
Five Ways
to Optimize
Data Lake
Costs for
Ad-hoc Analytics,
Streaming Analytics,
and ML
Qubole

Five Ways to Optimize Data Lake Costs for Ad-hoc Analytics, Streaming Analytics, and ML

Qubole

? Do You Know What You're Paying for Data Lake Compute in the Cloud?

1 Scale Infrastructure Without the Hassle
- Simplify infrastructure scaling
- Lower TCO with Workload-aware Autoscaling

2 Intelligently Manage Spot Nodes
- Increase Processing Efficiency and Reliability
- Mitigate the Risk of Spot Nodes

3 Increase Administrator Productivity by 10x
- Automate Cluster Lifecycle Management

4 Establish Boundaries for Cloud Finances
- Built-in Financial Governance

5 Work Smarter with Faster Time to Value
- Dramatically Increase Your Quality of Work
- Support More Data and Users with Existing Headcount

! Stop Overpaying for Data Processing in Data Lakes

Do You Know What You're Paying for Data Processing for Analytics in the Cloud?

Conducting ad-hoc analytics, streaming analytics and ML workloads in the cloud offers unique cost, performance, speed, time to value, and accessibility advantages. An open data lake platform adds increased scalability, optimized performance, and greater flexibility.

However, data in the cloud also means greater unpredictability of both workload sizes and the associated costs. Before you know it, your costs can spiral out of control — and you may not notice until the problem gets out of hand. Having the means to control costs and apply specified governance policies is critical for cloud based data lake platform users.

Popular vendors play an important role in your cloud infrastructure for ad-hoc analytics, streaming analytics and ML but these vendors can also end up costing your business considerable time, funds, and resources. These vendors don't offer the depth of controls and policy- based automation necessary to reign in costs.

Qubole helps customers save hundreds of thousands of dollars with built-in platform capabilities and sustainable economics that allow your infrastructure to automatically scale up or down as needed, increase the resiliency of your spot nodes, speed up your time to value, and automate management of the cluster lifecycle.

Qubole®

Read on to discover five reasons why Qubole is the superior open data lake platform for your needs.

Scale Infrastructure Without the Hassle

Ad-hoc analytics and data exploration workloads frequently fluctuate in size and type, making it virtually impossible to accurately predict how much compute they need to complete data jobs in time. Inefficient or manual oversight of infrastructure scaling will result in resource waste and additional costs incurred. Regardless of the type of workloads you're running, you need truly intelligent and automated scaling capabilities to provide as seamless of a transition as possible when demands on the infrastructure change.

Qubole is the only platform that provides true self-service autoscaling where clusters automatically scale to accommodate workloads while optimizing for TCO all time. Unlike the basic autoscaling functionality that many platforms offer, Qubole's workload-aware autoscaling (WAAS) upscales, downscales, and re-balances clusters with complete context of the workload, service-level agreement (SLA), and priority for each job.

SCALE WORKLOADS AUTOMATICALLY

"Our number one reason for choosing Qubole was we wanted to take advantage of cloud economics: only pay for what you use. Qubole's autoscaling and downscaling is definitely a huge cost saver, and the ability to isolate workloads to separate clusters is key to efficient operations."

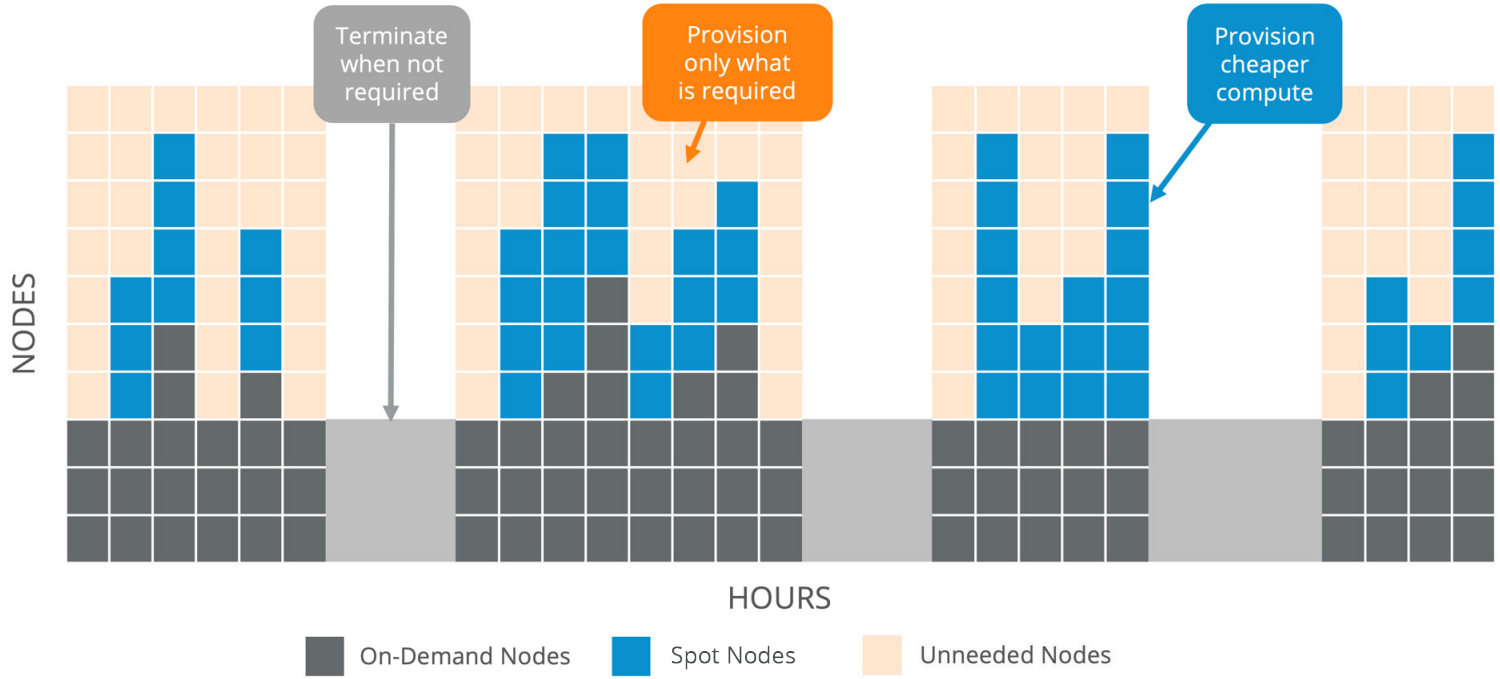
Oskar Austegard
Senior Director of Data Solutions
Gannett

GANNETT

How Qubole Simplifies Infrastructure Scaling

Qubole's workload-aware autoscaling adapts to the bursty and unpredictable nature of ad-hoc analytics and data exploration workloads without resorting to static autoscaling policies and configurations. Unlike common autoscaling that simply relies on CPU utilization or instance goups, Qubole's autoscaling uses machine learning (ML) algorithms that are workload-, SLA-, and job priority-aware. These proprietary ML models rely on several live runtime metrics to determine if and when to scale up or down, including predicting whether or not a job will complete is the allotted time (SLA) or with the assigned priority. Qubole's algorithm is also aware of the state of all nodes within a cluster, which allows for workload balancing within a cluster's nodes prior to scaling.

Qubole's aggressive downscaling leverages intelligent self learning algorithms to balance workloads across active nodes and decommission idle nodes without risk of data loss. This method of downscaling allows for the rapid termination of clusters at any time (rather than at predefined intervals). Aggressive downscaling prevents cluster performance degradation while delivering greater cost savings due to significantly reduced cluster idle time. In addition, Qubole's downscaling improves cluster productivity by drastically reducing the amount of time spent in a state of transition.



Workload-aware Autoscaling: Aggressive Downscaling and Workload Packing

The beauty of Qubole's workload-aware autoscaling is its awareness of workload, job priority, and SLA rather than simply relying on CPU utilization. WAAS automatically estimates the number of nodes required depending on the current workloads running in the cluster.

As a result, at any time the cluster can adaptively scale up or down based on workload demands, ensuring use of the same cluster for workloads of different characteristics.

Qubole's Workload Packing feature drastically improves workload efficiency, increases compute utilization, and reduces processing costs. With Workload Packing, workloads are efficiently packed into a few nodes instead of being spread across all available nodes (which is how autoscaling clusters typically function). The use of Workload Packing frees up the rest of the nodes for downscaling, resulting in higher cluster utilization and cost savings.

Intelligently Manage Spot Nodes

Spot nodes provide the opportunity for organizations to run ad-hoc analytics and data exploration workloads at a reduced price. However, they also come with a critical drawback: they can be taken away at any time depending on market demand. The loss of spot nodes like Amazon Spot instances or Google Preemptible VMs increases the risk of data loss or job restart delays for organizations looking to streamline big data costs without negatively impacting workload performance or productivity.

With Qubole, organizations gain the advantage of leveraging spot nodes without fear of job loss. Qubole increases the resiliency of these compute types with re-balancing, intelligent planning, and rapid job recovery. Qubole also mixes and matches nodes of different types within the same cluster, leading to more reliable clusters.

MANAGE COMPUTE NODES INTELLIGENTLY

"We're down one-third of what we were originally paying, so operations is very happy because they don't have to retune their jobs."

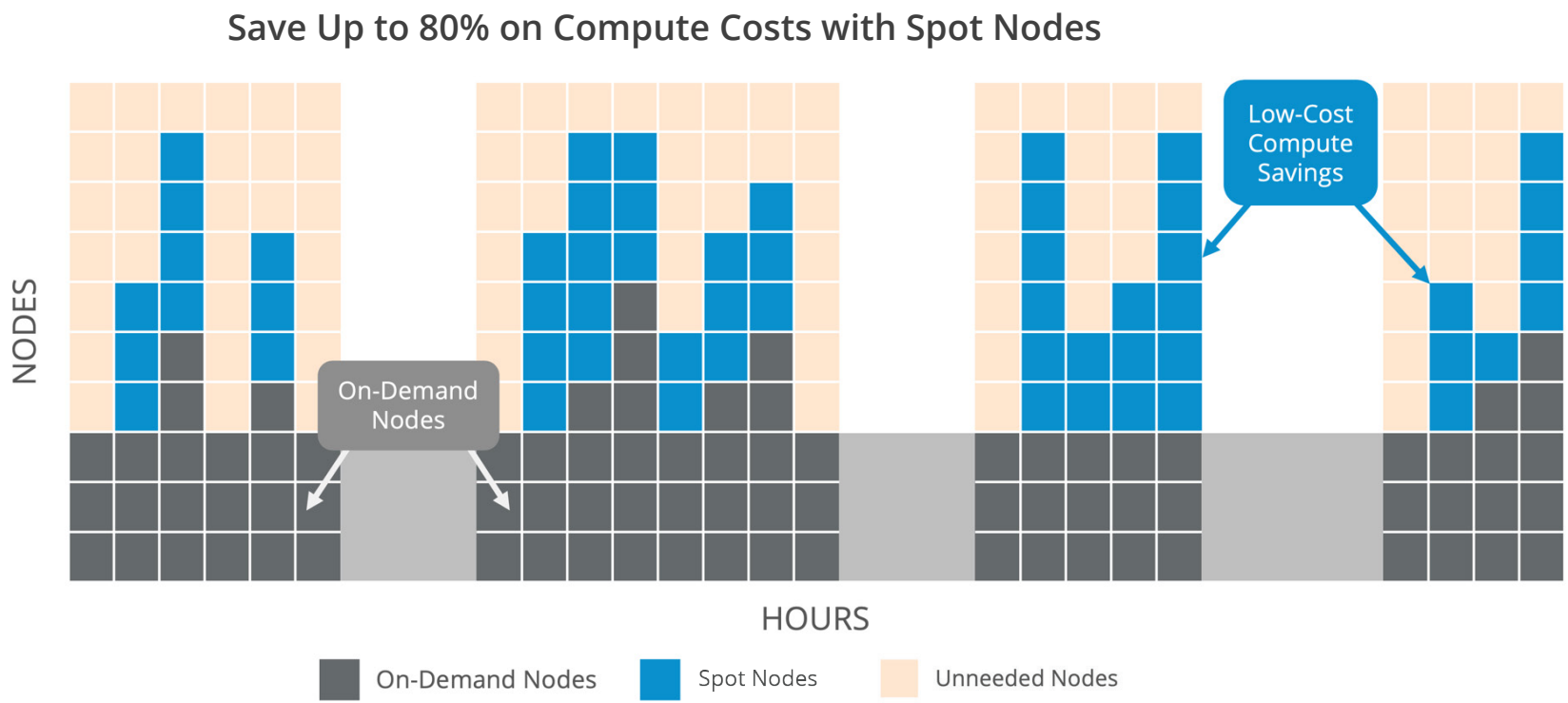
Craig Carl
Director of Solutions Architecture
Bare Metal Cloud Team
Oracle

ORACLE
DATA CLOUD

Mitigate the Risk of Spot Nodes

Spot nodes require careful management when bursty data workloads need infrastructure to scale at a moment's notice. Qubole provides policy-based automation of spot nodes to balance performance, cost, and SLA requirements. Organizations can automate and optimize their usage of spot nodes while maintaining reliability through re-balancing, proactive autoscaling, fault tolerance, and risk mitigation.

Qubole allows users to configure a desired percentage of spot nodes for a cluster — and makes the effort to maintain that ratio as nodes are added, deleted or lost, and as market availability varies. Qubole addresses the volatility of spot nodes by actively monitoring the market, halting the scheduling of new tasks, gracefully decommissioning spot nodes, and reverting spot nodes to on-demand instances when the former are unavailable. Qubole's opportunistic model ensures high resiliency, as the platform is built to bounce back to the desired percentage of spot node utilization despite temporary lapses in availability.

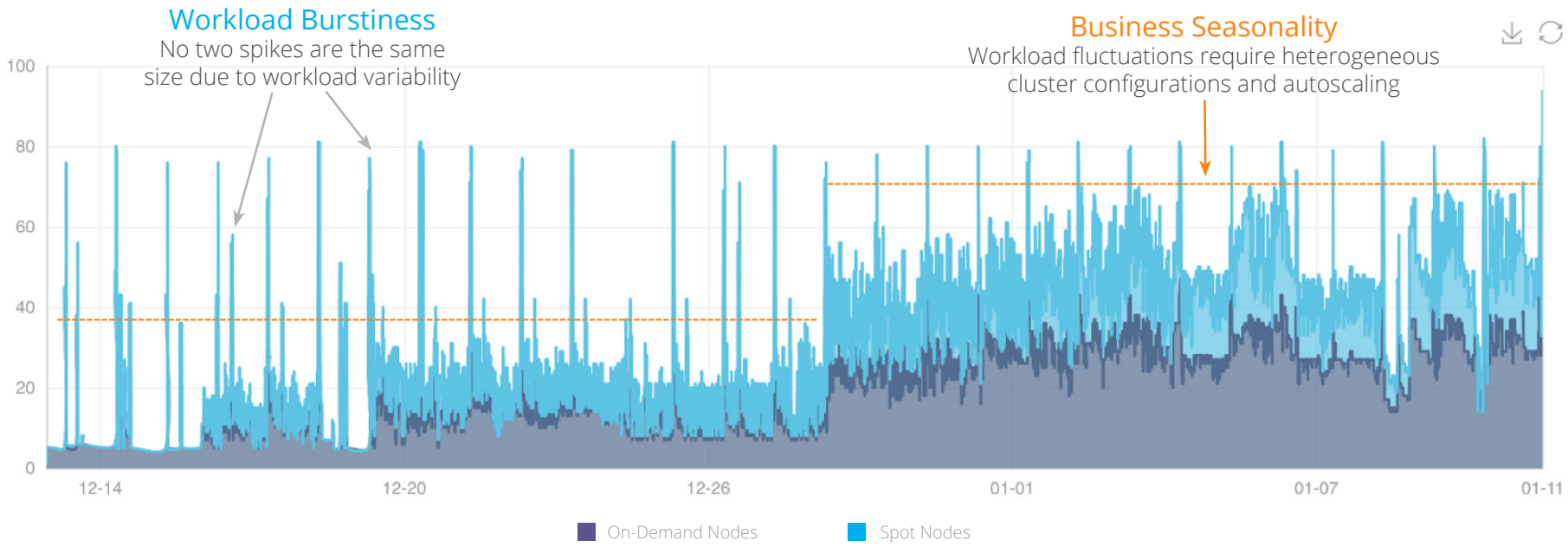


Increase Processing Efficiency and Reliability

Qubole's heterogeneous cluster configuration also increases data processing efficiency and reliability by allowing customers to leverage a mix of on-demand instances and low-cost compute nodes. Unlike homogeneous clusters, Qubole does not require the use of a single node type in a specific cluster. Qubole also delivers maximum cost savings on instance types by provisioning other instance types when spot nodes are unavailable, as opposed to immediately falling back to more expensive on-demand instances.

Qubole's handling of spot node interruption enables users to gain significant cost savings while minimizing the risk of job slowdown and failure. Customers of Qubole can save up to 80 percent on cloud compute costs by leveraging the features mentioned above.[1]

[1] APN Qubole Blog: Up to 80% Savings with AWS Spot Instances



Increase Administrator Productivity by 10x

Most platforms for data lake require detailed oversight, manual configurations, and significant technical support. But data teams must address demands using a different approach that allows them to rapidly scale and automate many burdensome platform administration tasks.

Qubole does just that by efficiently automating all major functions of a cluster's lifecycle. It eliminates manual hassles associated with configuring and managing cloud infrastructure and frees up data teams to focus on more impactful work. Customers who depend on Qubole experience significant improvements in the amount of data and users they can support. With Qubole's Cluster Lifecycle Management, one administrator can support 200 or more users and process 10 times the amount of data that their previous infrastructure supported.

"The savings from Qubole makes our data engineering team much more productive. Our data engineering team moved away from doing routine maintenance and management work to focusing on serving our customers' needs."

Lei Pan
Director of Engineering, Cloud Infrastructure
Nauto

Automate Cluster Lifecycle Management

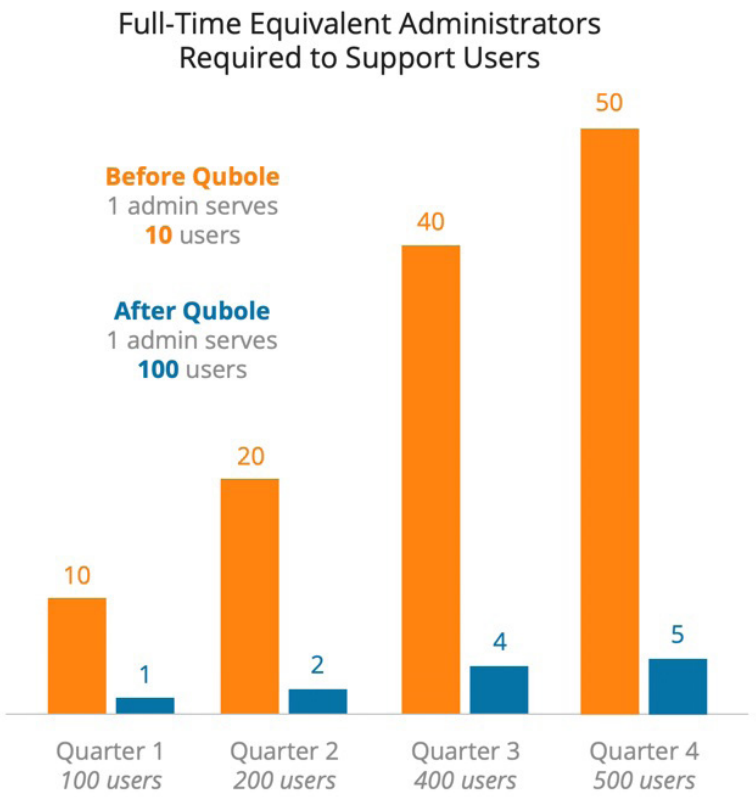
Qubole provides automated platform management for the entire cluster lifecycle: configuration, provisioning, monitoring, scaling, optimization, and recovery. The platform maintains cluster health by automatically self-adjusting based on workload size as well as proactively monitoring cluster performance.

Qubole also eliminates resource waste by automating the provisioning and deprovisioning of clusters, and automatically shutting down a cluster without risk of data loss when there are no active jobs. These decisions are based on granular-level details (like task progression) and occur autonomously in real time to avoid overpaying for compute, interrupting active jobs, or missing SLAs.

Customer in Action: Global Media Company

Eliminating Heavy-Touch Support for Analytics

- Data infrastructure issues due to slow or failing clusters and poor workload performance slow down data analytics
- Heavy-touch support is needed from data engineers, data admins, infrastructure admins (typically 1 admin per 10 users)
- Need more FTEs (e.g., experts on Apache Spark, data engineers for new data pipelines and engines)



Establish Boundaries for Cloud Finances

As ad-hoc analytics, data exploration and data engineering workloads increase in variety, companies are struggling to get a grip on out-of-control cloud costs. Resource waste has become a critical issue for organizations' increasing data workloads and making cloud data lakes more accessible to different teams. Despite allocated budget, organizations may find themselves with out-of-control cloud compute bills due to regularly fluctuating workload demands.

Predicting workloads that are inherently unpredictable is impossible — but with adequate financial governance, companies can set controls and policies that help make demands on their bud- gets as stable as possible. Qubole provides a rich set of financial governance capabilities like Qubole Cost Explorer that help you regain control of finances through policy controls and automation.

"Qubole helped prevent us from making bad decisions that cost the business tens or hundreds of thousands of dollars."

Robert Barclay
VP of Data and Analytics
ReturnPath

Return Path

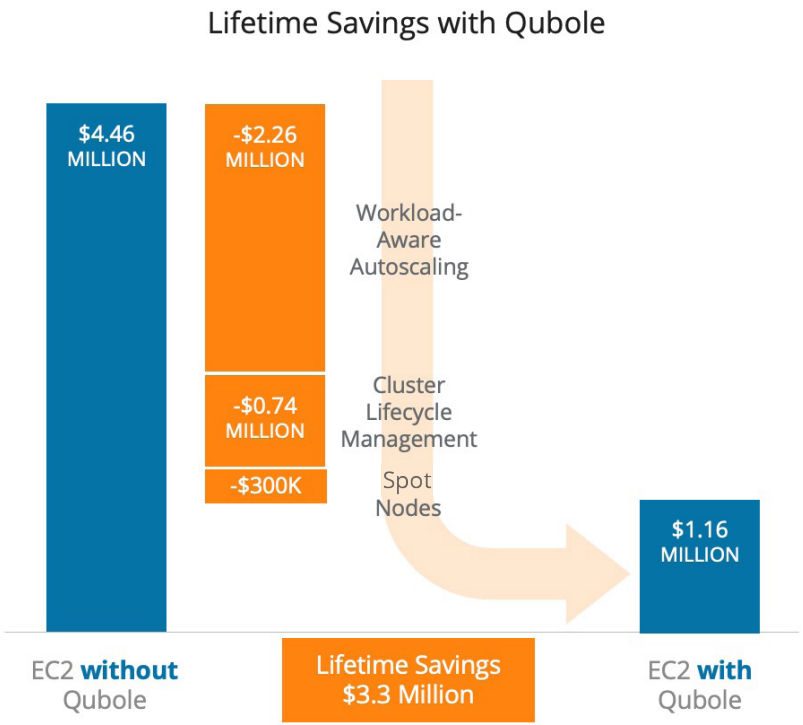
Built-in Financial Governance

To reduce the financial risk associated with unpredictable data processing expenses, businesses today require more advanced financial controls like those available with Qubole. Data teams can continually reduce costs based on policy, preference, and autonomous self-learning.

Qubole's built-in financial governance capabilities provide immediate visibility into platform usage costs with advanced tools for budget allocation, showback, and monitoring and controlling cloud compute spend.

Qubole provides powerful automation that allows administrators to control spend by optimizing resource consumption, using lower-priced resources, eliminating unnecessary resource consumption, and throttling queries based on monetary limits. Custom-configurable controls and insight into key sources of spend offer additional measures to oversee spend.

Qubole also provides governance through intelligent automation capabilities like workload-aware autoscaling, intelligent spot management, heterogeneous cluster management, Qubole Cost Explorer and automated cluster lifecycle management.



Customer in Action: Travel Industry Leader

Cloud Cost Savings with Qubole

Workload-aware autoscaling
+ Cluster lifecycle management
+ Spot nodes

$4.46 million without Qubole
$1.16 million with Qubole

Qubole saved $3.3 million
74% cloud cost savings

Work Smarter with Faster Time to Value

Data processing in the cloud delivers incredible value. Unfortunately, the process can be a daunting and tenuous one. Obstacles such as setup, customization, and use case prioritization frequently arise — delaying your team's ability to respond to users' needs.

To fully utilize data lake, organizations require a platform that allows data users to easily access the tools, engines, and frameworks they need to work productively.

With Qubole, you can address new requests in days — not weeks or months — and begin immediately leveraging data and insights to derive business value. The platform achieves this by delivering self-service access to different data sets in your data lake across groups with unique data demands, enabling many individuals to work simultaneously on a multitude of ad-hoc analytics, streaming analytics, and machine learning projects.

WORK SMARTER AND FASTER

"If it wasn't for Qubole, we would have probably been delayed months to a year in embarking on our big data journey. We would have missed all the insights from the data — insights that have been a strong driver of so many of our growth strategies."

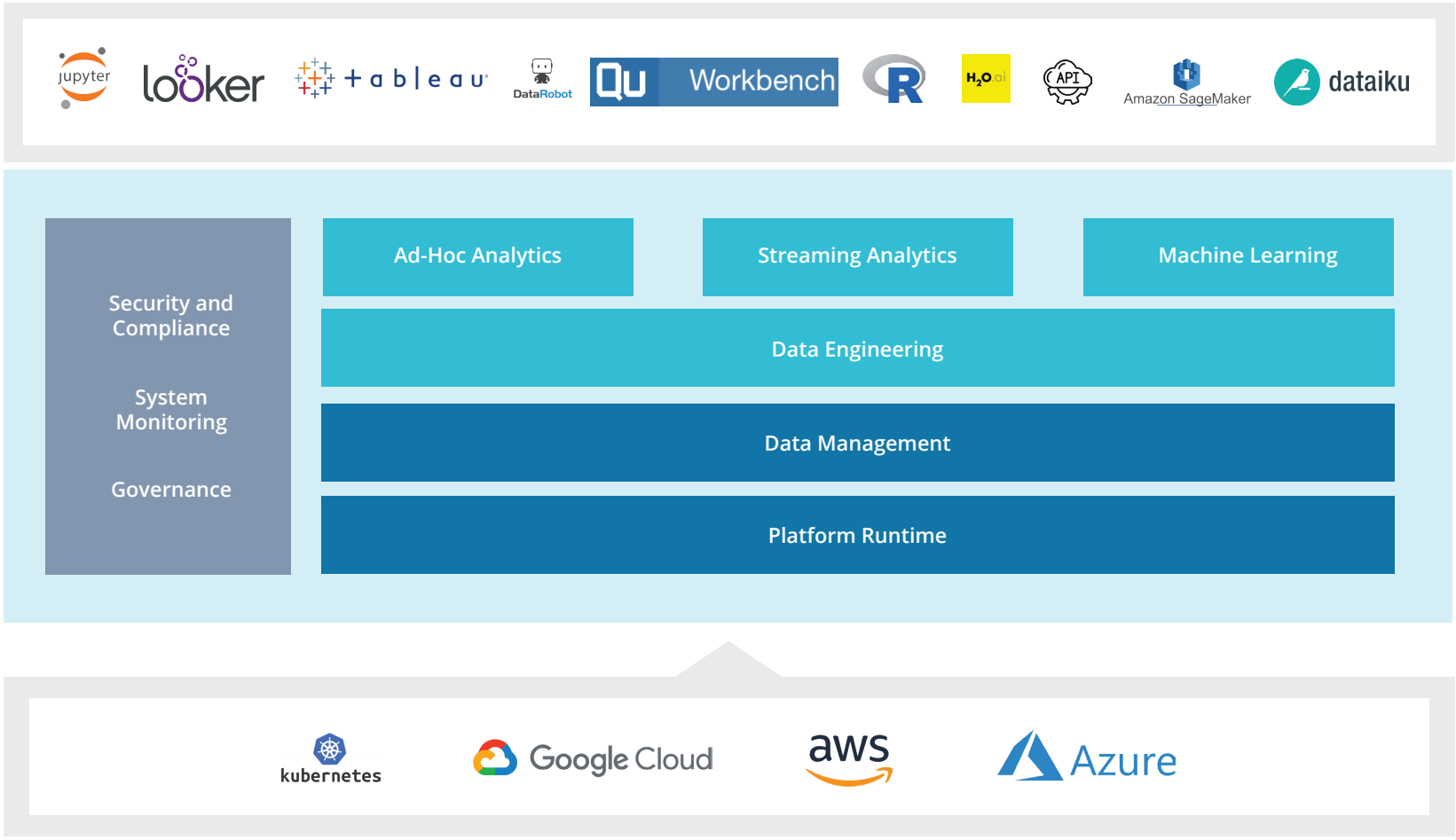
Barkha Saxena
VP of Data and Analytics
Poshmark

POSHMARK

Qubole Dramatically Increases Your Quality of Work

While other vendors cater primarily to only one type of user, Qubole offers a unified interface for data engineering, data analysis, data science, and administration. .

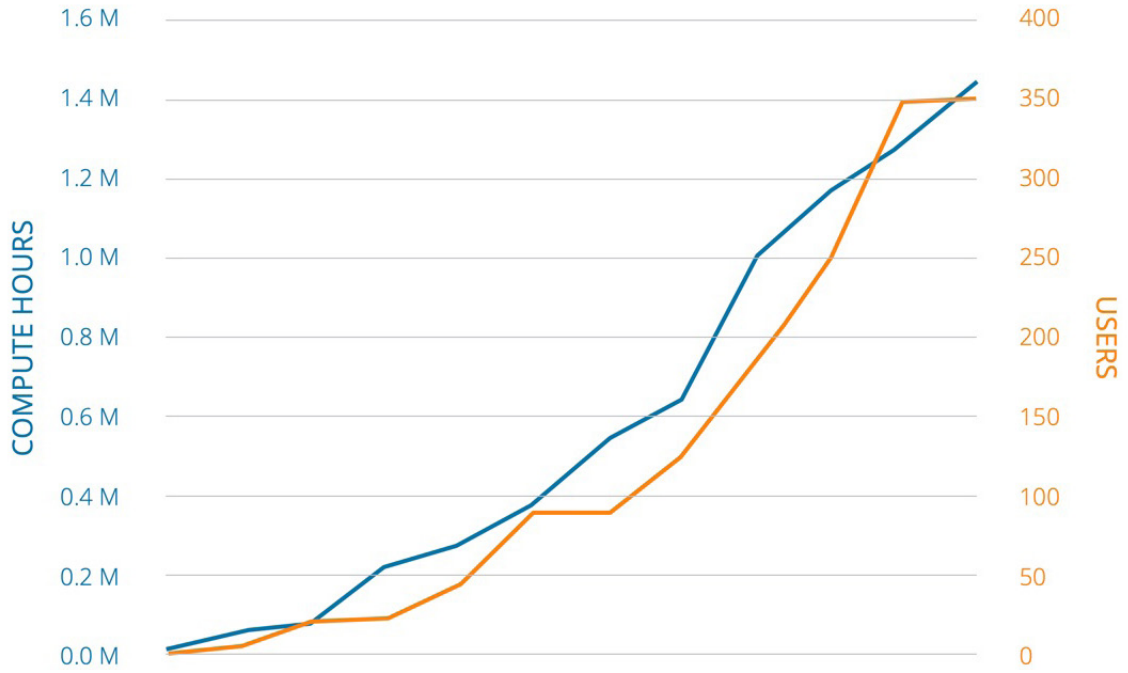
Qubole is the open data lake company that provides a simple and secure data lake platform for machine learning, streaming, and ad-hoc analytics



Support More Data and Users with Existing Headcount

In support of expedient access to data, no other platform provides the openness and data workload flexibility of Qubole while radically accelerating data lake adoption, reducing time to value, and lowering cloud data lake costs by 50 percent. Qubole's platform provides end-to-end data lake services such as cloud infrastructure management, data management, continuous data engineering, analytics, and machine learning with near-zero administration.

On average, Qubole customers are able to onboard over 350 users in months, and use an average of 1.5 million compute hours across multiple use cases. As shown in the diagram below, organizations who use Qubole see drastic increases in compute usage across multiple workloads.



With Qubole, workloads and access to them grows exponentially while maintaining existing budget and headcount

Stop Overpaying for Data Processing in Data Lakes

The growth of data usage for of ad-hoc data analytics, streaming analytics, and ML may be well understood, but what remains uncertain and thus completely unpredictable is when and how often a company's needs for data processing will spike or fall — and with it, their costs.

Instead, organizations must rely on controls, automation, and intelligent policies to govern data processing and attain sustainable economics. Qubole platform helps you regain control of costs and succeed at your initiatives without overpaying.

Qubole