

Reducing the Runaway Costs of a Hybrid Big Data Architecture

The Big Data Challenges of the Cloud Era

The rise of the cloud has changed the face of big data. Today, whether through lift-and-shift or re-architecting, almost every large enterprise has carried out some migration to the cloud. For the majority of large enterprises, privacy concerns, ongoing licenses, and large existing investments in on-prem systems mean that they have kept some of their workloads on-premises. In other scenarios, company acquisitions have landed an organization with a large cloud profile. However they have got there, most enterprises are now managing a hybrid – and usually multi-cloud – big data environment.

This pressure on IT operations to effectively manage a hybrid big data architecture will only increase.

35% of CIOs are decreasing their investment in their infrastructure and data center, while 33% are increasing their investment in cloud services or solutions (Gartner, 2019).

Big Data IT Operations teams continue to face the same challenges. They are running enormous big data clusters with thousands of nodes, and they require full stack visibility in order to optimize application performance, support SLAs, uncover infrastructure inefficiencies, and minimize MTTR (mean time to repair). They need to make Spark apps run faster and stop Hadoop clusters from blowing up. They need to deal with any malfunctioning workload as quickly as humanly possible.

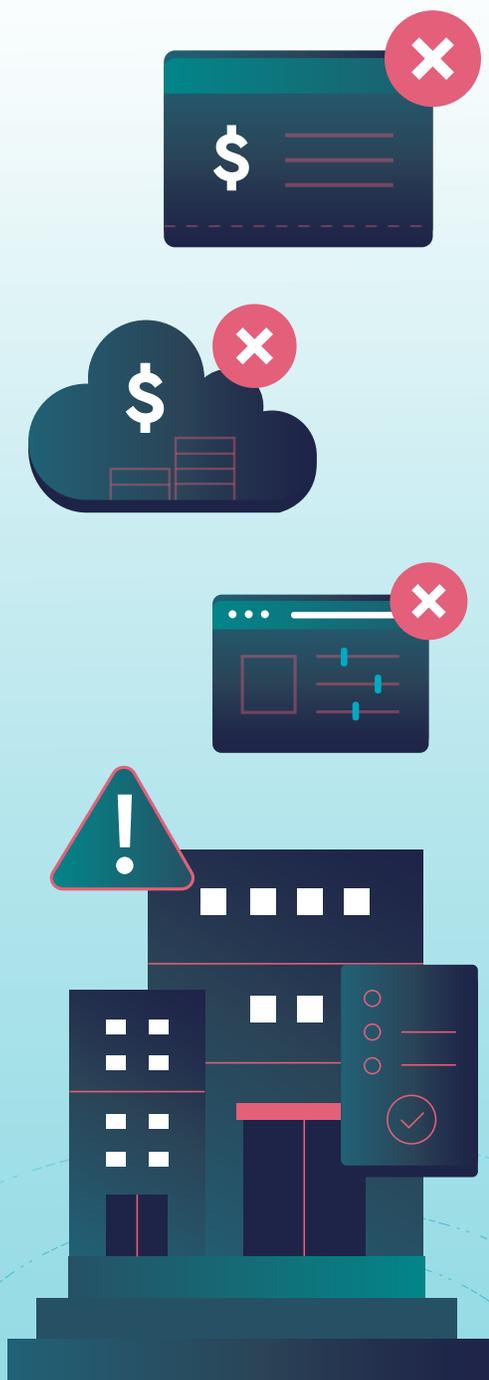
The problem? The management and optimization of big data in a hybrid cloud architecture is a new and evolving challenge. **As Gartner put it:**

“Changes in architectures, and specifically with funding moving away from the traditional data center with the advent of cloud, hybrid cloud, virtualization, SDN/SD WAN and disaggregation, change how infrastructure monitoring needs to be conducted.”



In the early days of cloud migration, it was all upside: Operating a data center in the cloud is always cheaper than on dedicated on-premises servers. Nothing to worry about. Fast-forward a few years, and IT operations is in a visibility crisis. Compared to their legacy stance, they suddenly cannot understand what they are spending or why.

“Through 2020, 80% of organizations will overshoot their cloud IaaS budgets due to a lack of cost optimization approaches” (Gartner, 2019).



And this visibility crisis is translating into a cost crisis. When enterprise IT organizations receive their first few cloud bills, many are shocked. Cloud invoices can run to thousands of lines, and can add up to hundreds of thousands more dollars than expected. When **Bain & Company** asked more than 350 IT decision-makers what aspects of their cloud deployment had been the most disappointing, the top complaint was that the cost of ownership had either remained the same or increased.

The sources of this crisis are made very clear by Gartner in a **2019 paper**, where they summarize the situation:

- 1 Complex multi-cloud environments are becoming commonplace, and billing details vary dramatically by provider.
- 2 Many I&O (infrastructure and operations) teams are operationalized for traditional data center principles rather than cloud IaaS, and they lack the organizational processes to manage costs in the cloud.
- 3 There are many options to address cloud expense management. As a result, I&O leaders may struggle to align options with the organizational cloud strategy.

How can we meet the challenge of this complexity, billing puzzle, and organizational struggle?

CapEx to OpEx: The Start of the Problem

Whatever their distribution of on-prem and cloud workloads, large enterprises with significant big data commitments share many of the same features: They are managing their big data using Hadoop systems, including Hadoop-in-the-cloud systems like Microsoft HDInsight and Amazon EMR. They are running databases, web portals, and the architecture that is available to a range of internal teams – almost like an internal cloud.

Large-scale cloud migration begins for a number of reasons. Contributing factors include an on-prem data center that's reached its capacity limit, aging hardware, and licenses that are expiring. But the most common reason is the frustration of internal, often customer-facing, teams.

Internal departments, pursuing their own KPIs, always want more computing resources, and they put pressure on the IT operations team to provide them. IT operations, meanwhile, are receiving a conflicting message from the CFO; they are being told to minimize spend and be as streamlined as possible. Legacy big data approaches often don't leave IT operations teams with many options. In an on-prem data center, there is an inherent and internal limit to compute capacity. An on-prem data center will never double its capacity overnight. Any utilization gains are hard-won, and big data specialists and ITOps teams can end up tearing their hair out attempting to free up resources to help meet requests from within the organization.

The cloud is seen as the obvious solution to this problem. With AWS, Microsoft Azure, or Google Cloud, you face none of the baked-in limitations of an on-prem data center. The technical and internal bottlenecks of the legacy architecture vanish. Compute capacity is theoretically unlimited.

The legacy, on-prem data center operated within a CapEx model. Though the tech was constrained, so was the budget. But as the infrastructure migrates to the cloud, a CapEx model is exchanged for an OpEx model. And here's where the trouble starts.

In the CapEx framework, the balance sheet was very clear and projections were simpler. Traditionally, the CFO would oversee strict cost control mechanisms. Though this translated to constrictions on compute capacity, the trade-off was watertight budgeting. But in the cloud-based OpEx paradigm, the control flows of how money is being spent suddenly become much looser and harder to define as there is no hard-coded capacity ceiling.

As cloud migration is usually a large-scale, exciting, multi-stakeholder internal project, enterprises throw resources at it. To unleash the full computing capacities of internal teams, the migration leaders are afforded a large budget. The engineers are delighted, as they love the idea of the new capacity. For every internal team, an all-you-can-eat approach to resources sounds like the promised land.

An OpEx spending model plus the infinite resources of the cloud equals a recipe for overspending.

A cloud-based architecture can be more streamlined and cheaper if the processes and controls are strict and clear. But this is almost never the reality when organizations migrate workloads to the cloud. Instead, in this brave new world, an engineer can spin up a hundred-node cluster in AWS on a Friday; forget about it and go home; and discover a month later that over the weekend it racked up thousands in cloud costs.

Why Controlling Cloud Spend is Like Flying Blindfolded

Of course, it is theoretically possible to put caps on cloud spending and control budgets within an OpEx paradigm. That the scenario described above is the norm is partially down to over-hasty or naïve project management and financial mapping.

There is also the issue of the cloud simply reiterating existing on-prem issues. Many companies perform a simple lift and shift, and the infrastructure they are lifting is already hampered by inefficiencies. As Bain & Company reveal, “we have found that 84% of on-premise workloads are overprovisioned. That means that when companies migrate a workload to the cloud, they're sending excess computing and storage capacity right along with it.”

The problem runs deeper than this though. Even with the best cloud migration strategy, and even the most dedicated attempts to curb cost, there are inherent features of the cloud landscape that make managing resources – and therefore cost – much more difficult.

Once you move to a hybrid model, you are never unwinding from the OpEx model, and you will always have to deal with potentially limitless provisioning. This – compounded by other forces such as the emergence of microservices and the Internet of Things – challenges traditional, on-prem-defined IT operations monitoring systems. **As Gartner puts it:**

IT operations monitoring has been around forever. The foundational goal has always been to gain insight into the availability and performance of systems, networks and applications, and to carry out root cause analysis of performance degradations. Once you move to the cloud, the basic mission is the same. But the technical problem is the same as the financial problem: unlimited scale.

In a large system, hundreds of thousands of instances will be supporting thousands of workloads, all of which are running big data computations for a range of internal customer teams. The range of ways to provision resources and compose the instance is much vaster in the cloud than in a legacy architecture. The users tweaking these provisioning parameters often have no insight whatsoever into how their actions will affect the next cloud bill. With so many live instances, the implications for cost can be very hard to track.

“Major changes in how applications’ infrastructure and networks are deployed and run introduce visibility gaps into traditional monitoring functions... Changes are coming to ITOps environments faster than monitoring strategies are evolving to handle them, leading to visibility gaps and performance challenges.”



Cost in the cloud is tied directly to ongoing consumption, so managing utilization is inextricable from managing expenses. But managing utilization requires full insight into your utilization. Visibility gaps start in the fact that IT operations teams no longer have access to the infrastructure and networking that underpin the big data services. Straight away, this is a challenge. On top of this, the cloud demands a formerly unnecessary range of skills. The hybrid big data environment expert needs to be familiar across both the cloud and on-prem infrastructures, at both the network and application layers.

The result? In the cloud, people often use considerably more compute than they anticipated, because they lack full and actionable visibility into their utilization and provisioning.

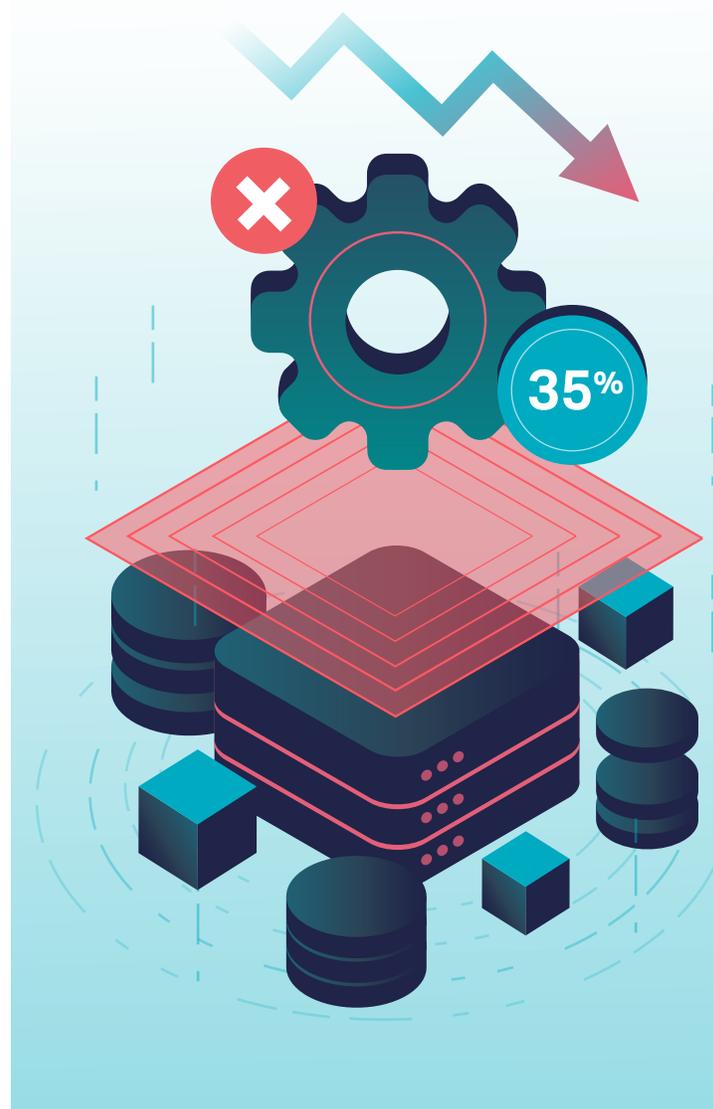
As if that weren't troubling enough, it isn't just overspend. The scale of the cloud environment, and the lack of visibility into how resources are being provisioned, also results in the opposite. Operators can't see what's going on, but they're terrified of underprovisioning, and so resources are oversized and left idling. Abandoned resources accrue charges without contributing value. Across the board there is an unnecessary over-provisioning of resources.

"Cloud services can have a 35% underutilization rate in the absence of effective management" (Gartner, 2019).

Cloud computing companies offer an à la carte menu of resources and options. There are thousands of different instances available on AWS alone. Because of this, cloud bills are complex, and run for thousands upon thousands of lines.

Parsing them can be difficult (and default Hadoop reporting has never been detailed enough). But every time a bill is woefully inflated when compared to organizational expectations, it is because of a damaging combination of overprovisioning and underutilization.

And compounding everything is the fact that AWS, Microsoft Azure, and Google Cloud have little incentive to keep their customers from overprovisioning and overspending. The visibility challenges, and thus potential for inefficiencies, will only increase with the growing complexity of the hybrid big data landscape: **In a Gartner survey**, 86% of respondents with cloud workloads said they "expect to have multiple public cloud IaaS providers by the end of 2020."



How Do Teams Get Spending Back Under Control?

What is the solution to all this? Firstly, budgets need to get back in line. And to meet budgets, organizations need transparency, so that the people who generate the cost are aware of what they are generating. AWS bills are often rolled up and sent to finance departments, but the developers and ITops teams never actually get much insight into how their actions translate into cloud spend. This should change. Other approaches, such as chargeback models, can help manage cloud spend.

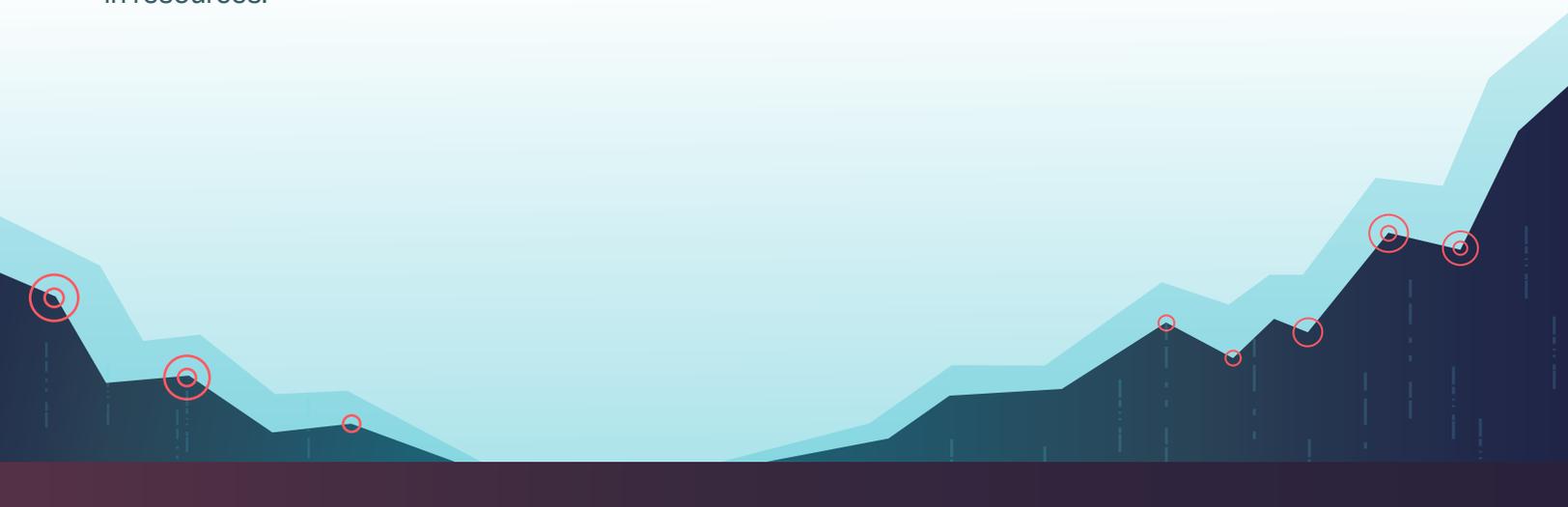
The key to rightsizing lies in visibility. You need to determine usage patterns; understand average peak computing demand; map storage patterns; determine the number of core processors required; treat nonproduction and virtualized workloads with care. To stay rightsized post-migration, you need full insight into the CPU, memory, and storage constituting every instance.

This sort of visibility is what can give people an understanding of what cloud costs they are generating. The only way to begin to bring down runaway cloud costs is by having a proper handle on resource use, spend, and how these two things interact. Users need to know what is actually going on with their big data jobs. They need the data and insights that will give them a clear picture of spend, wastage, and fluctuations in resources.

“Most monitoring solutions, while valuable in their own right, have not optimized the process of troubleshooting performance and availability problems. Users often complain of limited visibility with too few tools or too much complexity with too many tools, and everything in between” (Gartner, 2019).

However, the data and insights that IT operations teams need are almost impossible to acquire without the right tool. Even if they had the expertise, most organizations don't have the human resources or hours to dedicate to reducing cloud spend in a granular way. This would require expertise and time. Even someone with the skills would be playing a whack-a-mole of workload management.

The key to reducing the runaway costs of a hybrid big data architecture lies in efficiently analyzing and right-sizing on-prem and cloud resources to more closely match utilization. But the visibility that can empower an organization to do this can only come from dedicated software that offers powerful insights into jobs and usage from a bird's eye view. Not job by job, or individual user by individual user, but across the whole infrastructure. Hadoop couldn't offer this visibility before the cloud; it has no chance now.



Pepperdata Capacity Optimizer takes a unique approach to right-sizing by identifying wasted, excess capacity in big data cluster resources. By monitoring cloud and on-premises infrastructure in real-time, including hardware and applications, and leveraging machine learning with active resource management, Capacity Optimizer automatically re-captures wasted capacity from existing resources and adds tasks to those servers.

The net benefit is an increase in enterprise cluster throughput of 30 to 50 percent, or conversely, a 30 to 50 percent reduction in infrastructure resource requirements.

Ultimately, in the quest to control cloud spend, analytics are key. Without powerful, in-depth insights, big data teams simply don't have the information they need to do their job.

Ultimately, in the quest to control cloud spend, analytics are key. Without powerful, in-depth insights, big data teams simply don't have the information they need to do their job. For this, a dedicated solution is required – one that can:

- Visualize and optimize big data operations instantly, at scale;
- Offer a single dashboard for all big data environments, on-prem and in the cloud;
- Initiate automated infrastructure optimization



Pepperdata is built for IT operations teams who need visibility to recapture wasted resources and deploy capacity to maximize current infrastructure. Pepperdata is an analytics performance management platform that optimizes application and infrastructure performance on large enterprise multi-tenant and cloud systems. Unlike point solutions for individual apps or infrastructure that don't combine data, nor scale to thousands of nodes, Pepperdata automatically recovers resources and maximizes usability of hardware assets across tens of thousands of nodes. It does this by using hundreds of application and infrastructure metrics in five-second intervals, which includes CPU; memory, network, and disk utilization.