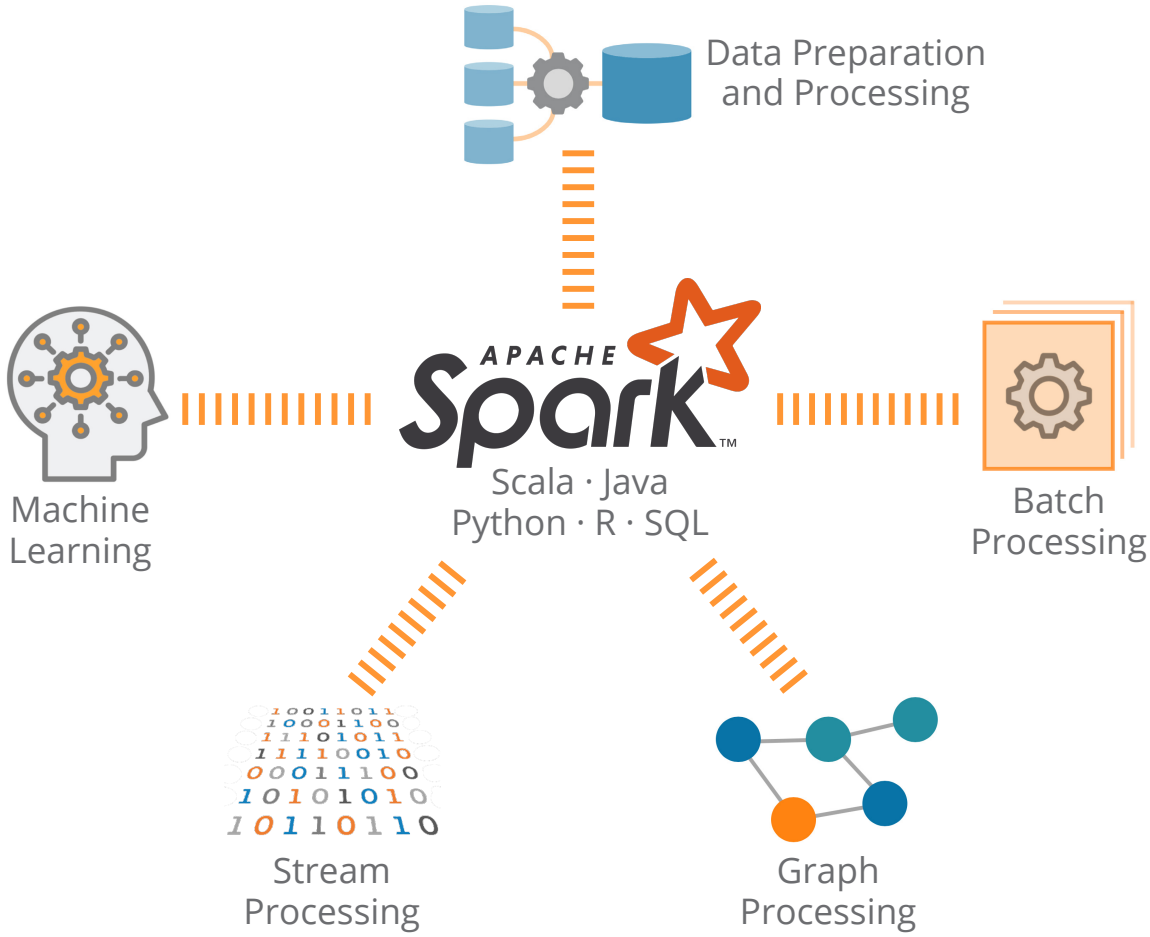

Qubole
APACHE SPARK AND QUBOLE
ACCELERATING TIME TO VALUE OF BIG DATA
January 2019

What Is Apache Spark?

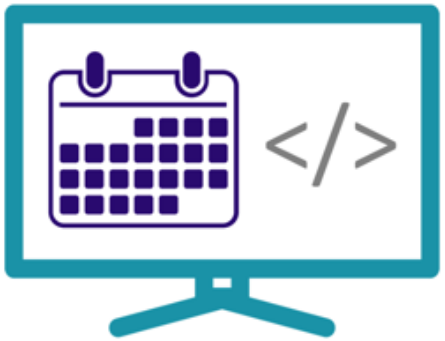
Apache Spark is a high-performance, distributed data processing engine that data practitioners prefer for handling big data workloads. First developed at the AMPLab at UC Berkeley, Spark has become a widely adopted framework for machine learning, complex data processing, advanced analytics, and other big data projects.



A Vibrant and Engaged Community Fuels Spark's Growth

Spark's vibrant community of contributors prioritizes agility, flexibility, and scalability through hundreds of code deployments per month. Spark is the leading big data framework in use, with a usage increase of 29% over 2017.

1,700+
CONTRIBUTORS

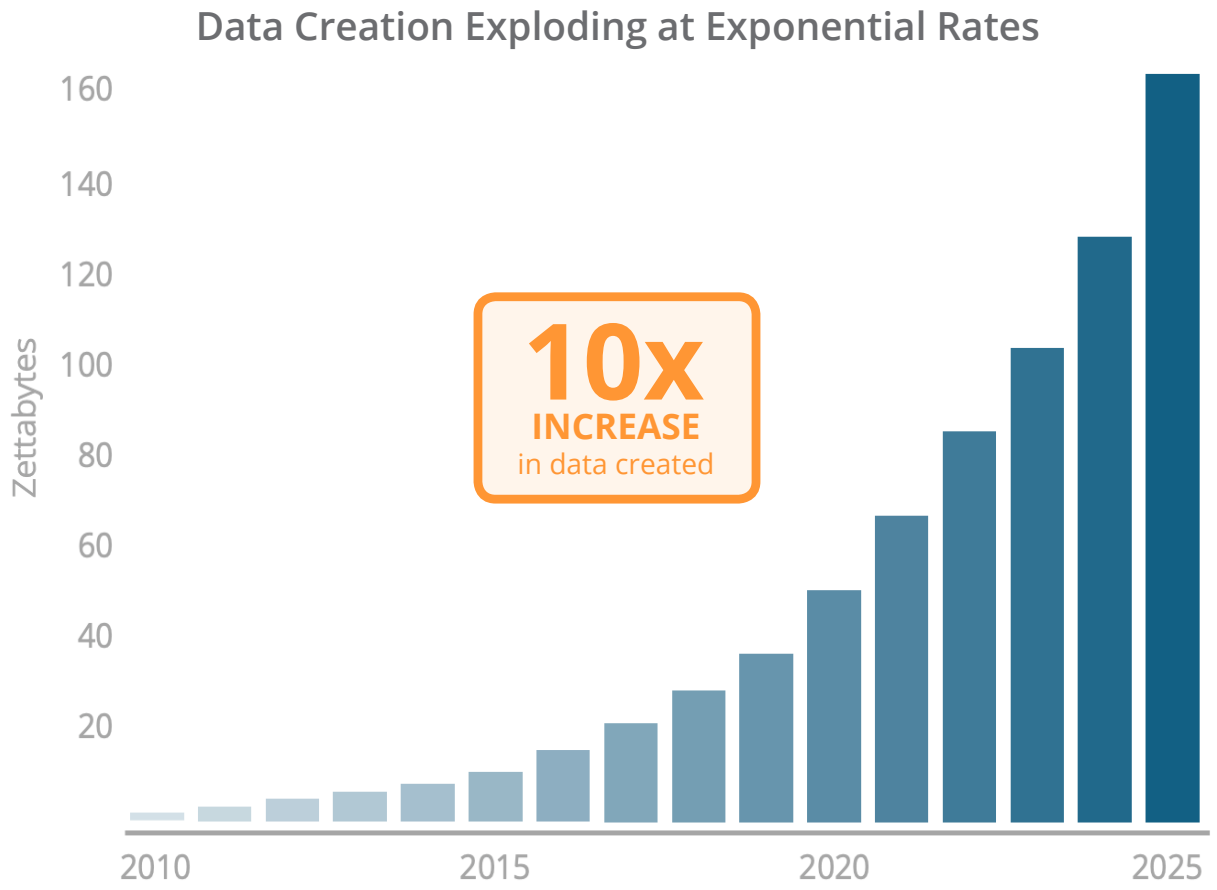


~400
CODE COMMITS MONTHLY

* Source: Big Data Survey, 2018

Data Growth Drives the Need for Spark

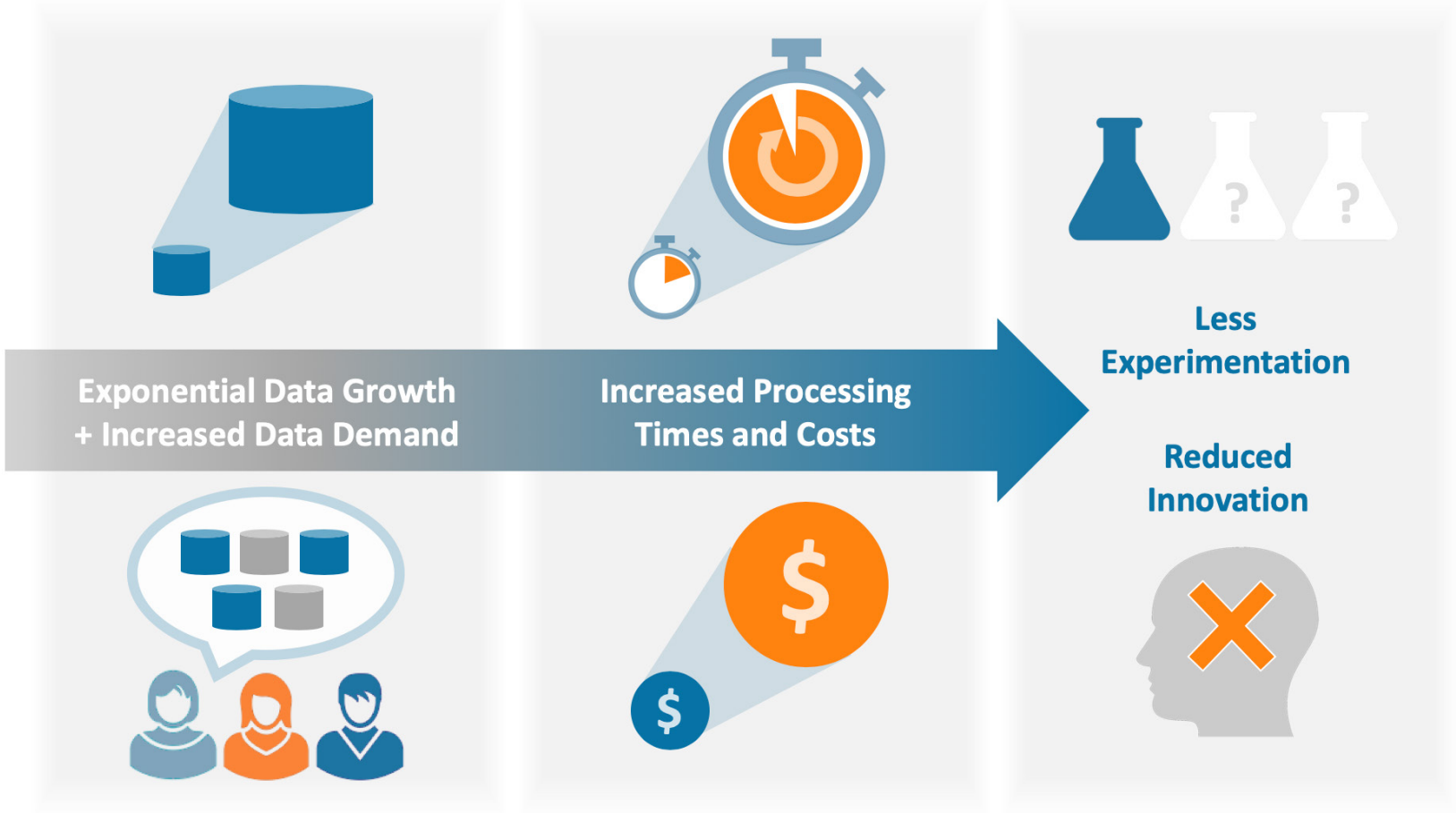
As networked devices, instruments, and applications proliferate globally, enterprises are capturing and processing ever-increasing mountains of data. Businesses must access and integrate information from numerous sources to develop the insights and analytics required to drive business strategies and make effective decisions.



Source: IDC's Data Age 2025 Study, sponsored by Seagate, April 2017

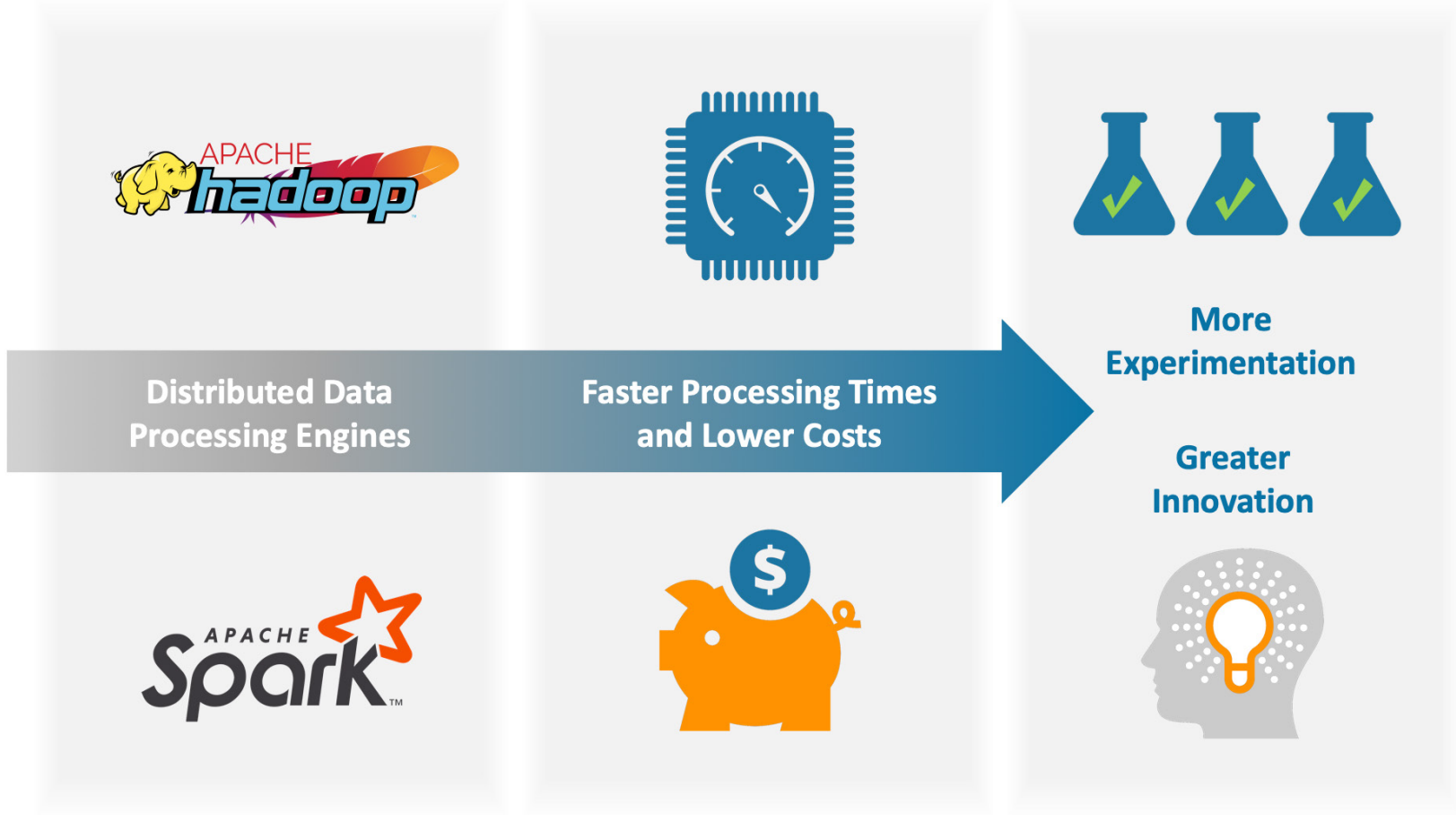
Intensive Processing Is Required to Create Usable Datasets

Traditional tools are not suited to meet user requirements of data today. Saddled with the additional cost of inefficient data processing, organizations have far less time and resources for experimentation. The additional cost of experimentation cripples innovation and reduces the return on investment on big data investments.



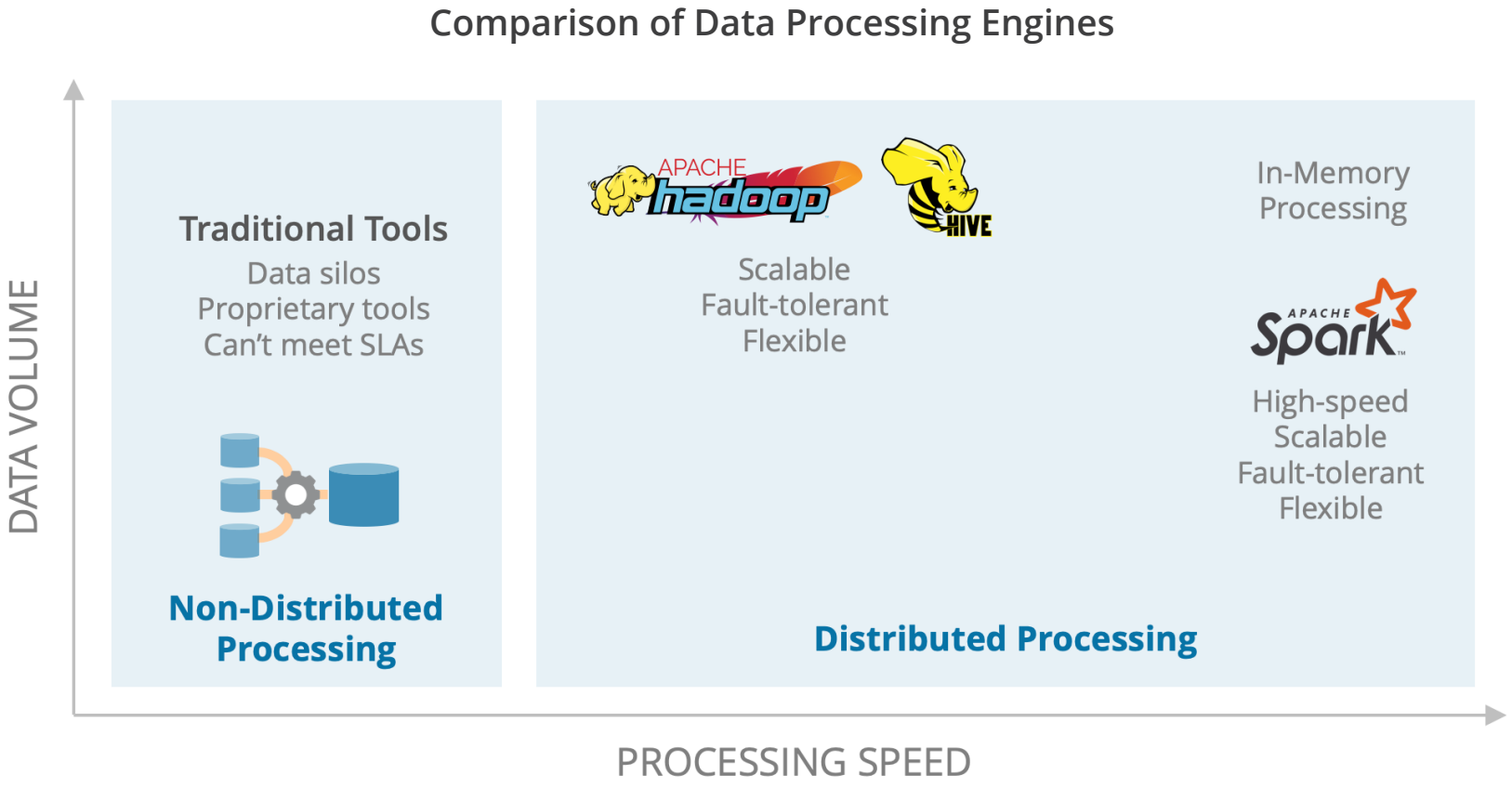
Distributed Processing to the Rescue

The inability of traditional tools to keep pace with the explosion of data has led to the emergence of distributed processing engines — such as Hadoop and Spark — that split the data into smaller, manageable chunks and process it across multiple computing nodes. Distributed engines greatly improve processing times and enable a wide spectrum of use cases in machine learning and big data analytics, which in turn lead to more experimentation and greater innovation.



The Evolution of Distributed Engines

Both Hadoop and Spark are distributed processing engines, but they're not interchangeable. While Spark improves the speed of data processing significantly, Hadoop can still process larger volumes of data. You should probably pick the one that most closely addresses your specific use case, workload type, and data volume.



Qubole

Apache Spark on Qubole

Customers trust Qubole to process and manage their Spark workloads. Qubole's enhancements improve cost efficiency, performance, and usability while ensuring the security and reliability required by enterprise applications.

Qubole

Apache Spark

Performance

Usability

Security

Reliability

Enterprise-Ready

Lower Costs

Qubole boosts the power of Spark to exceed the demands of enterprise workloads

Qubole

Qubole Reduces Computing Costs of Apache Spark

Cloud infrastructure costs can quickly spiral out of control. **Qubole's advanced cost controls have enabled our customers to see as much as a 50% reduction in their cloud computing costs.**

Workload-Aware Autoscaling

Advanced SLA-based scaling algorithm determines the exact number of executors to optimize resource utilization.

Aggressive Downscaling

Decommission idle and unneeded compute nodes. Use container packing to downscale with higher resource utilization.

Heterogeneous Clusters

Mix different instance types in the same cluster to create significant savings and more reliable clusters.

Intelligent AWS Spot Management

Automatically bids on Spot Instances and rebalances Spot Nodes. Utilizes low-cost compute resources without causing cascading failures.

Qubole Improves Apache Spark Performance for Big Data Workloads

Big data processing efficiency depends on performance, especially read/write and data operations speed. As a standalone, open-source solution, Spark is not inherently optimized for enterprise big data workloads.

Qubole adds performance optimizations and smart management tools that extend the power of Spark to even the most complex big data problems.

Direct Writes

Faster write throughput when writing to Amazon S3, alleviating the need to stage writes before committing them.

Join Optimizations

Significantly improve Spark performance of join operations on large datasets.

Fast Caching with RubiX

Platform-wide caching layer reduces I/O latency and speeds up data engines.

Performance Optimization

Added visibility and optimal configuration recommendations for Spark applications.

Qubole Improves the Usability of Apache Spark

The open-source version of Apache Spark is powerful, but extremely complex for anyone who's not an expert Spark developer.

By handling back-end configuration issues and automating day-to-day processes, Qubole makes the Spark world approachable for data engineers, data analysts, data scientists, and administrators.

Multiple Interfaces

Launch Spark jobs via Analyze, Notebooks, or API interfaces as required.

Workflow Automation

Schedule Spark jobs or leverage Airflow to build end-to-end pipelines.

Spark Clusters

Automate provisioning, scaling, and termination to optimize compute resources.

Package Management

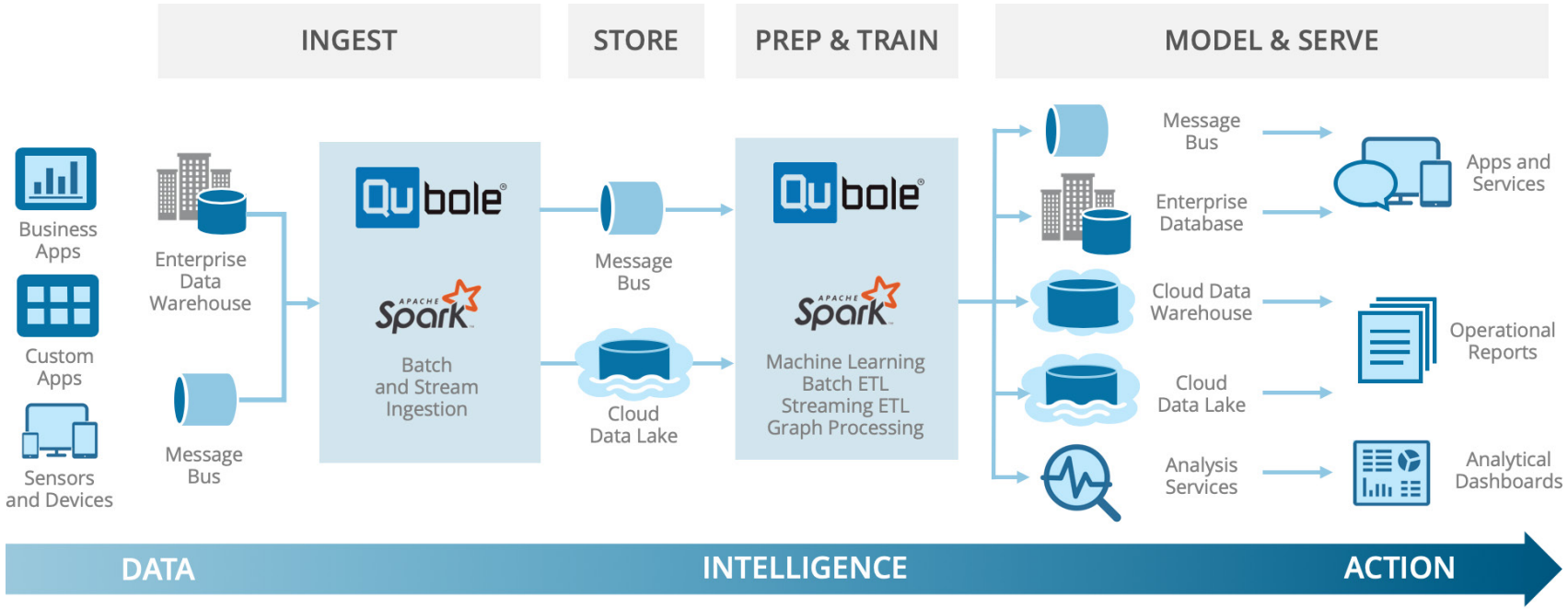
Auto-distribute Python and R packages using predefined dependencies.

Spark UI

Simplify and speed debugging of problematic jobs.

Apache Spark on Qubole: Data to Intelligence to Action

Apache Spark on Qubole supports the ingestion, preparation, integration, transformation, and analysis of data coming in from sources across your extended enterprise — and converts it into actionable intelligence that arms users with the insights they need to make decisions for maximum impact.



Users Trust Qubole for Their Spark Workloads

Qubole's industry-leading, distributed-computing platform delivers efficiencies for departmental applications to the largest of Spark clusters in distributed cloud-computing environments.

Qubole Provides Industrial-Strength Scale and Reliability

Support for even the largest clusters

750+

CONCURRENT NODE CLUSTERS*

Unbridled Growth in Spark Usage on Qubole

2018 annual increase in Spark commands

+439%

INCREASE OVER 2018*

* Source: 2018 Qubole Big Data Activation Report

Qubole®

Case Study: Return Path

Return Path uses Qubole to deliver self-service analytics, simplify infrastructure, reduce costs, improve team productivity, and accelerate time-to-value on data science projects.

Reduced cloud compute costs

Higher productivity

Increased innovation

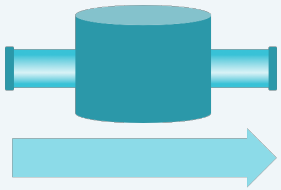
Greater customer satisfaction

"Qubole helped prevent us from making bad decisions that would have cost the business tens or hundreds of thousands of dollars."

Robert Barclay, VP of Data and Analytics, Return Path

Test Drive Apache Spark on Qubole Today

The #1 Cloud-Native Data Platform for Machine Learning and Analytics



Build data pipelines with ease

Bring machine learning to production

Analyze any type of data from any source

Take Qubole for a test drive today. See how data-driven industry leaders work smarter and slash cloud costs with Qubole.

Adobe® lyft malaysia airlines COMCAST

Start Your Qubole Test Drive Now

To learn more, visit qubole.com

Qubole®