DATA SHEET

Qubole Pipelines Service –
A Complete Stream Processing Service

Manage streaming ETL pipelines with zero overhead of installation, integration or maintenance

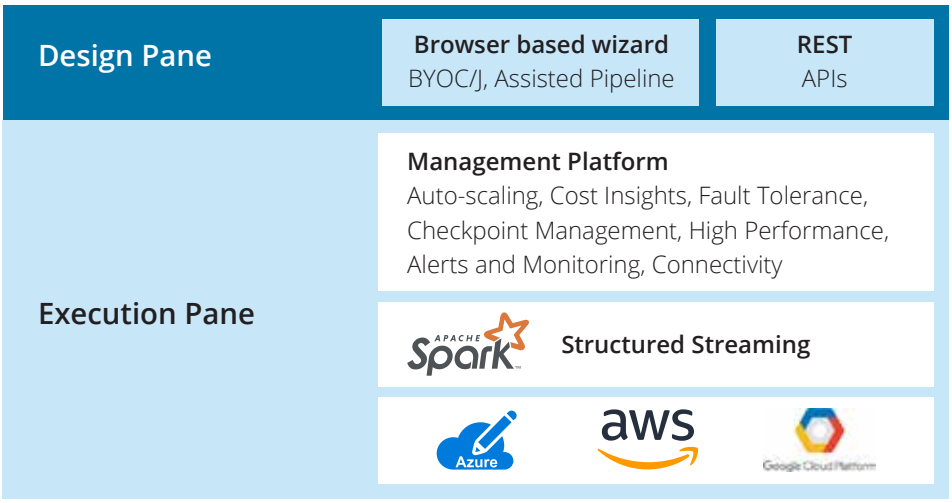
When it comes to data analytics, lowering end-to-end time-to-decision matters. Building pipelines to simply dump all data to a data lake and then indexing and warehousing is relatively easy but that, by itself, does not reduce the time-to-decision as there are additional processing steps required. Today, organizations are moving towards real-time processing solutions which help them deliver insights in a matter of seconds to minutes. Although this brings in significant business benefits it is poses additional challenges.

Qubole Pipelines Service solves the complexities in data processing that arise when dealing with data streaming. Qubole Pipelines Service is an enterprise-grade stream processing platform built on the highly fault-tolerant, scalable and performant Apache Spark Structured Streaming. It is supported on Apache Spark 2.3.x and above. Currently it is supported on AWS and Google Cloud.

Why Qubole Pipelines Service?

Lower time-to-market with productivity gains	Advanced machine learning platform completely under one roof	High Performance / Lower TCO
Entire event processing SDLC made easy. Lower time to value with rich user experience, REST APIs and connectors to real-time systems to quickly move from experimentation to testing to deployment to the management of long-running business-critical pipelines.	Data engineers and data scientists leverage the power of Apache Spark to build, train, and deploy MLlib models and create consumer applications that take advantage of derived inferences in real-time.	For a long-standing job where a steady-state is the norm rather than the exception, provide insights on optimizing Spark configurations to minimize cost without sacrificing SLA. For handling ebb and flow in data loads, the cluster will auto-scale based on workload. There are significant improvements in fault tolerance and performance when building stateful streaming applications such as de-duplication, pattern detection etc.

Qubole Pipelines Service Architecture



Qubole Pipelines Service allows data engineers to design streaming data pipelines either using a Wizard-based UI or through Qubole's REST API.

Qubole Pipelines Service offers additional productivity and reliability gains with fault tolerance, checkpoint management, alerting and monitoring while also taking advantage of Qubole's native workload aware autoscaling and cost insights.

Qubole Pipelines Service is built on Apache Spark Structured Streaming and is available on multiple public clouds.

The following table shows the enhancements that Qubole Pipelines Service provides over Apache Spark Structures Streaming.

Features	Apache Spark Structured Streaming	Qubole Pipelines Service
Operations at Scale		
End-to-end exactly-once semantics	✓	✓
Handle Late Arrival of data	✓	✓
Reliability of production pipelines Retry on Failures		✓
Management of error records Error sink to collate and reprocess errored out records		✓
Prevent Disk errors Logs rolling and aggregation on the underlying file system to prevent out-of-disk errors		✓
Security ACLs on streaming jobs		✓
Cluster Upgrade Reliability and zero downtime when upgrading cluster		✓
Alerts to multiple channels Alerts on failure to slack, email, Pagerduty, webhooks		✓
Real-Time Monitoring + Dashboard Integrated Monitoring with Prometheus and Grafana		✓

Features	Apache Spark Structured Streaming	Qubole Pipelines Service
Performance and TCO		
Compaction of small files Achieved through integration with HIVE ACID as sink		✓
Managed Streaming Cluster Balance it as per your streaming throughput requirements which are different than batch		✓
Cost and Health of pipeline helping to lower TCO Given SLA and cost, Indicate if a pipeline is balanced, needs attention, or at risk of failure		✓
Auto-Scale Qubole's Spark Auto Scaling cluster to handle burst and long inactivity		✓
Stronger Fault tolerance @ lower cost S3 eventual consistency Issues when checkpointing		✓
Pluggable State Storage Optimize performance of stateful streaming queries backed by RocksDB		✓

Features	Apache Spark Structured Streaming	Qubole Pipelines Service
Software Development Life Cycle (SDLC) Productivity		
Streaming De-Duplication	✓	✓
Stream-Stream Join	✓	✓
Operations on sliding event-time window	✓	✓
Input Sources	File Systems, Kafka	File Systems, Kafka, Kinesis Streams, S3-SQS
Output Sinks	File Systems, Kafka, Console, Memory	File Systems, Kafka, Console, Memory, HIVE, HIVE ACID, Snowflake, Big Query, MongoDB, ElasticSearch, Druid
Build Pipeline in few clicks Build pipeline using simple wizard to auto-generate code. Advance with your own code Out of Box Operators: Select, Filter, Aggregation with Window, Watermarking		✓
Upgrade application with zero disruption		✓
Pipeline Clone, Delete, Archive Ease of collaboration and management		✓
Schema Management Auto-fetch schema in AVRO format from Kafka Registry		✓
Test framework with strong debugging Time series view with results and appropriate logs		✓

About Qubole

Qubole provides a cloud-native data platform for analytics and machine learning that quickly activates large quantities of data for all users while lowering costs. Built by the team who built Facebook's data platform, Qubole serves some of the largest data-driven companies such as Lyft, Expedia, Box, and Oracle.

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

Try Qubole for free at
https://testdrive.qubole.com/