LEARNING MADE EASY

2nd Snowflake Special Edition

Cloud Data Lakes

for dummies®
A Wiley Brand

What a cloud data lake can achieve

How it enables a world of data collaboration

Tips for choosing a cloud data lake

Brought to you by

snowflake®

David Baum

About Snowflake

Snowflake delivers the Data Cloud — a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, Snowflake delivers a single and seamless experience across multiple public clouds. Snowflake's platform is the engine that powers and provides access to the Data Cloud, creating a solution for data warehousing, data lakes, data engineering, data science, data application development, and data sharing. Join Snowflake customers, partners, and data providers already taking their businesses to new frontiers in the Data Cloud. **snowflake.com.**

Cloud Data Lakes

2nd Snowflake Special Edition

by David Baum

for dummies®
A Wiley Brand

Cloud Data Lakes For Dummies®, 2nd Snowflake Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-119-89935-8 (pbk); ISBN 978-1-119-89936-5 (ebk)

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Development Editor: Brian Walls
Project Manager: Jennifer Bingham
Acquisitions Editor: Ashley Coffey
Editorial Manager: Rev Mengle
Business Development Representative: Molly Daugherty
Content Refinement Specialist: Tamilmani Varadharaj
Snowflake Contributors Team: Scott Teal, Vincent Morello, Dmytro Yaroshenko, Julian Forero, James Malone, David Gaule

Table of Contents

Introduction

According to a December 2021 report from TDWI on "Data Engineering and Open Data Lakes," the software industry is witnessing a massive shift from cloud data warehouses to cloud data lakes because of the data lake's superior flexibility. They fulfill a promise that has been long in the making: The need for a vastly scalable solution that can easily ingest, integrate, analyze, share, and secure any amount of data, in just about any format, without requiring the data to be modeled or stored in a predefined structure. This flexibility allows data professionals to "load data first and ask questions later," broadening the horizons of business intelligence, predictive analytics, application development, and other data-driven initiatives.

However, despite continued enthusiasm for the data lake paradigm, poorly constructed data lakes can easily turn into *data swamps* — unorganized pools of data that are difficult to use, understand, and share with business users. This has been happening for more than a decade. To mitigate this risk, the most advanced cloud data lakes are created on top of *cloud data platforms* — scalable solutions that combine everything great about data warehouses, data lakes, and other key workloads into one cohesively managed solution.

About This Book

Find out how a modern cloud data platform can ensure the success of your data initiatives based on four key principles:

- **Using all data together seamlessly:** A modern data lake allows you to integrate structured, semi-structured, and unstructured data, even across public clouds and regions.
- **Enabling fast, reliable performance:** Simplifying your architecture with an elastic compute engine for many workloads eliminates concurrency resource issues.
- **Collaborating on a common data set:** Securely sharing data across an organization; with customers, supply chain partners, and other companies; and externally with thousands of data providers and consumers can reveal new opportunities.

- **The essential aspects of data management:** Prioritizing data governance, data security, and data privacy.

Foolish Assumptions

In creating this book, I've made a few assumptions:

- You're a business user, data scientist, data engineer, data platform architect, data warehouse manager, or executive.
- You want to know how a data lake can store, integrate, analyze, visualize, or share data from a variety of sources.
- You want to enable data-driven business decisions and new business opportunities with a data lake.

Icons Used in This Book

Throughout this book, you'll find the following icons that highlight tips, important points to remember, and more:

TIP

This icon guides you to faster ways to perform essential tasks, such as better ways to put a cloud data lake to work.

REMEMBER

Here you'll find ideas worth remembering as you immerse yourself in the exciting world of data lake concepts.

CASE STUDY

Throughout this book, case studies provide best practices from organizations that have successfully operated cloud data lakes.

Beyond the Book

Visit `www.snowflake.com` to find loads of additional content about cloud data lakes and related topics. You'll also find contact information in case you want to get in touch with Snowflake or try Snowflake for free as your cloud data lake.

IN THIS CHAPTER

- Flowing data into lakes
- Acknowledging the limitations of traditional data lakes
- Discussing the pros and cons of cloud object storage
- Introducing modern cloud data lakes
- Looking at who uses modern data lakes and why

Chapter 1
Introducing Cloud Data Lakes

This chapter digs into the history of the data lake. It explains why this type of data repository emerged, what data lakes can do, and why traditional data lakes have fallen short of the ever-expanding expectations of today's data professionals.

Flowing Data into the Lake

What's behind the name *data lake?* Picture data streaming in from many different sources, all merging into one expansive pool.

Now, compare that vision to the function-specific "ponds" that characterize special-purpose data management systems, such as data warehouses and data marts designed explicitly for finance, human resources, and other lines of business. These siloed analytic systems typically load structured data into a predefined schema, such as a relational database, and easily accessible via Structured Query Language (SQL) — the standard language used

to communicate with a database. By contrast, the hope for data lakes was to store many types of data in their native formats to facilitate ad hoc data exploration and analysis. In addition to the orderly columns and rows of relational database tables, these data lakes would store *semi-structured* and *unstructured data,* and make that data available to the business community for reporting, analytics, data science, and other pressing needs (see Figure 1-1).

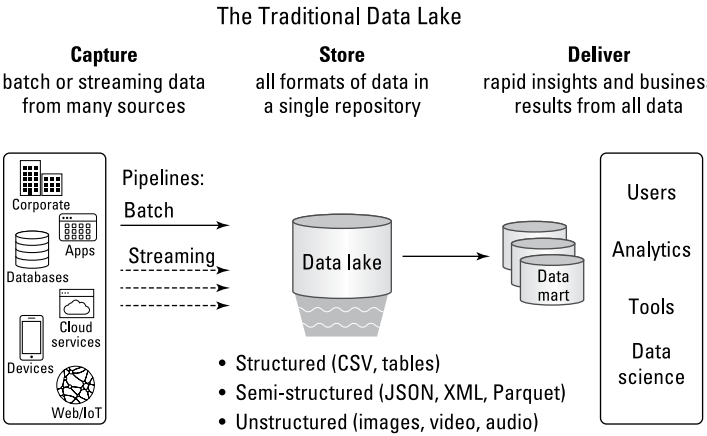


FIGURE 1-1: Data lakes were intended to capture data from many sources and store it in a versatile data repository that could serve many types of data users and data analytic workloads.

Understanding the Problems with Traditional Data Lakes

Data lakes arose to supplement data warehouses because the relational model can't easily accommodate today's diversity of data types and their fast-paced acquisition models. While data warehouses are generally designed and modeled for a particular purpose, such as financial reporting, data lakes don't always have a predetermined use case. Their utility becomes clear later on, such as when data scientists conduct data exploration for feature engineering and developing predictive models.

While this data discovery process opens up near-limitless potential, few people anticipated the management complexity, lackluster performance, limited scaling, and weak governance that characterized these open-ended data lake implementations.

Part of the problem was the inherent complexity of these early data lakes. The core technology was based on the Apache Hadoop ecosystem, an open source software framework that distributes data storage and processing among commodity hardware located in on-premises data centers.

Many of these data lake projects failed to fulfill the promise of data lake computing due to expensive infrastructure, slow time to value, and extensive system management requirements.

The inherent complexities of a distributed architecture and the need for custom coding for data transformation and integration, mainly handled by highly skilled data engineers, made it difficult to derive valuable insights and outcomes. It was easy to load and store huge amounts of data in many different formats but difficult to obtain valuable insights from that data.

Acknowledging Interim Solutions: Cloud Object Stores

In the years since data lakes were first introduced, cloud computing has evolved, and data storage technologies have matured considerably. Many organizations now leverage object storage services, such as Amazon Simple Storage Service (S3), Microsoft Azure Blob Storage, and Google Cloud Storage, as attempts to create their own data lakes from scratch.

Not having to create or manage compute clusters and storage infrastructure, as was necessary with Hadoop, is a big step forward. However, cloud object stores don't offer a total data lake solution either. For example, although customers no longer have to provision and scale a distributed hardware stack, they still have to create, integrate, and manage complex software environments. This involves setting up procedures to access and sometimes transform data, and establishing and enforcing policies for data security, data governance, identity management, and other essential activities. Finally, customers have to figure out how to achieve adequate performance for a variety of codependent analytic workloads, such as business intelligence, data engineering, and data science, all of which may compete for the same pool of compute and storage resources.

Other common problems include difficulty managing and scaling the environment, and inadequate procedures for managing data quality, security, and governance. Without attention to these complex issues, even well-constructed data lakes can quickly become data swamps. The greater the quantity and variety of data, the more significant this problem becomes. That makes it harder to derive meaningful insights, as depicted in Figure 1-2.

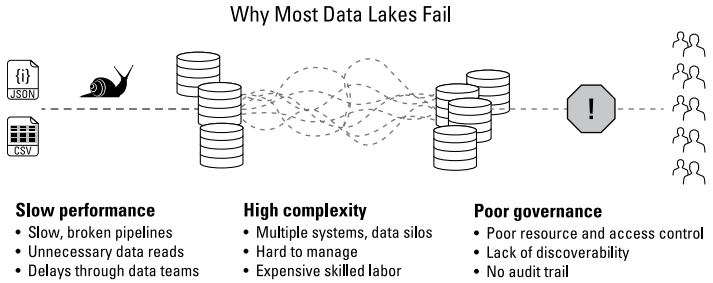


FIGURE 1-2: Slow performance, complexity, and poor governance are among the reasons traditional data lakes often fail.

REMEMBER

Cloud object stores allow organizations to store and analyze unlimited amounts of data in their native formats. However, that leaves organizations to take charge of data management, data transformation, data protection, data governance, data compliance, and many other complex activities.

Reviewing Modern Requirements

Despite these early failings, the original promise of the data lake remains: a straightforward and powerful way for organizations to collect, store, integrate, analyze, and share their data from a single repository. They want to explore, refine, and analyze petabytes of data without a predetermined notion of the data's structure.

Most of today's data lakes, however, can't effectively organize all of that data, let alone properly secure and govern that data.

REMEMBER

To be truly useful, a data lake must include a cohesive set of tools that reveal *what* is in the data lake, *who* is using which data, and *how* that data is being used, along with assurances that all data is protected. It must also store data in their native formats and

facilitate user-friendly data exploration, automate routine data management activities, and support a broad range of use cases and workloads, such as modern data sharing. What's more, the data lakes of today must be fed by a number of data streams, each of which delivers data at a different frequency, without imposing onerous requirements on the data engineering teams that build these data pipelines. And it must handle all of this without any storage or performance limitations.

To meet these needs, a third and far better data lake paradigm has arisen. These solutions have become the foundation for the modern data lake: A cloud-built repository where structured, semi-structured, and unstructured data can be staged in their raw forms — either in the data lake itself or in an external object storage service.

Anchored by a *cloud data platform,* these newer data lakes provide a harmonious environment that blends many different data management and data storage options, including a cloud analytics layer, a data warehouse, and a cloud-based object store. With the right software architecture, these data lakes provide near-unlimited capacity and scalability for the storage and computing power you need. They make it easy to derive insights, obtain value from your data, and reveal new business opportunities.

Explaining Why You Need a Modern Cloud Data Lake

This book reveals how to create innovative, cost-effective, and versatile data lakes — and extend existing data lakes created using Hadoop, cloud object stores, and other limiting technologies. It relies on a modern architecture that is secure, resilient, easy to manage, and supports many types of users and workloads.

In addition to anchoring a versatile data lake, standardizing on a well-architected cloud data platform has many other advantages. For example, it makes it easy to share data with authorized users without requiring database administrators to copy that data or establish a new data silo, all while upholding centralized data security and governance policies. It makes it easier to accommodate new design patterns, such as a data mesh, and integrate new data formats, such as Apache Iceberg tables. And yet, even

with this diversity, the entire environment can be operated with familiar SQL tools, while data professionals can also use their chosen languages, tools, open source libraries, and development frameworks. A consumption-based pricing model should accompany the data lake to ensure each user and team only pays for the precise compute and storage resources they use. Best of all, a modern cloud data platform should operate seamlessly across multiple public clouds via one consistent management interface, so your DevOps team can ensure maximum continuity, and your organization will never be limited to one single cloud provider.

REMEMBER

As on-premises data lakes decline in popularity and cloud object stores show their limitations, new architectural paradigms based on cloud data platforms are revealing their potential. Because all storage objects and necessary compute resources are internal to the platform, data can be accessed, analyzed, modeled, and manipulated quickly and efficiently. This is much different from the original data lake architectures, where data was always stored in an external data bucket and then copied to another loosely integrated storage-compute layer to achieve adequate analytics performance.

Looking at Which Industries Use Modern Data Lakes and Why

Modern cloud data lakes can play an important role in every industry. For example, ecommerce retailers use modern data lakes to collect clickstream data for monitoring web-shopping activities. They analyze browser data in conjunction with customer buying histories to predict outcomes. Armed with these insights, retailers can provide timely, relevant, and consistent messaging and offers for acquiring, serving, and retaining customers.

Oil and gas companies use data lakes to improve geological exploration and make their extraction operations more efficient and productive. Data from hundreds or thousands of sensors helps oil and gas companies discover trends, predict equipment failures, streamline maintenance cycles, and understand their operations at very detailed levels.

Banks and financial services companies use data lakes to analyze market risks and determine which products and services to offer. In much the same way, nearly all customer-focused organizations

can use data lakes to collect and analyze data from social media sites, customer relationship management (CRM) systems, and other sources, both internal to the company and via third-party data services. They can use all that data to gauge customer sentiment, adjust go-to-market strategies, mitigate customer support problems, and create highly personalized experiences for customers and prospects.

REMEMBER

Traditional data lakes fail because of their inherent complexity, poor performance, and lack of governance, among other issues. By leveraging the capabilities of a cloud data platform, modern data lakes overcome these challenges.

Foundational tenets of these versatile, high-performance data lakes include:

- **No data silos:** Easily store and access petabytes of structured, semi-structured, and unstructured data from a single platform, even across multiple clouds, in a cohesive way.
- **Fast and flexible:** Allow developers and other experts to work with data in their preferred languages. For example, data engineers can process data with Java, data scientists can run models in Python, and analysts can query with SQL.
- **Instant elasticity:** Supply nearly any amount of computing resources to any user or workload. Dynamically change the size of a compute cluster without affecting running queries, or scale the service to easily include additional compute clusters to complete intense workloads faster.
- **Concurrent operation:** Deploy to a near-unlimited number of users and workloads to access a single copy of your data, all without affecting performance. For example, you may merely want to run analytical queries, then later allow developers to build data-intensive applications.
- **Inherent control:** Present fresh and accurate data to users, focusing on data sharing and collaboration, data quality, access control, and metadata management.
- **Reliable:** Confidently combine data to enable multi-statement, ACID transactions.
- **Fully managed:** The data platform automates many aspects of data provisioning, data protection, security, backups, and performance tuning, allowing you to focus on analytic endeavors rather than on managing hardware and software.

INSTANT ELASTICITY FOR HEALTHCARE ANALYTICS

CASE STUDY

Scripps Health is a nonprofit healthcare system based in San Diego, California, that includes 5 acute-care hospital campuses and 28 outpatient centers and clinics. It has more than 16,000 employees and treats 600,000 patients annually via 3,000 affiliated doctors.

Previously, Scripps relied on an on-premises Hadoop cluster and a legacy data warehouse platform for healthcare analytics. The system supported several data warehouse use cases but required a specialized IT team to develop, administer, scale, and tune it for adequate performance. The team phased out the Hadoop cluster and subscribed to a modern cloud data platform and cloud blob storage.

Today, Scripps stores high-priority, or "hot," data in the cloud data platform, and stores archival "cold" data in cloud blob storage. By adopting this low-maintenance environment, Scripps achieved a 50 percent reduction in full-time equivalent (FTE) staff dedicated to database administration, and reduced its software licensing costs by 60 percent. Previously, users retrieved data and analyzed it on their own systems, creating data silos. Now, users access data through the intuitive interface of the cloud data platform, eliminating those silos.

The cloud data platform distinctly separates but logically integrates storage and compute resources into independently scalable entities, enabling Scripps Health to scale capacity up and down as needed. Each business unit pays only for the compute resources it consumes, and a data-masking feature masks plain-text data at query time for stronger security, an important factor when dealing with patient data and personally identifiable information (PII).

Now that Scripps Health has a robust modern cloud data lake, with repeatable processes for a variety of data-intensive workloads, it is experimenting with developing predictive analytics, building statistical models, and retrieving data using a standard ODBC connection. As a fully managed cloud solution, near-zero maintenance frees data professionals at Scripps Health to focus on revealing fresh insights to advance strategic business initiatives.

IN THIS CHAPTER

- Boosting team productivity
- Supporting popular data science tools
- Building on the right architecture
- Accommodating many different workloads

Chapter 2
Enabling Modern Data Science and Analytics

For many data science initiatives, data lakes are the repositories of choice. However, managing data in today's data lakes is fraught with difficulty. According to an Anaconda report titled "The State of Data Science 2020: Moving from Hype Toward Maturity," data scientists spend an average of 45 percent of their time preparing data in the data lake before they can use it to develop machine learning (ML) models and visualize the outcomes in meaningful ways.

This chapter describes the three fundamental attributes of a data lake that help ensure successful data science and other types of analytic endeavors:

- The capability to seamlessly *combine* and easily *access* multiple types of data, all stored in one universal repository
- The freedom for data scientists to collaborate using their chosen tools, frameworks, libraries, and languages
- An architecture that allows data scientists, business analysts, and other data professionals to collaborate productively over data without having to contend for compute and storage resources

Establishing a Data Foundation

Data lakes were born out of the necessity of big data analytics. These multipurpose repositories provide the technology organizations need to store data until data scientists discover potential uses and applications. However, traditional data lakes can be difficult to secure, govern, and scale. They may also lack the crucial metadata data scientists need to make sense of the information. Metadata is data about data.

A cloud data platform resolves these issues by providing a natural structure for many types of data. In addition to capturing raw data, as is common for a data lake, it stores and manages the metadata that allows data scientists to conduct meaningful analyses, such as tagging fields in a document and categorizing patterns within images. Having a common metadata layer also helps various data users collaborate with the data by ensuring accurate, consistent results when the data is displayed through dashboards and reports.

REMEMBER

The services layer is the linchpin of a modern cloud data platform. It manages metadata, transactions, and other operations. It performs these activities locally or globally across multiple regions and clouds, enforcing centralized security and governance as it tracks, logs, and directs access to every database element and object within the data lake, as shown in Figure 2-1.

The Architecture of a Cloud Data Platform

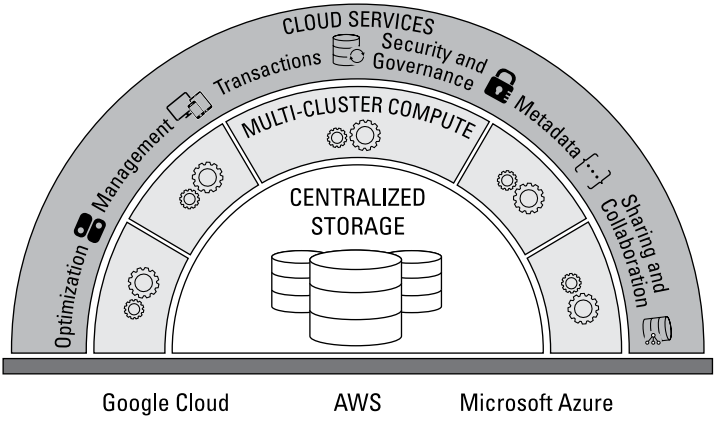


FIGURE 2-1: A data platform should support many data types and span public clouds to unite the work of data analysts, data scientists, and data engineers.

Boosting Team Productivity

A properly architected data lake supports multiple business units and workloads, with one centralized repository rather than multiple data silos serving discrete needs. The data platform enables a single dynamic copy of the data that can populate and update ML models, business intelligence (BI) dashboards, and predictive analytic apps. It also orchestrates analytics, data sharing, data ingestion, and data science.

This architecture allows data professionals to easily process data relevant to their sphere of operations. Whether creating data pipelines, conducting feature engineering, developing data applications, issuing queries, or setting up data-sharing relationships, all teams can collaborate on a unified, shared repository of data. This synergy is especially valuable for data science teams. Consolidating data into one central location streamlines the data science workflow by facilitating collaboration among all workflow participants, including data scientists, data engineers, and ML engineers, as depicted in Figure 2-2.

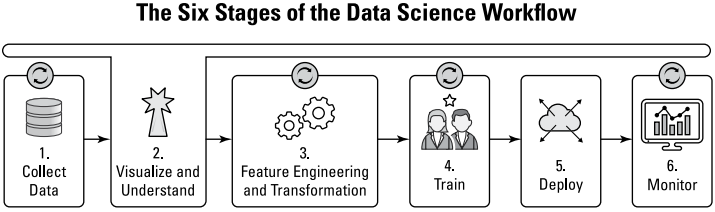


FIGURE 2-2: The data science workflow involves a multi-step process of populating, training, and deploying ML models. Data scientists, data engineers, ML engineers, and analysts should all be able to use their tools of choice.

REMEMBER

Having a complete services layer is what makes a data lake useful. It rationalizes differences among various data types, so people don't have to look for it in multiple places. It applies centralized security and governance, even when the data set spans multiple clouds and multiple regions. This eliminates the inconsistent results that arise when various work groups use different copies of the data.

Supporting Languages and Tools

Today's data science teams use a broad range of software tools, algorithms, open source libraries, and ML principles to uncover business insights hidden in vast volumes of data. Whether writing queries, building data pipelines, or embedding custom logic in a software program or procedure, it should be simple for data professionals to interact with the platform directly, without having to move data from one database to another. These highly paid workers are most productive when they can collaborate on a single shared version of data, upholding universal security constraints, even when they use multiple tools.

To allow all types of data professionals to work productively, your data lake must support popular ML frameworks and languages. Data engineers commonly use SQL, Python, and Java to prepare data. Data scientists use Python, Structured Query Language (SQL), and R to explore data relationships, conduct feature engineering, and train ML models. Ideally, your data lake should enable a *data frame* style of programming preferred by many technology experts, which aligns data into a two-dimensional array, much like the structured rows and columns of a relational database or spreadsheet.


TIP

When your data platform is architected to support multiple teams and workloads without forcing each team to contend for resources, the entire data science practice becomes timelier and more productive. Data scientists output the results of ML activities back into the data platform for general-purpose analytics, even as data engineers load data and business analysts consume it. A common repository allows BI apps to leverage the results of data science initiatives and put the data to work throughout the business. It also ensures reliable outcomes: All front-end apps reference the same back-end data definitions, ensuring consistent results for queries, forecasts, dashboards, and reports.

Accommodating Multiple Workloads and Communities

With a traditional data platform, fixed compute and storage resources limit concurrency — the capability for many users to deploy many data workloads simultaneously. A cloud data

platform built on a *multi-cluster, shared data architecture* scales compute and storage resources independent from each other and near-infinitely. This allows multiple users to query the same data without degrading performance, even as other workloads operate simultaneously, such as ingesting data or training an ML model.

A well-architected data lake also allows data users to combine data generated by an organization with third-party data sets, such as those acquired from its business partners or purchased from data marketplaces. By doing so, the organization gains previously unobtainable insights about its business and its customers. The enriched data and the insights it generates also create new market opportunities in the form of monetizing data and data applications that extend data science learnings to internal and external communities. An organization can offer these commercial data sets and data applications to customers and partners and also list them on a data marketplace. A modern cloud data platform should offer the capability to connect to a cloud data marketplace, where an ecosystem of third-party data, technology, and data service providers bring additional data, tools, and services into the ecosystem, broadening what's possible for data science teams.

REMEMBER

A cloud data platform makes a data lake more useful. It fosters collaboration and ensures the entire organization has a scalable data environment for data science and related analytic endeavors. For example, data scientists and ML engineers can access raw data straight from the data lake for feature engineering and modeling activities while business analysts generate reports via self-service dashboards — all without degrading performance.

START WITH THE BEST ARCHITECTURE

A multi-cluster, shared data architecture includes three layers that are logically integrated yet scale independently from one another:

- **Storage:** A single place for all structured, semi-structured, and unstructured data
- **Compute:** Independent computing resources dedicated to each workload to eradicate contention for resources
- **Services:** A common services layer that handles infrastructure, security, metadata, query optimization, and much more

OPTIMIZED PRICING, STORE LOCATION, AND SUPPLY CHAIN VIA DATA SCIENCE

CASE STUDY

Żabka owns the largest chain of convenience stores in Poland, with more than 7,000 stores visited daily by more than 2.5 million customers. With millions of daily transactions, the amount of data soon overwhelmed Żabka's on-premises data warehouse and data lake Hadoop cluster, making it impossible for data scientists to load and analyze data concurrently.

With its previous data lake environment, Żabka could analyze transaction data to optimize prices for each store, but data engineers and business analysts had to share a finite set of compute and storage resources. Meanwhile, Żabka's data scientists wanted to create new ML models to enable a more advanced product pricing strategy but could only work at certain times of the day.

Żabka switched to a modern cloud data platform to establish a modern data environment that anchors both the data warehouse and a new data lake, transforming its ability to make data-driven decisions. The new data platform allows each of these teams to instantly add additional computing power during high-traffic hours so they can load data, run queries, refine models, and generate reports as needed. For example, data scientists pulled data from the data lake to identify 14 consistent store segments, including internal data on transactions, marketing promotions, and assortments. They combined this data with dozens of external data sets containing the prices and locations of competitors, upcoming events, geographic coordinates, and demographic information.

These advanced data science models have allowed Żabka to optimize pricing for each product in each store, which has increased revenue and margins. In addition, a new revenue-estimation model allows team members to determine the most effective locations for new stores. Żabka can also share this near real-time data with suppliers to increase sales, personalize consumer communication, and perform market research. Insights gathered from its Poland stores will be valuable for expanding into other countries.

IN THIS CHAPTER

- Planning your data lake implementation
- Complying with privacy regulations
- Establishing comprehensive data security
- Improving data retention, protection, and availability

Chapter 3
Reducing Risk, Protecting Diverse Data

Your organization's data is incredibly valuable, and this book is all about maximizing that value with the latest technologies for storing, analyzing, and gaining useful insights from that data. However, your data is also valuable to bad actors who are continually unleashing malware viruses, phishing schemes, and other nefarious plots designed to steal or compromise your data assets. In the process, they may force your organization to pay a ransom to call off the attack. According to a recent report from Cybersecurity Ventures, ransomware costs are expected to reach $265 billion by 2031, while global cybercrime costs will grow 15 percent per year over the next five years, reaching $10.5 trillion annually by 2025.

This growing risk of malicious attacks is compounded by internal threats, mishaps, and compliance violations, often stemming from simple errors, omissions, or failure to apply software patches in a timely manner. This chapter discusses the need to plan carefully and deliberately as you set up your data lake to deliver the best data security, privacy, and regulatory compliance.

Facing Facts about Data Security

If you entrust your data to a cloud provider or software-as-a-service (SaaS) vendor, will they keep it secure? In the early days of cloud computing, this was a hotly debated topic. Today, the superiority of cloud security is one of the motivating factors that encourages organizations to put their data in the cloud. Cloud providers such as Amazon, Microsoft, and Google have established sophisticated security operation centers (SOCs) staffed by elite teams of IT professionals trained in the most current cybersecurity practices. Reputable SaaS providers have followed suit. As a result, a well-architected and properly maintained cloud data lake can be more secure than the data warehouses and data lakes that you host in your own data center.

All aspects of a data lake — its architecture, implementation, and operation — must center on protecting your data. Your data security strategy should include data encryption and access control, in conjunction with comprehensive monitoring, alerts, and cybersecurity practices. You must also monitor and comply with data privacy regulations that govern the use and dissemination of customer data.

However, ensure you understand precisely what your data platform vendor provides. Security capabilities vary widely among vendors. And although they might have good security, they differ in their degree of automation and assistance. Some cloud vendors automate only rudimentary security capabilities, leaving many aspects of data encryption, access control, and security monitoring to the customer. Others handle these tasks for you.


TIP

Effective security can be complex and costly to implement. Cybersecurity professionals are hard to come by. Instead of building an in-house security operations center from scratch, if you subscribe to a modern cloud data platform with automated security capabilities, you can achieve a high level of data protection as soon as you enable the data platform.

Encrypting Data Everywhere

Encrypting data, which means applying an encryption algorithm to translate the clear text into ciphertext, is a fundamental security feature. Data should be encrypted both "at rest" and "in

transit," meaning when the data is stored on disk when moved into a staging location for loading into the data lake, when it is placed within a database object in the data lake itself, and when it is cached within a virtual data lake. Query results must also be encrypted.

End-to-end encryption should be the default, with security methods that keep the customer in control, such as customer-managed keys. This type of "always on" security is not a given with most data lakes, as many highly publicized on-premises and cloud security breaches have revealed.

Managing Encryption Keys

After you encrypt your data, you'll decrypt it with an encryption key (a random string of bits generated specifically to scramble and unscramble data). To fully protect the data, you must protect the key that decodes your data. A robust data lake should handle data encryption and key management automatically, all the time, for all data, when it is in transit and at rest.


TIP

The best data lakes employ AES 256-bit encryption with a hierarchical key model rooted in a dedicated hardware security module to add layers of security, protection, and encryption. They also instigate key-rotation processes that limit the time during which any single key can be used. Data encryption and key management should be entirely transparent to the user but not interfere with performance.

Automating Updates and Logging

Cybersecurity is never static. The security measures you apply to your data lake must evolve to reflect today's dynamic threat landscape. That means always keeping up with security patches that address known threats.

Ideally, these security updates should be applied automatically to all pertinent components of the cloud data platform as soon as those updates are available. If you use a cloud provider, that vendor should also perform periodic security testing (also known as *penetration testing*) to proactively check for security flaws. These

safeguards should not impact your daily use of the cloud data platform.


TIP

As added protection, verify your data lake vendor uses file integrity monitoring (FIM) tools, which ensure critical system files aren't tampered with. All security events should be automatically logged in a tamper-resistant security information and event management (SIEM) system. The vendor must administer these measures consistently and automatically, and they must not affect query performance.

Controlling Access to Sensitive Data

All users must be authorized before accessing or manipulating data in the data lake. For authentication, ensure your connections to the data platform provider leverage standard security technologies, such as Transport Layer Security (TLS) 1.2 and IP whitelisting. (A *whitelist* is a list of approved email addresses or domain names from which an email-blocking program will allow messages to be received.) A cloud data lake should also support the SAML 2.0 standard so you can leverage your existing password security requirements and existing user roles. Regardless, multifactor identification (MFA) should be required to prevent users from logging in with stolen credentials. With MFA, users are challenged with a secondary verification request, such as a onetime security code sent to a mobile phone.


REMEMBER

After a user has been authenticated, it's important to enforce authorization to specific parts of the data based on that user's "need to know." A modern data lake must support multilevel, *role-based access control* (RBAC) functionality so users requesting access to the data lake are authorized to access only the data they are explicitly permitted to see.

In addition to this basic authentication, *fine-grained access control* allows database administrators to apply security constraints and rules to certain parts of each object, such as at the row level and column level within a database table. Access constraints can also be applied to compute servers to control which users can execute large data processing jobs. Another useful feature is *geofencing*, which allows the administrator to set up and enforce access restrictions based on the users' location.

COMMON WAYS TO STORE DATA

Data lakes use files, blocks, and objects to store and organize data.

- File storage organizes data as a hierarchy of files in folders. It is popular for unstructured data such as documents and images, especially when used in low-latency applications such as high-performance computing (HPC) and media processing.
- Block storage divides data into evenly sized volumes, each with a unique identifier. It is commonly used for databases that require consistent performance and low-latency connectivity.
- Object storage breaks files into pieces that can be spread out among hardware platforms, each object acting as a self-contained repository. It is useful for unstructured data, such as music, video, and image files.

As you add semi-structured and unstructured data to your data lake, other important stipulations apply. Granular access control becomes more difficult with the file-based storage often found in a data lake (see the "Common Ways to Store Data" sidebar), which doesn't conform to a tabular structure. With many of today's object stores, security may be "all or nothing": You either have access to the storage layer or don't. To bolster this basic security, your data lake provider should apply fine-grained RBAC measures to all database objects, including tables, schemas, and any virtual extensions to the data lake.

In some instances, you can also use *secure views* to prevent access to highly sensitive information most users don't need to see. This security technique allows you to selectively display some or all the fields in a table, such as only allowing HR professionals to see the salary fields in an employee table.

Complying with Data Privacy Regulations

For sensitive data, such as tables that populate financial reports or columns that contain personally identifiable information (PII), knowing where data resides within your data lake is critical to satisfying regulatory compliance requirements. Privacy regulations are increasingly rigorous, and organizations can't ignore

them. Leading the way are Europe's General Data Protection Regulation (GDPR), the United States' Health Insurance Portability and Accountability Act (HIPAA), and the California Consumer Protection Act (CCPA). Corporate data governance policies should verify data quality and standardization to ensure your data is properly prepared to meet these requirements. The types of information that fall under these specific guidelines include credit card information, Social Security numbers, names, dates of birth, and other personal data.

Certifying Attestations

Data breaches can cost millions of dollars to remedy and permanently damage customer relationships. Industry-standard attestation reports verify that cloud vendors use appropriate security controls and features. For example, your cloud vendors need to demonstrate they adequately monitor and respond to threats and security incidents and have sufficient incident response procedures in place.

In addition to industry-standard technology certifications, such as ISO/IEC 27001 and SOC 1/SOC 2 Type II, verify that your cloud provider also complies with all applicable government and industry regulations. Depending on your business, this could include Payment Card Industry Data Security Standards (PCI-DSS), GxP data integrity requirements, HIPAA/Health Information Trust Alliance (HITRUST) privacy controls, ISO/IEC 27001 security management provisions, International Traffic in Arms Regulations (ITAR), and FedRAMP certifications. Ask your providers to supply complete attestation reports for each pertinent standard, not just the cover letters.

REMEMBER

An important stipulation within these data privacy regulations is the *right to be forgotten*, which means consumers can opt out of communications from merchants or vendors. In these instances, all links to and copies of their PII must be erased from a vendor's information systems. When all your data is stored in one universal repository that automatically manages metadata and lineage, fulfilling these requests is much easier. With minimal manual intervention, the platform should automatically detect PII and apply the appropriate policies to that information, even as data is loaded, staged, and moved across multiple tables and objects.

Isolating Your Data

If your data lake runs in a multitenant cloud environment, you may want it isolated from all other data lakes. If this added protection is important to you, ensure your cloud data platform vendor offers this premium service. Isolation should extend to the virtual machine layer. The vendor should isolate each customer's data storage environment from every other customer's storage environment, with independent directories encrypted using customer-specific keys.

If your company must adhere to certain data sovereignty requirements, then investigate the regional penetration of your cloud provider's coverage. For example, will the provider enable you to maintain sensitive data in specific cloud regions? Can you store encrypted data in the cloud and the encryption keys on premises? These capabilities are especially important in Europe and other highly regulated regions.

TIP

Work only with cloud providers that can demonstrate they uphold industry-sanctioned, end-to-end security practices. Security mechanisms should be built into the foundation of the data platform. You shouldn't have to do anything extra to secure your data.

Finally, data security and compliance hinge on traceability. You must know where your data comes from, where it is stored, who has access to it, and how it is used, which Chapter 4 discusses.

REDUCING THE RISK OF SHARING DATA

CASE STUDY

Portland General Electric (PGE) is a fully integrated energy company with statewide operations in Oregon, serving 1.9 million people in 51 cities.

Previously, PGE managed a legacy, on-premises data warehouse that was expensive to maintain and had performance issues. The system's tightly coupled architecture was inflexible. In addition, multiple copies of the data proliferated across the organization, making it difficult to identify the authoritative source of data-driven insights.

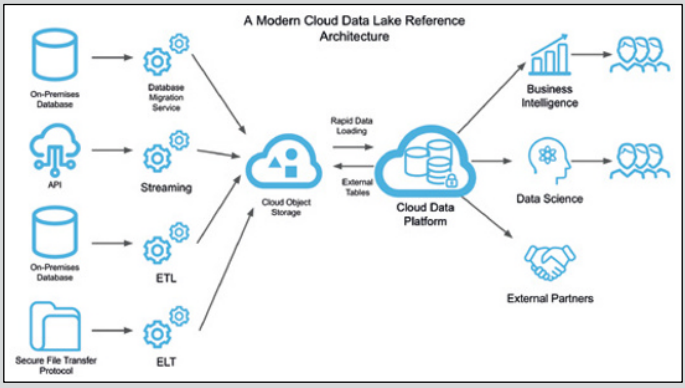
(continued)

(continued)

This environment also increased PGE's data storage costs. Realizing the need for a modern data environment, PGE selected a cloud data platform as a foundation for a modern data lake that features high performance, separation of storage from compute, near-zero maintenance, and an extensive security architecture.

Today, the cloud data platform increases security and governance capabilities for PGE's data lake, including data files stored in cloud object storage. Secure views on external tables keep data in place while providing row-level and column-level access to the data based on user IDs. Users are authenticated via a single sign-on into a Tableau business intelligence environment. Data requested from Tableau dashboards are access-controlled by user-level privileges managed within secure views.

The cloud data platform also supports secure data sharing. Previously, providing data to external partners was a complex process. Now, PGE can seamlessly provide data to external groups and internal groups, such as its data science team, without copying or moving data. Instead of making and sending static copies of the production data set, PGE enables data scientists with read-only access to the data that remains in its original location and is updated in near real time with a click of a button. PGE uses secure data sharing and secure views within its cloud data platform to maximize the accessibility of data while minimizing risks.



IN THIS CHAPTER

- Instituting robust data governance
- Cataloging and classifying your data
- Automatically detecting data schema
- Improving data quality
- Tracing data lineage

Chapter 4
Preventing a Data Swamp

How do you prevent your data lake from becoming a data swamp — a quagmire of unmanageable data? You must start with a data platform that automatically collects metadata and enforces systematic data governance.

Data governance ensures data is properly classified and accessed. Metadata helps you understand exactly what data you have and how people use it.

Your data platform should track who uploads data, when, and what type of data it is. It should also identify key fields and values — a capability that is especially important when dealing with personally identifiable information (PII).

This chapter explains how a modern data lake achieves these objectives with the following essential ingredients: a data catalog, automatic schema detection, capabilities to track data lineage, and provisions for classifying and adding business context to your data.

Understanding Metadata

A data lake achieves effective governance by keeping track of where data is coming from, who touches the data, and how various data sets relate to one another. A robust cloud data platform automatically generates this type of metadata for files in internal stages (that is, within the data lake) and external stages (such as Amazon S3, Google Cloud Storage, or Microsoft Azure Blob Storage). This metadata is often maintained in virtual columns and queried using standard commands, such as Structured Query Language (SQL) SELECT statements, and loaded into a table with the regular data columns.

Cataloging Data

Much of the data loaded into first-generation data lakes isn't usable because it hasn't been cataloged. Picture a storage unit where you stash numerous items — family heirlooms and treasures as well as old furniture, mismatched clothing, and castoffs of questionable value. As the years pass, you may not have a clear idea of what is even stored there.

And so it is with a traditional data lake. If no designations and identifiers are on the data, users have difficulty finding and retrieving relevant information — and don't even know what is there in the first place.

A data catalog helps users sort this out by empowering them to discover and understand the data. Most data catalogs include a self-service portal that makes it easy to view and understand metadata, which improves accuracy and enables more confident decision-making.

Whereas many organizations opt to use an external data catalog, modern data lakes are gradually evolving to include internal catalogs. Some solutions include directory tables that act as built-in file catalogs.


REMEMBER

If you don't catalog your data, you can quickly end up with a data swamp. A data catalog keeps track of what types of information you have, who can access it, and how popular it is, along with the lineage of the data and how it is used.

Detecting Data Schema

A *schema* defines how data is organized, structured, and related to other data. Schema objects can include table names, fields, data types, and the relationships among these entities. Whereas a data catalog keeps track of what data you have, a data schema helps you make sense of it.

A data lake that offers automatic schema detection can be helpful, especially to prepare semi-structured data for querying and analytics. For example, suppose a data engineer wants to create a table from Parquet files. In that case, the data lake can automatically register all the fields and data types, either by copying the file data into relational tables (schema on write) or by querying the file data in place (schema on read). To make it easier to query and join multiple data sets, look for a data platform that can detect schema in popular file types, such as Parquet, Avro, and ORC formats.

REMEMBER

A data lake can accommodate many types of data because it's not constrained by a predefined schema. However, data analysts need to know the schema of all the data sets and which tables and columns represent common entities. Schema detection automates these operations.

Classifying and Contextualizing Data

With such a high velocity and variety of data going into your data lake, how can you keep track of sensitive data and PII to preserve strong customer relationships and avoid compliance violations? For example, if a marketing team is collecting data about customers, they will likely acquire personal information, such as email addresses, phone numbers, and credit card numbers. Modern data lakes should use data classification tools to identify certain types of PII, helping database administrators to classify, control, and monitor its usage. These tools can identify where sensitive data is stored and ensure proper protection and monitoring for a growing number of data types.

Some data platforms can automatically understand the context of each part of the data set, such as when it was created, when it was last modified, and how it fits within the context of your business.

This can help auditors understand which database objects contain PII. Classifying data by department or business function can also help the business allocate costs to particular departments and cost centers.

To maximize the utility of your data lake, you need to know not just where data is located and what types of sensitive data it contains but also how and when it is used and by whom. Some data platforms maintain records and metrics that reveal how broadly the data is used. This helps data stewards manage the flow of data from inception to deletion to maximize its value and minimize data management costs. For example, frequently accessed data might be maintained in a high-performance storage environment ("hot" storage), then later placed in less expensive "cold" storage for archival purposes once it is being accessed less frequently.

Ensuring Data Quality

Data governance requires oversight to maintain the quality of the data your organization shares with its constituents. Bad data can lead to missed or poor business decisions, loss of revenue, and increased costs. Data stewards — people charged with overseeing data quality — can identify when data is corrupt or inaccurate, when it's not being refreshed often enough to be relevant, or when it's being analyzed out of context.

Ideally, you should assign responsibility for data quality efforts to the business users who own and manage the data because they're the people in the best position to note inaccuracies and inconsistencies. These data stewards should work closely with IT professionals and data engineers to establish data quality rules and processes with full transparency so users can see what changes are made as stewards cleanse the data.

REMEMBER

Considering that it's common to load raw data into a data lake, it's important to give users the capability to ensure the data is of the quality needed for the tasks at hand. For example, analysts generating reports need very clean data, which could mean removing duplicates and ensuring there are no missing values. However, a network security analyst querying event logs may want data in its rawest form to get a granular view of potential problems in source systems. You can meet these disparate needs

by curating data into logical "zones" based on how much the data has been transformed, mapped, modeled, and cleansed.

Building Trust with Data Lineage

With many types of users accessing various logical data zones, and many data pipelines refreshing them with new or transformed data, it is easy to lose visibility into the origin of information. Tracking the data's lineage helps users make sense of the data by revealing how it flows into the data lake, how it is transformed and manipulated, and where it goes when it flows out of the data lake.

Data lineage tools — either resident in the data platform or available through add-on services — help you understand the *journey* data follows through all your data-processing systems: what sources the data comes from, where it flows to, and what happens to it along the way. These technologies create a detailed map of all direct and indirect dependencies among data entities. This knowledge can help compliance officers trace the usage of sensitive data. It can also help data engineers foresee the downstream impact of changes.

Simplifying the Data Lake Architecture

Proper data governance involves many complementary capabilities and tools. With each "point" solution you add to the technology stack, the more challenging it becomes to aggregate all your metadata in a useful way for end users and administrators. A complete data platform should synthesize and integrate these components into one cohesive architecture. Ideally, your data lake should rest on a cloud data platform that integrates all metadata and data governance capabilities into a seamless experience.

REMEMBER

Implementing effective governance early in the data lake development process will help you avoid potential pitfalls, such as poor access control, unacceptable data quality, and insufficient data security.

ENABLING A COMPREHENSIVE DATA STRATEGY

CASE STUDY

Founded in 1906, CEMEX is a global building materials company that offers cement, ready-mix concrete, aggregates, and urbanization solutions in growing markets around the world, powered by a multinational workforce focused on providing a superior customer experience, enabled by digital technologies. Previously, CEMEX needed a dedicated IT team in each of these regions to manage software maintenance, dashboard updates, report requests, and month-end reporting. At the end of each month, a surge of reporting and other data-intensive workloads created performance bottlenecks.

To modernize its data management strategy, CEMEX chose a cloud data platform that stores structured and semi-structured data as the foundation for both a data lake and a data warehouse, enabling secure and governed access to all data. The new data platform powers CEMEX Go, a digital environment that automates order-to-cash workflows, supports online purchases, and tracks real-time orders in 21 countries. Each year, more than 500,000 payments and 2.5 million deliveries are completed through CEMEX Go, and nearly 90 percent of CEMEX's customers use the environment.

In the past, adding capacity required weeks of effort. Now, the cloud data platform scales automatically to meet short- and long-term needs cost-effectively, so CEMEX does not have to plan for infrastructure upgrades. Having all data in one location simplifies reporting, customer dashboards, advanced analytics, and application development. One application evaluates GPS and traffic data to determine the best routes for the company's ready-mix concrete trucks. Another calculates the optimal distribution of trucks based on the location of ready-mix concrete plants and the forecasted demand. CEMEX only pays for the compute and storage resources each user and application consumes.

Going forward, CEMEX plans to develop machine learning applications that leverage the data lake to identify upsell and cross-sell opportunities and generate recommendations on pricing strategy, including dynamic pricing.

IN THIS CHAPTER

- Unleashing the full potential of your data
- Simplifying data lake maintenance
- Sharing and enriching data
- Improving resiliency and business continuity
- Maintaining data in multiple clouds
- Optimizing time-to-value with an easy-to-use system

Chapter 5
Selecting a Modern Cloud Data Lake

By leveraging the unique attributes of a modern cloud data platform, your data lake can accommodate the needs of many types of data professionals. It can consolidate data across multiple public clouds with unlimited scale, exceptional reliability, minimal maintenance, and cost-effective pricing. This chapter describes some of the essential factors to consider as you identify the right data platform for your modern cloud data lake.

Empowering Many Users, Workloads, and Tools

Whether it's building new data applications or supporting new data science projects, a data lake must be able to keep up with the growth of your business. A modern cloud data lake should deliver all the resources you need, with instant elasticity and near-infinite scalability. You shouldn't have to overprovision resources

to meet peak demands. Storage and compute resources should be separate from one another yet logically integrated and designed to scale automatically and independently. This would allow the data lake to support a near-unlimited number of concurrent users and workloads and easily scale up and down to handle fluctuations in usage without adversely impacting performance or requiring the organization to purchase more capacity than it needs.

Of course, different data users have different language and tool preferences. For example, data analysts may prefer to work with data via SQL or a business intelligence tool, whereas a data scientist may prefer to use Python in Jupyter Notebooks. Given that a data lake is designed to be a one-stop shop for all data, it should enable many different types of data users and their data workloads productively.

Reducing Overhead

All modern organizations depend on data, but none want to be saddled with tedious systems management and database administration tasks. How easy is it to subscribe to the service, load your data, authorize users, and launch your most critical data workloads? After your data lake is up and running, how easy is it to provision more resources and ensure great performance?

A modern cloud data lake should enable you to leverage all your data without having to provision infrastructure or manage a complex environment. Your skilled data professionals shouldn't have to bother with infrastructure, such as expanding storage capacity, allocating computing resources, installing security patches, and optimizing query performance. Security, tuning, and autoscalability should be built into the cloud service, freeing up your skilled data professionals to focus on gaining the most value from your data.


TIP

Offload important but avoidable administrative chores with a fully managed cloud data platform so your IT professionals can shift their attention to value-added activities, such as discovering new ways to analyze, share, and monetize data. Free up your data professionals to maximize the value of the data lake and the utility of its data.

Using All Data Types

To accommodate all possible business needs, your data lake should be versatile enough to ingest and immediately query data of many different types. That includes unstructured data, such as audio and video files, and semi-structured data, such as JSON, CSV, and XML. It should also allow you to include open source data formats, such as Apache Parquet and ORC.

The promise of the modern data lake is to enable you to seamlessly combine these many types of data so that you don't have to develop or maintain separate silos or storage buckets. With all your diverse data sources and metadata integrated into a single system, users can easily put that data to work and obtain data-driven insights.

But look a little deeper: Your data lake vendor may claim to "support" multiple data types, but how easy is it to synthesize them? For example, if you're flowing structured relational data from a CRM application into your data lake, how easy is it to combine these CRM records with semi-structured JSON data from an ecommerce weblog? If you're creating a machine learning (ML) model that monitors purchasing trends and predicts buying activity, can you dictate a schema for the JSON data that models these diverse data sources? Can you integrate a function for processing image files, say to pull in pictures from a product catalog? Do you have to figure out how to extract information from those images, or can you simply embed that function into a SQL query?

A good data lake stores diverse types of data in their native formats without creating new data silos and imposes a schema to streamline access to all your data. You don't have to develop or maintain separate storage environments for structured, semi-structured, and unstructured data. It is easy to load, combine, and analyze all data through a single interface while maintaining transactional integrity.

TIP

Here are some guidelines for smooth data management:

- Establish a complete metadata layer to guide user analytics.
- Standardize on an architecture that supports JSON, Avro, Parquet, and XML data, along with leading Open Table formats, such as Apache Iceberg, as needed.
- Use data pipeline tools that allow for native data loading with transactional integrity.

Capturing Data of Various Latencies

Your data platform should include data pipeline tools to migrate data into your cloud data lake. *Bulk-load processes* work best for initial transfers, especially if you have many terabytes of data to load into the data lake. After that, you'll most likely want to integrate only the changes that have occurred since the last data load, a processing technique known as change data capture (CDC).

Increasingly, real-time and near real-time data feeds are used for *streaming data processes* that load data continuously. These processes are designed to capture IoT data, weblog data, and other continuous sources emitted by mechanical equipment, environmental sensors, and digital devices, such as computer hardware and mobile phones. Distributed publishing/subscribing messaging services represent a popular way to send and receive streaming data. These services act as publishers and receivers to ensure that data is received by the subscribers. Examples include open source technologies such as Apache Kafka and commercial technologies such as Amazon Kinesis, Microsoft Event Hubs, Google Cloud Pub/Sub, and Snowflake Snowpipe. Open source tools offer low-cost solutions, but they generally require more setup, tuning, and management than commercial tools.


TIP

Ensure your data pipelines can move data continuously and in batch mode. They must also easily support schema-on-read and handle the complex transformations required to rationalize different data types without reducing the performance of production workloads or hindering user productivity.

Sharing and Enriching Data

A modern data lake should not only simplify the process of storing, transforming, integrating, managing, and analyzing all types of data. It should also streamline how diverse teams *share* data, so they can collaborate on a common data set without having to maintain multiple copies of data or move it from place to place.

Modern data sharing enables any organization to share and receive live data, within minutes, in a governed and secure way — with almost none of the risk, cost, headache, and delay that continue to plague traditional data-sharing methods. It permits

organizations to share the data itself and services that can be applied to that data, such as data modeling services, data enrichment services, and even complete data applications.


REMEMBER

Traditional data lakes aren't capable of modern data sharing. These older architectures use file transfer protocol (FTP), application programming interfaces (APIs), email, and other repetitive methods to duplicate static data and make it available to consumers. Lack of security and governance prohibits these older data lakes from enabling unlimited, concurrent access by data consumers. They also produce static data that quickly becomes dated and must be refreshed with up-to-date versions. That means constant data movement and management.

Modern cloud data platforms enable you to easily share the data in your data lake and receive shared data across your business or with organizations external to your own in a secure and governed way — without moving it from place to place (see Figure 5-1).

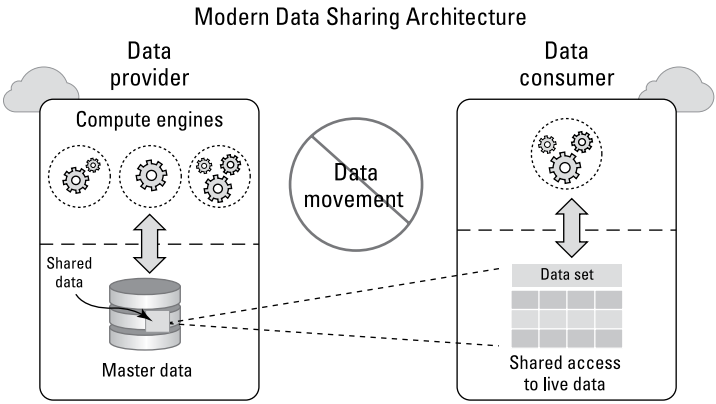


FIGURE 5-1: A modern data-sharing architecture enables access to live data between a data provider and a data consumer without moving data.

Your platform should also facilitate data sharing on a commercial scale by permitting organizations to tap into third-party data repositories, services, and streams. Using these modern data-sharing methods, organizations can share data with vendors, supply chain partners, logistics partners, customers, and many other constituents. They can also set up data-sharing services that turn their data lakes into profit centers. A multi-tenant architecture allows authorized members of the ecosystem to tap

into live, read-only versions of data and data functions within the data lake, and this ready-to-use data is immediately available all the time.

TIP

Look beyond the first-party data you own and consider second- and third-party data to improve your ML models and discover previously unknown patterns, correlations, and insights. Acquiring this external data can allow you to gain deeper insights, streamline operations, better serve customers, and discover new revenue streams based on data.

CASE STUDY

ELEVATING DATA ANALYTICS

Identity company Okta helps organizations securely connect people and technology. Thousands of organizations, including JetBlue and Slack, use the Okta Identity Cloud to manage access and authentication for employees, contractors, partners, and customers.

To enable data-driven decision-making across the company, Okta ingests and analyzes large amounts of product configuration and usage data. However, Okta's previous legacy cloud data architecture could not affordably scale to handle up to 500 million events per day from the Okta Identity Cloud. Resource contention led to multi-day data processing delays. Basic event stream queries took minutes to finish running, which negatively impacted the productivity of data analysts. Large, month-end processes took up to nine hours to complete, and Okta could not surface the insights that people were asking for.

Realizing the need for a modern data environment, Okta created a data lake on a cloud data platform, in conjunction with external storage on AWS. Today, ingesting data from numerous sources into the data lake provides Okta with a single source of truth for BI reporting and ad hoc analytics. The new data platform separates compute from storage resources, which has allowed Okta to near-instantly and near-infinitely scale resources to support more business units and create more reports.

For example, Okta's finance team can gather key metrics such as total customer count and net retention rate. Marketers have a unified view of advertising performance and attribution across all platforms, including Google, LinkedIn, and Facebook. Product teams monitor configuration and usage data to measure feature adoption and guide development decisions. And by combining product configuration and

usage data with CRM data, Okta surfaces pipeline opportunities that have resulted in millions of dollars in revenue. Okta's Director of Data and Analytics describes the platform as "a central nervous system that enables data sharing and self-service analytics."

Improving Resilience for Business Continuity

Even the most robust information systems can fail. In some cases, floods, fires, earthquakes, and other natural disasters can wipe out entire data centers. In other cases, cyberattacks can result in data loss, data inconsistencies, and data corruption. And don't forget the fallibility of internal personnel, such as when a database administrator inadvertently deletes a table from a database.

None of these crises or mishaps will cause lasting damage if your data lake architecture incorporates redundant processes and procedures to keep all data online, instantly available, and well protected so your critical workloads don't experience downtime.

To establish a workable strategy for data protection and business continuity, start by identifying the business impact of an outage for various workloads. From there, establish service level agreements (SLAs) that dictate your tolerance for downtime. What happens when a daily sales report is delayed? What is the downstream impact if an inventory dashboard isn't refreshed for several hours? Which databases are used by client-facing applications that drive revenue or customer experience? Answering these questions will help you understand user expectations and use those expectations to establish guidelines for data backups, data replication, data instance failover, and disaster recovery to ensure business continuity.


TIP

If avoiding downtime is critical to your operation, ensure your data lake provider uses data replication techniques, disaster recovery procedures, and instant failover technologies to insulate your operation from all these incidents. All cloud data lakes should protect data and ensure business continuity by performing periodic backups. Suppose a particular region of a public cloud provider experiences an outage, or even all its regions experience one. In that case, the analytic operations and applications that

need that data should automatically switch to a redundant copy of that data within seconds in another region or on another public cloud provider. Data retention requirements call for maintaining copies of all your data. It's important to replicate that data among multiple, geographically dispersed locations to offer the best possible data protection. The "triple redundancy" offered by some cloud vendors won't do you any good if all three copies of your data are in the same cloud region when an unforeseen disaster strikes.

A complete data protection strategy also considers regulatory compliance and certification requirements, which may stipulate that data be retained for a certain length of time for legal and auditing purposes.

Finally, pay attention to performance. Data backup and replication procedures are important, but if you don't have the right technology, these tasks can consume valuable compute resources and interfere with production analytic workloads. To ensure the durability, resiliency, and availability of your data, a modern cloud data lake should manage replication programmatically in the background without interfering with whatever workloads are executing at the time. Good data backup, protection, and replication procedures minimize, if not prevent, performance degradation and data availability interruptions.


TIP

Errors can come from many places, including human error, accidental deletions, and cloud infrastructure failure. You must be prepared for all of these. As explained in the next section, the data in your data lake, along with all the metadata, should be periodically replicated to multiple clouds and regions in conjunction with disaster recovery procedures that allow your data-driven operations to quickly failover to a replicated instance.

Supporting Multiple Clouds

A complete data protection strategy should go beyond merely duplicating data within the same physical region or zone of a cloud computing and storage provider. The data platform provider should be able to quickly shift all your production workloads from one region to a different region and ideally from one cloud provider to another cloud provider to uphold your SLAs.

Some data lake services are *multicloud,* meaning they can run on more than one major public cloud. Although this capability maximizes flexibility, it also propagates silos and negates the fundamental principle of centralizing all data in a data lake. If you transition a workload to a different region or cloud, will the data pipelines remain intact? Will all data security procedures and data governance policies be enforced?

Although multicloud capabilities can be useful, cross-cloud capabilities are superior. This not only means a system can run on any cloud but that it can also store and use data services among various clouds. For example, it can store data on one cloud and process it on another. This superior architecture enables you to leverage investments in Amazon Web Services, Microsoft Azure, or Google Cloud Platform and bring them all together in a cohesive way — or seamlessly transition from one to another.

REMEMBER

A modern cloud data lake should allow you to compile queries and coordinate database transactions across multiple regions and clouds, wherever data and users reside. It should also maintain transactional integrity for all data in any cloud worldwide. A common metadata layer should enforce consistency, even when data is stored in multiple clouds and across multiple regions.

Accommodating New Storage Paradigms

New types of data and new data storage paradigms are constantly appearing. For example, such table formats as Apache Iceberg are popular because they add a SQL-like table structure and ACID transactions to the unstructured and semi-structured data stored in files and documents. This allows computing engines, such as Spark, Trino, PrestoDB, Flink, Hive, Amazon EMR, and Snowflake, to easily manage and inspect the data. These newer data formats have tremendous momentum from the commercial and open source communities. Will your data platform support them if needed?

TIP

Whenever you adopt a storage paradigm or computing engine, opt for interoperability without compromising ease-of-use, enabling your technology professionals to work with their tools of choice, both now and in the future. For example, if you are currently using Avro files, you shouldn't be forced to change to Parquet

merely because your computing engine requires that storage format. Opt for a solution that works with multiple storage paradigms, data formats, and computing engines as necessary for your business.

Paying for What You Use

Your cloud data lake should offer a consumption-based pricing model. Each user and workgroup should pay only for the precise storage and compute resources used in per-second increments, so you never have to pay for idle capacity.

Contrast this pricing strategy with a subscription model, which requires customers to pay a recurring price for a set number of licenses or seats to use the SaaS provider's software. Although subscriptions work well for ensuring predictable revenue for SaaS vendors, this model can be challenging for customers. They must estimate upfront how many licenses they may need and pay a monthly fee without any guarantee of using all the licenses or features they contracted for.

To maximize your budget, ask yourself these questions: Are you paying for unused capacity? Have you purchased more storage and computing licenses than what's necessary? Are you oversubscribing to these resources to accommodate occasional but predictable surges in demand?

Work with a cloud data lake vendor that offers consumption-based pricing so you can "pay as you go" for resources actually consumed. You shouldn't agree to any multiyear licenses or service contracts, although you may get a better rate by committing to a minimum volume of usage.

TIP

To keep costs under control:

- Pay for usage by the second, not by the minute or month or a time frame affected by the busiest day of the year.
- Automatically increase and decrease data lake resources for daily, monthly, quarterly, or seasonal data surges.
- Eliminate onerous capacity-planning exercises by easily assessing your day-to-day requirements.

IN THIS CHAPTER

- Evaluating your needs
- Migrating data and workloads
- Establishing success criteria
- Setting up a proof of concept
- Quantifying value

Chapter 6
Six Steps for Planning Your Cloud Data Lake

Deploying a modern data lake requires careful planning and assessment of your current and future needs. Follow the steps in this chapter to get started.

Step 1: Review Requirements

Your data lake should allow you to store data in raw form, enable immediate exploration of that data, refine it in a consistent and managed way, and power a broad range of data-driven workloads. Consider these factors:

- **Data:** Identify the sources, types, and locations of the data you plan to load into your data lake. Will you gather new data? Will you stage data from an existing data warehouse or data store? Consider not only the data you have now but how other types of data could improve your operations, such as by powering new predictive models.
- **Users:** Determine who will be authorized to access the data, develop data-driven applications, and create new insights.

Compare the skills possessed by your current team against your plans for the business. If you plan to democratize access to business users, what tools or techniques will make the data accessible to them?

- **Access:** Do your business intelligence and data science tools use industry-standard interfaces and allow data professionals to work with popular frameworks and languages, such as Structured Query Language (SQL), Python, Java, and R? In particular, ensure your new data platform is ANSI-SQL compliant so you can discover value hidden within the data lake, and quickly deliver data-driven insights to all your business users.
- **Sharing:** Do you plan to share data across your organization or externally with customers or partners? If so, what types of data will you share, and will you use a data marketplace to monetize data? Identify archaic data sharing methods such as FTP and email, and consider how you can replace them with a modern data-sharing architecture.
- **Stewardship:** Determine who will be responsible for data quality, data governance, and data security, both for your initial data loads and continuously as new data is ingested.

Step 2: Migrate or Start Fresh

You may have an existing cloud data warehouse that you want to extend with new data types. Or perhaps you have an on-premises data lake created in Hadoop and use a cloud object store for additional data and files. Do you want to create a new data lake from scratch, using object storage from a general-purpose cloud provider or add to an existing object store? Do you have historical data sets you would like to migrate? If so, you will probably want to set up a one-time bulk transfer of this historical information to the data lake, then establish a pipeline to stream data, continuously or periodically, as your websites, IoT devices, data applications, and other apps generate new data.

Step 3: Establish Success Criteria

Identify important business and technical criteria, focusing on performance, concurrency, simplicity, and total cost of ownership (TCO). You should not have to install, configure, or maintain hardware and software. Backups, performance tuning, security updates, and other management tasks should be handled by the cloud solution provider. How will you define success once the data lake is in full production mode? Will your data-driven applications impact revenue? Do you plan to monetize data in your data lake?

Step 4: Evaluate Solutions

This book outlines the attributes you should be looking for in a modern cloud data lake. Popular choices include the following:

- Do-it-yourself open source platforms, such as Hadoop, Spark, Presto, and Hudi, which offer great flexibility and scalability, yet typically require complex infrastructure, custom coding, skilled engineering, and extensive system management requirements
- Object storage environments that use the near-boundless storage and compute services from Amazon, Google, Microsoft, and other vendors, on top of which you must develop and maintain your data lake environment
- Specialized cloud data platform solutions optimized for storing, analyzing, and sharing large and diverse volumes of data, driven by a common layer of services that simplify security, governance, and metadata management

TIP

Whatever type of cloud data platform you choose, ensure it can easily integrate many types of data in one universal repository to avoid creating data silos. If you plan to store data in a public object store, opt for a data platform that can accommodate these storage environments without the need to lift, shift, or copy data. Finally, the solution should support your existing skills, tools, and expertise and offer robust security and governance capabilities.

Step 5: Set Up a Proof of Concept

REMEMBER

A proof of concept (POC) tests a solution to determine how well it serves your needs and meets your success criteria. When setting up your POC, list all requirements and success criteria — not just the issues you're trying to resolve, but everything possible with a cloud solution. Ensure the new data lake overcomes the drawbacks of your current data management and analytic systems, such as making it easy to combine structured, semi-structured, and unstructured data. Can the storage layer accommodate multiple file formats in an efficient, cost-effective way? If you plan to deploy predictive use cases, does the platform make it easy to develop and apply data science models?

Step 6: Quantify Value

One platform is easier to manage than several, so consider the degree to which your new data platform can eliminate data silos and minimize your reliance on multiple solutions. Pay close attention to the services offered by the data platform vendor. Will the vendor handle data lake administration, security, management, and maintenance? If so, will you need fewer technology professionals than you did in the past? How will your month-to-month cloud usage fees compare to what you might have spent previously for on-premises software licenses and maintenance contracts?

TIP

Assuming you outsource everything to the vendor, you can calculate the TCO based on the monthly subscription fee or incremental usage fees. If you opt for an infrastructure-as-a-service (IaaS) or platform-as-a-service (PaaS) solution, you need to add the costs of whatever software, administration, and services the solution doesn't include. If you instead choose a cloud data platform, how much of this will the vendor manage for you? And don't overlook the savings possible when a cloud solution is scaled up and down dynamically in response to changing demand and when the vendor only charges by the second.

Transform your business with a cloud data lake

The data lake first emerged more than a decade ago as a single repository for easily storing, integrating, and analyzing all of an organization's data. Most data lake projects failed in the early days, but the goal remained. Fast-forward to today. The modern cloud data lake is that single repository and a whole lot more. Find out how your organization can use a cloud data lake for many types of workloads, such as data collaboration, allowing you to seamlessly and securely share and access live data across your enterprise, with your business partners and customers, and globally across major public clouds to create new business opportunities. Read on.

Inside...

- Avoid the mistakes of data lakes past
- Data science and other vital workloads
- The data lake for data collaboration
- Data governance, security, and privacy
- Evaluate different data lake options and learn practical steps to get started
- Read real-world data lake case studies

David Baum is a freelance business writer specializing in science and technology.

Go to Dummies.com™
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-119-89935-8
Not For Resale

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.