capacitas®

Cloud Cost Optimisation: A more thoughtful approach

Realising the hidden opportunities of a cost-efficient organisation

By Dr Manzoor Mohammed

Executive Summary

Cloud cost optimisation – a more thoughtful approach

Cloud is a powerful technology designed to help businesses grow, take advantage of new technologies, and gain a competitive advantage. However, many businesses find managing the costs of cloud a major challenge. Cost thus becomes a stumbling block to realising the value of cloud – that is: increased agility, cost-effectiveness, stability, and high levels of performance.

Getting a handle on cloud costs can be harder than it first seems. Businesses and their technology teams are tasked with maintaining service levels and availability, whilst organisational complexity increases along with demands to deliver innovation. As a result, managing the cost of cloud encompasses much more than just the name suggests.

When seeking to optimise cloud costs, the goal is to:

1. Understand what is driving the costs.
2. Identify where immediate savings can be made.
3. Embed organisation-wide practices that promote a culture of cost efficiency.

It is the embedding of culture that proves critical to achieving long-term, sustainable gains.

In this paper I explore the question: Are you getting best value from your cloud environment? And I will set out how to become a cost-efficient organisation to boost the value you get from cloud, covering:

- The core set of principles designed to guide you through building your cost-efficient organisation.
- What you must consider to get there.
- Examples against which you can benchmark your organisation.
- The challenges inherent in cost-efficiency, and how to overcome them.

The paper is based on more than a dozen years of experience delivering major cost and performance optimisation programmes for organisations as diverse as EasyJet, the UKHSA, and Ancestry.

Key takeaways

1. Cloud optimisation does not happen overnight and needs to be effective and sustainable to have long term benefits of new technologies.
2. Sustainable change requires cultural transformation – detailed in our 7 principles of a cloud cost-efficient organisation.
3. All stakeholders need to be aligned to the mission - including technology teams, the board, external regulators, etc.
4. A cost-efficient culture improves team agility, manages risks more efficiently, and ensures preparedness for future uncertainty.
5. Data driven conversations between business, technology, and performance/cost champions can turn business transformation into reality.

INDEX

Introduction

Global spending on public cloud services is predicted to hit $678.8 billion in 2024 – an increase of 20% from 2023, according to Gartner . To put that into context, 78% of US businesses are using cloud for some activity, from hosting and infrastructure, to storage and development. In EMEA, 54% of businesses are using cloud services in some or most parts of their organisation .

The question now is no longer about whether you should adopt cloud because, as the above statistics show, it is ubiquitous in today's business environment.

In 2024, the leading topic of discussion is **whether organisations are getting best value out of their cloud environment?**

This paper aims to address this question from a cost, performance, and opportunity point of view. Managing cloud costs isn't easy. Businesses must deal with organisational complexity, maintaining availability, and service delivery.

The cloud has been designed to be cost-effective, stable, and deliver high levels of performance. But if you're not using it correctly, then is it fulfilling its potential? Is it working as hard as it should for your business?

Cost optimisation is much more than simply managing cloud costs

Cloud costs can have a significant impact on business operations, from being able to invest in new technologies to fulfil innovation potential, to negatively impacting the bottom line. Not to mention internal cost pressures and challenges from the board or CIO.

Managing cloud costs becomes increasingly challenging the larger the business is. However, it is not cost alone, or the controlling of costs, that determines whether cloud is reaching its potential. It is having a deep understanding of what drives those costs. This could be nuances of product, new technologies, performance, business needs, and many more. Understanding these elements enables you to predict and explain the value provided by the different components of cloud cost, and how additional value can be delivered.

In essence, this is cloud cost optimisation - the focus of this paper - and how it can be properly implemented across an organisation to deliver long-term, sustainable advantage.

This should be of value to all stakeholders in the cloud ecosystem – CTOs, operating partners in private equity firms, head of operations or SVP. All of these stakeholders are concerned with cost to some degree, some in the long-term, some in the short-term. However, both short and long-term need to be aligned to ensure they are not working against each other.

The challenge is how to manage cloud costs with a view to getting the most value from your cloud environment. This paper introduces a more thoughtful approach to achieving cloud cost optimisation, along with the 7 principles of a cloud cost-efficient organisation that are better suited to the new cloud technologies.

https://www.gartner.com/en/newsroom/press-releases/11-13-2023-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-679-billion-in-20240
https://www.pwc.com/us/en/tech-effect/cloud/cloud-business-survey.html

https://www.pwc.com/gx/en/issues/technology/emea-cloud-business-survey.html

Dealing with cost optimisation

Engage in insightful conversations with key stakeholders like your CFO, the board, and senior leadership. This insight and transparency builds confidence in the value technology teams deliver and their ability to optimise costs and usage through scientific and fact-based data-driven insights.

A more thoughtful approach leads to sustained cloud cost optimisation

Cloud optimisation seeks to ensure an organisation has the right cloud resources, in the right place, at the right time, for the right price to meet business demands. It involves the practice of maximising performance and scalability, while maintaining service levels and controlling costs.

At Capacitas the approach we take goes one step further. We additionally focus on the long-term nature of cost optimisation so that it is never just a once-off project, but an ongoing, embedded organisational practice that continues to deliver value.

The benefits of being a cost-optimised organisation are threefold:

Achieving more with less in the long-term
Improved service reliability thereby minimising incidents that cause downtime
The uncovering of hidden opportunities in cost optimised systems

Ultimately, sustainable predictability of cost should result in service reliability for the future.

Why is this new approach important?

The technology landscape doesn't stand still. With new technologies always on the horizon, organisations must make sure they are correctly positioned to take advantage. The key here is investment. Having the resources – both money and capacity – to invest in new technologies that are often expensive, for example, large language models (LLMs). But if cash is tied up (needlessly) in the cloud, how can you leverage this new technology and remain ahead of competitors?

The answer is not easily or not at all!

At the risk of being left behind or being forced to take drastic measures e.g. cutting capacity in an uncontrolled way that will impact operations, you need to be on top of cloud spending.

While many businesses feel they have a handle on cloud costs, as mentioned, cost optimisation is more than just controlling the costs. It's also about looking at the long term –not just today but tomorrow too.
When it comes to cloud spending, the challenge for many CTOs or SVPs is balancing need with capacity. After all, most people feel it is better to have excess capacity than risk not having enough. But is this true?

While that sentiment is understandable – no-one wants their system suffering from service issues – there are hidden risks in excess overprovisioning, especially as cloud costs are an ongoing Operational Expense (OpEx) unlike their Capital Expenditure (CapEx) on-premise counterparts.

Over-provisioning hidden risks

The hidden risks of over-provisioning cloud include:

- Reduced agility: Less engineering capacity to innovate and improve efficiency.
- Masked issues: Excess capacity hides potential service problems for later.
- Uncontrolled cuts: Sudden cost pressures lead to hasty, disruptive cost reductions.

Cost Optimisation Benchmarks

- Cloud spend < 10% of technology spend for SaaS/technology firms and < 5% for enterprises
- Average estate CPU utilisation should be > 40% with CPU utilization peaking > 90% for key systems
- Non-production cloud costs are < 15% of all cloud spend

The journey to better cost-efficiency: Why the time is now!

Cloud costs are steadily rising, regardless of whether you 'lift-and-shift' what you have into the cloud, or build cloud natively. In addition, cloud providers are looking for ways to increase prices. Google Cloud led the trend, implementing significant price increases across its services – in some instances by up to 50% .

More recently, IBM increased prices for its cloud services by up to 26% starting in January 2024 . And AWS recently started charging customers for using IPV4 addresses which were previously free. Therefore, it's not surprising that managing cloud costs is the biggest challenge for technology leaders – more so than security – according to the Flexera's 2023 State of the Cloud Report .

With the immense data the cloud provides, inefficiencies often impact service reliability. Cost and performance challenges become more frequent and more apparent when using big data loads. This is something called out in an article on Infoworld by industry expert David Linthicum. In a nutshell, Linthicum says generative AI (a focus for 2024) needs an optimised cloud platform to be successful.

Removing inefficiencies provides an opportunity to improve response times and thus the customer experience. It also builds an organisation's capabilities to take advantage of current and future technologies that benefit from being in the cloud, for example AI LLMs, and to deliver them in a predictable manner that ensures optimised costs and thus service reliability.

> *"We pushed applications and data into public clouds without refactoring or other optimization procedures. This "lift and shift and hope for the best" approach has resulted in colossal cloud bills that can't continue. Enterprises need to take steps in 2024 to fix this.*
>
> *"…Generative AI systems need an enormous number of resources compared to other stuff.*
>
> *The problem is that we can make the same mistakes we did with lift-and-shift, putting generative AI systems into production that are hugely underoptimized and cost too much. If we need to run the existing applications in the cloud and want to take advantage of generative AI as a business force multiplier,* ***optimization and planning need to be on the 2024 to-do list.****"*

Traditional responses to increasing cloud costs

Initial reactions to increasing cloud costs include:

1. Looking for better discounts (e.g. savings plans, RIs, EDP etc.)
2. Investing in cloud cost management tools (Cloudability, CloudHealth, etc.)
3. Undertaking housekeeping and hygiene activities, such as turning off idle instances, rightsizing instances or moving to the latest generation technology.

Whilst these are all beneficial, what is less obvious is that the high costs themselves are hidden opportunities to upskill, streamline processes, and improve risk management. There can be double benefits of fixing both the costs and improving the underlying ways of working. In addition, this helps teams be better prepared to meet the challenges ahead with the emergence of new technologies, such as AI LLMs.

https://techcrunch.com/2022/03/14/inflation-is-real-google-cloud-raises-its-storage-prices/?guccounter=2
https://www.cio.com/article/651215/price-shock-ibm-to-increase-cloud-costs-by-up-to-26-from-2024.html
https://aws.amazon.com/blogs/aws/new-aws-public-ipv4-address-charge-public-ip-insights/
https://info.flexera.com/CM-REPORT-State-of-the-Cloud
https://www.infoworld.com/article/3710968/a-cloud-professionals-cloud-predictions-for-2024.html

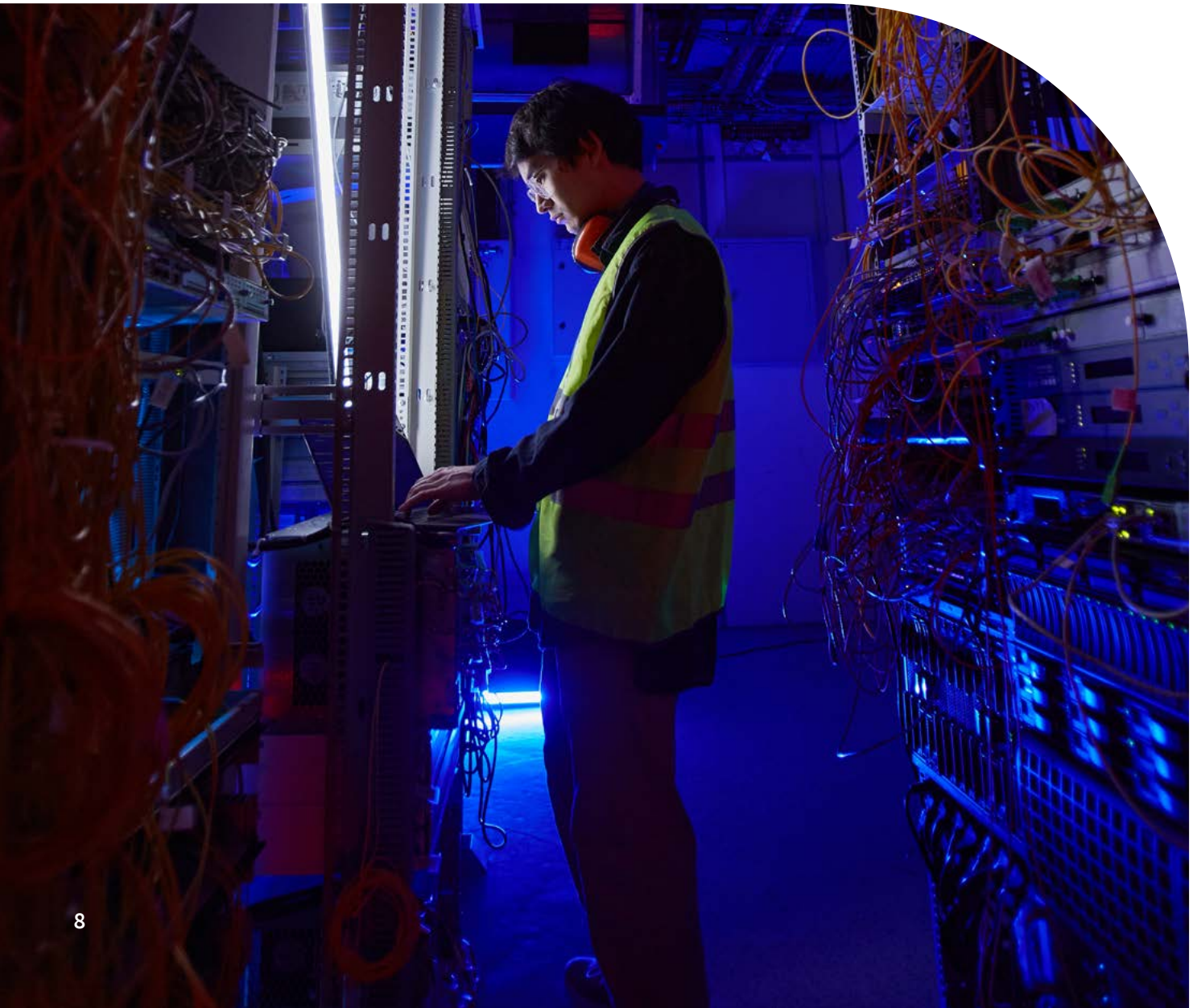
Optimisation begins with culture change

With the impetus on cost optimisation, it is important to recognise that it isn't just a one-off project. Cost optimisation is an ongoing set of practices and processes. In essence, effective cost optimisation is all about change – not just changing the way cloud is provisioned but transforming the behaviours around its use. As a result, it requires a cultural shift and buy-in from the entire organisation.

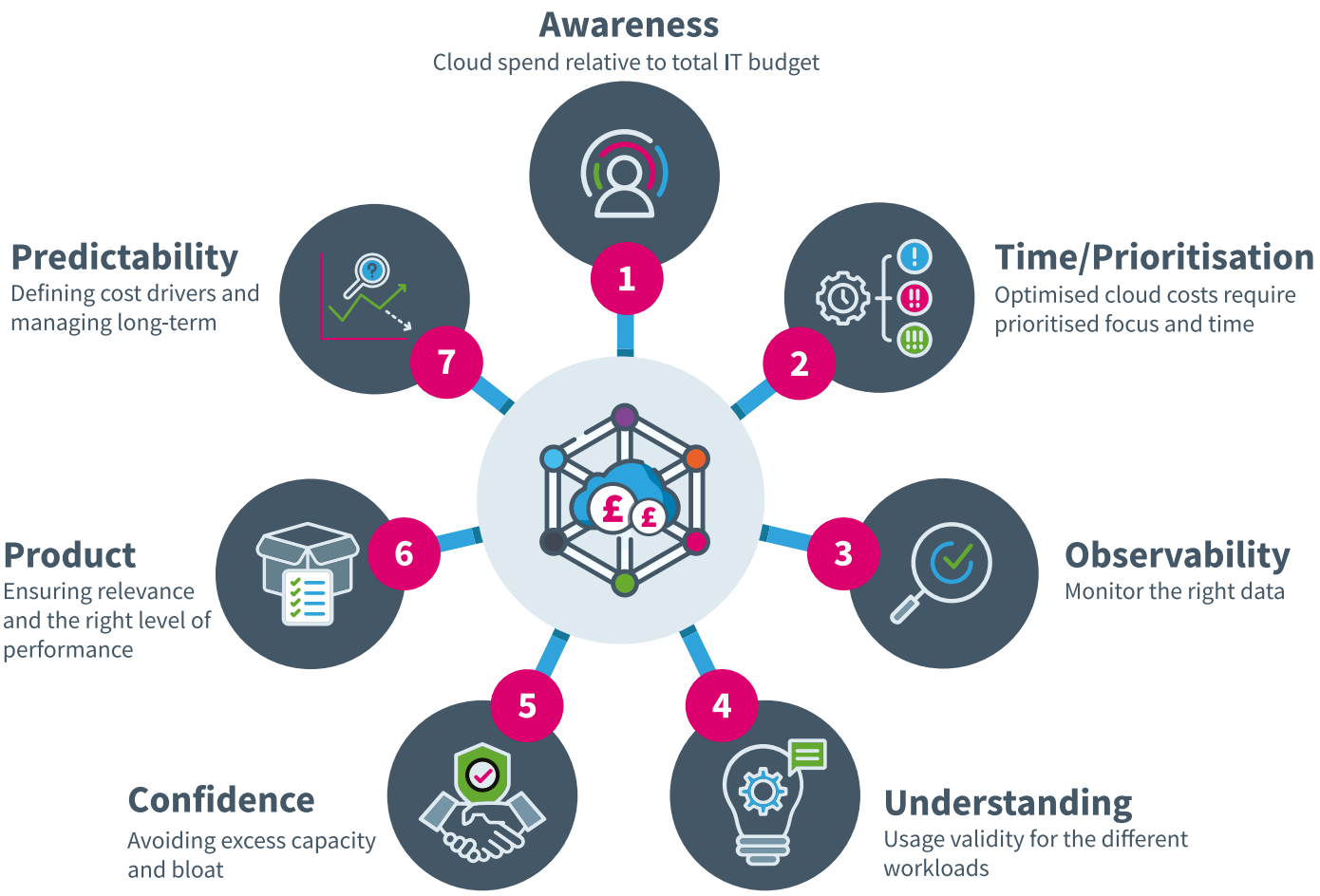
Taking from the Werner Vogels playbook - the CTO at Amazon who famously drafted the Seven Laws of Frugal Architectures - I have developed our own set of principles to guide successful cloud cost optimisation. Focused on fixing the problem of underlying drivers of cost increase, speeding up product development, and capitalising on the immense potential of the cloud.

In addition to setting out the principles, I discuss more than 20 real-world examples of how to achieve real change and transform ways of working into a more thoughtful, streamlined approach to managing cloud costs and service reliability. I provide practical insights into enhancing your cloud efficiency, particularly around understanding exactly what is driving your cloud spend.

This more thoughtful and deeper approach to cloud spend allows your engineers to make informed choices and better prepare the organisation for future uncertainty. Think of it as doing a regular fire drill so that when the time comes to make rapid reductions in costs, you can do so in a controlled manner.



The 7 Principles of a Cloud Cost-Efficient Organisation



The seven principles detailed on the following pages are not necessarily sequential. However, many organisations do start their cost optimisation journey with the first one. In addition, the end goal is the final principle, predictability in both costs and service reliability for the long term. In each of the following sections we provide tips on what to do, as well as benchmarks for you to compare how your organisation is doing.

The 7 principles are as follows:

1. Awareness

Teams must be aware of cloud costs relative to the organisation's technology spend and business revenues. Importantly, they need to know where the costs were previously, and where they are likely to be in the future, including their own contribution. In this principal, ownership is included but awareness is a prerequisite of ownership

2. Prioritisation

There needs to be dedicated, prioritised time set aside to manage and continually optimise cloud spend with the appropriate tooling and data investment.

3. Observability

Organisations need visibility at the correct granularity for cost, performance, and utilisation data over short and long timeframes.

4. Understanding

Teams should measure, quantify, understand and articulate the business drivers of capacity utilisation, cost, and performance of their services.

5. Confidence

Teams need the ability and the confidence to make cloud infrastructure changes quickly and without impacting service reliability.

6. Product

There must be a solid understanding of the value a software product effects or attributes bring to the users and/or business and appropriate level of service quality or performance required to deliver that value. In addition, even if these effects or attributes gave value in the past, organisations must ensure they will continue to do so.

7. Predictability

Teams need the ability to predict cloud spend over the next three years based upon predicted business demand to a high degree of accuracy. This time window gives organisation time to fix long term deviations. The business should have confidence in these predictions especially so that the CPO/CTO/Head of Operations is confident in delivering the right level of service reliability.

Principle #1 – Awareness

Typically, teams are not aware of the cloud budget and how much it is relative to the technology budget. In the FinOps approach to cost optimisation, this is the starting point to building a cost ownership culture – with the assumption that technical teams are responsible for cloud usage and optimisation. The tendency is to view cloud spend in the context of overall business revenue. However, this creates a false sense of efficiency when the reality is that cloud spend is increasingly eating into the technology budget.

The first step in developing awareness is to create context around your cloud spend and how large it is. You can further develop this context by identifying the figures from previous years, as well as predictions for the future.

The second step in developing awareness is for teams to identify how much of the overall cloud budget is consumed by their own service. Reports showing the spend of their systems should be provided on a regular basis to influence decisions.

The third step in developing awareness is recognising that cloud costs are important to the team. Too much cost relative to the business value delivered suggests that the service is not behaving as expected and increases the risk of reliability problems in the future.

Cloud cost management tools such as Cloudability, CloudHealth, AWS Cost Explorer, and Azure Cost Management can help, but engineers need the right level of awareness so that the tools can help deliver the expected outcomes.

Finally, the engineering teams should be accountable for their own service/ product budgets, as well as the service quality of the platforms. The team lead for that service would own the backlog, and interface with product manager to influence the product roadmap to minimise platform costs.

Benchmarks and examples of awareness in a cost-efficient culture

Inelastic Non-Production Test Costs
How to spot? Non-production costs over weekends are > 60% of weekday spend.

Idle or unused compute/VM instances
How to spot? Instances with very low CPU utilisation and network traffic

Idle or unused storage, e.g. unattached storage
How to spot? List of unattached storage volumes in the management console.

Solution: Many tools will provide you with a list. The effort of removing these is easy, you just need access to the management portal, or you can create a script to remove all unattached storage.

Unnecessary CloudTrail (AWS) costs
How to spot? CloudTrail costs should be close to zero or very low.

Solution: Significant costs indicate paid Management Events due to trail duplication (note: first trail is free). Configure an organisation-wide trail to capture all data in one trail and remove all other trails. Assign access to data from organisation-wide trail at S3 level.

Principle #2 Prioritisation

Teams need the time to prioritise to keep cloud costs optimal. This means teams spending time looking for unnecessary capacity usage or cost, which has no impact on the reliability of the service. This is often done by the cloud infrastructure or operations team but can also be done by the development teams who own the services or environments.

Teams should also ensure that they have a certain amount of time set aside to do these housekeeping activities – whether included within a sprint cycle or having a dedicated optimisation sprint cycle at regular intervals. The development teams may set aside a percentage of time for running their systems of which cost optimisation should be a part. For example, a figure of 5% for a five-person team would mean about 1.25 days per week focusing on managing the cost of the estate.

There is a temptation for product teams to be focused on features, neglecting high-value housekeeping activities. However, the housekeeping activity improves the teams understanding of the costs associated with cloud services and how they work.

Benchmarks and examples of awareness in a cost-efficient culture

Using unoptimized instances/capacity types

How to spot? Old instance types (e.g. AWS R5 or below, Azure Ev3 or below), or instance types that suit the workload (e.g. memory intensive workloads use memory optimised instances).

Benchmark: Old instances/capacity types should be < 1% of your environment.

Use of expensive storage classes

How to spot? There are a lot of GP2 EBS volumes (AWS), a lot of P-series/Ultra Managed Disks (Azure).

Solution: Move to cheaper storage class options without compromising performance (i.e. gp3 (AWS) or E-Series (Azure)).

Unnecessary snapshots

How to spot? Snapshots over one year old, or no snapshot lifecycle rules around data retention.

Solution: Delete old Snapshots and configure a retention policy to automatically detect and delete old or unnecessary snapshots.

Unnecessary backups on expensive storage

How to spot? Backup devices always written but are rarely read and are therefore likely to be eligible for placing on a cheaper storage class option.

Principle #3 Observability

Having the right level of observability helps manage costs and performance of cloud systems. It also provides insight of the team's understanding and confidence in their systems. For example, less confident teams will log everything, use logs rather than metrics, and keep data longer than required. Although these costs aren't the greatest, they are often simple to optimise.

Three areas to consider:

- Tagging: Cloud systems need to be tagged to ensure cost ownership can be tracked. This is the foundation of building awareness of your own costs and that of others (addressed in the first principle).
- Metrics vs logs: Many logs can be summarised as metric data. Metric data requires only a small fraction of ingestion and the storage costs that logs incur. An additional benefit is that metric data is easier to gain insights from compared to log data.
- Granularity and data retention: For system data, minute level granularity is sufficient in most instances. For cost data, hourly granularity is sufficient. There should be a summarisation strategy to reduce data retention.

Benchmarks and examples of awareness in a cost-efficient culture

Unnecessary monitoring costs (Cloudwatch for AWS, Monitor/Log Analytics for Azure)

How to spot? AWS Cloudwatch/Azure Monitor should be < 1% - 2% of cloud spend. These can be driven by custom metrics, VPC flow logs, etc. Environments with high monitoring costs often have low levels of observability. The effort of removing this can be challenging as you need to find the engineer who is using these specific metrics.

Solution: Turn-off the metrics and see if anyone complains.

Unnecessary Kinesis (AWS) monitoring costs

How to spot? A step change in Kinesis capacity results in a disproportionate increase in Cloudwatch costs. The default set-up is to collect data for each Kinesis shard – if there are a lot of shards this becomes a very expensive option.

Solution: Turn on only when troubleshooting and not for day-to-day observability.

High APM costs (Splunk/New Relic, etc.)

How to spot? APM costs are > 3% of cloud spend. This can be driven by unnecessary logging, collecting log data which could be collected as summarised metric data, data lifecycle rules that hold data for longer than necessary, etc.

Principle #4 Understanding

Prevent unnecessary processing and therefore cost by ensuring usage validity. This means understanding if the usage is valid for the different workloads. The latter is increasingly becoming the case for many organisations as data-driven workloads become more important.

In many organisations there is a clear relationship between the business demand signal and their key systems capacity, e.g. website orders and ecommerce systems. However, the usage and growth of data-driven workloads are not necessarily driven by business demand signals.

Data-driven systems may experience cost increases at a rate higher than revenue growth or demand signal. These investments are made with the intention that they enable process mining, business process optimisation, and the discovery of new business opportunities.

As the cost of these systems and services consume more of the total cloud budget, optimisation in analytics becomes more critical. However, it should be understood from both engineering and the business that this is an investment that may or may not impact the top line and the business should support this investment.

Benchmarks and examples of awareness in a cost-efficient culture

High background usage
How to spot? With non-elastic environments, CPU usage does not drop below 5% during non-peak periods.

Capacity usage doesn't follow the demand signal
How to spot? The minimum and peak capacity usages do not scale up with the daily business demand signal - it is either below or above the expected value. This can be applied to any transient cloud resource metric including CPU usage, network, disk IOPs, etc. In fully elastic resource types, the minimum and peak cost should scale up with demand signal, e.g. network traffic (egress, inter-zone etc).

Data-driven system costs increase faster than business revenue
How to spot? Look at the monthly cost of data over a three-year time window along with business growth.

Principle #5 Confidence

This is one of the largest opportunity areas. Cloud estates are bloated with excess capacity due to a lack of confidence in reliability of services. There is often organisational reluctance to remove capacity due to fear it will impact user experience. This is driven by a multitude of factors, the most common being:

- Service incidents
- Sizing based on assumptions
- Sizing based on overly simple models and/or flawed methodologies
- Using the wrong metric to size the peaks
- Inaccurate testing
- Technical debt

Increased confidence means greater speed in delivering quality code and products into production. Teams should develop a culture and supporting processes to determine the resources their services use, and to understand what the "safe" limits are for those resources. This enables resource optimisation over relying on over-provision. This becomes especially important in larger enterprises where delivery velocity slows down.

Benchmarks and examples of awareness in a cost-efficient culture

Oversizing of compute
How to spot? Average estate CPU utilisation should be >40% with utilisation peaking >90% for key systems.

Oversizing of memory
How to spot? Small or zero physical read disk activity on databases when there is high logical read activity.

Oversizing disk volumes
How to spot? Utilisation of EBS volumes is <90% and/or large amounts of available EBS space for large volumes (e.g. >100GB for 1 TB drive). Use of fixed thresholds irrespective of size of volume.

Oversized test environments
How to spot? Test environment spend >15% of all cloud spend suggests teams are oversizing test environments because they don't have confidence interpreting results from small environments.

Principle #6 Product

Performance and obsolescence/relevance are two elements of a product that can lead to increased cost and risk.

Performance

Ensure that both business and engineering teams are aligned on the level of availability and performance customers should get based on market conditions, i.e. understand the level of performance and resilience a system should have relative to its cost and value. Whilst the cloud provides gold-standard levels of service, reliability, and performance, your customers don't necessarily need them or are willing to pay for them. The business should deliver an appropriate level of quality and performance that is in line with that market, or the product/service strategy – whichever is the greater.

Obsolescence

Both the business and engineering are aligned on the value that effects / attributes deliver to users / business. Effects or attributes may have delivered value in the past but as the market changes these may not continue to deliver the same value as before.

Benchmarks and examples of awareness in a cost-efficient culture

Excess resilience

How to spot? Use of three availability zones for production systems, and more than one availability zone for non-production systems. AWS AZs are at least 30 miles apart and the use of three AZs would be considered overcautious for many systems or services especially if they also have Disaster ecovery (DR) capabilities. GCP/Azure are less clear on the distance between the AZs.

Excess performance

How to spot? IO2 rather than GP3 (AWS), premium SSD rather than standard SSD (Azure).

Inefficient Processing

How to spot? High levels of transactions relative to the business processing volumes. This level of detail will be found in your observability tools, including those provided by the cloud service provider.

Principle #7 Predictability

Having a long-term view on costs or service performance is key to success – as is understanding the drivers of those costs. In addition, understanding what value is delivered to the business by these costs allows for better conversations between the business and the engineering teams on the value of cloud services.

This is also important when negotiating minimum cloud commitments - AWS (EDP), Azure (Enterprise Agreement), GCP. Knowing exactly what you need will get you the best discount. Avoid using a cloud provider's forecasts as they always assume a simple straight-line forecast. Only you will know what you expect to spend based on business growth, optimisation plans, and technology changes planned over a three-year time window. Remember, cloud service provider (CSP) enterprise minimum commitment can be negotiated at any time.

Benchmarks and examples of awareness in a cost-efficient culture

Costs forecasts are driven by oversimplified model/linear trend

How to spot? The forecast cloud costs of all systems are following the same growth projections as the business. Systems grow at different rates – it is highly unlikely that all systems would follow the same growth trajectory.

Cloud spend driven by your CSP

How to spot? The CSP provides a linear trend on your cloud costs and teams agree to sign up to multi-year minimum commitment deal based on this calculation.

Cloud costs are higher than planned

How to spot? The cloud costs are tracking above the planned numbers and the explanations for the increase are attributed to features and functionality, etc. These explanations are not backed up with data.

Keeping costs within budget requires more effort than expected; a last-minute dash; or a freeze on spend

Solution: Understand the underlying driver of cost over a three-year timeframe, broken down by resource type, team, and environment. Create a simple predictive model based on your demand signal for the next 12 months. Track actuals against this model as when you see the differences it will make more sense. Look at different regions or new platforms, misconfiguration, or different usage patterns to really gain an understanding of these differences.

Overcoming the challenges of building a cost-efficient organisation

Cost planning and control is an ongoing commitment. Controlling costs is not just about reducing costs for a year – it is a permanent change in the way you work. A change in your culture. As a result, building a cost-efficient organisation doesn't happen overnight and there will be challenges along the way.

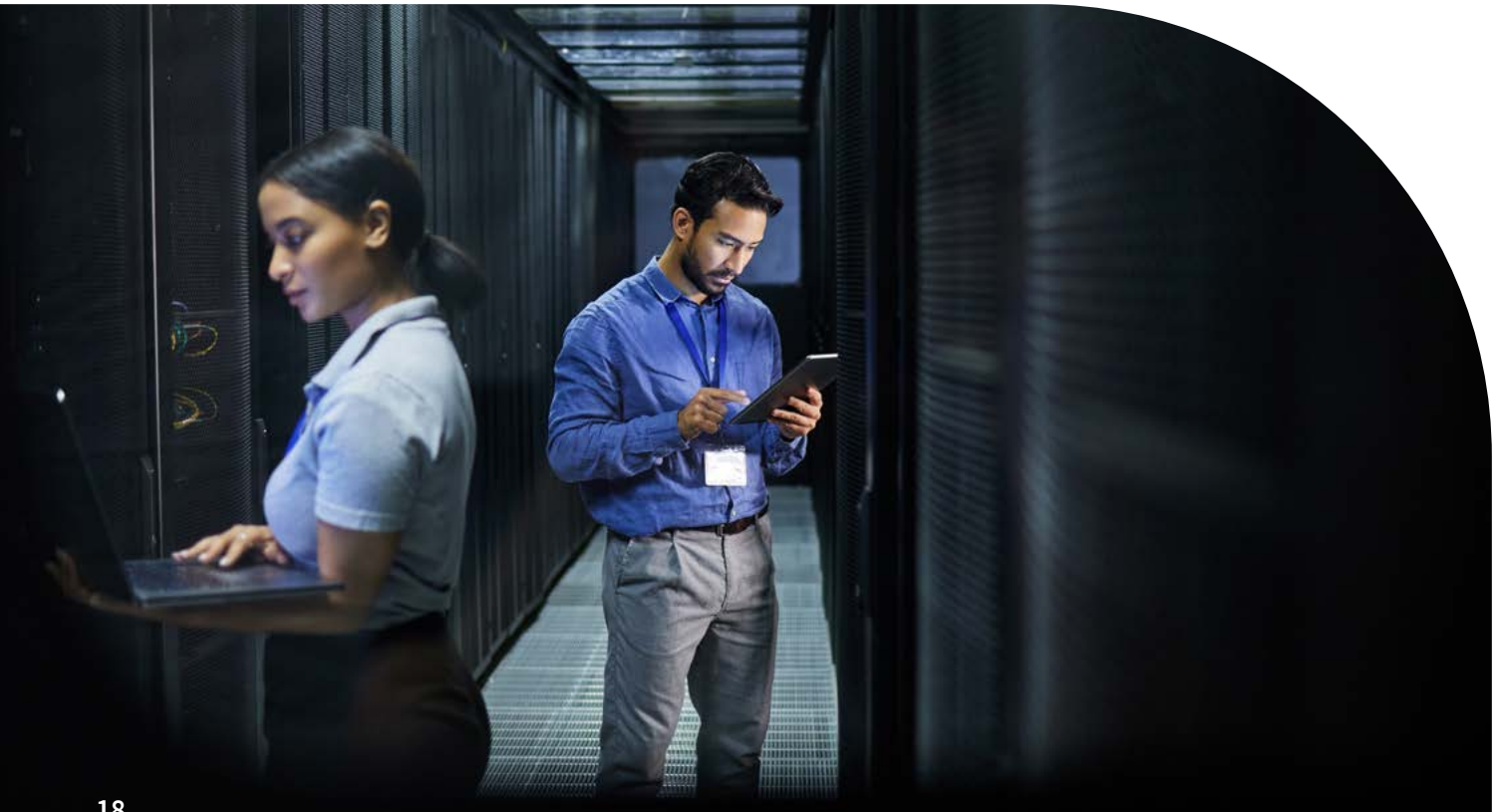
One of the biggest – which ties back to the discussion on cultural transformation – is people. At the heart of it, if your people are not onboard or comfortable with your approach to managing cloud costs and optimisation, it will be difficult to embed a cost-efficient culture in your organisation. Even a single individual can influence the process and hold back the entire organisation's transformation.

The approach to cloud cost optimisation requires a long-term approach to ensure sustainable gains, and to a lesser degree, there will be challenges to creating measurable goals that demonstrate that your organisation is becoming cost efficient.

Overcoming these challenges isn't insurmountable. There are key areas that can help including securing executive sponsorship with a leader demonstrably interested in cloud cost insights, and who is able to champion their importance within the organisation.

Other areas to consider, include:

- **A formal programme with an informed goal:** There is thinking behind the "number" to show that the goal, although a stretch target, is realistic and based on data.
- **Empower engineers and optimise systems:** Excess cost reduction isn't just about savings; it's about empowering engineers with deeper system understanding. Cloud cost becomes another vital metric alongside infrastructure data, revealing inefficiencies and opportunities for system health improvement.
- **Upskill and optimise:** Foster continuous learning through cost optimisation. Engineers gain valuable new skills while identifying cost reduction opportunities, leading to additional system health benefits and a change in mindset.



There will be mistakes and missteps. When starting your journey to becoming cost efficient, you need to accept that mistakes will be made; but that is not a bad thing as they serve as learning opportunities. This is especially important considering that all aspects of the IT delivery function will need to adapt – architecture, design, engineering, testing and operations.

At Capacitas we recommend building partnership – from senior leadership through to engineering teams on the ground. Show the various stakeholders how a cost-efficient organisation makes their life easier.

Partner with leaders, showcase deeper cost insights

Collaborate with leadership to unlock the potential of a deeper, data-driven approach to cost and capacity management. Demonstrate how these insights can optimise costs and improve decision-making.

Deep dive to identify massive savings

Partner with your team for a comprehensive cloud analysis using a proven methodology. Analyse cost, performance, utilisation across different timeframes,.. Go beyond tools to uncover substantial savings.

Deliver value with a phased programme

Work with senior management to implement a programme delivering identified opportunities. Prioritise and sequence optimisations to build trust and lay the groundwork for tackling complex solutions.

Empower engineers to improve reliability

Partner with engineers and use a standard data analysis model methodology that can be used across multiple teams and technologies (Capacitas uses Ramp Zero Flat (R0F) - which is available on request). This delivers safe optimisations, increases reliability, and establishes a common approach for large, multi-team organisations.

Myth busting with data

Use the methodology to debunk myths about incidents and excess capacity. We build confidence with data-driven insights.

Collaborative planning for finance and engineering

Guiding finance and engineering to establish a data-driven capacity planning process. This empowers accurate forecasting of future usage and costs.

Enabling cost optimisation culture change through team structure

As discussed, optimising cloud costs and reliability goes beyond simple tools to include cultural change. It also requires the right team structure and the fostering of deeper conversations between teams. This unlocks not just cost savings, but also benefits such as improved risk management and increased agility.

Product-focused organisation structure

Figure 1 below illustrates a sample team structure for larger, product-focused organisations. These showcase potential conversations each team should engage in. However, remember, every organisation is unique. Factors like company size, structure, and goals influence the ideal team configuration.

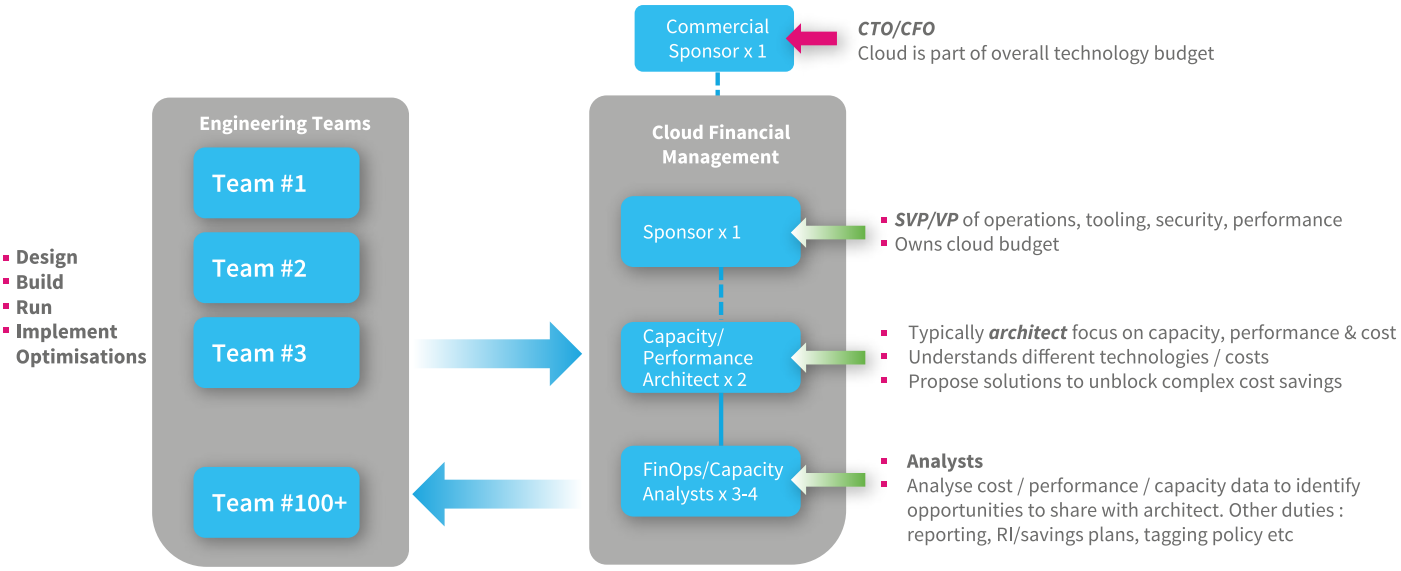


Figure 1 - Diagram

Centralised control for the non-product focused organisation

Enterprises without a strong product focus might prefer a centralised cloud governance team. This approach works well for companies with standardised workloads and less variation across departments.

In such environments, collaboration is key. Success lies in fostering meaningful conversations between teams, regardless of the chosen structure.

Consider these examples as starting points, not rigid models. Your optimal team configuration will depend on your specific needs and goals.

It's good to talk - championing cost efficiency

The cloud cost optimised organisation is reliant upon productive conversations. The capacity and performance architect needs to be a champion of cost efficiency. This requires a suite of capabilities, including:

- Knowing what good performance and optimised costs looks like using multiple data sources
- An experimental mindset that can be used for planning optimisations
- Critical thinking to challenge pre-existing views utilising both performance and cost data to support the challenge
- The ability to contextualise both monetary and performance value to teams, e.g. opportunity to learn from experiment more about the systems
- Help teams on the path of delivering optimisations
- Tracking and reporting on transformation progress, including unblocking and escalating as necessary

Figure 2 below summarises how this works in practice.

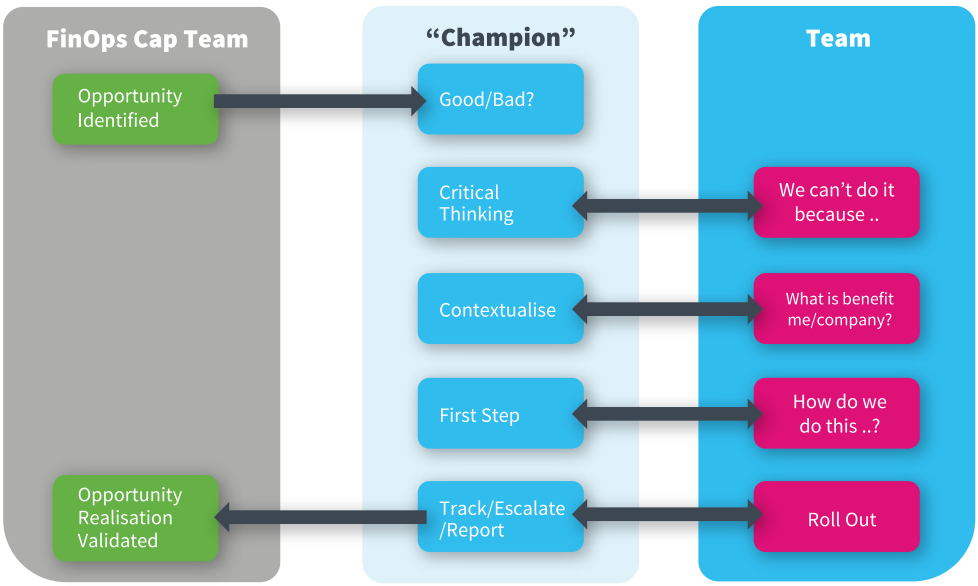


Figure 2 - Critical thinking

Organisations may be under the perception that they have the right structure in place and are following the 7 principles of a cloud cost-efficient organisation. You will discover if the seven principles are working as expected, with the self-diagnostic benchmark marked in green for each of the laws.

Conclusion

The premise of this paper was to provide a more thoughtful approach to cloud optimisation, and to do so by optimising cloud costs in order to maximise the value and opportunities from cloud.

Cloud is meant to bring agility and flexibility whilst being cost-effective, stable, and delivering high levels of performance. However, if you're spending all your time maintaining cost and stability status quo, then your cloud estate is most likely NOT delivering the value that you need or should expect.

Managing costs and understanding what is driving them go hand-in-hand when trying to boost the value of cloud. This forms part of cloud cost optimisation or building a cost-efficient organisation. As with many things, it is a journey that your organisation needs to embark on. It requires change, cultural transformation and buy-in from all stakeholders. The reason? Cost-efficiency isn't a one-off programme. It needs to be an embedded set of cultural practices that lead to ongoing gains and long-term benefits.

This is our experience. The best approach is to go beyond basic cost optimisation and look at the bigger picture – past, present, and future. You do this by analysing detailed cost data, capacity utilisation, business demand, and performance data in relation to your business requirements to build that view. From there you can develop recommendations to optimise and save, without compromising on performance.

Building a cost-efficient organisation that can capitalise on the opportunities of cloud while experiencing long-term, sustainable gains is more than possible. It is probable, with the right approach, understanding and culture.

Key takeaways

1. Cloud optimisation does not happen overnight and needs to be effective and sustainable to have long term benefits of new technologies.
2. There needs to be a cultural transformation – detailed in our seven principles of building a cost-efficient organisation.
3. All stakeholders need to be aligned to this mission (including technology teams, the board, external regulators, etc)
4. A cost-efficient culture improves team agility, manages risks more efficiently and prepares for future uncertainty.
5. Improved conversations with data between the 'performance/cost champion' and the business and technology teams makes transformation become real.

About Capacitas

Capacitas has over 20 years of experience partnering with organisations that have faced complex scaling and cost challenges. Capacitas challenges the cloud status quo by bringing a new way of thinking that allows organisations to be confident in removing excess cost, while maximising performance and scalability.

This paper is based on 12+ years of experience with many major cost/performance optimisation programmes, including dozens of private equity operating partners and CTOs helping companies save more than £100M in OpEx. Capacitas has worked with brands including Ancestry, Jaggaer, DHSC, Cegid, Skype, Smart Data Communications Company and JDSports, driving value and helping them realise the opportunities inherent in their cloud estate.

ancestry

"Within 2 months, a significant amount of AWS cost reduction has been delivered in live. Working together with Capacitas we have made our services more stable and are well positioned to deliver further cost optimisation."

Niraj Nagrani, SVP Products and Platforms, Ancestry

SILVER LAKE

"Capacitas have a uniquely analytical approach to modelling both cloud and Capex based infrastructure provision. Their thoughtful and actionable insights unlock both spend and performance-based interventions supporting both architects and DevOps teams in optimization and forward planning for global scale services."

Mark Gillet, Managing Director and Head of Value Creation, Silver Lake

Data Communications Company

"Working together (with Capacitas), we delivered 36% cost avoidance and identified a further additional avoidance opportunity of 17%. At the same time, the engagement pinpointed and removed six critical scalability and performance risks. The engagement also fostered great collaboration between our commercial & technical teams and our technology service providers.
As a result of this engagement, we are confident that the service can scale out cost effectively in the future. We are working with Capacitas to follow the same approach for other core services."

Alex Henighan, Director of Service Assurance, Data Communications Company (Smart DCC)

About the author

A 25+ year veteran of cloud, Manzoor Mohammed is co-founder and Chief Innovation Officer of Capacitas. The business has a proven track record of saving organisations millions in annual cloud spend and optimising the performance of their complex systems. Capacitas has helped companies achieve their financial goals through strategic cloud management.

Manzoor recognised the transformative power of cloud computing early on, understanding how it strengthens the link between performance and cost. This realisation led him to develop the 7 principles of a cloud cost-efficient organisation – an integral component of Capacitas's unique delivery methodology, which has empowered numerous clients to achieve significant cost reductions and performance improvements. These include easyJet, Skype, JAGGAER, Ancestry, Cegid, and BMC Software.

Key results include:

- Reducing Cegid's cloud spend by €7M p.a. - 27% increase in SaaS users.
- Scaling Skype from 60M to 90M subscribers & reduced their cloud spend by $26M p.a. & 96% reduction in incidents.
- Saving Ancestry $30M p.a. cloud spend & 30% increase in users.
- Reducing cloud costs by $2.7M p.a. for JAGGAER, developing the internal culture and building internal capabilities to sustain cost optimisation.

Manzoor is a trusted advisor and thought leader in the cloud computing space, passionate about helping businesses leverage the cloud to achieve their full potential.

Thanks

In addition to a host of resources drawn upon to draft this paper, we would like to thank the following people for their invaluable feedback, support and inputting their experiences into the document:

- Elinor MacKinnon, Startup Advisor, CIO, CTO, Partner StrataFusion
- Rajeev Dave, Consultant, StrataFusion
- Lars Rabbe, ex-CIO Skype/Yahoo, Partner StrataFusion
- Andre Brunetiere, CPO, Cegid
- Matthias Spycher SVP, Fanatics
- Matt Kane, Operating Partner, Silver Lake
- Julian Abad, VP, Sage
- Josh Fishman, VP, Sabre
- Brendan Farrell, VP, BMC
- Shankar Gomathinayagam, VP, BMC
- Theo Beack, Operating Partner, Covehill
- Benoit Marc, Head of Cloud, Cegid
- Charles Beadnall, CTO, Go Daddy
- Nigel Beighton, CTO, Digimgo Technology
- Ramadass Prabhakar, CTO, WPengine
- Mark Gillett, Head of Value Creation & Managing Director, Silver Lake
- Phil Scully, VP Digital & EMEA Technology, RS Group Plc.
- Martin Kersch, CTO, Jaggaer
- Andrew Gratton, Director of IT, Stonewater
- Charles Forde, COO, Nomura

References

https://www.gartner.com/en/newsroom/press-releases/11-13-2023-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-679-billion-in-20240

https://www.pwc.com/us/en/tech-effect/cloud/cloud-business-survey.html

https://www.pwc.com/gx/en/issues/technology/emea-cloud-business-survey.html
https://www.gartner.com/en/documents/4789331

https://techcrunch.com/2022/03/14/inflation-is-real-google-cloud-raises-its-storage-prices/
https://aws.amazon.com/blogs/aws/new-aws-public-ipv4-address-charge-public-ip-insights/

https://www.cio.com/article/651215/price-shock-ibm-to-increase-cloud-costs-by-up-to-26-from-2024.html

https://info.flexera.com/CM-REPORT-State-of-the-Cloud

https://www.infoworld.com/article/3710968/a-cloud-professionals-cloud-predictions-for-2024.html

https://thefrugalarchitect.com/

www.capacitas.co.uk