Google Cloud

Qubole

Migrating Your Data Lake to Google Cloud

Buyer's Guide

Overview

Creating a self-service data culture within an organization can seem daunting. You need to meet stringent technical and structural requirements while maintaining robust security and compliance controls. All these requirements become much easier to achieve with an Open Data Lake Platform and Google Cloud Platform. In this buyer's guide, we look at the efficiency and agility an organization can achieve by adopting the Qubole Open Data Lake Platform and Google Cloud Platform.

Whether you are evaluating cloud for your data lakes or already running analytics, streaming, or ML workloads in the cloud with more on-premises workloads going to cloud, this guide is for you. It gives you an overview and working checklist of key considerations for an Open Data Lake Platform for these workloads to migrate to Google Cloud Platform.

Data Lakes and Public Cloud

Organizations looking to implement a self-service data culture increasingly look to the public cloud's as-a-service model for their software infrastructure. Data lakes are no exception. Google's big data technology innovations plus next-generation breakthrough services and frameworks as-a-service help transform businesses with powerful data. A lot of this data and workloads are on-premises and require migration to data lake on Google Cloud.

However, the migration requires more than a simple lift and shift of existing on-premises applications and workloads. A cloud-first re-architecture is necessary for any organization that is looking to implement a self-service data-driven culture.

5 Reasons to Migrate Data Lakes To The Google Cloud

The Google Cloud Platform is uniquely suitable for building a self-service data analytics, streaming analytics, and ML workload data lake with an open data lake platform to manage it.

Scalability

One of the most significant advantages of the Google Cloud Platform is the ability to expand infrastructure to meet the needs of the organization quickly. Ad-hoc analytics, data exploration, streaming, and ML workloads can be huge and bursty. These workloads are challenging to run using on-premises infrastructure because the ability to scale is limited — workloads can grow only to the capacity of the physical infrastructure already available. It's challenging to grow the on-premises infrastructure quickly and cost-effectively. With limitations on infrastructure scalability, organizations often find themselves compromising on data. They use smaller data sets, resulting in inferior models and, ultimately, less valuable business insight. With the scalability of Google Cloud, organizations use very large representative data sets to test their hypotheses. The cloud eliminates the limitations that difficult-to- scale on-premises infrastructure places on the organization.

Elasticity

Elasticity in Google Cloud Platform means that organizations can provision or de-provision resources (compute, storage, network, and so on) to meet real-time demand of their data lake. The elasticity also considerably simplifies and speeds up operations. If users need more compute, they spin up more compute instances. They can change the capacity and power of these machines on the fly. Nothing is fixed, which leads to greater agility and flexibility. The operation's overhead dramatically decreases because infrastructure is altered on-demand in real-time.

Self-Service and Collaboration

Organizations can choose from a plethora of Google Cloud services for their data lakes and can turn them on/off depending on their immediate requirements. Google Compute, Preemptible VMs, Cloud Storage, Cloud SQL, and Google Networking services along with Big Data technologies such as BigQuery, Google Data Studio make organization incredibly agile and increases collaboration. These services are essential to the success of any organization by making data lake adoption faster, leading to faster time to value.

Lower Cost

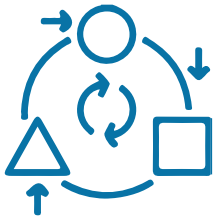
Google Cloud is more cost-effective than on-premises infrastructure for data lake initiatives. There are two reasons. First, the fee is calculated on a usage model on a per-second basis rather than a software-licensing one. Second, the operational costs are much lower because an IT operations staff that traditionally managed and maintained the infrastructure can now focus on other important initiatives. Moving to the Google Cloud Platform boosts the productivity of IT personnel.

Near-zero Administration

Organizations require near-zero administration for their data lakes. Google Cloud provides infrastructure orchestration tools that automate the installation, configuration, and maintenance of clusters. Also, organizations get complete visibility into how Google Cloud resources utilization like machine computing hours, network and storage usage, I/O, and so on. In addition to ensuring fair sharing of multi-tenant resources, the Google Cloud monitoring tools allow organizations to tie usage costs to business outcomes and therefore gain visibility into their return on investment (ROI).

Six Benefits of Using Open Data Lake Platform with Google Cloud

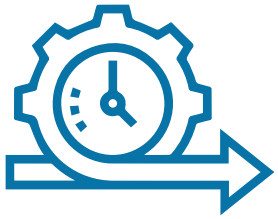
Google Cloud's openness combined with Open Data Lake Platform leads to the following main benefits for ad-hoc, data exploration, streaming analytics, and ML data lake workloads:



Adaptability

Qubole Open Data Lake Platform on Google Cloud adapts seamlessly to changing workloads and business requirements. The elasticity of Google Cloud allows data teams to focus more on managing data and spend less time managing the data platform. Data teams can scale clusters up and down as needed or rely on Qubole Open Data Lake Platforms for complete cluster lifecycle management to automatically scale clusters up and down to match query workloads.

Qubole Open Data Lake Platform allows users to select the instance type that is best suited for a given workload and gives them access to an assortment of different engines — Apache Hive, Apache Spark, Presto, and more — depending on the use case.



Agility

On-premises solutions frequently require six to nine months to implement, while Qubole customers on Google Cloud begin querying their data on average within 2.7 days.* With such a low startup time, business teams spend a majority of their time and resources building applications, running queries, and extracting actual business value as opposed to setting up and managing infrastructure.

Qubole, Open Data Lake platform, allows teams to iteratively determine the best performance and cost and make adjustments as needs change. By moving to Qubole's platform, users can adjust and optimize the configuration, such as the machine type or cluster size. On-premises solutions do not give users this option, meaning they're stuck with what they bought and deployed.

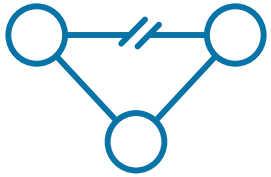
Cost Optimization on an Ongoing Basis

Data Lake workloads are compute-intensive and quickly become very expensive as data lakes expand year over year. With on-premises solutions, organizations have to plan and buy infrastructure (build capacity) for peak usage. Whereas on the Google cloud, organizations can scale compute as needed and only pay for what they use on a per-second basis.

The Open Data Lake Platform helps you optimize cost by leveraging the preemptible VMs offered at a discount compared to on-demand VMs. It significantly reduces the cost of running data workloads or increases an existing application's compute capacity.

Geographic Reach

With Open Data Lake Platform and Google Cloud, organizations have a choice in where they can store their data. The decision is based on various factors, such as data origination, organizational policies, and regulatory and compliance frameworks.



Fault Tolerance, Resilience, Disaster Recovery

Open Data Lake Platform, combined with Google Cloud, is more fault-tolerant and resilient than on-premises solutions. The combination allows enterprises to recover more quickly in the event of a disaster as the number and types of VMs available are at the scale of tens of thousands. If there's a node failure or issue with a cluster, teams can seamlessly provision a new node or spin up a cluster in an alternate location. By built-in failsafe mechanisms in Qubole Open Data Lake Platform, data teams spend less time on maintenance and can focus more on business needs.

Enterprise Grade Security

Open Data Lake Platform provides enterprise-grade security, regulatory compliance for enterprise data. It provides granular read-write capabilities and granular access to data across the data lake on Google Cloud.

When organizations migrate workloads from on-premises to Google Cloud Platform, they get to take advantage of all Google Cloud has to offer without suffering any of the traditional limitations of an on-premises solution. This newfound freedom allows data teams to get to the real work of expanding the number of active users, thereby enhancing the analytic value of the data.

Data Storage, Cost Savings, & Ecosystem Considerations

Before we discuss the different architecture that organizations take to move data lakes to Google cloud, we need to understand cloud data storage better. Reduced data storage costs, elastic computing, and ecosystem are critical reasons for running big data on the Google Cloud.

Storage

One typical storage pattern is to store HDFS data blocks in Hadoop clusters using local instance storage. The issue with using local instance storage is that it's ephemeral. If a server goes down, whether it is stopped or due to failure, data on instance storage is lost. This can be metadata, schemas, and result sets, which can ultimately set back the user's job completion schedule and create risk. Qubole's Open Data Lake Platform on Google Cloud protects users against these common storage problems.

Autoscaling

Autoscaling in a big data environment is different from autoscaling in a transactional, short-running job in a web server environment. Look closely at how autoscaling works for long-running, bursty big data jobs.

Automated Support of Preemptible VMs

Google Cloud Preemptible VMs represent excess capacity and are priced at up to 80 percent discount from on-demand instance prices. By setting a simple policy, such as "bid up to 100 percent of the on-demand price and maintain a 50/50 on-demand to Preemptible VM ratio", automatically manage the composition and scaling of the cluster, look for heterogeneous cluster support, enabling the inclusion of multiple instance types for nodes within a cluster. By casting a wider net of instance types, you can take advantage of the broader Preemptible VMs market and pricing efficiencies, for example, substituting one extra-large node for two large nodes if it costs less.

Ecosystem

Organizations not only require infrastructure resources from their cloud provider but also look for specific use case services perceived as the cloud providers. Google Cloud provides a plethora of services for data analytics such as BigQuery, Google Data Studio, AI/ML services for organizations looking to run their Ad-hoc Analytics, and ML workloads.

Data Lake Architectural Approaches

Typically, data lake migrations to the Google Cloud fit into one of these three architectural styles:

Lift and Shift

In this style of migration, the organization simply replicates its on-premise clusters into the cloud and continues to own its software stack. The cloud is used to achieve OPEX vs. CAPEX financial advantages of rental vs. purchase and to relieve the business from purchasing, operating, and supporting hardware. None of the agility advantages of Google Cloud computing and underlying architectural efficiencies are achieved.

Lift and shift strategies make sense for always-on workloads as long as the organization owning the data lake are tracking the cloud resource consumption and ensuring that it is provisioning the right amount of resources.

However, many organizations utilize multiple engines on top of Hadoop for different use cases of data science, ETL or BI (Hive, MapR, Spark, Presto, etc.). With a lift and shift architecture, each software must be run on its cluster sized for peak capacity making sharing of resources impossible and making this a very high-cost option. Further, with Lift and Shift, clusters are not optimized for BI query environments, which 75% of big data organizations now support. Sophisticated scheduling is not available, opening up issues of individual users or queries consuming the clusters' resources without regard to service agreements.

Lift and Shift architecture for migration works in the short term when an organization is making its initial move to the cloud. However, this architecture does not fully take advantage of the scalability, flexibility, and cost efficiencies the cloud has to offer. Over time as more users are on-boarded, cloud costs can significantly add-up -- substantially inhibiting experimentation.

Lift and Reshape

In this style, Google Cloud's underlying infrastructure efficiencies are adopted. This is the minimum "right approach" architecture for most. The full benefits of the cloud begin to materialize when an organization adopts a workload-driven approach rather than a capacity-driven approach that takes full advantage of the cloud's elasticity. With lift and reshape, IT can move from the role of provisioning expensive "what if" capacity to become a facilitator of business impact.

With lift and reshape, the organization migrates its data lake to Google Cloud. They achieve the scalability and cost benefits of separating compute from storage. They can control and manage cluster costs and take advantage of the wide range of managed compute and storage options available. They can take advantage of Google Cloud's rules-based autoscaling, which is based on CPU utilization and other pre-configured metrics but is not optimized from a workload perspective. Preemptible bidding for clusters can be performed but is neither optimized nor automated.

With the lift and reshape architecture, IT is responsible for ensuring support for all tools and technologies the data teams need, while continuously optimizing the cloud infrastructure as new users are on-boarded. This process helps get started faster but can get very cumbersome within a short time. This happens as the number of users and their requirements grow while tools and technologies are continually changing.

Rip and Replace with Open Data Lake Platform

This style builds on top of the lift and reshapes cloud data lake adoption. It adds advanced features explicitly built to optimize costs and cloud computing for data lake operations. Open Data Lake Platform with Google Cloud ensures workload continuity, high performance, and more significant cost savings.

Automation of lower-level tasks makes engineering teams less reactive and more focused on improving business outcomes. An Open Data Lake Platform provides greater visibility into performance, usage patterns, and cloud spend by analyzing metadata about infrastructure (cluster, nodes, CPU, memory, disk), platforms (data models and compute engines), and applications (SQL, reporting, ETL, machine learning).

Four key areas addressed by an Open Data Lake Platform on Google Cloud are cluster lifecycle management, autoscaling clusters, automated optimization of Preemptible VMs bidding, and support for heterogeneous clusters.

Automated Cluster Lifecycle Management

Cluster lifecycle management automates the management of the entire lifecycle of Ad-hoc Analytics, Streaming Analytics, and ML clusters. This simplifies both the user and administrator experiences. Users such as data analysts and data scientists can simply submit jobs to a cluster label, and an automated cloud platform like Qubole will automatically bring up clusters. There is no dependency on an administrator to ensure cluster resources. Similarly, administrators no longer need to spend time manually deploying clusters or developing scripts or templates to automate this action. Furthermore, administrators do not need to worry about shutting down clusters to avoid charges when jobs are complete, as this occurs automatically.

Workload-aware Autoscaling

Autoscaling in an Open Data Lake Platform goes Google Cloud autoscaling to optimize for price and availability across available node types. Autoscaling does this while ensuring data integrity and that the required compute resources are applied to meet service agreements. Using autoscaling optimized for data lakes compared to generic approaches has been shown to save as much as 33 percent on compute costs and lower the risks of data loss described earlier.

Automated Optimization Of Preemptible VMs

Preemptible VMs provide an opportunity to save on compute costs. With automated Preemptible bidding, an agent 'shops' for the best combination of price and performance based on the policy you provide. It achieves this by shopping across different instance types, by dynamically rebalancing Preemptible and on-demand nodes, and by considering different availability zones and time-shifting work. Also, replicas of one copy of data are stored on stable nodes to prevent job failures when Google Cloud reclaims Preemptible nodes. Automated Preemptible bidding for Ad-hoc Analytics, Streaming Analytics, and ML has been shown to achieve costs 90 percent lower than Preemptible bidding with on-demand clusters.

Heterogeneous and Multi-tenant Clusters

Google Cloud offers multiple node instance types. Each instance type is priced differently based on availability and demand. Typically, users pick a default node instance type and set up homogeneous clusters. Homogeneous clusters are not very optimal for bursty data lake workloads. The availability of these instances varies considerably and could result in significant delays. Heterogeneity in on-demand and Preemptible nodes allows users to pick the most cost-effective combination for their job.

Functional Areas Migration Checklist

Organizations choosing to migrate workloads from on-premises to Google Cloud should consider an Open Data Lake Platform that activates all available data for all data users and meets all functional criteria listed below.

Functional Area	Qubole + GCP		Current Solution	
Unified experience for Data Science,	Fully integrated: Notebooks; Workbench; Dashboards; Airflow; Table/ Storage Explorer; Scheduler; Presto notebooks; Qubole Drivers;	✓		
Engineering and Analytics	Expert Engineering & Tech Support teams by OSS engine included in pricing when purchased through GCP Marketplace	✓		
Enterprise Open Source Software Support	Workload SLA-based Autoscaling; Cluster auto-start/auto-terminate; Container packing; Rebalancing; Intelligent Preemptible mgmt; Heterogeneous clusters; Granular Cost Reporting	✓		
Automated Financial Governance	Fine-grained Resource ACLs, Dual IAM, Ranger, Hive Authorization, Allows for GCP IAM support	✓		
Enterprise Security & Controls	Sharing of commands, notebooks, clusters;	✓		
Easy Collaboration	Dedicated workbench for simplified collaboration	✓		
Multi-cloud	GCP; AWS; Azure; Oracle	✓		
Day 1 Self Service access	GCP Marketplace (GCP Console coming soon); Data connectors; Example notebooks & commands; Built-in tools & interfaces, less than 90 minutes to deploy	✓		
Easy to use Templates	Notebooks & command samples; Bootstrap scripts; Full Package management	✓		
Cost-effective	Automated Financial Governance; Per-second billing; Pay-per-use;	✓		
Google Cloud integrations	Marketplace; BigQuery storage; GCS; GCE; Shared VPC; Google Console (soon)	✓		
Enhanced Open Source Software	RubiX; Spark, Presto, Apache Beam, Air Flow, Jupyter, Zeppelin, Optimized engines	✓		
Easy Administration & Maintenance	Zero-downtime upgrades; Easy debugging; Cluster labels; Spark, RM, Tez, Presto UIs; Online/Offline Logs; Command Status & History, Autocaling clusters	✓		

About Qubole

Qubole is the open data lake company that provides a simple and secure data lake platform for machine learning, streaming, and ad-hoc analytics. 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 is trusted by leading brands such as Expedia, Disney, Oracle, Gannett and Adobe to spur innovation and to transform their businesses for the era of big data. For more information visit us at **www.qubole.com**

TRY QUBOLE FOR FREE ON GOOGLE CLOUD

Start 30 Day Free Trial

Google Cloud Qubole