CASE STUDY

Building a Self-Service Datalake to Enable Business Growth

Overview

In order to help our users easily earn cash back rewards, Ibotta analyzes receipts and partners with popular retailers and manufacturers. This analysis enables Ibotta to seamlessly pay users and create personalized experiences when they interact with the app.

Jon King (Manager of Data Engineering at Ibotta): I have the responsibility of orchestrating the company's data infrastructure and providing the right teams with the right data. My team is focused on ensuring that the company's data is getting to where is needs to be reliably and cost-effectively; including virtually each data user in the company (Data Science, Analytics, Developers, Support, Sales and Finance).

David McGarry (Director of Data Science at Ibotta): My responsibility is to lead the Data Science teams. We are focused on optimizing internal analytics through augmenting our customer intelligence data, as well as engineering Machine Learning features within the mobile application.

Together this means we have a treasure trove of data. Processing it at scale to meet our customers' needs also means productionalizing new features in parallel, which requires some serious firepower. Enter Qubole, a technology platform for big data in the cloud with managed autoscaling for Spark, Hadoop, and Presto.

> *"Our data users were changing. We needed to support not only professionals interested in descriptive reporting, but data scientists and analysts seeking to derive value from the data."*
>
> **Nathan McIntyre**
> Data Engineer, Ibotta

Company

Ibotta is a mobile technology company transforming the traditional rebates industry by providing in-app cashback rewards on receipts and online purchases for groceries, electronics, clothing, gifts, supplies, restaurant dining, and more for anyone with a smartphone. Today, Ibotta is one of the most used shopping apps in the United States, driving over $5 billion in purchases per year to companies like Target, Costco and Walmart. Ibotta has over 23 million total downloads and has paid out more than $250 million to users since founding in 2012.

Maintaining a competitive edge in the eCommerce and retail industry is extremely difficult because it requires engaging and unique shopping experiences for consumers. Ibotta solves this with a smartphone app that provides seamless ways to instantly register purchases by scanning receipts. This simplified purchase registration drives immediate rebates for consumers, as they receive easy cash back rewards by shopping at their preferred stores and restaurants. As a result, Ibotta's app delivers unique shopping experience that increases customer satisfaction and brand engagement, while gathering a wealth of insights from the consumer's buying experience.

From 2012 to 2017, Ibotta's data volume was no larger than 20 TB (terabyte) total. Today with new data acquisition and Machine Learning features, our company is pushing over 1 PB (petabyte) of data assets. Using Qubole we are now able to deliver these quality features and manage scale according to the ROI of each data project -- at the same time tuning up both storage or compute as necessary based on volume, velocity, and variety of data.

Business Need

Ibotta's ability to track consumer engagement at the point of sale allows us to provide a 360 degree view of analytics on purchase attribution back to our partners. Our app provides Business Analytics that enables retailers and brands make more informed buying decisions in-store and online. This information helps retailers and brands engage with their customers at a very personal level, as well as optimize future investments in new products and marketing campaigns.

The insights we provided were so valuable, our partners kept requesting more detailed information and features to better engage with existing and new customers. As a result, the company decided to focus on expanding advertising and eCommerce segments in the product. This, in turn, has created new revenue streams that we can reinvest back into the business and increase our consumer's savings. It was at this point Ibotta began to see a huge growth in the data. We needed a solution to help scale the company's goals.

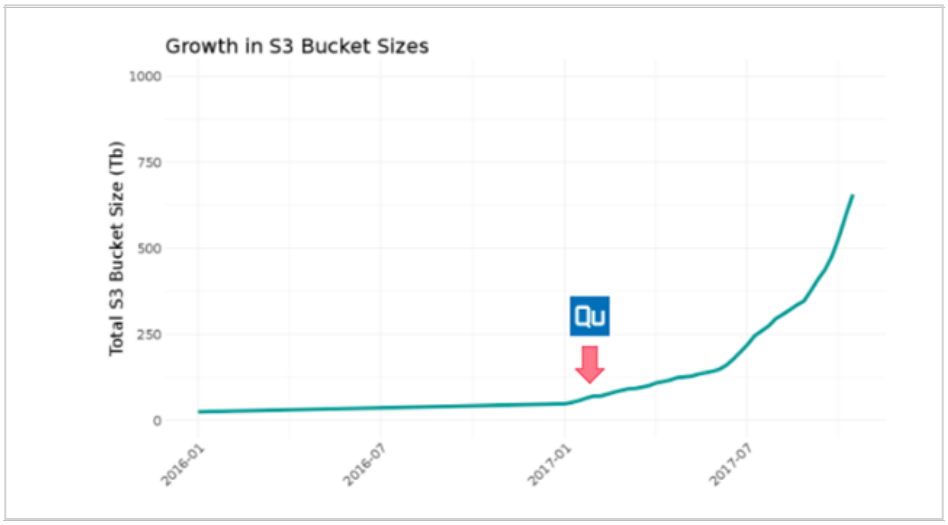
Challenges

While the growth in data has helped improve insight, it also had a significant impact on the data infrastructure and was a key driver for us to change technology operations.

Since March of 2017, Ibotta's data has grown by over 70x, to nearly 1PB, with over 20 TB of new data coming in daily. The biggest driver of data growth has come from generating first-party data features in order to improve our users' personalized experiences. Ibotta is now able to fully leverage our data asset, including:

- **SKU-level data from receipts and loyalty cards:** 226 million processed receipt images managed by organizing, modeling, and consuming the parsed results; supplemented with over 1.5 million loyalty cards from 100+ integrated retailers.
- **In-App User Activity:** Content users have viewed and interacted with, as well as geofence breaks captured from 700k+ stores.
- **Self-Reported User Data:** Basic demographic information such as age, gender, ethnicity, income, education, family; and is later decorated thru surveys, custom questions and other engagement opportunities.

Prior to moving to a big data platform with Qubole, Ibotta's Data and Analytics infrastructure based on a cloud data warehouse (AWS Redshift) was static and inflexible. This meant our teams were limited in scope and the technical capabilities to meet the business' growth. Scaling the system due to the surging data volumes became cost prohibitive and we were seeing a diminishing return on costs. Furthermore, this constrained the team's ability to truly leverage the data across the business. Scaling out any new product data features or consumer insights was impossible.



The data infrastructure constantly was running into scalability problems, raising another challenge. The problem in Redshift is that compute is concurrently tied to storage, so there wasn't a straightforward way to scale up storage without concurrently scaling compute. End-users increasingly were forced to compete for shared compute resources and the data infrastructure was burdened with the weight of any added workloads. Ibotta's data acquisition began to outpace the compute requirements. This became increasingly cost inefficient.

More importantly, this data warehouse was not an ideal solution for iteratively scaling the memory intensive predictive and prescriptive analytics workloads that the Data Science team was getting ready to start putting into production.

The Team

We needed to grow beyond business analytics that was complementary to our products, into a pure data-driven company. This meant the organization needed to be segmented so we could staff the right teams and people in order to help accomplish these aspirations:

- **Data Engineering & DevOps:** Architect data lake, manage technologies, provide data services, and create automated pipelines that feed into various data marts.
- **Data Science:** Enhance data and insights while creating new product features—ranging from use cases such as category predictions, item text classification, a recommendation engine, and in-app A/B Testing.
- **BI and Consumer Insights:** Develop and deliver insights for internal customers and retail partners.

Building a self-service platform is easier said than done, and could not be simply done as a 'lift and shift.' Data needed to be separated from compute and accessible by all users, so we had to start building a new infrastructure from scratch with a cloud data lake.

> *"When I started at Ibotta our data was exclusively accessible via Redshift and our infrastructure limited my team to in-memory solutions to the problems we were solving."*

> *"Within my first two weeks at Ibotta, Qubole enabled my team to immediately move data from Redshift into S3 and use distributed frameworks to train and deploy our models. This led to my team building new products within a matter of days."*

David McGarry
Director of Data Science, Ibotta

‘Data Lake’ Plan: Building a New Plane While Flying the Old One

To address the problems faced by the data teams, we built a cost-efficient self-service data platform. Ibotta needed a way for every user -- particularly Data Science and Engineering -- to have self-service access to the data; and to be able to use the right tools for their use cases with big data engines like Apache Spark, Hive, and Presto. The data team needed to be able to complete the tasks of preparing data for those Data Science and Engineering Teams at the same time. Qubole simultaneously provided an answer to the demands of both teams, those perfecting operations as well as analyzing the data.

“*Qubole increases the speed and agility of democratizing data to end users and services once data is in the cloud by providing a unified and collaborative data platform*,” said King.

Bijal Shah, SVP of Analytics & Data Products, brought in David McGarry to run the Data Science operation. Ron White, VP of Engineering, hired Jon King to lead Data Engineering and the architecture plan. Along with this new leadership, we had to set expectations to the business:

- **Architecture:** Building the new plane (‘Data Lake’) while fixing the old one (‘Redshift’) mid-flight.
- **User Enablement:** Training a new toolset with distributed computing on Hadoop, Presto and Spark. Relying heavily on both external training and support (through Qubole), and internal tech talks and support channels.
- **Executive Support:** Scaling organizational supply with demand by showing value in the short-term helped significantly increase buy-in from upper management to approve the needed headcount and external support.
- **Control of “Unlimited” Resources:** Working with Qubole we are able to make data operations straightforward for developers and users alike. Qubole makes it easy to administer and scale clusters without expert knowledge of distributed computing and AWS, and teams like Data Science to manage their own costs and operations per workload.

It was at this point our data lake initiative was in full force. We started to point the data ingest buses (AWS Kinesis and Apache Kafka) and pushing all new data to AWS S3 object store. The outcome here was the baseline starting point of the data lake. We began to use Qubole as the data operations layer around which to build our data platform.

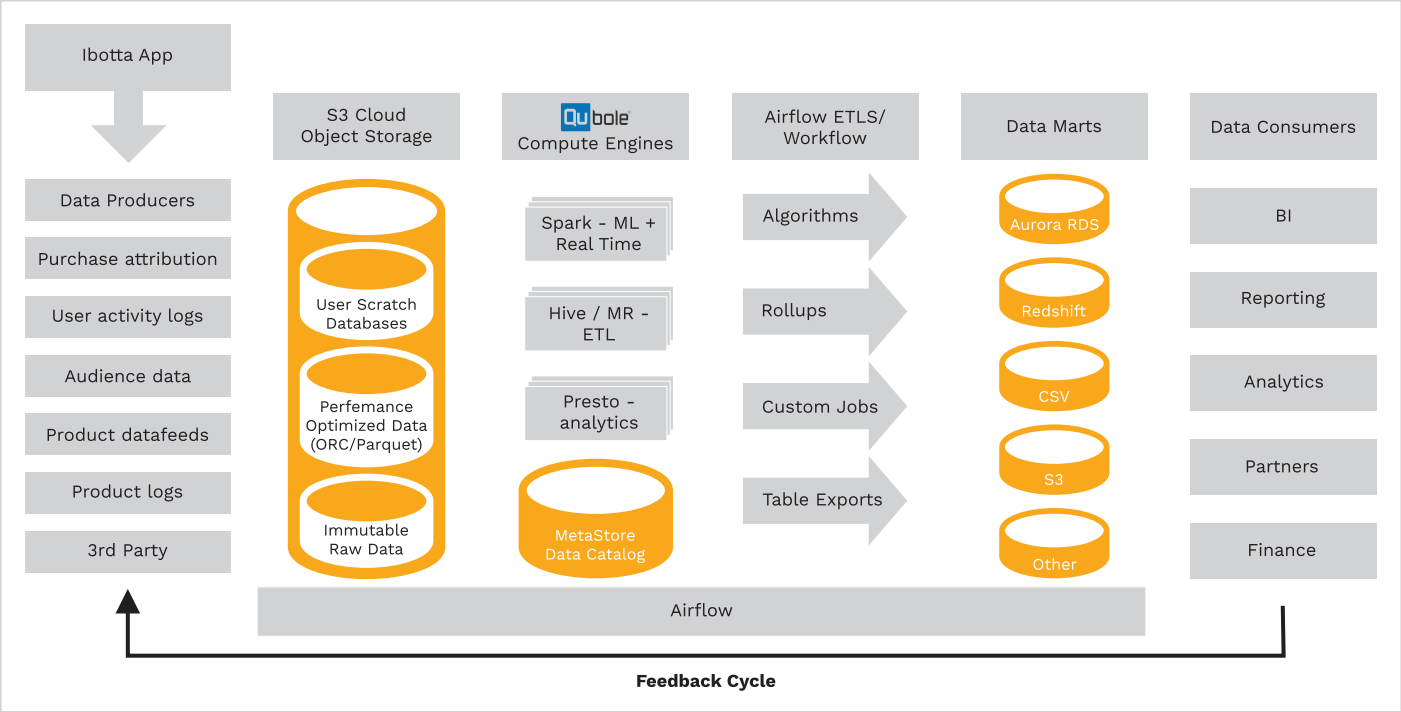
Amazon S3 storage is an extremely low-cost and nearly infinitely scalable storage solution. Also it’s high availability allows Qubole to be our workhorse and start working on the data immediately with as much compute as we need.

Solution - Scalable & Cost-Effective Data Platform

"Qubole allowed us to focus on finding the value in our data with less management of the system," explained Jon King. "Qubole's write once, read anywhere paradigm allows users the ability to try Hive, Spark, Presto and Mapreduce against our data and choose the best overall solution."

To mitigate the legacy data warehouse constraints, Ibotta now has ETL jobs loading data from Hive into Redshift for consumption by our BI tool, Looker. Ibotta is also moving to transition some of the larger Looker reports towards using Apache Presto as the SQL engine and data in S3 as its backend. Ibotta utilizes Hive and Spark jobs for processing raw data into production-ready tables used by Analytics, orchestrated using Apache Airflow.

Using Airflow's hooks into Qubole to ease automating jobs via the API. Airflow gives more control over orchestration than cron and AWS Data Pipeline. It also provides performance benefits from parallelization and the flexibility of scheduling jobs as a DAG instead of assuming linear dependency.



The data lake architecture no longer uses Redshift as the one stop solution. The app and internal tools that keep track of users, clients and campaigns utilizes an operational data store based on AWS Aurora DB. High volume tracking data is also ingested through Kinesis and written to Firehose, which is then stored on S3 object stores.

This data is standardized in JSON and periodically dropped to our raw S3 buckets in gzip format, which is our DR environment. Third-party data is synced between S3 buckets or delivered via SFTP. From there, data is formatted into ORC and Parquet for performance optimizations and pushed to our separate production S3 bucket used by the Data Science and Analytics teams.

> *"Qubole allows us to make more useful features for our customers by allowing us to combine multiple data sources and faster iterations."*

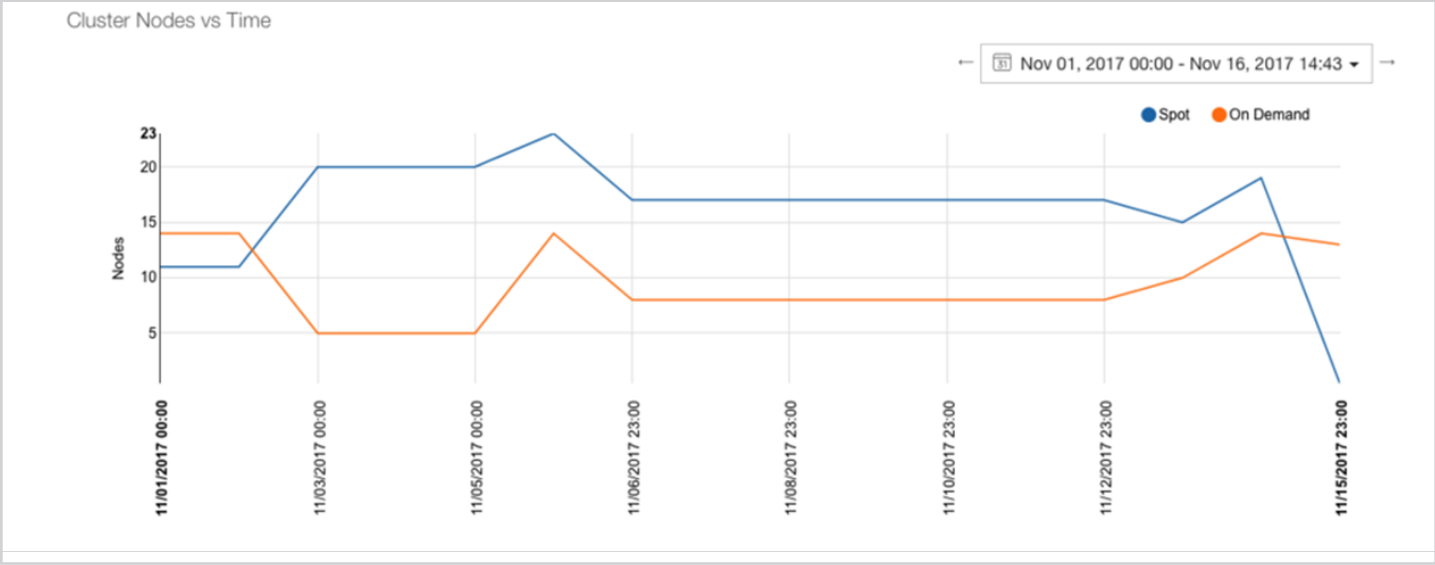
Jon King
Manager of Data Engineering & DevOps, Ibotta

Impact: Affordable Data-Driven Applications

"Qubole has made us more innovative," King stated, "by allowing us drive greater value from our data in less time and with greater collaboration."

Utilizing this platform, Ibotta has empowered the Data Science teams to build products and Business Intelligence to produce real-time dashboards for hundreds of users. Since instituting our new data platform, Ibotta has increased the volume of processed data by over 3x within 4 months of getting started; and we are passing over 30k queries per week through Qubole.

Ibotta uses Qubole provisioning and automating our big data clusters. Specifically, Spark is used for machine learning and other complicated data processing tasks; Hive with ETL; and Presto for ad-hoc queries like exploratory analytics.



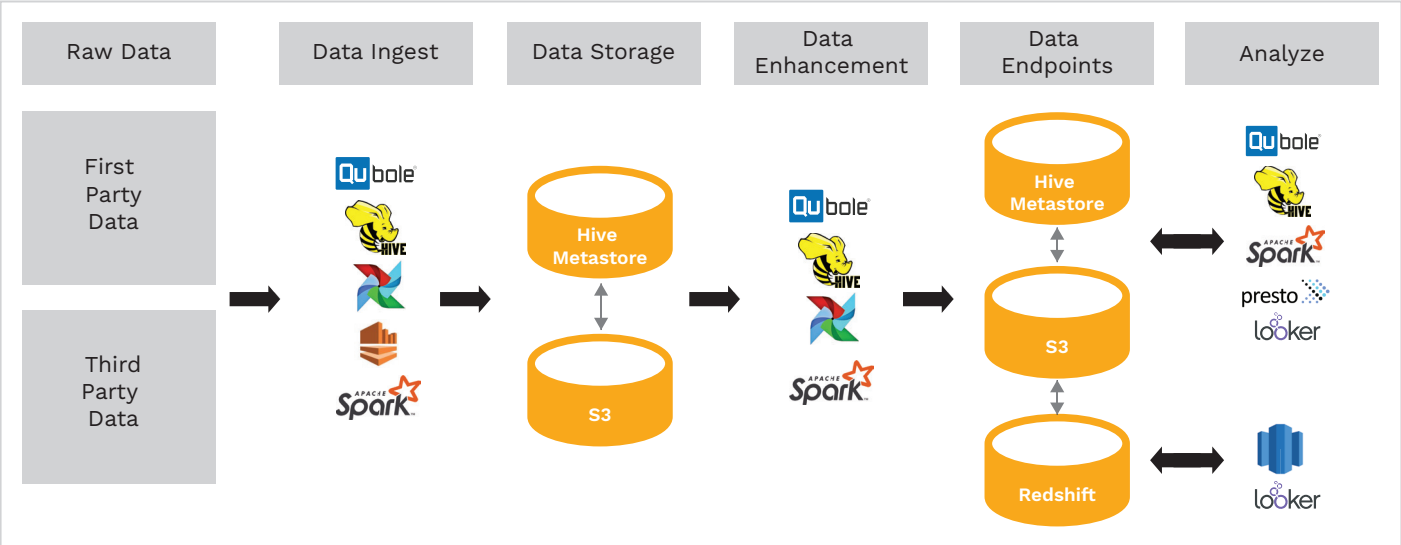
Above is a two week running ETL cluster with Hive on YARN with Qubole's managed autoscaling and Spot Instance Shopper that scales according to workload demand. As you can see in the blue horizontal line, when EC2 Spot Instances become unavailable Qubole will seamlessly reprovision these "lost nodes" to other AWS node-types with Spot Instances or on-demand EC2 instances gracefully.

Building a Better Product

Ibotta's Data Science and Engineering teams were immediately empowered once Qubole was in place. They achieved the goal of self-service access to the data and efficient scale of compute resources in AWS EC2 for big data workloads. Within a month Data Science was launching new prescriptive analytics features in the product that included a recommendation engine, A/B testing framework, and an item-text classification process.

Data Platform: Lifecycle

When combined, Ibotta's data platform becomes the foundation for creating a data-driven culture



Ibotta now can provide even better user experiences by delivering personally relevant content and unique customer experiences. Operationally, Qubole enables the Data Science and Analytics teams to focus less on mundane tasks and more on what matters.

Savings

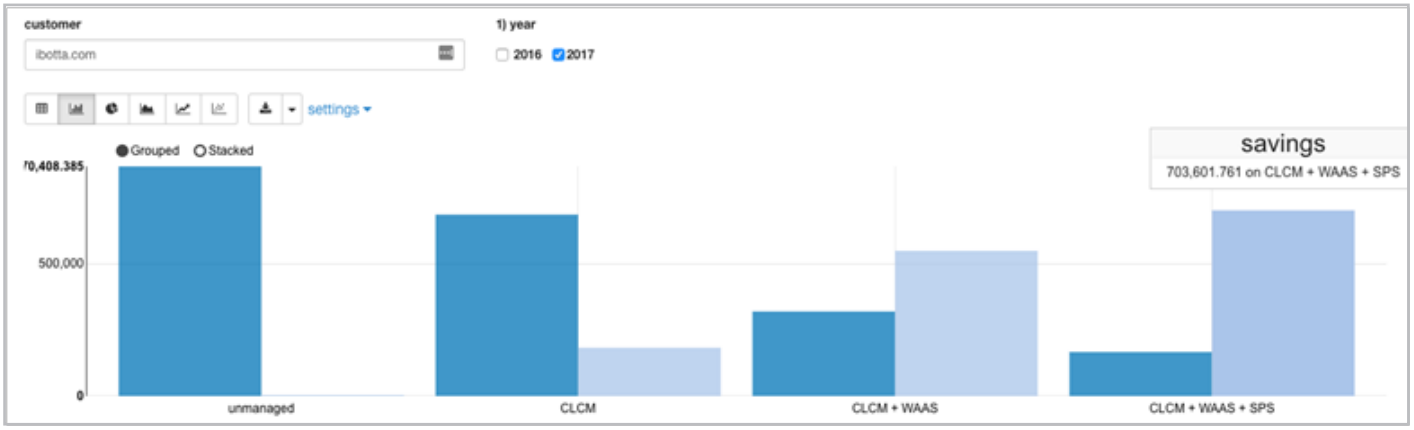
Qubole provides near-immediate access to data so Ibotta can now perform big data operations in hours rather than days or weeks. Additionally, as we have grown **Ibotta has realized savings of 70-80% of our big data costs on Amazon EC2.**

Ibotta's Big Data Cost Estimates on AWS EC2 from May-December '17:

- Saved* an estimated $1.2 Million
- Spent an estimated $270k

***Note:** Savings are a measurement of TCO relative to Ibotta's cluster configurations in Qubole and managed automation for Hadoop and Spark clusters: Workload Aware Autoscaling (WAAS), Cluster Lifecycle Management (CLCM), and Spot Instance Shopper (SPS).

Comparison of costs from big data clusters run through Qubole

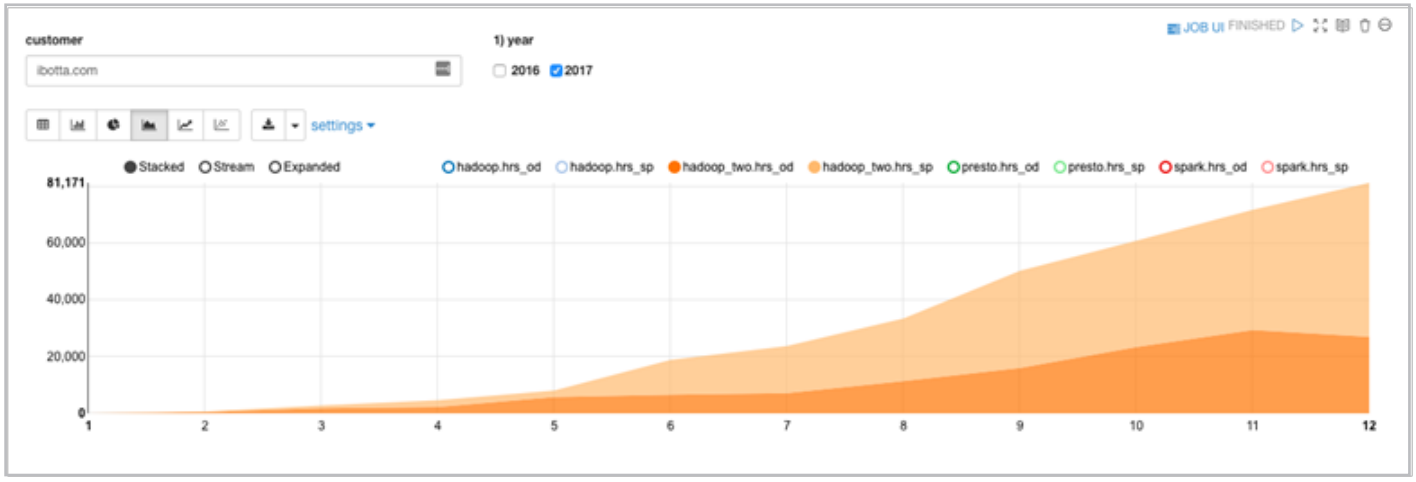


> *"The great thing about Qubole is that it allows me to tell the story of smarter spending for big data."*

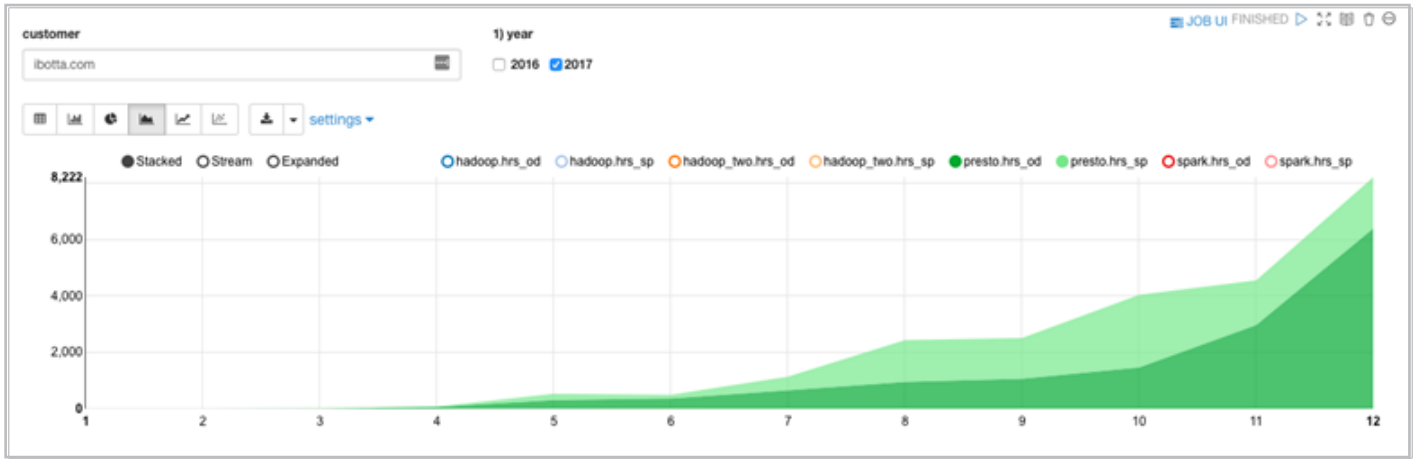
> *"We were able to achieve 70% spot usage with only a couple hours of work, so the ROI was well worth the effort. Qubole support is also able to help us maximize our spot usage even further."*

Jon King
Manager of Data Engineering & DevOps,
Ibotta

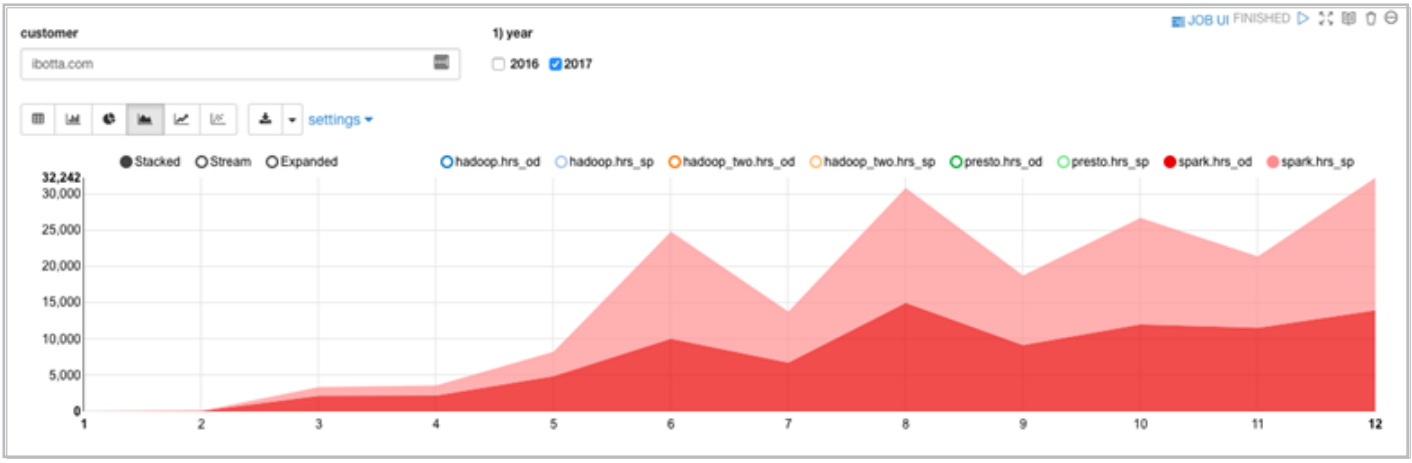
Apache Hadoop 2: Hive on YARN + Tez (ETL and concurrent analytics) - average 64% Spot Utilization across all workloads



Apache Presto: (interactive SQL analytics) - average 63% Spot Utilization across all workloads



Apache Spark: Spark on YARN (ML, ETL & ad-hoc analytics) - average 52% Spot Utilization across all workloads



Our big data clusters are using 60-90% mix of Spot Instances with on-demand nodes, which combined with use of Qubole's heterogeneous cluster capability makes it really easy and reliable to achieve the lowest running cost for big data workloads.

Having Qubole at Ibotta, each of our teams can have as many resources as we need (within limitation), but are also able show concrete cost savings based on the workloads we are running in AWS. This means managing budget and ROI is much easier and to manage, and allows us to forecast how we scale different features and projects accordingly. On top of this, saving over what would've cost millions of dollars to build on AWS, we can make the case to our executives of spending smarter and not less, which allows our teams to focus on the value of the data and iterate faster.

Next Steps

Ibotta is well on its way to building the world's starting point for rewarded shopping by partnering with Qubole and building out our cloud data lake. More than ever, we are focusing on delivering next generation ecommerce features and products that help drive both a better user experience and partner monetization. Qubole allows us to spend time developing and productionalizing scalable data products, more importantly concentrating on bringing value back to our users and company:

- **Data-Driven Culture:** Continuing to make sure that technology, projects and company culture work together seamlessly. Through training and a community of thought, sharing among teams has been helping everyone adapt to new infrastructure.
- **Product Innovations:** Leveraging Qubole to drive new eCommerce and digital media features within the retail industry that lead to even greater actionable insights for our partners.
- **Real Time Stream Analytics:** Leveraging stream processing to deliver realtime unique user experiences and increase the quality of our product.
- **Increased Performance:** Query tuning, code reviews, and optimizing data structure. We're also leveraging the new class of intelligence features in Qubole with **AIR Data Intelligence** (Alerts, Insights, and Recommendations) to improve operational efficiencies and performance across queries and workloads.

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. 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

Try QDS for Free:
qubole.com/pricing

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