

azure.microsoft.com

What is a Data Lake? Data Lake vs. Warehouse | Microsoft Azure

55-70 minutes

Learn about the difference between data lakes and data warehouses. Discover how to build a scalable foundation for all your analytics with Azure.

Data lake definition

This introductory guide explores the many benefits and use cases of a data lake. Learn what a data lake is, why it matters, and discover the difference between data lakes and data warehouses. But first, let's define data lake as a term.

A data lake is a centralized repository that ingests and stores large volumes of data in its original form. The data can then be processed and used as a basis for a variety of analytic needs. Due to its open, scalable architecture, a data lake can accommodate all types of data from any source, from structured (database tables, Excel sheets) to semi-structured (XML files, webpages) to unstructured (images, audio files, tweets), all without sacrificing fidelity. The data files are typically stored in staged zones—raw, cleansed, and curated—so that different types of users may use the data in its various

forms to meet their needs. Data lakes provide core data consistency across a variety of applications, powering [big data analytics](#), [machine learning](#), predictive analytics, and other forms of intelligent action.

Why are data lakes important for businesses?

Today's highly connected, insights-driven world would not be possible without the advent of data lake solutions. That's because organizations rely on comprehensive data lakes platforms, such as Azure Data Lake, to keep raw data consolidated, integrated, secure, and accessible. Scalable storage tools like Azure Data Lake Storage can hold and protect data in one central place, eliminating silos at an optimal cost. This lays the foundation for users to perform a wide variety of workload categories, such as big data processing, SQL queries, text mining, streaming analytics, and machine learning. The data can then be used to feed upstream data visualization and ad-hoc reporting needs. A modern, end-to-end data platform like Azure Synapse Analytics addresses the complete needs of a big data architecture centered around the data lake.

Data lake use cases

With a well-architected solution, the potential for innovation is endless. Here are just a few examples of how organizations across a range of industries use data lake platforms to optimize their growth:

- **Streaming media.** Subscription-based streaming companies collect and process insights on customer behavior, which they may use to improve their recommendation algorithm.
- **Finance.** Investment firms use the most up-to-date market data, which is collected and stored in real time, to efficiently manage portfolio risks.
- **Healthcare.** Healthcare organizations rely on big data to improve the quality of care for patients. Hospitals use vast amounts of historical data to streamline patient pathways, resulting in better outcomes and reduced cost of care.
- **Omnichannel retailer.** Retailers use data lakes to capture and consolidate data that's coming in from multiple touchpoints, including mobile, social, chat, word-of-mouth, and in person.
- **IoT.** Hardware sensors generate enormous amounts of semi-structured to unstructured data on the surrounding physical world. Data lakes provide a central repository for this information to live in for future analysis.
- **Digital supply chain.** Data lakes help manufacturers consolidate disparate warehousing data, including EDI systems, XML, and JSONs.
- **Sales.** Data scientists and sales engineers often build predictive models to help determine customer behavior and reduce overall churn.

Data lake vs. data warehouse

Now you know what a data lake is, why it matters, and how it's used across a variety of organizations. But what's the difference between a data lake and a data warehouse? And when is it appropriate to use one over the other?

While data lakes and [data warehouses](#) are similar in that they both store and process data, each have their own specialties, and therefore their own use cases. That's why it's common for an enterprise-level organization to include a data lake and a data warehouse in their analytics ecosystem. Both repositories work together to form a secure, end-to-end system for storage, processing, and faster time to insight.

A data lake captures both relational and non-relational data from a variety of sources—business applications, mobile apps, IoT devices, social media, or streaming—without having to define the structure or schema of the data until it is read. Schema-on-read ensures that any type of data can be stored in its raw form. As a result, data lakes can hold a wide variety of data types, from structured to semi-structured to unstructured, at any scale. Their flexible and scalable nature make them essential for performing complex forms of data analysis using different types of compute processing tools like [Apache Spark](#) or [Azure Machine Learning](#).

By contrast, a data warehouse is relational in nature. The structure or schema is modeled or predefined by business and product requirements that are curated, conformed, and optimized for SQL query operations. While a data lake holds data of all structure types, including raw and unprocessed

data, a data warehouse stores data that has been treated and transformed with a specific purpose in mind, which can then be used to source analytic or operational reporting. This makes data warehouses ideal for producing more standardized forms of BI analysis, or for serving a business use case that has already been defined.

	Data lake	Data warehouse
Type	Structured, semi-structured, unstructured	Structured
	Relational, non-relational	Relational
Schema	Schema on read	Schema on write
Format	Raw, unfiltered	Processed, vetted
Sources	Big data, IoT, social media, streaming data	Application, business, transactional data, batch reporting
Scalability	Easy to scale at a low cost	Difficult and expensive to scale
Users	Data scientists, data engineers	Data warehouse professionals, business analysts

	Data lake	Data warehouse
Use cases	Machine learning, predictive analytics, real-time analytics	Core reporting, BI

Data lake vs. data lakehouse

Now you know the difference between a data lake vs. a data warehouse. But what's the difference between a data lake and a data lakehouse? And is it necessary to have both?

Despite its many advantages, a traditional data lake is not without its drawbacks. Because data lakes can accommodate all types of data from all kinds of sources, issues related to quality control, data corruption, and improper partitioning can occur. A poorly managed data lake not only tarnishes data integrity, but it can also lead to bottlenecks, slow performance, and security risks.

That's where the data lakehouse comes into play. A data lakehouse is an open standards-based storage solution that is multifaceted in nature. It can address the needs of data scientists and engineers who conduct deep data analysis and processing, as well as the needs of traditional data warehouse professionals who curate and publish data for business intelligence and reporting purposes. The beauty of the lakehouse is that each workload can seamlessly operate on top of the data lake without having to duplicate the data into another structurally predefined [database](#). This ensures that

everyone is working on the most up-to-date data, while also reducing redundancies.

Data lakehouses address the challenges of traditional data lakes by adding a [Delta Lake storage layer](#) directly on top of the cloud data lake. The storage layer provides a flexible analytic architecture that can handle ACID (atomicity, consistency, isolation, and durability) transactions for data reliability, streaming integrations, and advanced features like data versioning and schema enforcement. This allows for a range of analytic activity over the lake, all without compromising core data consistency. While the necessity of a lakehouse depends on how complex your needs are, its flexibility and range make it an optimal solution for many enterprise orgs.

	Data lake	Data lakehouse
Type	Structured, semi-structured, unstructured	Structured, semi-structured, unstructured
	Relational, non-relational	Relational, non-relational
Schema	Schema on read	Schema on read, schema on write
Format	Raw, unfiltered, processed, curated	Raw, unfiltered, processed, curated, delta format files

	Data lake	Data lakehouse
Sources	Big data, IoT, social media, streaming data	Big data, IoT, social media, streaming data, application, business, transactional data, batch reporting
Scalability	Easy to scale at a low cost	Easy to scale at a low cost
Users	Data scientists	Business analysts, data engineers, data scientists
Use cases	Machine learning, predictive analytics	Core reporting, BI, machine learning, predictive analytics

What is data lake architecture?

At its core, a data lake is a storage repository with no set architecture of its own. In order to make the most of its capabilities, it requires a wide range of tools, technologies, and compute engines that help optimize the integration, storage, and processing of data. These tools work together to create a cohesively layered architecture, one that is informed by big data and runs on top of the data lake. This architecture may also form the operating structure of a data lakehouse. Every organization has its own unique configuration, but most data lakehouse architectures feature the following:

- **Resource management and orchestration.** A resource

manager enables the data lake to consistently execute tasks by allocating the right amount of data, resources, and computing power to the right places.

- **Connectors for easy access.** A variety of workflows allow users to easily access—and share—the data they need in the form that they need it in.
- **Reliable analytics.** A good analytics service should be fast, scalable, and distributed. It should also support a diverse range of workload categories across multiple languages.
- **Data classification.** Data profiling, cataloging, and archiving help organizations keep track of data content, quality, location, and history.
- **Extract, load, transform (ELT) processes.** ELT refers to the processes by which data is extracted from multiple sources and loaded into the data lake's raw zone, then cleaned and transformed after extraction so that applications may readily use it.
- **Security and support.** Data protection tools like masking, auditing, encryption, and access monitoring ensure that your data remains safe and private.
- **Governance and stewardship.** For the data lake platform to run as smoothly as possible, users should be educated on its architectural configuration, as well as best practices for data and operations management.

Additional resources

Explore

Solution ideas

Guides

Webinars

Frequently asked questions

- A data lake is a centralized repository that ingests, stores, and allows for processing of large volumes of data in its original form. It can accommodate all types of data, which is then used to power big data analytics, machine learning, and other forms of intelligent action.

[Learn more about data lakes](#)

- Organizations across a range of industries, including retail, finance, and entertainment, use data lake platforms to store data, gather insights, and improve the overall quality of their services. Investment firms, for example, use data lakes to collect and process up-to-market data, allowing them to manage portfolio risks more efficiently.
- Data lakes store all types of raw data, which data scientists may then use for a variety of projects. Data warehouses store cleaned and processed data, which can then be used to source analytic or operational reporting, as well as specific BI use cases.

[Explore data lakes vs. data warehouses](#)

- A data lakehouse combines elements of a data lake and a data warehouse to form a flexible, end-to-end solution for [data science](#) and business intelligence purposes.

[Learn more about data lakehouses](#)

- Absolutely. Major organizations across all industries rely on the massive amounts of data stored in data lakes to power intelligent action, gain insights, and grow.

[Discover the benefits of data lakes](#)

- Large volumes of data, including raw and unstructured data, can be difficult to manage, leading to bottlenecks, data corruption, quality control issues, and performance issues. That's why it's important to maintain good governance and stewardship practices to help you run your data lake platform smoothly.
- Data lake architecture refers to the specific configuration of tools and technologies that helps keep data from the data lake integrated, accessible, organized, and secure.

[Explore best practices for data lake architecture](#)

Free Account

Pay as you go

Get started with pay-as-you-go pricing. There's no upfront commitment—cancel anytime.