Qubole®

# OPEN DATA LAKE REFERENCE ARCHITECTURES

## FOR AD-HOC ANALYTICS, STREAMING ANALYTICS, AND MACHINE LEARNING

# Winning Open Data Lake Architectures

Data-driven businesses use data in three ways:

- To understand the past
- To deal with the present
- To predict the future

These organizations win by building a modern data architecture that brings ad-hoc analytics, streaming analytics, machine learning, and continuous data engineering together to understand the present and to predict the future. They leverage an open data lake platform for these use cases and data warehouses for standard BI reporting. Data warehouses offer limited functionality for streaming analytics and machine learning as they are purpose-built and optimized primarily for SQL-based access. They are constrained by the ETL requirement to pre-process data prior to storing it compared to open data lakes where data is stored in standard file formats, is accessible via standard APIs, and processed via open-source frameworks. An open data lake platform is ideal for:

**Ad-hoc Analytics:** Open Data Lake platform provides a self-service UI to develop and deliver ad-hoc SQL analytics through ANSI/ISO-SQL (Presto, Hive, SparkSQL) and 3rd party tools such as Tableau, Looker, and Git. Further, it eliminates the underlying process and resource bottlenecks by optimizing infrastructure for these queries automatically.

**Streaming Analytics:** Open Data Lake platform supports native streaming where data streams are processed and made available for analytics as they arrive. The data pipelines in the data lake platform transform this data from the data stream and trigger computations required for analytics.

**Machine Learning (ML):** Open Data Lake platform not only enables SQL-based access to data but also provides native support for programmatic distributed data processing frameworks like Apache Spark, MLflow, and languages such as Python, Scala, R, Java and more.

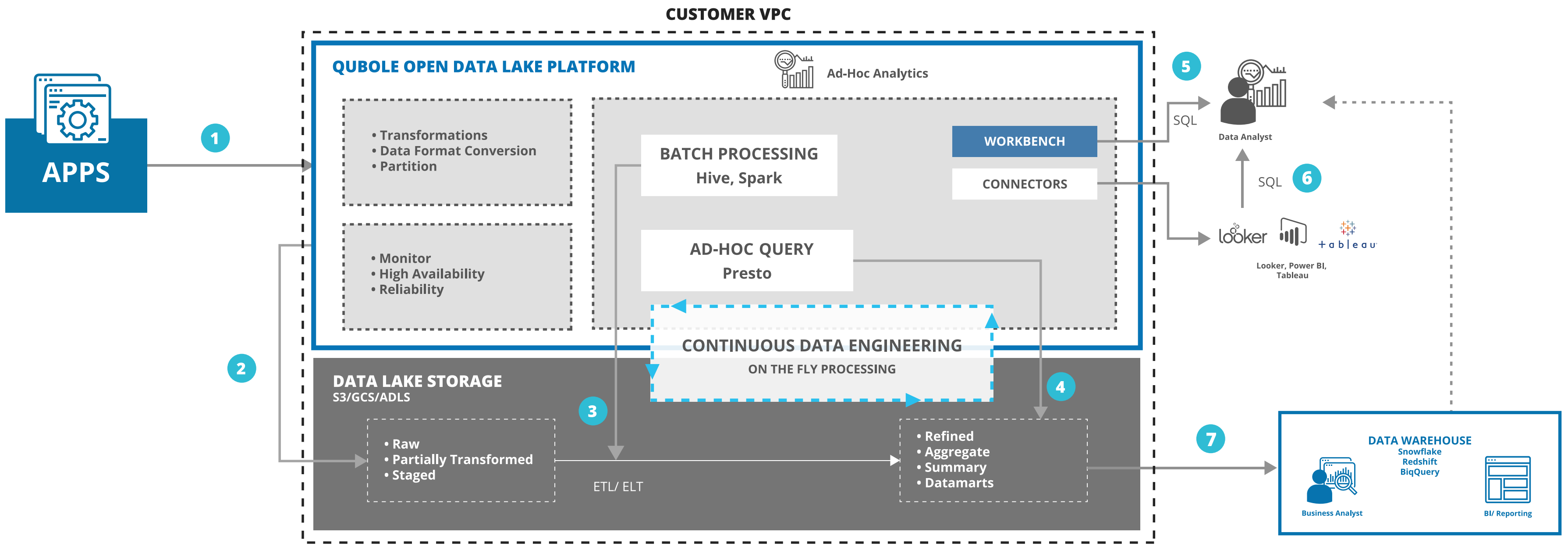
The raw data for these use cases are processed to create specific custom trusted datasets. For Ad-hoc Analytics and Machine Learning use cases, users continually refine data sets for their analysis needs. As a result, every data lake implementation must enable users to iterate between data engineering and use cases such as ad-hoc/interactive analytics and Machine Learning. This is commonly referred to as "Continuous Data Engineering".

Continuous Data Engineering involves the interactive ability to author, monitor, and debug data pipelines. The open data lake platform accelerates continuous data engineering. In an Open Data Lake, data pipelines are authored using standard interfaces and open source frameworks such as SQL, Python, Apache Spark, and/or Apache Hive. The waterfall approach of ETL is replaced by the iterative approach of continuous data engineering. The raw data that lands in a data lake can be accessed and transformed iteratively via SQL and programmatic interfaces to meet the changing needs of the use case. Continuous data engineering support is critical for ad-hoc analytics, streaming analytics, and machine learning.

This document provides you solution reference architectures for all these use cases based on Qubole's Open Data Lake Platform as part of your modern data architecture. It minimizes the effort required to operationalize your data initiative, get faster time to value from all your data, and optimize your TCO along the journey.

# Ad-Hoc Analytics

1. Applications generate all types of data for ad-hoc analytics and data exploration
2. Platform provides easy to use and reliable ingestion
3. Qubole powers fully automated batch processing with optimized infrastructure
4. Qubole enables ad-hoc search query on refined data
5. Users gets self-service access to data lake
6. Platform has native integrations with leading data visualization platforms such as Tableau
7. Users can access data via Data Warehouse for Traditional BI reporting



- Author, share, save and collaborate ad-hoc queries and reports
- Fully optimized infrastructure for ANSI SQL (Presto, Hive, SparkSQL)
- Pre-built Looker, Tableau and Git integration
- Zero downtime, migration and upgrades
- Unified metastore

# Streaming Analytics

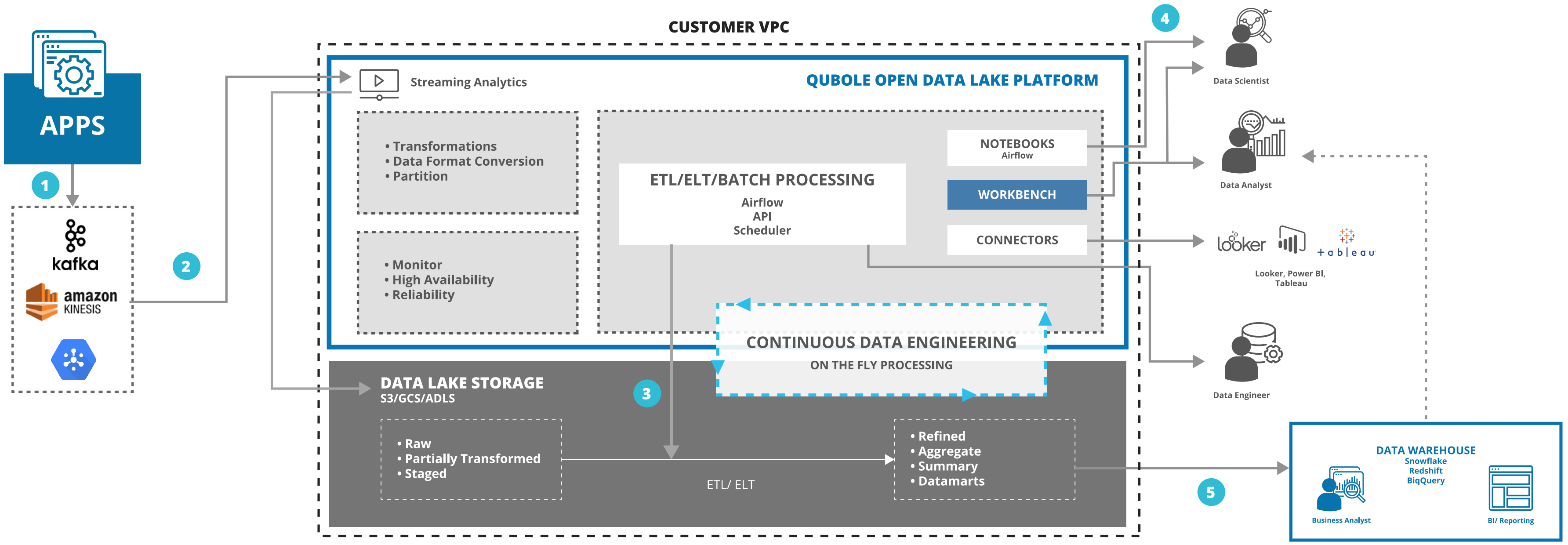
1. Application generate real-time events data for streaming analytics

2. Qubole provides easy to use, reliable and UI-driven ingestion with real-time streaming analytics

3. Users do continuous data engineering with ETL/ELT pipeline and optimize Airflow, Spark and Hive

4. Platform provides unified interface for Data Engineers, Data Scientists, Data Analysts, Workloads & App

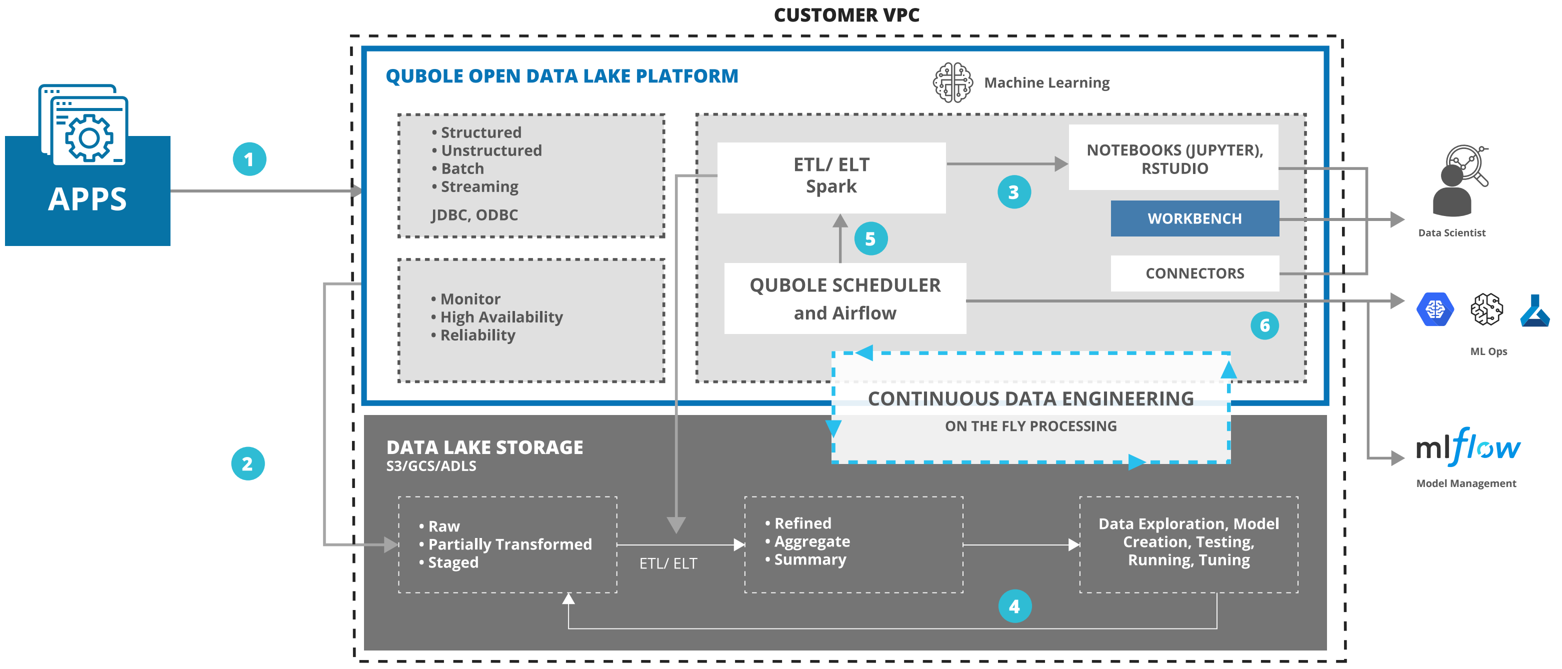
5. Users can access data via Data Warehouse for Traditional BI reporting



- Combine Streams and Batch datasets
- Collect, process stateful events

# Machine Learning

1. Applications generate all type of data for ML initiatives
2. Platform provides easy to use and reliable ingestion
3. Platform gives unified self-service access to data lake with commonly used interface like Jupyter Notebooks
4. Users can do iterative modeling for Data Exploration, Model Creation, Testing, and Tuning
5. Platform enables end-to-end feature engineering
6. Platform has native integrations with Sagemaker, ML Flow



- Programmatic Access
- Integrated Package Management
- Multi-Language Interpreter

- Offline upgrades
- Jupyter/ Qubole Notebooks to monitor job status

# Conclusion

Enterprises are increasingly adopting and sharing the know-how of their modern data architecture based on an open data lake platform. The increase in volume, velocity, and variety of data, combined with new types of analytics and machine learning has created the need for an open data lake platform. Qubole customers including market leaders like Expedia, Disney, Adobe, Lyft, Grab, Swiggy, Glassdoor and more, utilizes variations of these solution reference architectures to accelerate their data-driven business initiatives. The 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.

**Explore Qubole free for 30 days,
start your open data lake journey today.**

**START FREE TRIAL**

## About Qubole

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.

**www.qubole.com**