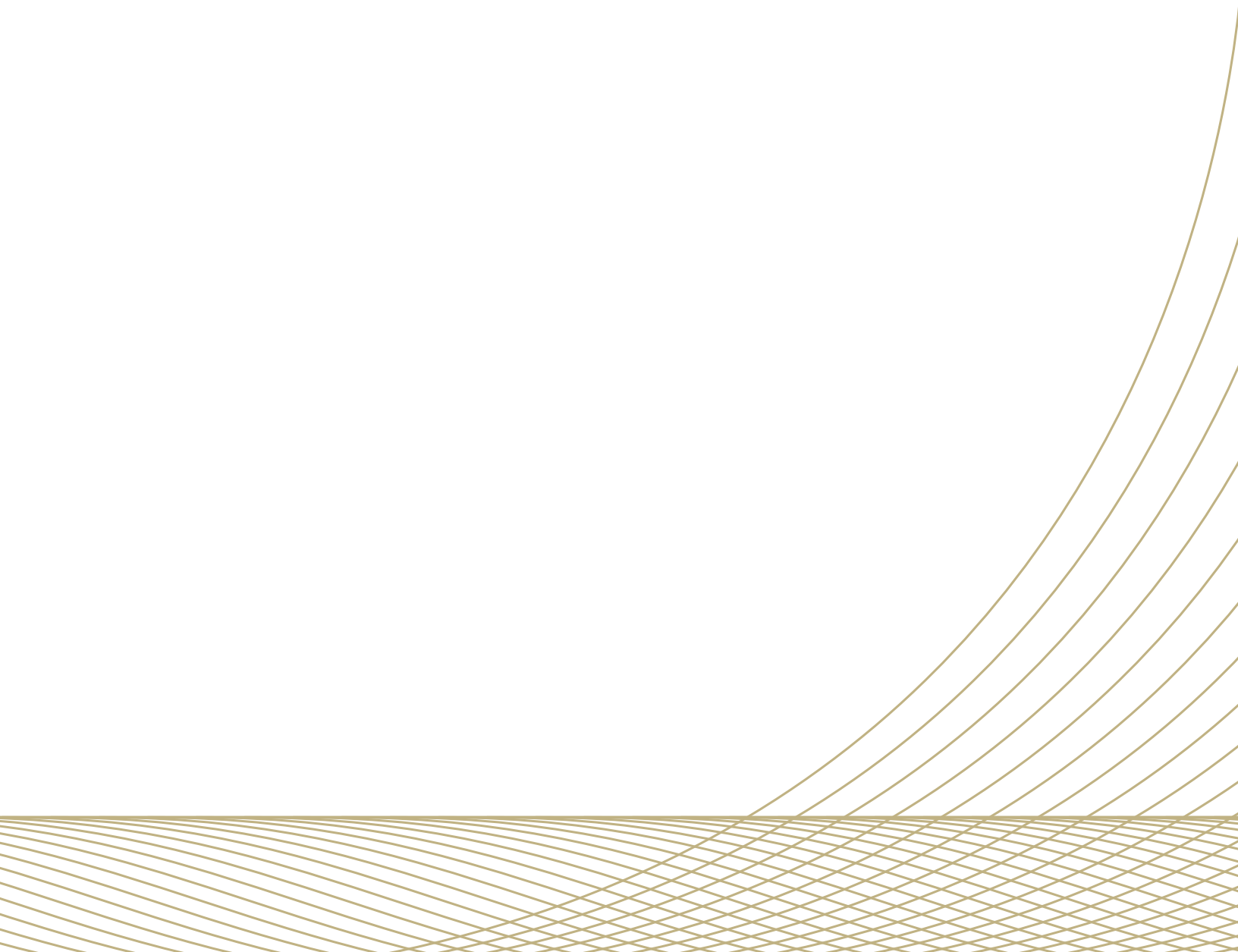
Skytap

Preventing Cloud Sprawl Using Transparency and Control



INTRODUCTION:

Public cloud adoption is in hypergrowth mode and, like any major technological transformation, the massive shift to a new model introduces challenges while delivering previously unachievable benefits. For enterprises embracing cloud-first strategies and the leaders who are tasked with executing those strategies, achieving the same visibility and control over cloud resources and spend that existed on-premises is a major barrier to a successful cloud implementation.

Governance over cloud solutions is one of the top three concerns among CIOs moving to the cloud (2016 Harvey Nash/KPMG CIO survey). Forrester research supports this finding, reporting that the complexity of managing the cloud and lack of monitoring tools and services is one of the top five concerns for enterprises that are implementing cloud services (Business Technographics Survey, Q3 2016).

In this white paper, we will discuss why it is so important for enterprises to ensure transparency and control of their cloud usage, the challenges they often face, and how to overcome them.

THE UNFULFILLED PROMISE OF CLOUD INFRASTRUCTURE: WELCOME TO CLOUD SPRAWL

Many companies take up cloud initiatives expecting to replace CapEx with OpEx. Only paying for what you use sounds simple and implies "no waste" and "cheaper," but the reality is often quite the opposite. Organizations can end up with bills larger than expected, because they don't fully understand the cloud purchasing model.

Traditional enterprise IT purchasing model

The traditional IT department purchasing model is generally a centralized process involving few people in charge of a fixed budget. For example, the IT infrastructure manager is given a budget for the coming year, and plans ahead for new capital equipment purchases, software upgrades, and other costs. While nothing ever goes exactly to plan, this model maximizes predictability and minimizes unexpected expenditures.

Cloud purchasing model

In contrast, purchasing cloud capacity and services is decentralized. Anyone with access to their organization's cloud account or a credit card can spin up services. For example, a developer setting up a test environment or a sales engineer prepping for a demo might use the same account. Easy access to cloud resources makes their jobs easier, but at the end of the month, the organization

can receive an unexpectedly large bill and not know exactly who used which services for what purpose.

The problem is compounded when organizations encourage infrastructure teams to rapidly scale up or grow cloud adoption with the expectation of getting bulk-pricing discounts. While some bulk discounts may apply, they forget that cloud pricing is a la carte. While you may save on infrastructure and capacity costs, you will start paying extra for things like data transfer fees, PaaS, and other routine operational costs like monitoring and alerting.

As cloud adoption grows, the billing becomes increasingly complex and difficult to track. IT, Finance, and Accounting managers receive statements lacking the detail necessary to map cloud expenses to specific projects and divisions. The ever-increasing costs along with lack of transparency and control is frustrating for everyone involved. We call this "cloud sprawl."

	Enterprise Purchasing Model	Cloud Purchasing Model
Buyers	Centralized, single or few buyers	Decentralized, anyone in the company can buy
Buying timing	Buy capacity first, then use it	Use capacity and services first, then pay later (monthly payment)
Budget	Fixed, documented budget with some wiggle room	Often no budget or hard to enforce budget
Tracking	Easy to map purchases to cost centers and projects	Difficult or impossible to map purchases to cost centers and projects
Visibility	Top line procurement metrics are easily measured and reported	Cloud can offer more detail data on consumption, but makes it more complex to understand and control

To better understand how companies get lost in cloud sprawl, we need to take a closer look at the distinctions between cloud infrastructure consumption and traditional models:.

- Self-Service Nature One of the biggest reasons people move to the cloud is for self-service provisioning of infrastructure and resources. Teams that previously waited days for hardware and environment stacks can now provision their resources in hours resulting in streamlined application development lifecycles and faster releases. A dream situation for development teams, however, results in a nightmare scenario for enterprise finance and accounting

departments. What used to be a centralized, planned, well-tracked, and monitored purchasing process managed by a few is now a de-centralized, unplanned, out-of-sight, and opaque situation operated by tens or hundreds. But, the efficiency and acceleration gains resulting from self-service, on-demand infrastructure are too great to ignore; there is simply no going back. Instead, businesses must evolve how they control and optimize usage in this new world.

- Performance Doubts Result in Overprovisioning--Fearing that cloud resources will perform inconsistently is often cited as a top reason why IT managers overprovision cloud infrastructure, according to TechTarget and Cloudyn. In general, any seasoned operations professional will opt for overprovisioning to some extent for peace of mind. In cloud, overprovisioning takes on a new meaning and becomes a "wastage" problem. Elastic or auto-scaling, as offered by many cloud providers, is touted as the solution, but humans still create the rules that auto-scaling mechanisms must follow, and the people making the auto-scaling rules rarely have the experience necessary to optimize the system. Auto-scaling is neither easy nor automatic to set up...there is no magic "auto-elastic" button. Instead, each organization must monitor demand going up and down and determine its own rules. Through trial and error, IT organizations can figure out how to be elastic, but this takes time, especially for traditional applications that are not designed to auto-scale. Unless you are rewriting or refactoring your application to take advantage of auto-scaling, overprovisioning is likely and you'll spend more than is necessary to run the application.

- Scale Leads to More Complexity and Cost, Not Less--This is the most counter-intuitive part of cloud adoption. According to RightScale's 2016 "State of the Cloud" report, the major benefits from cloud adoption result from faster access to infrastructure, scalability, availability, and faster time to market — not cheaper infrastructure costs. Despite fierce price competition among cloud service providers, cloud costs seem high to enterprises, due to directly incurred "new" charges that were never separate line items in the enterprise datacenter pricing model. For example, as an enterprise scales up cloud infrastructure, it also increases other associated line items, like charges for data transfer or higher disk IOPS; associated service charges, such as alerts and monitoring, queue, and messaging; and new costs like training employees on cloud technologies and infrastructure management. Furthermore, migration costs often leave enterprises stuck with their existing datacenters until they can move all applications to the cloud, which can take years. The result?

Increased scale actually can drive costs up rather than down — at least in the near term. Many enterprises get caught off guard and experience rising expenses without proper visibility and control of their cloud infrastructure.

- Predictability Versus Uncertainty--In traditional, pre-cloud enterprises, infrastructure cost was predictable. Honed over years, these purchasing and provisioning processes were transparent and controlled, providing predictability, but at the expense of efficiency. The new cloud world flips the tables—offering faster provisioning and software delivery but without good visibility and control. Now, enterprises struggle to manage the unpredictable costs for cloud services. Even worse, the lack of visibility into what's driving the increased cloud spend makes it nearly impossible to put an effective management strategy in place. l

- Control and Waste Reduction are Not Top Priorities--Since Cloud is a relatively new industry, cloud solution providers are still trying to innovate at blistering speeds. They tend to invest heavily in developing core functionality and exciting new features to attract new customers, leaving less sexy (yet critical) visibility and control features off the priority list. As a result, enterprises often lament that they don't receive enough data and analytics to understand their spend. They often try to purchase 3rd-party cloud spend management solutions to ease their pain, but that adds extra cost to an already increasing cloud budget, and often fails to meet their unique needs.

BUILDING BLOCKS FOR PREVENTING CLOUD SPRAWL

The art of cloud spend management is balancing agility and control. Enterprises need to enable speed and flexibility while effectively managing infrastructure adoption and the associated costs.

To get started, think about laying the building blocks:

1 Monitoring and Measuring

Managing your cloud spend starts with efficient cost and usage monitoring. As Peter Drucker said, "what you can't measure, you can't manage." You have to start measuring things before you can make intelligent decisions on how to control or optimize them. In addition to measuring your cloud usage and costs, you also need to identify and measure what drives those metrics:

- Timeliness — You need to know what you are spending right now. Cloud spend, by its "as a service" nature, is often a "use now, pay later" model. This makes it unpredictable, especially given how decentralized purchasing quickly can become across the company. As a result, you need to be able to access the spend/usage data in near real-time. But, unfortunately, many cloud providers don't supply the data in real-time (e.g. AWS updates the Cost and Usage reports files only once a day). This can create a problem for enterprises that are accustomed to planning their budgeting and spending cycles in advance.

- Categorization — One of the best tools to understand cost and usage drivers is categorization. Just knowing how much you have spent for compute versus storage, or for large servers versus small servers is not enough. You need to know which people and departments are incurring the usage, which projects or product teams are using your infrastructure, or which cost centers or internal initiatives are consuming at what rate. Once you establish a good categorization scheme, it's easy to catch rogue spending fast, because each line item has its proper tag. If something is left untagged, it likely falls into a new category, or it is rogue spending. In either case, any given line item should be only temporarily uncategorized, awaiting proper tagging. Unfortunately, some categorization tools offered by cloud platforms are inadequate. For example, AWS and Azure tags are completely free-form, so organizations are unable to mandate or enforce usage. As a result, you may end up with usage that is not categorized or is mis-categorized.

- Visualization — Once you have real-time data in clean, meaningful categories, you can start visualizing it to understand inconsistencies and trends. Visualization tools help turn large data sets into actionable business insights.

- Instant, automated feedback — While visualization tools provide easily consumable visual data, as usage grows, manually looking at updated charts can be time consuming. To manage this problem, once you have analyzed the patterns, you can set up feedback loops,

policies, quotas, and budgets. For example, when spend in a certain category reaches 80% of the threshold as determined by a custom rule, stakeholders are automatically notified. Automating tasks like this is critical for coping with the ever-increasing data as you scale.

2 Control

Effective activity monitoring empowers you to manage spend more efficiently.:

- Proactive control by using quotas and limits — The best way to control unpredictable spend is to plan ahead and restrict usage based on your budget. If your infrastructure is elastic, your budget should be too, but you can't afford to have costs pile up without approval.

  - Quotas: Think of quotas as hard limits. — When the quota is reached, the user or department is unable to spin up new resources. For example, you should be able to set the following quotas on compute, storage, and other services:
    - Account-wide quota
    - Department (or sub-account) quota
    - User-level quota
    - Tag/label-based quota

  - Limits: While the quotas are "hard" stops, limits are "soft," and allow bursting, but only up to a certain upper threshold. The more granular and flexible controls that you have, the more in-charge you are. This is where you make your budget truly elastic by allowing infrastructure adjustments based on need, but with controls in place.

- Reactive, assertive control through automation — Most cloud providers do not provide tools for stopping rogue usage or lost instances. However, automation can help stop waste through policies. For example, you can set up a policy that suspends a VM instance after 30 minutes of inactivity to avoid excessive charges should someone start a machine and forget to shut it down. Similar mechanisms, like enforcing a tag or label every time users of a particular group try to start an expensive instance type, will ensure that resources are only used for allowed purposes.

- Reactive soft control by notification — Notification is another tool that can be used to build awareness and reduce waste. Of course, notifications do not provide direct control to stop or suspend resources, but you can build a workflow to receive notifications (either via email or messages) and take actions using APIs.

3 Optimize

For most companies, keeping costs down is an important and obvious goal, but making the most of every dollar spent is just as critical. Ask yourself questions like: Are my resources well-utilized? Are there unused or underused resources that can be used more efficiently?

Common areas to look for optimization improvements:

- Underused resources — By looking at standard performance metrics (CPU utilization, disk I/O, etc.) over time, you can spot potential unused or underused resources. Some cloud providers provide VM-level or resource-level performance metrics, or you can install third-party software in your virtual machine to monitor these metrics.

- Runaway resources — Sometimes people spin up machines for temporary projects like testing, customer demos, or teaching students. When machines are accidentally left running, you may be able to identify them from utilization metrics, but they can be hard to spot in cloud infrastructure.

- Right-sizing — Cloud providers generally offer flexible purchasing options, letting you choose various sizes of resources (e.g. various VM sizes or storage/database blocks), performance (e.g. higher IOPS or network bandwidth), and options like reserved vs. on-demand ("buy now, use later" vs. "pay-as-you-go"). Purchasing the right mix can help optimize your spend, and also inform whether cost-saving tactics like buying in bulk are right for you. For example, you may want to purchase a subscription or reserved amount to cover your normal and expected usage levels, and then buy additional, on-demand capacity for handle random surges. Another example is resizing or scaling your resources horizontally as your demand rises and falls.

4 Plan

Budgeting and planning cycles were longer in the pre-cloud world. Many companies had annual budgets and tracked against those. Now with the agility that cloud offers, IT managers need better tools to plan budgets for shorter timelines, track against those budgets, and also have the ability to manage project-specific budgets.

- Budget and forecasting — Based on historical usage and user-defined inputs, it's possible to develop an intelligent budget for the future. For finer control, cloud providers need to provide granular budget

preparation based on resources, services, and tags or labels. These pre-defined budgets can be tracked using either soft reminders (notifications) or hard rules (action taken based on budget threshold) to give customers the ability to both plan and adhere to the plan.

- Capacity planning — Many companies that have moved to the cloud now employ a "cloud capacity planner." This person estimates and predicts peak workloads and ensures the company has purchased enough cloud resource capacity. This often means cross-region load balancing inside a single cloud or distributing loads using a multi-cloud/hybrid strategy to maintain business continuity. But whatever the approach, data and analytics drive the decision, and sophisticated users are building their own tools or using third-party tools to help them completely automate this process business continuity. But whatever the approach, data and analytics drive the decision, and sophisticated users are building their own tools or using third-party tools to help them completely automate this process.

HOW SKYTAP CLOUD REDUCES CLOUD SPRAWL

Skytap Cloud provides one of the most user-friendly IaaS offerings in the industry. It is exceptionally easy to use and manage, while offering the fine control needed to prevent cloud sprawl.

Skytap Cloud capabilities:

- Real-time visibility (Monitoring) — Unlike many cloud providers, Skytap Cloud maintains and provides customer access to their usage and audit data in near real-time (in contrast, AWS usage data is usually updated only once a day). Skytap Cloud also provides real-time visibility into the current usage of an entire account and by each region, so that customers can stay informed of their actual usage around the world and put controls whenever necessary to make sure they don't go over budget. Even better, they can monitor usage at the user and department levels providing further insights into usage patterns. Using simple dashboards and APIs, customers can also download and analyze historical data and then perform custom analysis.

- Categorizing with custom labels (Monitoring) — In Skytap Cloud, you can create your own label categories and attach them with custom values to resources. As the resources incur different usage types (e.g. compute or storage), Skytap Cloud creates usage records that are annotated with custom labels. Later, you can download usage data

using these custom categories to track and analyze your spend. Other providers have similar features, but Skytap Cloud's functionality is more powerful, because label categories are determined by admin users, which results in cleaner reporting. In other clouds, users often create duplicate categories and values, resulting in noisy reports and garbage data. In Skytap Cloud, using curated categories and auto-suggested values, you can generate and track clean reports.

- Subscription-based purchasing model (Control) — Skytap Cloud lets you buy cloud resources with a defined quota. You can buy capacity for your entire company (in terms of peak VM and storage amounts) to make sure that you are well covered for peak demand, while still setting bursting limits that allow you to surpass the quotas up to a certain limit. This unique purchasing model is useful for those who like usage predictability and control. Skytap Cloud also offers pre-paid and pay-as-you-go plans.

- Granular usage controls (Control) — Skytap Cloud lets you set up departments to organize users and mimic your organizational structure inside an account. Even better, you can create usage quotas for each department and user. When a quota is reached, that department or user is unable to run more resources. Admins can set up notifications so that users and department owners are alerted when they are approaching these limits.

- Regional quotas (Control) — In addition to granular department and user-specific quotas, you can also set region-specific quotas and bursting limits (how much more you want to consume when quotas are exceeded). These help you stay in line with your regional target spend or budget.

- Auto-suspend (Control) — In Skytap Cloud, you can set a universal rule that all VMs will be automatically suspended after a chosen period of time. These settings can also be managed at the environment level. In this case, all VMs inside the environment are in scope for the setting. Skytap Cloud customers often select a longer universal auto-suspend time (say 1 hour) and individually select shorter auto-suspend times (say 5 or 10 minutes) for individuals based on their use cases.

- Sharing portal timers (Control) — Another easy way to control resource usage in Skytap Cloud is setting times of Sharing Portals. This feature is unique to Skytap Cloud – a sharing portal provides URL-based access to one or more VMs in an environment. This lets you share a working environment with an external user for classroom

training, a sales demo, a support case, or collaboration with other developers. These portals are highly customizable, including runtime, expiration date, and even limiting access to only a specific time of day.

- Find unused environments or templates (Optimize) — Skytap Cloud also provides you with an easy way to find environments or templates that you have not used for a while. Using search filters, you can look for potentially unnecessary resources that were created x months ago and not used for the last y days and take action. This helps you free up space to run more workloads.

All these capabilities are designed to give you full control of your cloud environment while retaining cloud agility.

CONCLUSION

In a world where cloud spend is out of control for large enterprises, Skytap Cloud avoids cloud sprawl by keeping costs down while still providing you with benefits of clouds, such as agility, flexibility and scalability. Skytap Cloud is easy to use, monitor, and govern. This simplicity makes it easy to ensure resources are used wisely and efficiently and to measure the right metrics to plan for the future.

Skytap Headquarters
719 2nd Ave., Suite 800
Seattle, WA 98104
1-888-759-8278

Skytap UK
60 Cannon Street
London, EC4N 6NP
+44 20 3790 9062