Boris Zibitsker
Alex Lupersolsky

Predicting Cloud Data Platforms Carbon Footprint for Large Data Warehouses

This white paper focuses on the BEZNext methodology and technology, comparing the power consumption and carbon footprint of cloud data platforms. The case study illustrates how modeling and optimization are used to compare Snowflake, Vantage and Redshift cloud data platforms, optimize cloud migration, dynamic capacity management and DevOps decisions. It helps evaluate different tradeoffs and determine how to meet sustainability targets for your business workloads.

Table of Contents

2 Organization Sustainability

3 How CSP Decisions Affect Power Consumption and Carbon Emission

4 How Cloud Data Platforms Affect Power Consumption and Carbon Emission

4 How Customers' IT Decisions Affect Power Consumption and Carbon Emission

5 Power Consumption and Carbon Footprint Estimation Process

5 Case Study

- **5 Workload Characterization**
- **6 Workload Forecasting**
- **6 Modeling and Optimization**

9 Summary

10 Additional Resources

Organization Sustainability

The perception of a company's sustainability affects its investors, customer purchasing decisions, and bottom line. Sustainability reporting is becoming a corporate necessity. Companies are doing their best to appear sustainable in their annual reporting since it has become an increasing selling point.

The United Nations announced 17 goals of sustainability by 2030 [1]. Power consumption and carbon footprint of cloud data platforms are the significant parts of sustainability reporting.

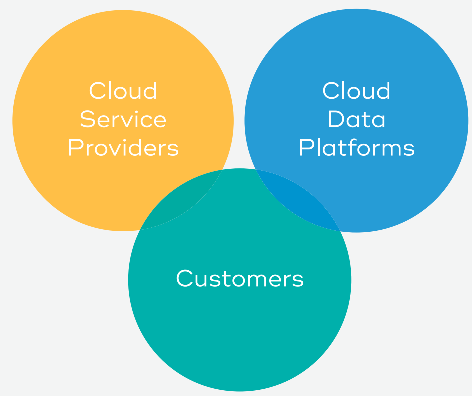


Figure 1. Shared responsibility for power consumption and carbon footprint

An organization's sustainability depends on qualitative and quantitative factors. We will focus on predicting the quantitative factors, including how Cloud Service Providers (CSP), Cloud Data Platforms (CDP) and Customers' IT departments (Figure 1) affect the power consumption and carbon emission while running business applications in the cloud.

CSPs are responsible for optimizing the sustainability of the cloud by delivering efficient, shared infrastructure, water stewardship, and sourcing renewable power.

DBMS optimizers, workload management, and resource allocation strategy by CDPs affect the resources needed to support business workloads in the cloud.

Selecting appropriate CDP, CSP, tuning customer's applications and databases, optimizing cloud migration decisions, dynamic capacity management and DevOps decisions can significantly reduce power consumption and carbon emission.

Several methods and software products are available for carbon footprint estimation [2], but no universal sustainability reporting format has been accepted yet. B-Corp provides a popular standard of sustainability reporting and certification [3]. Gartner expects carbon footprint measurement technologies to see significant adoption as organizations broaden their focus to all emission types and increase reporting transparency [4].

BEZNext (www.beznext.com) developed performance and financial management technology for the hybrid multi-cloud environment, and cloud carbon footprint estimation is a part of that. This white paper describes the BEZNext methodology and technology of the power consumption and carbon footprint estimation.

The case study illustrates how to predict the minimum CDP configurations needed to support business Service Level Goals (SLGs) for large data warehouses on Snowflake, Teradata Vantage and AWS Redshift. Then we calculate the power consumption and carbon footprint of the recommended configurations.

How CSP Decisions Affect Power Consumption and Carbon Emission

CSPs use electrical power and emit carbon while supporting cloud workloads. Selection of architecture, instance types available, CPU and storage options, power and water suppliers, capacity management decisions by CSP affect the power consumption and carbon footprint.

Power Usage Efficiency (PUE) is a ratio of the total data center electricity consumption to the power used to run only the IT equipment. This separates out the air conditioning costs, which are also substantial. For example, the global average PUE for data centers is 1.57, according to the Uptime Institute [5]. Google's PUE in 2021 was 1.10 [6].

We built the regression model that reflects the power consumption dependency on CPU utilization using the recent results of power benchmarks [7] (this year only) shown on Figure 2. Based on this data, AMD processors used two times less power per vCPU than Intel processors.

As we see on Figure 2, an instance with CPU utilization of 80% uses about 3.7 Watts per vCPU on Intel and 1.6 Watts on AMD processors. It will generate approximately 1.48 and 0.64 grams of carbon, respectively.

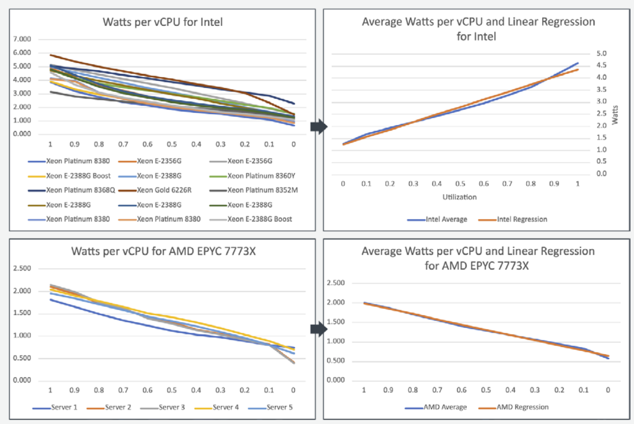


Figure 2. vCPU power consumption depending on the CPU utilization

Over the next three years, cloud providers like AWS, Azure and GCP will come under increasing pressure to have a transparent climate strategy and a clear roadmap to carbon reduction.

Gartner predicts that by 2025 the carbon emissions of CSPs will be one of the top criteria in cloud purchase decisions [8]. For example, Amazon is on the path to powering its operations with 100% renewable energy by 2025 and is committed to achieving its net-zero carbon footprint by 2040 [9].

Even if the CSPs have a lower carbon footprint and are more energy efficient than typical on-premises alternatives, cloud customers do care about the efficiency with which their workloads execute in the cloud. Sustainability in the cloud is a continuous effort focused primarily on energy reduction and efficiency across all components of a workload to achieve the maximum benefit from the resources provisioned and minimize the total resources required. Power consumption optimization will stay an important sustainability measure after the CSP becomes carbon net-zero.

How Cloud Data Platforms Affect Power Consumption and Carbon Emission

CSPs support CDPs like Vantage, Snowflake, BigQuery, Redshift, etc.

The differences in CDP architecture, elasticity and scalability implementation, the efficiency of optimizer, and workload management capabilities of CDP affect power consumption and carbon emissions.

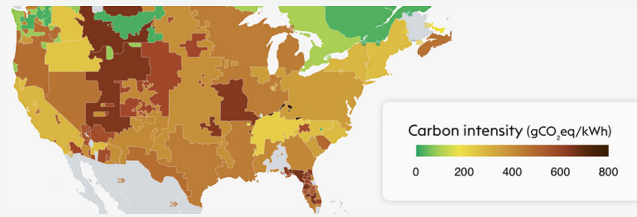


Figure 3. Data center location and energy sources used by the data centers affect the carbon emission [10].

How Customers' IT Decisions Affect Power Consumption and Carbon Emission

Application and DB design and tuning, cloud platform selection, selection of rules affecting allocation of resources and workload management affect the resource utilization and resources needed to satisfy performance requirements of growing workloads.

Well-designed software and energy-efficient hardware can reduce the carbon footprint by 30%-60% [11].

To support workload and volume of data growth in Vantage™, the number of node instances can be increased on-demand. For Redshift, the number of instances can be changed dynamically through API, for example, every shift [12]. Snowflake changes the number of instances automatically based on the number of concurrent queries [13].

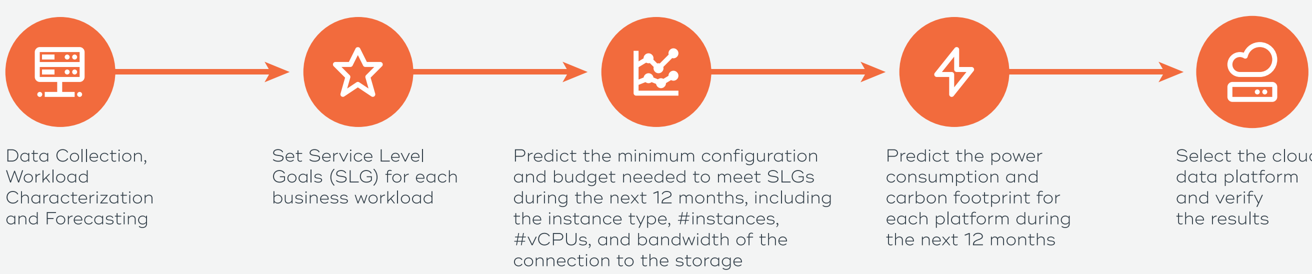


Figure 4. Predicting the minimum configuration, budget, power consumption, and carbon footprint

Power Consumption and Carbon Footprint Estimation Process

The number of vCPUs, memory size, and bandwidth of the connection to the storage are specific to the cloud instance (compute node). BEZNext's methodology finds the smallest configuration necessary to meet SLGs for each data platform on each cloud.

Performance and resource utilization measurement data are collected in production on-premises and cloud environments.

We rely on the customer's provided SLGs and expected increase in the number of users and volume of data for each workload.

With all this data in hand, BEZNext applies its modeling and optimization technology. Foremost, the iterative queueing network models and gradient optimization determine the minimum configuration required to meet SLGs for each workload. This reveals the cloud instance type, number of instances, and the budget needed to meet SLGs for each business workload on each vendor product. To estimate the power consumption and carbon footprint, we use the coefficients presented in Table 1. The process of estimating the power consumption and carbon footprint during the selection of the cloud data platform is shown in Figure 4.

Table 1. Case Study Assumptions

Hardware Component	Wh
vCPU during 1 hour at 45% average utilization	2.1
SSD TB during 1 hour	1.5
HDD TB during 1 hour	0.9
Average carbon intensity of electricity consumed (gCO_2eq/Wh)	0.4

Power consumption in Watts-hour (Wh) per virtual CPU and terabyte stored (SSD and HDD) is averaged over all CPU and disk models [14]. The average PUE is assumed to be included. Herein to simplify the cloud data platforms comparison we assume that all cloud data platforms use the same AWS data center and apply the average carbon intensity of electricity consumed in the USA in grams of CO_2 per Wh.

Case Study

The objective of this case study is to predict the minimum configuration, cost, and carbon footprint needed to meet SLGs for well-tuned Teradata workloads accessing tuned databases, with the optimized workload management controlling workloads priorities and throughput in a large environment. Customer wants to compare the power consumption and carbon emissions on different cloud data platforms, including: Teradata Vantage, Snowflake, and AWS Redshift.

Workload Characterization

BEZNext software aggregates collected measurement data into business workloads and builds hourly performance, resource and data usage profiles.

Business workloads include transaction-oriented and analytic queries.

An example of CPU utilization by well-tuned business workloads during night, day and evening shifts and different hours of the day is shown in Figure 5.

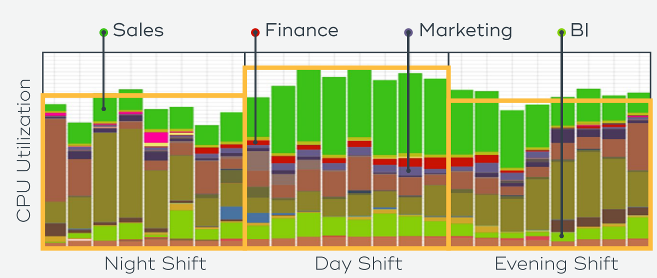


Figure 5. Example of daily CPU utilization by business workloads

BEZNext software determines the performance and financial anomalies on-prem and on each cloud platform (an example of the report is shown in Figure 6). This information is used to create corrective actions.

Workload Forecasting

The number of users and volume of data will be growing. The customer expects a 12% growth in the number of users and a 10% growth in the volume of data.

Modeling and Optimization

BEZNext software converts On-Prem CPU service time and the number of I/O operations for each workload into the same parameters for each cloud platform. Conversion coefficients reflect the efficiency of optimizers and the difference in CPU speed. Coefficients are determined based on benchmark tests.

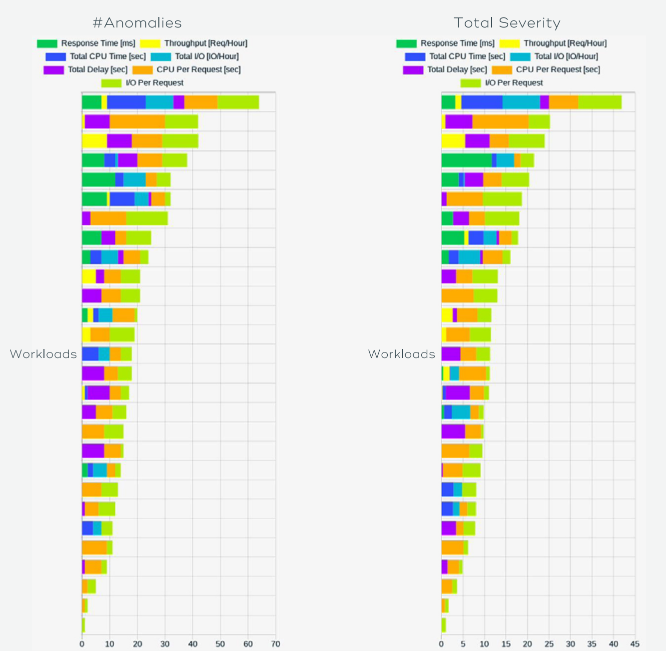


Figure 6. Example of daily CPU utilization by business workloads

BEZNext software [15] applies iterative queueing network models and gradient optimization (Figure 7) to find the minimum configuration and budget needed to meet SLGs for each growing business workload. Modeling results determine instance type and number of instances needed during different times of the day and different months of the year, and corresponding change of workload management rules.

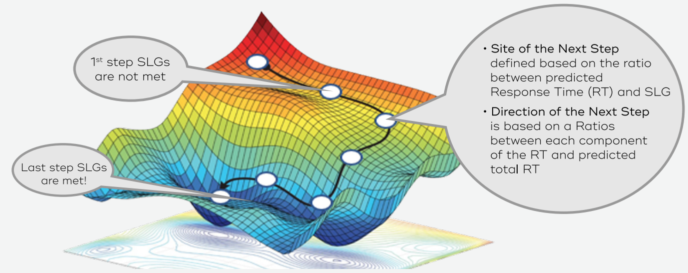


Figure 7. BEZNext Iterative queueing network modeling and gradient optimization

The **budget** is calculated based on predicted minimum configurations and pricing models available for each cloud provider. The instance type, number of instances, and storage space are used to estimate the expected cost of each cloud platform, including cost of the cloud instances, storage, and data management software. These costs are all based on publicly available prices for IaaS instances and CDP software.

The total **power consumption** is calculated based on the total number of vCPUs needed, their utilization, and storage capacity. **Carbon footprint** values are proportional to the power consumed with the average coefficient shown in Table 1. AWS instances used by Vantage, Redshift and Snowflake work on Intel processors. Redshift AQUA has been built on AWS-designed processors with AWS Nitro chips adapted to speed up data encryption and compression, and custom analytics processors implemented in field-programmable gate arrays (FPGA) to accelerate operations such as scans, filtering, and aggregation [16]. As no information about its power consumption was available at the time of the case study, we did not include AQUA power consumption in the Redshift results.

Table 2. Predicted minimum configuration, cost, power, and carbon footprint for Teradata Vantage

4 Workloads on Vantage / Month	1	2	11	12	Total
Night Time Min # Instances	9	9	11	11	
Night Time CPU utilization % on min	72.1	73.1	70.9	71.9	
Night Time CPU utilization % on max	20.3	19.4	20.5	20.8	
Night Power kW*h	742	777	884	888	9,803
Night queries per month	2,911,680	2,928,960	3,265,920	3,283,200	37,160,640
Night W*h per query	0.25	0.27	0.27	0.27	0.26
Day Time Min # Instances	32	34	38	38	
Day Time CPU utilization %	92.5	91.7	91.8	91.8	
Day Power kW*h	1,591	1,680	1,878	1,878	20,920
Day queries per month	44,219,520	45,014,400	49,351,680	49,792,320	564,010,560
Day W*h per query	0.04	0.04	0.04	0.04	0.04
Evening Time Min # Instances	13	13	14	14	
Evening Time CPU utilization % on min	70.80	72.02	77.72	78.89	
Evening Time CPU utilization % on max	28.76	27.54	28.63	29.06	
Evening Power kW*h	841	879	997	1,003	11,091
Evening queries per month	11,085,120	11,197,440	12,225,600	12,329,280	140,408,640
Evening W*h per query	0.08	0.08	0.08	0.08	0.08
Monthly cost with storage, $	$342,730	$351,782	$383,467	$383,467	$4,352,654
Total power kW*h with storage					57791.3
Average power W*h per query					0.08
Annual Carbon Footprint (Kg)					23,117
Average Carbon Footprint g per query					0.03

Table 3. Predicted minimum configuration, cost, power, and carbon footprint for AWS Redshift

4 Workloads on Redshift / Month	1	2	11	12	Total
Night Time Min # ra3 Instances	52	52	58	60	
Night Time CPU utilization %	56.6	57.5	60.1	59.4	
Night Power kW*h	1,424	1,436	1,644	1,689	18,538
Night queries per month	2,911,680	2,928,960	3,205,440	3,265,920	36,944,640
Night W*h per query	0.49	0.49	0.51	0.52	0.50
Day Time Min # ra3 Instances	130	130	150	150	
Day Time CPU utilization %	88.2	88.7	88.9	89.5	
Day Power kW*h	4,691	4,709	5,444	5,466	61,405
Day queries per month	44,392,320	44,763,840	49,014,720	49,317,120	563,967,360
Day W*h per query	0.11	0.11	0.11	0.11	0.11
Evening Time Min # ra3 Instances	72	74	82	82	
Evening Time CPU utilization %	57.5	57.1	60.3	61.3	
Evening Power kW*h	1,990	2,037	2,330	2,351	26,164
Evening queries per month	10,955,520	11,085,120	12,139,200	12,234,240	139,345,920
Evening W*h per query	0.18	0.18	0.19	0.19	0.19
Monthly cost with storage, $	$824,410	$830,669	$937,075	$943,334	$10,675,315
Total power kW*h with storage					122,085
Average power W*h per query					0.16
Annual Carbon Footprint (Kg)					48834.1
Average Carbon Footprint g per query					0.07

Table 4. Predicted minimum configuration, cost, power, and carbon footprint for Snowflake

4 Workloads on Snowflake / Month	1	2	11	12	Total
Night Time					
WH Size	2XL	2XL	2XL	2XL	
Min # WH	5	5	6	6	
Min # Instances	160	160	192	192	
Night Time CPU utilization %	42.6	43.1	42.6	43.2	
Night Power kW*h	640	644	769	774	8,779
Night queries per month	2,980,800	2,998,080	3,352,320	3,369,600	38,568,960
Night W*h per query	0.21	0.21	0.23	0.23	0.23
Day Time					
WH Size	4XL	4XL	4XL	4XL	
Min # WH	3	3	3	3	
Min # Instances	384	384	384	384	
Day Timr CPU utilization %	80.9	81.3	85.6	86.1	
Day Power kW*h	2,211	2,218	2,294	2,304	27,068
Day queries per month	45,645,120	46,111,680	50,645,120	50,492,160	577,316,160
Day W*h per query	0.05	0.05	0.05	0.05	0.05
Evening Time					
WH Size	3XL	3XL	3XL	3XL	
Min # WH	5	5	5	5	
Min # Instances	320	320	320	320	
Evening Time CPU utilization %	30.4	31.0	36.3	36.9	
Evening Power kW*h	1,102	1,110	1,188	1,197	13,782
Evening queries per month	11,197,440	11,318,400	12,389,760	12,510,720	142,300,800
Evening W*h per query	0.10	0.10	0.10	0.10	0.10
Monthly cost with storage, $	**$774,096**	**$774,096**	**$801,744**	**$801,744**	**$9,565,632**
Total power kW*h with storage					**65,607**
Average W*h per query					**0.09**
Annual Carbon Footpirint (Kg)					**26,243**
Average Carbon Footprint g per query					**0.03**

Power consumption and carbon emission are customer workload dependent. This is especially true if cloud instances can be turned on and off whenever necessary.

Modeling output for each cloud data platform for night, day and evening shift during the next 12 months includes: Instance type, Minimum number of instances needed to meet SLGs, CPU Utilization, Electrical power kW*h consumed by CPUs, Electrical power kW*h consumed by storage, Average power W*h per query, Annual carbon footprint (kg), Average carbon footprint per query (g).

Table 2, Table 3 and Table 4 contain the modeling results for Teradata Vantage, AWS Redshift and Snowflake during the night, day and evening shifts over the next 12 months. The architecture of each platform evolved since this analysis was done. For example, Redshift Aqua was introduced, and the performance of Redshift significantly improved. It also changed the cost.

Figure 8 shows the predicted relative cost and carbon footprint for the three CDPs. In our case study, with complex well-tuned workloads running on a large data warehouse, the Vantage minimum configuration's cost is 2.45 times lower than for Redshift and 2.2 times lower than for Snowflake. The predicted carbon footprint of the Vantage is about 1.14 times lower than on Snowflake and 2.11 times lower than on Redshift.

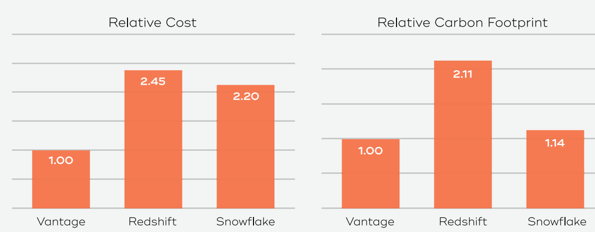


Figure 8. Predicted relative cost and carbon footprint for cloud database platforms in a large environment.

In this case study, we compared Teradata Vantage, Snowflake and Redshift. The same approach can be applied to other cloud service providers and cloud data platforms.

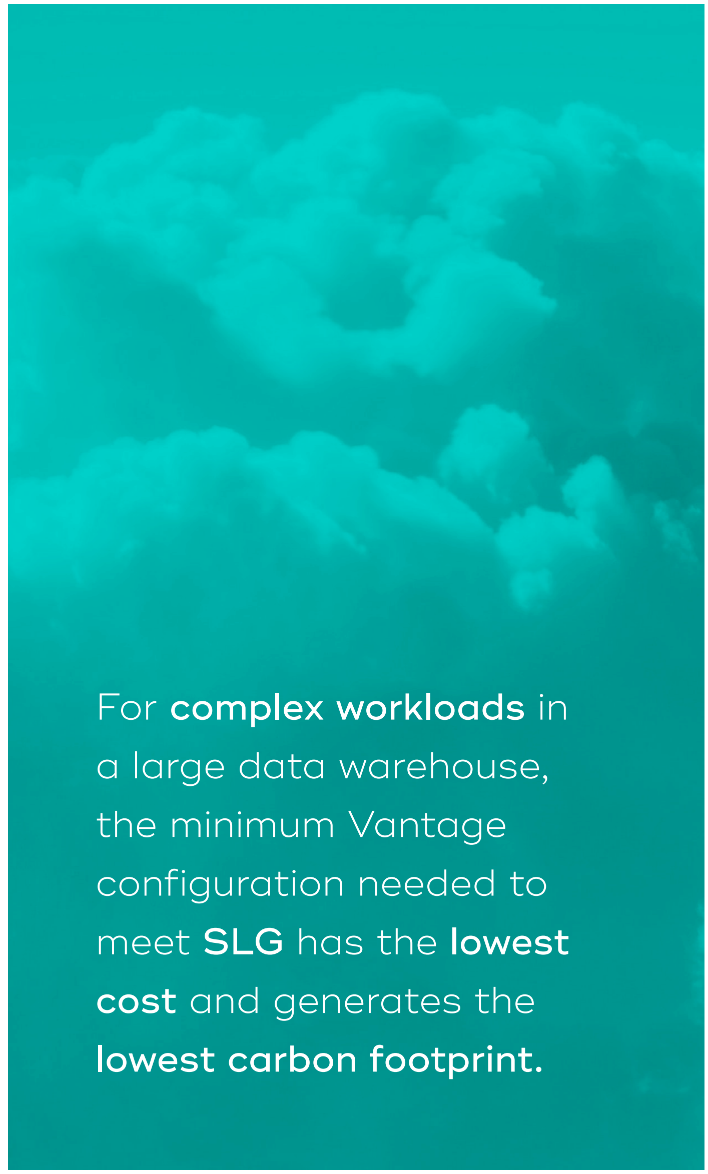
Summary

Customers can significantly reduce the use of resources and carbon footprint by determining the minimum amount of hardware needed to meet SLGs and automatically scale out as loads increase and scale in as loads decrease [17].

We reviewed the process of power and carbon footprint optimization, which includes several steps:

- Measure and predict the impact of your cloud decisions on power consumption and carbon footprint and estimate the impact of proposed changes over time.
- For each cloud workload, predict the minimum configuration, budget and carbon footprint expected to support growing business needs.
- Implement predictive scheduling to reduce the vCPU hours consumed by under-utilized or unused instances.
- Optimize workload placement on cloud data platforms to increase energy efficiency.
- Scale infrastructure to continually match user load and ensure that only the minimum resources required to support users are deployed.
  - Analyze the effect of users on load and capacity utilization over time and respond to changes in demand by scaling in resources during periods of low utilization.
  - Evaluate your workload for predictable patterns and proactively scale as you anticipate predicted and planned changes in demand.
  - Align scaling with cyclical utilization patterns (for example, end-of-month processing activities).

BEZNext modeling and optimization technology predicts the minimum configuration needed to meet business SLGs for business workloads and resource utilization.



Contact Information

inquiry@beznext.com

Additional Resources

[1] United Nations announced 17 goals of sustainability by 2030

- https://sdgs.un.org/goals

[2] Methods and software products available for carbon footprint calculation:

- https://normative.io/book-a-demo-lp
- https://sustainability.aboutamazon.com/carbon-methodology
- https://sustainability.aboutamazon.com/reporting_standards.pdf

[3] Popular standard of sustainability reporting and certification

- https://en.wikipedia.orgwiki/B_Corporation_(certification)
- https://en.wikipedia.org/wiki/Carbon_accounting

[4] https://www.goodbalancer.org/

[5] Uptime Institute survey

- https://uptimeinstitute.com/2021-data-center-industry-survey-results

[6] Google's PUE

- https://www.google.com/about/datacenters/efficiency/

[7] SPECpower methodology and results

- https://www.spec.org/power/docs/SPEC-Power_and_Performance_Methodology.pdf
- https://www.spec.org/power_ssj2008/results/

[8] Gartner conference on sustainability. Gartner Tech Growth & Innovation Conference, July 12-13, 2022

- https://www.gartner.com/en/newsroom/press-releases/2022-04-21-gartner-says-three-emerging-environmental-sustainability-technologies-will-see-early-mainstream-adoption-by-2025

[9] https://sustainability.aboutamazon.com/environment/the-cloud?energyType=true

[10] Carbon intensity of electricity consumed in grams CO_2 per Wh

- https://app.electricitymaps.com/map

[11] An example of efforts to reduce carbon footprint in AWS

- https://sustainability.aboutamazon.com/carbon_reduction_aws.pdf

[12] Amazon Redshift features

- https://aws.amazon.com/redshift/features/

[13] Overview of Snowflake Warehouses

- https://docs.snowflake.com/en/user-guide/warehouses-overview.html

[14] Power consumption estimation in the cloud

- https://www.etsy.com/codeascraft/cloud-jewels-estimating-kwh-in-the-cloud

[15] Which Platform Is Best for Your Cloud Data Warehouse; Zibitsker; 2021

- https://www.beznext.com/wp-content/uploads/2021/07/BEZNext-White-Paper-Which-Platform-is-Best-for-your-Cloud-Data-Warehouse-2-17-2021.pdf

[16] https://aws.amazon.com/redshift/features/aqua/

[17] https://docs.aws.amazon.com/wellarchitected/latest/sustainability-pillar/wellarchitected-sustainability-pillar.pdf

[18] Journey to the Cloud: Performance and Financial Governance Optimization. Zibitsker; 2022

- https://www.beznext.com/wp-content/uploads/2022/02/220225-BEZNext-White-Paper.pdf

Optimizing your AWS Infrastructure for Sustainability

- **[19] Part I: Compute**
  http://aws.amazon.com/blogs/architecture/optimizing-your-aws-infrastructure-for-sustainability-part-i-compute/
- **[20]Part II: Storage**
  http://aws.amazon.com/blogs/architecture/optimizing-your-aws-infrastructure-for-sustainability-part-ii-storage/
- **[21] Part III: Networking**
  http://aws.amazon.com/blogs/architecture/optimizing-your-aws-infrastructure-for-sustainability-part-iii-networking/

[22] Renewable Energy Methodology

- https://sustainability.aboutamazon.com/amazon-renewable-energy-methodology

[23] Predictive Scaling for EC2, Powered by Machine Learning

- http://aws.amazon.com/blogs/aws/new-predictive-scaling-for-ec2-powered-by-machine-learning/

[24] AWS Well-Architected

- http://aws.amazon.com/architecture/well-architected

[25] Sustainability in the Cloud

- https://sustainability.aboutamazon.com/environment/the-cloud

[26] United Nations Sustainable Development Goals

- https://www.un.org/sustainabledevelopment/

[27] Greenhouse Gas Protocol

- https://ghgprotocol.org

[28] https://aws.amazon.com/redshift/features/aqua/