Get more value from your analytics fast

Optimized AI and analytics in the cloud

In the cloud computing world, performance is literally money: cost control, revenue and time to value. Establishing the most performant foundation for your organization is critical to meeting current needs and preparing for the future.

The research presented in this paper shows that SAS® Viya® is a more cost-performant and cost-conscious analytics platform available to enterprise cloud users. That translates directly into a productive workforce and powerful analytics.

The conventional computing wisdom is that performance equals time. The ideal performance state is using your resources to their peak without affecting runtime.

Can all your CPUs immediately run at full capacity to respond to demand? This requires perfect harmony between the math, the data and the chip to parallelize CPU usage for the shortest amount of time. This is a theoretical ideal because it's impossible to achieve perfection in practice. But moving those three elements as close to perfection as possible is the goal for every workload.

On the right, you'll see an overly simplified view of the results. In **Figure 1**, runtime is cut in half when cores are doubled. In **Figure 2**, CPU is used to maximum capacity while the workload is executing. In a cloud environment, an analytics platform must operate within these extremes to be as cost-performant as possible.

Figure 1: Runtime and cores provisioned

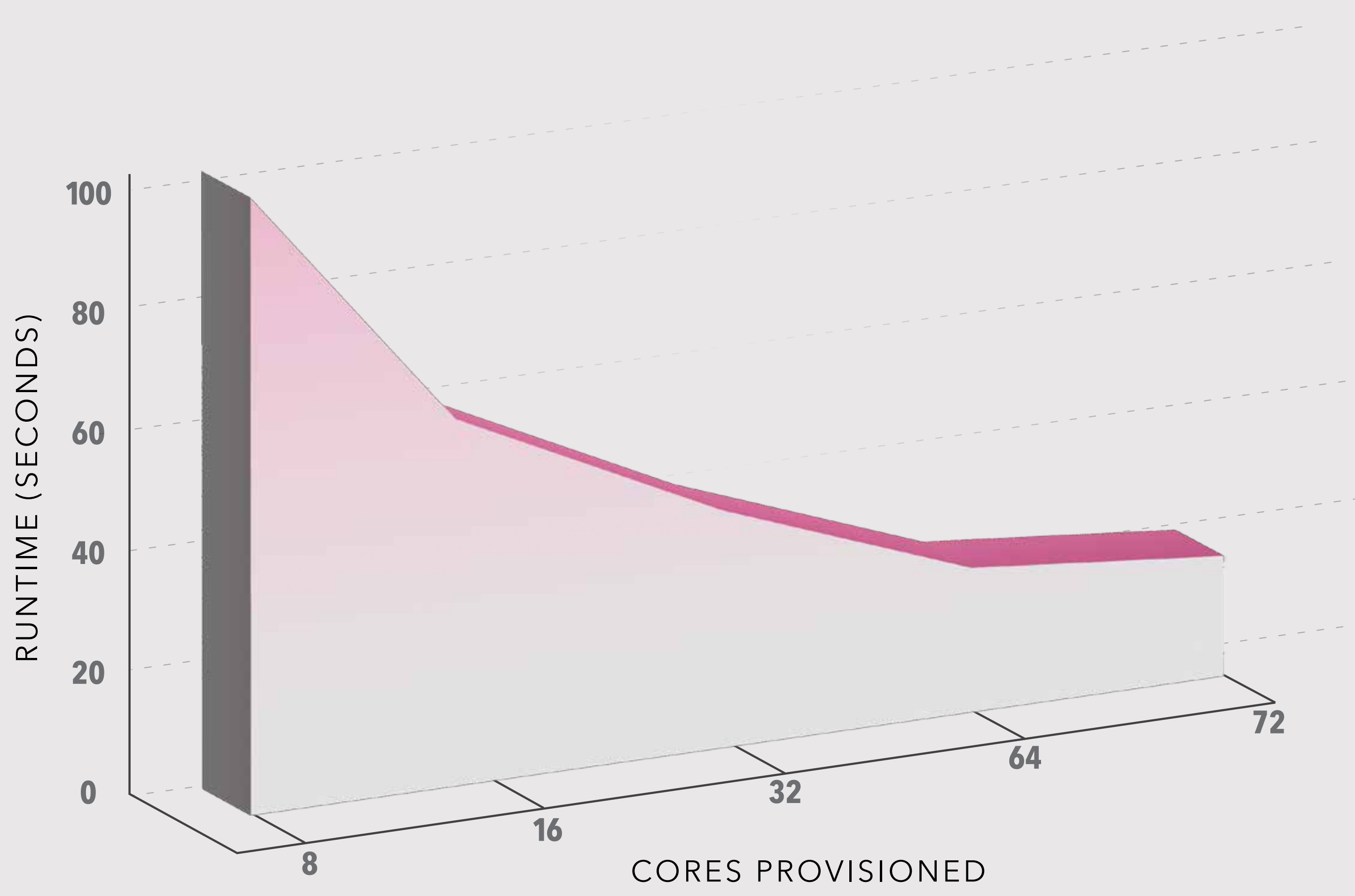
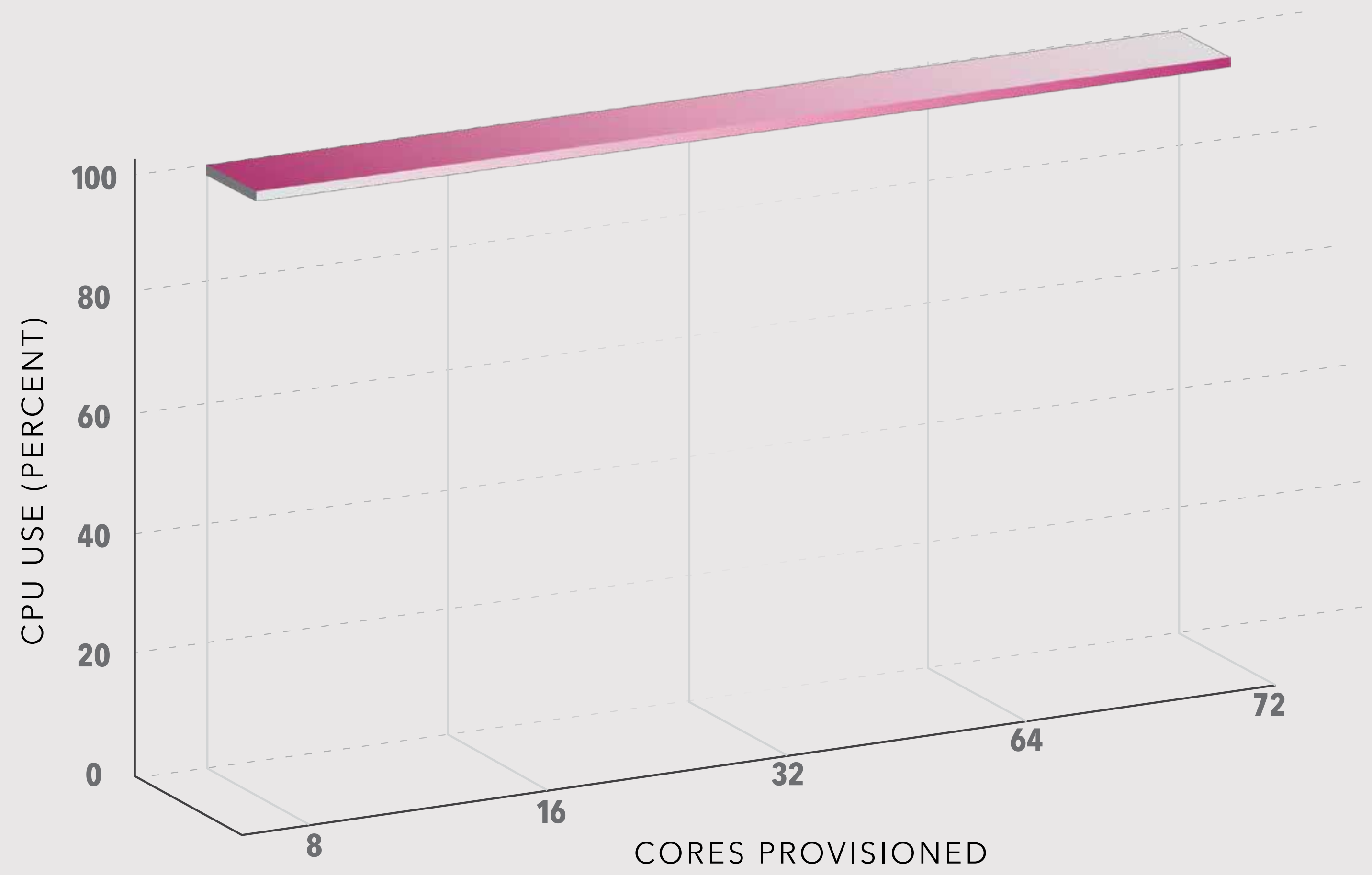


Figure 2: CPU use, with cores provisioned



The top reasons companies back away from the cloud

Falling far short of this ideal state is a key reason that companies abandon cloud strategies they pursued without fully understanding the costs of executing their workloads in the cloud.

Industry analyst IDC states that 80% of public cloud customers have repatriated some workloads off-cloud. Security issues are the leading driver (19%), but close behind are:

PERFORMANCE (14%) **COST (12%)** **CONTROL (12%)**

Why? Organizations without a cloud optimization process tend to overspend by 40%, due to unmanaged costs, unexpected usage, suboptimal design and implementation, and wrong-sized production and waste in development and testing environments, among other reasons.

SAS can help you avoid this scenario by accurately scoping your analytics transformation before implementing your cloud migration strategy or supplementing the cloud strategy already underway.

Applying ecosystem diagnostics enables you to see which workloads consume the most resources and runtime; generate cloud environment and workload sizing; and determine business justifications for cloud and data strategies that consider interdependencies between data, locations, routines and users.

Having detailed knowledge of your consumption metrics provides two potent insights. The first is a reasonable forecast of cloud cost and performance expectations and where they can be improved. The second is creating a validated map of which workloads and data benefit from a cloud environment (and which should not migrate at all).

Cloud spending is growing year over year and becoming a more substantial portion of a company's costs. As a result, infrastructure expenses are a line item of growing importance to senior executives. Estimates are that 80% of organizations will overshoot their budgets for infrastructure as a service (IaaS) as a direct result of poor cloud optimization governance and overspending in advance on cloud commitments that are too small to support their analytic ecosystem's consumption.

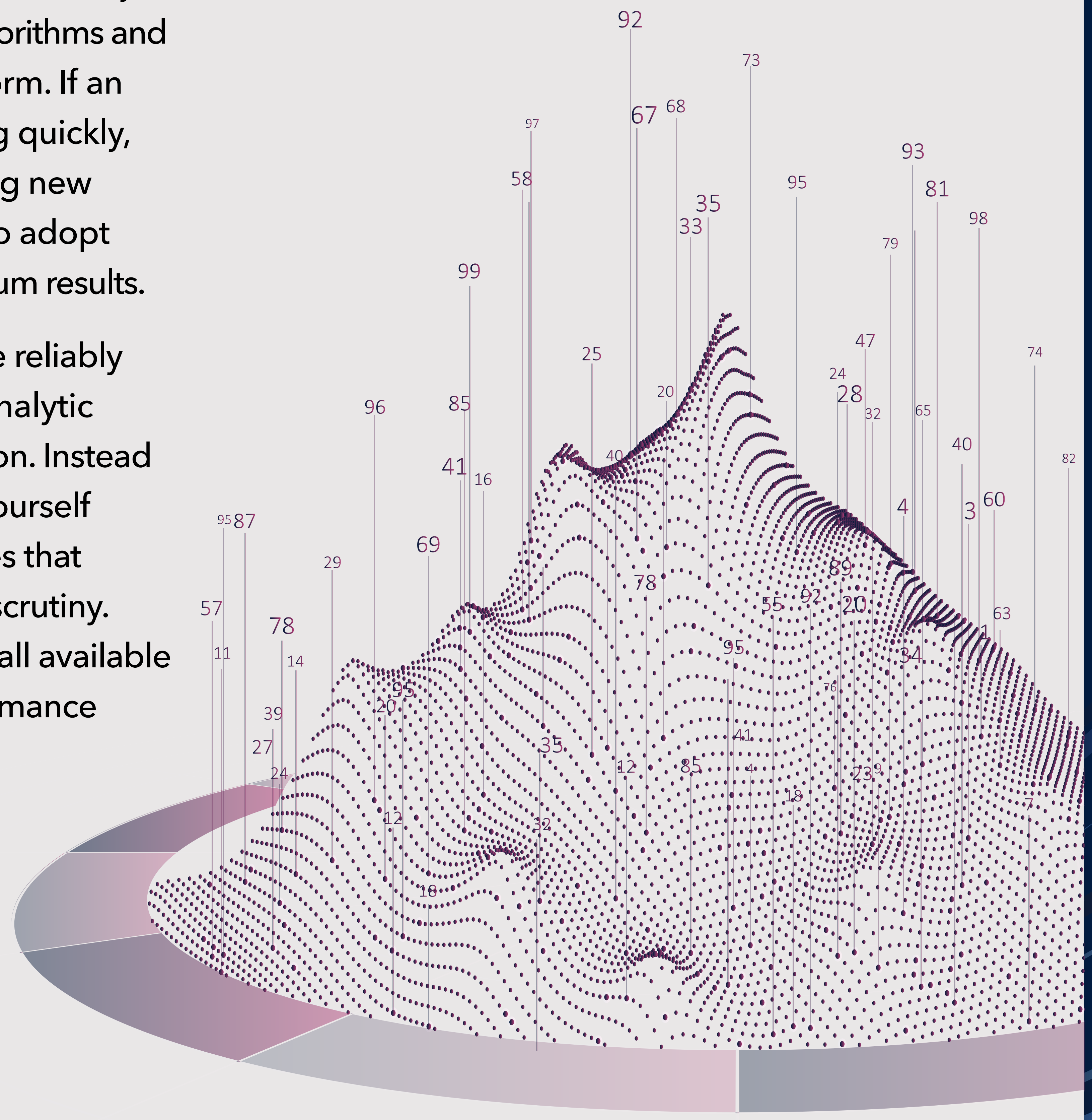
Cloud spending is growing year over year and becoming a more substantial portion of a company's costs. As a result, infrastructure expenses are a line item of growing importance to senior executives.

Finding cost-performant harmony despite ever-changing requirements

Computations change as constraints are introduced. The data which feeds the math is constrained by the chip and the distribution of the math. Changes to the data and how that data is distributed to the chip can alter performance, cost and accuracy.

SAS Viya helps you move toward the perfect harmony of math, data and chip hardware in several ways:

- SAS has been proven to be a highly performant language. Our algorithms optimize CPU usage, runtime, memory usage and I/O speed, often eliminating the need to purchase additional infrastructure to compensate for inefficient CPU usage or high memory requirements – frequent issues when running large jobs that translate to unnecessary additional expense. In this way, the cost of compute is married to performance for the best results.
- The SAS platform is inclusive; it embraces and improves open source algorithms. Many data scientists who write in Python, R or other languages can readily take advantage of the stability, performance, affordability and model governance of the SAS platform. You can also use its ability to run any new algorithms as part of existing model tournaments. SAS actively identifies leading open source algorithms and incorporates them into our platform. If an algorithm is better at parallelizing quickly, maximizing core usage or offering new advances, the platform is ready to adopt and refine its approach for maximum results.
- Optimal cost performance can be reliably pursued by calling SAS Cloud Analytic Services (CAS) actions from Python. Instead of performing the optimization yourself in Python, employing CAS ensures that parallelization undergoes heavy scrutiny. SAS assists you in choosing from all available options to help ensure the performance and repeatability of the pipeline.



The SAS platform is inclusive; it embraces and improves open source algorithms. Many data scientists who write in Python, R or other languages can readily take advantage of the stability, performance, affordability and model governance of the SAS platform.

What does performance harmony look like?

Consider this comparison of workload executions for logistic regression (a common analytics algorithm) between SAS Viya and open source packages (OSPs) when monitoring technology was deployed to the sessions to observe their performance. Workload executions used the same data set, code, driver instances and worker nodes per test. SAS executions were automatically mapped to the available cores. Open source packages required manual changes to use all available cores. Each test was performed five times per package to identify anomalies. No anomalies were observed or removed. The aggregate of the five tests is graphed across the various worker nodes.

SAS Viya provides the expected outcome: Adding cores showed that SAS runtime was 12% to 700% faster than the OSPs across the worker nodes.

OSP 1, which is closest to SAS in eight cores, didn't keep up with SAS runtimes as cores were added. This leads to higher costs with no improvement in runtime (or sometimes longer runtime).

SAS Viya provides the expected outcome: Adding cores showed that SAS runtime was 12% to 700% faster than the OSPs across the worker nodes.

Figure 4 shows how well the CPU was used during the workload execution. Because SAS Viya has efficient execution times, the CPU requirements across all cores are highest for eight and 16 cores in this test. The blue line declining indicates that the workload is complete, and those cores can now be reallocated for other work as needed. The OSP packages remain high in CPU usage across worker nodes but are not vastly improving the runtimes. This is akin to driving a car in second gear without going any faster.

The net result is a significant difference in the cost of computing for the same work. Because the SAS runtime is faster and needs fewer cores to achieve those faster runtimes, you can see why SAS Viya is a cost-performant leader for analytics.

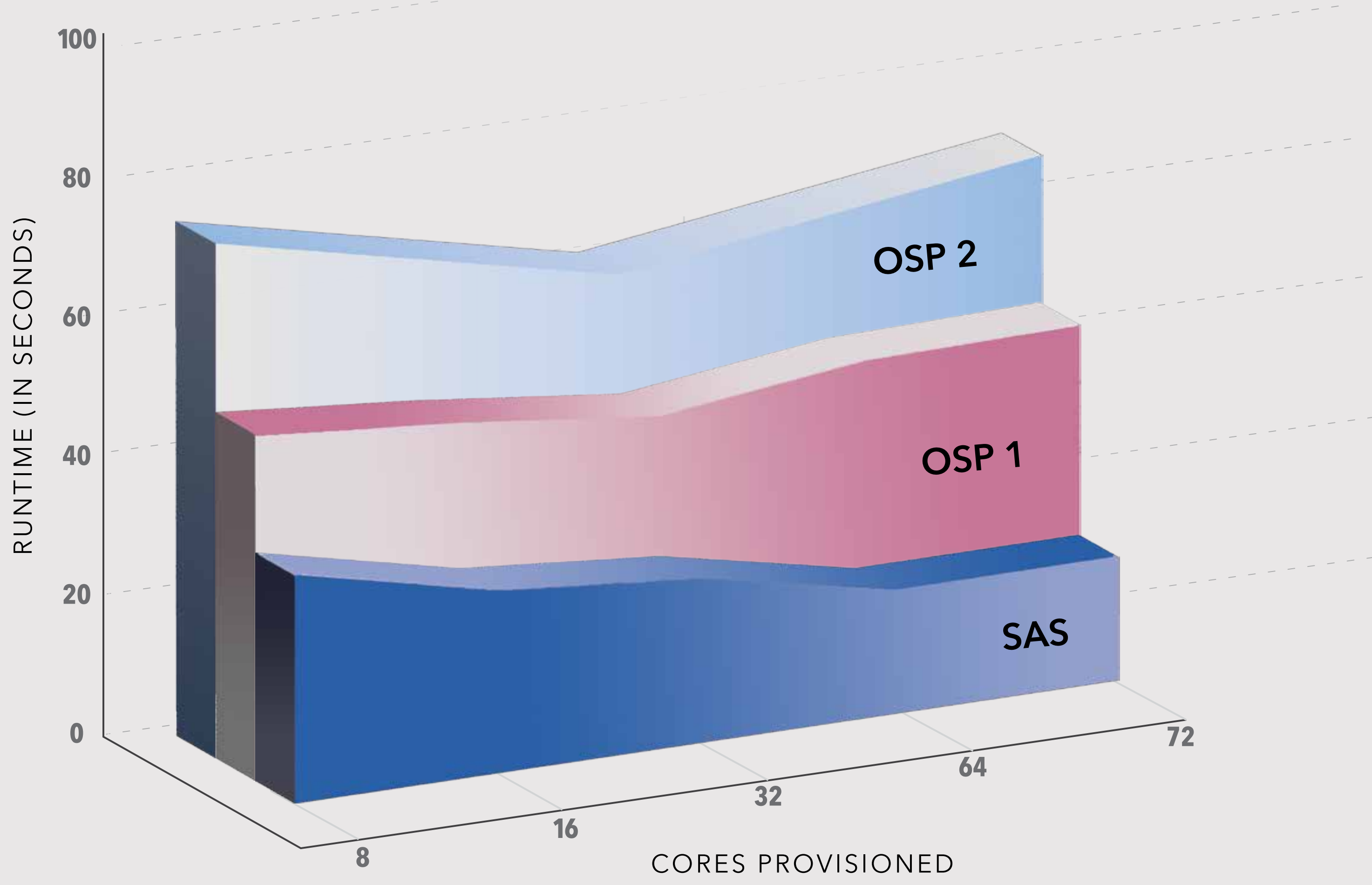


Figure 3: Runtimes vs. available cores

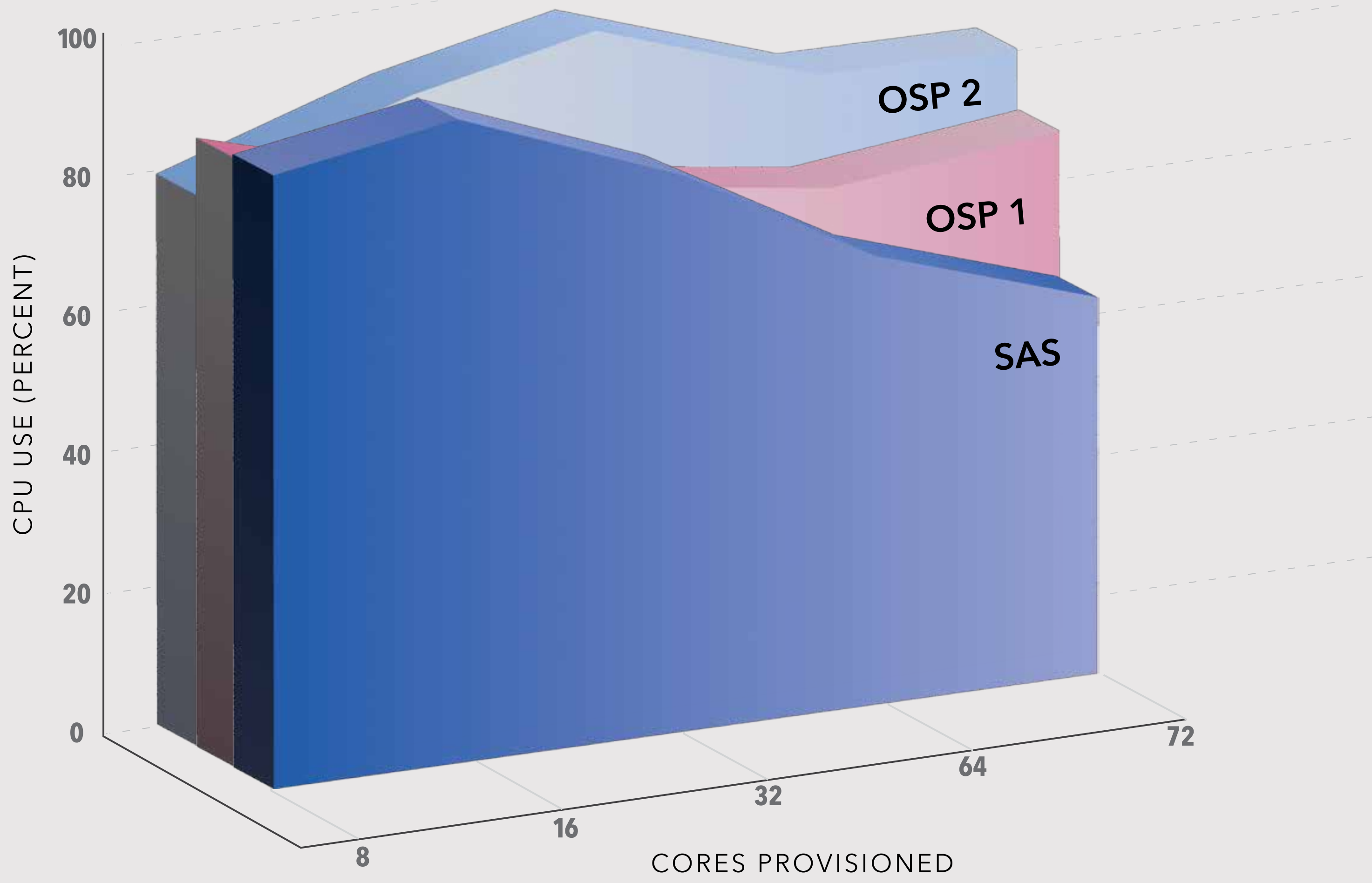


Figure 4: CPU utilization vs. available cores

GET MORE VALUE FROM YOUR ANALYTICS FAST

Optimized AI and analytics in the cloud

The development community of open source is vast, but being the best depends on providing the right assurances

The cloud and distributed computing introduce new constraints that affect performance, computing costs, and the repeatability and accuracy of results. Comparing those SAS and OSP workloads is just one example of how SAS is evaluating leading open source algorithms against SAS Analytics.

By embracing open source algorithms, SAS Viya enables your analysts to pursue the best outcome every time regardless of algorithm. However, when it comes to putting analytics into production, these assurances need to be provided:

- Effective and efficient.
- Cost-conscious.
- Repeatable.
- Accurate.
- Ethical and responsible.
- Able to generalize to different data set sizes, types and shapes while managing model drift.

A SAS algorithm is tested for these qualities. SAS also continues to test open source algorithms to evaluate their ability to provide the same assurances. If an open source algorithm delivers on most of those assurances, SAS will improve it to the point where it can be adopted on the Viya platform. This enables users of Viya algorithms to spend less time coding, testing, interpreting and trying to achieve better harmony between the math, data and chip. This frees them to focus on delivering results.

New Consideration No. 1:
It's accurate, but is it repeatable?

A growing concern about unmanaged cloud costs led to the creation of the FinOps Foundation in 2019. A cloud finance best practices hub and certification provider, it provides guidance on this new financial management discipline and guides the cultural shift precipitated by the cloud. The goal is to maximize the business value of cloud spending that, according to the FinOps Foundation, requires "engineering, finance and business teams to collaborate on data-driven spending decisions." It's a new world in which each role takes responsibility for its cloud usage, keeping cloud costs in mind and collectively making the best use of the variable cost advantages offered by cloud computing.

These financial considerations lead back to the need for analytics-driven enterprises to improve accuracy and repeatability. When the "perfect" harmony of math, data and chip is less than perfect, errors and inaccuracies proliferate. A single processor can easily generate a single average with accuracy. But with distributed data, the ability to produce averages is limited by the amount of data that can fit on a specific chip. Unless a central organizing principle is applied, the distributed averages will produce an incorrect macro average. It might be fast, but it is neither accurate nor repeatable.

Each time you distribute the math to the chips in a different way, it will produce a different answer. Python manages this variability by requiring appropriate configuration of the math to the chip. While this can be effective, it requires hand-coding by a user highly adept at the mapping of math, data and chip – reducing productivity, increasing cloud costs and putting the analytic results at risk.

SAS alleviates the user of these burdens. The platform flags that the process will produce an inaccurate answer if it is distributed, indicating that the best route would be to run a single-threaded process to execute that workload, reducing the risk of inaccuracy.

In this way, SAS Viya creates a solid foundation for the cornerstones of predictive modeling: cost performance, productivity, accuracy and repeatability. The platform helps to eliminate common risks faced by users who adopt open source to take advantage of perceived low start-up costs.

Running an off-the-shelf analytics package and execution engine while paying for bulk cloud commitments routinely costs more and can produce inferior answers. Whether migrating from traditional on-premises computing or reassessing your existing cloud economics, SAS Viya can help you balance how cloud commitments are applied to variable workloads with greater flexibility and speed. We believe this is the ideal state of analytics performance in the cloud.

The goal is to maximize the business value of cloud spending, which requires engineering, finance and business teams to collaborate on data-driven spending decisions. It's a new world in which each role takes responsibility for its cloud usage, keeping cloud costs in mind and collectively making the best use of the variable cost advantages offered by cloud computing.

New Consideration No. 2:
Does cost constrain access and innovation?

Performance is more than saving time – better performance means your organization can produce more. Unsupervised machine learning results in many "blind" models that can generate good answers, but paying for each iteration of that large project is far less than ideal. This is especially true when the technology is slow: The data scientists will limit their tournament to fewer models to optimize the cost.

In examples like the one shared in **Figure 4**, our research demonstrated that SAS can run twice as fast as leading OSPs that use nine times as much infrastructure. This enables data scientists to run more models without driving up costs.

In short, bigger tournaments equal better answers. But the cloud cost structure requires an additional framework that has not been a consideration in prior computing ecosystems: achieving the desired goals while ensuring value within cost constraints.

Unsupervised learning and modeling time happen while the meter is running in the cloud. This offers a new dynamic for users who are now responsible for the cost of their analytic design time.

For example, what if you need to run a 200-model tournament but have a $100 budget? How can you fit that tournament into your budget while still generating the best answer? How can you maximize your analytic design time to build the most models without going over budget?

The same considerations apply to production runtime in addition to design time. If you want to put your model into production to run multiple times a day, how do you stay within your $100 budget? Run it fewer times to stay within that constraint? Make the model more efficient to run it more often?

And if you need to build 200 models, how quickly can you build meaningful models? Building a good model is time-intensive. Achieving the cost-performant harmony between math, data and chip requires additional coding overhead.

Furthermore, citizen data scientists and business analysts need to be enabled to perform higher-level analytics without being overburdened by cost constraints or writing a large

Performance is more than saving time – better performance means your organization can produce more.

amount of code. This is why no-/low-code interfaces with SAS algorithms can be incredibly powerful. They avoid the coding and overhead. They lower time to code models, offer approachability to new users and eliminate the need to ensure the data, math algorithm and chip are as close as possible to perfect harmony.

This new dynamic highlights the importance of cost performance and why SAS is highly attuned to this issue. In the on-premises world, constraints are addressed by buying more cores and capitalizing the expense, a long-term commitment with predictable IT spending. In the cloud world, IT spending happens at the edge, with resources scaled according to need. This creates challenges for forecasting IT spending and hampers the organization's ability to benefit from heavily discounted long-term cloud commitments. This change in buying behavior increases the risk of overpaying for cloud computing.

The other element in this dynamic is that the cost decisions are now in the hands of the teams running the models – not with the IT or purchasing departments. In the cloud, system engineers procure the infrastructure, choose the algorithm, select the amount of data – and now carry the fiscal responsibility when they run their workloads.

As a result, organizations are trying to find FinOps unicorns to guide executives, finance, system engineers and product owners toward more fiscally responsible engineering in the cloud.

Analytics users want to spend their time effectively to find better business solutions. Instead, they're spending time writing code, ensuring it is tailored to be cost-performant, or finding themselves limited to building and testing fewer algorithms and models to stay within cost constraints. Some users may even be excluded entirely from pursuing analytics due to these new considerations.

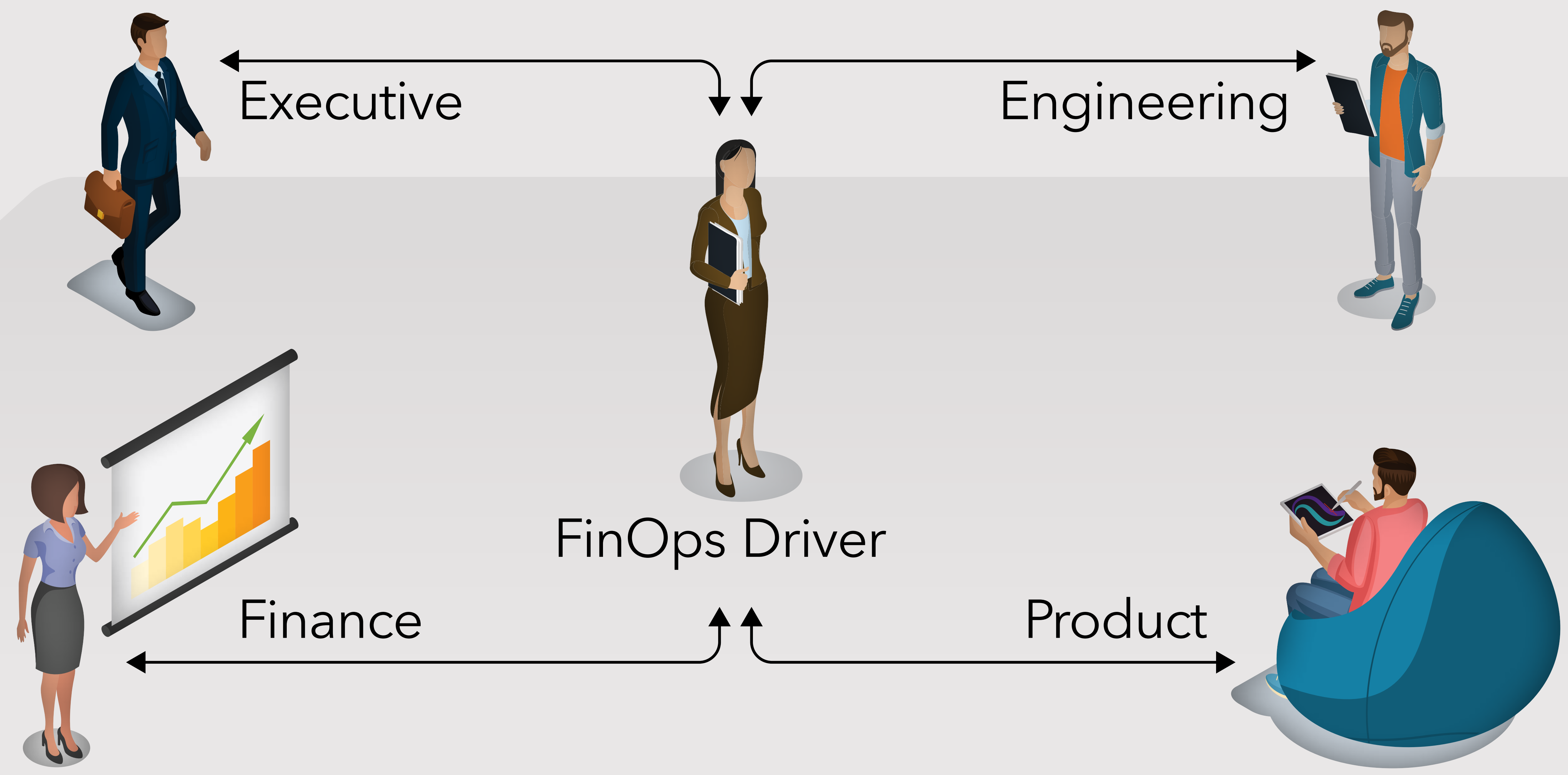


Figure 5: A FinOps unicorn coordinates cloud finance between stakeholders
(Source: FinOps Foundation)

A cost-performant approach

Becoming cost-performant buys the freedom to do more. This influences the outcomes you can achieve. Improvement in runtime means you are producing more outcomes instead of a single outcome.

SAS proposes a cost-performant approach:

1. **Be more flexible and innovative.** Put the data in the user's hand, and they will manage their behavior to squeeze more miles per gallon (aka the Prius effect). The cultural change of cloud financial management needs to happen in analytics-driven organizations, regardless of programming language preference, in order to mitigate runaway cost overages, and ultimately stem the tide of needless repatriation of workloads from the cloud. SAS offers a solution of ecosystem diagnostics, which does just this by identifying code that is eating resources and restructuring how it executes to improve its cost performance.
2. **Convert an ungoverned cost into a measured consideration.** Our research showed that two identical approaches can have vastly different outcomes. Being able to easily assess performance allows cloud costs to be brought into line with expectations. Tight harmony between math, data and chip allows algorithms to run faster and reduces cost considerations by several orders of magnitude.

The shift to cloud has increased the potential to mine untapped value from data, but also the risk of driving costs into uncharted territory. Changing the behavior of data engineers and the code they run is a new challenge and an evolving process. The potential to get 12% to 700% more value out of your cost equates to a continued, compounding return on your investment: Users can access more runtime by computing for less, devoting more computing power to the most pertinent problems and generating better results. What incremental value could you drive by running more analytics at the cost you pay today?

Want to know how you can improve your cost performance? Click here to learn more.

The potential to get 12% to 700% more value out of your cost equates to a continued, compounding return on your investment: Users can access more runtime by computing for less, devoting more computing power to the most pertinent problems and generating better results.

To contact your local SAS office, please visit sas.com/offices