Cloud benchmarks aren't enough

The use of modeling to address the limitations of benchmark tests

Boris Zibitsker, PhD, BEZNext CEO
Alex Lupersolsky, PhD, BEZNext CTO

Abstract

This white paper explores the value and constraints of industry-standard and customized benchmark tests in making informed cloud decisions. We present the BEZNext approach to overcoming these limitations by leveraging modeling and optimization technology. Our method adds value to results of the benchmark tests by evaluating options, optimizing cloud performance, reducing cost, and refining decisions related to platform selection, cloud migration, dynamic capacity management, DevOps choices, and carbon footprint estimation within a Hybrid Multi-Cloud environment.

Introduction

Standard benchmarks like TPC-DS and TPC-H [1], along with custom benchmarks, help companies assess resource utilization and scalability of cloud data platforms. However, such benchmarks have limitations.

Benchmarks use specific query types that are different from the customers' production workloads. Benchmark queries access benchmark data that are different from a live production environment [2].

While benchmarks can simulate varying concurrent user scenarios [3], challenges arise due to disparities in instance types, hardware capacity. Consequently, benchmark outcomes may not necessarily identify the optimal platform type and size necessary to meet business performance needs in a real-world production environment. Relying solely on benchmark results for cloud decisions risks unexpected financial and performance challenges [4].

In this paper we will review the value and limitations of benchmarks and present BEZNext cloud performance optimization and cost control technology based on measurement data collected at the customer environment. We incorporate results of the benchmark tests done on different cloud platforms.

The BEZNext approach includes organizing the observability of the customer cloud environment, modeling and optimization to compare options, and assisting with cloud data platform selection, the migration of workloads to the cloud, dynamic capacity management of the Hybrid Multi-Cloud environment, DevOps, carbon footprint estimation, and, finally, continuous performance and cost control.

Factors Affecting the Performance of Cloud Data Platforms

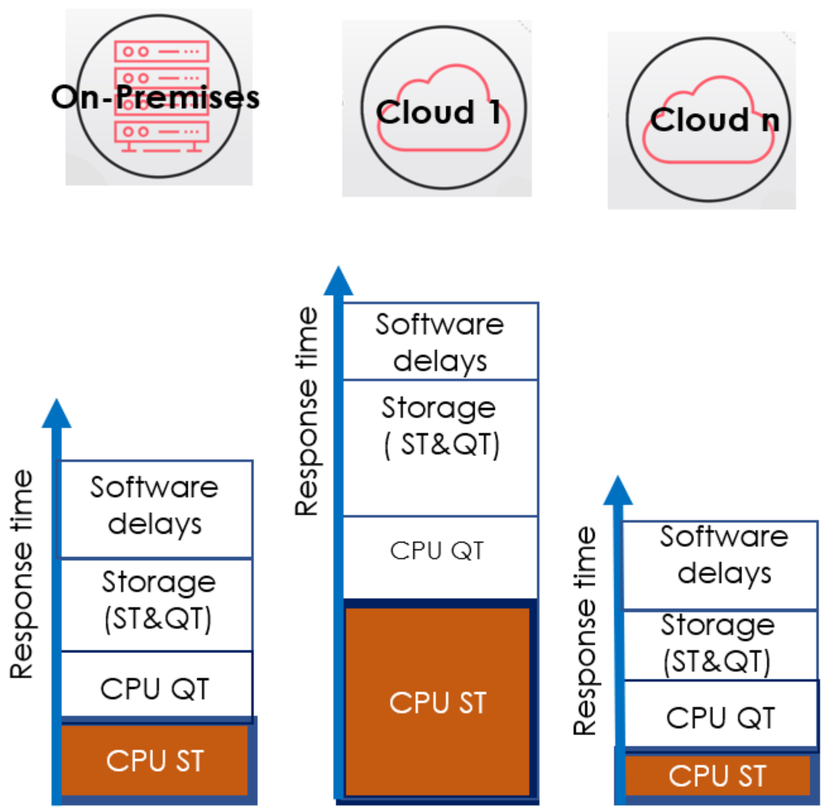

Figure 1: *Factors affecting CPU and storage service time (ST), queueing time (QT) and response time (RT) per query on different cloud platforms*

Major factors affecting the performance of the same queries in different cloud platforms include the hardware architecture, configuration (instance types, instance count, storage configuration), DBMS optimization techniques, workload management and resource allocation rules.

For instance, on-premises DBMS optimizers exhibit more proficient utilization of indexes and materialized views compared to their cloud counterparts.

Migration to the cloud frequently necessitates modifications to query structures and database design to align with platform specifications.

Benefits of Benchmark Tests in a Cloud Environment

TPC-DS benchmarks show the difference in resource consumption for the same queries on different cloud platforms

To assess CPU service times and megabytes processed per query, BEZNext executed the TPC-DS benchmark across diverse platforms, including Teradata on-premises and cloud data platforms, including Vantage Enterprise, VantageCloud Lake, Databricks, Snowflake, and BigQuery.

We divided 99 TPC-DS queries into short, medium, and long categories. Table 1 illustrates the ratios of CPU time and physical IO MB per query on cloud platforms compared to on-premises.

Results were obtained by sequentially running benchmark queries and collecting CPU service time and MB processed per query data measured on each platform. Notably, CPU service times were not normalized by CPU speed.

For instance, the value 2.087 in the first cell of the table signifies that executing short queries on Vantage EE consumes over two times the CPU resources compared to Teradata. In the following column, a comparison between queries on VantageCloud Lake Edition and Teradata on-premises highlights a twofold increase in CPU resource usage for short queries, while less time is required for longer queries. The ratios of physical IO megabytes per query are detailed in columns 6 to 10 of Table 1.

In terms of volume data processed by query, Vantage Enterprise exhibited 11 times more operations than Teradata on-premises for short queries. Conversely, VantageCloud Lake Edition showed only 0.49 times the operations, while Snowflake 2XL recorded a substantial 400-fold increase in MB per query for short queries compared to Teradata on-premises.

CPU Time ms				INTERVAL	Physical IO MB				
Vantage to TD	VCLE to TD	Databricks to TD	BigQuery to TD	by TD CPU Time	Vantage to TD	VCLE to TD	Snowflake 2XL to TD	Databricks to TD	BigQuery to TD
2.087	2.025	0.151	5.431	<130,881	11.405	0.488	406.688	1.394	6.934
1.652	1.580	0.100	1.412	130,881-311,093	14.708	0.586	357.033	1.311	6.567
0.758	0.756	0.019	1.022	>311,093	5.504	0.577	111.957	0.331	1.182

***Table 1**: Comparison of the resource utilization by TPC-DS benchmarks on Databricks, Snowflake, Vantage cloud data platforms and Teradata on-premises*

Benchmark tests used for assessing tokenization and detokenization overhead

Enterprises implement diverse information security measures to safeguard data during its transition from secure on-premises environments to public clouds. A common approach involves tokenizing data during upload to the cloud and detokenizing data upon retrieval. Tokenization and detokenization processes consume CPU resources, consequently impacting costs. Employing benchmark tests, both with and without tokenization, furnishes essential insights to enhance the precision of cost and performance predictions.

Leveraging benchmark tests to assess the effects of SQL and database tuning

Comparing CPU service time and the number of MB per query before and after tuning offers insights into resource consumption changes. However, it falls short of facilitating a cost-centric comparison between various tuning options.

Scalability analysis for cloud data platforms during customized tests

In addition to the standard TPC benchmarks, BEZNext conducted a customized proof-of-concept test to assess scalability. This test involved progressively increasing the number of concurrent users to evaluate response time (RT) and throughput across Redshift, Snowflake, and Vantage data warehouses. Concurrent users were doubled at intervals of 10, 20, 40, and 80, with results visualized in Figure 2.

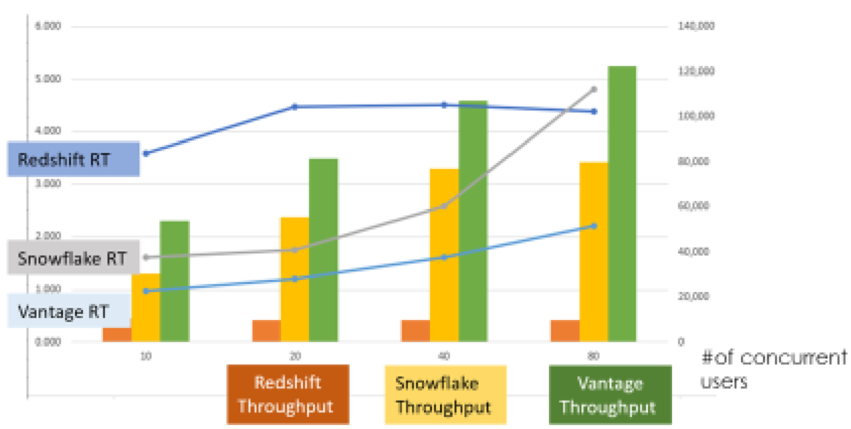

Figure 2: *Response time and throughput of Teradata Vantage, Snowflake and Redshift measured during the customized tests significantly differ depending on the number of concurrent users.*

During the customized test, throughput on Vantage is consistently higher than on Snowflake and Redshift. The average Vantage RT is lower than on Snowflake and Redshift with the increase in number of concurrent users.

Values of the Industry Standard Benchmark Tests (e.g., TPC-Ds and TPC-H) for Cloud Data Platform Evaluation:

1. Standardization: These benchmark tests provide a standardized framework for evaluating the performance of various cloud data platforms. The benchmark goal is to provide fair and consistent comparisons across different solutions.
2. Performance comparison: By running the same benchmark on multiple cloud data platforms, users can compare their performance.
3. Tokenization: Benchmark tests can be used to estimate the overhead of tokenization and detokenization in the cloud.
4. Vendor neutrality: Industry-standard benchmarks are typically designed to be vendor-neutral. This helps prevent bias and provides an objective basis for evaluating different cloud data platforms.
5. Trend analysis: Over time, as new versions of cloud data platforms are released, benchmark results can reveal trends in performance improvements or declines. This helps users track the evolution of platforms and make upgrade decisions.

While benchmark results offer insights for comparing cloud platforms scalability, they may not always accurately represent real-world workload, potentially leading to misleading conclusions.

What Can’t Benchmarks Show You?

Diversity of real-world workloads

To underscore the limitations of benchmark tests, we undertook an analysis encompassing the performance variability and resource utilization of business workloads across 48 companies. These organizations employed a mix of Teradata, DB2, and Oracle on-premises solutions, along with cloud data platforms such as BigQuery, Vantage, Azure, Redshift, Snowflake, and Oracle.

	Platform	Aero-space	Energy	Entertain-ment	Insurance	Manufac-turing	Oil and Gas	Pharma	Real Estate	Telecom
Cloud	Big Query	1	2			1				1
	Vantage		2			2		1		
	Azure		1	1	1			1	1	
	Redshift					2	2			1
	Snowflake	1	1			1	1	1		
	Oracle Cloud									
On Prem Analytics	Teradata	1	2			1				1
	DB2		2			2		1		
	Oracle		1	1	1			1	1	
On Prem Mixed	Teradata				2					1

Table 2: *The number of customers of cloud data platforms and on-premises Data warehouses in different industries selected for analysis*

Table 2 details the count of data platform installations investigated across diverse industries. Metrics examined are illustrated in Table 3, while Table 4 showcases the spectrum of configurations and performance data observed within customer environments. The results from characterizing real-world workloads among the 48 customers indicate substantial diversity in configurations, performance levels, and resource utilization. This emphasizes the challenge of applying benchmark tests within production environments, which can potentially lead to misleading conclusions.

Platform	# vCPU		DW Size (TB)		Concurrency		CPU Time per query (s)		Execution Time (s)		Industries
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	
Azure	160	480	20	300	2	68	70	80,741	4	801	Energy, Entertainment, Real Estate, Insurance, Pharma
Big Query	160	480	16	450	2	22	90	115,546	7	1,003	Telecom, Aerospace, Manufacturing, 2 Energy,
Snowflake	64	256	16	41	2	30	19	58,353	3	786	Energy, Manufacturing, Oil and Gas, Aerospace, Pharma
Oracle cloud	48	384	20	125	5	20	40	27,740	7	301	Oil and Gas, Finance, Real Estate, Pharma, Manufacturing
Vantage	32	128	10	24	2	16	21	15,375	6	572	Pharma, 2 Manufacturing, 2 Energy
Redshift	360	2304	20	512	2	16	207	353,802	4	912	2 Manufacturing, Telecom, 2 Oil and Gas
Teradata on prem	96	864	8	120	2	9	23	52,203	3	318	Real Estate, Insurance, 3 Finance
DB2 on prem	32	128	6	40	2	8	9	5,932	2	331	2 Insurance, 2 Financial, Real Estate
Oracle on prem	128	576	80	200	8	21	113	51,610	8	448	3 Finance, Energy, Oil and Gas

Table 3: *The range of the configuration and variance of the performance measurement data observed on each cloud and on-premises platforms*

Query Type	% Queries	# Queries	Category (Execution Time, sec)	Avg. Data Processed (GB)	Avg. Execution Time (sec)	Avg. Concurrency	Avg. Tables per Query	CPU Time used (sec)
Select	8	124	<= 20	23	16	4	2	246
Select	19	295	>20 <= 120	65	98	18	4	2,634
Select	18	279	> 120 <= 300	167	167	26	4	5,130
Select	4	62	> 300	412	786	28	4	28,170
Insert	8	124		171	342	18	3	9,193
Update	6	93		114	267	19	2	5,468
Copy	5	78		94	142	16	3	2,363
Merge	4	62		112	291	18	3	4,097
Sum	10	155		72	121	21	2	1,549
Min/Max	8	124		66	47	18	2	602
Show	4	62		32	25	19	1	256
Other	6	93		78	139	17	2	1,957
X-Large			**Average**	**108.9**	**168.7**	**19.0**	**3.0**	**4,247**
16 nodes	100	1551	**Total (GB, hours)**	**168,857**	**72.7**			
Energy Company. Fortune 500. Across the major tables in the datastore, there are approximately 242,000,000 rows.								
Warehouse size		28	TB					
Descriptive -> 20%, Predictive -> 40%, Prescriptive -> 40%								

Table 4: *An example of the metrics for a specific customer on Snowflake platform.*

Limitations of Industry Standard Benchmark Tests:

1. Benchmarks simplify real-world scenarios: Benchmark queries are designed to mimic real-world data processing scenarios but they do not work well, because business workloads are very different.
2. One-size doesn't fit all: Industry benchmarks offer a single set of queries and workloads that might not accurately represent all use cases.
3. Focus on performance: Benchmark tests primarily focus on performance metrics, such as query response times and throughput. They might not consider other important factors like data governance, and ease of management.
4. Limited coverage: Benchmarks may not cover all aspects of cloud data platform functionality. For instance, they might not fully address features like data transformation, integration, real-time processing, or support for certain data formats.
5. Vendor-specific enhancements: Some vendors might optimize their platforms to perform exceptionally well on specific benchmark queries, leading to skewed results that may not translate to better overall performance in real-world scenarios.

BEZNext Modeling and Optimization Make the Difference

Actual business workloads represent lines of business and that differ from the standard benchmark. BEZNext incorporates observability—including data collection, workload aggregation and characterization on each platform—and builds customer's workload performance, resource utilization, data usage and cost profiles. Results of the observability feed modeling and optimization engines to optimize cloud performance and control cost.

BEZNext uses modeling and optimization to evaluate options and optimize decisions during the journey to the cloud.
We implemented a three-part process that begins with the preparation step, including workload aggregation and characterization by line of business and building hourly performance, resource utilization, and data usage on-premises profiles for short, medium and large groups of queries of each workload.

The benchmark tests are run on each cloud platform. CPU service times and MB per query are collected and segregated into small, medium and large categories based on results. Change ratios are calculated to allow for a comparison of each cloud platform in each category.

Finally, the analytic model that was created and calibrated using on-premises data is modified to reflect the differences in configuration, CPU service time, MB per query, table sizes, and user concurrency of each cloud data platform. The gradient optimization and queueing network models predict the minimum configuration and cost required to meet business SLGs of each business workload on each cloud data platform after migration and in the future.

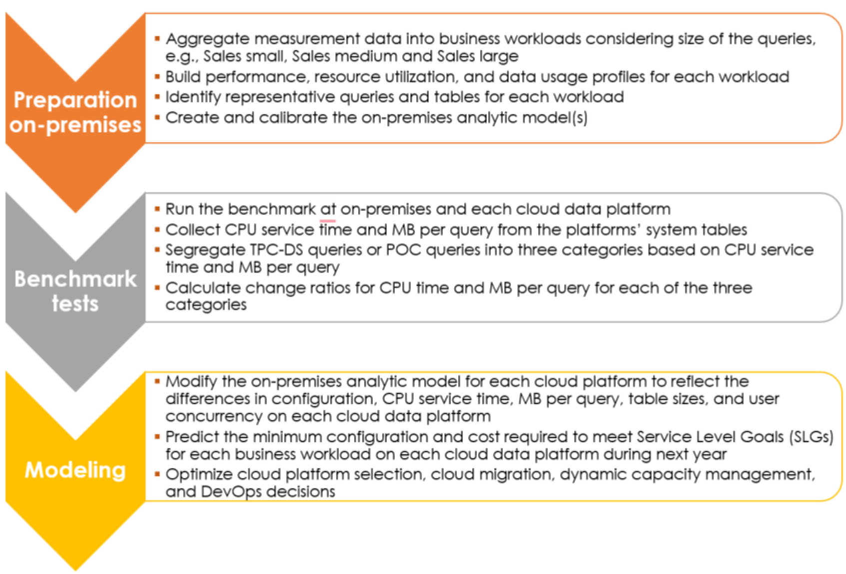

***Figure 3**: BEZNext modeling and optimization process complementing the results of the benchmark tests*

Predicted minimum configuration and pricing models are used to estimate the monthly and yearly budget needed to support SLGs for all business workloads on all platforms.

Predicted utilization of instances, storage is used to estimate power consumption and carbon footprint for each workload on each cloud data platform.

Platform	Instance Type	Shift	# Instances (Clusters) / Month											
			JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Teradata Vantage	m4.16xlarge	1st	10	10	10	10	10	10	10	10	10	10	11	11
	m4.16xlarge	2nd	32	34	34	34	34	34	36	36	36	36	38	38
	m4.16xlarge	3rd	13	13	13	13	13	13	13	13	14	14	14	14
Amazon Redshift	ra3.16Xlarge	1st	52	52	52	54	54	54	56	56	58	58	58	60
	ra3.16Xlarge	2nd	130	130	130	140	140	140	140	150	150	150	150	150
	ra3.16Xlarge	3rd	72	74	74	76	76	78	78	80	80	82	82	82
Snowflake	2XL	1st	5	5	6	6	6	6	6	6	6	6	6	6
	4XL	2nd	3	3	3	3	3	3	3	3	3	3	3	3
	3XL	3rd	5	5	5	5	5	5	5	5	5	5	5	5

Table 5: *Predicting the minimum configuration needed to meet SLG for each business workload on Vantage, Redshift and Snowflake cloud data platforms. The same approach can be used for BigQuery, Databricks, VantageCloud Lake and other platforms.*

The most frequent and severe performance and cost anomalies and their root causes are detected. The most critical users, SQL, databases, and tables accessed are identified to determine tuning recommendations.

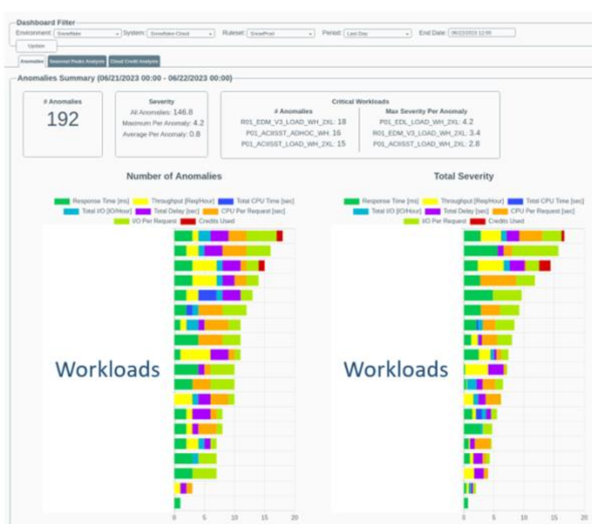

Figure 4*: Performance and cost anomalies and root cause determination as a part of the dynamic capacity management.*

Our unique approach combines advanced modeling and optimization technology with benchmark data to evaluate options, predict the minimum configuration and budget needed to meet business SLGs, provide actionable insights for better decision-making, and set realistic expectations prior to implementing recommended changes.

Moreover, our modeling capabilities extend to evaluating cloud migration scenarios, organizing dynamic capacity management, DevOps decision optimization, and estimation of each business workload's carbon footprint on different cloud data platforms.

For example, we can use modeling to determine the minimum configuration and budget required for a faster data load after migrating to the cloud. This means we can predict the minimum configuration and other changes needed to complete the data load process in just 8 hours instead of the previous 12 hours on-premises.

In the realm of DevOps, our models offer valuable recommendations to both developers and operations teams. This includes suggesting the right platform, size, and budget to meet specific performance targets before deploying new applications.

These examples showcase how our modeling approach goes beyond the limitations of traditional benchmark tests. It empowers businesses to make critical cloud decisions with greater confidence and accuracy.

Furthermore, modeling and optimization enable the organization of continuous closed-loop feedback control based on comparing the performance and cost measurement results with the expected (predicted) and development corrective actions when needed.

These examples illustrate how modeling expands the limitations of using benchmark tests to make critical business decisions related to cloud platform selection, migration, and management.

BEZNext modeling and optimization solve the challenges that can't be addressed by benchmark tests alone:

1. Modeling and optimization technology uses results of the cloud observability of customer workloads done automatically in real-time at on-premises and each cloud data platform.
2. ML technology identifies trends, anomalies, and detects root-causes.

3. Modeling and optimization technology determines the minimum configuration and budget needed to support SLGs for all business workloads in the Hybrid Multi-Cloud environment.
4. Modeling addresses many of the performance and financial challenges during the journey to the cloud and operating the Hybrid Multi-Cloud environment.

- Modeling helps select the appropriate cloud platform configuration needed to meet SLGs for each business workload with the lowest cost.
- Modeling optimizes cloud migration decisions
- Modeling optimizes dynamic capacity management decisions in the Hybrid Multi-Cloud environment.
- Modeling optimizes DevOps development and operational decisions prior to new application deployment.

Call to Action

Industry-standard benchmark tests offer valuable insights into the performance of cloud data platforms. However, benchmarks have limitations. BEZNext developed modeling and optimization technology to evaluate options and optimize decisions during the journey to the cloud.

BEZNext offers:

One-month **performance and cost control assessment service**. The price depends on the type and number of cloud data platforms.

- We use BEZNext software to organize observability on-premises and on cloud data platforms used by your organization.
- We determine performance and cost anomalies and their root causes to present performance tuning recommendations.
- We use modeling and optimization to evaluate options, including
  - Appropriate cloud data platform selection.
  - Cloud migration decision optimization.
  - Organizing continuous dynamic capacity management.
  - DevOps decision optimization before deployment of new applications.
  - Carbon footprint estimation.

- **Four weeks hands-on workshop**.
  - If your organization is starting the journey to the cloud or actively seeking solutions to enhance cloud performance and reduce the cost of running applications within a Hybrid Multi-Cloud environment, utilizing cloud platforms such as Snowflake, Teradata Vantage, Redshift, BigQuery, Databricks, or other similar platforms, then this workshop is for you!

- **Continuous performance and cost optimization service**. The price depends on the type and number of on-prem and cloud data platforms.
    - We organize a continuous automated observability, modeling and optimization, and a cost control process in your Hybrid Multi-Cloud environment.

References and Links

1. TPC. 2023. *TPC Benchmarks Overview*. https://www.tpc.org/information/benchmarks5.asp
2. Kounev, S., Lange, K., & von Kistowski, J. (2020). *Systems Benchmarking: For Scientists and Engineers*. Springer.
3. Gillam, L., Li, B., O'Loughlin, J. *et al.* Fair Benchmarking for Cloud Computing Systems. *J Cloud Comp* **2**, 6 (2013). https://doi.org/10.1186/2192-113X-2-6
4. Hwang, K., Bai, X., Shi, Y., Li, M., Chen, W. -G., & Wu, Y. (2016). Cloud Performance Modeling with Benchmark Evaluation of Elastic Scaling Strategies. *IEEE Transactions on Parallel and Distributed Systems*, *27*(1), 130-143. https://doi.org/10.1109/TPDS.2015.2398438

If you have questions, please email us at inquiry@beznext.com or leave a voicemail at 847-926-7237.