Cloud Data Lake Comparison Guide

AWS, Azure, Google, Cloudera, Databricks, and Snowflake

LEAD WITH DATA™ Qlik

Data lakes make their move to the cloud.

First created to overcome the limitations of the traditional data warehouse, data lakes offer the scalability, speed, and cost effectiveness to help you manage large volumes and multiple types of data across your various analytic initiatives – AI, machine learning, streaming analytics, BI, and more.

While early on-premise solutions like Hadoop paved the way for data lakes and created their initial architectural framework, cloud platform providers and other cloud-native innovators have further enhanced the possibilities. Today, with data lakes increasingly migrating to the cloud, you can take advantage of new benefits – avoiding the high upfront costs of setting up and maintaining a lake, and focusing on getting the most value out of your data.

However, with a variety of cloud data lake providers, from traditional vendors to newer data lakehouse offerings, selecting the right solution can be a challenge. This eBook provides a better understanding of the core differences between platforms, so you can make the right decision for your needs.

LIMITATIONS OF ON-PREMISE LAKES

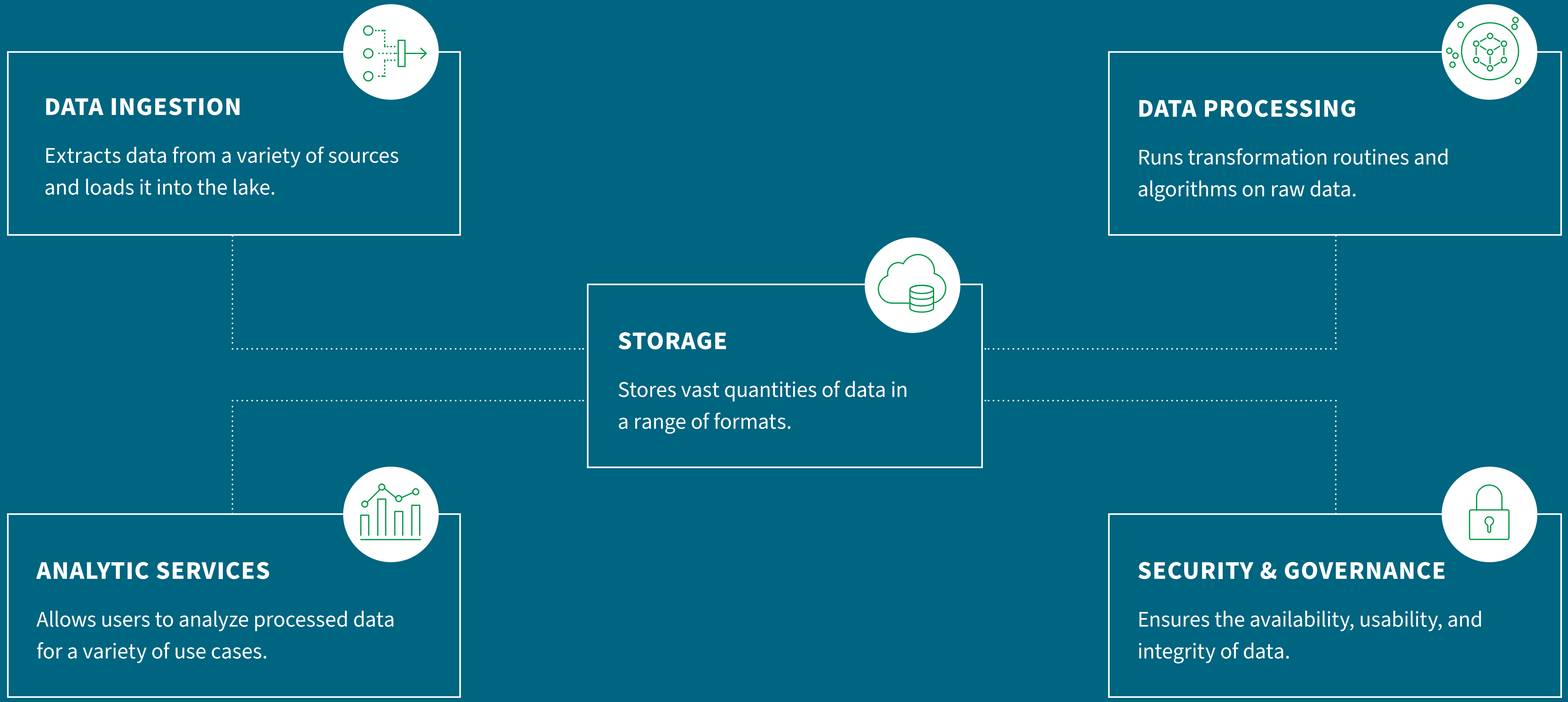
- ✗ Elasticity
- ✗ Lack of security and governance
- ✗ High maintenance costs

ADVANTAGES OF CLOUD-BASED LAKES

- ✓ Decoupled storage and compute
- ✓ Built-in security and encryption
- ✓ Transparent scaling
- ✓ Flexible on-demand infrastructure
- ✓ Consumption-based pricing

What makes a cloud data lake?

Every cloud data lake provider has unique features, functionality, and capabilities to offer your business. But, at their core, all data lakes consist of a few key core components, with attributes that vary between vendors.



DATA INGESTION

Extracts data from a variety of sources and loads it into the lake.

DATA PROCESSING

Runs transformation routines and algorithms on raw data.

STORAGE

Stores vast quantities of data in a range of formats.

ANALYTIC SERVICES

Allows users to analyze processed data for a variety of use cases.

SECURITY & GOVERNANCE

Ensures the availability, usability, and integrity of data.

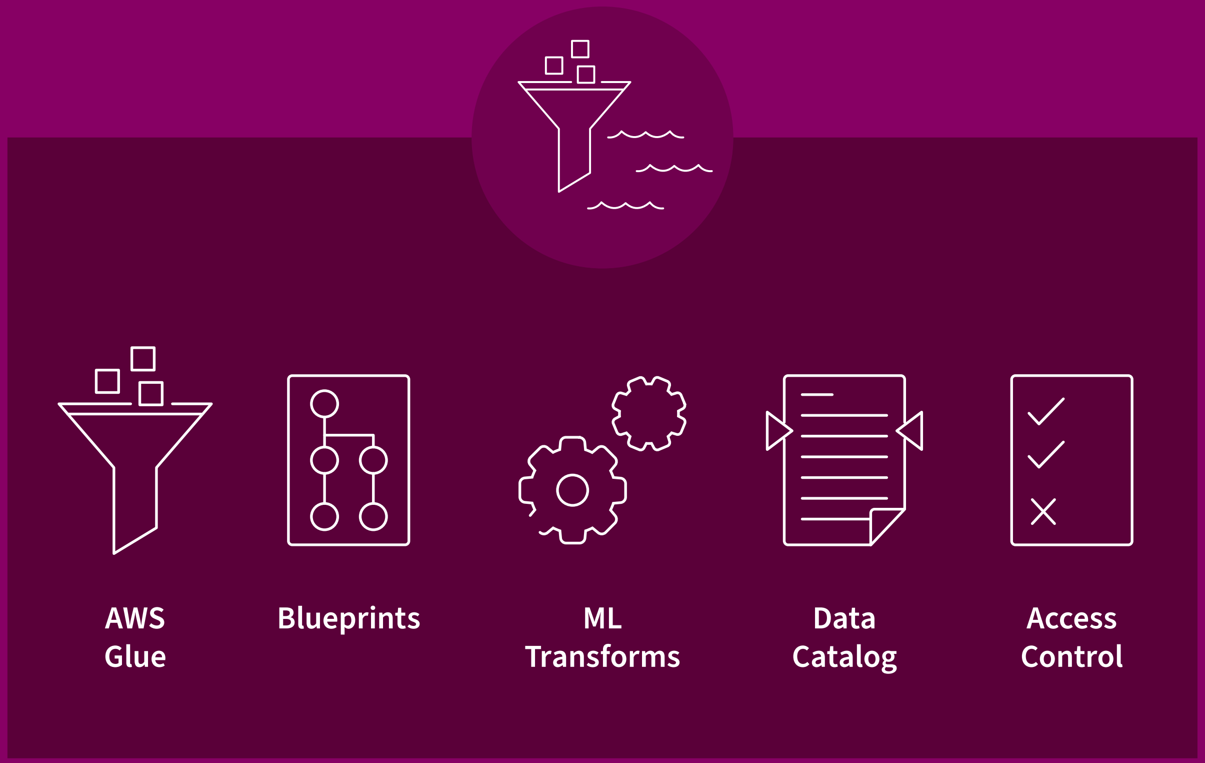
AWS Data Lake

Amazon Web Services (AWS) offers multiple services for building secure, flexible, and cost-effective data lakes. The core services that make up AWS-based lakes are:

- Amazon Simple Storage Service (S3), which provides general-purpose storage. In some instances, Amazon DynamoDB, a NoSQL database, is also used to store low-latency data such as clickstream or IoT data.
- Amazon Elastic MapReduce (EMR), the open-source-tools-based (e.g., Apache Spark, Apache Hive, Presto) processing engine, which automates batch and streaming data processing

AWS provides multiple web services (e.g., Kinesis Stream, Kinesis Firehose, Database Migration Service [DMS]) as well as partner solutions to help ingest and migrate data from cloud and on-premise sources into S3.

Additionally, AWS offers several fully managed analytic services like Elasticsearch and Athena to help analyze log data and run interactive queries.



AWS Lake Formation

To help you create data lakes more easily, Amazon offers AWS Lake Formation, a fully managed service designed to automate the setup and creation of data lakes in S3.

While it has multiple components, the heart of Lake Formation is AWS Glue, Amazon’s serverless ETL and cataloging service, which helps users search, register, and merge data.

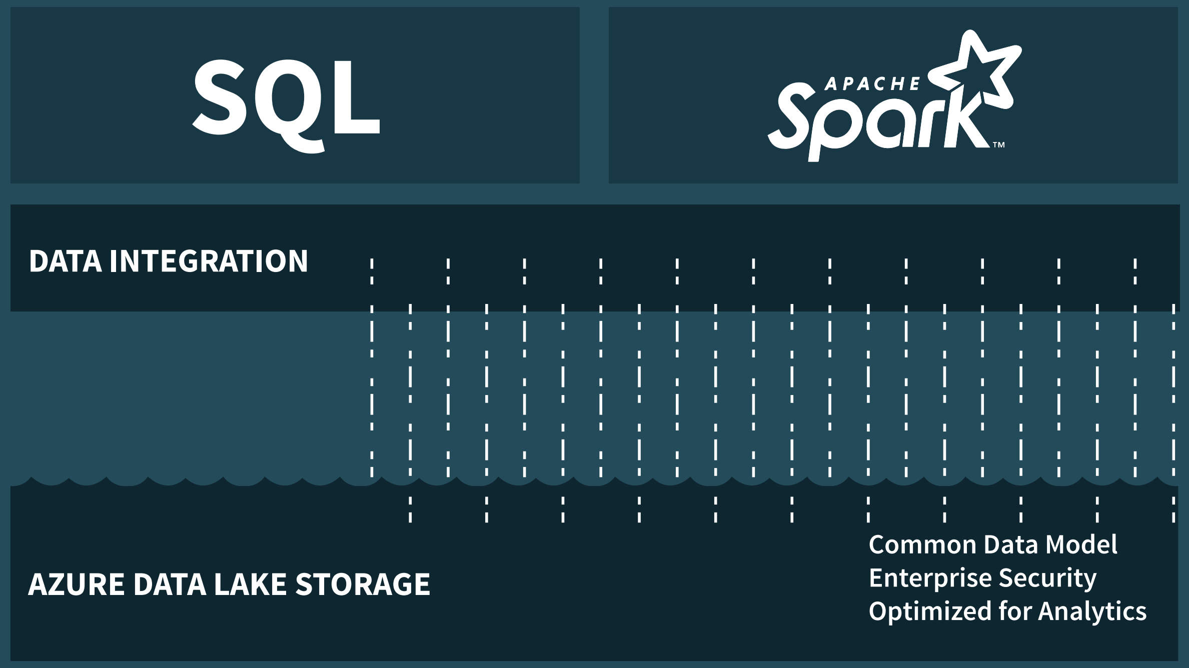
Primarily focused on data access and security, Lake Formation includes its own finer-grained authorization layer on top of the Identity and Access Management (IAM) capability of S3.

Azure Data Lake

Part of the Microsoft Azure Cloud Platform, Azure Data Lake provides scalable storage and the ability to perform all types of processing and analytics across multiple platforms and programming languages. The key components include:

- Azure Data Lake Storage (ADLS) Gen 2, which combines the file system storage of ADLS Gen 1 with Binary Large Object (BLOB) storage to provide improved scalability, analytic workload performance, and cost
- Azure HDInsight, an open-source-tools-based managed service, and Azure Synapse, which combines SQL querying with Apache-Spark-based large-scale data processing
- Azure Data Lake Analytics, an on-demand platform that lets you develop your own code and provides multi-language support, including U-SQL, R, Python, and .NET

Azure Data Lake includes disaster-recovery features and integrates with other Azure services like Azure Active Directory to provide role-based access controls and single sign-on capabilities. You can also extend your on-premise security controls to the Azure cloud environment.



Azure Synapse Analytics

Azure Synapse Analytics is Microsoft's nod to data lakehouse architecture – an increasingly popular hybrid approach that brings together data lake and data warehouse constructs.

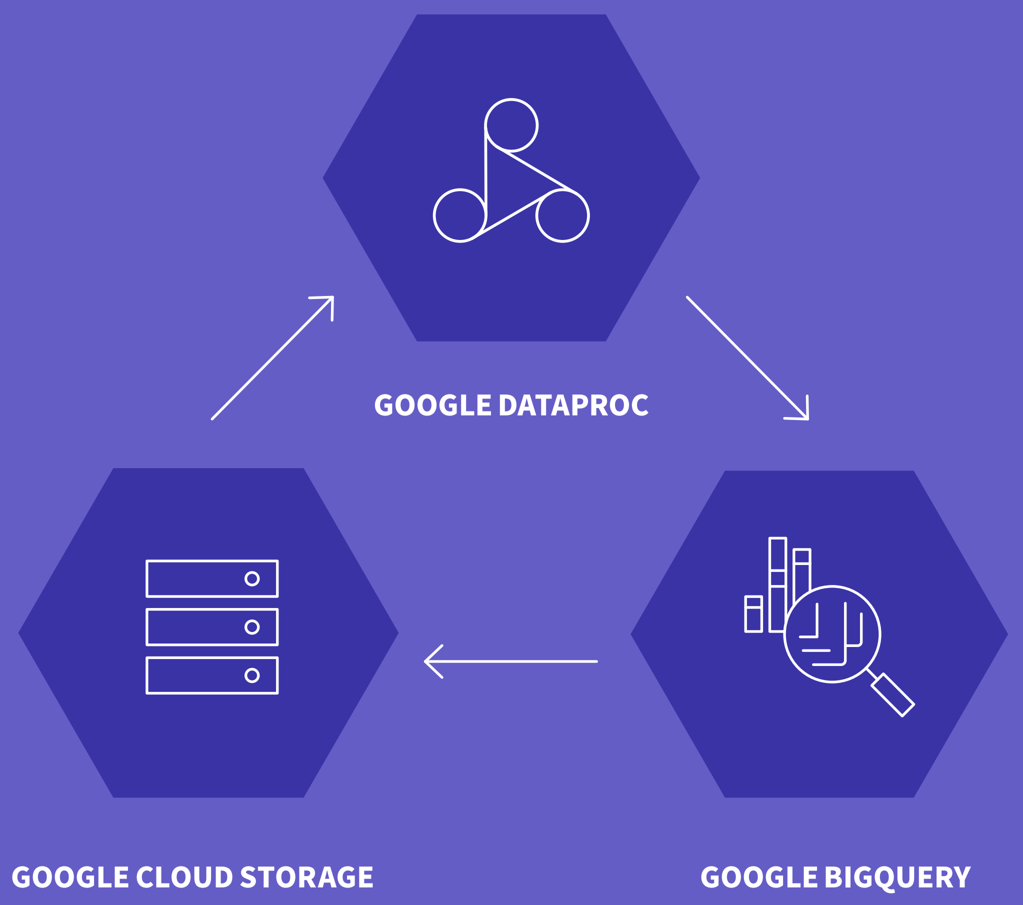
Based on ADLS Gen 2, Azure Synapse combines an SQL engine and Apache Spark in one platform to help you process and query large amounts of data.

Google Data Lake

Google Cloud Platform (GCP) offers its own data lake to help you securely ingest, store, and analyze large volumes of diverse data. Well integrated with other GCP services, Google Data Lake includes the following key elements:

- Google Cloud Storage (GCS), a general-purpose storage service that provides a low-cost option for companies of all sizes
- Google Dataproc, a fully managed open-source-tools-based service (e.g., Apache Hive, Apache Spark), which processes and analyzes cloud-scale data sets. Additionally, Google's serverless data warehouse service, Google BigQuery, allows users to run native queries on GCS data for lakehouse-like functionality.

Google provides a variety of other services that natively integrate with GCS. For the ingestion and migration of both real-time and stored data, Google offers tools like Pub/Sub, Transfer Services, and Transfer Appliance. For data processing and analysis, it includes Dataflow, for serverless processing of real-time and batch data, and Cloud Datalab, for data exploration, analysis, visualization, and machine learning.



Google BigQuery

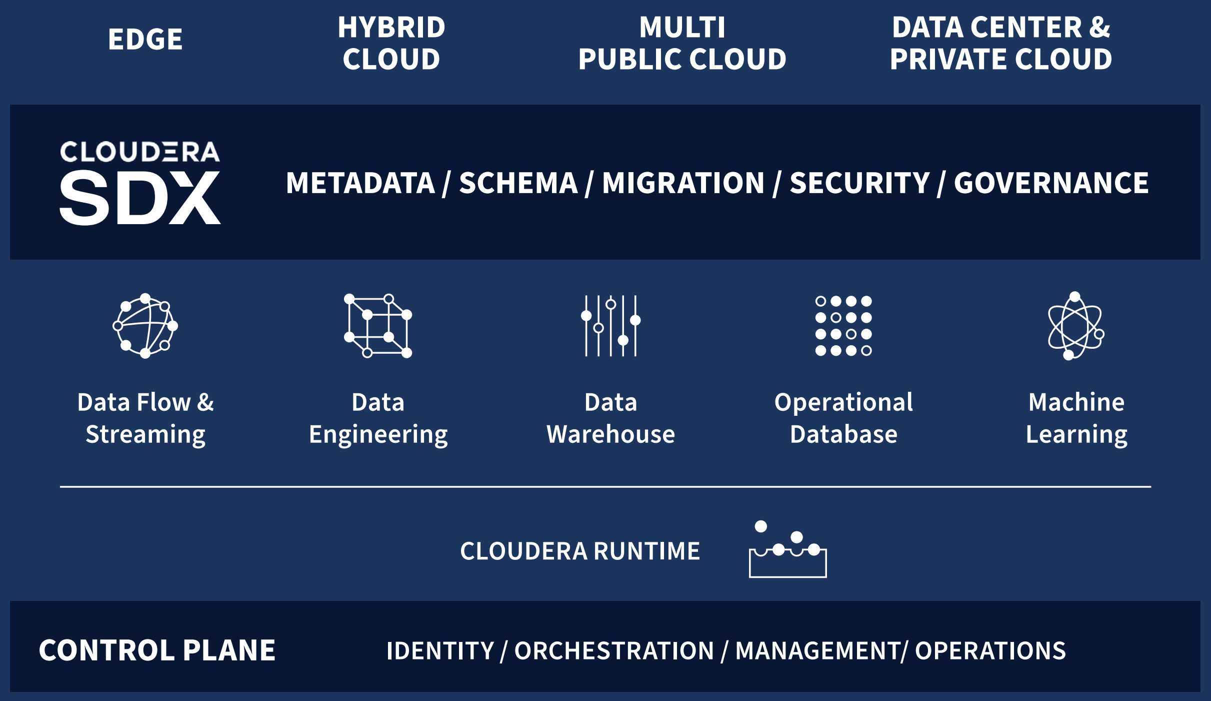
Google BigQuery not only gives SQL users high-performance native querying capabilities for data stored in GCS, it's also a perfect companion to Google Data Lake.

With cost-free movement of data between GCS and Google BigQuery, and a compatible security model, Google gives users consistent access across both services. Also, data moved from GCS to BigQuery is automatically registered with a data catalog, eliminating the need to do so yourself.

Cloudera Data Platform (CDP)

Cloudera Data Platform (CDP) is a cloud-agnostic data platform that lets you manage your infrastructure, data, and analytic workloads across every environment your business uses – public, private, hybrid, and multi-cloud. CDP brings the capabilities of Cloudera and Hortonworks together, moving Cloudera into lakehouse territory by providing both data lake and warehouse services in one platform. The core components and services of CDP are:

- Data Hub, a workload service that allows you to deploy an entire cluster in the cloud with just a few clicks, without any manual intervention
- Shared Data Experience (SDX), which helps you consolidate all your data in one place and share it securely across teams and services
- Self-service analytics services for data warehouse and machine learning use cases
- A management console that let you centrally manage, monitor, and orchestrate users and services across environments with a single interface



EDGE | HYBRID CLOUD | MULTI PUBLIC CLOUD | DATA CENTER & PRIVATE CLOUD

CLOUDERA SDX — METADATA / SCHEMA / MIGRATION / SECURITY / GOVERNANCE

Data Flow & Streaming | Data Engineering | Data Warehouse | Operational Database | Machine Learning

CLOUDERA RUNTIME

CONTROL PLANE — IDENTITY / ORCHESTRATION / MANAGEMENT/ OPERATIONS

Cloudera Data Platform

CDP includes both data lake and data warehouse services, as well as analytic tools, giving you the option to support multiple analytic workloads at the same time.

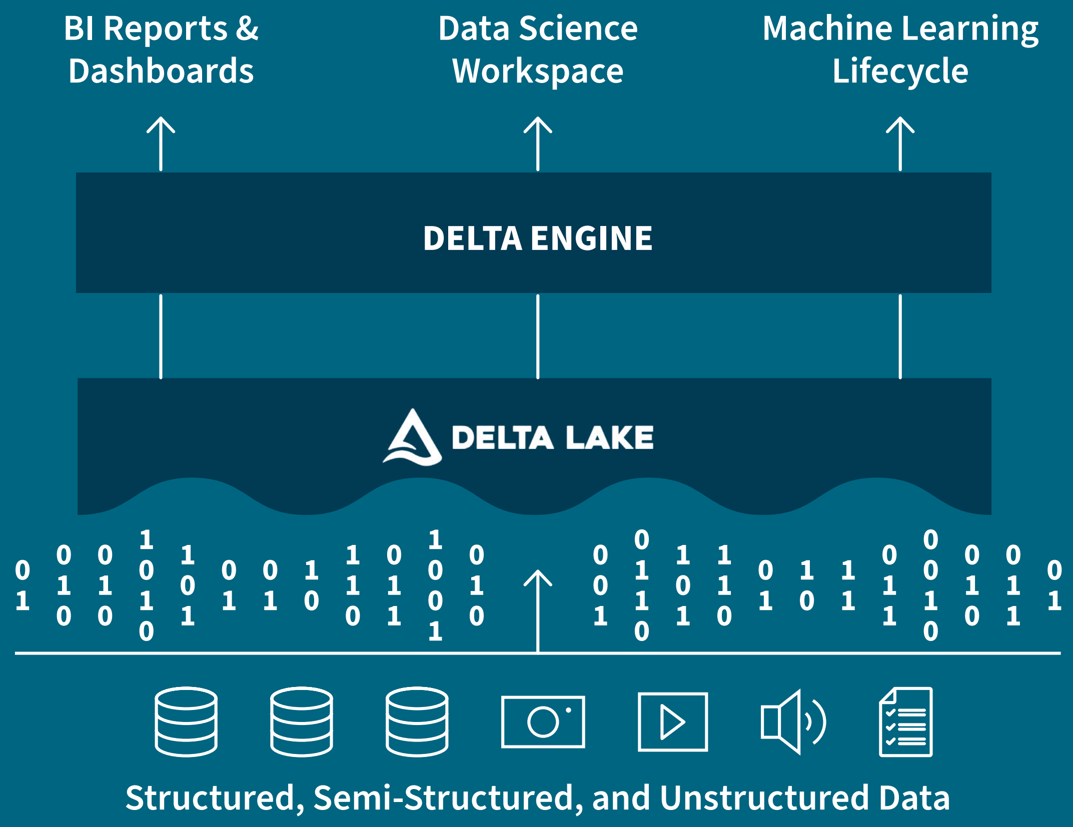
While CDP's data lake service lets you create secure and governed data lakes and then share that data across all your services and workloads, their data warehouse service is auto-isolated, only giving users access to appropriate data.

Databricks Unified Analytics Platform

Originally focused on modernizing data lakes, Databricks now positions itself as a data lakehouse – an open, unified platform designed to store and manage all your data for all the analytic needs of your business. The multi-cloud platform – available on AWS, Azure, and GCP – includes the following key elements:

- Delta Lake, an open-source storage layer that sits on top of your existing data lake on your preferred cloud platform – eliminating the need to change your current architecture. Focused on data reliability, Delta Lake includes ACID transactions, schema enforcement, auto-compacting, and auto-optimization features for improving data lake reliability and performance.
- Delta Engine, an Apache-Spark-compatible query engine that processes the data in Delta Lake. Unlike Spark, Delta Engine is optimized for lakehouse data and supports a variety of workloads, from large-scale ETL processing to ad hoc interactive queries.

Additionally, Databricks provides native support for a variety of common programming languages, including R and Python, as well as a collaborative data science and machine learning platform.


BI Reports & Dashboards
Data Science Workspace
Machine Learning Lifecycle
DELTA ENGINE
DELTA LAKE
Structured, Semi-Structured, and Unstructured Data

Databricks SQL Analytics

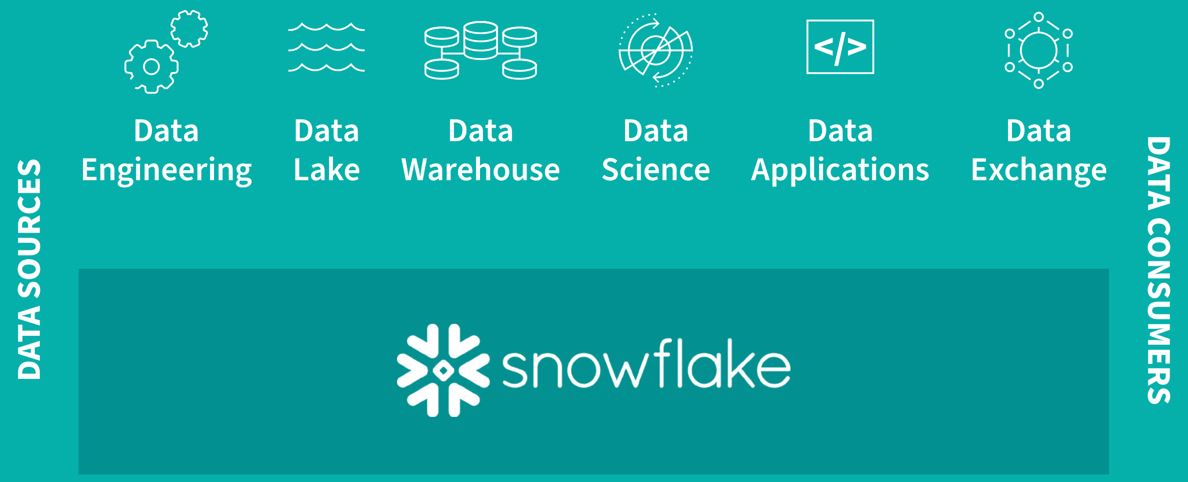
Databricks' SQL Analytics Service is the company's latest step in establishing itself as a lakehouse – a single unified platform for all analytic initiatives. Designed to support your unique BI and reporting needs, the service gives SQL users a familiar interface to easily query data and build dashboards.

Snowflake Cloud Data Platform

Better known as a cloud data warehouse, Snowflake has been increasingly blurring the lines between data lake and data warehouse. Built on a flexible platform, Snowflake provides the scalability, elasticity, and low-cost storage of a lake, along with the security, governance, and performance of a warehouse.

Available in AWS, Azure, and GCP, Snowflake allows you to load a diverse array of data in its native format, without having to transform it, giving you the flexibility and agility of a data lake. Users can also leverage Snowflake's MPP architecture to spin up multiple virtual warehouses and run multiple queries at the same time.

Snowflake also enables you to share data with partner tools like Apache Spark, using ODBC and JDBC connectors for real-time large-scale data processing.



Snowpark

Snowpark, Snowflake's recently launched developer tool, further supports their lakehouse approach by allowing data scientists, engineers, and programmers to develop and deploy custom code on Snowflake using a variety of programming languages including Java, Scala, and Python.

Cloud Data Lakes at a Glance

CRITERIA	aws	Microsoft Azure	Google Cloud Platform	CLOUDERA	databricks	snowflake
Primary Storage Service	Amazon S3	ADLS Gen 2	Google Cloud Storage	Cloudera Data Platform	Data Lake atop AWS, GCS, or ADLS	Snowflake Cloud Data Platform
Processing Engine	Amazon EMR	Azure HDInsight, Azure Synapse	Google Dataproc, Data Flow	CDP Data Engineering	Delta Engine	Snowflake
Hive and Spark Support	Yes	Yes	Yes	Yes	Spark	N/A
Decoupled Storage & Compute	Yes	Yes	Yes	Yes	Yes	Yes
Pipeline Service	AWS Glue	Azure Data Factory	Cloud Data Fusion, Data Flow	Cloudera Data Engineering	Delta Engine	Snowpark
SQL Support	Amazon Athena, Redshift, Spectrum	Azure Synapse	Google BigQuery	Self-Service Analytic Services for Data Warehouse	SQL Analytics Service	Snowflake
Support for Multiple Programming Languages	Yes	Yes	Yes	Yes	Yes	Yes
Catalog	AWS Glue	Azure Data Catalog	Google Data Catalog	Cloudera Data Platform	Partner Solutions	Partner Solutions
Lakehouse Architecture	Yes	Yes	Yes	Yes	Yes	Yes
Multi-Cloud	No	No	No	Yes	Yes	Yes

Great lakes start with great data integration.

As you begin to move your data to the cloud, more and more vendors are ready to meet you there, with different solutions built to fit your unique needs. But, no matter which platform you choose, the one must-have is robust data integration, to get data where it needs to go.

Whether it's with a lake or a lakehouse, data integration is needed to not only ingest and migrate a variety of data from a plethora of sources, but also to process and refine that data so that it's readily available for all your analytic use cases.

The Qlik® Data Integration Platform can automate the data lake pipeline to accelerate and streamline the availability of analytics-ready data, allowing your data engineering teams to deliver continuously updated data to the data scientists, business analysts, and other data users across your organization.

Qlik for cloud data lakes.

Qlik's Data Integration Platform can help you get more out of your cloud data lake investment, sooner, by continuously delivering the accurate, timely, and trusted data you need. The platform provides the unparalleled ability to automate data streams from any source – including legacy mainframes, enterprise applications like SAP, databases, data warehouses, and more – into your lake. Qlik also delivers analytics-ready datasets, without any coding.

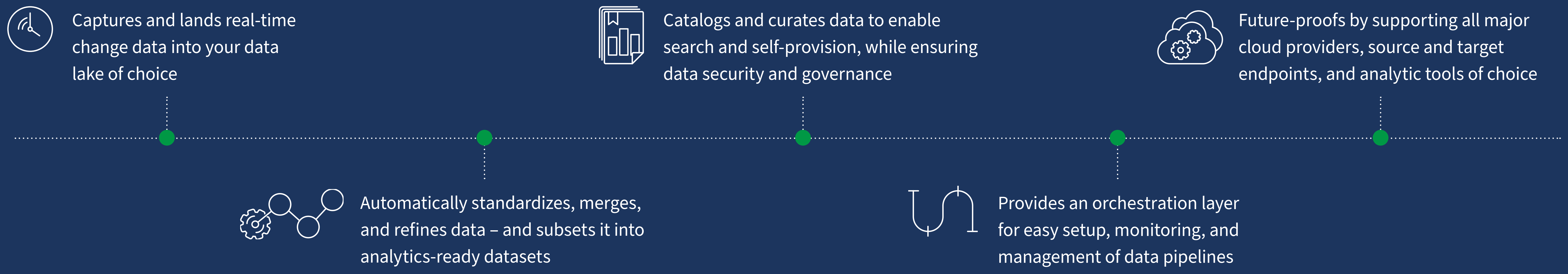
Unlock the full value of all your data by accelerating the delivery of analytics-ready data and maximizing the efficiency, agility, and flexibility of your analytic programs.

Want to see how Qlik can make an impact on data delivery in your business?

Start with a free trial.

Test Drive

Qlik automates the complete data lake pipeline:



Captures and lands real-time change data into your data lake of choice

Automatically standardizes, merges, and refines data – and subsets it into analytics-ready datasets

Catalogs and curates data to enable search and self-provision, while ensuring data security and governance

Provides an orchestration layer for easy setup, monitoring, and management of data pipelines

Future-proofs by supporting all major cloud providers, source and target endpoints, and analytic tools of choice

ABOUT QLIK

Qlik's vision is a data-literate world, where everyone can use data and analytics to improve decision-making and solve their most challenging problems. Qlik provides an end-to-end, real-time data integration and analytics cloud platform to close the gaps between data, insights, and action. By transforming data into Active Intelligence, businesses can drive better decisions, improve revenue and profitability, and optimize customer relationships. Qlik does business in more than 100 countries and serves over 50,000 customers around the world.

Qlik® LEAD WITH DATA™

qlik.com