Qubole®

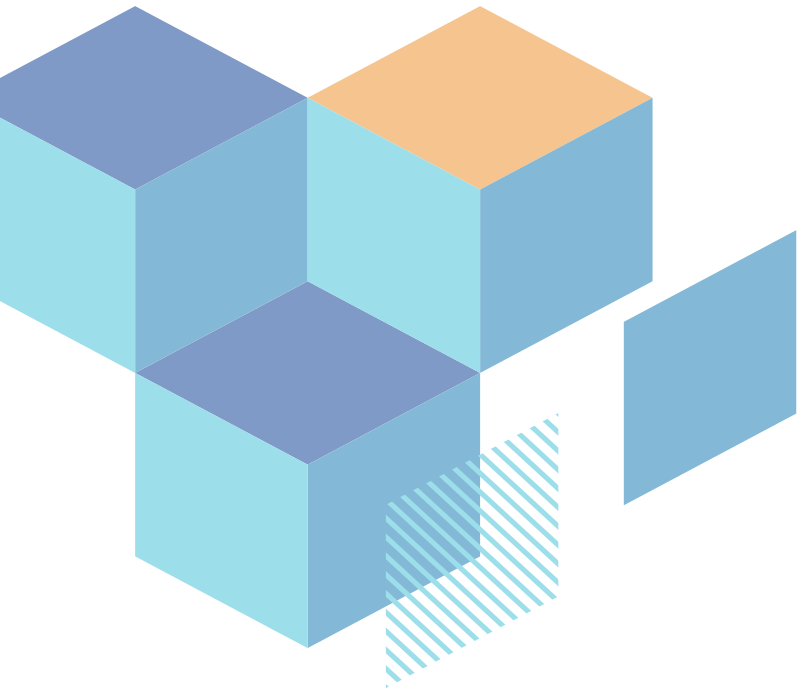
Cloud Data Lake Platforms

Buyer's Guide

Overview

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

Whether you are evaluating cloud big data platforms or already running big data workloads in the cloud, this guide is for you. It will give you an overview of key considerations for an Open Data Lake Platform.




DATA
LAKE

The Rise Of Cloud Computing

Cloud architecture and the as-a-service model is becoming the new norm for software infrastructure and applications in almost every category — and data lakes are no exception. Organizations looking to implement a self-service data culture can do so with greater ease on the cloud.

However, adopting the cloud deployment model 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.

The shift to the cloud is well underway, with 51 percent of companies actively increasing their investment in supporting big data in the cloud according to Forrester Research. The majority of new big data programs are being built in the cloud today (compared to on-premises) on Infrastructure as a Service Providers (IaaS) such as Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform, Oracle, and others.

Data Lakes On The Cloud

The cloud has five properties that make it a uniquely suitable infrastructure for building a self-service data platform:

1. Scalability

One of the biggest advantages of cloud computing is your ability to quickly expand infrastructure to meet the needs of your organization. Today, because of the cloud, huge amounts of infrastructure are now instantaneously available to any organization. Big data workloads can be very large and very bursty. These workloads are difficult to run using an on-premises infrastructure, because the ability to scale is limited — you can grow only to the capacity of the physical infrastructure you have already put into place. It's difficult to grow your 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 the cloud, you can use very large, representative data sets to test your hypotheses. The cloud eliminates the limitations that difficult-to-scale on-premises infrastructures place on you.

2. Elasticity

Elasticity in cloud infrastructure means that you can provision or de-provision resources (compute, storage, network, and so on) to meet real-time demand. Cloud elasticity also considerably simplifies and speeds up operations. If you need more compute, you spin up more compute instances. You can change the capacity and power of these machines on the fly. Nothing is fixed, which leads to greater agility and flexibility. Your operations overhead decreases dramatically because you can alter your infrastructure on demand.

3. Self-Service and Collaboration

In the cloud model of software and infrastructure, everything is API-driven and users can choose compute, storage, and network resources as well as other resources without having to request someone else do it for them. This self-service model makes organizations incredibly agile and increases collaboration. These capabilities are essential to the success of any data lake initiative.

4. Cost Effectiveness

Another important property of the cloud is that it's significantly more cost-effective than on-premises infrastructure. There are two reasons for this: one, the fees are calculated on a usage model rather than a software-licensing one; and two, your operational costs are much lower because you don't need to maintain an IT operations staff to manage and maintain the infrastructure. In fact, moving to the cloud generally boosts the productivity of IT personnel.

5. Monitoring and Usage Tracking

To take full advantage of the cloud and all aforementioned properties and benefits, organizations need complete visibility into how their cloud resources are being used. This brings us to the final cloud property: monitoring and usage tracking. Monitoring capabilities include tools for monitoring machine computing hours, network and storage usage, I/O, and so on. In addition to ensuring fair sharing of multi-tenant resources, the monitoring tools allow organizations to tie usage costs to business outcomes, and therefore gain visibility into their return on investment (ROI).

What makes these properties possible is the cloud architecture. The big data architecture for the cloud is different, and must be built according to very specific architectural characteristics such as separation of compute and storage, multitenancy, security, and more.

Benefits of Cloud Computing for Data Lakes

The cloud's architecture leads to six main benefits for data lake workloads:

Adaptability

A data lake infrastructure on the cloud adapts seamlessly to changing workloads and business requirements. The elasticity of the 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 advanced Open Data Lake Platforms like Qubole that offer complete cluster lifecycle management to automatically scale clusters up and down to match query workloads.

The cloud also allows you to select the instance type that is best suited for a given workload, and gives you access to an assortment of different engines — Apache Hive, 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 the 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.

The cloud also allows teams to iteratively determine the best performance and cost, and adjust as needs change. By moving to the cloud, teams can adjust and optimize the configuration, such as the machine type or cluster size. On-premises solutions do not give teams this option, meaning they're stuck with what they bought and deployed.

Geographic Reach

On the cloud, organizations have a choice in where they can store their data. This decision can be based on factors such as overall convenience, where the data originates from, and any legal issues with how the data is being used.

*Average time between account sign-up to execution of a query

Lower Overall Cost

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 buy infrastructure (build capacity) for peak usage. Whereas on the cloud, organizations are able to scale compute as needed and only pay for what they use.

You can also use computing instances offered at a discount compared to on-demand pricing (called Spot instances in AWS and Azure, Preemptible VMs in Google Cloud) to significantly reduce the cost of running big data applications or increase an existing application's compute capacity.

Fault Tolerance, Resilience, Disaster Recovery

The cloud is more fault tolerant and resilient than on-premises solutions. It allows enterprises to recover more quickly in the event of a disaster. 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 automating this process, data teams spend less time on maintenance and can focus more on business needs.

Enterprise Grade Security

It's a commonly held myth that the cloud is not as secure as an on-premises solution. On the contrary, the cloud is often the more secure option. Cloud providers typically dedicate much more time and resources to security and are able to adopt best practices faster.

MAKE YOUR MOVE TO THE CLOUD

When companies migrate workloads from on-premises to the cloud, they get to take advantage of all the 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 And Cost Containment Strategies

Before we can begin our discussion of the architectural approaches that companies take to move data lakes to the cloud, we need to better understand cloud data storage and approaches. Reduced data storage costs and elastic computing are critical reasons for moving big data to the cloud, so make sure your platform will provide the scalability you seek.

1. Storage

One common 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 your job completion schedule and create risk. Your Open Data Lake Platform should protect against this loss.

2. Autoscaling

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

3. Automated Support of Lower Priced (Spot, Preemptible) Instances

AWS or Azure Spot, and Google Cloud Preemptible instances 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 Spot ratio", some platforms will automatically manage the composition and scaling of the cluster while making bids for Spot instances automatic instead of manual. One way to advance the cost-savings potential of Spot instances further is to 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 greater advantage of the broader Spot market and pricing efficiencies, for example substituting one extra large node for 2 large nodes if it costs less. By taking advantage of these efficiencies as Qubole does, customers can save up to 90 percent from on-demand instance pricing.

Data Lake Architectural Approaches

Typically, data lake migrations to the cloud fit into one of these 3 architectural styles:

1. Lift and Shift

In this style of migration, the business simply replicates their on-premise clusters into the cloud and continues to own their 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 cloud computing are achieved.

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

However, many businesses 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 own 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 -- greatly inhibiting experimentation.

2. Lift and Reshape

In this style, true generic cloud computing is adopted and 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 their data lake to Amazon, Azure, Google Cloud, Oracle, or another IaaS provider. They achieve the scalability and cost benefits of separating compute from storage. They can control and manage cluster cost and take advantage of the wide range of managed compute and storage options available. They can take advantage of the cloud provider's rules-based autoscaling, which is based on CPU utilization and other preconfigured metrics, but is not optimized for big data workloads. Spot/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 can get very cumbersome, as tools and technologies are constantly changing.

4. Open Data Lake Platform

This style builds on top of lift and reshape Cloud Data Lake adoption and adds advanced features built specifically to optimize costs and cloud computing for Data Lake operations. Using a combination of heuristics and machine learning, Cloud Data Lake automation ensures workload continuity, high performance, and greater 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 that should be addressed by an Open Data Lake Platform are cluster lifecycle management, autoscaling clusters, automated optimization of Spot/Preemptible bidding, and support for heterogeneous clusters.

A. Cluster Lifecycle Management

Cluster lifecycle management automates management of the entire lifecycle of Hadoop, Spark, and Presto 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.

B. Auto-scaling

Autoscaling in an Open Data Lake Platform goes beyond generic cloud provider 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.

C. Automated Optimization Of Spot/Preemptible Bidding

Discounted Spot/Preemptible instances provide an opportunity to save on compute costs. With automated Spot/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 re-balancing Spot/Preemptible and on-demand nodes, and by considering different availability zones and time shifting work. In addition, replicas of one copy of data is stored on stable nodes to prevent job failures when the Public Cloud provider reclaims Spot/Preemptible nodes. Automated Spot/Preemptible bidding for big data has been shown to achieve costs 90 percent lower than Spot/Preemptible bidding with on-demand clusters.

D. Heterogeneous Clusters

Cloud service providers offer 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 Spot/Preemptible nodes allows you to pick the most cost-effective combination for your job.

HOW TO EVALUATE A CLOUD DATA LAKE PLATFORM

Companies choosing to migrate workloads from on-premises to the cloud should consider an Open Data Lake Platform that activates all available data for all data users.

Below are a few important criteria to help you evaluate a cloud data platform:

CLOUD-NATIVE

A platform architected for the cloud will take full advantage of the scalability, elasticity, and flexibility of the cloud. Qubole's Open Data Lake Platform is optimized for the cloud.

AUTOMATED OPERATIONS

Automation of repetitive mundane tasks related to administration of the data infrastructure ensures that the data team is no longer the bottleneck. Automation is key to data activation for all end-users of data. Qubole automates the mundane tasks associated with managing infrastructure on the cloud with a simple, easy-to-use interface.

SUPPORT FOR MULTIPLE ENGINES

Data users prefer different engines depending on the workload and tasks they are trying to accomplish. For instance: data engineers prefer to use Hive, whereas data scientists prefer to use Spark. A single cloud data platform with native support for multiple engines allows users to collaborate and ensures that people are not siloed into different tools. Qubole supports more than eight open source big data engines and provides users with the right tool for every job.

ECONOMICS OF SCALING

The infrastructure costs and personnel requirements are something that need to be evaluated for each platform, and planners need to know how costs vary as new users are on-boarded. In addition to actual compute and storage costs, metrics such as the administrator-to-user ratio should also be considered. Qubole customers such as Lyft have dramatically improved their administrator-to-user ratio after moving to Qubole.

PORTABILITY ACROSS CLOUD VENDORS

Cloud technologies are constantly evolving, and you need to be able to pick the best-suited tools and technology for your business. Qubole is available on multiple clouds including Amazon Web Services, Google Cloud, Oracle Cloud, and Microsoft Azure.

EASE OF ON-BOARDING USERS

Greatly reduce time-to-value on data lake projects by ensuring a rapid on-boarding process. Qubole users typically get started in one to two days.

Summary

We hope this buyer's guide to Cloud Data Lake Platforms has been useful at providing valuable information about why data lakes are moving to the cloud, how savings and agility are achieved, and how data platforms offer specific advanced cloud capabilities.

Qubole was founded by the creators of Hive and the leaders of the data lake team at Facebook, who used automation to achieve mass adoption across usage types and data lake technologies enterprise-wide. We invite you to learn more about Qubole — the Open Data Lake Company — at **www.qubole.com.**

READY TO TAKE A TEST DRIVE?

Try Qubole for Free

About Qubole

Qubole is passionate about making data-driven insights easily accessible to anyone. Qubole customers currently process nearly an exabyte of data every month, making us the leading, and industry first cloud-agnostic Open Data Lake Platform. Qubole's Open Data Lake Platform self-manages, self-optimizes and learns to improve automatically and as a result delivers unbeatable agility, flexibility, and TCO. Qubole customers focus on their data, not their data platform. Qubole investors include CRV, Lightspeed Venture Partners, Norwest Venture Partners and IVP.
For more information visit **www.qubole.com**

For more information:

Contact:
sales@qubole.com

469 El Camino Real, Suite 205
Santa Clara, CA 95050
(855) 423-6674 | info@qubole.com

WWW.QUBOLE.COM