CASE STUDY

DataXu Uses Qubole to Make Big Data Cloud Querying, Highly Available, and Efficient

About DataXu

DataXu develops and delivers a suite of cloud-based marketing applications that enable marketers to better understand and engage their customers. Working with enterprise customers across the globe, the company is no stranger to leveraging Big Data; having lots of experience with an on-premise Hadoop cluster as well as Amazon EMR. In fact, DataXu's ability to put Big Data to work is one of the things that have propelled the company's growth, earning it the Inc. 500 award for the fastest growing advertising and marketing company.

But managing these on-premise deployments has never been easy. Provisioning clusters, maintaining Hadoop distributions, and adding machines for additional capacity, and upkeep of adhoc clusters is very time consuming. And configuring the Hadoop system can result in run-time issues which negatively impact availability of the system resources.

Managing adhoc cluster availability is challenging. For example, when a user issues a query, it can take as long as 15 minutes to start a cluster and requires manual engineering assistance. "We have business teams spread across the world with varied skillset who want to query data live at any time" comments DataXu's Vice President of Technology, Yekesa Kosuru.

Auto-scaling with Qubole requires no manual intervention. Otherwise it would be tedious to monitor workloads and call an API to obtain more nodes or jobs might run out of capacity or slow down.

DataXu needs high performance for its Big Data queries, and Qubole optimizes performance several ways including MapReduce split computations, and S3 I/O optimization.

> *Simply put, Qubole is awesome. It's put our cluster management, auto-scaling and ad-hoc queries on autopilot. Its higher performance for Big Data queries translates directly into faster and more actionable marketing intelligence for our customers.*

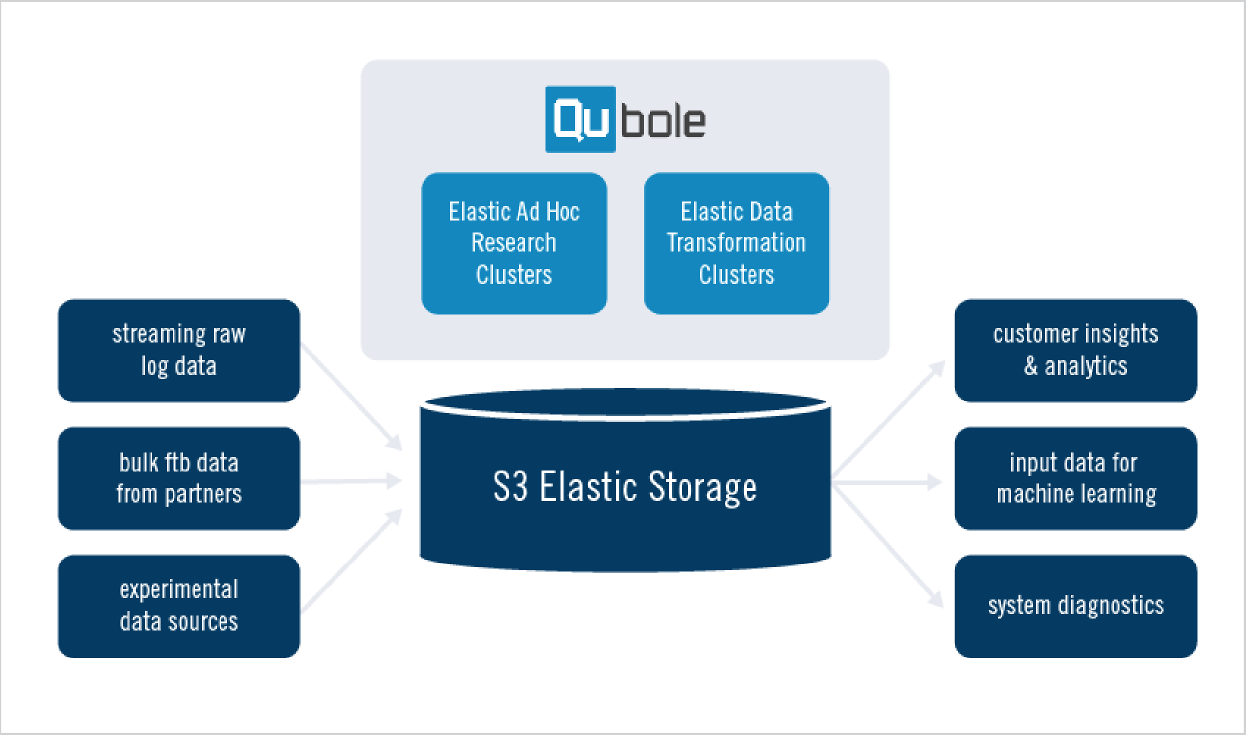
YEKESA KOSURU
Vice President, Technology,
DataXu

Why Qubole Data Service?

DataXu had heard about Qubole's Hive as a Service and wanted to give it a try to see if could help automate cluster management, auto-scaling, and ad-hoc queries. The company was also interested in Qubole's extensive Hive performance optimizations. The company left its existing on-premise and Amazon EMR deployments in place, adding Qubole for ad hoc analytics on Hive with an eye on eventually using Qubole for other Big Data processing needs.

Customer Profile

- DataXu Big Data and Cloud Querying in a box
- Tens of users across engineering, data science, operations and business development
- Applications: ad-hoc queries and data science
- Data: 200 terabyte cluster, double-digit terabytes daily
- Higher availability with auto-scaling, reliable cluster configurations, and automated cluster starts
- Highly optimized Hive processing with faster split computations, S3 I/O, and queries
- No dedicated staff to setup and manage clusters and Hadoop distributions


Qubole
Elastic Ad Hoc Research Clusters
Elastic Data Transformation Clusters
streaming raw log data
bulk ftb data from partners
experimental data sources
S3 Elastic Storage
customer insights & analytics
input data for machine learning
system diagnostics

Results

By using QDS to put its Big Data processing tasks on auto-pilot, DataXu now achieves:

- Higher availability with auto-scaling, reliable cluster configurations, and automated cluster starts when queries are executed
- Highly optimized Hive processing with faster split computations, S3 I/O, and queries
- No dedicated staff to setup and manage clusters and Hadoop distributions

Qubole Data Service (QDS) is in production at DataXu with a 200 terabyte cluster that grows daily by double-digit terabytes. QDS makes DataXu's Hive implementation faster, more available and easier while helping the company save money on processing and engineering support. Adhoc users are very satisfied with the Qubole performance, replicated metastore capability, cluster startup time, ease of use and technical support.

Adding machines for more capacity is also automated with QDS' auto-scaling. QDS scales DataXu's nodes up and down based on workloads without the need for engineers to monitor them and manually request additional nodes when needed.

"Qubole has put our cluster management, auto-scaling and ad-hoc queries on autopilot," says Yekesa Kosuru, Vice President Technology at DataXu. "Its higher performance for Big Data queries translates directly into faster and more actionable marketing intelligence for our customers"

QDS automates DataXu's queries and runs them a lot faster. Users can issue queries whenever they want without engineering's assistance since QDS automatically starts a cluster when the query is executed and maintains cluster size appropriately based on load. DataXu also benefits from the QDS user interface and its Python SDKs to make querying Big Data much more intuitive. DataXu finds that setting up and managing clusters is very simple using Python SDK. Data scientists specify the Hadoop distribution and the number of machines and QDS automatically sets up the cluster and runs the workload. In addition to saving time, this has virtually eliminated all manual configuration errors, giving DataXu the automated end-to-end execution.

And, QDS' extensive Hive optimization gives DataXu faster split computations, Amazon S3 I/O performance and Hive query processing.

Without QDS, DataXu estimates that it would have had to have hired additional Hadoop engineers, and operations personnel to manage, monitor and start clusters

The Future

DataXu has been so successful with QDS that its next steps are to move additional workloads to QDS for improved performance, optimization, cluster management and failover. DataXu is also very excited about QDS' new ability to run a job in another cluster if a cluster goes down so that it can meet its time to report requirements without having to invest in making clusters highly available.

DataXu is also exploring the R programming language integration offered by QDS for its machine learning team. The teams want to leverage R and Hadoop for computational statistics, visualization and data science.

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 cloud-agnostic big-data-as-a-service provider. Customers have chosen Qubole because we created the industry's first autonomous data platform. This cloud-based data 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.