Modeling expands load testing and benchmarking results

Performance and financial decisions optimization in the hybrid multi-cloud world

Boris Zibitsker, PhD and Alex Lupersolsky, PhD,
BEZNext, www.beznext.com

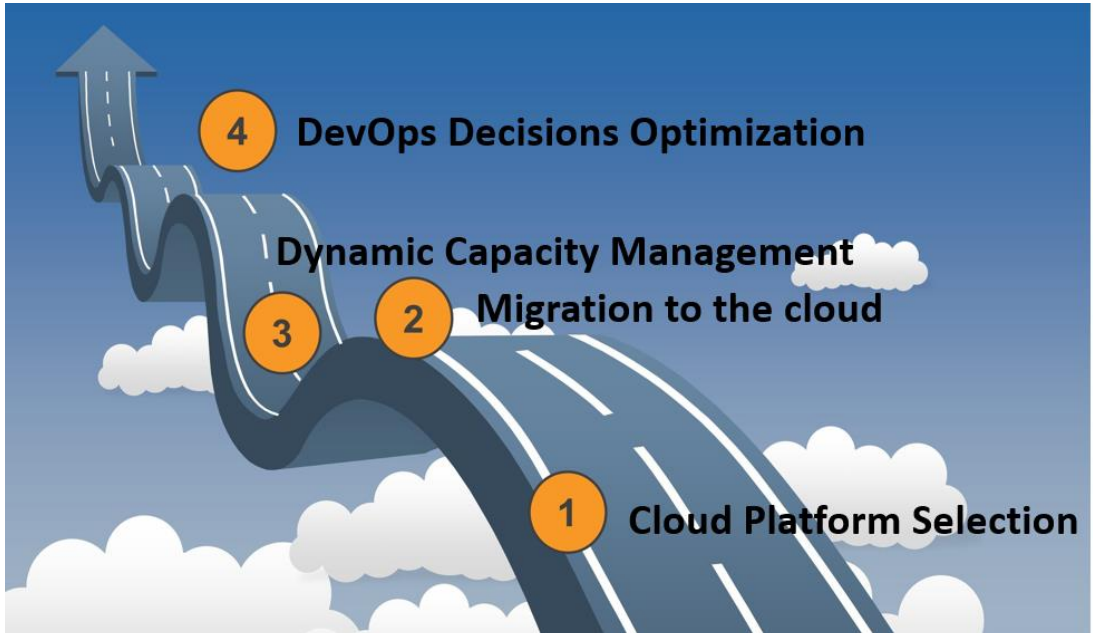
LTB 2023
Load Testing & Benchmarking

The Eleventh International Workshop on
Load Testing and Benchmarking of Software Systems (LTB 2023)

April 16, 2023
Coimbra, Portugal

Co-located with ICPE 2023
The 14th ACM/SPEC International Conference on Performance Engineering

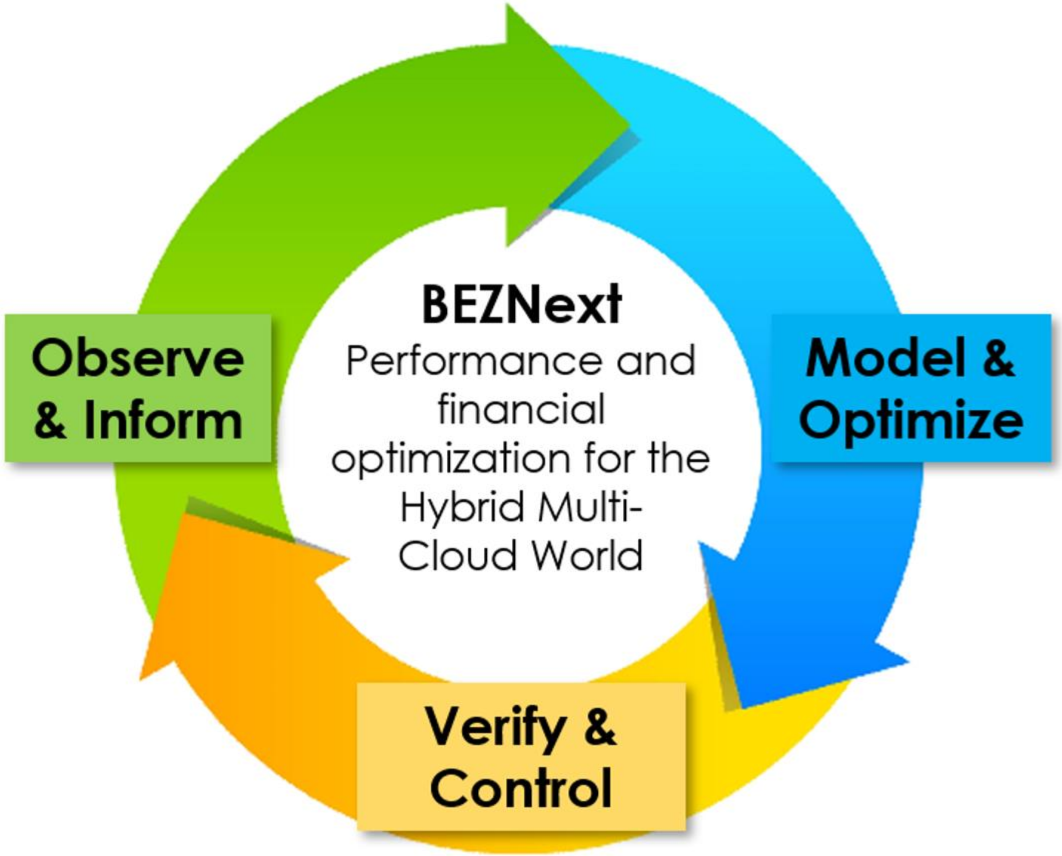
Challenges
Performance and Cost



- How to meet business performance goals in the cloud at the lowest cost?
- Mix workloads with different resource demand and performance requirements
- Many options of resource allocation and workload management
- Uncertainty and risk of performance and financial surprises

Our Goal

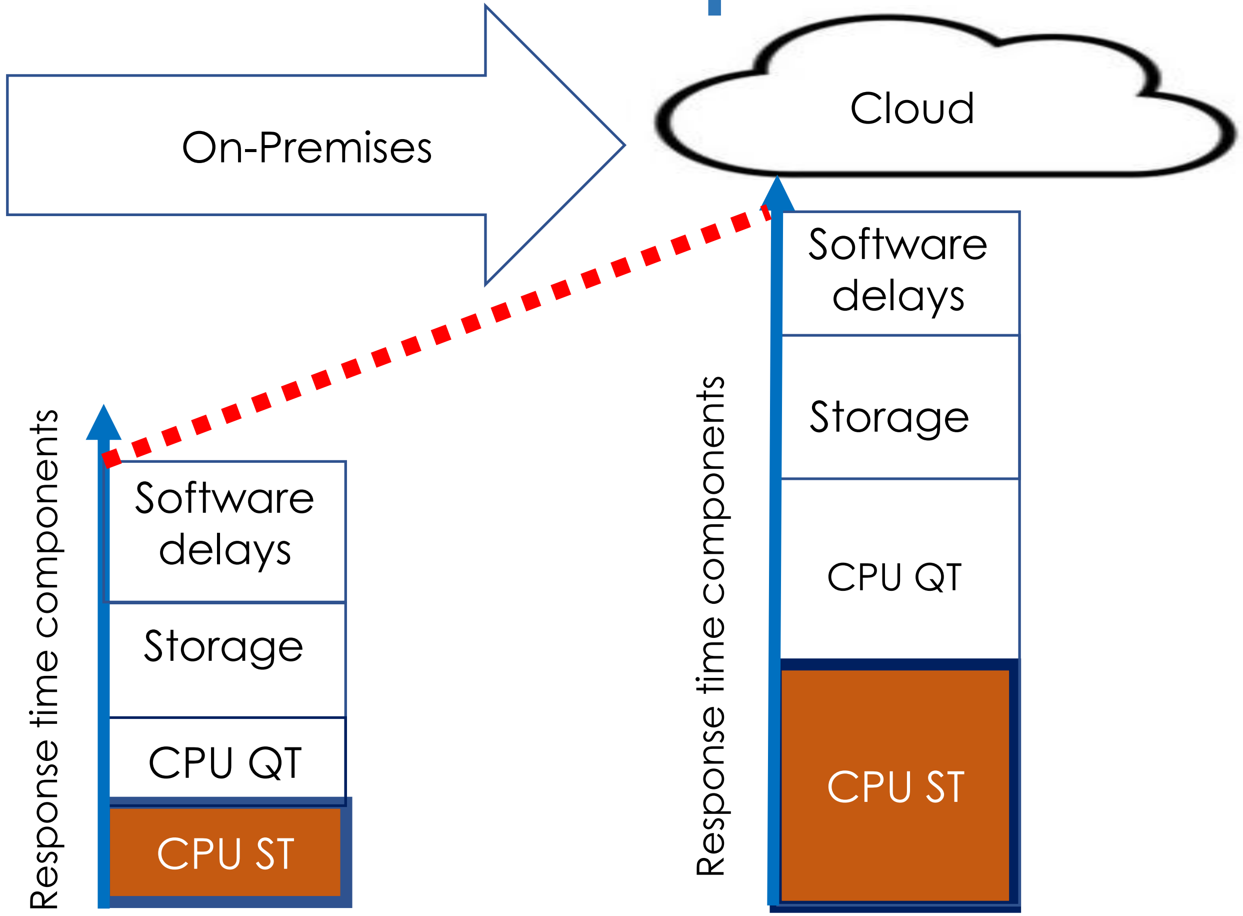
Automate hybrid multi-cloud performance and financial decisions optimization



Observe and Inform

See details in our whitepapers accessible through out website www.beznext.com

We use benchmarks and load testing to compare CPU service time and MB per request in clouds vs on-premises

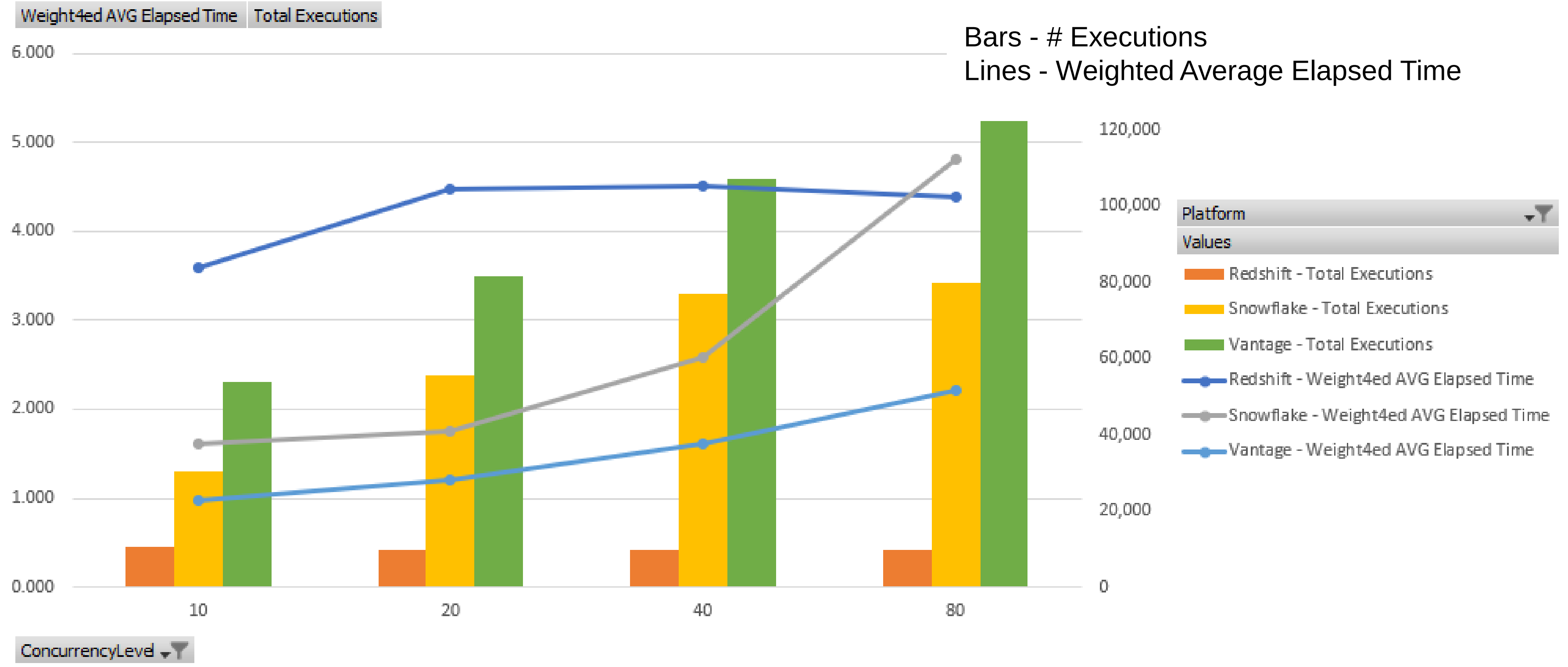


Differences in the CPU type, number of CPUs, software, sophistication of DBMS optimizers, affect CPU Service time

Response time, CPU service time, #I/Os and MB/Query, and queueing time change after migration to the cloud

Load testing results

	Column Labels					
	Redshift		Snowflake		Vantage	
Row Labels	Weight4ed AVG Elapsed Time	Total Executions	Weight4ed AVG Elapsed Time	Total Executions	Weight4ed AVG Elapsed Time	Total Executions
10	3.592	10,655	1.618	30,377	0.976	53,967
20	4.467	9,817	1.748	55,343	1.199	81,385
40	4.501	9,661	2.588	76,943	1.609	107,073
80	4.385	9,886	4.817	79,705	2.213	122,489
Grand Total	**4.222**	**40,019**	**3.008**	**242,368**	**1.626**	**364,915**



Relative CPU Time per Query and MB per query measured during TPC DS Benchmarks on different platforms for short, medium, and complex queries

	CPU Time (ms)						INTERVAL
Teradata on-prem	Vantage to TD	VCL to TD	Redshift 1 to TD	Redshift 2 to TD	Redshift Aqua to TD	BigQuery to TD	by TD CPU Time
1	2.087	2.025	0.226	0.113	0.023	5.431	<130,881
1	1.652	1.580	0.050	0.023	0.007	1.412	130,881-311,093
1	0.758	0.756	0.047	0.031	0.009	1.022	>311,093

Physical IO (MB)							
Teradata on-prem	Vantage to TD	VCL to TD	Redshift 1 to TD	Redshift 2 to TD	Redshift Avg to TD	2XL to TD	BigQuery to TD
1,107	11.405	0.488	0.889	0.962	0.416	406.688	6.934
1,818	14.708	0.586	0.715	0.728	0.430	357.033	6.567
8,799	5.504	0.577	0.191	0.165	0.101	111.957	1.182

Process

- Run benchmark queries serially
- Measured CPU Service Time #IOs and MB/IO

Limitations

- Does not answer business questions

Automation of Inform functions in Hybrid Multi-Cloud environment

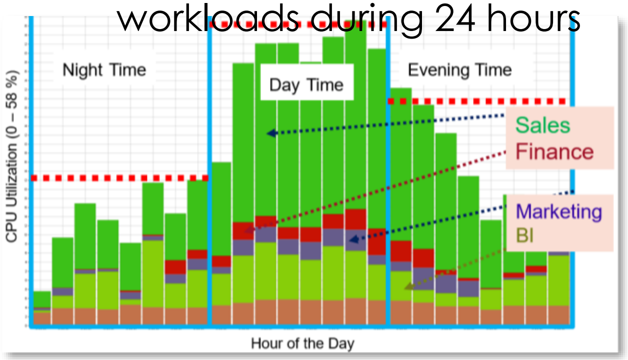
Data Collection

- Performance
  - Response time
  - Throughput
- Resource utilization
  - CPU
  - Storage
  - etc.
- Data usage
  - Databases accessed by workloads
  - Level of parallelism during accessing data
  - Etc.
- Configuration
- Cost

Workload Characterization

- Workload Aggregation
  - Aggregation measurement data into business workloads
- Workload characterization
  - Build performance, resource allocation, data usage, cost and carbon emission profiles for each workload

Example of the CPU Utilization by business workloads during 24 hours



Anomaly detection

- Detect most frequent and severe anomalies
- Root causes
- Critical SQL
- Databases accessed
- Candidates for tuning

Value of Observability and Inform functions

- Automation reduce efforts
- Detect anomalies to focus the performance efforts on the he most frequent, sever problems
- Generate regular FinOps reports

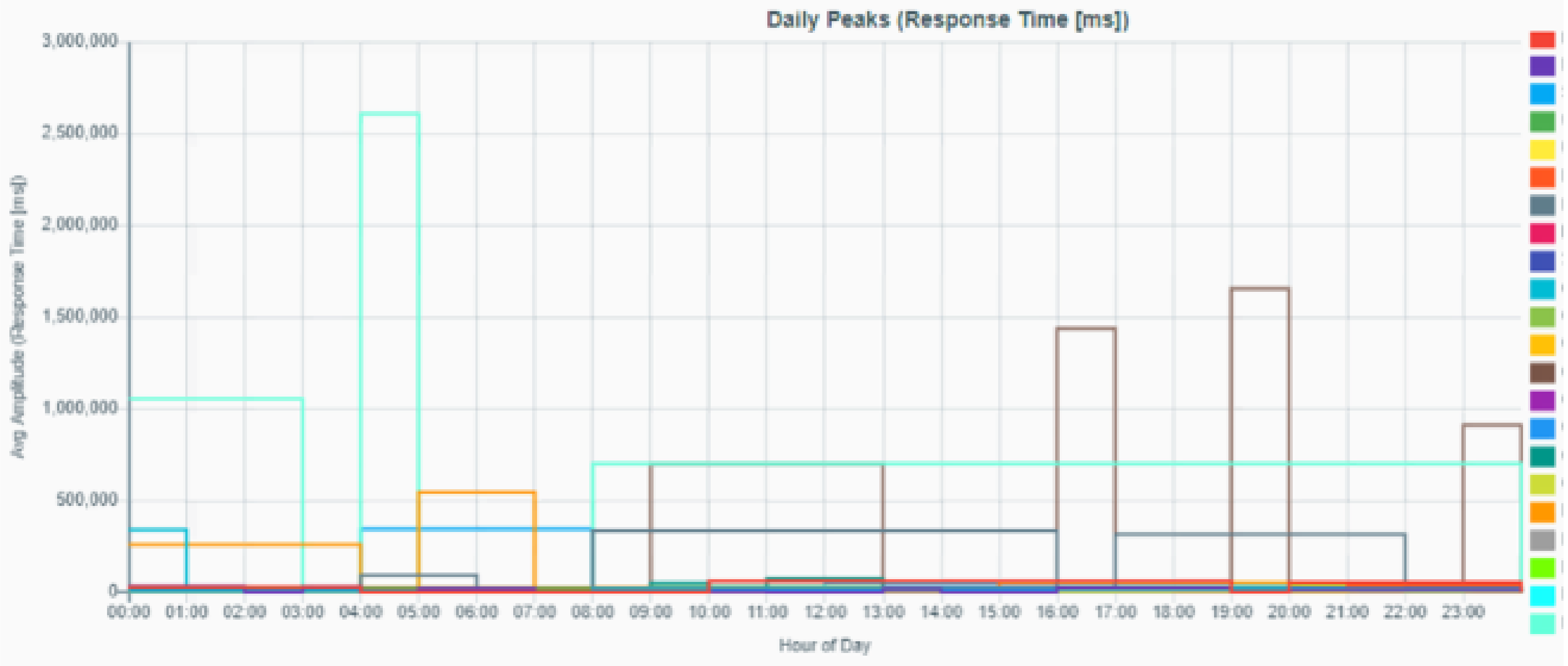
Determine the frequency of accesses to databases by virtual warehouses / workloads

Customer Specific

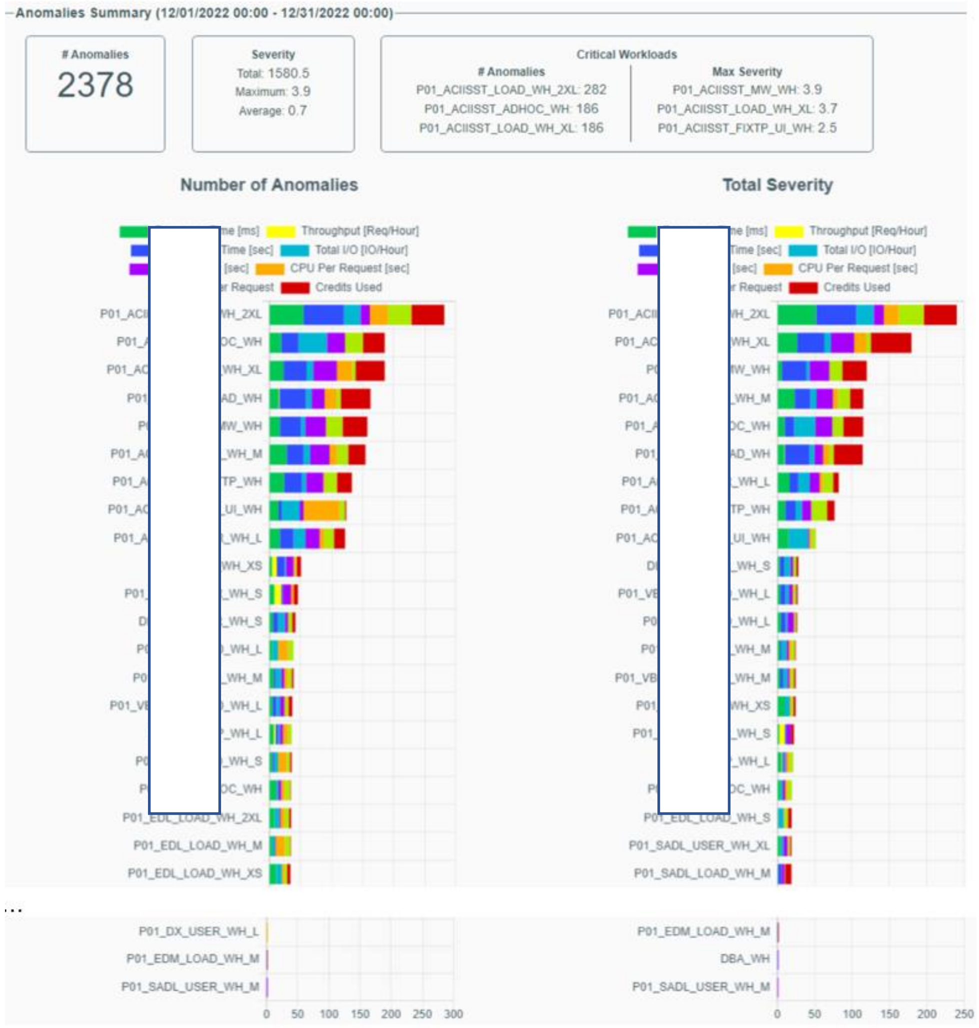
Row Labels	# QUERIES	TOTAL ELAPSED TIME SEC	TOTAL MB PROCESSED
C01_A	**523**	**15,468**	**5,919,596**
P01_AC D_WH	475	3,856	1,235,928
P01_AC _WH_XL	1	1	1,561
P01_ED _WH	29	10,098	3,657,447
P01_ED H_M	1	1	0
P01_VB H_L	17	1,512	1,024,659
C01_A S	**23**	**51**	**66,292**
P01_VB H_L	23	51	66,292
C01_ED	**2,230**	**630,238**	**106,448,025**
DBA_W	15	8	0
P01_DX _WH_L	228	730	596,724
P01_DX _L	48	445	293,849
P01_ED _WH	72	74	1,677
P01_ED H_2XL	127	62,151	10,797,767
P01_ED H_3XL	18	12,546	19,145,243
P01_ED H_L	96	38,448	7,323,541
P01_ED H_M	589	255,647	13,237,666
P01_ED H_S	304	17,296	402,705
P01_ED H_XL	104	218,088	35,534,391
P01_ED H_XS	11	556	17,541
P01_OF WH_L	9	1,606	750,921

- Dozens of Databases with Thousands of Tables, accessed by dozens of business workloads
- Different frequency of accesses with different access time and different MB processed

Business workloads seasonality used during building the resource allocation rules



Determined frequency and severity of performance and financial anomalies is used to determine the root causes and critical SQL needed tuning on of the cloud platform



BEZNext

Determined top queries causing most frequent and severe anomalies are candidates for tuning

Top 10 queries of all root causes ordered by CPU execution time

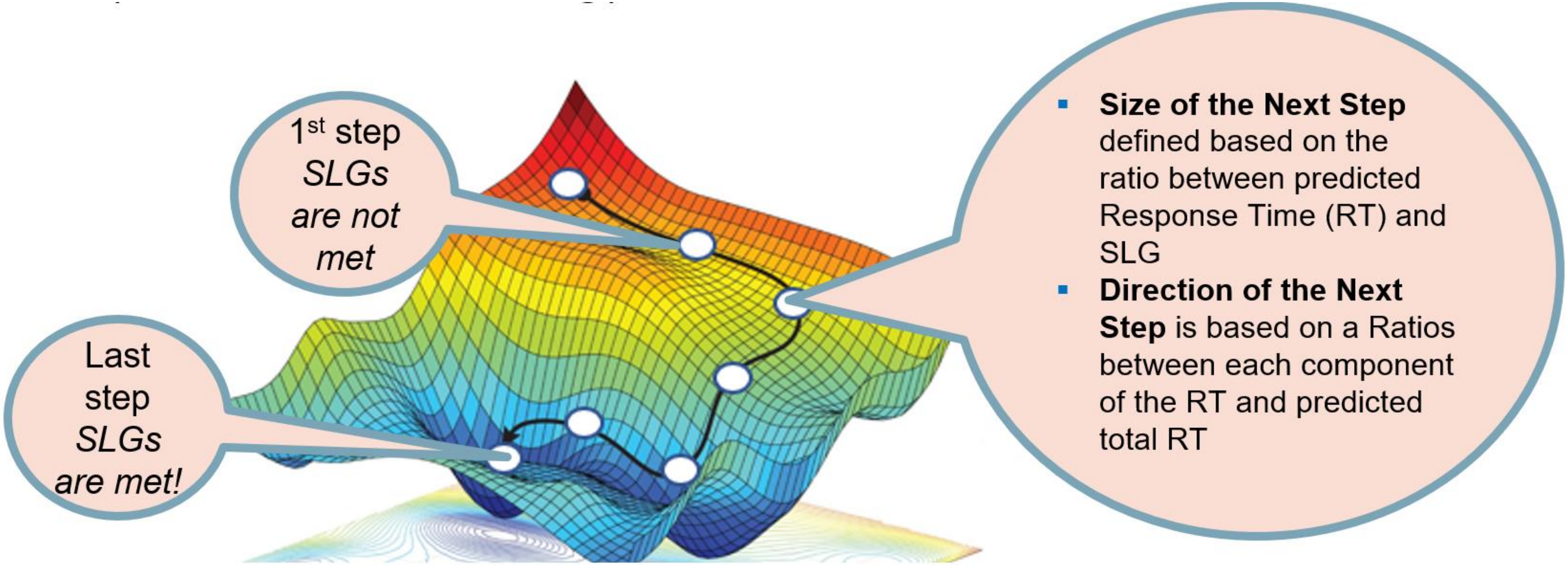
QueryID	applicationId	userName	executionTime	elapsedTime	cpuExecutionTime	totalCpuTime	logicalIOs	logicalIOsKB	physicalIOs	physicalIOsKB
306063201135690131			0:26:07	0:26:08	720,969	720,969	1,248,626,170	143,201,301,856	175,634,683	143,201,301,856
306193201135964541		ML	0:22:56	0:23:13	176,674	176,674	170,822,327	7,634,657,048	12,791,779	7,634,657,048
306473201135772402		ML	0:08:26	0:08:26	166,326	166,326	185,208,046	8,749,232,808	17,569,488	8,749,232,808
306423201138154090		Customer Specific	0:12:13	0:12:20	124,430	124,431	568,270,099	61,955,713,784	85,742,757	14,767,239,040
306333201135806108			0:09:39	0:09:40	119,957	119,958	573,344,173	62,639,315,760	83,872,277	14,614,633,576
306193201135964791		ML	0:10:56	0:10:56	114,067	114,067	102,277,216	4,565,186,396	10,535,569	4,565,186,396
306473201135772443		ML	0:05:44	0:05:48	113,631	113,631	111,353,325	5,258,605,636	12,463,626	5,258,605,636
306423201138153781			0:06:36	0:06:37	113,163	113,164	196,042,613	22,727,423,684	39,656,802	5,749,523,068
306023201136113060			0:14:04	0:14:05	105,219	105,220	668,027,682	55,407,807,244	399,434,868	35,784,833,160
306673201135533582			0:14:51	0:14:52	101,251	101,251	655,230,691	54,082,685,484	384,211,070	34,061,062,252

Query IDs are used to find and tune critical queries

Model and Optimize

See details in our whitepapers accessible through our website www.beznez.com

We developed Iterative modeling and gradient optimization software to optimize performance and financial decisions for the Hybrid Multi-Cloud environments



Examples of the performance and financial optimization projects performed in large Hybrid Multi-Cloud environments

Most of the functions performed by our software are automated

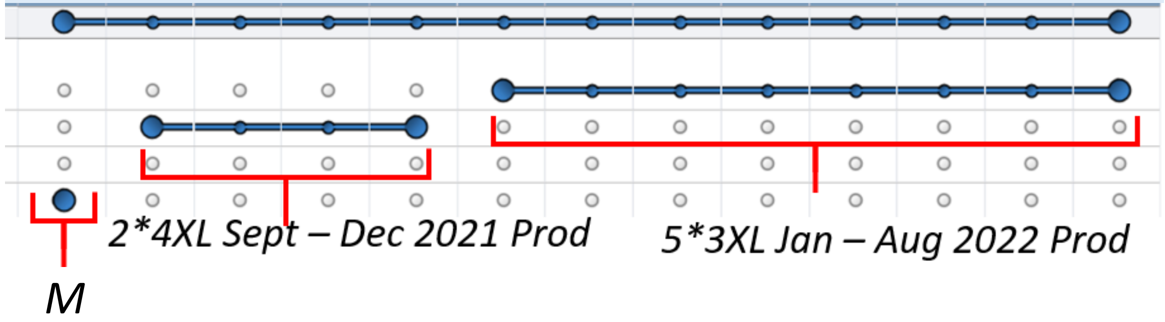
Project 1. DevOps decisions optimization for new applications before deployment

DevOps, MLOps

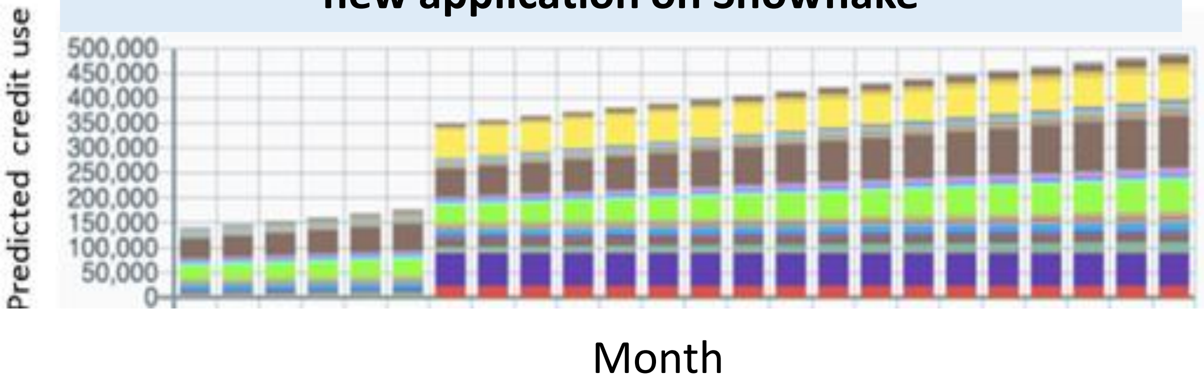
- Performance measurement data collected during testing of new application after each major build
- Detected anomalies and their root causes presented to application developers for tuning them prior to deployment of new application

Recommendation to operations on minimum configuration and budget needed to meet business performance goals on different cloud platforms

Predicted minimum Snowflake configuration needed for new application to meet SLGs



Predicted monthly cost increase after deployment of new application on Snowflake



Project 2. Cloud Data Platform Selection

Objective

Use modeling and optimization technology to determine the minimum configuration, and budget needed to meet performance goals for on-premises and cloud Vantage, Redshift, or Snowflake platforms

Many options

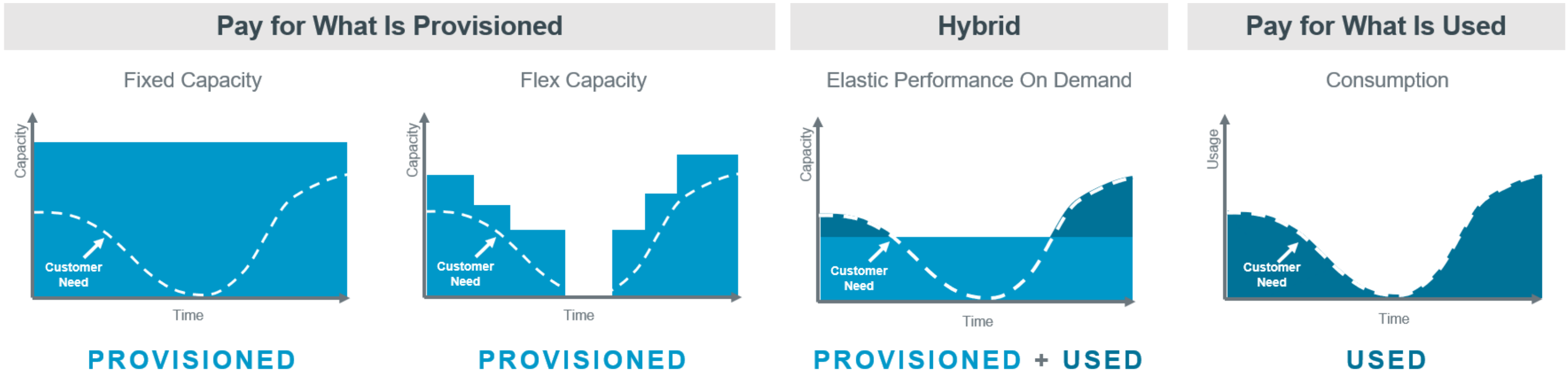
AWS	Azure	GCP Cloud
Snowflake Teradata, Redshift, Aurora, IBM Db2, Neptune, Hadoop	Snowflake Teradata, SQL DB, Synapse, CosmosDB, Hadoop	Snowflake Teradata, BigQuery, Spanner

Predicted minimum configuration needed to meet performance goals

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

BEZNext

Example of selecting the best pricing model for Vantage



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
Consumption	$134,011	$138,408	$138,408	$139,581	$139,581	$139,581	$144,271	$144,271	$146,616	$146,616	$152,772	$152,772	$1,716,890
EPOD	$140,474	$146,640	$146,640	$146,640	$146,640	$146,640	$152,806	$152,806	$155,889	$155,889	$162,056	$162,056	$1,815,177
Flex	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	**$1,591,334**
Fixed	$132,753	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$132,611	$1,591,476

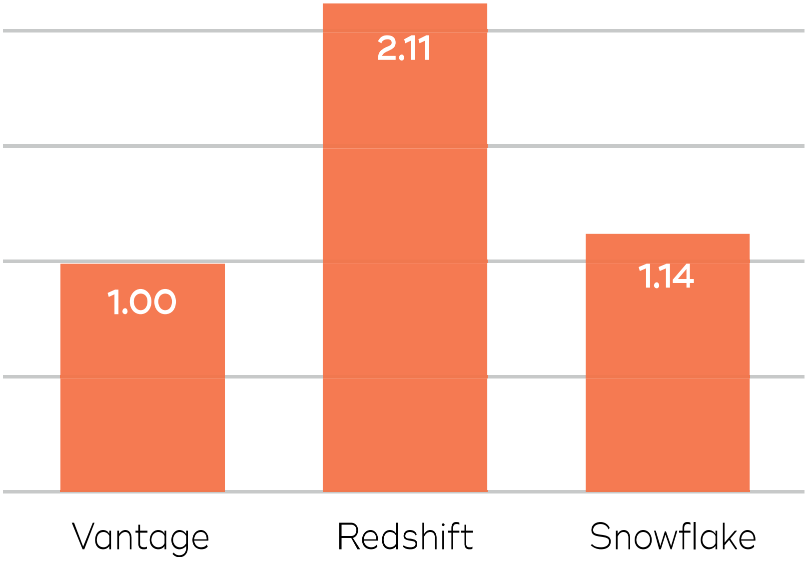
- Predict the monthly and yearly cost for each pricing model needed to meet SLG of each workload
- In this example the Flex pricing model will allow to meet SLGs for business workloads with the lowest-cost

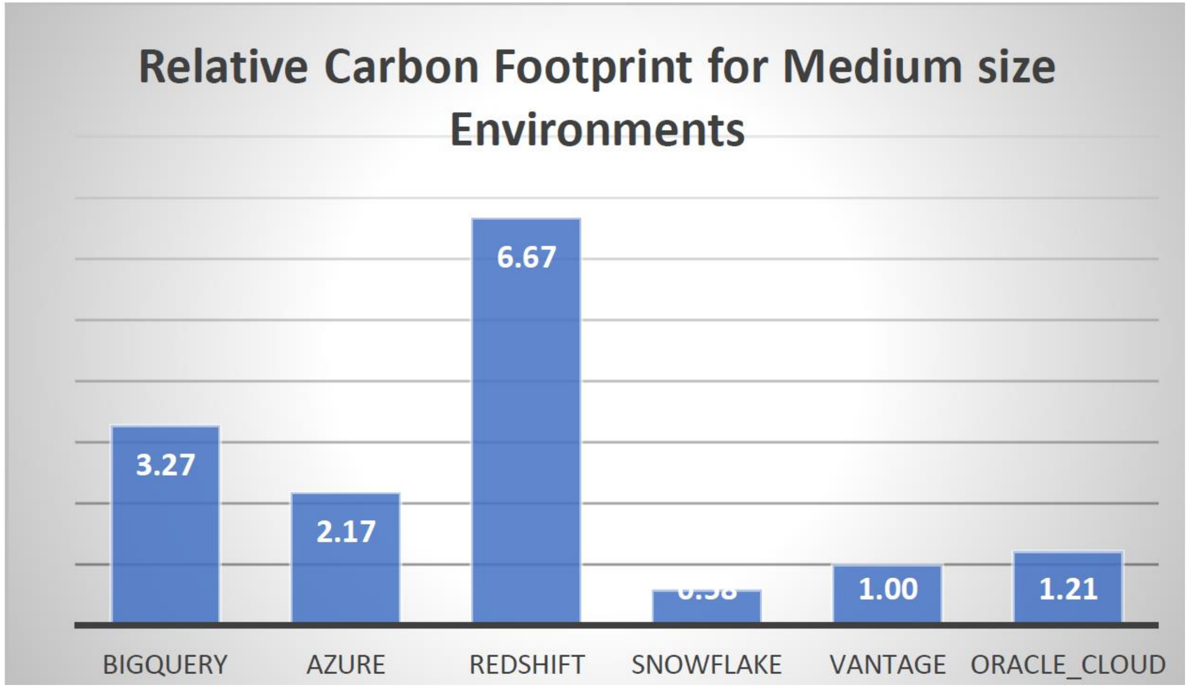
Predict the budget needed to meet SLGs

		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	

Predicted power consumption, carbon emission for production workloads on different cloud data platforms

The relative carbon footprint for large, well-tuned cloud data platforms





BEZNext

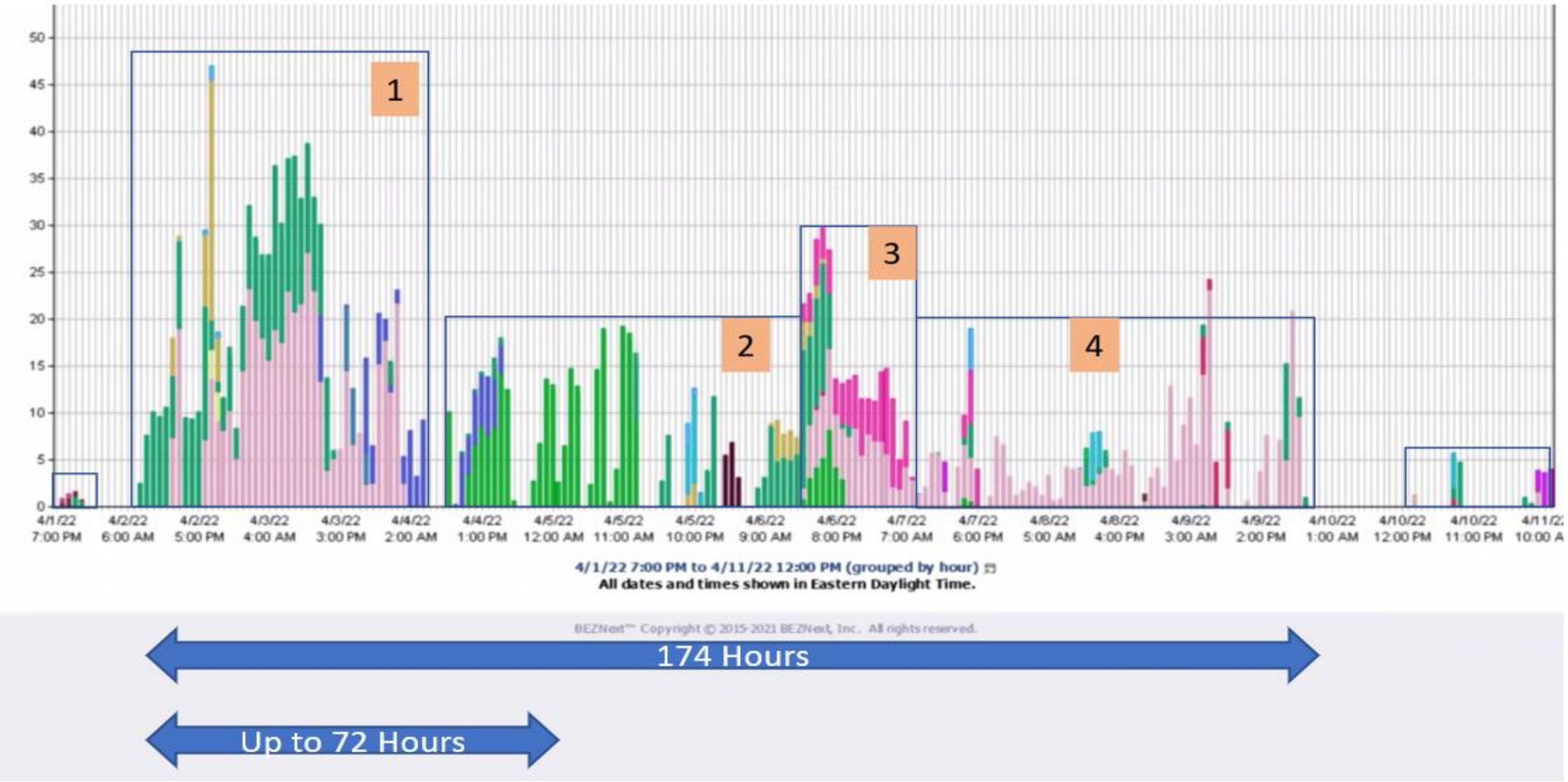
Value of decisions optimization during cloud data platform selection

- Determine **the minimum** configuration and budget needed to meet business SLGs on different cloud platforms
- Set realistic expectations
- Reduce risk of performance and financial surprises
- Duration of the project was 3 weeks

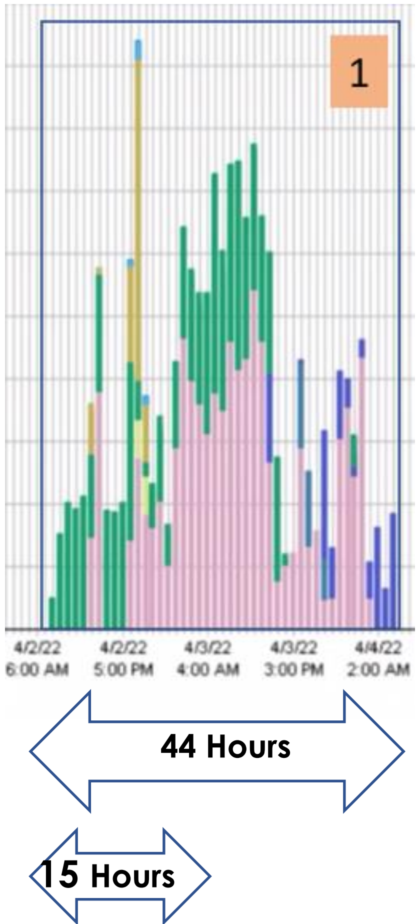
BEZNext

Project 3. Examples of the applying modeling and optimization to determine how to reduce ETL time from 13 to 3days

Each phase of ETL processing has a different demand for CPU resources

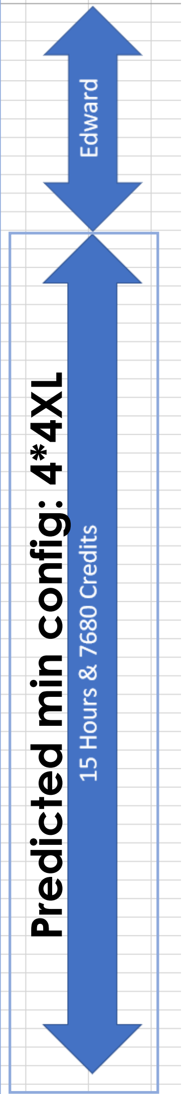


CPU Utilization during ETL 1sr phase of ETL load & Predicted min configuration needed to meet SLG



Measured CPU Utilization-4/02 8am till 4/04 4am

Capitation (%	Claim (%)	Clinical (%)	DRGGrouper (%)	HCC (%)	Member (%)	NST (%)	Pharmacy (%)	Provider (%)	Restated (%)	Revenue (%	Specialty (%)	Trend (%)	UM (%)	All Other(s) (%	SubTotal CPU Util per Hour
0.01	0.01	0.03			0.12						0.35		0.01	0.35	0.88
	0.01				0.03			0.84			0.5			0.02	1.4
					0.94			0.65							1.59
					0.51	0.01		0.2							0.72
					2.44										2.44
					7.64										7.64
					10.15										10.15
					9.58	0.01									9.59
					10.6										10.6
	7.3				6.57		4.08								17.95
	18.83				9.37		0.67								28.87
	0.01				9.47										9.48
					9.38										9.38
					10.08										10.08
	6.94				14.29		7.63						0.61		29.47
	13.67	2.93			3.21		25.51						1.65		46.97
	9.18	2.97			1.06		4.66						0.69		18.56
	8.06				3.53								0.09		11.68
	10.12				6.9										17.02
	4.99				3.37	0.01									8.37
	14.37				7.03										21.4
	23.1				9.01										32.11
	19.73				9.05										28.78
	17.84				9.05										26.89
	15.47		0.01		11.37										26.85
	18.74				17.58					0.01					36.33
	17.41				12.83										30.24
	22.84				14.29										37.13
	20.57				16.81										37.38
	21.46				11.38										32.84
	27.03				11.77										38.8
	22.85				10.13	0.01									32.99
	13.25			7.07	9.86										30.18
	3.78				9.99										13.77
	4.96				1.06										6.02
	6.12														6.12
	14.37		6.77	0.33											21.47
	6.53		6.1												12.63
	7.82														7.82
	2.24		3.31	10.36											15.91
	2.42			4.13											6.55
	15.17			5.5											20.67
	17.67			2.27											19.94
	12.14			0.73	2.56	0.01									15.44
	21.57			1.58											23.15
	2.42			3.01											5.43
				8.18											8.18
				3.3											3.3
				9.25											9.25



Predicted cost, and carbon emission to reduce load time from 13 to 3 days on Snowflake

Predicted **monthly** Snowflake credits / budget needed to meet SLGs, power consumption and carbon emission for Load and analytic workloads

Workload	Credits	Cost	Power KWH	CO2 kg	Objective
Load	15,744	$56,678	270	108	Reduce monthly time from 12 days to 3 days
AdHoc	22,176	$79,834	446	179	RT should be the same as on Prem
Application	16,896	$60,826	290	116	RT should be the same as on Prem
Subtotal	54,816	$197,338	1,006	403	

Value of BEZNext cloud migration decision optimization

- Cloud migration decisions based on gut feelings have a high risk of performance and financial surprises
- Almost 80% of the cloud migration project do not finish on time and within the budget
- Modeling and optimization evaluates options to optimize migration decisions

Project 4. Dynamic Capacity Management

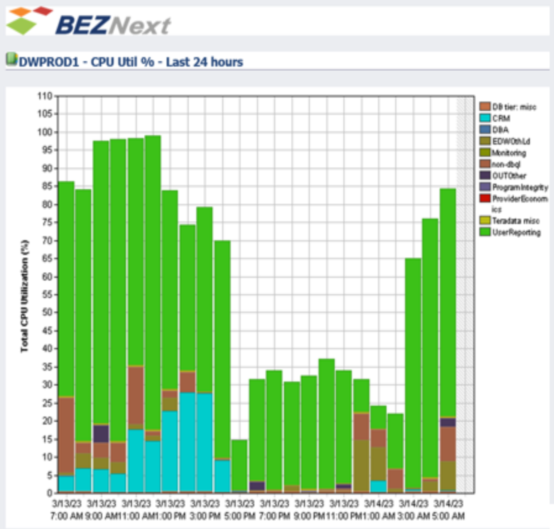
Objective

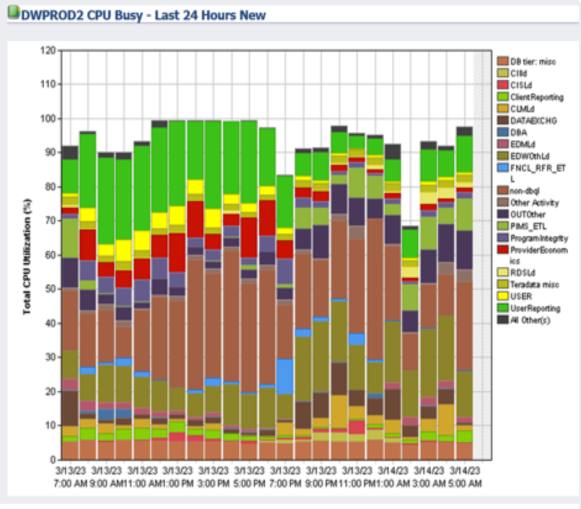
- Organize dynamic capacity management in a Hybrid Multi-Cloud environment to continuously meet SLGs for all growing and changing workloads during different times of the day and the next 12 months with the lowest cost

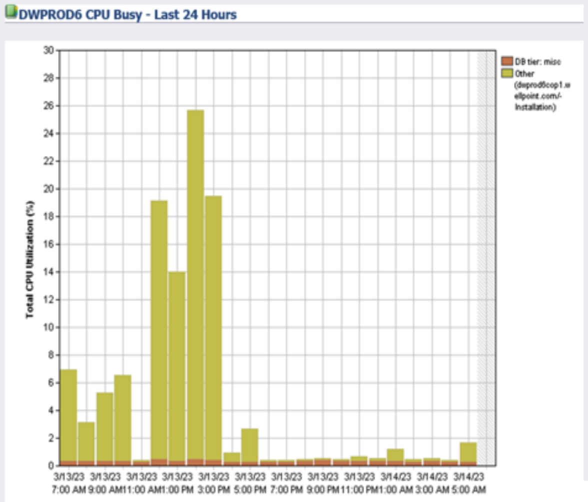
Major Functions

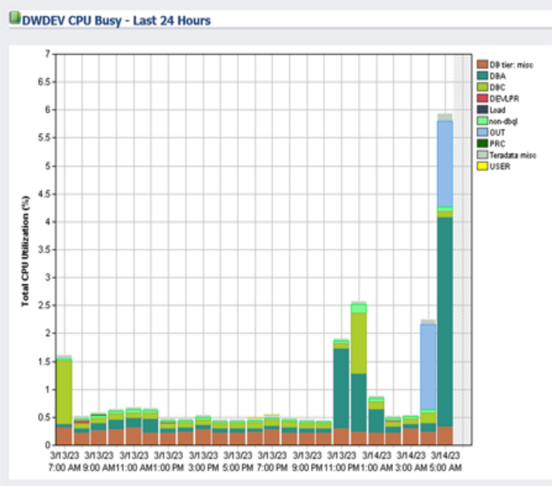
- **Predict performance and budget** required to continuously meet SLGs for all workloads and set realistic expectations
- Determine the performance and financial **anomalies, root causes and seasonality**
- Evaluate alternatives and develop **recommendations** with performance and financial expectations
- **Verify** results
- Organize the continuous feedback **control**

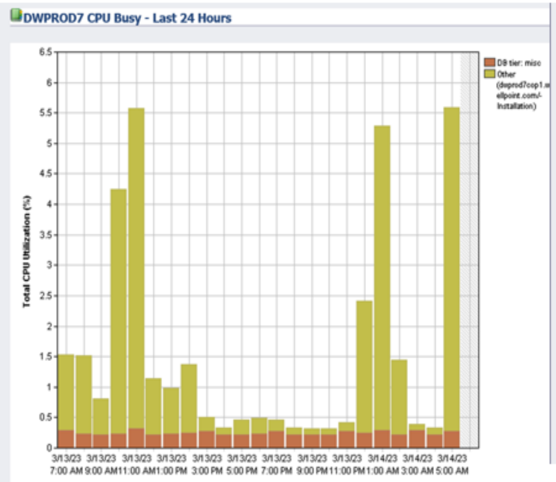
Daily CPU Utilization at on-prem Data Warehouses

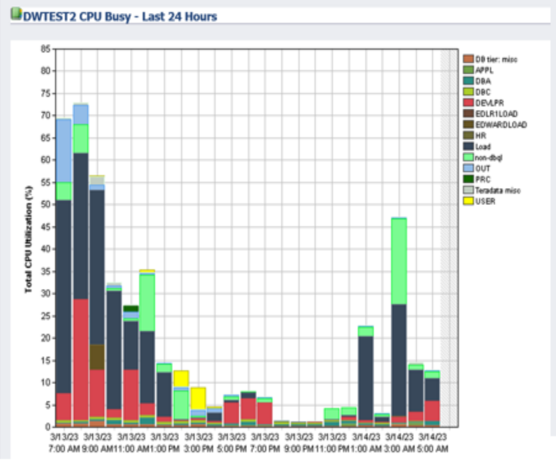


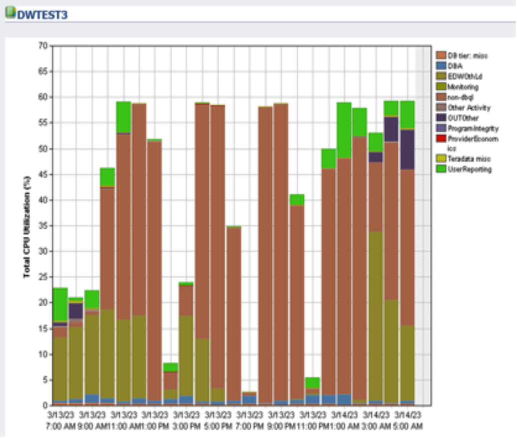


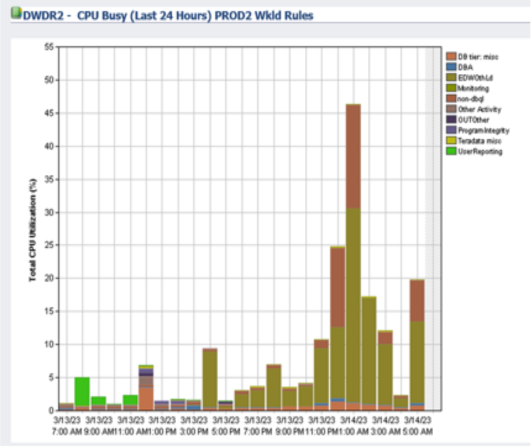












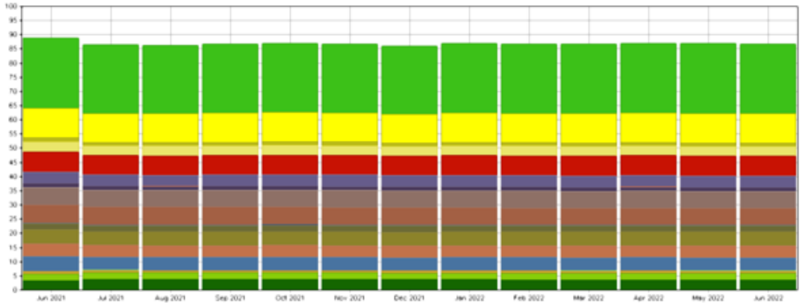
Predicted results of tuning and resource allocation optimization

The example of the recommendations and expectations
- the type of AWS EC2 instance
- the number of instances for each month,
- the expected change in the average query response time and CPU utilization for each workload

Months after migration	AWS Instance Type	# Instances
1	m5.12xlarge	86
2	m5.12xlarge	88
3	m5.12xlarge	89
4	m5.12xlarge	90
5	m5.12xlarge	92
6	m5.12xlarge	95
7	m5.12xlarge	95
8	m5.12xlarge	97
9	m5.12xlarge	99
10	m5.12xlarge	100
11	m5.12xlarge	102
12	m5.12xlarge	104

Relative Response Time (% change)

CPU Utilization (%)



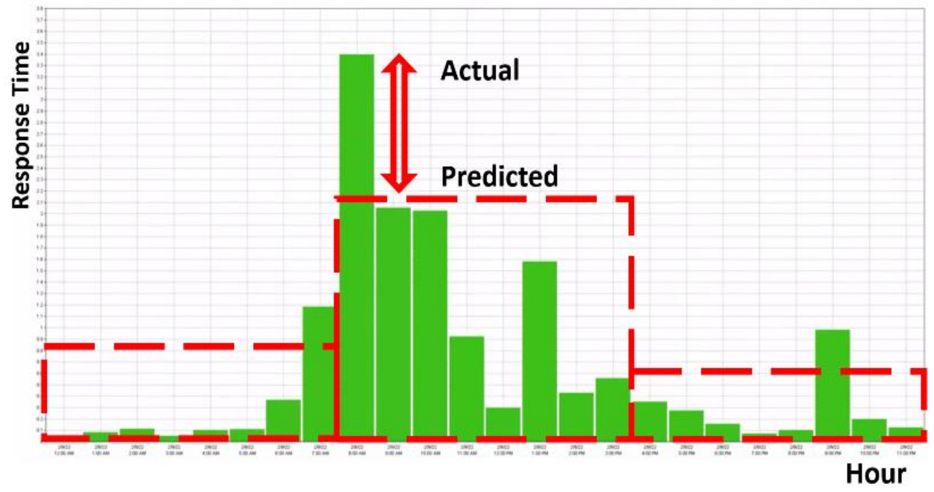
Value of the Dynamic Capacity Management

- Automatic performance and financial anomalies, root cause and seasonality detection
- Performance and financial control for all business workloads on all platforms of the Hybrid Multi-Cloud environment
- Provide realistic performance and financial expectations enabling results verification
- Reduce risk of performance and financial surprises

Results verification and organizing of the closed-loop feedback control

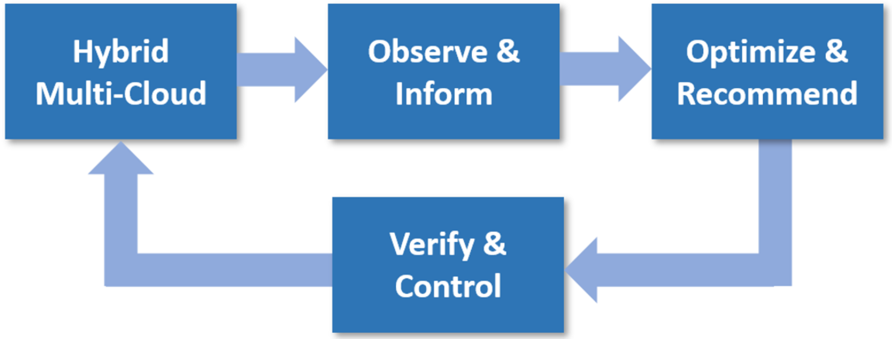
Compare the actual performance and financial results with expected. If results are significantly different, use modeling and optimization to recommend corrective actions

Actual Response Time vs Predicted



Value of closed loop feedback control

- Performance and financial results verification
- Continuous process of performance and financial governance



Summary

Uniqueness

- BEZNext modeling and optimization technology compliment results of load testing, benchmark tests and results collected during the Observe and Inform phase
- Focus analysis and optimization on business workloads and their performance goals
- Apply modeling to optimization performance and financial decisions (FinOps) during journey to the cloud
- Modeling set realistic performance and financial expectations and enable results verification
- A semi-automatic, continuous, closed-loop performance and financial control for the Hybrid Multi-Cloud Environment

Value

- Reduces the risk of performance and financial surprises
- A vendor-neutral approach
- Enables better collaboration between business, IT, and financial leaders
- Our modeling and optimization technology complement other FinOps tools in building and automating enterprise-wide FinOps process

For more information:

1. Read our white papers:

- https://www.beznext.com/wp-content/uploads/2022/02/BEZNext-White-Paper-Which-Platform-is-Best-for-your-Cloud-Data-Warehouse-2-17-2021-1.pdf
- https://www.beznext.com/wp-content/uploads/2022/02/220225-BEZNext-White-Paper.pdf
- https://www.beznext.com/wp-content/uploads/2022/09/220919-Carbon-Emission-Evaluation.pdf

2. Visit our websites:

- www.beznext.com, www.beznextworkshop.com

3. Contract us:

- Inquiry@beznext.com