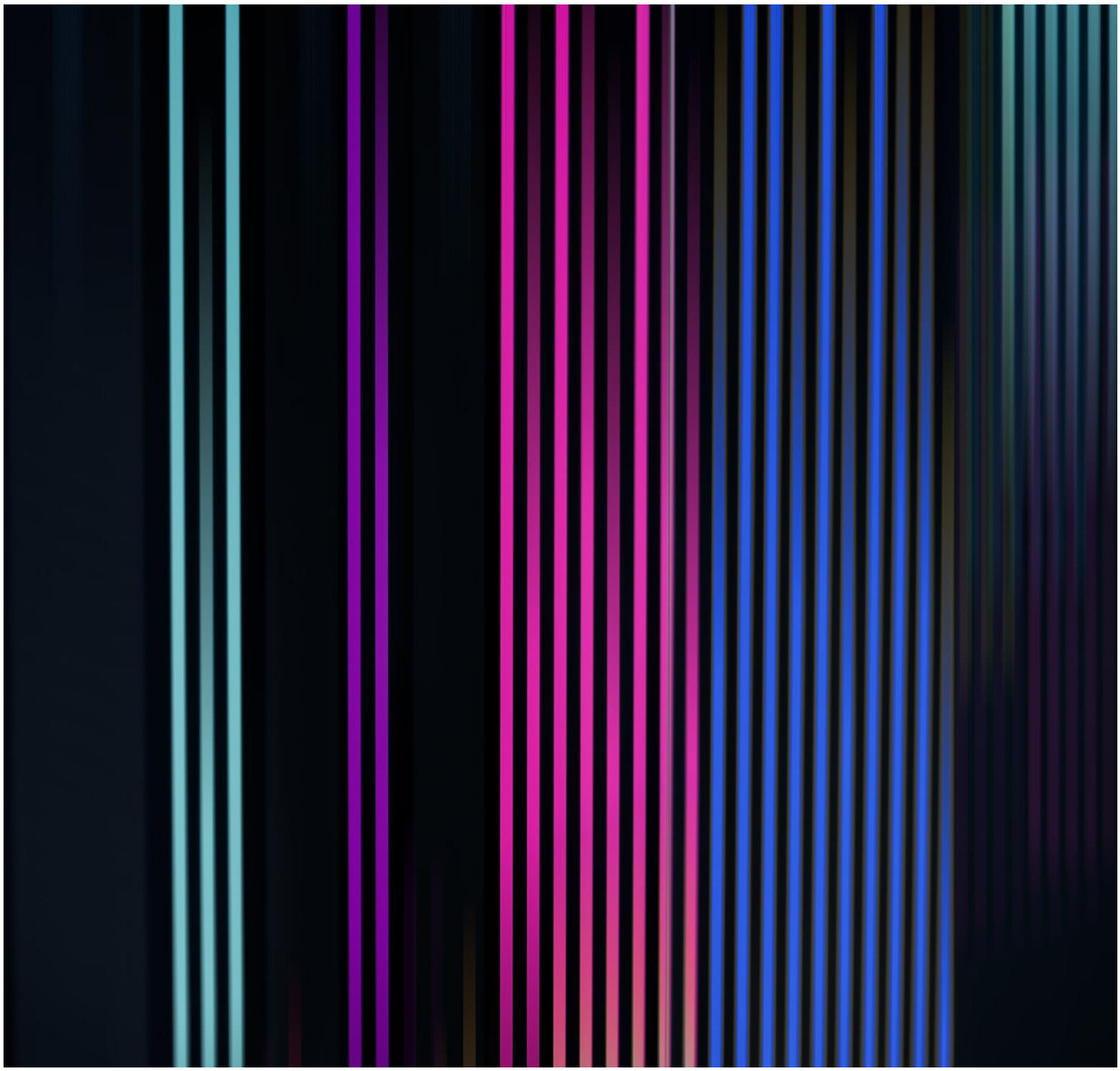

Which Platform Is Best for Your Cloud Data Warehouse?

BORIS ZIBITSKER

BEZNext | bzibitsker@beznext.com | www.beznext.com | 2/11/2021

Summary

In this white paper we review a case study illustrating how BEZNext modeling and optimization technology is used to determine the minimum configuration and cost required to meet Service Level Goals for Data Warehouse production workloads on Snowflake, Teradata Vantage, and Amazon Redshift. This approach can be used to evaluate Google BigQuery, Azure Synapse Analytics, Oracle Autonomous Data Warehouse, and other platforms on Amazon Web Services, Microsoft Azure, Google Cloud, and other clouds as well.

Table of Contents

Introduction

Customers are attracted to cloud platforms for a variety of reasons, including faster development and deployment at lower costs. This decision can be made for a new analytic application, an application currently running on-premises, or an application currently running on another cloud platform. This paper outines a framework for improved selection of a cloud data analytics platform.

These decisions are challenging because there are so many options are available, and one may achieve similar levels of performance on any cloud but with very different costs. So, organizations are interested in estimating the costs to achieve an acceptable level of performance (known as Service Level Goals, or SLGs) for their workload across different clouds.

BEZNext's approach to this challenge begins with an analysis of performance, resources, and data usage measurements to characterize the customer's data-intensive production workload. Our modeling and optimization technology is designed to predict the minimum configuration and budget required to meet the performance goals for each business workload on a variety of cloud-based database services. These predictions factor in the variability of resource demands during the day, expected increases in the number of users and data volume, as well as the impact of additional applications which will be deployed. For new applications, we collect and analyze data during the DevOps process.

We developed this approach to help IT organizations currently running analytic applications on-premises, using data warehouse technology from Teradata, Oracle, IBM, and others, who are considering moving workloads to the cloud.

We illustrate our approach with an actual enterprise customer's case study showing how our modeling and optimization engines are used to find the smallest cloud platform configuration required to meet the SLGs for all workloads on each cloud platform. We then select the platform with the lowest cost.

Background

Whether the decision to rely on cloud computing resources is strategic or tactical, the trend toward greater IT spending with cloud providers is evident. Migrating data-rich, analytic applications from on-premises to the cloud is a complex effort involving both business and technological factors that resist easy characterization. Many organizations find the inherent flexibility of cloud computing very attractive because it allows them to pay only for resources that they actually use. This flexibility is appealing in the face of uncertainty with the varying rates of demand for resources growth.

Public cloud platforms may also be appropriate for applications associated with geographically distributed data acquisition, such as Internet of Things (IoT). Globally distributed computing resources on a potentially massive scale, combined with uncertainty about the rate of data growth and the timing of that growth, can present a compelling business case for partnering with a major cloud vendor willing and able to guarantee delivery of those resources as required.

Growth scenarios of all types bring further complications. Many data-intensive, cloud-based applications start small, but success leads to more users and larger data stores. These applications are eventually absorbed into the IT organization's strategic planning processes, which abruptly finds itself responsible for gaining control of the application's spiraling costs.

Early in the decision-making process, IT organizations frequently rely on industry standard Transaction Processing Council (TPC) benchmarks or conduct pilot or Proof-of-Concept (POC) projects with a sampling of current, on-premises data and associated queries to a cloud computing data platform. These studies investigate the performance of the on-premises workload in the cloud.

POC projects can help to clarify the operational requirements of moving production workloads to the cloud for the IT organization. (For the deployment of new applications, the DevOps teams are also involved.) While POC projects are quite extensive and sometimes do clarify the technical requirements associated with a new cloud computing initiative, they do not answer the critical question about the cost required to meet SLGs for production workloads in different cloud environments.

Fortunately, the comprehensive, end-to-end, cloud computing capacity management framework that we describe here can reliably estimate the minimum configuration and cost required to meet the SLGs of growing workloads.

Our process begins with understanding the customer workload and its resource requirements.

Understanding the Customer Workload

Choosing workloads for migration to the cloud is challenging for customers. A major part of that challenge is the lack of workload metrics. For each workload, customers should know:

- How many queries are processed during different times of the day by each workload?
- What is the distribution of response times by workloads?
- How much CPU resources are used by different workloads?
- How much I/O is requested by different workloads?

Without these metrics, it is impossible to evaluate if a new platform is likely to meet expectations and what are the associated costs.

BEZNext eliminates this quandary with tools and techniques to report on key performance metrics, including response time and throughput, resource consumption, and data usage by each business workload. These workload profiles are used as input for modeling to evaluate options and enable informed decisions.

An example of an actual customer's production workload characterization study is shown in Figure 1, where the response time data for all candidate workloads for a representative 24-hour period is on Figure 1A and variability of the response times for the Finance workload is on Figure 1B.

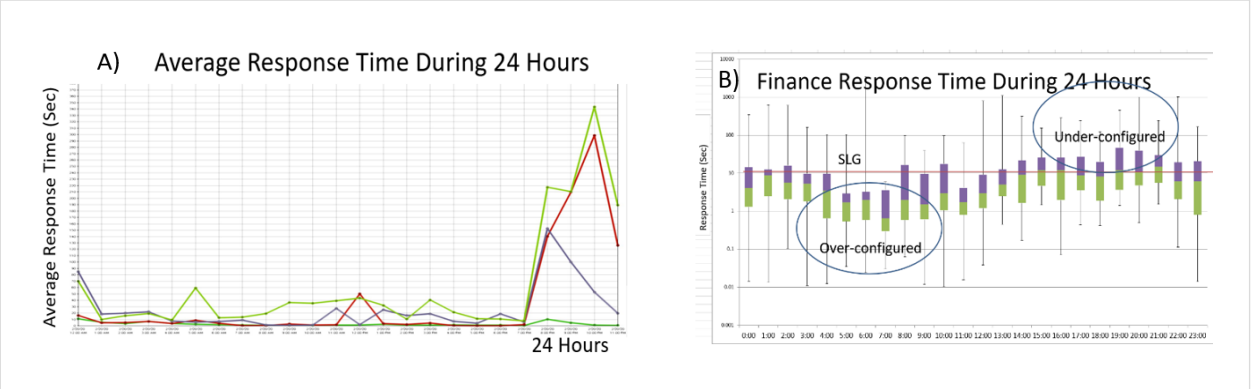

Figure 1. *A) Hourly average Response Time of production workloads on-premises and B) Variability of the response times for the Finance workload. The goal is to avoid over-configuring and under-configuring by allocating just enough resources to meet the SLGs during different hours of day.*

The response time distribution data is mainly considered in the context of understanding how representative any particular sample interval is out of the overall workload resource consumption profile. We also use it here to highlight periods where the current configuration is over-configured for the actual demand.

Figure 2 shows application CPU utilization and disk I/O activity for the same 24-hour period as in Figure 1A. The stacked bar charts break out the resource consumption for each of the major production workloads.

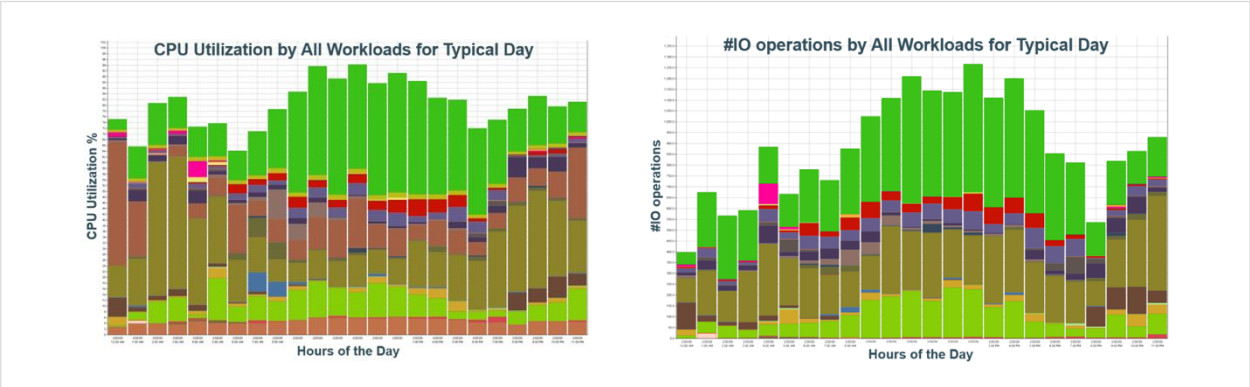

Figure 2. *Example of the CPU Utilization and IO operations by workload during 24 hours for an on-premises system.*

In the case study, the client was mainly interested in evaluating the cost of moving four major workloads to the cloud. CPU usage for these workloads – Sales, Marketing, Finance, and Business Intelligence (BI) – is broken out in Figure 3. It is evident that CPU usage peaks during the second shift (the day shift), especially for the Sales workload. These workloads also have different monthly usage patterns. Moving such workloads to a cloud platform where allocation and deallocation of resources varies according to the actual service demand should result in significantly lower cost.

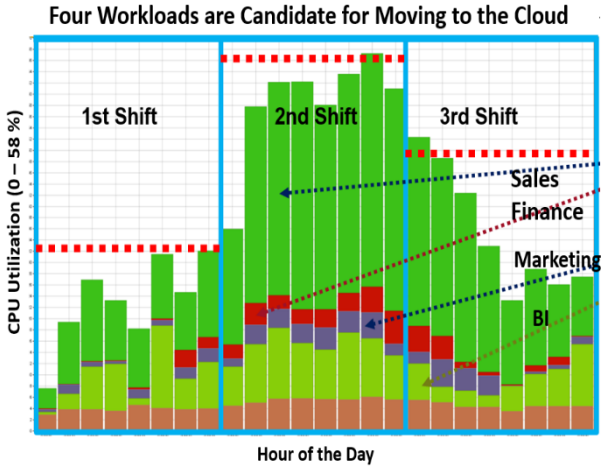

Figure 3. *Daily CPU Utilization profile of the four production workloads chosen for migration to the cloud. CPU consumption during the second shift peak load is twice as high as in the first (night) and third (evening) shifts.*

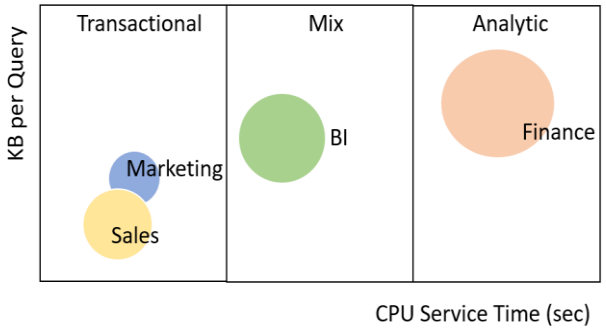

Figure 4. *Comparison of the four workloads along CPU Service Time and KB per Query. Analysis was done using a k-means clustering of the CPU Service Time per Query and KB per Query.*

Our research has found that the analytic and transaction-oriented workloads scaled differently in the cloud. The results of cluster analysis of queries belonging to the four major workloads that were considered the best candidates for migration are summarized in the bubble charts shown in Figure 4.

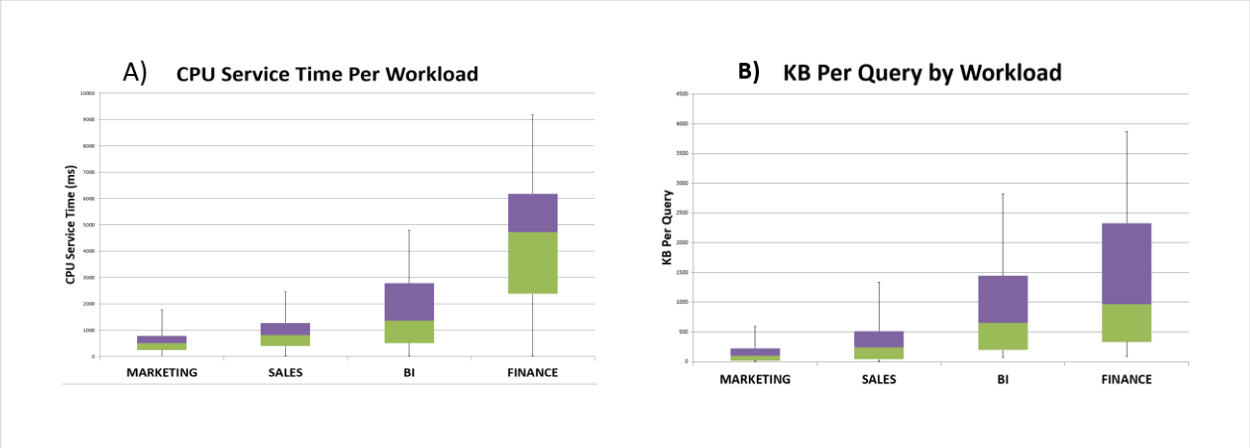

Figure 5. *A). CPU Service Time and B). KB per Query for Marketing and Sales - Transactional workloads, BI - Mix workload - and Finance - Analytic workload.*

We simplified the analysis down to three categories of query complexity:

- **Transaction-oriented queries** processed using less than 1 CPU second,
- **Mixed queries** that required 1-3 seconds of CPU time and scanned 500 KB – 1 MB of data, and
- **Analytic queries** that consumed more than 3 CPU seconds and scanned several MB of data

We assigned the Marketing and Sales workload to the “Simple” category, the BI workload to the “Medium” category, and the Finance workload to the “Complex” category.

We also analyzed variability for each of these workloads by the number of concurrent queries executed and then looked deeper into their patterns of data usage, including the internal parallelism associated with the Query Plan Optimizer. These key characteristics were found to affect the scalability and performance of database queries on the different cloud data platforms.

We examined the major architectural differences between these platforms and factors related to elasticity, DBMS optimization technology, use of indexes to improve query performance, and workload management approaches and rules. The modeling technology we used to predict the performance of the applications across platforms reflects these additional factors.

In production on-premises environments, we forecast the number of users and volume of data. We considered information from the business plan characterizing expected workload growth by each line of business, and we also considered entry of new applications. We collected measurement data during testing of these applications and applied modeling to predict how new applications will perform and how they will affect performance of other workloads in Production; see Figure 6.

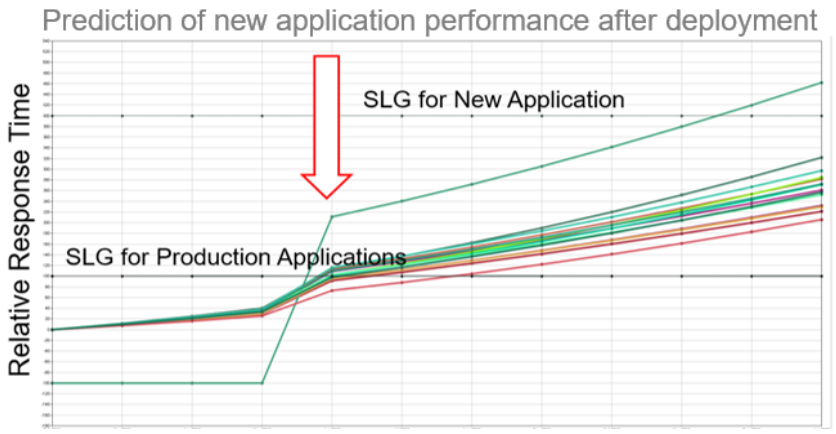

Figure 6. *Prediction of how a new application will perform after deployment on-premises and how it will affect the performance of existing applications.*

In all of this, the purpose of modeling is to determine if a cloud platform can meet the required performance goals, and at what cost. The intent is to always have enough resources, but not too much, to meet the goals.

Cloud and Data Analytic Options

Cost and performance are the major criteria for cloud platform selection. Selection is complicated by the range of vendors and the various database management and data

warehouse systems they support. The Appendix lists the major cloud providers and the data management options they support.

Amazon Web Services leads this market, while Microsoft Azure, Google Cloud, and IBM are among the favorites. Alibaba and Oracle are also runner-ups.

Cloud database platforms employ different architectures, different processing nodes and storage configurations, different DBMS optimizers, and use indexes differently. If you run the same query on-premises and on various cloud platforms, you will encounter different performance and resource consumption results.

Why We Don't Use Standard Industry Benchmarks Results

To compare performance and scalability of the cloud platforms, IT organizations may try to apply published benchmark results. Among the most common to understand the cost and performance of data-intensive analytic applications are TPC-H and TPC-DS, which both try to simulate Decision Support workloads. The Transaction Processing Council (TPC) is an independent industry group; it defines the benchmark workloads and audits the results of benchmark experiments run by hardware and software vendors submitted to it for publication. TPC benchmarks explicitly incorporate both performance and cost dimensions to facilitate comparisons across vendors' offerings.

Results of TPC-compliant benchmarks are sometimes available outside the official TPC audit and certification process where they are presented to compare cloud data platform vendor alternatives. For example, the set of results, published by GigaOm in 2019 and based on the TPC-DS specification, compare the cost/performance of Amazon Redshift, Google BigQuery, and Snowflake against Microsoft Azure SQL Data Warehouse (now a part of Azure Synapse Analytics). Microsoft sponsored the comparison project, which, not surprisingly, shows the Azure solution to be consistently faster and less expensive to run than its competitors. When the consultant who performs the benchmarks has direct access to the sponsor's considerable expertise in how to best configure and run its product, but not similar access to expertise for the other platforms, comparisons inevitably suffer some loss in objectivity.

But the biggest issue extrapolating standard industry benchmark results to specific customer workloads is the uncertainty on how well the standard benchmark specifications align with a real-life customer's specific workloads.

How We Test Performance and Scalability

Instead of TPC benchmarks, some organizations develop and perform POC tests based on selected representative queries and data sets. In BEZNext's POC tests we collect measurement data from each cloud platform – Teradata Vantage, Snowflake, and Amazon Redshift – using queries selected from the actual on-premises applications and a subset of the actual data. Vantage, Snowflake, and Redshift were all configured to run on AWS infrastructure acquired for testing. The on-premises data platform providing a baseline was the customer's on-premises Teradata database management system (DBMS).

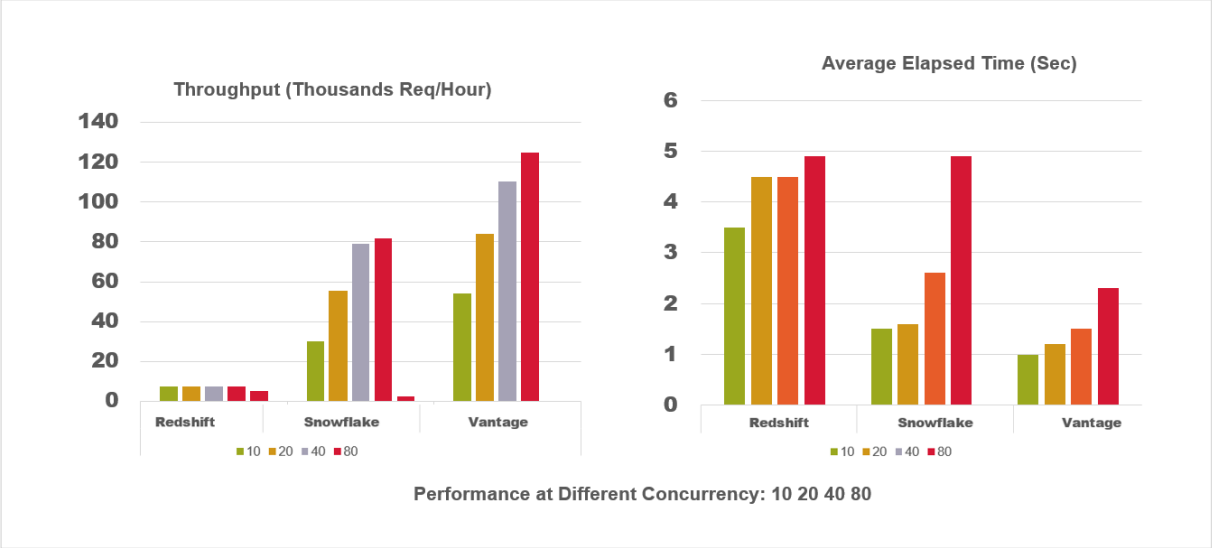

Figure 7. *During a POC test where the on-premises database was replicated to Vantage, Snowflake, and Redshift, throughput and average elapsed time for a representative set of simple, medium and complex requests were measured. Vantage outperformed Snowflake and Redshift.*

BEZNext software agents measured the average elapsed time (response time) and throughput for these representative queries, scaling up the testing successively to 10, 20, 40, and 80 concurrent users. Figure 7 shows that for this customer, both Vantage and Snowflake outperformed Redshift, while Vantage achieved higher throughput with lower response times than Snowflake.

We then analyzed the resource consumption by the test queries and grouped the queries into transactional, mixed, and analytic classes. Table 1 shows the CPU time and data per query for each cloud data platform relative to the resource usage on the original on-premises application.

Comparing the resource consumption for three workload classes (transactional, mixed, and analytic) is also revealing. Snowflake processed the complex requests very efficiently but required 3.3 times ~~of~~ the CPU time for simple transactions, compared to the on-premises baseline. Redshift's resource consumption looks very erratic. Vantage's resource consumption was the clear leader; it produced the most consistent results across all three workload classes using a fraction of the resources required for the queries in the original on-premises environment.

		On Prem	Vantage	Redshift	Snowflake
Transaction	CPU Service Time per Query	1	0.3	9.0	3.3
	KB per Query	1	0.2	54.3	15.1
Mix	CPU Service Time per Query	1	0.3	1.9	0.2
	KB per Query	1	0.3	24.4	1.5
Analytic	CPU Service Time per Query	1	0.3	12.4	0.5
	KB per Query	1	0.1	20.7	0.4

Table 1. *The ratio of CPU Service Time per Query and Read/Write KB per Query for each cloud data platform, broken out by workload class, compared to the original on-premises environment. This ratio characterizes how differences in architecture, hardware, software, and the use of indexes affect the CPU Service Time and KB per Query of each cloud data management option analyzed.*

Selecting the Cloud Configurations

The measurement data that BEZNext agents captured during the POC tests provide valuable information for executive decision-making. However, measurements alone leave open some practical questions:

- What is the minimum cloud configuration required to provide performance comparable to that of the original, on-premises production workloads?
- What is the initial cost of the cloud configuration, and what will it cost in the future to support workload growth in both the number of requests and the volume of data?
- How will new applications going through the DevOps process perform on different cloud platforms?
- What will it cost to support a new application after deployment, and what will it cost to support an expected increase in the number of users and volume of data on different cloud platforms?

There are two software components involved in finding the minimum cloud configuration necessary to meet Service Level Goals (SLGs) for each workload: a prediction engine based on iterative queuing network model and a gradient optimization engine.

On each step the difference between the predicted performance of the modeled configuration and the SLGs determines the size and direction of the next configuration change. Steps are repeated until the predicted performance is better than the SLG. Figure 8 illustrates how gradient descent iteratively determines the optimal platform configuration. Our queuing network models of cloud data platforms can predict the impact of new applications added to existing ones running in the cloud, as illustrated previously in Figure 6.

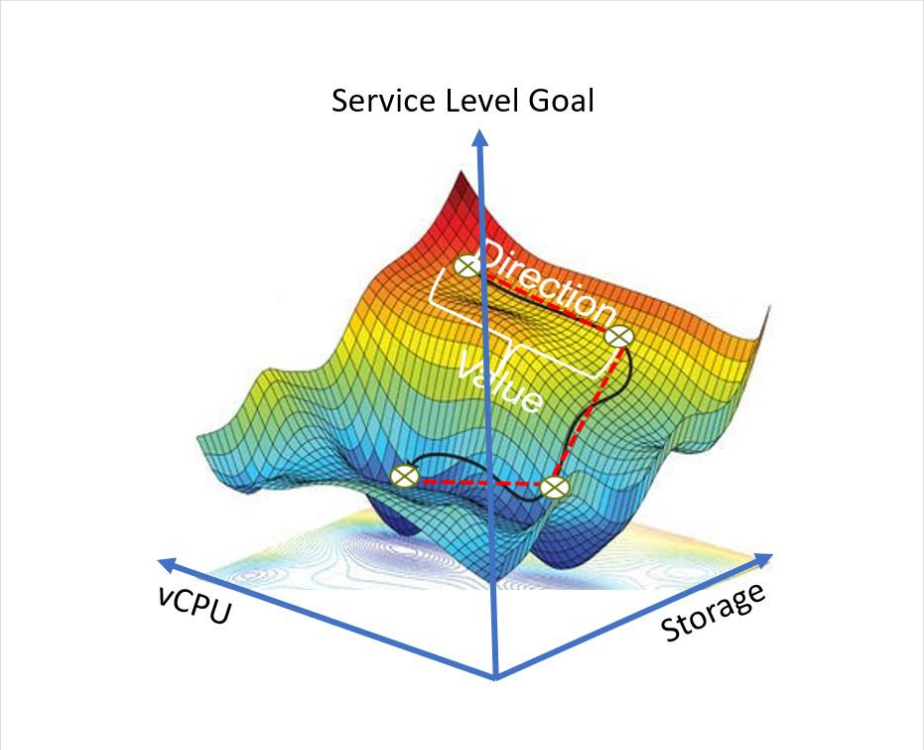

Figure 8. *Iterative modeling and gradient optimization predict the minimum hardware configuration and corresponding workload management parameters required to meet SLGs for all workloads during cloud selection and subsequent dynamic capacity management. The direction and value of each step depend on the difference between predicted performance and SLG.*

A key aspect of configuration selection is to consider daily, monthly, and seasonal patterns of workload activity that can potentially be translated into cost-savings using the flexible pricing models of the public cloud services.

The output of the modeling and optimization is a set of the minimum configurations required to meet the SLG for each workload at different hours of the day and different months of the year for each cloud platform. The cost of each configuration is calculated using corresponding platform pricing models, and the platform with the minimum total cost is recommended.

Table 2 illustrates the cloud configurations selected in the case study for consolidation of the four growing workloads. Platform requirements are broken out by shift.

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 2. *Summary of the recommended configurations of each cloud platform required to meet SLGs for four growing production workloads. It includes the Instance Type and minimum Number of Instances which will be required during the night, day, and evening shifts for the next 12 months. Color coding of the table cells is used to emphasize a change in the number of instances.*

The pricing models available for each data platform are summarized in Table 4. We found four different pricing models in use, which were all variations of provisioning-based pricing or consumption-based pricing. The consumption-based pricing model is the simplest to understand, but it is only offered by one of the compared vendors. Consumption-based pricing charges the customer for the amount of logical IO data processed by customer-initiated, successfully completed database queries.

The Fixed Capacity schemes are straightforward once one has translated compute resource requirements to satisfy SLG into units of computing capacity that each vendor employs. In Flexible Capacity pricing it is necessary to break out daily resource usage by hour or shift, something that was illustrated back in Table 2. Each of the cloud computing vendors provides a hybrid pricing alternative, which charges for a combination of fixed capacity and flexible capacity.

Analytic Platform	Fixed Capacity	Flexible Capacity	Hybrid Capability	Consumption (Usage-based)
Teradata Vantage	3-year commit per TCore • Base Tier: $1.04 • Advanced Tier: $1.53 • Enterprise Tier: $1.85	1-year commit per TCore-Hour • Base Tier: $1.83 • Advanced Tier: $2.69 • Enterprise Tier: $3.26	3-year commit per TCore-Hour • Advanced $2.62 (reserved) + $4.27 (on-demand) • Enterprise: $3.09 (reserved) + $5.18 (on-demand)	Per logical IO of completed queries • $3/TB logical IO for 3-year commit
Amazon Redshift	Reserved instance, depending upon instance type • 3-year or 1-year commit • Upfront option		Compute: billed per second of active node in the cluster • dc2.large, $0.25/hr; dc2.8xlarge, $4.80/hr; ra3.4xlarge, $3.26/hr; ra3.16xlarge, $13.04/hr Managed storage (ra3): billed by GB-hour: $0.024/GB-mo Spectrum: $5/TB scanned	
Snowflake	Multiple compute sizes • AWS storage: $23/TB/month • Azure storage: $23/TB/month • GCP storage: $20/TB/month		Compute: billed per second by tier • AWS: standard, $2/hr; enterprise, $3/hr, bus. critical, $4/hr; VPS, $6/hr • Azure: standard, $2/hr; enterprise, $3/hr; bus. critical, $4/hr • GCP: standard, $2/hr; enterprise, $3/hr, bus. critical, $4/hr Storage: billed per TB-month • AWS, $40; Azure, $40; GCP: $35	
Google BigQuery	Flat rate: slots x commitment time Monthly: $2K per 100 slots Annual: $1.7K per 100 slots/mo	Hourly: $4 per 100 slots billed per second, subject to capacity availability when purchased	On Demand: • Bytes processed (read): $5/TB, with 1TB/month free • Storage: Active ($0.02/GB) or Long-term ($0.01/GB), with 10GB/mo free • Storage API: $1.10/TB • Streaming inserts: $0.01/200MB	

***Table 3**. Pricing Models for different cloud platforms.*

	Jan	Feb	• • •	Dec	Annual
Consumption	$134,011	$138,408		$152,772	$1,716,890
EPOD	$140,474	$146,640		$162,056	$1,815,177
Flex	$132,611	$132,611		$132,611	**$1,591,334**
Fixed	$132,753	$132,611		$132,611	$1,591,476

***Table 4.** Flexible Capacity is the most cost-effective pricing model for the four selected workloads in the Vantage environment for this particular customer.*

Table 4 contains an example of comparing Provisioned (Fixed and Flex Capacity), Hybrid (Elastic Performance on Demand, or EPOD), and Pay Only for What Is Used (Consumption) pricing models for the four workloads in the Vantage environment. As we can see, the Flex(ible) pricing model is the most appropriate in this case.

With the predicted minimum configurations, pricing models are used to estimate the budget required for each platform to meet SLGs for each of the four growing workloads and increasing volume of data processed by each workload during different hours of the

day for the next 12 months. Table 5 shows that the Vantage platform provides the lowest total cost and cost per query for the selected, on-premises production workloads.

		Jan	Feb	. . .	Dec	Annual Cost	Relative Cost
Teradata Vantage	Cost per month	$234,778	$241,453		$261,479	$2,964,189	1
	Cost per **query**	$0.0040	$0.0041		$0.0040	$0.0040	
Amazon Redshift	Cost per month	$807,206	$813,466		$926,131	$10,468,877	3.53
	Cost per **query**	$0.0139	$0.0138		$0.0143	$0.0141	
Snowflake (1 system)	Cost per month	$1,255,660	$1,255,660		$1,301,740	$15,528,720	5.24
	Cost per **query**	$0.0210	$0.0208		$0.0196	$0.0205	
Snowflake (4 systems)	Cost per month	$1,658,880	$1,670,400		$1,877,760	$21,519,360	7.26
	Cost per **query**	$0.0287	$0.0286		$0.0290	$0.0292	

Table 5. *Predicted Monthly and Annual cost of supporting growing workloads on the Vantage, Snowflake, and Redshift data analytic platforms.*

Conclusion

We described the BEZNext framework for dynamic capacity management that incorporates workload characterization, analytic modeling, optimization, benchmarking, and business forecasting. We extend traditional forms of computer capacity management to decision-making in selecting and managing a cloud-based data management platform. Additional examples not discussed in this paper compare major cloud providers Amazon Web Services (AWS), Google Cloud, Microsoft Azure, and others.

Our goal is to provide a plan for cloud deployments that aligns the performance necessary to meet stringent business requirements with the associated costs of running that cloud configuration. To take full advantage of the flexible pricing models, one also

needs to predict patterns of daily, monthly, and seasonal activity of the applications. The planning process also needs to factor in growth expectations for both the amount of data being managed and the number of users accessing that data. It should be understood that new applications might also impact current operations.

In the case study described, modeling and optimization results were used to compare the cost-performance for Vantage, Snowflake, and Redshift data analytic platforms, all running on AWS. This approach can also be applied with other cloud computing alternatives to evaluate the migration of current production workloads from Teradata, Oracle, and IBM to AWS, Azure, and Google Cloud hosting Vantage, Snowflake, Redshift, BigQuery, Oracle Autonomous Data Warehouse, and others.

Compared to our approach, the common methods that many IT organizations employ for adopting and planning cloud configurations often fall short. Standard TPC benchmarks do not accurately represent actual customer workloads that they intend to migrate to the public cloud. On their own, these benchmarks cannot reliably determine the minimum cloud computing configuration that is required to support an organization's performance goals. Similarly, proof-of-concept projects that help clarify the technical feasibility of migrating current operations to the public cloud also fall short. Adopting these simplistic approaches creates a higher risk that the cost to support a business' Service Level Goals in the cloud will be significantly higher than expected.

The BEZNext modeling and optimization methodology can provide comprehensive answers to these questions of capacity and cost within 1-2 weeks, versus the 6-9 months typically required to conduct a customized POC test of the applications under consideration. And for organizations with Hybrid Multi-Cloud platforms, we also offer Dynamic Capacity Management solutions to optimize resource allocation and workload management to continuously meet SLGs for each workload with the lowest cost.

Useful Links

https://iamondemand.com/blog/the-basics-of-cloud-capacity/;
https://gigaom.com/report/data-warehouse-cloud-benchmark/
https://clouddamcdnprodep.azureedge.net/gdc/gdcGFnQ2v/original
https://azure.microsoft.com/en-us/services/synapse-analytics/compare/
https://techbeacon.com/enterprise-it/devops-cloud-10-steps-success

Appendix

Cloud Provider (market share)	**AWS** (34%)	**Azure** (13%)	**Google Cloud** (6%)	**IBM** (8%)	**Oracle**	**Alibaba** (4%)
Vantage	x	x	x			
Snowflake	x	x	x			
Redshift	x					
DynamoDB	x					
Aurora	x					
SQL DB		x				
Synapse		x				
Cosmos DB		x				
BigQuery			x			
BigTable			x			
Spanner			x			
SAP Hana	x	x	x			
DB2 OLTP				x		
Netezza				x		
DB2 DW				x		
Sailfish				x		
Oracle TP					x	
Oracle DW					x	
Apsara						x

Table 6. *A list of major cloud providers and data analytic platforms they support (Source: Ventana Research in Cloud Based Architecture for Data and Analytics).*