Qubole Open Data Lake Platform on AWS

Accelerate your data lake adoption, reduce time to value, and lower cloud data lake cost by 50%

Table of Contents

Qubole is an open, simple, and secure data lake platform for machine learning, streaming, and ad-hoc analytics. Qubole on AWS provides end-to-end data lake services such as AWS infrastructure management, data management, continuous data engineering, analytics, and machine learning with near-zero administration. Qubole on AWS delivers:

Unified experience for data science, data engineering, ad-hoc analytics	Native workbench that includes notebooks, dashboards, and a common interface for all commands and tasks. This enables data engineers and data scientists to collaborate using familiar tools, languages (SQL, Python, R, Scala), and data processing frameworks (Apache Spark, Presto, Hive and Airflow).
Low cost and high reliability	Workload-aware autoscaling for optimized upscaling, rebalancing, and aggressive downscaling of clusters with a complete context of the workload, SLA, and priority of each job. Includes intelligent policy-based management of On-demand and Spot Nodes.
Enterprise-grade security	Fine-grained predefined or custom identity and access management roles to separate compute and data access. Qubole also offers role-based access controls for secure collaboration in notebooks and commands.
AWS Native Integrations	Native integration with AWS services like EC2, S3, Sagemaker, Redshift, AWS FSx for Luster.

Why do users choose Qubole on AWS?

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.

Users adopt Qubole Open Data Lake platform on AWS for the following reasons:

Reduce Data Lake Cost by more than 50%

- Automated cluster lifecycle management
- Intelligent spot management
- Heterogeneous cluster management

Near-Zero Administration

- Automated platform management
- Workload-aware Autoscaling
- Insights and reporting
- Built-in AWS specific optimizations

Fast Adoption of Their Data Lakes

- Self-service platform for all users
- Out-of-box tools for data science, data engineering, and analytics
- APIs and pre-built integrations with 3rd party solutions

Unify on Simple, Open and Secure Platform

- Single platform for data ingestion, processing, management, and consumption
- Open & standard file formats, languages and APIs
- Secure and granular access

User Spotlight 1: Malwarebytes

About Malwarebytes

Malwarebytes is a cybersecurity company that produces anti-malware software for a variety of platforms. The company offers consumers free, premium, and enterprise-grade versions of Malwarebytes, which detect, remove, and remediate computer malware. Malwarebytes uses machine learning (ML) and artificial intelligence (AI) to identify and predict emerging threats before they infect machines.

Business Problem

To predict, detect, and neutralize emerging threats, Malwarebytes processes billions of threat telemetry records daily. The company then performs advanced analytics on this data to identify potential threats and runs ML and AI models to determine what action to take to protect its customers.

Malwarebytes formerly relied on a third party on-premises deployment to ingest and process this data. But this system proved inadequate. For example, the pipeline took a few days to complete Extract-Transform-Load (ETL) on one data stream alone. And, queries on the ingested data were painfully slow.

Malwarebytes needed some way to modernize its big data processing to improve turnaround time while also keeping costs down.

Improved Processing Speed and Lowered Costs

Malwarebytes adopted Qubole in concert with Kafka (for ingesting data streams), and an AWS S3 data lake (for data storage). First, it de-coupled compute and storage. Second, “playing by the rules of the game of the cloud,” says Kulkarni— leveraging things like autoscaling (scaling out and scaling up, and being elastic & ephemeral in nature), low-cost compute instances (AWS Spot), and storage (an AWS S3 data lake)— significantly improved the efficiency of the data platform. Today, Malwarebytes uses Qubole to process its data. About 60 to 70% of it is logs, telemetry, and other types of unstructured and semi-structured data that is being processed in Qubole.

Platform’s ability to add and remove compute resources on-demand based on the workload or SLA, and without human intervention—in a matter of minutes has greatly increased the speed at which Malwarebytes processes critical data, directly affecting the company’s ability to detect, predict, and remediate emerging threats.

Qubole aggregates and processes between 20 and 48 terabytes of raw data per day but delivers just 2 to 3 terabytes of meaningful and actionable data. Qubole provides a single framework for processing data more quickly, whether for use in ML models for predictions, in BI applications for business reporting or for GDPR compliance—all with just one full-time administrator plus three senior engineers—a few times per quarter. The result is more powerful insights because they involve better data.

> *“Qubole has really mastered the elasticity component of the cloud. Qubole helped us run our ETL at night, spinning up and spinning down clusters when we needed them.”*
>
> **Manju Vasishta**
> Director of Data Science and Engineering,
> Malwarebytes

Quick ROI

The platform's ROI was quickly revealed: by the meaningful data it helps to discover, which yields more powerful insights. These insights—for example, predictive insights about emerging threats; marketing lead conversion propensity using ML algorithms; behavioral clustering of malware; sentiment analysis of reviews about Malwarebytes products and features on various social media platforms using advanced natural language libraries; drive key decisions that serve the business and its customers well.

Key Takeaway

- Greater data-processing capacity at much lower costs
- Improved efficiency to produce meaningful data and more powerful insights
- Easy user onboarding resulting in high adoption
- Quick, tangible ROI

> *"Qubole has really mastered the elasticity component of the cloud. Qubole helped us run our ETL at night, spinning up and spinning down clusters when we needed them."*
>
> **Manju Vasishta**
> Director of Data Science and Engineering,
> Malwarebytes

User Spotlight 2: Neustar

About Neustar Unified Analytics

Neustar Unified Analytics is an integrated marketing measurement, analytics, and attribution solution from Neustar Information Services, Inc. Neustar Unified Analytics is not a marketing campaign management tool. Rather, it runs alongside its clients' campaign management platforms to measure and attribute overall marketing spend across campaigns. More than 90 Fortune 200 companies depend on Neustar Unified Analytics to assess and improve their marketing investments.

Business Problem Overview

The Neustar Unified Analytics platform helps marketers understand the impact of marketing on key business outcomes, and provides tools to enable them to optimize the allocation of their marketing investments. First, it ingests large volumes of client marketing data from a variety of sources. Then, it applies proprietary algorithms to build a predictive spend attribution model on top of that data. This reveals how the client's marketing spend correlates to revenue—enabling marketers to determine which marketing channels are working, which ones aren't, and what to do next.

To meet the demands of its growing client roster, the Neustar Unified Analytics team needed to confront the issues of variety, volume, velocity, and veracity—often called the "four Vs." At the same time, the team needed to keep operational costs down. For this, it turned to Qubole.

Ensuring Data Veracity

"Models are only as good as the input data," said Peterson. "If your data has lots of gaps, then the model won't be good, no matter what algorithm you use." But most data scientists fail to detect "dirty" data until after they run the model—a typical data science pipeline has data processing, modeling, and scoring stages. Data scientists must then fix the data and rerun the many of their processes—a task that might take weeks or even months, depending on processing speed and capacity. Often, this cycle repeats, compounding the delay. Indeed, "the main reason these things take a long time is because of the reruns," said Peterson.

Neustar Unified Analytics is different. Its machine learning models include a series of pre-checks and post-checks to validate data. "Because we have a very comprehensive set of validation routines that run on Qubole, we're able to isolate problems earlier and avoid these reruns," Peterson explains. As a result, data validation jobs require just one to one and a half run cycles. This allows Neustar to deliver insights to its clients much faster and with the highest degree of confidence.

Keeping Costs Down

In a given month, the heavy compute time needed for most machine learning jobs is 80 to 90 hours on average for each customer. The rest of the time is typically consumed running reports, tuning parameters, and so on—tasks that require considerably less compute power. For this reason, before Neustar Unified Analytics partnered with Qubole, its 400-odd compute nodes per customer were frequently underutilized—with no adjustment in cost. Now, Qubole aggressively—and automatically—shuts down excess capacity during slow periods, efficiently "packing" workloads in fewer nodes. This dramatically reduces operating costs, without compromising performance or delivery times.

Neustar Unified Analytics team has reduced its costs to the tune of 85 to 95 percent over its prior use of other vendor tools with reserved compute instances and administrator-led scaling.

> *"From a performance aspect, we want to be faster and faster...and Qubole fits right into this."*
>
> **Dan Peterson**
> Vice President of Systems Engineering,
> Neustar

Key Takeaway:

- Decreased machine learning model turnaround from six months to three weeks, end to end.
- Reduced model data validation cycle time by more than 62 percent.
- Cloud cost savings by 85 to 95 percent.

> “Qubole is cheaper and much more economical than other vendors...but more importantly, it’s much more stable, and much more high-performing. Qubole offered us the best price for performance, and outstanding support.”
>
> **Dan Peterson**
> Vice President of Systems Engineering,
> Neustar

User Spotlight 3: Publicis Media

About Publicis Groupe

Publicis Groupe is one of the four solutions hubs of Publicis Groupe [Euronext Paris FR0000130577, CAC 40], alongside Publicis Communications, Publicis Sapient and Publicis Health. Led by Steve King, CEO, Publicis Media and COO, Publicis Groupe, is comprised of Starcom, Zenith, Digitas, Spark Foundry, and Performics, powered by digital-first, data-driven global practices that together deliver client value and business transformation. Publicis Media is committed to helping its clients navigate the modern media landscape and is present in more than 100 countries with over 23,500 employees worldwide.

Business Problem Overview

Few years ago multiple teams from the various media agencies merged to form Publicis Media. This merger revealed the need for a central data and analytics platform. "We wanted our agency teams to be able to mine data, but not to have to deal with the operational overhead of managing data infrastructure," explains Darren Smith, who leads the engineering and data teams. "Our intent was to democratize data."

According to Smith, the team's existing data infrastructure "was a bunch of bespoke solutions" that combined AWS Redshift, large monolithic on-premise servers, and various unwieldy traditional technologies. Offering a central data and analytics platform would require both a complete overhaul of this infrastructure and some way to tie all of its pieces together.

A Centralized Platform for Democratizing Data

"The focus of our team was to build a data architecture and infrastructure that would allow our agency teams to move forward in a big data world," says Joe Tan, director of products at Publicis Media. The resulting infrastructure couples a global data lake —which stores large volumes of multiple types of data—with a framework to ingest and process data. In addition to building this data infrastructure, Tan's team had another job: "to provide tools that allow agency teams to really focus on doing analytics for their clients instead of having to worry about data ops and data engineering." Qubole enables agency teams to "work with the data they're used to in the tools and languages they're used to, like Tableau and Presto, or SQL, Python, R, Scala, etc." says Tan. It also helps Publicis Media make data available to users with different skill sets. "It even," says Tan, "gives users the ability to learn how to do more with minimal additional effort." As more and more clients have grasped the potential power of Publicis Media's platform, Qubole has played a key role in helping increase its adoption.

Scaled with Customer Data Demands

Publicis Media handles lots of data for its agency clients. Its data lake stores close to a petabyte of it. Agency clients use this data to run machine learning models for analytics purposes. Scaling to process larger data sets posed a challenge before Qubole. "I regularly walked into offices and ran into someone who'd had a model running for six hours," recalls Tan. Qubole solved this problem by enabling agencies to automatically scale up compute infrastructure for large jobs and to aggressively scale back down when a job is complete to keep costs low. So, jobs that once took six hours to complete can now be finished in mere minutes, with almost 10,000 queries per month on average. In addition, Qubole also supports multi-region data availability without latency—further improving the performance and consistency of Publicis Media's data globally.

> “ *We have had a steady growth rate of one to two agency clients onboarding onto our platform per month," says Tan. "That might not sound like a lot, but a lot of those teams service multiple clients of their own, so it's pretty impactful."*
>
> **Joe Tan**
> Director of products, Publicis Media

Secure Enterprise-grade Data Lake Platform

Data security is top of mind for Publicis Media. Qubole addresses its requirements with regard to single sign-on, strict role-based access control, and agency data isolation, among other security issues. While both Smith and Tan see these features as “table stakes,” Smith acknowledges that, “A lot of vendors don’t support them.”

Key Takeaway:

- A central data and analytics platform that democratizes data
- Ability to manage nearly 1 petabyte of data
- Reduction in the model run time from six hours to mere minutes, with almost 10,000 queries per month on average
- Multi-region data availability without latency
- Easy administration with an administrator-to-user ratio of 3:100
- Support for robust security and compliance requirements

> “*Qubole really meshed well with the overall architecture and design of our data lake. I don’t think we could have found a better platform.”*
>
> **Joe Tan**
> Director of products, Publicis Media

AWS and Qubole Native Integrations

Amazon EC2 Spot

Qubole Open Data Lake platform provides a policy-based way to automate the spot bidding process, allowing data teams to take full advantage of spot instances without devoting resources to managing it.

Qubole uses AWS spot nodes when dynamically adding cluster nodes or as part of the core minimum nodes for a cluster. Users select a maximum bid they are willing to pay for a spot instance. The platform then automatically places bids for them, making the process easy to use. Qubole clusters begin with nodes at on-demand instances and rebalance automatically by switching on-demand instances for spot nodes when spot availability is higher. With this ease of use, the Qubole platform is used for advanced provisioning strategies. Those strategies come in three categories:

- **On-Demand Only:** Auto-scaled nodes that are added will only be On-Demand instances.
- **Spot Instances Only:** Auto-scaled nodes that are added will only be Spot nodes.
- **Hybrid:** Auto-scaled nodes combine On-Demand and Spot nodes. Users are able to choose what the maximum percentage of Spot nodes is.

The platform also has additional built-in intelligence to maximize spot instance usage for the workloads:

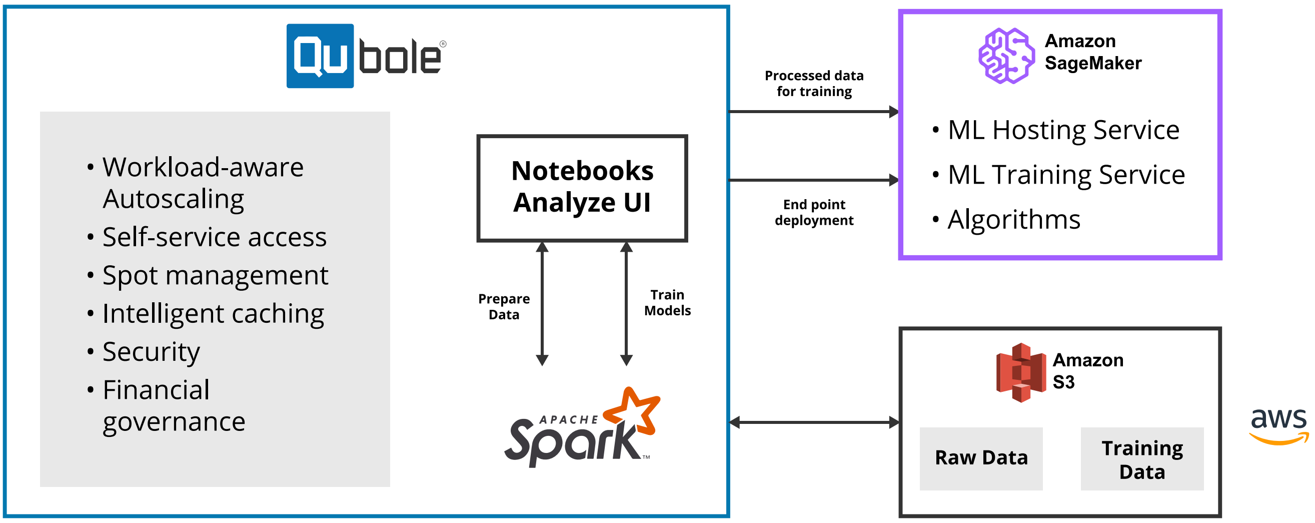
- **Qubole Placement Policy:** Qubole has multiple pricing options for stable spot nodes and volatile spot nodes. Via the placement policy, Qubole spreads out underlying storage across stable and volatile nodes, thereby minimizing the risk of job loss due to loss of a Spot instance.
- **Fallback to on-demand instances after a configurable timeout:** Qubole can automatically fall back to requesting on-demand nodes if spot nodes cannot be provisioned within a configurable timeout period.
- **Intelligent AZ Selection:** Spot pricing can vary by AZ (availability zone), sometimes by up to 15-20%. Qubole can automatically select an optimal AZ based on Spot pricing for the cluster instance type chosen. Currently, AZ selection is only supported for non-VPC clusters.

Amazon SageMaker

The SageMaker and Qubole integration allow enterprise users to leverage Qubole Notebooks and Apache Spark on Qubole to explore, clean, and prepare data in the format required for Machine Learning algorithms. Once the raw data is cleansed and prepared in Qubole, it is used to train ML algorithms in SageMaker. There are 2 ways for users to leverage this integration.

- **Prepare Data and Initiate Training from Qubole**
  Qubole loads data from multiple data sources such as Transactional databases, Data Warehouses, Streaming data, interaction data such as clickstreams, social media feeds sensor data, log files, and more. Users read their data into Qubole Spark data frames, use Qubole Notebooks to transform, cleanse, and prepare the data. Once the data is stored back on Amazon S3, the users initiate model training — from Qubole — using the estimator in the SageMaker Spark library. This initiates ML training in SageMaker, builds the model, and creates the endpoint to host that model.

- **Prepare Data in Qubole from SageMaker Notebook**
  Alternatively, SageMaker users enhance the SageMaker data processing capabilities by connecting a SageMaker Notebook instance to Qubole. Data scientists use Apache Spark to process and prepare data at scale with Qubole. Qubole Open Data Lake Platform greatly reduces the cost of computing by consuming less compute and/or consuming cheaper compute. With this integration, data scientists use Qubole to cleanse and prepare (transform, featurize, join, etc.) prior to ML training in Amazon SageMaker.



Qubole High Level Architecture

AWS FSx for Lustre

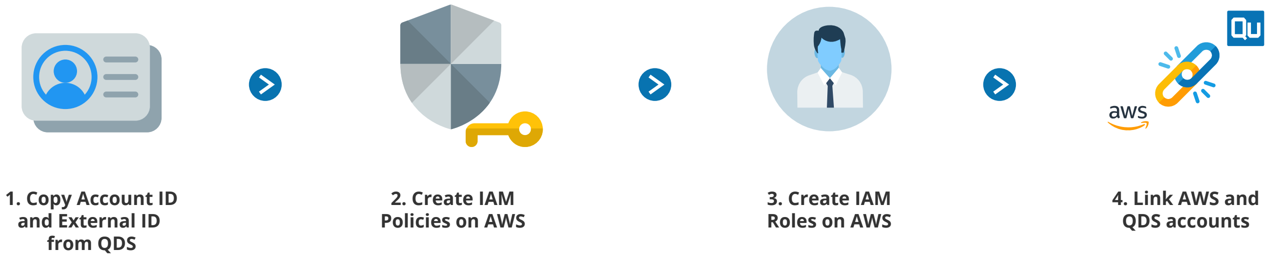
AWS FSx for Lustre and Qubole Open Data Lake Platform together reduce user's compute cost and minimize intermediate data loss while running workloads. Users do not pay to maintain idle AWS EC2 instances and also not worry about intermediate output (shuffle data) loss due to spot nodes interruption. Qubole uses Amazon FSx for Lustre to store and process intermediate data through its parallel, high-speed file system. By doing so, users no longer need to retain idle EC2 instances to store this intermediate data. Instead, Amazon FSx for Lustre allows them to re-use the data otherwise normally held within EC2 local storage.

AWS Glue

Qubole and AWS Glue provide users with flexibility and choice of a unified shared metastore with a cdata lake platform. Users use Glue's data crawlers to scan and classify data, extract schema details, and build the data catalog. Qubole's platform is configured with this catalog as the metastore and shared across your AWS accounts, applications, and services. With Qubole's multiple open source frameworks support, users run Hive, Presto queries, and Spark jobs leveraging this catalog. Alternatively, users can continue using their existing or Qubole-hosted metastore and synchronize it with the Glue Data Catalog.

Accessing via AWS Marketplace

Qubole makes it easier for users to access, manage, monitor and govern their data in S3 data lake with Open Data Lake Platform. Users can subscribe and access the platform through AWS marketplace with automatic account setup, AWS authentication and simplified user onboarding in less than an hour with their data.



1. Copy Account ID and External ID from QDS
2. Create IAM Policies on AWS
3. Create IAM Roles on AWS
4. Link AWS and QDS accounts
aws
Qu

Learn More

For the latest information about our product and services, please see the following resources:

- Qubole Whitepapers
- Qubole Case Studies
- Qubole Technical Documentation

You can visit the AWS Marketplace anytime to get up and running with Qubole!

TRY QUBOLE IN AWS TODAY!

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**