Informatica

What Is a Data Lake?

Exploring what data lakes are and how they help in business

Table of Contents

What Is a Data Lake?

A data lake is a centralized repository that stores data regardless of source or format. Data lakes let you store data in multiple forms — structured, semi-structured or unstructured, raw or granular. Data lakes help organizations manage their petabytes of big data. In a data lake, companies can discover, refine and analyze data with batch processing for AI, machine learning (ML) and data science use cases.

Data lakes power analytics for better business insights and decision-making. No matter what industry you're in, a data lake can enhance customer experiences and give your business a competitive edge. In the retail sector, cloud data lakes let you leverage your customer data to create a differentiated and highly personalized experience for each customer. In financial services, a cloud data lake can help you make trusted data actionable so it can be accessed by users, applications and ML processes.

The Evolution of Data Warehouses and Data Lakes

Traditional databases and on-premises storage can only go so far when it comes to data management. This is especially true with the immense amount of data generated by today's organizations. In the early internet era, data silos helped companies manage several different types of data. But data silos were not organized together in a way that led to meaningful insights. And data silos could not make the most of data for organizations seeking to modernize their data to the cloud.

Accelerate ROI from Your Data Lakes and Cloud Data Warehouses

From data silos to data warehouses

A data warehouse is an enterprise infrastructure that allows businesses to bring together and access various structured data sources. These data sources include the kind that were historically managed with different silos. In data warehouses, structured data is standardized, formatted and organized. This structure makes it easier for search engines and other tools to read and understand the data. Examples of structured data include addresses organized into columns or phone numbers and health records all coded in the same way. In short, data warehouses are organized, making structured data easy to find. However, data warehouses have a harder time understanding unstructured data.

These days, data stored in an enterprise comes from a variety of sources — both structured and unstructured. Unstructured data includes clicks on social media, input from IoT devices and user activity on websites. All this information can be valuable to commerce and business, but it is more difficult to store and track than structured data.

Hadoop and data lakes

In the early 2000s, Apache Hadoop, a collection of open-source software, allowed for large data sets to be stored across multiple machines. The data sets could be treated as if they were a single file. Companies could more easily handle and analyze large amounts of unstructured data. This marked the beginning of data lakes.

How Do Data Lakes Work?

You can use a data lake for both data storage and compute. The data lake architecture can use a combination of cloud and on-premises locations. Unlike a data warehouse, data lake or cloud data lake architectures are designed to be effective for structured, semi-structured and unstructured data. A data lake can manage structured data much like databases and warehouses can. But data lakes can also handle unstructured data that is not formatted or organized in a predetermined way.

As the volume of unstructured data has grown in the enterprise, effective data management has become a business imperative. Data lakes are an effective way to store diverse data. They can scale data management up to petabytes and beyond. And you don't need a specific structure schema for data to flow into the data lake. Just as rivers, streams and other waterways flow into a lake, data from across the business environment can easily flow into a data lake.

What platforms can support a data lake?

Hadoop was the first platform to support data lakes with an on-premises and cost-effective model. Early data lake platforms were not scalable and were limited in what they could accomplish. Since the earliest days of on-premises data lakes, various models and platforms have expanded to cloud storage.

Amazon Web Services (AWS) was the first cloud-based data lake. AWS allows customers a greater degree of scalability and flexibility. Other services like Azure Data Lake were quick to follow. They all took advantage of cloud storage and computing to offer businesses quality data management and data preparation.

Different platforms can offer specific services for different data types. Informatica for Google Cloud Storage (GCS) optimizes the value and insight of Google Analytics. It integrates easily with other Google properties such as Google Ads and YouTube. It allows users to manage metrics from the full range of the Google ecosystem.

Data Lakes Challenges & Solutions

Data governance

Today's data users access data lakes for querying, consuming and reporting on data. Without solid data governance, even well-constructed data lakes can turn into data swamps. These data swamps often are disorganized and difficult to use, understand and share. When you onboard a comprehensive data governance and catalog solution, you can monitor and observe what's in your data lake, who is using your data and how it's being used.

Data collection provisioning

Collecting and ingesting data from different sources into a data lake can create a bottleneck for analytics and AI/ML initiatives. A code-free unified data ingestion solution can efficiently ingest any data, any pattern, at any latency into a cloud data lake in minutes for processing, reporting, real-time analytics, data science and AI/ML usage.

Disparate data silos

In multicloud environments, data can be siloed, and the needs of data consumers may not be met. In addition, data repositories are often not coordinated. Data lakes help organizations overcome these challenges, bringing a unified view of their data. Petabytes of structured, semi-structured and unstructured data can be ingested into a single repository. This includes both streaming and batch data. Then all data users can seamlessly access data for reporting, analytics, data science and other needs.

Data quality

Poor data quality can lead to missed business opportunities, poor decisions, loss of revenue and increased costs. In a data lake, data quality solutions enable data stewards to better oversee quality and identify when the data is corrupt or inaccurate. Data engineers and IT professionals can establish data quality rules and processes so users can see what changes are made as the data is being cleansed.

Data drift

Unexpected and undocumented changes to the data structure, or data drift, can break processes and corrupt data. Detecting changes in the data structure at the source are a key pain point for many organizations. Your data integration solution should be able to handle data drift intelligently and automatically propagate changes to your data lake.

Data protection, availability and retention

Data loss, data inconsistencies and data corruption can occur in a cloud data lake. This is due to human error or cyber-attacks. All aspects of a data lake — its architecture, implementation and operation — must center on protecting your data in transit and at rest. To ensure data durability, resiliency and availability, a modern cloud data lake should have robust data replication that operates programmatically in the background.

Cost overruns

Cloud environments can help eliminate data center expenses. Organizations presume this will lower ongoing costs. Applying a flexible, data consumption-based pricing model with a data lake will help. This model ensures that each user and team pays only for the precise compute

and storage resources they use. Autonomous data management (AutoDM) can help ease bottlenecks, using metadata, automation and AI to standardize and accelerate data delivery with minimal human intervention.

Complex operations

Cloud data lakes often suffer from the same operational issues as data warehouses. Custom hand-coded or point solutions increase the total ownership cost, are hard to maintain and lack enterprise scalability. With an enterprise-scale DataOps solution, organizations can quickly build, manage and operationalize data pipelines. They can onboard new data sources into a cloud data lake for driving cloud analytics and AI.

Difficult-to-handle large metadata

With distributed data from different departments, applications, data warehouses and data lakes, it's difficult to know what data you have and where it is located. The volume of metadata in a cloud data lake increases proportionally. Metadata management is critical when building a data-driven business. A portfolio of best practice processes and technologies enables users of all kinds to search, understand and access the data they need to do their jobs.

AI/ML projects making it from pilot to production

According to Accenture's report, "AI: Built to Scale," 84% of business executives believe they need to use AI to achieve their growth objectives. However, 76% acknowledge struggling with how to scale AI across their business.[1] Companies need the ability to deploy machine learning models automatically and govern them in a data lake. Otherwise, companies cannot gain measurable value from AI. No-code/low-code machine learning operations (MLOps) help data engineers operationalize ML models developed in any workbench or framework. MLOps helps eliminate ad-hoc scripting and manual processes.

Structure and governance

Storing large amounts of unstructured data in one place has its challenges. If a data lake lacks standards or governance, it can quickly become a data swamp. Data swamps may be rich with information but lack data insights. Dirty data can hold a lot of information, but it's not useful until it's cleaned with good data management. Because of the lack of structure, it's difficult to glean value from a data swamp — leaving useful insights buried in its depths. Building a data lakehouse on your data lake can help.

A data lakehouse provides structure and governance to data. Cloud lakehouses have curated zones that enable data to move easily from the lake to the warehouse. This makes trusted data available to more users. But the data lake can still ingest unstructured, semi-structured or raw

data from a variety of sources. A data lakehouse brings together both the strengths of the data lake and the data warehouse on one platform. This makes the contents of a data lake more accessible to data scientists, AI and any other person or resource that can use it.

Data Lake Key Features

Key features of modern data lakes include:

Data ingestion

Data lakes should be able to handle batch, streaming and change data capture (CDC) data. In a data lake, you can move data in bulk from the source to the target with minimum or no change. You can automatically capture data changes in the source system and replicate your data in real time in the target data lake with CDC. Ingest any data — real-time streaming data or bulk data — from on-premises systems, such as file systems, mainframes or data warehouses, using scalable streaming and mass ingestion methods. This is done with comprehensive and high-performance connectivity for batch, streaming or near real-time processing.

Connectivity

Provide access to practically any data type, such as databases, on-premises systems, SaaS applications, and all data lakes, using high performant intelligent connectivity.

Data integration

Integrate data from various disparate data sources at any latency and rapidly develop extract, load, transform (ELT) or extract, transform, load (ETL) data flow. Plus, apply data, transformations into a single, unified view for business consumption.

Data quality

Provide a holistic data quality experience that delivers trusted, timely and relevant information to support any analytics or AI/ML initiatives. This is critical to cloud data lakes, especially when dealing with a variety of data.

Metadata management

Manage metadata to allow data lake users in an enterprise to gain visibility into their data inventory. Metadata management also answers questions about business context, lineage and value. This enables data engineers to manage expanding data volumes and data sources in a self-service manner. It also helps data stewards to define processes and ensure data consistency across multi-cloud environments.

Data Lake Benefits

Comprehensive data from an enterprise's multiple data sources including IoT deliver many benefits to businesses, such as:

Quickly and flexibly ingest data

Large amounts of unstructured data are a reality for nearly all industries, and data lakes provide the means to quickly and easily store and manage petabytes of disparate data.

Scalability

Industries that dealt in terabytes just a decade ago now verge on dealing with petabytes. Data lakes can handle colossal volumes of data — and, since they live in the cloud, they can expand with the needs of your business.

Productivity and accessibility

Good data and analytics can inform better policies, illuminate opportunities and demonstrate how resources can be efficiently used. Data lakes provide a means to rapidly store unstructured data prior to it becoming available for analytical tasks data-driven decision-making. In addition, many data lakes technologies are open source, making them affordable.

Sharper decision-making

If you don't store data, you can't glean insight from it. Data lakes allow decision-makers to make better decisions from insights derived from both structured and unstructured data.

Better data science

Scientists and engineers need access to data, and data lakes give them more information to work with and analyze than traditional forms of data storage. AI and machine learning can benefit from data lakes, as they rely on the quality of data input into them.

Data Lake Use Cases by Industry

Every industry relies on data. Here's how you can benefit from using a data lake to store and manage data.

- Healthcare

  Healthcare data ranges from a single patient's heartbeat or oxygen levels to large-scale studies of cancer and other diseases. Whether in a clinical or research situation, healthcare data comes from a variety of sources, in a variety of formats and needs to be accessed by a variety of users. Data lakes give organizations the ability to ingest unstructured data.

- Government

Government and public sector organizations collect and organize a variety of data. Information includes census data, public records and data regarding public services like electrical grids. Much of this data does not have a unifying schema. Data lakes make storage of this unstructured data much easier. Public officials gain insights about population dynamics, utilities, crime rates, migration and services with access to good data. Data lakes allow policymakers and experts to make informed decisions about laws, regulations and standards.

- Manufacturing

  Manufacturing relies on big data and real-time insights about supply chains, electricity costs, transportation and countless other activities. These data flows translate into billions of dollars worth of activity. Data lakes can turn a flow of unstructured data into a valuable source of insights and analytics. This allows manufacturers to make decisions based on good business intelligence on a routine basis.

- Financial Services

  Banking and capital markets increasingly incorporate AI and machine learning for everything from business planning to customer engagement. Algorithmic trading requires data sets that inform traders about which stocks to buy and sell. Data helps traders understand where potential value will grow. Their decisions happen in fractions of a second and constantly draw on the data contained within a data lake. Every trade and transaction generates new data that flows into a data lake.

Data Lake Examples

Here's a real-world success story where data lakes play a key role in driving business differentiation:

Sunrun

Hailing from San Francisco, Sunrun has been a leader in the solar power industry since 2007. The company needed to update, streamline and simplify its data architecture, reporting and visualization. Using a data lake and data warehouse model, Informatica helped Sunrun migrate from on-premises data storage to cloud infrastructure. With the new model, their IT professionals save a great deal of time. Reporting and visualization tasks that once took multiple quarters are now executed 3x faster. Read the full customer success story.

Conclusion

Data lakes are an important part of a large cloud data ecosystem. They provide a powerful way to manage and glean insights from data. Modern data lakes work well alongside an automated data management platform such as Informatica's AI-driven Intelligent Data Management Cloud™ (IDMC). Today, data management incorporates intelligent, AI-powered data integration,

data quality, data governance and master data management. Together, they prevent data lakes from degrading into data swamps. Data management and governance are crucial elements of business success, whatever your industry.

Data Lakes Resources

Informatica solutions for data lakes are vendor-agnostic. They provide an intuitive, UI-based approach that requires no hand coding. Learn more about how Informatica can help you make the most of your data and turn information into insight:

- Explore Informatica's data lake solutions.
- Learn more with the CDO's Guide to Intelligent Data Lake Management.
- Get started with an Informatica free trial.

[1] https://www.accenture.com/us-en/insights/artificial-intelligence-summary-index