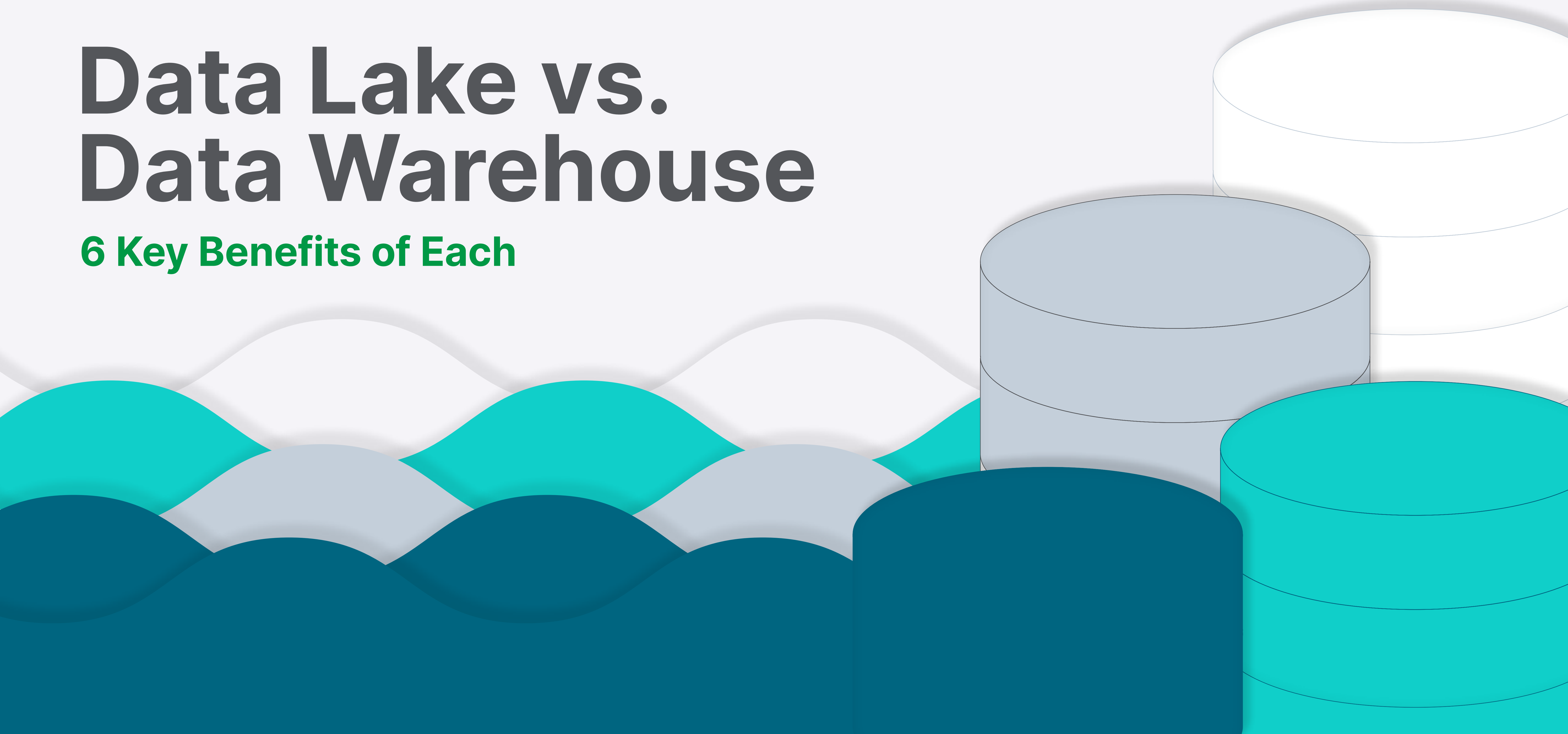
COMPARISON GUIDE

Data Lake vs. Data Warehouse

6 Key Benefits of Each

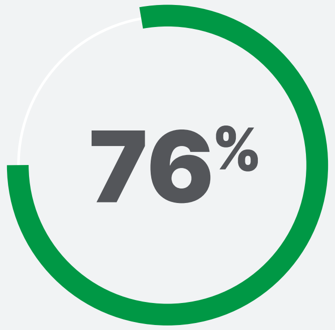


What’s the difference between a data lake and a data warehouse?

On the surface, they may seem similar. Both are repositories that store data for use in data science and analytics. Both house data that can inform all your business analysis and reporting. And in recent years, both lakes and warehouses have moved to the cloud to accommodate the massive influx of data, the increasing demand for speed, and the need to extend insights to as many users as possible.

But data lakes and data warehouses are not the same, and there are important differences to be aware of as you evaluate your organization’s needs.

This eBook looks at the differences between data lakes and data warehouses, the six key benefits of each, the top vendors offering them, and options for optimizing your experience with both.



In a 2021 survey, over 76% of IT executives stated that their organization was investing more in analytics infrastructure such as data platforms, data warehouses, and more.[1]

What’s a data lake?

A data lake is a massive pool that can store huge volumes of data in its raw state. Instead of predefining the schema and data requirements, you use tools to assign unique identifiers and tags to data elements, so only a subset of relevant data is queried to analyze a given business question. This analysis can include real-time analytics, big data analytics, machine learning, dashboards, and data visualizations.

First created to overcome the limitations of traditional data warehouses, data lakes offer the scalability, speed, and cost-effectiveness to help you manage large volumes and multiple types of data for all your analytic initiatives.

Data lakes in the cloud.

While early on-prem solutions like Hadoop paved the way for data lakes, cloud innovators have 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.

Six key benefits of a data

Because the large volumes of data in a lake aren't structured, both skilled data scientists and end-to-end self-service BI tools can provide access far faster than in a data warehouse. Six key advantages include:

1 **AGILITY**

You can easily configure queries, data models, and applications without planning ahead. In addition to SQL queries, the data lake strategy is well suited to support real-time analytics, big data analytics, and machine learning.

2 **REALTIME**

You can import data in its original format from multiple sources in real time. This allows you to perform real-time analytics and machine learning as well as triggering actions in other applications.

3 **SCALE**

Data lakes can handle massive volumes of both structured and unstructured data, including ERP transactions and call logs.

4 **SPEED**

You don't have to perform time-intensive tasks like transformation and schema development until you define the business question(s) that need to be addressed.

5 **BETTER INSIGHTS**

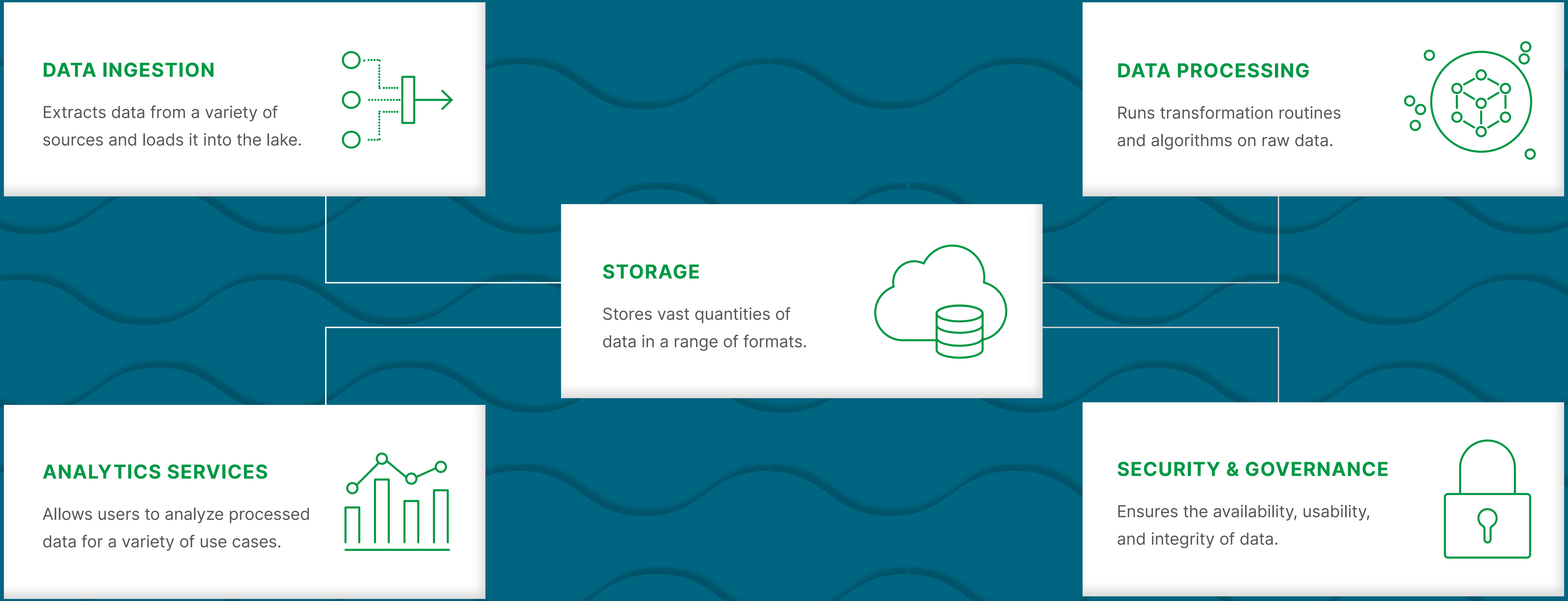
You can gain unexpected and previously unavailable insights by analyzing a broader range of data in new ways.

6 **COST SAVINGS**

Easier management makes for lower operational costs. And storage costs are lower than in warehouses, too, because most of the management tools are open-source and run on low-cost hardware.

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.

Top cloud data lakes at a glance.

Features	aws	Google Cloud	Microsoft Azure	CLOUDERA	databricks	snowflake
Primary Storage Service	Amazon S3	Google Cloud Storage	ADLS Gen2	Cloudera Data Platform	Data Lake atop AWS, GCS, or ADLS	Snowflake Cloud Data Platform
Processing Engine	Amazon EMR	Dataproc, Dataflow	Azure HDInsight, Azure Synapse	CDP Data Engineering	Delta Engine	Snowflake
SQL Support	Amazon Athena, Redshift, Spectrum	Google BigQuery	Azure Synapse	Self-Service Analytics Services for Data Warehouse	SQL Analytics Service	Snowflake
Catalog	AWS Glue	Data Catalog	Azure Data Catalog	Cloudera Data Platform	Unity Catalog	Partner Solutions
Pipeline Service	AWS Glue	Cloud Data Fusion, Dataflow	Azure Data Factory	Cloudera Data Engineering	Delta Engine	Snowpark
Apache Hive and Apache Spark Support	✓	✓	✓	✓	Apache Spark	N/A
Support for Multiple Programming Languages	✓	✓	✓	✓	✓	✓
Decoupled Storage & Compute	✓	✓	✓	✓	✓	✓
Lakehouse Architecture	✓	✓	✓	✓	✓	✓
Multicloud	⊗	⊗	⊗	✓	✓	✓

What's a data warehouse?

Like a data lake, a data warehouse aggregates large volumes of data from multiple sources into a single repository. Unlike a data lake, a warehouse stores only highly structured and unified data to support specific business intelligence and analytics needs.

As in an actual warehouse, contents are first processed and then organized into sections and placed onto "shelves" (called data marts) to be accessed by consumers. Data from a warehouse is ready to support historical analysis, reporting, artificial intelligence, and machine learning to inform decision-making across an organization's lines of business.

By 2025, the global market for data warehousing is projected to be $30 billion USD.[2]

Warehousing in the cloud.

Data warehouses have been staples of enterprise analytics and reporting for decades. But they weren't designed to handle today's explosive data growth or to keep pace with end users' ever-evolving needs.

All that changed with the arrival of cloud data warehousing. No longer constrained by physical data centers, companies can now grow or shrink their warehouses to rapidly meet dynamic requirements.

Six key benefits of a data warehouse.

A data warehouse offers enormous benefits to organizations, especially for BI and analytics.

The major advantages of a data warehouse include:

1 **BETTER DATA QUALITY**

Data from a warehouse has been cleansed, de-duplicated, and standardized.

2 **MORE TRUST**

Having a consistent, single source of truth builds confidence in the insights and decisions derived from any analysis.

3 **USER-FRIENDLINESS**

Because the data is already structured, there's little or no data prep needed on the business side, making it far easier for analysts and other consumers to access it and put it to use.

4 **DATA HARMONY → FASTER ANALYSIS**

A warehouse unifies and harmonizes data from a wide range of sources, including operational databases, transactional systems, and flat files. This allows you to more quickly leverage BI capabilities like reporting, dashboarding, and ad-hoc analysis.

5 **SPEED**

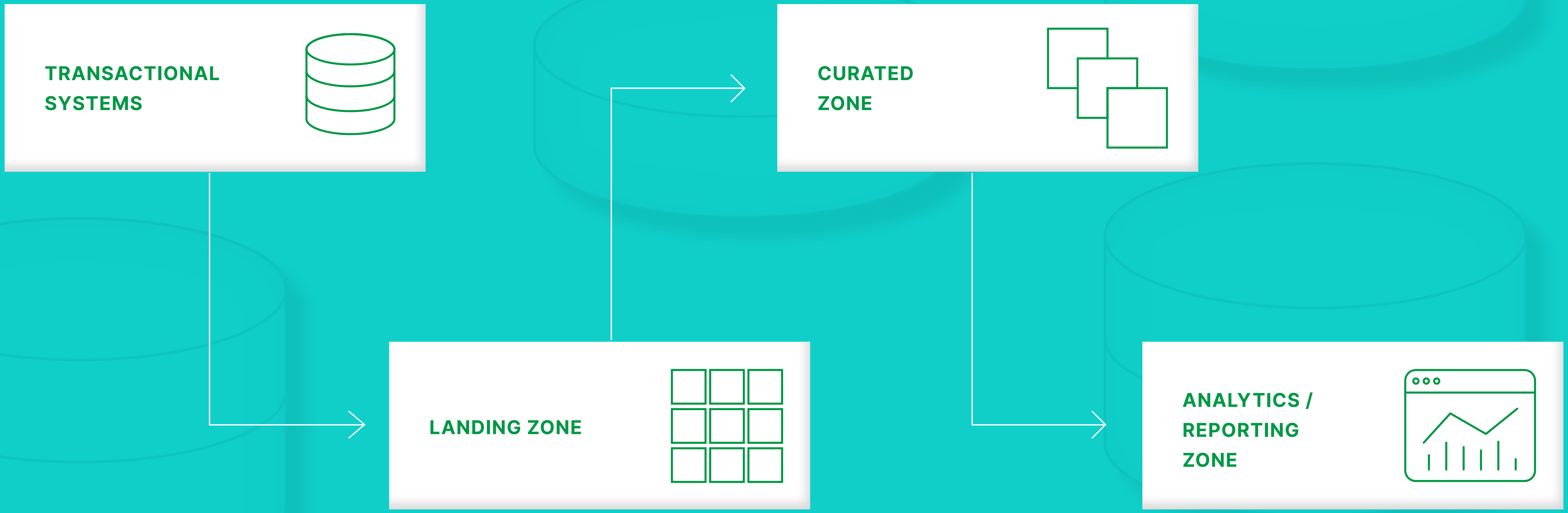
Accurate, complete data is available more quickly, so you can turn information into insights faster.

6 **SECURITY**

Data warehouses offer advanced security features like encryption, role-based access control, and auditing.

What makes a cloud data warehouse?

Your data warehouse architecture will be determined by your organization's needs. Here's a high-level diagram of the typical structure.



TRANSACTIONAL SYSTEMS

LANDING ZONE

CURATED ZONE

ANALYTICS / REPORTING ZONE

Top cloud data warehouses at a glance.

Features	amazon REDSHIFT	Azure Synapse Analytics	Google BigQuery	snowflake®
Transaction	ACID	ACID	ACID	ACID
Elasticity	Manual	Manual and Automatic	Automatic	Automatic
Query Language	Amazon Redshift SQL	TSQL	Standard SQL 2011 & BigQuery SQL	Snowflake SQL
Initial Release	2012	2016	2010	2014
Concurrency	✓	✓	✓	✓
Durability	✓	✓	✓	✓
MPP	✓	✓	✓	✓
Columnar	✓	✓	✓	✓
Foreign Keys	✓	✓	✗	✓
Separates Storage and Compute	✗	✓	✓	✓
Automation	✗	✗	✗	✗
Multicloud	✗	✗	✗	✓

You can use both.

Most organizations use both a data lake and a data warehouse to cover the spectrum of their data storage needs. Here's a side-by-side look at the two options so you can see how they can work in tandem.

	Data Storage	Users	Analysis	Schema	Processing	Cost
Data Lake	The data lake contains all of an organization's data in a raw, unstructured form and can store the data indefinitely – for immediate or future use.	Data from a data lake – with its large volume of unstructured data – is typically used by data scientists and engineers who prefer to study data in its raw form to gain new, unique business insights.	Predictive analytics, machine learning, data visualization, BI, and big data analytics.	The schema is defined after the data is stored in a data lake, making the process of capturing and storing the data faster.	ELT (extract, load, transform). In this process, the data is extracted from its source for storage in the data lake and structured only when needed.	Storage costs are fairly inexpensive in a data lake versus a data warehouse. Data lakes are also less time-consuming to manage, which reduces operational costs.
Data Warehouse	A data warehouse contains structured data that has been cleaned and processed, ready for strategic analysis based on predefined business needs.	Data from a data warehouse is typically accessed by managers and business-end users looking to gain insights from business KPIs, as the data has already been structured to provide answers to predetermined questions for analysis.	Data visualization, BI, and data analytics.	The schema is defined before the data is stored. This lengthens the time it takes to process the data, but once complete, the data is ready for consistent, confident use across the organization.	ETL (extract, transform, load). In this process, data is extracted from its source(s), scrubbed, and then structured so it's ready for analysis.	Data warehouses cost more than data lakes and require more time to manage, resulting in additional operational costs.

Building real-time data pipelines for data-informed action.

However you store your data, there's one must-have for successful data delivery: data integration. You need data integration not only to ingest a variety of data from a multitude of sources, but also to process and refine that data so it's readily available for analytics.

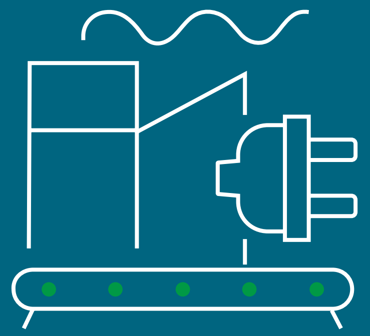
Qlik Data Integration® automates and accelerates your data pipeline, allowing you to continuously deliver up-to-the-minute data.

Qlik for Cloud Data Lakes

Get to ROI sooner by automating data streams from any source – legacy mainframes, enterprise applications (including SAP), databases, data warehouses, and more – into your lake. Qlik® also delivers analytics-ready data sets without any coding.

Qlik for Cloud Data Warehouses

Automate your entire data warehouse lifecycle to accelerate the availability of analytics-ready data. Give your data engineers the agility to create a data model, add new sources, and provision new data marts. And ensure success at every step of the pipeline, from data modeling and real-time ingestion to data marts and governance.



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.

See the Factory

Accelerate, automate, and govern data delivery with Qlik Data Integration.

To succeed in the real-time economy, you need to deliver the latest and most accurate data to as many users as quickly as possible. Qlik Data Integration enables a DataOps approach to accelerate the discovery and availability of real-time, analytics-ready data by automating every stage of the data-delivery pipeline.

1

Real-time data streaming (CDC)

Extend enterprise data into live streams to enable modern analytics and microservices with a simple, real-time, and universal solution.

2

Managed data lake creation

Automate complex ingestion and transformation processes to provide continuously updated and analytics-ready data lakes.

3

Agile data warehouse automation

Quickly design, build, deploy, and manage purpose-built cloud data warehouses without manual coding.

4

Enterprise data catalog

Enable analytics across your enterprise with a single, self-service data catalog.

Which data lake or data warehouse is right for your business? Get an overview of the top solutions in these guides:

Qlik

COMPARISON GUIDE

Top Cloud Data Lakes for the Enterprise

How do the big six stack up?

COMPARISON GUIDE: Top Cloud Data Lakes for the Enterprise

Qlik

COMPARISON GUIDE

Top Cloud Data Warehouses for the Enterprise

How do the big four stack up?

COMPARISON GUIDE: Top Cloud Data Warehouses for the Enterprise

About Qlik

Qlik transforms complex data landscapes into actionable insights, driving strategic business outcomes. Serving over 40,000 global customers, our portfolio leverages advanced, enterprise-grade AI/ML and pervasive data quality. We excel in data integration and governance, offering comprehensive solutions that work with diverse data sources. Intuitive and real-time analytics from Qlik uncover hidden patterns, empowering teams to address complex challenges and seize new opportunities. Our AI/ML tools, both practical and scalable, lead to better decisions, faster. As strategic partners, our platform-agnostic technology and expertise make our customers more competitive.

qlik.com