Top Reasons for Migrating Data Lake Workloads from On-premises to Cloud

Dhiraj Sehgal, Vihag Gupta, Ashish Kumar

Data-driven enterprises are incorporating the power of data processing, analytics, and ML frameworks like Presto, Airflow, Apache Spark, RStudio and others for their ad-hoc analytics, streaming analytics and ML workloads. But they are also discovering some of the challenges of operating these technologies in on-premises data lake environments. On-premise Data Lake environments have a fixed amount of computing resources. Cloud based data lakes provide the best ROI when underlying compute resources are elastic and enable auto scaling up and down depending on the type, volume and SLAs of workloads. Although cloud data lakes solve data migration challenges with quick to start and easy to use storage migration services, job and workload migration challenges still remain there. Teams still need to put additional effort into provisioning resources and handling uneven workload demand at large scale in real-time for creating data pipelines. Therefore teams need data lake platforms for workload and job migration without disrupting their data lake users. Following reasons should be kept at the forefront as the migration is done.

Near-Zero Administration

Keeping Admin:User ratio to minimal and leveraging open-source framework innovation faster

When users are on-premises, they manage the data lake infrastructure provisioning, deployment and teardown for each workload. This might be manageable when demand is limited and planned well ahead in months. Also, on-premises software may have requirements for certain replicas of the data which leads to having enough capacity for the HDFS and the cluster nodes. This problem is addressed automatically with the cloud based data lake platform. Near-zero administrative overhead is maintained for adopting more frameworks and running the pipelines at an accelerated pace of data processing. The end-to-end data processing is automated starting from data ingestion to data storage and lastly data output for business decisions. The scale to run pipelines while being tailored to each user's requirements is taken care by data lake management platform in the cloud. Data Lake platform addresses pain points of cloud services in following manner from job and workload perspective:

Storage: Cloud Object Stores provided by cloud service providers (CSP) such as AWS S3 or GCP GCS are the storage fabric of the cloud based open data lake. These object stores are infinitely scalable, durable and highly available and eliminates administrative overheads of maintaining storage capacity.

Access: Users have unified access to the data on the data lake using a central metastore with choice of tools and interfaces with the platform. This enables self-service access to the data without continuous administrative intervention.

Tools & Interfaces: Workloads are unique. Different workloads require different tools and interfaces. Open Data Lake platforms abstract the hosting, maintenance and upgrades of such tools and interfaces allowing teams to focus on using them to deliver business value instead of grappling with maintenance and administration.

Compute: Open Data Lake platform provides full automated cluster lifecycle management. Platform uses the CSP's services to provide the best combination of performance and cost as required by the workloads. This eliminates administrative overheads of provisioning VMs and maintaining them. Users also benefit from zero-downtime upgrades as part of the CLCM capabilities. The compute elasticity mapped to the workload requirements is must-have for keeping business continuity during demand surge and realizing cost savings automatically by switching off services during lean times. For example: AI/ML algorithms which need heavy computing in matter of hours instead of weeks to help with end user demand for application should be satisfied with running an open data lake platform on cloud based data lakes. The platform should handle large workloads where data sizes vary and need different compute powers on different days.

With such near-zero administrative overhead, teams can realize faster time to value with new data pipelines and use cases. The openness of the data lake architecture allows quick adoption of new frameworks. Teams can focus on delivering business value with the tools and frameworks of their choice while the data lake platform seamlessly scales the infrastructure to suit the workloads.

Lowering the Cost

Aligning with Workload Need of Storage, Compute, Memory Utilization Curves

Every workload has differing requirements on disk-capacity, processing-power and memory. On-premises environments running Apache Hadoop or Apache Spark directly tie together the compute and storage resources in the same servers, creating an inflexible model where they must scale in lock step. As such, on-premise environments are over-provisioned to cater for growth, processing bursts and resource contention. This means that almost any on-premise environment ends up paying extra dollars for underutilized disk-capacity, processing power and system memory. To attain maximum resource utilization, workloads should not be a function of fixed computing resources; instead the computing resources used should be a function of the workload currently running. Workloads should be free to run whenever and however is most efficient, while still accessing the same shared underlying storage or data lake.

Combination of cloud-native data lake platforms and cloud provider services like Qubole Open Data Lake Platform, AWS Glue, AWS LustreFX and Amazon S3 enable you to decouple and scale your compute and storage independently, while providing an integrated, well-managed, highly resilient environment, immediately reducing so many of the problems of on-premises approaches. This approach leads to faster, more agile, easier to use, and more cost-efficient big data and data lake initiatives. For data lake driven workloads cloud-native data lake platform, it translates to

- **Billing**
  - Pay only for what you actually use with aggressive downscaling and optimized upscaling
  - Automatically use spot instances and heterogenous configurations to reduce cost by up to 50%

- **Adaptability**
  - Best machine configuration for the workload
  - Best engine for the workload

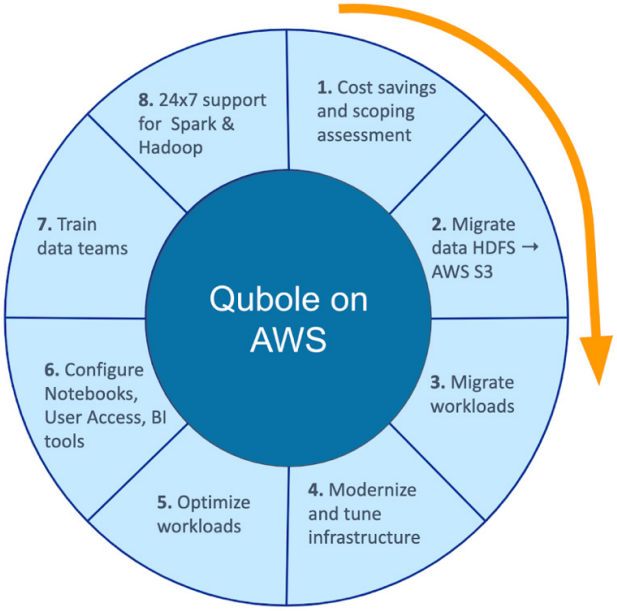
Future proofing with Single Platform

Building underlying platform synergies for ad-hoc analytics, streaming analytics & ML use cases instead of maintaining siloed solutions

With faster innovation around open source frameworks and data being a common foundation for ad-hoc analytics, streaming analytics and ML, siloed on-premises approach is not future proof. On-premises solutions are often architected and deployed with one workload or one open source framework in mind. Cloud resources and cloud-native data lake platforms together provide a single, scalable, future proof platform. It should be able to do workload-aware autoscaling, provide developer tool sets for each type of workload, integrate with use case specific solutions such as Looker, Tableau, RStudio. For cloud-native data lake platforms, it translates to taking care of intermediate steps to get the data to the user seamlessly and always available at fingertips.

How to Approach Migration

Steps to do data lake migration from on-premises to cloud should be bucketed in three main phases with the following steps in them.



Phase 1: Assessment & Migration

Assess

- Initial self-driven, 10-minute cost savings assessment (web-based)
- Live 1hr spend review meeting
- Follow up custom business value assessment:
- Evaluate environment
- Assess data stores, structures, volume
- Review all workloads
- Prepare a migration plan and success metrics

Migrate Data

- On-prem to Cloud:
- HDFS → AWS S3
- Leverage trusted systems integrator (SI), or cloud provider solution architects

Migrate and modernize workload

- Includes data pipelines, commands, scripts, notebooks, dashboards
- Leverage Open Data Lake platform and engage professional services

Phase 2: Modernization

Tune Infrastructure

- Test migrated workloads in Qubole Open Data Lake Platform
- Setup configuration parameters and realize cluster tuning
- Optimize for performance or capacity

Optimize Workloads

- Run performance tests
- Fine-tune engine parameters (Spark, Presto, Hive)
- Optimize workloads for specific cloud
- Do cost analysis and optimization

Configure User Access

- Provide RBAC based access to users
- Configure notebooks and dashboards
- Enable other type of user and application access (e.g., APIs and SDK)
- Configure third- party reporting tools (e.g., Tableau, Looker)

Phase 3: Post Migration

Train users virtually or in-person

- Train data administrators
- Train end users like data engineers, scientists, & analysts
- Keep up to date with documentation and online courses and labs

Support

- Be supported 24X7 with a multi-region technical support
- Tap into technical know-how of subject matter experts in data lake world and cloud to keep abreast with new developments or solve unique problems.

Users who follow the phased approach outlined above with Qubole Open Data Lake platform and cloud provider of their choice have realized the three benefits discussed in the paper.

Qubole can help you get started with Free Trial

CLICK HERE