Qubole on Google Cloud Platform

Rapidly deploy analytics and machine learning with Apache Spark, Hadoop, and Presto

Qubole on Google Cloud Platform is a self-service, collaborative, enterprise platform for data engineering, predictive analytics, and machine learning. The platform enables data scientists and data engineers on Google Cloud Platform (GCP) to collaborate on building scalable data pipelines using native interfaces, built-in tools, and optimized data processing engines.

Qubole and Google have partnered to build an enterprise-grade platform that delivers the unique set of benefits described below.

A unified and rich experience with built-in end-user tools such as native notebooks, an integrated workbench for commands, and built-in connectors to multiple data sources.

Highly optimized versions of data processing engines such as Apache Spark, Hive, and Airflow for better performance and efficiency.

Enterprise support for data processing engines such as Apache Spark, Hive, and Airflow by specialized engineering teams focused on each engine.

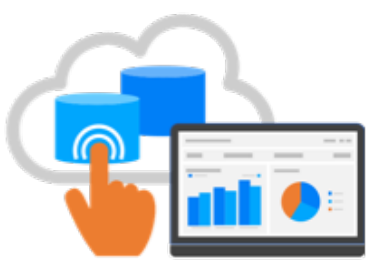
Advanced automation and cluster lifecycle management that enables any scale deployments while maintaining low administrative overhead.

Modern, container-based Qubole control tier deployed on Kubernetes, which autoscales to handle high volumes of user sign-ups and concurrent commands.

Enterprise-grade security with fine-grained access controls to data and platform resources such as clusters, notebooks, commands, and more.

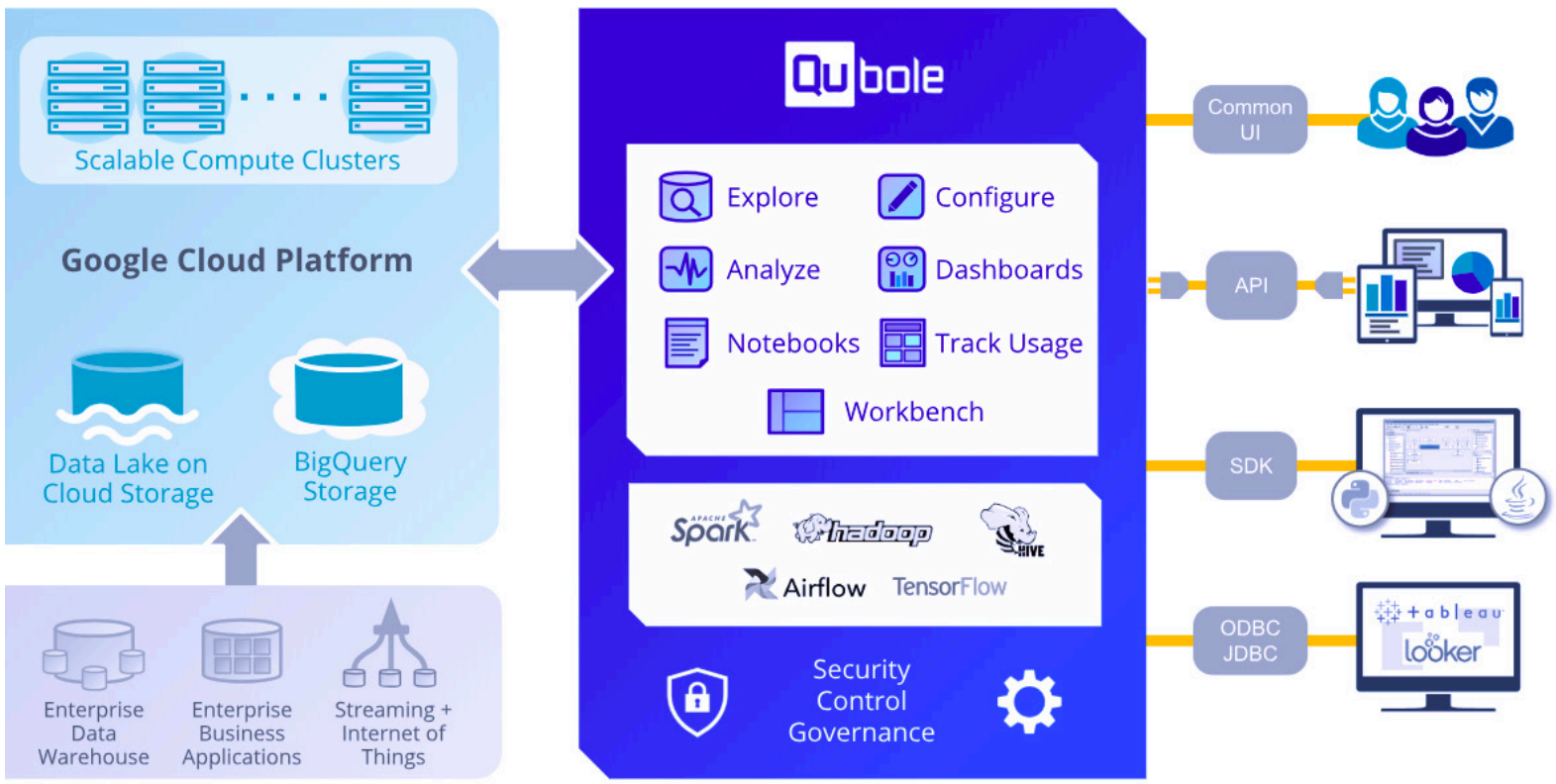
Deep integration with Google services such as Google Compute Engine, Google Cloud Storage, BigQuery storage, preemptible virtual machines (VMs), and Cloud Identity and Access Management (IAM).

Fast access from GCP marketplace with flexible purchase plans, automated service setup, and simplified user onboarding.

A Modern Scalable Architecture

Qubole is built on a modern scalable architecture that leverages Kubernetes. The control tier authenticates users into the service using their Google Cloud account and provides access either through a user interface (UI), APIs, or SDK. It provisions big data clusters within a customer's virtual private cloud (VPC); manages access to data in Google Cloud Storage and BigQuery storage; and delivers workload-aware autoscaling of all compute resources.



The platform offers a choice of optimized open-source engines, such as Apache Spark, Hive, Presto, and Apache Airflow. In addition, Qubole includes a unified Hive metastore that acts as a central metadata store for all data sources.

Users can submit, monitor, and manage their data processing jobs from either the Qubole UIs, APIs, or SDK. For example, they can use Qubole notebooks to build and deploy machine learning models; or the Workbench to create and execute queries; or use Airflow to build and deploy scalable data pipelines.

CUSTOMER SPOTLIGHT

AgilOne: Taking Machine Learning to Enterprise Scale

AgilOne operates complex machine learning (ML) models and stores vast quantities of data for its customers, including major brands like Lululemon, Travelzoo, and Tumi.

AgilOne Cortex is a very robust and flexible machine learning framework built into a customer data platform. AgilOne Cortex uses supervised machine learning models to predict customer events such as purchase, subscription, and engagement. It also intelligently segments customers together based on interest and behavior using unsupervised learning techniques. AgilOne Cortex's recommender models allow the orchestration of offers and messages to customers on a 1:1 basis.

AgilOne operates Cortex on both Amazon Web Services (AWS) and Google Cloud Platform (GCP) and performs close to one billion predictions every day, averaging dozens of millions of customer predictions for each client across all its models.

To meet the challenges of such vast amounts of data and millions of predictions, AgilOne partnered with Qubole to better automate the provision of machine learning data-processing resources based on workload, while allowing for portability across cloud providers; eliminating prototyping bottlenecks, supporting the seamless orchestration of jobs, and automating cluster management.

AgilOne now runs a variety of workloads for querying data, running ML models, orchestrating ML workflows, and more on Qubole—all on a single platform with optimized versions of Apache Spark, Apache Airflow, Zeppelin Notebooks, and leveraging Qubole's APIs to automate tasks.

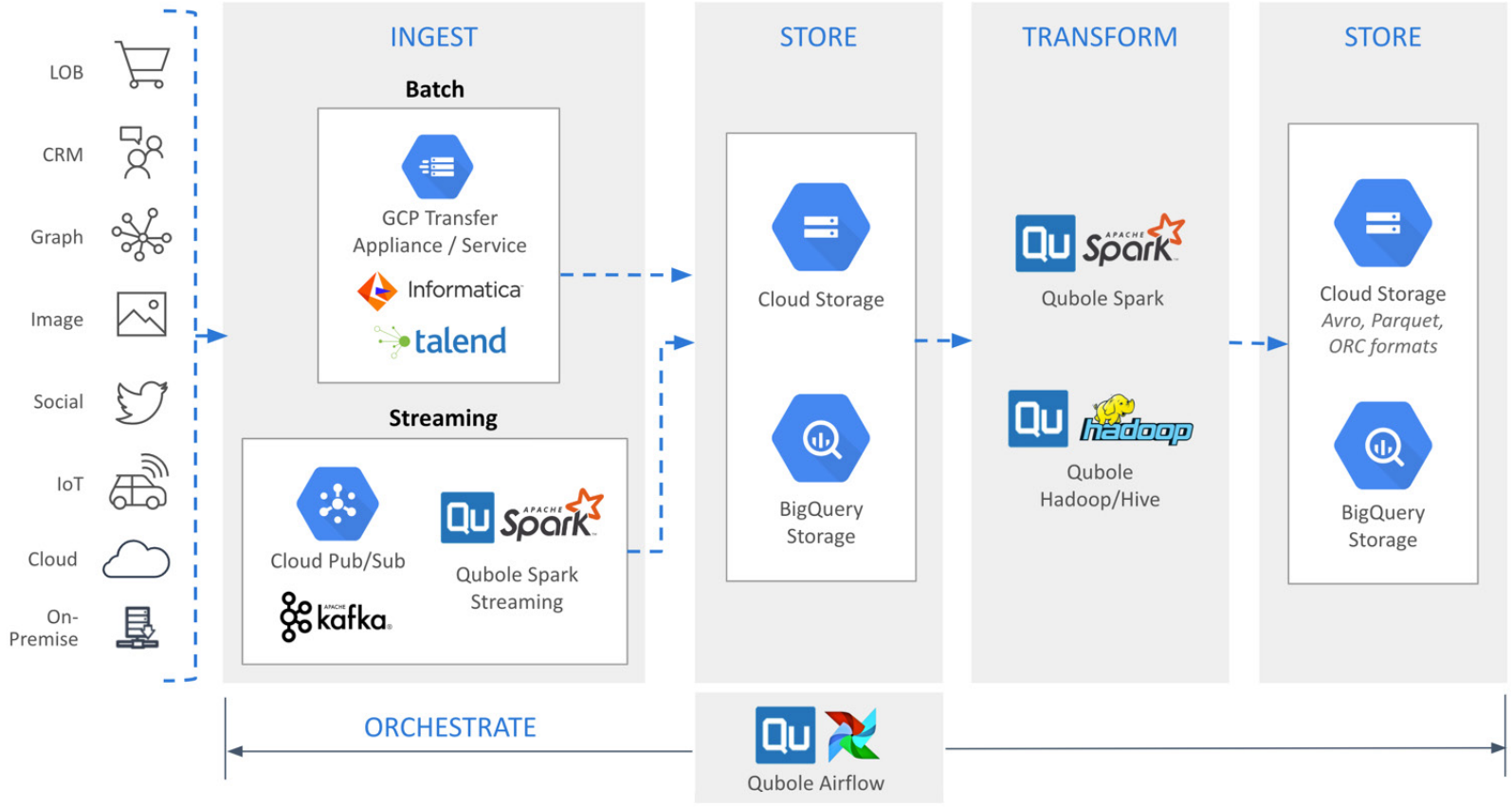
Learn more about how AgilOne takes machine learning to enterprise scale.

AGILONE

Architecture Patterns and Use Cases

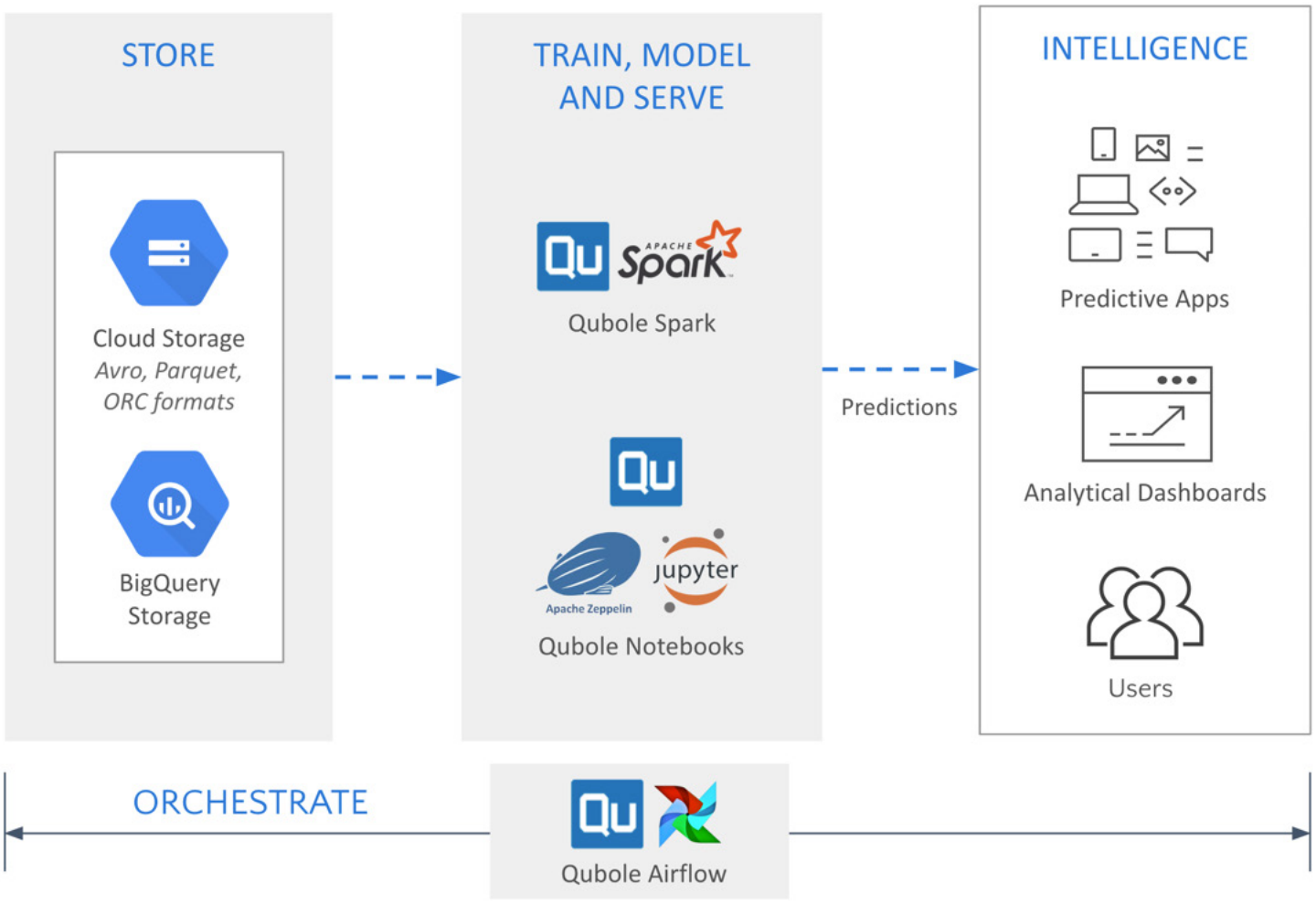
Data Ingestion and Transformation

With Qubole, data engineers can build scalable pipelines for batch and streaming ingestion of data into Google Cloud Storage or BigQuery storage. They can use Qubole's optimized versions of Spark or Hive to perform data transformations, including complex joins or conversions to Apache Avro, Parquet, or ORC formats for downstream use cases, such as machine learning (ML) or data exploration. Qubole's built-in workbench enables engineers to execute Spark and Hive queries, or use Qubole's optimized version of Apache Airflow to build and orchestrate scalable data pipelines on GCP.



Machine Learning and Predictive Analytics

Data Scientists can use Spark on Qubole and its native notebook interface to build and train ML models using data from Google Cloud Storage or BigQuery. They can develop models in multiple languages such as Python, Scala, or R and queries in SQL. They can add additional Conda packages and manage dependencies between packages using Qubole Package Management. Qubole's unified interface built for collaborative development enables data scientists and ML engineers to share notebooks, schedule jobs, and publish results to dashboards for easy visualization and consumption by business users.

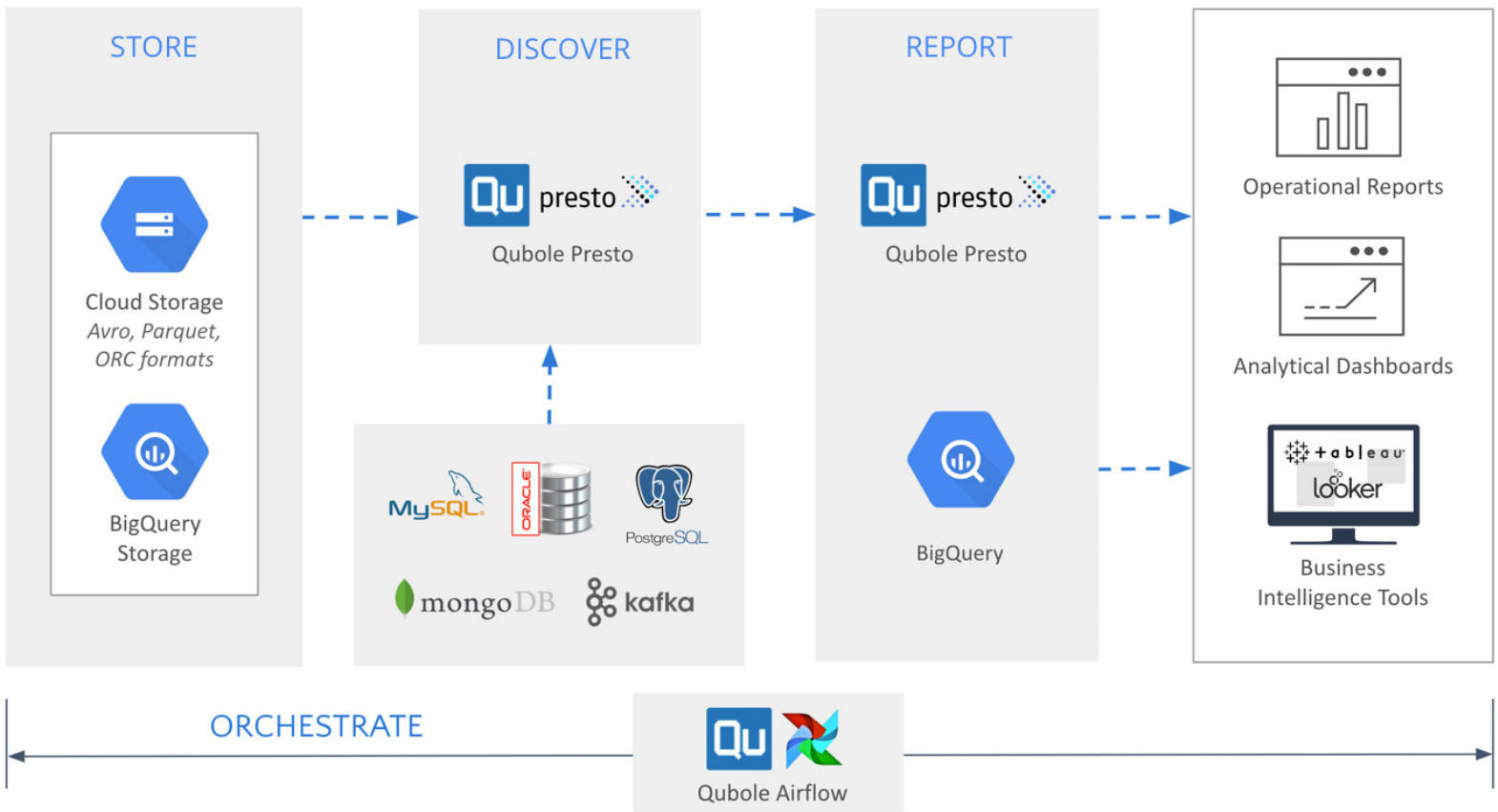


Data Discovery and Business Intelligence (BI)

Data analysts use Presto on Qubole for data discovery and BI, including reporting and dashboards. Presto is based on a Massively Parallel Processing (MPP) design that can process interactive queries at high speeds and scale. What makes Presto unique is its ability to run federated queries across multiple sources such as Google Cloud Storage, BigQuery storage, databases such as MySQL, Oracle, Postgres, and MongoDB, and real-time data streaming applications such as Kafka.

Qubole's pricing model, based on computing hours used instead of data scanned, combined with its capability to autoscale and automatically terminate clusters, provides a cost-effective option for data analysts to perform exploratory analytics and data discovery at petabyte scale.

Data analysts can also use Qubole's optimized version of Presto to run reports directly off a data lake, or load periodic aggregates of raw data into BigQuery for low latency reporting in Looker or Tableau, leveraging certified connectors. In addition, Qubole offers native notebooks for Presto to provide data analysts with another option for easily building interactive queries.



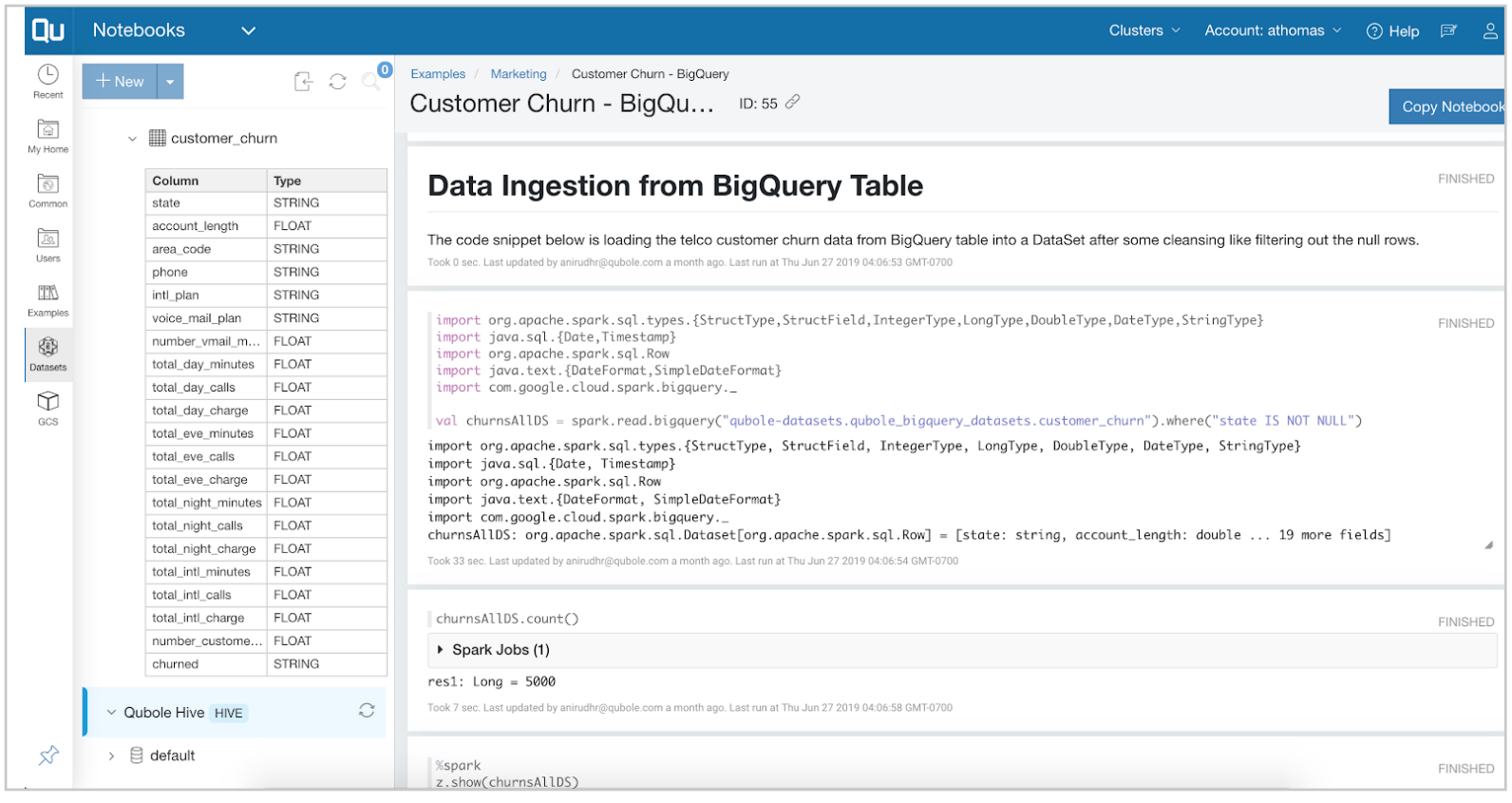
Highlights of Qubole on GCP

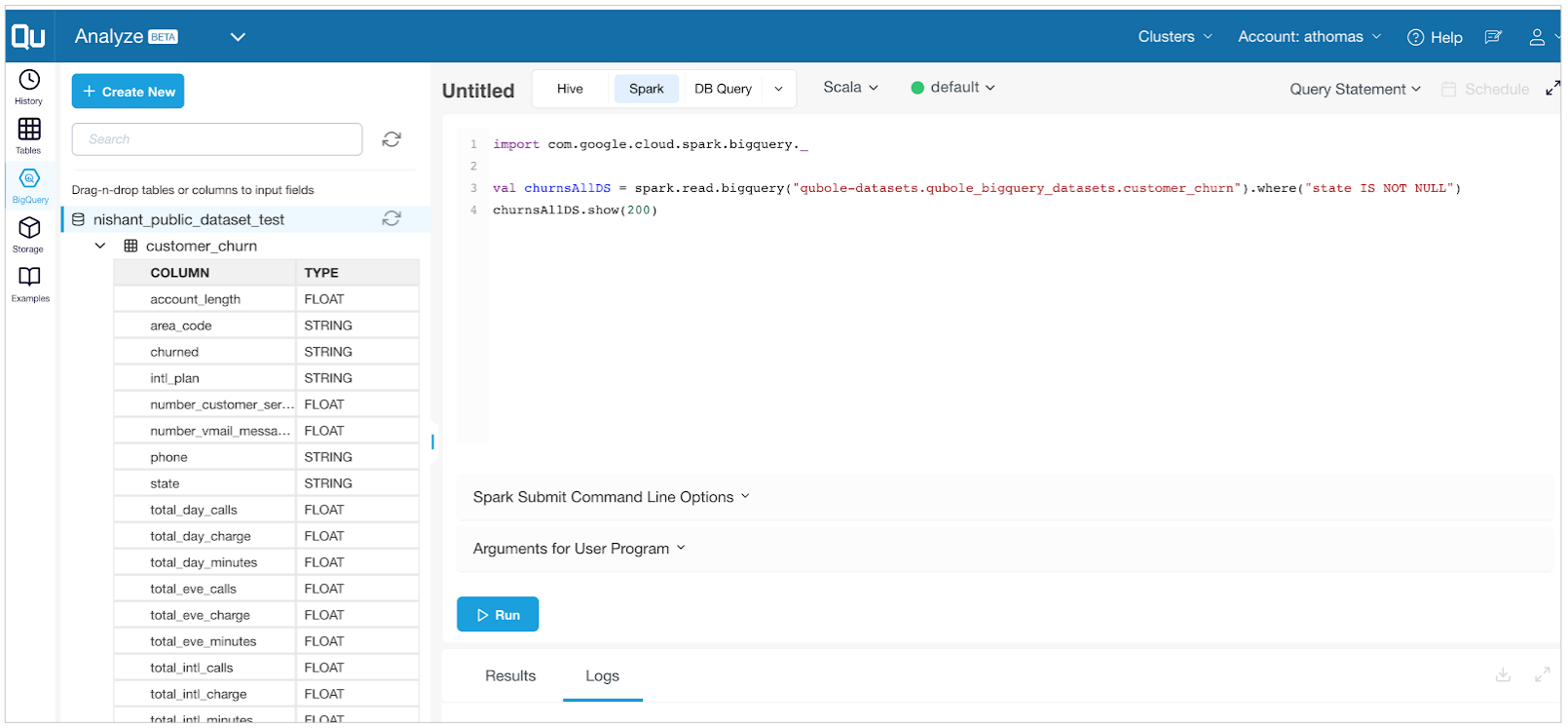
Integration with BigQuery

Spark on Qubole is integrated with BigQuery to enable direct reads of data from BigQuery storage into Spark Dataframes. This allows data engineers to easily explore BigQuery datasets or join data from Google Cloud Storage and BigQuery to perform complex data transformations and queries. Data scientists can now easily look up BigQuery datasets and build machine learning models using Spark on Qubole through notebooks, resulting in significant time savings and efficiency gains for both data engineers and scientists.

The data is read in Apache Avro format using parallel streams with dynamic data sharding across streams to support low latency reads. The connector for BigQuery eliminates the need to export data from BigQuery to Google Cloud Storage, thereby significantly improving data processing times. For more information about the BigQuery Storage API and its advantages, such as direct reads, dynamic sharding, predicate filtering, column filtering and more, refer to the Google documentation.

In addition to Qubole's Spark connector for BigQuery storage, the platform also shows BigQuery datasets directly in Qubole's workbench and notebooks interfaces, enabling data scientists and engineers to easily discover BigQuery tables and datasets from within Qubole.





Autoscaling with Preemptible VMs

Qubole allows data teams to run clusters at scale while still maintaining low cloud costs and high reliability through strong integration with GCP Preemptible VMs. Administrators can now select a mix of regular and preemptible instances and specify the percentage of preemptible VMs desired on a per cluster basis. Qubole handles the acquisition and management of preemptible instances autonomously, and efficiently handles the loss of these instances by minimizing the risk of incomplete jobs or data loss without any intervention needed from platform administrators.

Qubole delivers unique, autonomous capabilities that reduce costs and minimize the risk of using preemptible VMs.

Fallback to On-Demand VMs

Qubole can fall back to using regular VM instances in a cluster when preemptible instances are not available during upscaling. This ensures that jobs complete within a reasonable time, while still making the best effort to contain costs by using preemptible instances when they become available.

Cluster Rebalancing

Qubole can natively and automatically rebalance clusters and retire excess regular (on-demand) nodes when the ratio of regular to preemptible exceeds the desired ratio set by a user. This typically happens when the system is either unable to use enough Preemptible VMs during cluster upscaling, or when it has removed more preemptible instances than regular ones during downscaling. Rebalancing works by automatically swapping out regular instances with Preemptible VMs at the right time when the availability of preemptible ones is higher.

Resiliency and Preemption Handling

Qubole uses several measures to provide a reliable integration with Preemptible VM instances.

Placement Policy

Qubole uses an HDFS block placement policy, which ensures that at least one replica of a file local to a Preemptible VM resides on a stable, regular instance. This ensures that if the cluster loses a preemptible node, a replica of the data continues to exist on regular instances. This not only provides resiliency, but also allows all Preemptible VMs to participate as data nodes in the cluster without compromising the availability and integrity of HDFS. As a result, the storage capacity of the cluster increases and user jobs are no longer constrained to using only the core cluster nodes.

Intelligent Task Placement

Important jobs like Application Master and shell commands that often launch other long running jobs are never scheduled on preemptible VM instances.

Reliable Preemption Handling

When instances get preempted by Google, Qubole immediately stops scheduling new tasks on these nodes. It also stops further HDFS writes to these nodes, backs up container logs, and tries to replicate any state left on such nodes to other surviving nodes. Application masters are automatically notified to restart interrupted tasks on other nodes in the cluster, and cluster autoscaling logic kicks-in to replace the lost node with a new Preemptible VM, should it be available and within policy.

Presto on Qubole

Using Presto on Qubole, GCP customers can perform data discovery, profiling, and reporting and analysis directly off their data lake on GCP. Qubole's optimized version of Presto not only includes a connector to Google Cloud Storage, but also the ability to autoscale using Preemptible VMs—in addition to supporting Qubole's JDBC driver for external application integration.

Data engineers, scientists, and analysts can use Presto to perform discovery of datasets in Google Cloud Storage to figure out patterns and trends, profile datasets to identify quality or integrity issues, query data to answer specific questions, or publish results to BI tools like Looker or Tableau for consumption by business users.

Native Jupyter Notebooks

Qubole offers a JupyterLab notebook interface for data scientists and ML engineers. JupyterLab is a web-based interactive development environment for Jupyter notebooks, code, and data. In addition to the rich and usable JupyterLab interface, Qubole's notebooks provide additional features like per-user folder with automatic code persistence, library dependence management for clusters, easy access to Spark resources, automatic Spark application lifecycle management, and enhanced UI widgets like application progress meter.

Custom Commit Plan on GCP Marketplace

Qubole Data Service is available on the GCP Marketplace with integrated billing, which means you receive a single bill from Google for all GCP services including Qubole. In addition, Qubole on GCP is available on a Custom Commit Plan through Marketplace, aside from the existing, flexible pay-as-you-go Essentials Plan. This enables large enterprises to purchase Qubole using annual multi-year subscriptions, with the ability to customize the terms of purchase with personalized discounts and private quotes.

Get Started Easily with Qubole

Qubole enables data teams to run big data analytics at scale using a self-service model with built-in tools, enterprise support for open source engines, and advanced automation that lowers administrative burden and cloud costs.

To learn more, visit our Qubole on GCP webpage or sign up for Qubole Data Service on the GCP marketplace.