Qubole®

2018

BIG DATA

ACTIVATION REPORT

FOREWORD

Welcome to the 2018 edition of the Qubole Big Data Activation Report, which presents insights from anonymous platform-level[1] data of over 200 organizations using Qubole Data Service (QDS). Activation is the process of putting data into active use. Specifically, activation puts data into the hands of any user who needs it, exactly when they need it, exactly in the way they're used to working with it.

This report represents a very exciting milestone, as it demonstrates how these companies have become data-driven worldwide, giving us a unique perspective and a benchmark for evolving our own organizations.

It's been a fantastic journey from the "old days" of hackathons at Facebook where Joydeep and I developed a *breakthrough* SQL-based declarative language that would allow data engineers to plug in their scripts and programs to obtain insights from "new" object data stores, when traditional SQL queries against data warehouses couldn't get us the insights we needed. Those were the beginnings of Facebook's transformation into a data-driven company.

Fast-forward 10 years, and we now live in a true big data world—according to IDC, the amount of data in the world doubles every two years, and by 2020 the digital universe will reach 44 zettabytes, or 4 trillion gigabytes. This growth isn't linear, nor is it of one type of data, but an ever-morphing mix of traditional enterprise alphanumeric data with video, audio, geolocation, social, streaming, IoT / machine data, and more.

What's so exciting—like our experience at Facebook—is to see how quickly organizations have understood the immense value of data and are building data-driven applications in search for new insights and revenue opportunities. It's no wonder why many believe we should treat data as an intangible asset.

This report provides clear evidence of how companies are activating their big data and continuously reducing costs by leveraging the compute power of the cloud and multiple open source big data engines and tools. For example:

- **Using the right tool for the job** — 76% of organizations actively leverage at least three big data open source engines for data preparation, machine learning, and reporting and analysis workloads (e.g. Apache Hadoop/Hive, Apache Spark, and Presto).
- **Self-service is the norm** — In aggregate, the number of commands run by users nowadays is vast, with over 58 million commands processed by users in the three main engines (Apache Hadoop/Hive, Apache Spark, and Presto) in 2017. The year-over-year increase in the number of users that ran commands is 255% for Presto, 171% for Spark, and 136% for Hadoop/Hive.
- **Greater productivity and automation is a priority** — The data shows that as the size of implementations grow, data-driven companies constantly seek to improve the ratio of users per administrator, thereby reducing the incremental cost of growth and increasing user self-sufficiency. For example, for small-scale implementations, on average there are 16 users per administrator; for medium implementations the ratio is 48 to 1; and for large-scale implementations the ratio rises to 188 to 1.

At Qubole we are passionate about helping organizations activate their big data, and with the first edition of this annual report we expect to continue that journey. We encourage you to dive into these numbers and use them as benchmarks for your organization. I would love to hear your comments and questions, so please contact me at bigdataactivationreport@qubole.com

Ashish Thusoo
CEO and Co-Founder
Qubole, Inc.

[1] Refer to section on Data Privacy and Security

REPORT HIGHLIGHTS

The findings in this report are based on anonymized data from over 200 Qubole customers. While the report does not represent the entire big data industry as a whole, the size of our customer base and their disbursement across both regions and verticals creates a representative sample size by which to draw conclusions.

Highlights from our report include:

- In 2017, 76% of companies deployed multiple big data engines.
- Since January 2017, total usage[2] across the three major engines has grown by 162%.
- Presto usage grew 420% year-over-year and Apache Spark grew 298%. Apache Hadoop/Hive 2 showed the greatest number of compute hours—over 45.4 million—and grew 102%.
- The magnitude of big data workloads is vast, having reached over 58 million commands processed only in Presto, Apache Hadoop/Hive, and Apache Spark in 2017.
- Comparing January 2017 to January 2018, the number of commands run on Apache Hadoop/Hive grew 129%, while the number of commands run on Apache Spark grew 439% and Presto by 365%.
- Customers in aggregate are running 24x more commands per hour in Presto than Apache Spark and 6x more commands than Apache Hadoop/Hive.
- Total usage of Amazon EC2 Spot Instances grew 4.6 times across all three engines since January 2016.

[2] Based on compute hours.

ABOUT DATA PRIVACY AND SECURITY

We are not only passionate about helping our customers activate their big data, but are equally focused on helping companies protect their intangible data assets and providing the peace of mind and trust that their deployments have the utmost security available today.

During the collection of data for this report Qubole anonymized all data to only reflect patterns in the utilization of open source data processing engines or tools, with the purpose of illustrating industry trends. Our data collection methods are designed to look at the aggregate usage of the Qubole Big Data Activation platform—Qubole Data Service (QDS)—and in no way specific customer or user data.

QDS does not store data, which remains in the customer's cloud provider account—data lake, data warehouse, or any other storage mechanism. QDS provides a single-tenant, serverless environment that isolates all data processing within customers' virtual private clouds or virtual cloud networks, regardless of the engine or tool used.

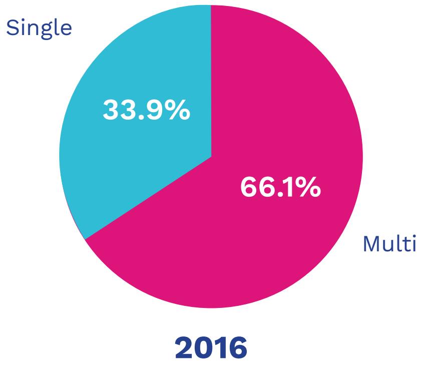
Qubole strives to maintain a secure and available Big Data Activation platform for our customers at all times. Our security team works around the clock to ensure the confidentiality and integrity of all QDS services, tools, and the overall platform.

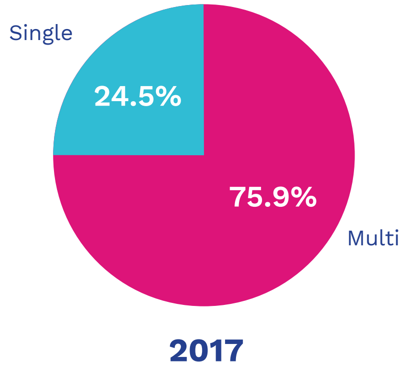
For more information on Qubole security, visit https://trust.qubole.com

USING THE RIGHT TOOL FOR THE RIGHT JOB

Beyond scale, the value of big data analytics in the cloud is its ability to address a variety of needs and types of data processing jobs. Emerging big data engines like Apache Spark and Presto have become increasingly popular for solving specific uses cases and, as a result, companies are moving toward multi-engine deployments. In 2016, the percentage of companies using multiple engines was 66.1%, and in 2017 it grew to 75.9%. Today, we are seeing that trend surpass 80%.

OPEN SOURCE BIG DATA ENGINES
Percentage of Companies Who Use Single Engine vs. Multiple Engines





Whether you have structured data, semi-structured data, unstructured data, or binary data, different engines such as Hadoop/Hive, Spark, and Presto can be used in different ways to process data. As you're evaluating engines, it's important to understand the characteristics of each and where they're most commonly found.

- **Structured data** has a defined schema and can be stored in a relational, columnar or linear formats. For example, data stored in a spreadsheet, or in tables of a RDBMS,[3] such as MySQL, Postgres, SQL Server, or Oracle.
- **Semi-structured data** has a loosely-defined structure that can change from one data tuple to another. JSON[4] logs collected to measure user engagement on a website is an example of semi-structured data.
- **Unstructured data** is any data that doesn't have a pre-defined data model or is not organized in a pre-defined manner. Common examples are human-readable text files, social media posts, the body of an email, or raw machine data that can't just be put into a database and analyzed.
- **Binary data** typically comes from information such as images, audio, and video that are decoded into numbers so they can be analyzed further. Binary data most commonly appears in deep learning and ML.

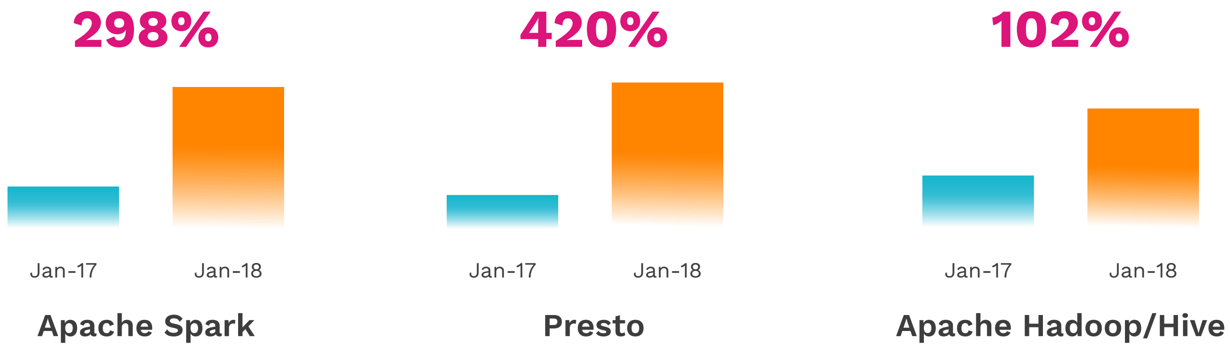
Each data type is unique and processing it comes with its own challenges. Picking the right tool becomes very important—not just related to its ability to process the data—but how efficiently it can be done. Therefore, it is important to understand the characteristics and strengths of each big data processing engine as well:

- **Apache Hadoop/Hive** is a workhorse for handling massive volumes of data for ETL (Extract, Transform, Load), ELT, or data preparation on structured and semi-structured information. The Hive Metastore also serves as a unified schema for data lakes, providing users with schema-on-read capabilities in order to query data easily.
- **Apache Spark** is powerful for processing complex and memory-intensive workloads. Whether you are creating and processing data pipelines or implementing machine learning, Apache Spark is great for both structured and unstructured analysis. Most importantly, running this engine on a cloud data activation platform enables data scientists to use various machine learning packages and distribute their models for training or production across petabyte data sets.
- **Presto** shines in interactive analysis such as with business intelligence (BI) or data discovery tools, when data is in a semi-structured or structured form. Presto can join terabytes of data in seconds, or cache queries intermittently for a rapid response upon later runs. The engine's in-memory scale-out capabilities are very powerful, particularly for analytics workloads of data that is transformed into a columnar formats such as ORC or Parquet.

[3] Relational Database Management System
[4] JavaScript Object Notation

COVERING ALL BIG DATA ANALYTICS USE CASES

Total Engine Usage Globally by Compute Hours (YoY Change Jan ‘17–Jan ‘18)



As companies are collecting more and more data, they’re also increasingly looking for ways to implement a variety of analytics and mining in order to extract the most value from their data. In fact, 76% of respondents in our Qubole DataOps Survey said their company currently has a big data initiative, and another 20% said they plan to soon. This trend explains why we see a growing usage of multiple big data engines.

Since January 2017, usage (based on compute hours) across all three engines has grown by 162%, with Hadoop/Hive remaining the most used engine, followed by Spark and then Presto.

As discussed previously, each big data engine has both strengths and weaknesses based on the type of data you’re working with and what you’re trying to do with it. This is why the data shows different growth rates when you compare commands alone, compute hours alone, or commands per compute hour. Year-over-year usage of Presto grew 420% and Apache Spark grew 298%. Apache Hadoop/Hive 2 showed the greatest number of compute hours—over 45.5 million—and a growth rate of 102%.

With the increasing importance and business value of machine learning (ML) today, organizations are using various technologies for this purpose, ranging from open source ML libraries and tools, as well as workflow and scheduling tools. What we hear is of particular importance, is the ability to combine the right tools for the job in a single platform, in order to cover all use cases.

One particular framework—TensorFlow—has become quite popular for machine learning, as enterprises and startups alike embed more deep learning and lazy training models into their product / service offerings, and internal analytics. It is based on numerical computations that use data flow graphs to represent mathematical operations, while graph edges represent tensors that flow between them.

BIG DATA TOOLS TO WATCH

Workflow & Scheduling:

Over 29.6% of companies have used Apache Airflow for orchestrating sophisticated data preparation pipelines and operationalizing machine learning using Python code. It offers unique capabilities for monitoring jobs, handling failures, and managing dependencies.

Machine Learning & Artificial Intelligence:

XGBoost is one of the most popular ML libraries for building models used for predicting everything from real estate prices to efficacy of medical treatments.

Pandas is a Python-based data science tool used for statistical analysis, predictive modeling, and many other ML applications. The library can support either batch or real-time workloads with Apache Spark.

MLLib is Apache Spark’s ML library based on Scala. MLlib fits natively with Spark’s APIs and can operate with NumPy in Python and R libraries. It also works with many Hadoop data sources, making it easy to add into other big data workflows.

TensorFlow makes it easier to distribute workloads, and is commonly used in image recognition or speech detection use cases.

SELF-SERVICE IS THE NORM

When Efficiency Matters

New and ever-growing data means different engines and analytical use cases are needed to meet the development requirements of the business. As discussed in the previous section, certain engines perform better for specific use cases. Presto works well for interactive queries, data discovery and BI; Spark is great for machine learning and data mining; and Hadoop/Hive performs well in ETL and interactive analytics.

These days, big data resources are at a premium at most companies, and data teams are stretched thin. That's when efficiency matters. We're seeing companies follow the "right engine for the right job" approach to become as efficient as possible on every workflow so that their data teams can focus on the most high-value projects.

THROUGHPUT: Increase in Monthly Average of Commands per Compute Hour (2018 over 2017)

140%	**101%**	**124%**
Apache Spark	**Presto**	**Apache Hadoop/Hive**

A good way to evaluate efficiency is to compare the change in the monthly throughput. Between 2017 and 2018, the average monthly number of commands per compute hour grew by 140% for Apache Spark, 101% for Presto, and 124% Apache Hadoop/Hive. When we looked at the total number of commands run, Apache Hadoop/Hive grew YOY by 129%; Apache Spark 439% and Presto 365%. Even more impressive still, customers in aggregate are running 24x more commands per hour in Presto than Apache Spark and 6x more commands than Apache Hadoop/Hive.

YOY Change in Total Number of Commands Run (YOY Jan '17–Jan '18)

439%	**365%**	**129%**
Apache Spark	**Presto**	**Apache Hadoop/Hive**

These figures not only demonstrate significant growth, but also the differences in usage by engine. For example, the impressive growth in Presto suggests greater efficiency and self-sufficiency, as it is a powerful tool for use cases with interactive and ad hoc SQL analytics where joins and highly concurrent simple queries are common. This engine doesn't require a user to constantly tune cluster configurations, thereby reducing drastically the time-to-insights. Compute is continuously growing to keep up with the data and, in order for data teams to provide consistent answers, companies must have the right infrastructure in place in order to meet high and unexpected volumes of information in order to service their customers reliably.

Increasingly Self-Sufficient Users

Today virtually every organization has an analytics group or a data team that supports other functional areas with data insights. Traditionally, these groups have been the sole conduit to data analyses, particularly with greater uses of big data where specialized skills, multiple processing engines, and a variety of tools are required.

These teams are proven to be more productive when they have direct access to a modern data activation platform that provides self-service access to multiple big data engines for as many users as possible. Cloud data lakes have also become a key element in companies' big data infrastructure strategy, because they can store all of their data in inexpensive storage buckets with high availability. This separation of storage from compute results in efficient use of resources, since compute resources are used only when data needs to be processed.

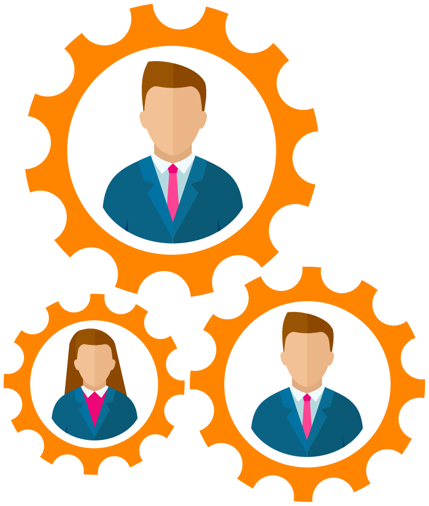
In the last year we have seen a significant increase in self-service, measured by the number of users that run data processing or query commands in different engines.

So not only are we seeing high growth rates in adoption and efficiency of data processing in multiple engines, but a significant increase in self-service access for many users.

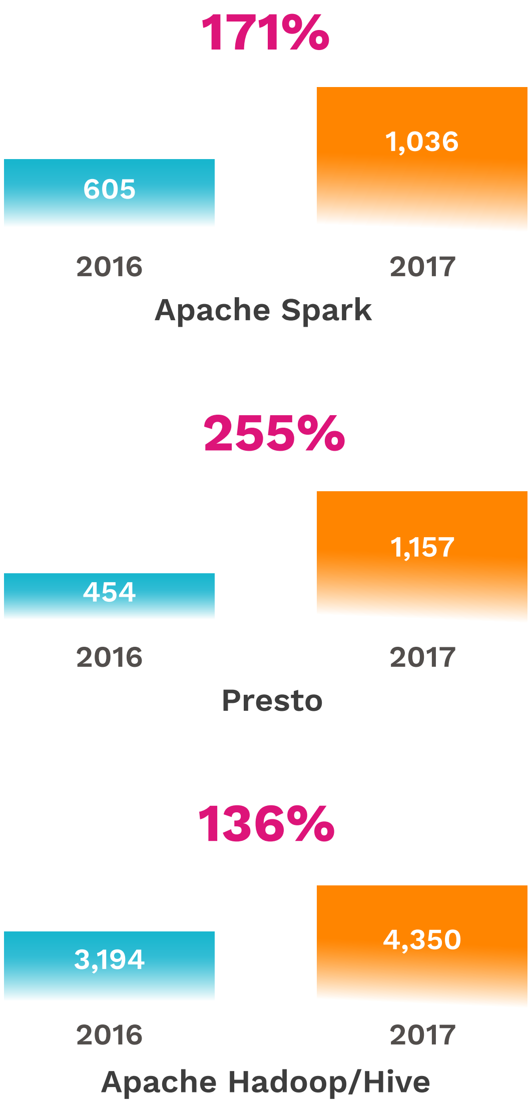
These findings are consistent with the trend in hiring data scientists, which has grown over 650%[5] since 2012 to approximately 35,000 people in the US with data science skills. However, there are still over 190,000 unfilled data-related jobs in the US alone and hundreds of companies hiring. This huge skills gap results in companies constantly looking for ways to enable as many users as possible users with self-service access to big data analytics engines.

With self-serve data access, analytics and data teams are able to spend more time on higher-value tasks, such as uncovering previously hidden insights, identifying new revenue streams, improving the user experience, or modernizing their processes, with minimal intervention from the DataOps or DevOps teams.

Having a cloud big data activation platform with unlimited scale also enables data engineering and DataOps teams to focus on improving the performance and reliability of their data pipelines, rather than maintaining and patching legacy on-premises systems. In addition, it enables more and more people to identify, define, and solve business problems using real data-based insights, instead of guesswork or not taking action because they lack visibility into a problem.



Increase in No. of Users That Ran Commands in Each Engine



[5] Source: https://goo.gl/1ShvHx

Greater Productivity and Automation are Priorities

Today we generate and capture more and more data daily from many different sources. This exponential growth in volume of data and data types means that use cases and user expectations are ever growing and changing. But IT budgets and availability of big data skills certainly don't follow the same trends, so productivity and automation are top priorities—for data preparation, processing, discovery, and analysis alike.

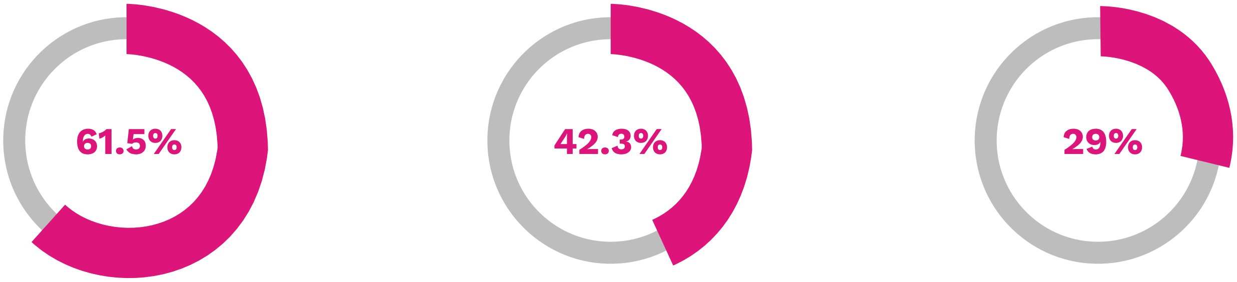
Increased pressure to drive down costs and more demand for data from more users are forcing data teams to look for ways to make users more self-sufficient. Looking at Qubole's customer base as a benchmark, small-scale implementations have an average of 16 users per administrator; medium implementations have ratios of 48 to 1; and for large-scale implementations the ratio rises to 188 to 1.

> **54% of all Amazon EC2 compute hours used were spot instances**

Another important way to reduce operational costs is to use spot instances[6], which allow data teams to bid for and purchase unused Amazon EC2 computer capacity at a highly-reduced hourly rate, set by supply and demand. Given the seasonality and unpredictable bursty nature of big data workloads, this type of infrastructure enables data ops to quickly scale up or down cost-effectively as analytics needs change.

It's no surprise that in 2017, 54% of all Amazon EC2 compute hours used were spot instances. This has resulted in an estimated *$230 million in savings of Amazon EC2 costs* through spot instances leveraged by Qubole's workload-aware autoscaling of big data workloads.

Percentage of Spot Instances per Cluster for each Engine



Apache Hadoop/Hive
Spot instance usage increased 6.33x in 2017

Apache Spark
Spot instance usage increased 2.66x in 2017

Presto
Spot instance usage increased 4.1x in 2017

Consider a hypothetical scenario where you are running a large cluster with 100 nodes for extraction-transformation-loading (ETL) workloads using Apache Hadoop/Hive, which you run for ten hours a day. Your consumption would amount to approximately 1,000 compute hours. The Qubole data shows that the greatest cost optimization and reliability is achieved by leveraging 60-85% of spot instances in a given cluster. If we apply these guidelines to our hypothetical example, your organization would be able to consistently cut costs in the range of 50-80% for your ETL workload[7].

It is important to note that having spot instances available is not a guarantee of cost optimization unless the data platform is intelligent enough to autoscale, first, to take advantage of different capacities of nodes within a heterogeneous cluster, and second, to use spot instances when available[8]. In addition, the cost optimizations will vary by type of data processing job and the engine or tool used.

For example, compared to other engines, Hadoop/Hive 2 jobs can benefit from spot instances because tasks can be reprovisioned easily to other on-demand or spot instances if the original spot node is taken away. Whereas Presto and Apache Spark require different heterogeneous cluster and spot configurations for distributing jobs efficiently since they run in memory.

[6] A service offered by Amazon Web Services, https://aws.amazon.com/ec2/spot/

[7] Refer to the Qubole blog for further information on cost savings https://www.qubole.com/blog/the-value-of-auto-scaling/

[8] For detailed information on heterogeneous clusters and spot instances, refer to http://docs.qubole.com/en/latest/user-guide/clusters/heterogeneous-clusters.html

CONCLUSION

Big data is no longer just about corporate data from enterprise systems behind or outside the firewall. On average, a person will interact with or contribute to big data at least 30 times per day. It's coming from your smartphone, favorite show, your car, home, online accounts, security systems, cab rides, devices, and much more. Being competitive in today's age means incorporating the right technology to uncover new insights and make critical business decisions that leverage all of this ever-growing and new information.

The promise of big data technologies is to allow people to process of all kinds of data sets to be brought together and analyzed—whether unstructured, semi-, or structured. The goal is to derive rich new insights for informing critical business decisions, enhancing products, devising new ways to increase revenues, or optimizing costs. This is what we call "big data activation"—the process of putting data into active use.

However, there is a huge gap in skills and the number of big data professionals, which puts great pressure on analytics and data teams to continue meeting their service level agreements (SLAs) for activating their data. This is where a big data activation platform comes in—its job is to unlock the full utility of all data, making it accessible by all users for any use case, exponentially amplifying the work of data scientists, data engineers, and end users alike.

The facts and insights from this report clearly illustrate how a big data activation platform enables self-sufficiency by providing the right tool for the job for as many people as possible within the same environment, combined with the highest productivity and lowest TCO possible. The inability to afford access and use of all your data is no longer a barrier to business velocity.

AUTHORS: José Villacís, Holden Ackerman, Steve Biondollilo, Ashish Thusoo

CONTRIBUTORS: Daniel Leybzon, Farid Mehovic, Kristin Crosier, Jorge Villamariona, Matheen Raza, Kevin Kennedy, Balaji Mohanan

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 Activation 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's intelligent automation and self-service supercharge productivity, while workload-aware auto-scaling and real-time spot buying drive down compute costs dramatically. 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, Norwest Venture Partners, and Singtel Innov8. For more information visit **www.qubole.com**.