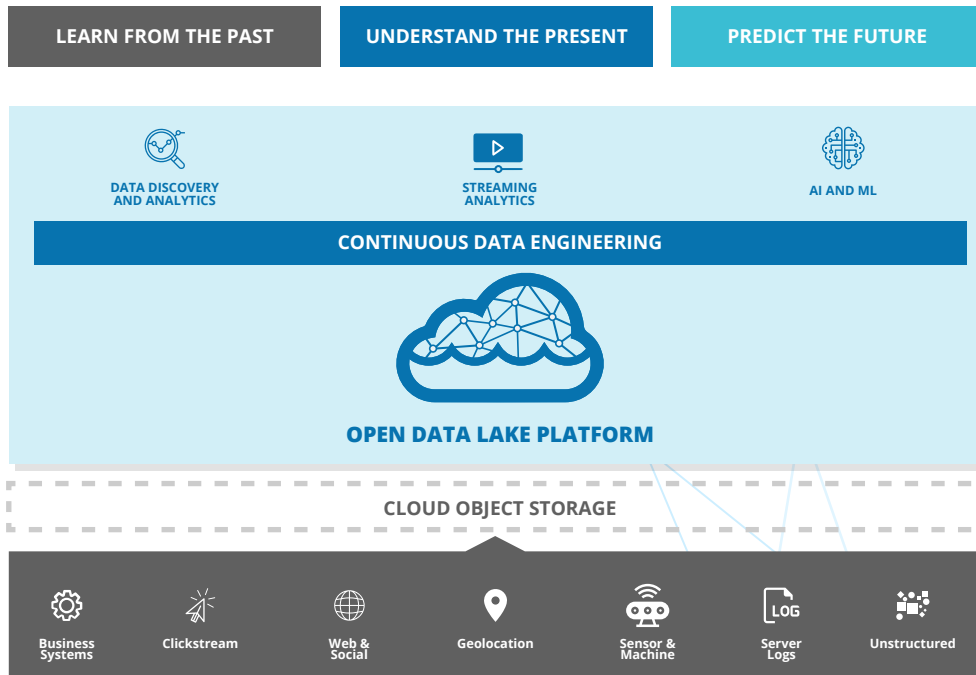# THE OPEN DATA LAKE PLATFORM

The exponential growth of data, combined with analytics and machine learning (ML) applications, calls for an open data lake architecture. One that ensures openness to data storage, data management, data processing, operations, data access, governance, and security while supporting a diverse range of analytics. An open data lake provides a robust and future-proof data management paradigm to support a wide range of data processing needs, including data exploration, ad-hoc analytics, streaming analytics, and machine learning.

LEARN FROM THE PAST	UNDERSTAND THE PRESENT	PREDICT THE FUTURE

DATA DISCOVERY AND ANALYTICS **STREAMING ANALYTICS** **AI AND ML**

CONTINUOUS DATA ENGINEERING

OPEN DATA LAKE PLATFORM

CLOUD OBJECT STORAGE

Business Systems Clickstream Web & Social Geolocation Sensor & Machine Server Logs Unstructured

Qubole is an open data lake platform that is simple and secure for enterprises working with their trusted and raw datasets in data lakes.

OPEN

Qubole stores data in multiple formats that are accessible through open standards-based connectors and APIs. It is agnostic to cloud platforms (AWS, GCP, Azure), open-source frameworks (Presto, Apache Spark, Apache Hive, Apache Airflow), or data file formats such as ORC or Parquet.
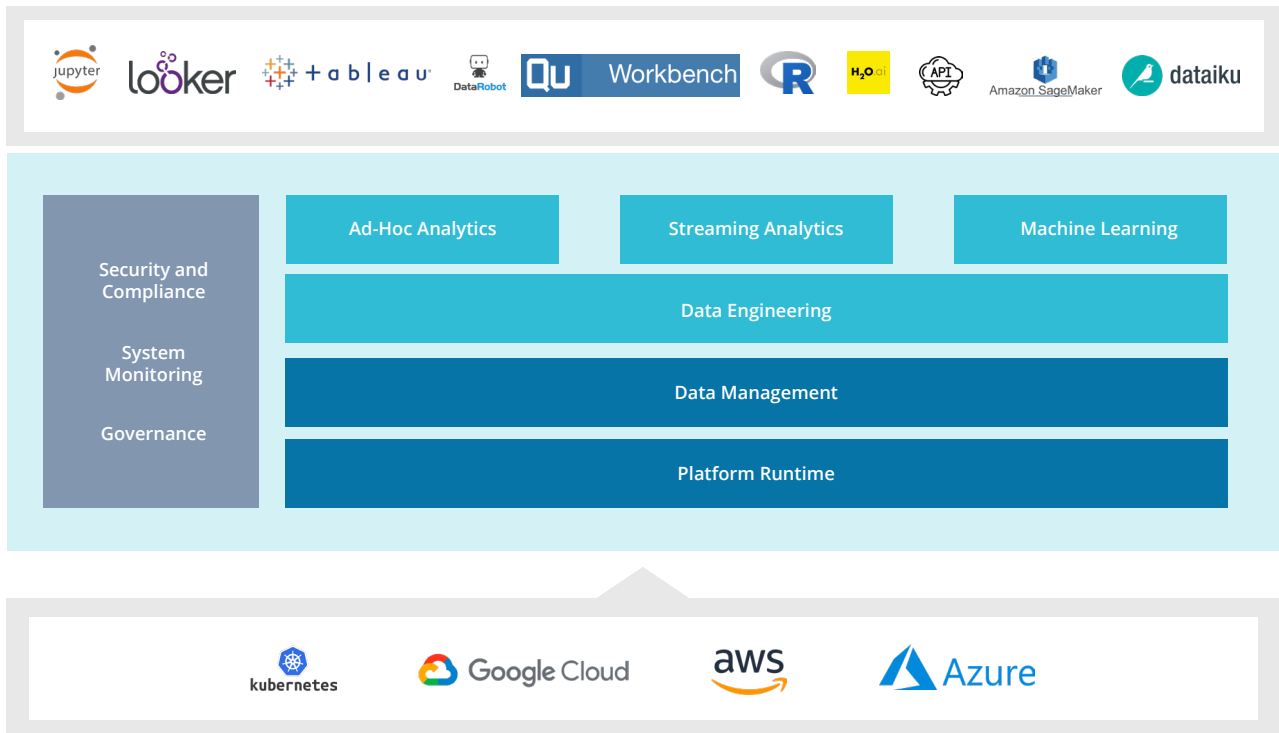
SECURE

Qubole provides robust data access controls and security features through non-proprietary technologies and APIs. The platform facilitates table, row, and column level granular security enabling regulatory compliance (GDPR and CCPA). Security and policy administrators can grant permissions against already-defined user roles in enterprise directories such as Active Directory, Google Cloud Identity Management.
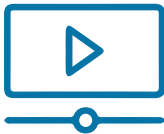
SIMPLE

Qubole enables near-zero administration, a unified environment for data pipeline creation, and robust orchestration for continuous data engineering. It allows collaboration through a common workbench, shareable notebooks, dashboards for data scientists, data engineers, and data analysts.

Qubole Open Data Lake Platform reduces time, effort, and cost for Data Exploration, Ad-hoc Analytics, Streaming, and AI/ML. The platform abstracts away the underlying complexities at multiple levels shown here.

Data Exploration and Ad-hoc Data Analytics

Utilize the built-in Workbench to author, save, collaborate, and share reports and queries. Develop and deliver ad-hoc SQL analytics through optimized ANSI/ISO-SQL (Presto, Hive, SparkSQL) and 3rd party tools such as Tableau, Looker, and Git.

Streaming Analytics

Build streaming data pipelines, combine with multiple streaming, and batch datasets to gain real-time insights. Collect and process stateful events, replay and reprocess data, and integrate with monitoring and alerting solutions.

Machine Learning

Build, visualize, and collaborate on ML models. Use offline editing, multi-language interpreters, and version control to deliver faster results. Leverage Jupyter or Qubole notebooks to monitor application status, job progress, and use the integrated package manager.

Data Engineering

Work with the raw data in the data lake and build data pipelines with the processing engine and language of choice - Apache Spark, Hive, Presto using SQL, Python, R, Scala. Easily create, schedule, and manage data engineering workloads for continuous data engineering.

Data Management

Leverage and manage automatic metadata creation, discover and explore data dependencies, and leverage indices and data views to improve workload performance and response time. Avoid lost updates, dirty reads, stale reads, and enforce app-specific integrity constraints with built-in ACID transactionality support across engines.

Platform Runtime

Capitalize near-zero administration on a highly available platform with 20+ runtime services. Plus, reduce cloud infrastructure cost by 50% using automated platform-wide capabilities like Workload-aware Autoscaling, Workload Packing, Heterogeneous Cluster Management, Intelligent Spot Management, and Automated Lifecycle Management.

Data Governance, Security and Compliance

Protect data and secure access with encryption and RBAC controls. Leverage native integration with IAMs, AD, and LDAP implementations to provide consistent data access privileges. Comply with the right to forget and right to erase guidelines in GDPR and CCPA with efficient updates and deletes in your data lake.

Qubole's Open Data Lake Platform includes Premium Support; 24x7 access to support engineers, business hour access to solution architects via online case submission, email correspondence, and phone or live. Qubole provides a range of professional services to obtain the most value from your data lake from intensive task-specific sessions. Qubole is trusted by leading brands such as Expedia, Disney, Oracle, Gannett, and Adobe to spur innovation and to transform their businesses.

To learn more, go to **www.qubole.com**