COMPARISON GUIDE

Top Cloud Data Lakes for the Enterprise

How do the big six stack up?
For managing multiple types of data in large volumes, data lakes are fast, scalable, and cost-effective. And while on-prem solutions like Hadoop paved the way, today there’s an option with more agility and efficiency: cloud data lakes. In the cloud, you can avoid the upfront costs of setting up and maintaining a lake and focus on getting the most value out of your data.

Among the variety of cloud data lake providers in the market, which solution is right for you? In this eBook, we explain the core differences among the top six platforms.

**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

---

**The Cloud Hyperscalers**
- AWS
- Google Cloud
- Microsoft Azure

**The Multicloud Solutions**
- Cloudera
- Databricks
- Snowflake
What makes a cloud data lake?

Every cloud data lake provider offers unique features. But at their core, all data lakes consist of a few key components.

- **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.
- **ANALYTICS SERVICES**: Allows users to analyze processed data for a variety of use cases.
- **SECURITY & GOVERNANCE**: Ensures the availability, usability, and integrity of data.
The Cloud Hyperscalers

- AWS
- Google Cloud
- Microsoft Azure
Amazon Web Services (AWS) Data Lake

AWS offers multiple services for building secure, flexible, and cost-effective data lakes. Two core services make up AWS-based lakes:

- **S3**: 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.

- **EMR**: Amazon Elastic MapReduce (EMR) – the processing engine based on open-source tools like Apache Spark, Apache Hive, and Presto – 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 analytics services like Elasticsearch and Athena to help analyze log data and run interactive queries.

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 fine-grained authorization layer on top of the identity and access management (IAM) capability of S3.
Google Cloud Platform (GCP) offers a 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 service – based on open-source tools like Apache Hive and Apache Spark – that processes and analyzes cloud-scale data sets.

Google’s serverless data warehouse service, Google BigQuery, allows users to run native queries on GCS data for lakehouse-like functionality.

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 it yourself.

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 Service, 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.
Microsoft Azure Data Lake

Part of the Microsoft Azure cloud platform, Azure Data Lake provides scalable storage as well as processing and analytics across multiple platforms and programming languages. The key components include:

1. Azure Data Lake Storage (ADLS) Gen2, which combines the file system storage of ADLS Gen1 with binary large object (Blob) storage to provide improved scalability, analytic workload performance, and cost.

2. Azure HDInsight, a managed service based on open-source tools, and Azure Synapse, which combines SQL querying with Apache Spark-based large-scale data processing.

3. Azure Data Lake Analytics, an on-demand platform that lets you develop your own code and provides multi-language support, including for 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.
The Multicloud Solutions
Cloudera Data Platform (CDP)

Cloudera Data Platform 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 multicloud. 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:

1. **Data Hub**, a workload service that allows you to deploy an entire cluster on the cloud with just a few clicks – no manual intervention required.
2. **Shared Data Experience (SDX)**, which helps you consolidate all your data in one place and share it securely across teams and services.
3. **Self-service analytics for data warehouses and machine learning**.
4. **A management console** for centrally managing, monitoring, and orchestrating users and services across environments with a single interface.

**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, governed lakes and share the data across all services and workloads, their data warehouse service is auto-isolated, giving users access only to appropriate data.
Databricks Unified Data 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 analytics. The multicloud platform – available on AWS, Azure, and GCP – is built with the components shown on the right.

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.

**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 analytics initiatives. Designed to support your unique BI and reporting needs, the service gives SQL users a familiar interface for easily querying data and building dashboards.
Snowflake Cloud Data Platform

Known primarily as a cloud data warehouse, Snowflake has increasingly edged into data lake territory. 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 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.

Snowflake’s developer tool further supports their lakehouse approach by allowing data scientists, engineers, and programmers to develop and deploy custom code using a variety of programming languages – including Java, Scala, and Python.
# Cloud Data Lakes at a Glance

<table>
<thead>
<tr>
<th>Criteria</th>
<th>THE CLOUD HYPERSCALERS</th>
<th>THE MULTICLOUD SOLUTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AWS</td>
<td>Google Cloud</td>
</tr>
<tr>
<td>Primary Storage Service</td>
<td>Amazon S3</td>
<td>Google Cloud Storage</td>
</tr>
<tr>
<td>Processing Engine</td>
<td>Amazon EMR</td>
<td>Google Dataproc, Dataflow</td>
</tr>
<tr>
<td>SQL Support</td>
<td>Amazon Athena, Redshift, Spectrum</td>
<td>Google BigQuery</td>
</tr>
<tr>
<td>Catalog</td>
<td>AWS Glue</td>
<td>Google Data Catalog</td>
</tr>
<tr>
<td>Pipeline Service</td>
<td>AWS Glue</td>
<td>Cloud Data Fusion, Dataflow</td>
</tr>
<tr>
<td>Apache Hive and Apache Spark Support</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Support for Multiple Programming Languages</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Decoupled Storage &amp; Compute</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Lakehouse Architecture</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Multicloud</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

Top Cloud Data Lakes for the Enterprise | 11
Great data lakes start with great integration.

Rapidly design, deploy, and manage data lakes with no coding.

As you begin to move your data to the cloud, more and more vendors are ready to meet you there, with a variety of solutions built for differing needs. But no matter which platform you choose, there’s one thing you can’t go without: robust data integration.

Whether you’re using a data lake or lakehouse, you need integration – not only to ingest data from a wide range of sources but also to process that data so that it’s readily available for any and every type of analytics.

That’s where we come in. Qlik Data Integration® automates the continuous delivery of updated, accurate data sets for analytics. It empowers data engineers to ensure success at every step of the pipeline, from real-time ingestion to refinement, provisioning, and governance. And in the process, Qlik® makes it possible for all your users to get reliable, analytics-ready data when they need it most.

Need APIs? Meet the Qlik Connector Factory.

With an in-house R&D team dedicated to developing standard APIs, we’re continually expanding access to and delivery of data from hundreds of SaaS applications and data sources. Our customers already benefit from over 250 existing connectors, and throughout 2023, we’ll be adding 100 more.

Learn More
Qlik for cloud data lakes.

Qlik Data Integration 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. And all of that without coding, too. Learn more about data lake creation today.

Curious to see how Qlik can revolutionize data delivery in your business? Try it free.

Qlik automates the complete data lake pipeline:

- Captures and lands real-time change data into your data lake of choice
- Catalogs and curates data to enable search and self-provision, while ensuring data security and governance
- Future-proofs by supporting all major cloud providers, source and target endpoints, and analytic tools of choice
- Automatically standardizes, merges, and refines data – and subsets it into analytics-ready data sets
- Provides an orchestration layer for easy setup, monitoring, and management of data pipelines
About Qlik

Qlik is the global leader in data integration, data quality, and analytics solutions. Its comprehensive cloud platform unifies data across cloud and hybrid environments, automates information pipelines and data-driven workflows, and augments insights with AI. Qlik enables users to make data more available and actionable for better, faster business outcomes. With more than 40,000 active customers in over 100 countries, Qlik is committed to providing powerful data solutions to meet the evolving needs of organizations worldwide.