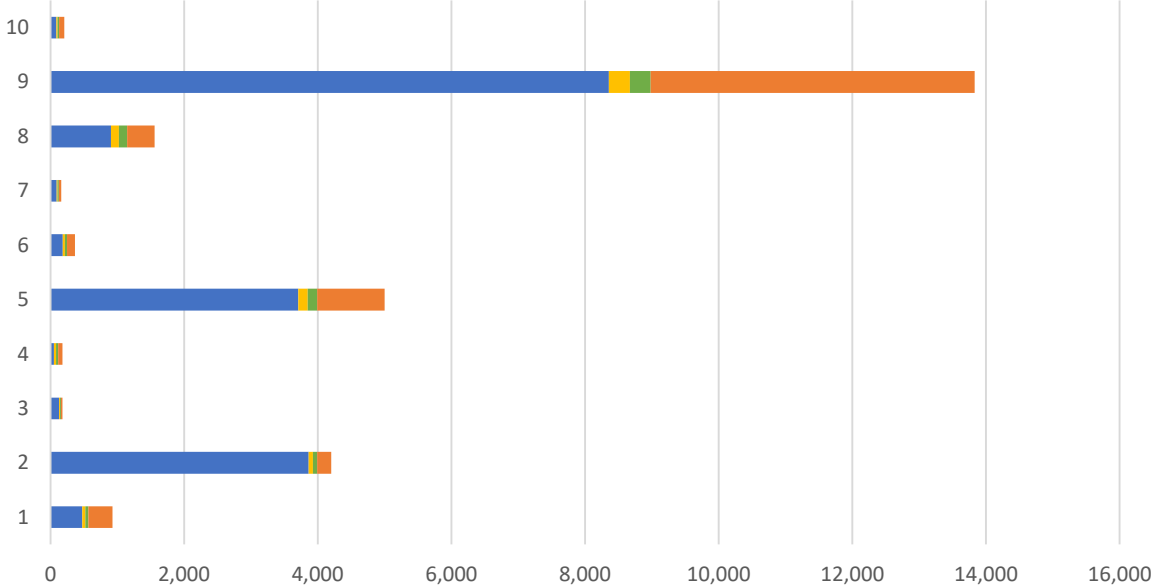
Test Sponsor	Node(s)	Operating System
Inspur Cloud	20x Data-Cloud Server	InLinux 23.12 (LTS-SP1)

Metrics Overview

Total System Cost	Performance	Price/Performance	Availability Date
\$610,450 USD	8,990.07 AIUCpm@3000	67.91 USD \$/AIUCpm@3000	May 25, 2025

Executive Summary

The [Executive Summary](#) follows on the next several pages.

		InspurCloud Data-Cloud		TPCx-AI 2.0.0	
				TPC Pricing 2.9.0	
				Report Date May. 25, 2025	
TPCx-AI Performance		Total System Cost	Price/Performance	Availability Date	
8,990.07 AIUCpm@3000		\$610,450 USD	\$67.91 USD/AIUCpm@3000	May 25, 2025	
Framework	Operating System	Other Software	Scale Factor	Streams	
InspurCloud Data Cloud Platform 5.2.0	InLinux 23.12 (LTS-SP1)	N/A	3,000	20	
Use Case Time (sec.) by Phase Training Serving 1 Serving 2 Throughput (Avg) 					
Physical Storage / Scale Factor		Scale Factor / Physical Memory		Main Data Redundancy Model	
183.47		0.29		Replication 3	
Servers:		20			
Total Processors/Cores/Threads		40 / 1,280 / 2,560			
Server Type	20x Data-Cloud Server				
Processors	2x AMD EPYC 9374F 32-Core Processor GHz				
Memory	512 GiB				
Storage Controller	1x Broadcom / LSI SAS3008				
Storage Device	2x 960 GB SATA SSD; 8x 3.2 TB NVMe SSD				
Network Controller	1x Intel Corporation Ethernet Controller X710 (12 nodes)				
	1x Intel Corporation I350 Gigabit Network Connection				
	2x Mellanox-MCX556A-ECAT 100 Gb 2-port				
Connectivity	1x Huawei CloudEngine 8850-64CQ-EI; 1x H3C S5560 Series				

		InspurCloud Data-Cloud			TPCx-AI 2.0.0		
					TPC Pricing 2.9.0		
					Report Date May. 25, 2025		
Description		Part Number	Source	List Price	Qty	Extended Price	1-Yr. Maintenance
Hardware							
InspurCloud Data-Cloud Server		P54199-B21	1	\$5,566.00	20	\$111,320.00	
AMD EPYC 9374F 3.85GHz 32-core 320W Processor		P54199-B21	1	\$2,514.00	40	\$100,560.00	
2U Passive CPU Heat Sink for AMD Socket SP5 Processors		SNK-P0083P	1	\$42.00	40	\$1,680.00	
Middle Cooling Fan for 2U Hyper-S Systems 80x80x38mm 13.5K RPM		FAN-0209L4-1	1	\$28.00	80	\$2,240.00	
32GB DDR5 RECC 4800B 2R*8 (M321R4GA3BB6-CQKMS)		M321R4GA3BB6	1	\$140.00	320	\$44,800.00	
SSD 960G SATA 6Gbps 2.5in (7mm) PM893 (MZ7L3960HCJR-00B7C)		MZ7L3960HCJR	1	\$140.00	40	\$5,600.00	
SSD 3.2T U.2 PCIe 2.5in D7-P5620 (SSDPF2KE032T1N1)		SSDPF2KE032T1N1	1	\$420.00	160	\$67,200.00	
1600W redundant single output power supply with inp		PWS-1K63A-1R	1	\$210.00	40	\$8,400.00	
Intel Corporation Ethernet Controller X710		Intel-X710	1	\$360.00	12	\$4,320.00	
Intel Corporation I350 Gigabit Network Connection		Intel-I350	1	\$84.00	20	\$1,680.00	
Mellanox-MCX556A-ECAT 100Gb 2-port Adapter		MCX556A-ECAT	1	\$698.00	40	\$27,920.00	
Maintenance - 7x24x4 Care Pack (1-yr)			1 (included)		20		\$0.00
				Subtotal		\$375,720.00	\$0.00
Software							
InspurCloud Data Cloud Platform 5.2.0 Subscription Edition			1	\$10,475.00	20	\$209,500.00	
InspurCloud 7x24*4 On-site Service (1-yr)			1 (included)		20		\$0.00
				Subtotal		\$209,500.00	\$0.00
Other Hardware							
HUAWEI CloudEngine 8850-64CQ-EI		CE8850	1	\$23,040.00	1	\$23,040.00	
H3C S5560 Series Switch		S5560	1	\$1,120.00	1	\$1,120.00	
Rack 48U Advanced Pallet			1	\$418.00	1	\$418.00	
Mellanox 100Gb 5m Direct Attach Copper Cable			1	\$11.00	20	\$220.00	
H3C S5560 10m Network Cable			1	\$6.00	20	\$120.00	
Keyboard and mouse			1	\$32.00	1	\$32.00	
Monitor			1	\$280.00	1	\$280.00	
				Subtotal		\$25,230.00	\$0.00
				Total		\$610,450.00	\$0.00
Pricing: 1 = Inspur Cloud			Total System Cost (USD):		\$610,450		
Audited by Doug Johnson, InfoSizing			AIUCpm@3000:		8,990.07		
			\$/AIUCpm@3000:		\$67.91		
Prices used in TPC benchmarks reflect the actual prices a customer would pay for a one-time purchase of the stated Line Items. Individually negotiated discounts are not permitted. Special prices based on assumptions about past or future purchases are not permitted. All discounts reflect standard pricing policies for the listed Line Items. For complete details, see the pricing section of the TPC Benchmark Standard. If you find that the stated prices are not available according to these terms, please inform the TPC at pricing@tpc.org. Thank you.							

		InspurCloud Data-Cloud		TPCx-AI	2.0.0
				TPC Pricing	2.9.0
				Report Date	May. 25, 2025
<i>Numerical Quantities</i>					
AIUCpm@3000	8,990.07	T_{Load}	1,907.61		
Scale Factor	3,000	T_{LD}	1,907.61		
Streams	20	T_{PTT}	467.97		
		T_{PST1}	48.82		
Kit Version	2.0.0	T_{PST2}	48.43		
Execution Status	PASS	T_{PST}	48.82		
Accuracy Status	PASS	T_{TT}	36.87		
Test Times					
Overall Run Start Time		2025-04-24 15:24:16.572			
Overall Run End Time		2025-04-25 00:43:48.745			
Overall Run Elapsed Time		33,572.173			
Load Test Start Time		2025-04-24 16:35:24.596			
Load Test End Time		2025-04-24 17:07:14.729			
Load Test Elapsed Time		1,910.133			
Power Training Start Time		2025-04-24 17:07:14.733			
Power Training End Time		2025-04-24 22:04:59.497			
Power Training Elapsed Time		17,864.764			
Power Serving 1 Start Time		2025-04-24 22:04:59.499			
Power Serving 1 End Time		2025-04-24 22:18:35.428			
Power Serving 1 Elapsed Time		815.929			
Power Serving 2 Start Time		2025-04-24 22:18:35.431			
Power Serving 2 End Time		2025-04-24 22:32:07.272			
Power Serving 2 Elapsed Time		811.841			
Scoring Start Time		2025-04-24 22:35:35.070			
Scoring End Time		2025-04-24 22:40:39.817			
Scoring Elapsed Time		304.747			
Throughput Start Time		2025-04-24 22:40:39.824			
Throughput End Time		2025-04-25 00:43:48.743			
Throughput Elapsed Time		7,388.919			

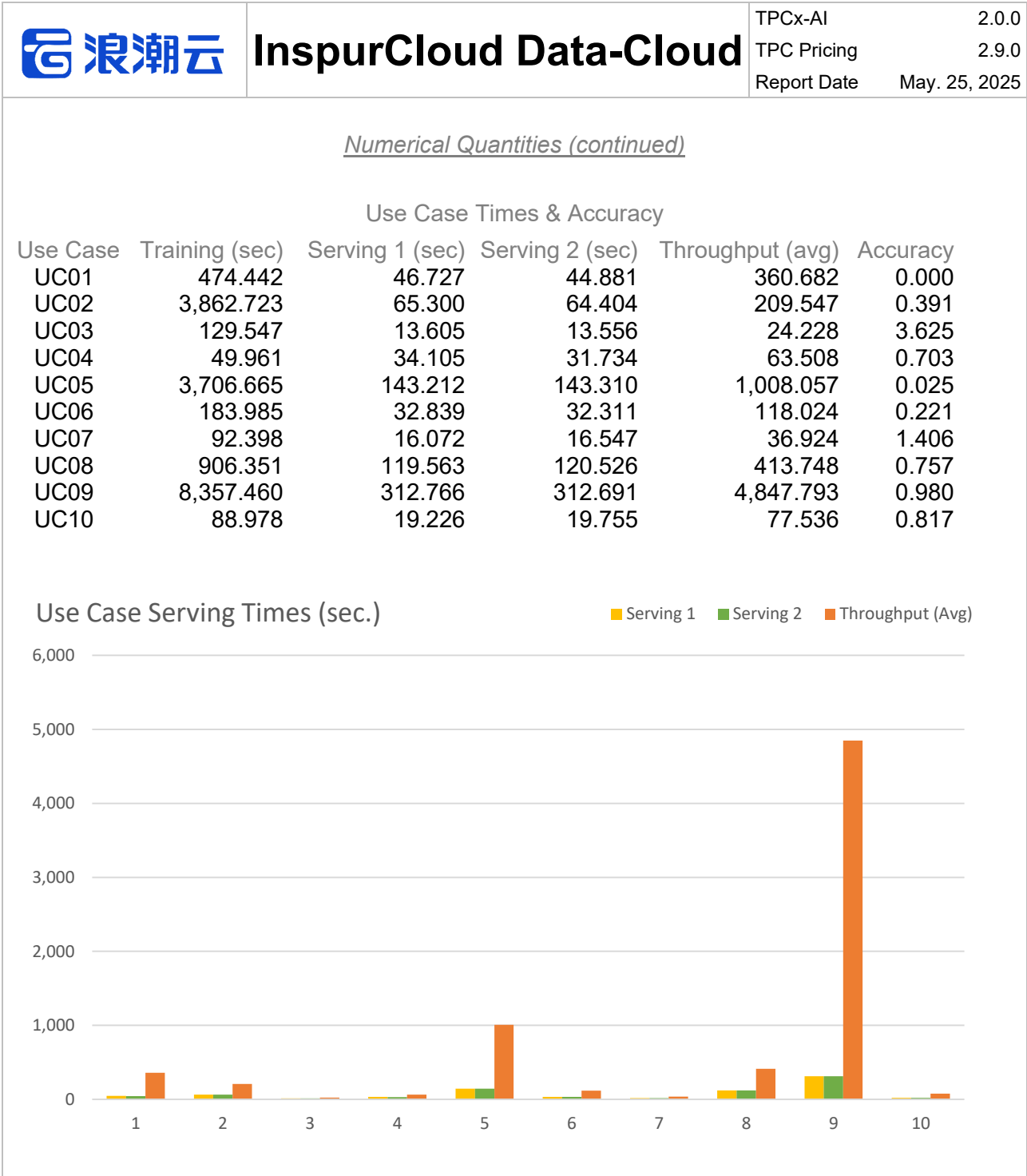


Table of Contents

Abstract.....	3
Executive Summary	3
Table of Contents.....	8
Clause 0 – Preamble	10
0.1 TPC Express Benchmark™ AI Overview	10
Clause 1 – General Items	12
1.1 Test Sponsor	12
1.2 Parameter Settings	12
1.3 Configuration Diagrams	12
1.3.1 Measured Configuration.....	13
1.3.2 Differences Between the Measured and the Priced Configurations.....	13
Clause 2 – SW Components & Data Distribution	14
2.1 Roles and Dataset Distribution	14
2.2 File System Implementation.....	14
2.3 Execution Engine, Frameworks, Driver & Libraries	14
2.4 Applied Patches.....	14
Clause 3 – Workload Related Items.....	15
3.1 Hardware & Software Tuning.....	15
3.2 Kit Version & Modifications	15
3.3 Use Case Elapsed Times	16
3.4 SUT Validation Test Output	18
3.5 Configuration Parameters	19
Clause 4 – SUT Related Items.....	20
4.1 Specialized Hardware/Software.....	20
4.2 Configuration Files.....	20
4.3 SUT Environment Information.....	20
4.4 Data Storage to Scale Factor Ratio	20
4.5 Scale Factor to Memory Ratio	20
4.6 Output of Tests	20
4.7 Additional Sponsor Files	20
4.8 Model Optimizations	20
Clause 5 – Metrics and Scale Factor	21
5.1 Reported Performance Metrics	21

5.2 Throughput Test Stream Times22

Auditor’s Information23

Third-Party Price Quotes.....26

Supporting Files Index27

Clause 0 – Preamble

0.1 TPC Express Benchmark™ AI Overview

Artificial intelligence (AI) has become a key transformational technology of our times. Advances in neural networks and other machine learning techniques have made it possible to use AI on a variety of use cases. From the public sector to aerospace, defense and academia, new and improved ways to use AI techniques are changing the way we harness data and analytics. This along with advances in compute, interconnect and memory technologies have made possible to solve complicated challenges that will ultimately benefit customers in production datacenter and cloud environments.

Abundant volumes of rich data from text, images, audio and video are the essential starting point for creating a benchmark that would represent the myriad of use cases and customers. TPC Express Benchmark™ AI (TPCx-AI) is created in keeping with the TPC tradition of emulating real world AI scenarios and data science use cases. Unlike most other AI benchmarks, the TPCx-AI uses a diverse dataset and is able to scale across a wide range of scale factors. TPCx-AI may later expand with additional use cases and add additional flexibility for a greater variety of implementations.

The benchmark defines and provides a means to evaluate the System Under Test (SUT) performance as a general-purpose data science system that:

- Generates and processes large volumes of data.
- Trains preprocessed data to produce realistic machine learning models.
- Conducts accurate insights for real-world customer scenarios based on the generated models.
- Can scale to large scale distributed configurations.
- Allows for flexibility in configuration changes to meet the demands of the dynamic AI landscape.

The benchmark models real-life examples of companies and public-sector organizations that use a range of analytics techniques, both AI and more traditional machine learning approaches, as well as the potential application of these techniques in situations like those in which they have already been successfully deployed. In addition, the benchmark measures end to end time to provide insights for individual use cases, as well as throughput metrics to simulate multiuser environments for a given hardware, operating system, and data processing system configuration under a controlled, complex, multi-user AI or machine learning data science workload.

The purpose of TPC benchmarks is to provide relevant, objective performance data to industry users. To achieve that purpose, TPC benchmark specifications require benchmark runs be implemented with systems, products, technologies and pricing that:

- Are generally available to users.
- Are relevant to the market segment that the individual TPC benchmark models or represents (e.g., TPCx-AI models and represents complex, high data volume, decision support environments).
- Would plausibly be implemented.

The TPCx-AI kit is available from the TPC website (see www.tpc.org/tpcx-ai/ for more information). Users must sign up and agree to the TPCx-AI End User Licensing Agreement (EULA) to download the kit. All related work (such as collaterals, papers, derivatives) must acknowledge the TPC and include the TPCx-AI copyright. The TPCx-AI kit includes: TPCx-AI Specification document (this document), TPCx-AI Users Guide (README.md) documentation, scripts to set up the benchmark environment, code to execute the benchmark workload, Data Generator, use case related files, and Benchmark Driver.

The use of new systems, products, technologies (hardware or software) and pricing is encouraged so long as they meet the requirements above. Specifically prohibited are benchmark systems, products, technologies or pricing (hereafter referred to as "implementations") whose primary purpose is performance optimization of TPC benchmark results without any corresponding applicability to real-world applications and environments. In other words, all "benchmark special" implementations that improve benchmark results but not real-world performance or pricing, are prohibited.

The rules for pricing are included in the TPC Pricing Specification.

Further information is available at www.tpc.org.

Clause 1 – General Items

1.1 Test Sponsor

This benchmark was sponsored by Inspur Cloud Information Technology Co., Ltd.

1.2 Parameter Settings

The [Supporting Files Archive](#) contains the parameters and options used to configure the components involved in this benchmark.

1.3 Configuration Diagrams

The measured configuration diagram is shown below. In addition, any differences between the measured and the priced configurations are described.

Clause 2 – SW Components & Data Distribution

2.1 Roles and Dataset Distribution

Table 2-1 describes the distribution of the dataset across all media in the SUT.

Server	Host Name	SW Services	Storage	Contents
20x Data-Cloud	indata-10-108-8-[76,78]	NameNode, ResourceManager	2x 960 GB SATA SSD 8x 3.2 TB NVMe SSD	OS, Kit, Models Data
	indata-10-108-8-79	Spark Master		
	indata-10-108-8-[70-91]	DataNode, NodeManager, Spark Worker		

Table 2-1 Software Components and Dataset Distribution

2.2 File System Implementation

A distributed file system provided by InLinux 23.12 (LTS-SP1) / InspurCloud Data Cloud Platform 5.2.0 was used for data generation and the Load Test. The data set was not relocated after generation and before the Load Test.

2.3 Execution Engine, Frameworks, Driver & Libraries

InspurCloud Data Cloud Platform 5.2.0 consisted of the following components.

Component	Version
HDFS	3.4.1
YARN	3.4.1
MapReduce2	3.4.1
Spark	3.1.2

Table 2-2 Software Components

For a detailed listing of installed libraries, please see the envInfo logs in the [Supporting Files](#).

2.4 Applied Patches

No additional vendor-supported patches were applied to the SUT.

Clause 3 – Workload Related Items

3.1 Hardware & Software Tuning

The [Supporting Files](#) archive contains all hardware and software configuration scripts.

3.2 Kit Version & Modifications

Table 3-1 shows the version of the TPCx-AI used to produce this result along with any kit files that were modified to facilitate system, platform, and framework differences.

TPCx-AI Kit Version	2.0.0
<u>Modified File</u>	<u>Description of Changes</u>
tools/parallel-data-load.sh	Group commands into a single background job. Ensure that the "mkdir" command is completed before proceeding. Reduce the number of SSH connections per file.
tools/tpcxai_fdr.py	Modify the placeholder "%.5f" used for formatting floating-point numbers in the SQL statement. This is to resolve the following error. Traceback (most recent call last): File "tools/tpcxai_fdr.py", line 604, in <module> main() File "tools/tpcxai_fdr.py", line 589, in main report = make_report(connection, benchmark_id, include_clean) File "tools/tpcxai_fdr.py", line 200, in make_report uc_result = cursor.execute(phase_use_case_query, (benchmark_id, phase_name, rec['stream'], rec['phase_run'])) sqlite3.OperationalError: no such column: "%.5f" - should this be a string literal in single - quotes?

Table 3-1 Kit Version & Modifications

3.3 Use Case Elapsed Times

Below are the elapsed times for each use case. Use cases are grouped based on whether they use Deep Learning or Machine Learning techniques.

Type	UC ID	P1	P2	T1	T2		
Deep Learning	2	P1	P2	T1	T2	T3	T4
	5	65.300	64.404	144.447	183.965	224.648	196.509
	9	143.212	143.310	1,153.648	1,156.088	758.115	856.194
Machine Learning	1	312.766	312.691	4,858.890	4,847.980	4,860.874	4,816.340
	3	46.727	44.881	360.592	385.049	334.067	400.595
	4	13.605	13.556	15.335	25.721	25.071	29.518
	6	34.105	31.734	63.399	91.277	70.304	54.325
	7	32.839	32.311	133.910	129.027	122.731	122.289
	8	16.072	16.547	36.016	40.484	40.169	38.530
	10	119.563	120.526	395.532	365.667	444.082	421.546

Type	UC ID	T5	T6	T7	T8	T9	T10
Deep Learning	2	203.607	155.341	200.738	233.009	231.402	221.543
	5	861.934	1,135.176	987.983	789.777	1,182.592	1,229.991
	9	4,872.133	4,840.423	4,870.884	4,825.803	4,846.539	4,843.694
Machine Learning	1	343.782	372.732	384.268	354.129	390.876	363.903
	3	33.566	29.730	17.604	15.228	15.008	29.711
	4	62.694	78.457	43.649	73.897	39.960	75.901
	6	131.784	78.341	92.972	130.776	127.566	105.850
	7	42.431	24.158	23.514	45.803	21.040	15.052
	8	403.321	431.575	448.079	424.517	402.680	378.376
	10	81.275	80.783	75.541	74.507	63.488	82.397

Type	UC ID	T11	T12	T13	T14	T15	T16
Deep Learning	2	226.414	251.147	277.168	221.185	161.043	130.521
	5	1,204.753	787.237	865.398	1,226.910	1,114.764	1,222.490
	9	4,802.508	4,845.691	4,856.276	4,835.220	4,832.379	4,826.768
Machine Learning	1	381.542	333.054	299.203	354.870	352.033	384.252
	3	23.169	26.048	28.525	28.922	30.236	14.488
	4	67.862	55.315	66.513	77.698	71.986	43.362
	6	128.401	132.277	107.860	134.989	83.383	111.603
	7	31.579	30.398	38.421	56.088	40.528	48.400
	8	416.756	406.682	396.869	364.180	464.589	434.024
	10	74.035	75.314	73.443	73.542	75.637	74.330

Type	UC ID	T17	T18	T19	T20
Deep Learning	2	236.474	237.465	211.612	242.710
	5	855.790	830.089	1,165.100	777.109
	9	4,852.219	4,881.249	4,859.683	4,880.303
Machine Learning	1	382.005	335.984	369.344	331.355
	3	31.339	23.096	15.357	26.889
	4	48.544	54.474	64.199	66.335
	6	99.001	132.803	134.895	120.030
	7	43.744	40.628	42.986	38.513
	8	445.158	426.078	375.972	429.273
	10	87.125	81.952	75.729	76.231

Table 3-2 Use Case Elapsed Times

3.4 SUT Validation Test Output

<u>Validation Run Report</u>			
AIUCpm@1	27.26	T _{Load}	85.75
Scale Factor	1	T _{LD}	85.75
Streams	20	T _{PTT}	80.91
		T _{PST1}	23.40
Kit Version	2.0.0	T _{PST2}	23.14
Execution Status	PASS	T _{PST}	23.40
Accuracy Status	PASS	T _{TT}	1.45

Test Times	
Overall Run Start Time	2025-04-24 14:19:58.260
Overall Run End Time	2025-04-24 15:23:44.986
Overall Run Elapsed Time	3,826.726
Load Test Start Time	2025-04-24 14:24:36.025
Load Test End Time	2025-04-24 14:26:04.211
Load Test Elapsed Time	88.186
Power Training Start Time	2025-04-24 14:26:04.214
Power Training End Time	2025-04-24 15:01:32.518
Power Training Elapsed Time	2,128.304
Power Serving 1 Start Time	2025-04-24 15:01:32.519
Power Serving 1 End Time	2025-04-24 15:06:01.816
Power Serving 1 Elapsed Time	269.297
Power Serving 2 Start Time	2025-04-24 15:06:01.819
Power Serving 2 End Time	2025-04-24 15:10:30.205
Power Serving 2 Elapsed Time	268.386
Scoring Start Time	2025-04-24 15:13:51.531
Scoring End Time	2025-04-24 15:18:41.452
Scoring Elapsed Time	289.921
Throughput Start Time	2025-04-24 15:18:41.459
Throughput End Time	2025-04-24 15:23:44.985
Throughput Elapsed Time	303.526

(continued on next page)

Validation Run Report (continued)

Accuracy Metrics					
Use Case	Metric Name	Metric	Criteria	Threshold	Status
1	N/A	0.000	N/A	0.00	Pass
2	word_error_rate	0.425	<=	0.50	Pass
3	mean_squared_log_error	6.633	<=	5.40	Fail*
4	f1_score	0.698	>=	0.65	Pass
5	mean_squared_log_error	0.085	<=	0.50	Pass
6	matthews_corrcoef	0.223	>=	0.19	Pass
7	median_absolute_error	1.696	<=	1.80	Pass
8	accuracy_score	0.699	>=	0.65	Pass
9	accuracy_score	1.000	>=	0.90	Pass
10	accuracy_score	0.817	>=	0.70	Pass

*Because of the small dataset size used for the Validation Test, Spark-based implementations may not be able to satisfy the accuracy threshold for Use Case 3. The TPCx-AI Subcommittee is aware of this issue and has decided that this failure does not invalidate the test.

3.5 Configuration Parameters

The [Supporting Files](#) archive contains all Global Benchmark Parameter and Use Case Specific Parameter settings.

Clause 4 – SUT Related Items

4.1 Specialized Hardware/Software

No Specialized Hardware/Software was used in the SUT.

4.2 Configuration Files

The [Supporting Files](#) archive contains all configuration files.

4.3 SUT Environment Information

All envInfo.log files are included in the [Supporting Files](#) archive.

4.4 Data Storage to Scale Factor Ratio

The details of the Data Storage Ratio are provided below.

Node Count	Disks	Size (GB)	Total (GB)
20	2	960	38,400
20	8	3,200	512,000
Total Storage (GB)			550,400
Scale Factor			3,000
Data Storage Ratio			183.47

4.5 Scale Factor to Memory Ratio

The details of the Memory to Scale Factor Ratio are provided below.

Nodes	Memory (GiB)	Total (GiB)
20	512	10,240
Scale Factor		3,000
Total Memory (GiB)		10,240
SF / Memory Ratio		0.29

4.6 Output of Tests

The [Supporting Files](#) archive contains the output files of all tests.

4.7 Additional Sponsor Files

The [Supporting Files](#) archive contains any additional files that were used.

4.8 Model Optimizations

The [Supporting Files](#) archive contains any model optimization files that were used.

Clause 5 – Metrics and Scale Factor

5.1 Reported Performance Metrics

Metric Overview

TPCx-AI Performance Metric	8,990.07	AIUCpm@3000
TPCx-AI Price/Performance Metric	67.91	\$/AIUCpm@3000
TPCx-AI Scale Factor	3,000	
TPCx-AI Stream Count	20	

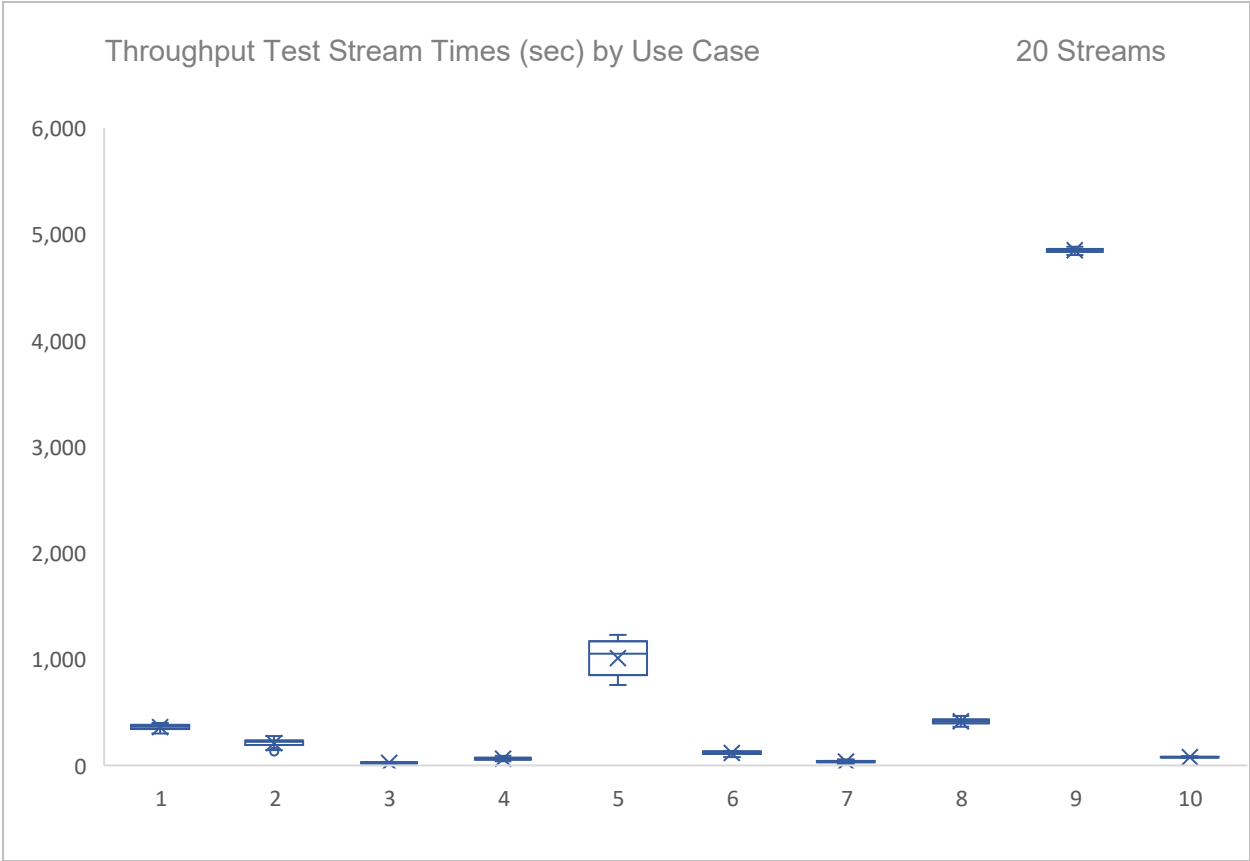
Test Times

Overall Run Start Time	2025-04-24 15:24:16.572
Overall Run End Time	2025-04-25 00:43:48.745
Overall Run Elapsed Time	33,572.173
Load Test Start Time	2025-04-24 16:35:24.596
Load Test End Time	2025-04-24 17:07:14.729
Load Test Elapsed Time	1,910.133
Power Training Start Time	2025-04-24 17:07:14.733
Power Training End Time	2025-04-24 22:04:59.497
Power Training Elapsed Time	17,864.764
Power Serving 1 Start Time	2025-04-24 22:04:59.499
Power Serving 1 End Time	2025-04-24 22:18:35.428
Power Serving 1 Elapsed Time	815.929
Power Serving 2 Start Time	2025-04-24 22:18:35.431
Power Serving 2 End Time	2025-04-24 22:32:07.272
Power Serving 2 Elapsed Time	811.841
Scoring Start Time	2025-04-24 22:35:35.070
Scoring End Time	2025-04-24 22:40:39.817
Scoring Elapsed Time	304.747
Throughput Start Time	2025-04-24 22:40:39.824
Throughput End Time	2025-04-25 00:43:48.743
Throughput Elapsed Time	7,388.919

<u>Accuracy Metrics</u>					
Use Case	Metric Name	Metric	Criteria	Threshold	Status
1	N/A	0.000	N/A	0.00	Pass
2	word_error_rate	0.391	<=	0.50	Pass
3	mean_squared_log_error	3.625	<=	5.40	Pass
4	f1_score	0.703	>=	0.65	Pass
5	mean_squared_log_error	0.025	<=	0.50	Pass
6	matthews_corrcoef	0.221	>=	0.19	Pass
7	median_absolute_error	1.406	<=	1.80	Pass
8	accuracy_score	0.757	>=	0.65	Pass
9	accuracy_score	0.980	>=	0.90	Pass
10	accuracy_score	0.817	>=	0.70	Pass

5.2 Throughput Test Stream Times

The following chart shows the minimum, 1st quartile, median, mean (X), 3rd quartile, and maximum stream times by use case for the Throughput Test. Outliers are marked with “o”.



Auditor's Information

This benchmark was audited by Doug Johnson, InfoSizing.

www.sizing.com
63 Lourdes Drive
Leominster, MA 01453
978-343-6562.

This benchmark's Full Disclosure Report can be downloaded from www.tpc.org.

A copy of the auditor's attestation letter is included in the next two pages.



Zheng Wei
 Inspur Cloud Information Technology Co., Ltd.
 No.1036 Inspur Road
 Jinan City
 China

May 24, 2025

I verified the TPC Express Benchmark™ AI v2.0.0 performance of the following configuration:

Platform: 20x InspurCloud Data-Cloud Servers
 Operating System: InLinux 23.12 (LTS-SP1)
 Additional Software: InspurCloud Data Cloud Platform 5.2.0

The results were:

Performance Metric 8,990.07 AIUCpm@3000

Secondary Metrics	T _{LD}	1,907.61
	T _{PTT}	467.97
	T _{PST}	48.82
	T _{TT}	36.87

System Under Test 20x InspurCloud Data-Cloud Servers with:

CPU	2x AMD EPYC 9374F 32-Core Processor		
Memory	512 GiB		
Storage	Qty	Size	Type
	2	960 GB	SATA SSD
	8	3.2 TB	NVMe SSD

In my opinion, these performance results were produced in compliance with the TPC requirements for the benchmark.

The following verification items were given special attention:

- All TPC-provided components were verified to be v2.0.0.
- All checksums were validated for compliance.
- Any modifications to shell scripts were reviewed for compliance.
- No modifications were made to any of the Java code.
- The generated dataset was properly scaled to 3,000 GB.

63 Lourdes Dr. | Leominster, MA 01453 | 978-343-6562 | www.sizing.com

- The generated dataset used for testing was protected by Replication 3.
- The elapsed times for all phases and runs were correctly measured and reported.
- The Storage and Memory Ratios were correctly calculated and reported.
- The system pricing was verified for major components and maintenance.
- The major pages from the FDR were verified for accuracy.

Additional Audit Notes:

Because of the small dataset size used for the Validation Test, this Spark-based implementation was not able to satisfy the accuracy threshold for Use Case 3. The TPCx-AI Subcommittee is aware of this issue and has decided that this failure does not invalidate the test.

Respectfully Yours,

A handwritten signature in cursive script, reading "Doug Johnson", followed by a horizontal flourish line.

Doug Johnson, Certified TPC Auditor

63 Lourdes Dr. | Leominster, MA 01453 | 978-343-6562 | www.sizing.com

Third-Party Price Quotes

All components are available directly through the Test Sponsor (Inspur Cloud).

Supporting Files Index

The Supporting Files archive for this disclosure contains the following structure.

Supporting Files Directory	Description
CheckIntegrity/...	Output of CHECK_INTEGRITY test (if the phase is not done as part of the Validation and Performance Test).
PerformanceTest/...	Performance Test output files.
ValidationTest/...	Validation Test output files.

Additional files used by Inspur Cloud

Sponsor/ModelOptimization/...	Details of model optimization.
Sponsor/ModifiedKitFiles/...	2 modified file(s).
Sponsor/Tuning/...	All tuning files used.