CASE STUDY

Industry: Real Estate

Powering the Complete Online Home Lifecycle with Qubole

The Challenge of Eliminating Data Silos

Zillow's portfolio of brands is, quite literally, data-driven. Their business model depends on putting up-to-date and accurate real estate information instantly at the fingertips of both consumers and real estate professionals. Keeping their businesses running smoothly – and improving them – depends on putting vast amounts of actionable information at the disposal of their business analysts.

About Zillow

Founded in 2006, Zillow is an online real estate marketplace created by former technology executives to find houses, apartments, mortgage rates and more to help consumers land their next home.

As Zillow grew, the data volume of its various business units and the need to consolidate information across its brands has risen dramatically. In response, Zillow has been gradually migrating its business data from brand and application data silos to a group-wide data lake built on the Amazon AWS S3 platform.

But the Data Infrastructure team soon ran into several challenges processing data and managing its clusters due to the technology it had originally selected. The inefficient compute autoscaling and Spot buying from the prior platform made it impossible to efficiently scale clusters for Zillow's highly variable Presto workloads and its huge clickstream batch jobs in Spark. Plus, frequent stability issues and lagging performance with Presto were putting a strain on the small Data Infrastructure team trying to meet its SLAs to Zillow's large user community.

Seeking Greater Performance, Stability and Economy

To address these challenges, Zillow turned to Qubole. The Qubole data platform supercharges productivity through intelligent automation, enabling a 1:100+ admin-to-user ratio, and delivers faster cycle times with a self-service infrastructure for all users. There is zero waste as Qubole's workload-aware autoscaling optimizes compute usage based on the task, SLA, or priority, thereby ensuring that a job will have the appropriate resources to complete. Both, aggressive downscaling and container packing also ensure there is no wasted dollars in compute when it becomes idle. Plus the platform is always able to find the best (low cost) Spot instance buy to increase capacity temporarily while further driving down compute costs.

Shifting to Qubole also allowed Zillow greater flexibility. The platform is optimized for multiple clouds, including AWS, Microsoft Azure, Google, and Oracle, and it supports not only Presto and Spark, but also Hive, Airflow, Hadoop, Tensorflow and more. Plus, Qubole's customer-comes-first culture and industry-leading experience in both Big Data and the cloud meant they could get expert help in overcoming their cluster stability and performance problems.

> *"We've been able to lower our Presto compute costs quite a lot with Qubole, because we can depend on Qubole's autoscaling and automated Spot buying to adjust our cluster size quickly and appropriately.*
>
> **Russell Rhodes**
> Senior Manager, Big Data and head of Data Infrastructure Team, Zillow

Lowering Infrastructure Costs

The Data Infrastructure team within Zillow's Big Data group serves a community of 200-300 data analysts, data engineers, and ML engineers; plus, over a thousand users running queries—through third-party tools like Tableau or Mode Analytics. Every day, some 300 to 400 of those users are running terabyte-size queries against an always-running Presto cluster. With such large queries being submitted randomly, the cluster's workload varies widely and rapidly throughout the day.

Zillow quickly realized it was not able to scale its Presto cluster efficiently with its prior solution. "In the past, we were forced to scale it sort of manually – that is, in a scheduled fashion," says Russell Rhodes, Senior Manager, Big Data and head of Zillow's Data Infrastructure team. "That led to inefficiencies where the cluster was larger than it needed to be."

Improving Presto Reliability and Performance

Zillow business analysts use the Presto on Qubole cluster to generate weekly reports on the performance of Zillow's various products and services. This activity intensifies when tracking product launches, with these core metrics reports becoming critical and daily.

"Because we have such a small infrastructure team, I would say somewhere between 25 and 50 percent of the team's time (while running Presto on a prior solution) was spent trying to figure out how to get Presto to remain stable against the load that we had," says Rhodes. "Switching to Qubole got us to a point where Presto is stable. So, in effect, we've been rendered much more productive. I'd say we've also increased user trust in queries, another area where we were looking to improve."

Qubole's dedicated, expert support made a huge difference in stabilizing Zillow's analytics infrastructure based on Presto clusters. Improved stability plus better autoscaling also greatly improved the performance of those clusters.

"The reason that we're saving this much in terms of productivity is because Qubole helped us identify key configurations that needed to be adjusted – stuff you can't just look up in the Presto documentation," says Rhodes. "It took somebody with quite a bit of expertise digging into it with us to find what the issue was."

Saving on Clickstream Processing Costs

Zillow also uses Qubole to process the clickstream data from all its websites. This is a huge batch job on the order of three terabytes (compressed) daily, amounting to approximately 90 terabytes of data per month. It runs on a large, expensive Apache Spark cluster that only runs for a short period each day, so jobs have to be highly efficient. Zillow wanted to leverage Qubole's autoscaling, aggressive downscaling, and automated Spot buying to lower the costs of this massive job.

"By switching to Qubole, we've been able to be more aggressive with Spot instances and more aggressive with our scaling up and down," says Rhodes. "Because of the rate of autoscaling and the reliability of getting good Spot pricing with Qubole, we've been able to reduce costs associated with that workload as well."

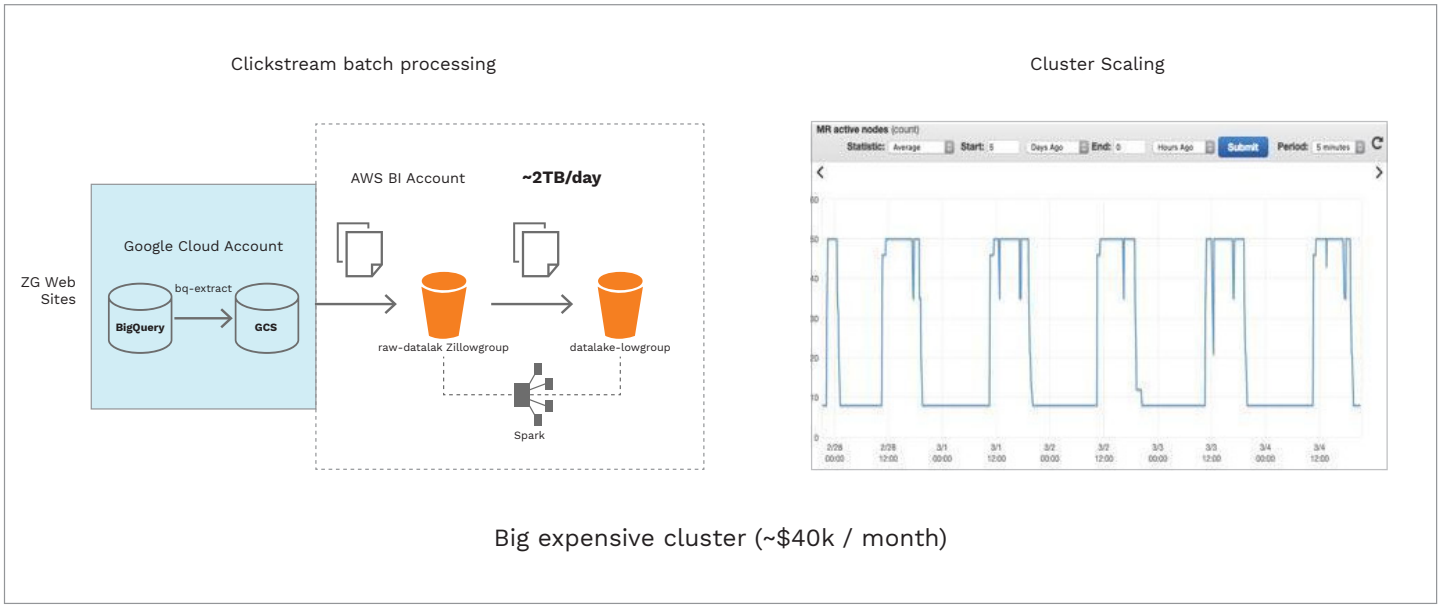

Figure 1: Autoscaling Zillow's clickstream Spark cluster on Qubole

Looking Ahead: A Quest for Operational Excellence

Adopting Qubole and solving the problems just described were only a few of the milestones in Zillow's quest to make their currently-7 Petabyte data lake as reliable and trustworthy as it can be, given its growth of about 10 Terabytes per day.

"A focus for us is operational excellence, meaning getting to scale in all senses of the word," says John Cruchon-Dupeyrat, Zillow's Principal Product Manager for Big Data. "Not just in terms of the number of users, but also in performance, reliability, security and financial governance. It's all about making our data lake a truly enterprise-grade solution."

To increase trust and reliability in its data, Zillow is currently rolling out LDAP integration for every way that users can query the data lake. Their biggest immediate goal, though, is achieving full compliance with the California Consumer Privacy Act – the CCPA – by the end of 2019. "Our desire at Zillow is to go beyond mere compliance with the law," says Cruchon-Dupeyrat. "We want to be a trusted platform for our consumers and users, and trust includes privacy."

Qubole aligns with Zillow's security goals by providing advanced role-based security and authentication designed to detect and avoid gaps that could potentially lead to security or confidentiality concerns.

Qubole offers a rich set of custom-configurable controls that provide Zillow insights into key sources of spend — including locations, time, people, and processes — as well as the ability to apply unprecedented levels of customization.

The Business Value of Qubole for Zillow

- Reduced Costs
  - Inefficiencies of "manual" compute infrastructure scaling eliminated
  - Workload-aware autoscaling, aggressive downscaling, and intelligent Spot buying have lowered operational costs of Presto and Apache Spark dramatically
- Improved productivity
  - Data Infrastructure team now 25% to 50% more efficient due to improved infrastructure stability
  - Better query performance using fewer nodes, trust in the data, and faster insights from the data lake
- Increased satisfaction and reliability in information
  - Clusters now remains stable, whether always-on (Presto) or daily hourly use (Spark) ones
  - Fewer missed SLAs at peak and low demand periods
  - Improved user trust in query results and analytic insights

About Qubole

Qubole is revolutionizing the way companies activate their data — the process of putting data into active use across their organizations. With Qubole's cloud-native big data platform, companies exponentially activate petabytes of data faster, for everyone and any use case, while continuously lowering costs. Qubole overcomes the challenges of expanding users, use cases, and variety and volume of data while constrained by limited budgets and a global shortage of big data skills. Qubole offers the only platform that delivers freedom of choice, eliminating legacy lock in — use any engine, any tool, and any cloud to match your company's needs. Qubole investors include CRV, Harmony Partners, IVP, Lightspeed Venture Partners, Northwest Venture Partners, and Singtel Innov8. For more information visit www.qubole.com.

FOR MORE INFORMATION

Contact:
sales@qubole.com

Try Qubole for Free:
https://www.qubole.com/products/pricing/

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

WWW.QUBOLE.COM