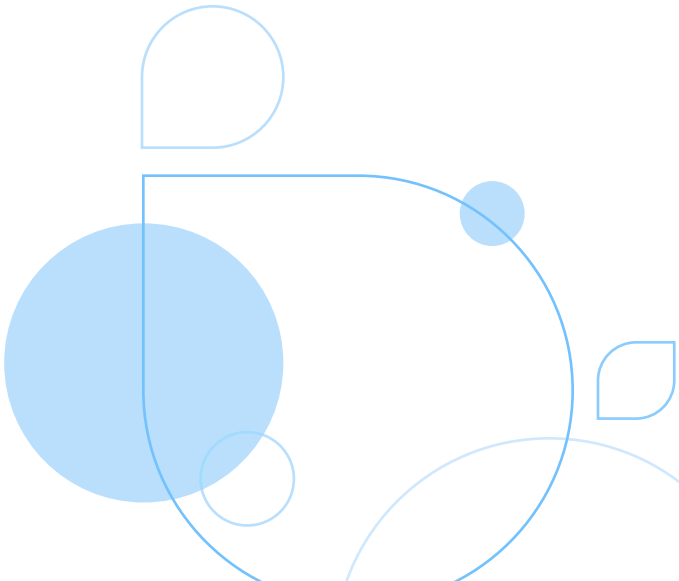
WHITE PAPER

SAS® Viya®: Optimizing cloud costs with new sizing, auto-scaling and workload management capabilities

Contents



Migrating IT workloads to the cloud is a transformational technology change that can result in significant benefits, including greater business agility and lower IT costs. But simply shifting IT costs from a fixed capital expenditure to a variable operating expenditure enabled by the cloud's pay-for-what-you-use model doesn't automatically translate into lower IT costs. It can also make infrastructure costs unpredictable. In fact, many organizations today are struggling to manage escalating cloud spending. Everything from architecture choices to varying service and performance levels can result in skyrocketing cloud costs, leading to sticker shock for many CTO and CFO organizations.

Optimizing cloud costs requires changing how you think about workloads and IT resources because performance is literally money in the cloud computing world. When performance declines, workloads take longer and consume more cloud resources. This, in turn, drives up costs and delays time to value. So, it's incumbent upon IT and finance leaders to architect a performance-optimized cloud technology foundation – one that can meet current and future IT processing demands in a cost-optimized way.

This paper explores the drivers of cloud costs and explains how the new auto-scaling and workload management capabilities of SAS Viya empower you to make proactive, cost-optimized decisions.

The impact of overprovisioning on cloud costs

Total cloud spending is tied to compute, storage, data IO and time spent on running analytic workloads on the cloud. Costs accrue incrementally based on the cloud services you actually use, minute by minute, which means they can be highly variable. In contrast, on-premises data centers require large, up-front costs for a fixed amount of server infrastructure and software licensing. To account for business growth and variability in workloads (such as unexpected spikes), IT departments typically overprovision compute and storage on on-premises data centers, which means your business has to pay for extra infrastructure even when it's not being used.

The cloud's highly scalable, pay-for-what-you-use model eliminates the need for costly overprovisioning. But to realize this benefit, IT teams need to think about resources differently. Too often, when on-premises workloads are migrated to the cloud, this overprovisioned design is carried forward to the cloud infrastructure to handle workloads that far exceed actual peak demands. As shown in Figure 1, cloud resources are provisioned to match the peak compute resources required on-premises, as depicted by the orange line (see "Overprovisioned" diagram on left). Our customer data shows that provisioning compute resources for peak consumption in the cloud, using on-premises as a guideline, can reduce cloud spending only by a nominal amount (see "Peak-Provisioned" diagram on the right), leaving significant room for optimization.

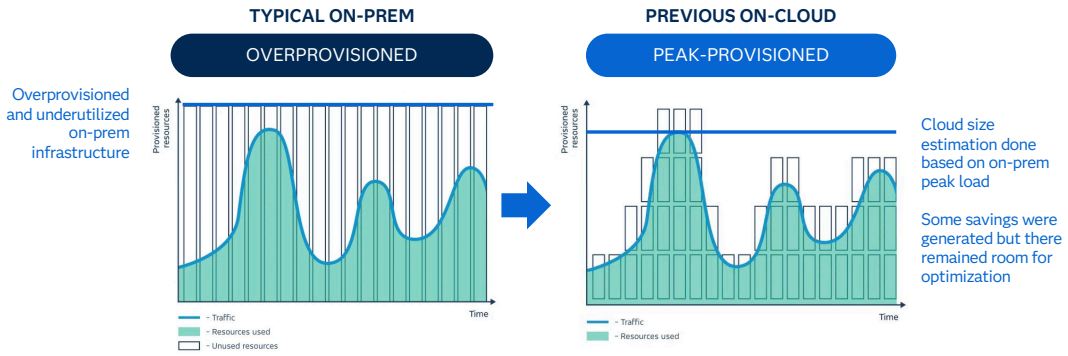


Figure 1: Provisioning based on peak workloads has only a nominal impact on cloud spending.

Rightsizing your cloud environment based on current data about actual workloads can eliminate these costs. As explored in the following sections, this allows you to use inherent cloud benefits – such as auto-scaling – to optimize cloud costs.

New SAS Viya cloud auto-scaling and workload management capabilities can simplify rightsizing, making SAS deployments scalable while reducing the overall cost of ownership. Designed to work with all major cloud service providers (including Azure, AWS and GCP), these capabilities reduce CapEx by recommending that your IT teams start with a smaller infrastructure footprint – one provisioned for baseline workloads rather than peak requirements. Additionally, these capabilities help reduce OpEx by enabling you to scale up and down compute resources to optimally align with workloads running at different times and for different users.

Viya allows for a small initial compute footprint in the cloud. The compute nodes can be auto-scaled up and down with demand, reducing both the initial CapEx and the ongoing OpEx cost of the cloud infrastructure underlying Viya workloads (as shown in the "Rightsized, Scalable" diagram on the right in Figure 2).

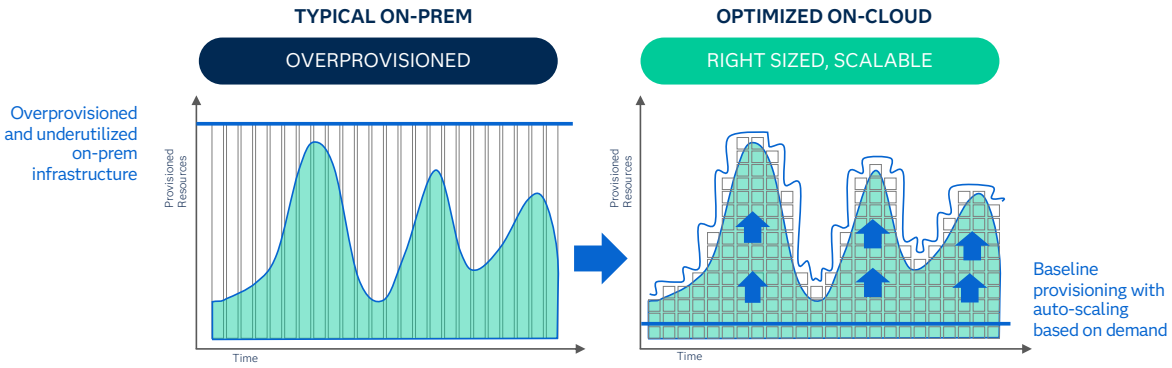


Figure 2: The small compute footprint and auto-scale capabilities of SAS Viya reduce cloud costs.

SAS Viya capabilities

Before the release of Viya, which included auto-scaling, sizing for future-state infrastructure needs was based on estimating peak demand modeled on the current on-premises server and storage architecture. On-premises infrastructure was typically overprovisioned to cater to expected growth, and infrastructure could not be expanded at short notice. SAS code and objects were lifted and shifted to a cloud provider with the same on-premises infrastructure requirements, resulting in an overprovisioned cloud infrastructure. There was also no visibility into how the current on-premises infrastructure was being utilized, which prevented future-state cloud optimization. Given these constraints, future-state infrastructures were designed so that client workloads ran with great performance and availability – but were not optimized for cost.

Viya addresses these issues with sophisticated new capabilities designed to prevent cloud overprovisioning and optimize costs. Let's explore three primary ways Viya does this – and unlocks the potential for significant cloud cost savings.

DEPLOYING VIYA

Viya is a cloud-native and cloud-agnostic data and AI platform. When you work with SAS Managed Services, we help you consume SAS how you want to while meeting your needs and realizing the most value from your digital spending.

SAS Managed Services provides the best of SAS: our software, services, support and expertise. SAS Managed Services shorten the distance between data and value by delivering maximum uptime with minimal administrative overhead. You will see additional advantages like reduced costs, retirement of technical debt and increased speed and security so you can focus on what matters: innovation and real results.

Reap the benefits of SAS expertise from start to finish, with SAS installing, configuring, securing, operating and maintaining its platforms and solutions in the cloud.

#1: Rightsize infrastructure for baseline workloads

The new optimized compute footprint of Viya enables you to start small for your baseline workloads. It may also be possible, in the future, for the separate stateful and stateless nodes to be combined into a services node pool, enabling an additional reduction in the infrastructure footprint. The Compute and Services nodes can be reduced from the earlier 32 and 16 core nodes, respectively, to four core nodes. Additionally, by using an SMP (Symmetric Muti-Processing) CAS server in alignment with your organization's computing requirements, you can reduce CAS infrastructure requirements from a 6-node, 16-core server to a single-node with fewer cores. This can result in at least a 21% reduction in a cloud infrastructure footprint and associated costs, even for the smallest node pools.

For larger environments, the savings could be much more. For example, in a recent client situation, a legacy SAS environment that was costing $450K per year due to significant overprovisioning of infrastructure was reduced to $104K per year due to rightsizing achieved using new Viya architecture recommendations. The result was a 77% savings in infrastructure spending, even on a reserved pricing basis.

A legacy SAS environment that was costing **$450K per year** due to significant overprovisioning was reduced to **$104K per year** due to rightsizing achieved using new SAS Viya architecture recommendations. The result was a 77% savings in infrastructure spending, even on a reserved pricing basis.

#2: Diagnostics-based architecture recommendations help you plan for future cloud requirements

The SAS Ecosystem Diagnostics collection of software utilities (specifically, SAS Content Assessment and SAS Enterprise Session Monitor) can be used to observe user, workloads and storage patterns over a specific time and record a snapshot of all SAS programs, data sets and object inventories created by your organization. This creates a rich set of data points that the diagnostics tools can use to analyze server and storage utilization and then recommend an optimized, future-state cloud architecture. In practice, this means that instead of estimating future-state infrastructure needs on current, on-premises and overprovisioned infrastructure, these tools can analyze current utilization, calibrate future infrastructure needs based on this data and generate recommendations on optimizing your cloud spending.

To illustrate the power of this approach, SAS collected and analyzed data generated by SAS Ecosystems Diagnostics from more than 40 multi-deployment customer environments across different industries and a broad range of SAS estate sizes. The analysis shows that the median CPU saturation is 9.82% (including holidays), and the standard deviation is 21.13%. This implies that CPU hours currently consumed are around 52% (median + 2 standard deviations) of the total CPU hours available on a 24x365 basis. This helps in estimating the number of virtual cores that will be required in an optimized cloud environment, even when accounting for peak usage patterns (given the two standard deviations that were considered). The diagram below shows output from SAS Ecosystem Diagnostics for CPU utilization, which is used to generate a customized analysis for the client.

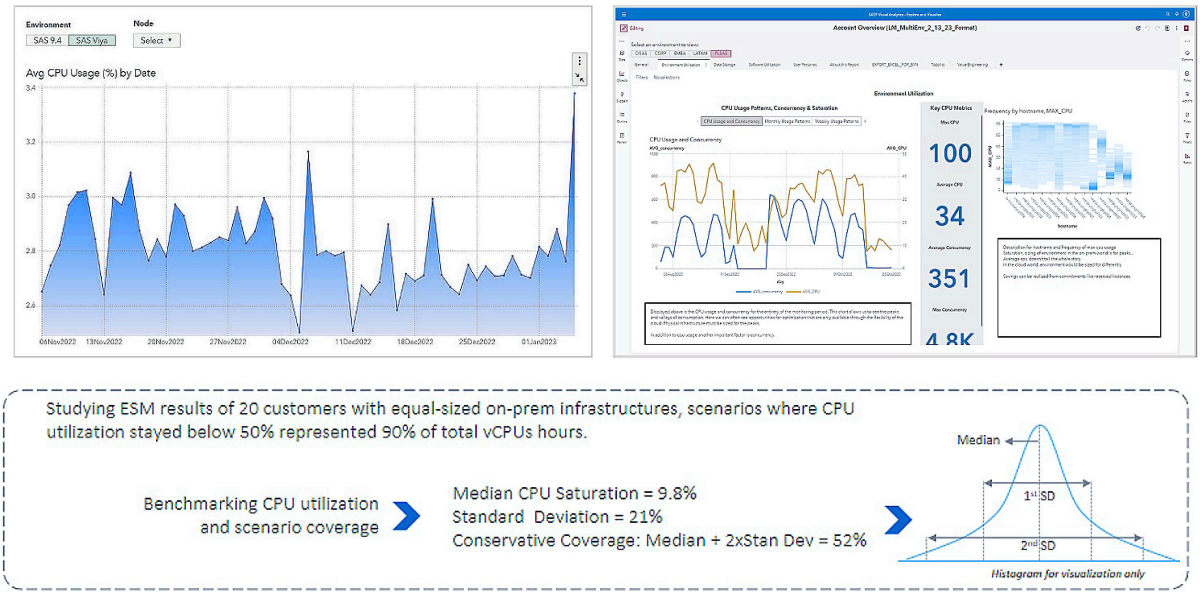


Figure 3: CPU utilization snapshots from Ecosystem Diagnostics used to create customized analyses.

Based on this analysis of Ecosystem Diagnostics data (and associated benchmarks), SAS can recommend a "Reserved Instance" or "Savings Plan" – based cloud pricing that accommodates the client's average cloud usage and recommends additional capacity as "pay-as-you-go" to account for demand spikes.

#3: Auto-scaling and workload management provide additional optimization.

The Viya auto-scaling capability works with Kubernetes to automatically add nodes and queue jobs when waiting on a node and scale down nodes when they are idle. This not only allows SAS Viya users to seamlessly scale up for higher demand periods (such as nightly batch jobs and month-, quarter-, year-end jobs), but also allows them to select the node type (or machine type) based on workload needs. This capability is enabled by an external-only PostgreSQL, which has no performance impact.

Users can also select the node type (or machine type) appropriate for different types of demand. For heavy demand periods, they can spin up smaller 4 vCPU Compute instances based on their requirements and spin them down once the task is completed. In addition,

they can label and associate different queues to different node types based on the nature of a given workload, as well as use a max node value to limit how many nodes can be deployed concurrently or how quickly the nodes can be scaled up. These limits keep a check on unnecessary cloud costs.

For example, SAS recently conducted a benchmarking "day-in-the-life study" for a health care organization running a combination of 3,190 long ETL workloads and short analysis reporting jobs. The workload used in this scenario replicated real-world data complexity and volumes and used Standard_D4s_v5 nodes both in the fixed node scenario (12 nodes running 24x7) and the auto-scale scenario (scaling from 1 – 12 nodes based on jobs available in the queue to process). The study's results show that auto-scaling the compute node pool provides **53% cost savings** per weekday over the fixed compute node deployment. Over a week's time using this same scenario during weekdays – specifically, one auto-scale compute node over the weekend and 12 fixed nodes 24x7 – the difference in cost compared to fixed nodes showed a **64% cost reduction**.

The auto-scale capability, which is enabled by SAS Workload Manager, allows you to test various pricing strategies for your workloads and determine the most cost-efficient cloud pricing strategy. For example, in general:

- **For customers that run Viya up to 12 – 16 hours per day** – it is best to start with a basic compute node (as a savings plan) and then scale using it as a pay-as-you-go strategy. Customers have reported up to 46% savings using this pricing strategy.
- **For customers that run Viya 12 – 24 hours per day** – it is recommended to have a three-year savings plan (i.e., a pricing structure provided by a cloud provider with long-term discounts built in) for 12 – 24 usage hours per day; customers have reported 49% savings using this approach. Alternatively, a one-year savings plan for 22 hours per day has resulted in up to 23% in savings over a pay-as-you-go strategy.
- **For customers that run SAS 12 – 22 hours per day and do not want to opt for a savings plan** – they should consider choosing a pay-as-you-go option, which customers report can result in 13% – 36% savings.

Without the auto-scaling functionality of SAS Workload Manager, you would have to manually prepare compute instances to scale manually. These instances couldn't be launched by tracking workloads, so this effort would be both effort-intensive and cost-inefficient. But with SAS Workload Manager, you can choose the best pricing model based on actual usage and ensure that the Compute nodes are launched on demand for true cost optimization.

In addition to enabling auto-scaling, SAS Workload Manager ensures optimal infrastructure spend by reducing job overflows, poor infrastructure utilization and missed jobs that can cause disruption and result in penalties. In a separate SAS benchmarking study, priority queues were used to allocate 60 batch jobs (20 high priority, 10 medium priority and 30 low priority) across five Compute nodes (8 cores; 64 GB RAM), mirroring a real production scenario. The goal was to test how the priority-configured queues of SAS Workload Manager impact performance. As shown in the following table scenarios 1 – 4, using SAS Workload Manager priority queues (compared to the default scenario 5) results in a significant performance improvement. Additionally, a constrained scenario using low maximum concurrent jobs was tested (scenario 6), which showed a longer time to complete compared to all other scenarios.

Table 1. Performance impacts of various default and priority cue scenarios.

Scenario	Duration (mm:ss)	Condition	Remarks
1	12:24	Submit the 60 batch jobs in the priority order (20-high, 10-medium, 20-low).	This is the best scenario. Prioritizing jobs can have a significant impact on performance.
2	13:02	While the 60 batch jobs are running in priority order, additional concurrent SAS Studio or SAS Model Studio sessions are also executed by users in an ad hoc manner.	Ad hoc users have a low impact on prioritized jobs.
3	13:13	Submit 60 batch jobs all into the default queue with 40 max concurrent jobs to reduce pagination or overloading of nodes.	Setting an optimal concurrency limit has no major impact on performance.
4	13:14	Preempt some lower-priority jobs to make room for higher-priority jobs. SAS Workload Manager decides which jobs to run next from the pending queue based on priority.	Putting higher-priority jobs to the front of the execution queue and resuming lower-priority jobs later has some impact on performance.
5	13:33	Submit 60 batch jobs all into the default queue with no priority queues and no max concurrent job limit.	This is the baseline scenario where the default queue is used, with no priority or concurrent job limits. This works reasonably well, but it is not optimized.
6	28:07	Submit 60 batch jobs all into the default queue with a 10 max concurrent job limit.	Setting a non-optimal concurrency limit has a major impact on performance and is not recommended.

Assuming a standard D4s 8 vCPU node, the compute-only cost for five nodes works out to $2,250 per month. This study shows that there is an approximate 8% gain for scenario 1 (the best) compared to scenario 5 (the default). So, if the nodes are utilized, for instance, for three months in a year, the result is $540 in savings ($2,250x 3 x 8%) due to the efficiency gains enabled by SAS Workload Manager. Most importantly, SAS Workload Manager allows you to set and test these parameters to arrive at the most optimized scenarios that work for different workloads and different time periods.

LOOKING AHEAD

Scaling in Viya is extended to the CAS (Cloud Analytic Services) nodes as well, although this will currently have to be done manually by the SAS admin. This is important because CAS nodes had the largest cost in the proposed baseline cloud architecture. By starting small and scaling CAS nodes only when required, the potential to reduce cloud costs is significant. The reduction is based on the following factors that SAS Ecosystems Diagnostics can capture to help recommend the best strategy:

- The type of pricing agreements you have in place with cloud providers.
- How smartly you can scale up and scale down CAS nodes at appropriate times of day.
- Your ability to identify which workloads will benefit from CAS.

Key takeaways

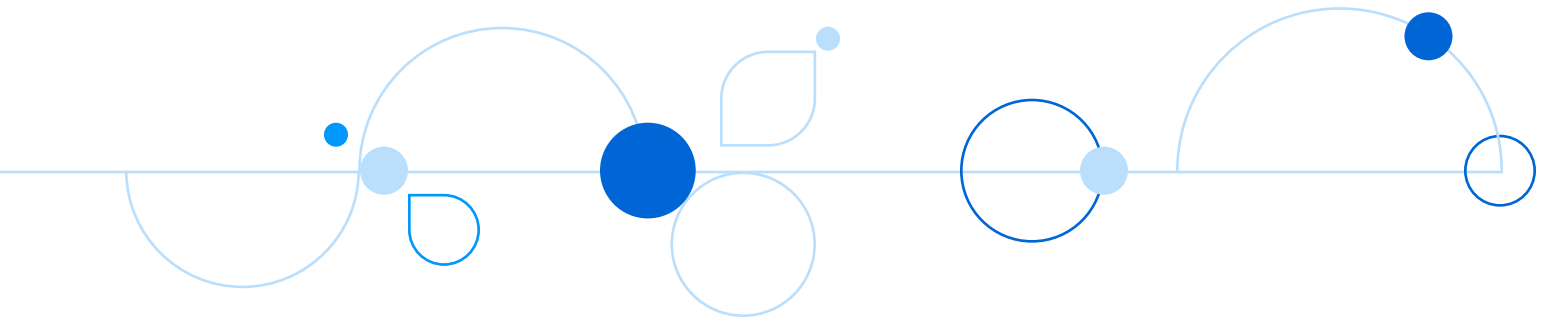
With high and unpredictable cloud spending becoming a major issue for software-as-a-service (SaaS) customers, it's more important than ever to understand the drivers of these costs and use analytical tools to determine – and optimize – your total cloud spending. As explored in this paper, the new Ecosystem Diagnostics utilities of Viya provide visibility into key metrics so you can understand bottlenecks and rightsize your future-state infrastructure.

In addition, with its new baseline sizing recommendations, the auto-scaling capabilities and workload management features of Viya give your IT admins and users powerful controls to optimize cloud spending. The result is a scalable and cost-optimized deployment that was not previously available for SAS workloads.

The results are real and significant. In a separate study , Viya is shown to be 30 times faster and 86% more cost-effective than commercial and open source alternatives.[1] This high performance significantly reduces the time for analytical workloads to run on costly cloud infrastructure. The speed of Viya, together with the platform's rightsizing, auto-scaling and workload management features, unlocks even greater spend optimization potential.

To learn more, visit **sas.com/cloud**.

[1] **Faster AI & Analytics: SAS Viya Outperforms the Competition - The Futurum Group**



Learn more about how SAS can help with optimizing your cloud costs.