LEARNING MADE EASY

Snowflake Special Edition

Cloud Data Lakes

for dummies®
A Wiley Brand

What is a modern cloud data lake?

How it compares to other analytics solutions

Tips for choosing a cloud data lake

Brought to you by

snowflake™

David Baum

About Snowflake

Snowflake started with a clear vision: Make modern data warehousing effective, affordable, and accessible to all data users. Snowflake enables the data-driven enterprise with instant elasticity, secure data sharing, and per-second pricing, across multiple clouds. Because traditional on-premises and cloud solutions struggle at this, Snowflake developed a new product with a new built-for-the-cloud architecture that combines the power of data warehousing, the flexibility of big data platforms, and the elasticity of the cloud at a fraction of the cost of traditional solutions. Snowflake: Your data, no limits.

For more information, visit Snowflake at **snowflake.com**.

Cloud Data Lakes

Snowflake Special Edition

by David Baum

for dummies®
A Wiley Brand

Cloud Data Lakes For Dummies®, Snowflake Special Edition

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

ISBN 978-1-119-66624-0 (pbk); ISBN 978-1-119-66648-6 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. For details on how to create a custom *For Dummies* book for your business or organization, contact info@dummies.biz or visit www.wiley.com/go/custompub. For details on licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

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

Development Editor: Steve Kaelble
Project Editor: Martin V. Minner
Executive Editor: Steve Hayes
Editorial Manager: Rev Mengle
Business Development Representative: Karen Hattan
Production Editor: Tamilmani Varadharaj
Snowflake Contributors Team: *Vincent Morello, Michael Nixon, Clarke Patterson, Leslie Steere*

Table of Contents

Introduction

Data lakes emerged more than a decade ago to solve a growing problem: the need for a scalable, low-cost data repository that allowed organizations to easily store all data types from a diverse set of sources, and then analyze that data to make evidence-based business decisions.

But what about the data warehouse, which was the de facto solution for storing and analyzing structured data and preceded the data lake by 30 years? It couldn't accommodate these new, big data projects and their fast-paced acquisition models, many of which envisioned easily storing petabytes of data in structured and semi-structured forms. With the future of big data looming large, the data lake seemed like the answer: an ideal way to gather, store, and analyze enormous amounts of data in one location.

Interest in data lakes skyrocketed for one simple reason: Most organizations consider data a very important asset, and the systems of the time couldn't handle the variety. For decades, organizations have collected structured data from enterprise applications, and now they're supplementing it with these newer forms of semi-structured data from web pages, social media sites, mobile phones, Internet of Things (IoT) devices, and many other sources, including shared data sets. Data scientists, business analysts, and line-of-business professionals still need a way to easily capture, store, access, and analyze that data.

The initial data lakes were deployed on premises, mostly using open source tools from the Apache Hadoop ecosystem. In the decade since data lakes were first introduced, cloud computing has evolved and data storage technologies have matured. Better and easier ways to create data lakes have emerged that leverage the power, flexibility, and near-infinite scalability of the cloud.

About This Book

Cloud Data Lakes For Dummies is your guide to modern data lakes that combine the power of analytics with the flexibility of big data models and the agility and limitless resources of the cloud. Whether you're considering your first data lake or wish to update an existing one, this book offers ideas to help you achieve your business and technology goals.

Foolish Assumptions

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

- You're a business user, data scientist, data platform architect, data warehouse manager, or perhaps a company executive.
- You want to store, analyze, visualize, or share data from a variety of sources, and your existing solution is coming up short. Or perhaps you need access to massive data sets to train a machine learning model.
- You want to understand how a data lake can create new business opportunities and help your organization make better, more complete, and more timely decisions.

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 applied cloud data lakes.

Beyond the Book

Visit `www.snowflake.com` to find loads of additional content about cloud data lakes and related topics. Read other ebooks, view webinars, and get the scoop on upcoming events. 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 or for another business need.

IN THIS CHAPTER

- Flowing data into lakes
- Limitations of traditional data lakes
- Introducing modern data lakes
- Looking at who uses modern data lakes, and why

Chapter 1
Diving into Cloud Data Lakes

Back in 2010, James Dixon, who founded and served as chief technology officer for Pentaho, coined the term *data lake* to describe a new type of data repository for storing massive amounts of raw data in its native form, in a single location.

This chapter digs into the history of the data lake, why the idea emerged, and how it has evolved. It explores what data lakes can do, and where traditional data lakes have fallen short of ever-expanding expectations. It spells out data lake strategies and explains where the cloud data lake fits in.

Flowing Data into the Lake

What's behind the metaphoric name *data lake?* According to Dixon, think of a large body of water, into which new water streams from many channels, and from which samples are taken and analyzed.

It was a revolutionary concept. Prior to the data lake, most analytic systems stored specific types of data, using a predefined database structure. For example, data warehouses were built primarily for analytics, using relational databases that included a

schema to define tables of structured data in orderly columns and rows. By contrast, the hope for data lakes was to store many data types in their native formats and make that data available to the business community for reporting and analytics (see Figure 1-1). The goal was to enable organizations to explore, refine, and analyze petabytes of information without a predetermined notion of structure.

REMEMBER

The most important thing to understand about a data lake is not how it is constructed, but what it enables. It's a comprehensive way to explore, refine, and analyze petabytes of information constantly arriving from multiple data sources.

One petabyte of data is equivalent to 1 million gigabytes: about 500 billion pages of standard, printed text or 58,333 high-definition, two-hour movies. Data lakes were conceived for business users to explore and analyze petabytes of data.

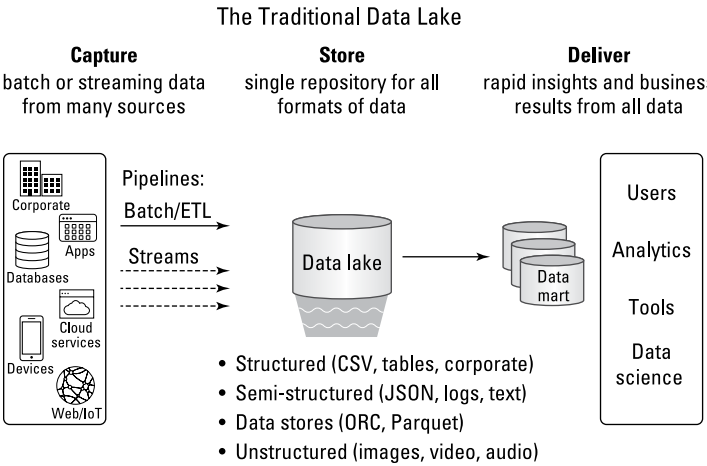

FIGURE 1-1: The original goal of the data lake, which failed to deliver the desired rapid insights.

Understanding the Problems with Data Lakes

The initial data lake concept was compelling, and many organizations rushed to build on-premises 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. Hadoop includes a file system called HDFS that enables customers to store data in its native form. The ecosystem also includes open source equivalents to Structured Query Language (SQL) — the standard language used to communicate with a database, along with batch and interactive data processing technologies, cluster management utilities, and other necessary data platform components.

Unfortunately, many of these on-premises data lake projects failed to fulfill the promise of data lake computing, thanks to burdensome complexity, slow time to value, and heavy system management efforts. 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 useful analytics and contributed to Hadoop's demise. In fact, some estimates place the failure rate for Hadoop data lake projects as high as 70 percent.

Although many Hadoop-based data lake projects aren't delivering their promised value, organizations of all types still want all the insights from all their data by all their users.

REMEMBER

The original promise of the data lake remains: a way for organizations to collect, store, and analyze all of their data in one place. Today, as cloud computing takes center stage and legacy technology models fade into the background, this new paradigm is revealing its potential. As this book shows, modern cloud technologies allow you to create innovative, cost-effective, and versatile data lakes, or extend existing data lakes created using Hadoop, cloud object stores (computer data storage architecture that manages data as objects), and other technologies.

Reviewing the Requirements

To be truly useful, a data lake must be able to easily store data in native formats, facilitate user-friendly exploration of that data, automate routine data management activities, and support a broad range of analytics use cases. Most of today's data lakes, however, can't effectively organize all of an organization's data.

What's more, the data lakes of today must be filled from a number of data streams, each of which delivers data at a different frequency.

REMEMBER

Without adequate data quality and data governance, even well-constructed data lakes can quickly become *data swamps* — unorganized pools of data that are difficult to use, understand, and share with business users. The greater the quantity and variety of data, the more significant this problem becomes. That makes it harder and harder to derive meaningful insights.

Other common problems include lackluster performance, difficulty managing and scaling the environment, and high license costs for hardware and software. Take a look at Figure 1-2.

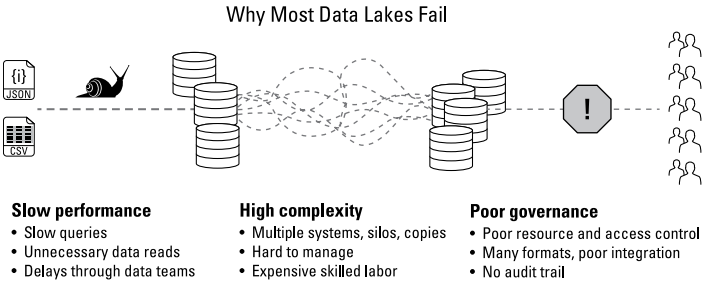

FIGURE 1-2: Slow performance, complexity, and poor governance are among the reasons traditional data lakes often fail.

Introducing the Cloud Data Lake

When the data lake emerged back in 2010, few people anticipated the management complexity, lackluster performance, limited scaling, and weak governance. As a result, Hadoop-based data lakes became data swamps, a place for dumping data. These early data lakes left many organizations struggling to produce the needed insights.

As the cloud computing industry matured, object stores from Amazon, Microsoft, and other vendors introduced interim data lake solutions, such as Amazon Simple Storage Service (S3),

Microsoft Azure Blob, and Google Cloud Storage. Some organizations leveraged these data storage environments to create their own data lakes from scratch. These highly elastic cloud storage solutions allow customers to store unlimited amounts of data in their native formats, against which they can conduct analytics. However, although customers no longer have to manage the hardware stack, as they did with Hadoop, they still have to create, integrate, and manage the software environment. This involves setting up procedures to transform data, along with establishing policies and procedures for identity management, security, data governance, and other essential activities. Finally, customers have to figure out how to obtain high-performance analytics.

Considering the Rise of the Modern Data Lake

As cloud-built versions of the data warehouse emerged in recent years, a third and far better data lake paradigm has arisen that blends today's popular object stores with a cohesive, flexible, high-performance cloud-built data warehouse. These solutions have become the foundation for the modern data lake: a place where structured and semi-structured data can be staged in its raw form — either in the data warehouse itself or in an associated object storage service. Modern data lakes provide a harmonious environment that blends these object storage options to easily store, load, integrate, and analyze data in order to derive the deepest insights to inform data-driven decision-making.

This combination of a cloud analytics layer, the data warehouse, and a cloud-based object store comprises the modern data lake. These data lakes provide near-unlimited capacity and scalability for the storage and computing power you need. A modern data lake dramatically simplifies the effort to derive insights and value from all that data and ultimately produces faster business results.

Explaining Why You Need a Modern Cloud Data Lake

The ability to store unlimited amounts of diverse data makes the cloud particularly well-suited for data lakes. And the entire environment can be operated with familiar SQL tools. Because all storage objects and necessary compute resources are internal to the modern data lake platform, data can be accessed and analytics can be executed quickly and efficiently. This is much different than 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 Who Uses Modern Data Lakes, and Why

Modern data lakes have the potential to 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 geologic 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. They can use that data to gauge customer sentiment, adjust go-to-market strategies, mitigate customer-support problems, and extend targeted offers to customers and prospects.

REMEMBER

Traditional data lakes fail because of their inherent complexity, poor performance, and lack of governance, among other issues. Modern cloud data lakes overcome these challenges thanks to foundational tenets, such as:

- **No silos:** Easily ingest petabytes of structured, semi-structured, and unstructured data into a single repository.
- **Instant elasticity:** Supply 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.
- **Embedded governance:** Present fresh and accurate data to users, with a focus on collaboration, data quality, access control, and metadata (data about data) management.
- **Transactional consistency:** Confidently combine data to enable multi-statement transactions and cross-database joins.
- **Fully managed:** With a software-as-a-service (SaaS) solution, the data platform itself largely manages and handles provisioning, data protection, security, backups, and performance tuning, allowing you to focus on analytic endeavors rather than on managing hardware and software. You just set the modern data lake platform and go.

KIXEYE ANALYZES SEMI-STRUCTURED DATA IN THE CLOUD

CASE STUDY

KIXEYE is a mobile and online gaming company based in San Francisco.

Analytics plays a central role in helping the company meet its revenue goals, allowing KIXEYE to continually experiment with new features, functionalities, and platforms. KIXEYE's analytics team captures event data from gaming activities in JavaScript Object Notation (JSON), which is a language-independent data format, and makes it available for exploration, reporting, and predictive analytics.

Previously, the event data had to be converted to Structured Query Language (SQL, the language typically used by databases). That's what KIXEYE's analytics platform required, and it was a convoluted, multi-step process:

- Funnel JSON event data into an Apache Kafka pipeline, which stores streams of records in categories called *topics.*
- Process the data on a Hadoop cluster.
- Import the results into Hive tables, providing a SQL-like interface for querying data.
- Transform and load data into relational tables for analysis.

In order to simplify this data-transformation and processing cycle, KIXEYE migrated its analytics environment to a cloud-built data lake that natively supports semi-structured data. In this new architecture, JSON event data is output directly from Kafka to an Amazon S3 storage service, where it is loaded into the cloud data lake to make it accessible for processing and analysis — with each business group at KIXEYE able to operate independent virtual warehouses (compute clusters) for transforming, slicing, aggregating, and analyzing. The process is an order of magnitude faster than before, and the company pays by the second for the storage and compute power it needs.

IN THIS CHAPTER

- Creating a data lake solution
- Understanding the difference: Data lake versus data warehouse
- Surviving the data deluge
- Moving data to the cloud
- Spotting trends in data and analytics

Chapter 2
Explaining Why the Modern Data Lake Emerged

Data-driven businesses rely on data from traditional sources, such as enterprise resource planning (ERP), customer relationship management (CRM), and point-of-sale systems. They also pull from more modern data sources such as weblogs, clickstreams, Internet of Things (IoT) devices, and social media networks. Beyond that, they mix in data from third-party sources — from traffic information to weather forecasts to financial market feeds.

This chapter explores how to eliminate the headache associated with loading, integrating, and analyzing such diverse data types. It also spells out how data lakes differ from data warehouses, explains why the cloud data lake is an ideal solution, and reveals what's possible by democratizing analytics across your organization.

Differentiating the Data Warehouse from the Data Lake

Before going any further, it's worth underscoring the differences between a data warehouse and a data lake.

Data warehouses emerged decades ago as a method for organizations to store and organize their data for analytics — that is, to ask questions (queries) about the data to reveal answers, trends, and other insights. Data warehouses also orchestrate data marts for different work groups. Data marts often involve moving a copy of the source data, stored in the data warehouse, to the different data marts.

Data warehouses and their associated data marts handle thousands or even millions of queries a day. Vital queries range from reporting order trends, to uncovering common demographics information about customers, to forecasting probable business trends.

As discussed in Chapter 1, each source of data inside the traditional data warehouse has its own *schema,* or a table row-column definition, that dictates how the data is organized. Therefore, the attributes of the data must be known upfront. Data lakes, by contrast, store newer semi-structured data types in their native format, without requiring a schema to be defined upfront.

The tables used by traditional data warehouses can't easily contain newer, semi-structured data types from such sources as weblogs, clickstreams, mobile applications, or social media networks. These types of data must be transformed into a more structured format before they can be stored in a data warehouse and analyzed.

Traditional data lakes, at a minimum, are capable of storing these mixed data types. That's just the start, though. In order to analyze that data, you need deeply technical data analytics and data science professionals, who are in short supply. If you can hire these experts, they may end up spending an inordinate amount of time deriving usable insights from the data. If you're relying on either a traditional data warehouse or a traditional data lake, you'll rarely gain all insights possible.

TIP

With a modern, cloud-built data lake, you get the power of a data warehouse and the flexibility of the data lake, and you leave the limitations of both systems behind. You also get the unlimited resources of the cloud automatically.

Behind the scenes, the cloud data lake provisions the necessary infrastructure and platform services immediately. These implementations are not only less expensive and easier to scale, but they're also virtually trouble-free. Security, tuning, and software updates are automatic and handled by the cloud vendor.

Staying Afloat in the Data Deluge

Your company is likely gathering data in increasing quantities, and from a more diverse set of sources than ever before. Read on for some leading examples.

Harnessing data from many enterprise applications

In the past, an organization may have relied on a few significant enterprise applications. Now, organizations often rely on dozens or hundreds of enterprise applications, including ubiquitous software-as-a-service (SaaS) solutions.

These business applications generate mountains of valuable data if you can harness it for analysis. Examples include credit card transactions for fraud detection and prevention; social media data for recruiting and talent management; supplier data for manufacturing resource planning and supply chain management; and order-fulfillment data for inventory control, revenue forecasting, and budgeting.

Unifying device-generated data

Petabytes of data also come from people using websites and software applications on their computers, tablets, phones, and smart watches, as well as from many other types of digital and mechanical devices. More devices mean more data.

REMEMBER

Billions of equipment sensors, meters, gauges, and other IoT devices are now connected to the Internet, from thermostats and security systems in the home to industrial sensors on oil rigs and in factories. Utility meters, traffic monitors, parking sensors,

and many other devices gather and transmit data to enhance processes, monitor fluctuating conditions, and streamline maintenance cycles.

Whether you're importing log data, IoT data, or social media data, the volume and complexity of these semi-structured data sources can quickly overwhelm a conventional data warehouse. Bottlenecks in data processing cause analytics jobs to hang, or even crash the system. To streamline these diverse data management activities, you need a unified repository to easily and efficiently store all your data and make it useful. That's the heart of the modern, cloud-built data lake.

Keeping Your Data in the Cloud

REMEMBER

With the majority of data now in the cloud, the natural place to integrate this data is also in the cloud. Astute organizations are now easily creating cloud-built data lakes that can weave these various threads of information into a cohesive fabric. Modern cloud data lakes allow them to capture, store, and facilitate analysis to discover trends and patterns.

You're no longer worrying about whether your data center has the capacity to keep up with the needs of your data warehouse or data lake. In the cloud, you can focus on scaling cost-effectively, without friction, on the order of magnitude necessary to handle massive volumes of varying data.

But remember: Not all cloud systems are created equal. It's not just a matter of taking the same technologies and architectures from inside an on-premises data center and moving them elsewhere. Back when managed cloud services first emerged, the biggest change was shifting management of the technology from the enterprise to a cloud software provider, which was nice, but not enough.

As the cloud matured, it became clear it was also a new development platform — one with unique capabilities and potential. As will become clear in the chapters that follow, the best data lake solutions have been designed first and foremost for the cloud.

REMEMBER

Most on-premises data warehouse and data lake offerings are simply copied to the cloud, or what the industry calls "cloud washed." To take full advantage of what the cloud offers, a solution must be built for the cloud, from the ground up.

WHY MOVE TO A CLOUD DATA LAKE?

Motivating factors for moving to a cloud-built data lake include the business need to:

- Minimize capital expenses for hardware and software.
- Get new analytic solutions to market quickly.
- Eliminate data silos by consolidating multiple data types into a single, unified, and infinitely scalable platform.
- Capture batch and streaming data in a common repository with robust governance, security, and control.
- Simultaneously execute multiple workloads — data loading, analytics, reporting, and data science.
- Establish a robust, fully managed, and extensible environment.

Democratizing Your Analytics

REMEMBER

Executives and professionally trained analysts thrive on data exploration and data-driven decision-making. But to realize the full potential of your data, your organization needs to make analytic activities available to the other 90 percent of business users. Here are four benefits of cloud-built analytics:

- **Data exploration:** This entails discovering trends and patterns, but it's difficult to know in advance precisely what amount of computing resources you'll need to analyze huge data sets. The cloud offers on-demand, elastic scalability ideal for this type of exploratory analysis.
- **Interactive data analysis:** This pursuit provides the answer to a single, random business question, which may lead to other questions. The dynamic elasticity of the cloud gives you the flexibility and adaptability to perform these additional queries without slowing down other workloads.
- **Batch processing:** This refers to scheduling and sending a specific set of queries to the data lake or data warehouse for execution. Batch jobs can be huge, and can be a drain on

performance. With a traditional data lake that doesn't scale, you often must schedule these large batch jobs at off-hour times, when more compute resources are available.

- **Event-driven analytics:** These demand constant data. They incorporate new data to refresh reports and dashboards on a continual basis, so managers can monitor business processes. Ingesting and processing streaming data requires an elastic data lake to handle variations and spikes in data flow.

TIP

A growing trend is to build analytics into cloud business applications, which serve many types of users and the queries (workloads) users run to analyze that data.

CASE STUDY

REPLACING HADOOP

Accordant Media makes advertising investments more successful for marketers by unlocking the value of audience data. Its business model involves analyzing billions of data points about audience profiles, ad impressions, clicks, and bids to help clients determine what works for targeted audience segments.

Speed and accuracy are crucial to optimizing ad visibility and response rates, but load performance was difficult to achieve with Accordant Media's Hadoop environment, especially when ingesting data sets approaching 100 terabytes. Formerly, the analytics team used MapReduce (a program model within the Hadoop framework used to access big data) to evaluate, translate, and load queries into a SQL table and had to manually verify data quality.

Now, Accordant Media no longer needs the resource-intensive and error-prone SQL-to-MapReduce translation activities and subsequent data-duplication processes. Its more nimble cloud-based data warehouse uses standard SQL database comparison methodologies instead, boosting performance and improving accuracy.

By shifting analytic workloads from Hadoop into a modern cloud-built environment, the company has increased its analytic performance a hundredfold, can manage five times the client environments with the same staff, has eliminated error-prone data-translation processes, and has minimized data quality risks.

IN THIS CHAPTER

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

Chapter 3
Reducing Risk, Protecting Data

Your organization's data is incredibly valuable, and this book is all about maximizing that value with the latest technologies for storing data, analyzing it, and gaining useful insights. But if your data is valuable to you, it's also valuable to those who entrust you with their data and to other, malevolent actors. This chapter explores how one of your most valuable assets can also be one of your greatest risks, and what to do about it. It discusses the need to plan carefully and deliberately as you set up your data lake to deliver the best in data quality, security, governance, and legal and regulatory compliance. As recent news indicates, sensitive information can get into the wrong hands or be improperly expunged. That can lead to regulatory penalties, lawsuits, and even jail time. You may also lose your valuable customers.

If you're planning to build or deploy a data lake, it is important to identify several key issues:

- What data should your data lake store?
- How much effort will be required to secure and govern it?
- How will you meet the European Union's General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and other data sovereignty and data protection regulations?

- What data retention policies are important, and how do you manage the data in the lake to meet the requirements of those policies?
- How will you enforce data governance/access controls and ensure business users can access only the data they're authorized to see?
- How will you mitigate data quality or integrity problems that can compromise vital business processes?

Implementing Compliance and Governance

Privacy regulations are increasingly rigorous, and organizations can't ignore them. Leading the way are Europe's GDPR, the United States' Health Insurance Portability and Accountability Act (HIPAA), and the CCPA, which PwC has called "the beginning of America's GDPR."

Data governance ensures data is properly classified, accessed, protected, and used. It also involves establishing strategies and policies to ensure the data lake processing environment complies with necessary regulatory requirements. Such policies also verify data quality and standardization to ensure the data is properly prepared to meet the needs of your organization. For example, data governance policies define access and control of personal identifiable information (PII). The types of information that fall under these specific guidelines include credit card information, Social Security numbers, names, date of birth, and other such data.


TIP

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


REMEMBER

Data governance isn't a technology. Rather, it's an organizational commitment that involves people, processes, and tools. There are five basic steps to formulating a strong data governance practice:

1. Establish a core team of stakeholders and data stewards to create a data governance framework. This begins with an

audit to identify issues with current data management policies and areas needing improvement.

2. Define the problems you're hoping to solve, such as better regulatory compliance, increased data security, and improved data quality. Then determine what you need to change, such as fine-tuning access rights, protecting sensitive data, or consolidating data silos.
3. Assess what tools and skills you will need to execute the data governance program. This may include people with skills in data modeling, data cataloging, data quality, and reporting.
4. Inventory your data to see what you have, how it's classified, where it resides, who can access it, and how it is used.
5. Identify capabilities and gaps. Then figure out how to fill those gaps by hiring in-house specialists or by using partner tools and services.

REMEMBER

A data lake achieves effective governance by following proven data management principles, including adding context to metadata (data about the data) to make it easier to track where data is coming from, who touched that data, and how various data sets relate to one another; ensuring quality data is delivered across business processes; and providing a means to catalog enterprise data.

Ensuring Data Quality

Like regulatory compliance, data security hinges on traceability. You must know where your data comes from, where it is, who has access to it, how it's used, and how to delete it when required. Data governance also involves oversight to ensure 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.

TIP

Ideally, you'll assign these tasks 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 can work with IT professionals and data scientists to establish data quality rules and processes.

Incorporating Protection, Availability, and Data Retention

Cloud infrastructure can fail, accidental data deletions can occur, other human error can happen, and bad actors can attack — resulting in data loss, data inconsistencies, and data corruption. That's why a cloud data lake must incorporate redundant processes and procedures to keep your data available and protected. Regulatory compliance and certification requirements may also dictate that data is retained for a certain minimal length of time, which can be years.

REMEMBER

All cloud data lakes should protect data and ensure business continuity by performing periodic backups. If a particular storage device fails, the analytic operations and applications that need that data can automatically switch to a redundant copy of that data on another device. Data retention requirements call for maintaining copies of all your data.

That's not all. A complete data-protection strategy should go beyond merely duplicating data within the same physical region or zone of a cloud compute and storage provider. It's important to replicate that data among multiple geographically dispersed locations to offer the best possible data protection.

This is because the much-vaunted "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.

REMEMBER

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.

Protecting Your Data with End-to-End Security

All aspects of a data lake — its architecture, implementation, and operation — must center on protecting your data in transit and at rest. This should be part of a multilayered security strategy that considers both current and new security threats.

Your protection strategy should address external interfaces, access control, data storage, and physical infrastructure, in conjunction with comprehensive monitoring, alerts, and cyber security practices. Read on to find out more about the primary aspects of data lake security.

Encrypting everywhere


REMEMBER

Encrypting data, which means applying an encryption algorithm to translate the clear text into cipher text, is a fundamental aspect of security. Data should be encrypted when it is stored on disk, when it is 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's 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 the key

Once you encrypt your data, you'll decrypt it with an encryption key (a random string of bits generated specifically to scramble and unscramble data). In order to fully protect the data, you also have to protect the key that decodes your data.


REMEMBER

The best data lakes employ AES 256-bit encryption with a hierarchical key model rooted in a dedicated hardware security module. This method encrypts the encryption keys and instigates 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


REMEMBER

Security updates should be applied automatically to all pertinent software components of your modern cloud data lake solution as soon as those updates are available. If you use a cloud provider, that vendor should perform periodic security testing (also known as *penetration testing*) to proactively check for security flaws.

As an added protection, file integrity monitoring (FIM) tools can ensure that 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


REMEMBER

For authentication, make sure your connections to the cloud provider leverage standard security technologies such as Transport Layer Security (TLS) 1.2 and IP whitelisting. (A *whitelist* is a list of 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 as well as existing user roles. Regardless, multifactor identification (MFA) should be required to prevent users from being able to log in with stolen credentials. With MFA, users are challenged with a secondary verification request, such as a one-time security code sent to a mobile phone.

Once a user has authenticated, it's important to enforce authorization to specific data based on each user's "need to know." A modern data lake must support multilevel, *role-based access control (RBAC)* functionality so each user requesting access to the data lake is authorized to access only data that he or she is explicitly permitted to see. Discretionary and role-based access control should be applied to all database objects including tables, schemas, and any virtual extensions to the data lake. As an added restriction, secure views can be used to further restrict access — for example, to prevent access to highly sensitive information that most users won't need.

Certifying compliance and attestations

Data breaches can cost millions of dollars to remedy and permanently damage relationships with customers. 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 they 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 your cloud provider also complies with all applicable government and industry regulations. Depending on your business, this could include PCI, HIPAA/Health Information Trust Alliance (HITRUST), and FedRAMP certifications. Ask your providers to supply attestation reports to verify they adequately monitor and respond to threats and security incidents and have sufficient incident response procedures in place. Make sure they provide a copy of the entire report for each pertinent standard, not just the cover letters.

Isolating your data

If your data lake runs in a multi-tenant cloud environment, you may want it isolated from all other data lakes. If this added protection is important to you, make sure your cloud vendor offers this premium service.

Isolation should extend to the virtual machine layer. The vender should isolate each customer's data storage environment from every other customer's storage environment, with independent directories encrypted using customer-specific keys.

Work only with cloud providers that can demonstrate they uphold industry-sanctioned, end-to-end security practices (see Figure 3-1). Security mechanisms should be built into the foundation of the cloud-built data lake-as-a-service. You shouldn't have to do anything extra to secure your data.

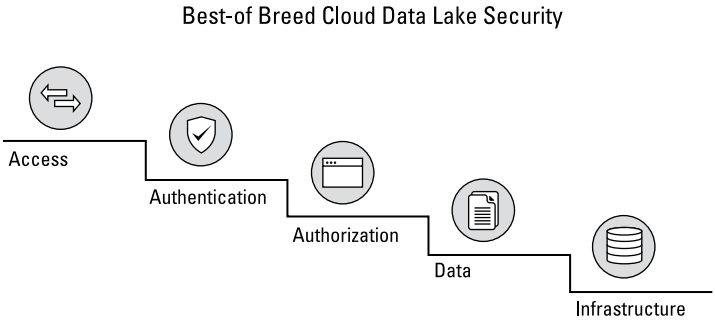

FIGURE 3-1: A complete security strategy considers both data and users.

Facing Facts about Data Security

Effective security is complex and costly to implement, and cybersecurity professionals are hard to come by. Cloud-built data lakes shift the responsibility for data center security to the SaaS cloud vendor. A properly architected and secured cloud data lake can be more secure than your on-premises data center.

But beware. Security capabilities vary widely among vendors. The most basic cloud data lakes provide only rudimentary security capabilities, leaving things such as encryption, access control, and security monitoring to the customer.

COMPLYING WITH HIPAA GUIDELINES

CASE STUDY

Amino helps people find the best possible health care by sharing detailed insights about providers, estimates, and costs. The company deals with sensitive information and must adhere to HIPAA guidelines governing the storage, processing, and exchange of patient data.

Amino maintains more than 1 petabyte of data, including information about 15 million people, 900,000 providers, and 5 billion patient-doctor interactions. Previously, Amino used Apache Hadoop for batch processing, in conjunction with MySQL and Postgres for interactive analytics. To boost security and improve its analytic capabilities, the IT team consolidated a its Hadoop/Hive cluster into a cloud-built data lake.

A batch job that took seven days to execute with Amino's Hadoop/Hive cluster takes less than an hour in the cloud, and the company's new data lake secures all data, at rest or in motion, to comply with government mandates.

IN THIS CHAPTER

- Beginning with the right architecture
- Supporting all your data requirements
- Scaling storage capacity, with no limitations
- Accommodating users and workloads
- Simplifying access to applications and data

Chapter 4
Strategies for Modernizing a Data Lake

Some of today's most valuable data doesn't come with a predefined structure, and that's where data lakes shine. In this chapter, you learn how you can continuously load raw data into a modern data lake, without parsing or transforming it first, and immediately query it via familiar SQL tools.

This chapter explains how you can ingest semi-structured and unstructured data and then analyze it immediately, accommodating many types of users and workloads. It shows how external tables and *materialized views* (database objects that contain the results of a query) enable you to efficiently query data from an object store. And it shows how the separation of compute and storage helps you to cost-effectively store massive volumes of data and securely share it with your ecosystem of business partners via your data lake.

Beginning with the Right Architecture

First, a bit of review. Some vendors originally offered cloud data lakes that were architected for on-premises environments, such as Hadoop. These traditional architectures were created long before the cloud emerged as a viable option, and because they don't fully tap into the magic of the cloud, these data lakes often failed to produce their anticipated results.

In response, organizations began creating their data lakes in cloud-based object stores, accessible via SQL abstraction layers such as Presto, Impala, and Hive. Progress? Not entirely. These cobbled-together data lakes soon became data swamps, requiring custom integration and expert administration, and the result was poor analytic performance.

Today's more versatile and modern data lakes often take the form of a cloud-based analytics layer that optimizes query performance against data stored in a data warehouse or an external object store for deeper and more efficient analytics.

For example, a cloud-built data warehouse is a fine place to store all your data. Or, you can store it in a secure object store, such as Amazon S3, which is also accessible from your cloud analytics layer.

However, to do this well you need specialized technology in your analytics layer such as materialized views, which give you exceptional performance on the external tables (read-only tables that can be used for query and join operations) in an object store — often comparable to what you will see with data that has been ingested directly into your data lake.

This versatile architecture enables seamless, high-performance analytics and governance, even when the data arises from more than one location. You don't have to transform your data to fit into a set of predefined tables, and you can immediately analyze raw data types, thanks to versatile technologies known as *schema-on-read*. You can ingest unstructured and semi-structured data, but you don't have to do any pre-transformations, as is required with a structured data warehouse. Data transformation happens automatically inside the data lake once the data lands there.

Collecting and Integrating a Range of Data Types

A complete data lake strategy should accommodate all types of data, including JSON, tables, CSV files, and Optimized Row Columnar (ORC) and Parquet data stores. If you're also storing and analyzing unstructured data, you may need a separate repository, such as an external object store (that's a data storage architecture that manages data as objects rather than as file hierarchies or data blocks). A separate repository may also be necessary if you have specialized needs for data sovereignty and data retention, or if you must comply with certain industry regulations that govern where your data is stored.

REMEMBER

Whether your data is stored in one location or multiple locations, having an integrated cloud analytics layer reduces risk and makes life simpler. You don't have to move data among multiple physical data marts (smaller but static versions of your data lake), and you won't have multiple potential points of failure. The entire environment can be queried as a *single source of truth* via SQL and other familiar tools.

Continuously Loading Data

Data-driven organizations need real-time analytic systems that can continuously ingest data into cloud storage environments. For example, analysts need up-to-date data to observe trends and identify opportunities. Data scientists require current data to develop machine learning models. Executives need up-to-the-minute data to guide their organizations.

Data pipeline tools can migrate on-premises application data into a cloud data lake. Bulk-load processes work best for initial transfers, especially if you have many terabytes of data you wish to move from storage devices within your enterprise. After that, you'll most likely want to capture incremental changes. Increasingly, real-time data feeds and streaming applications with low latency are becoming the industry norm in many data streaming architectures.

BLACKBOARD MOVES TO HEAD OF CLASS WITH DATA PIPELINE SERVICE

CASE STUDY

Blackboard Inc., an education technology company, deployed a cloud-built data lake with learning analytics to help students succeed. For example, Blackboard Predict improves student retention by proactively identifying at-risk students and suggesting how faculty and staff members can help turn the situation around. Blackboard's data lake also helps the organization to analyze how people use the system, and to continuously improve its products.

To ensure a continuous stream of student data for these analytic endeavors, Blackboard established a data pipeline into its data lake, which loads data from more than 1,000 educational sites. This sophisticated data-ingestion pipeline orchestrates multiple streams and tasks and automatically updates hundreds of thousands of tables.

Having an economical data pipeline has significantly reduced Blackboard's IT infrastructure and management costs. Built-in data quality procedures have made data corruption a thing of the past, while the infinite scalability of the cloud has eliminated resource bottlenecks.

TIP

Make sure your data pipeline can move data continuously as well as in batch mode. It also must handle the complex transformations required to rationalize different data types without reducing the performance of production workloads.

REMEMBER

A continuous data pipeline automatically and asynchronously detects new data as it arrives in your cloud storage environment and then continuously loads it into the data lake (see Figure 4-1).

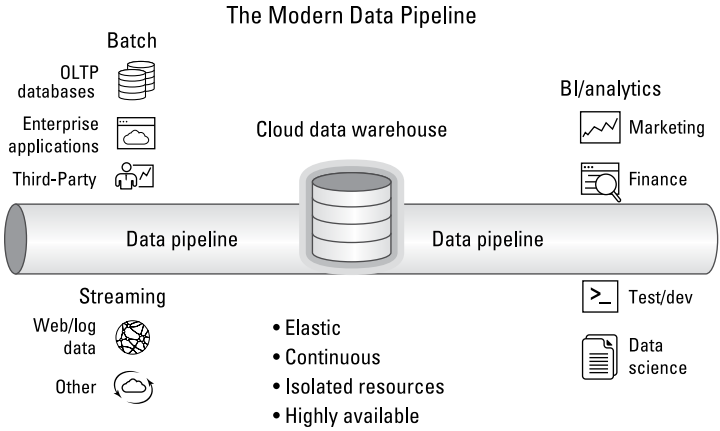

FIGURE 4-1: Modern data pipelines ingest all types of data at scale, across clouds, and query it.

Enabling Secure Data Sharing

There are numerous sources of data, both inside and outside an organization, and many businesses enhance their operations by tapping into third-party data repositories, services, and streams for even deeper insights. How does that data get from one place to another? Traditional data sharing methods, such as FTP, APIs, and email, require you to copy the shared data and send it to your data consumers.

Unfortunately, these methods are cumbersome, costly, and risky. 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.

REMEMBER

Modern data sharing technologies have emerged, enabling organizations to easily share slices of their data, and receive shared data, in a secure and governed way. This simplifies data sharing within their organizations and within an ecosystem of business partners. You're able to gain deeper insights, streamline operations, better serve customers, and discover new market opportunities.

These more robust data sharing methods allow you to share live data without moving it from place to place (see Figure 4-2). This method requires a multi-tenant, metadata-driven architecture,

allowing authorized members of the ecosystem to tap into live, read-only versions of data within a data lake. Ready-to-use data is immediately available, all the time.

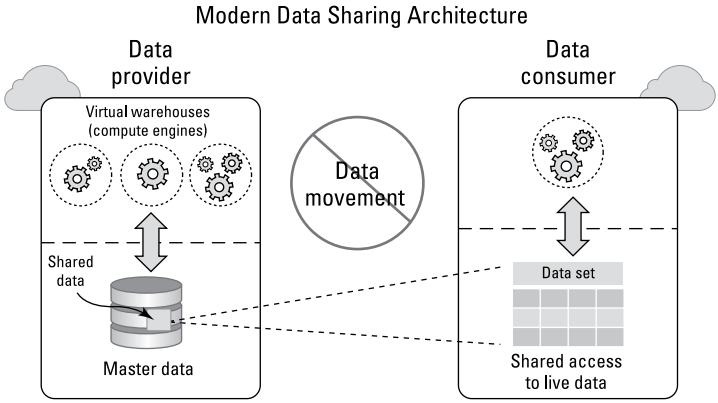

FIGURE 4-2: A modern data sharing architecture enables access to live data between a data provider and a data consumer, without moving data.

Using these methods, you can share data with vendors, supply chain partners, logistics partners, customers, and many other constituents. You can also set up data-sharing services that turn your data lake into a profit center. For more information, please read the ebook *Data Sharing For Dummies*.

REMEMBER

A multi-tenant cloud-built data lake enables organizations to share live data and receive shared data from diverse sources without having to move that data. And there's no contention or competition for resources.

Customizing Workloads for Optimal Performance

REMEMBER

A modern cloud data lake can deliver all the resources you need, with the instant elasticity to scale with demand. You no longer have to over-provision resources to meet peak demands, and cloud storage typically costs only a fraction of what you've probably been paying for traditional systems. You can meet user demands for data volumes, velocity, and variety on a scale unimaginable only a few years ago.

Enabling a high-performance SQL layer

What do you do if a workgroup starts a new project, and instead of querying data in one-week or one-month increments, analysts suddenly start querying data sets that include a quarter's worth of data, or more? For example, a supply chain analyst who normally evaluates day-to-day performance might want to suddenly access a rolling set of data for an entire month or quarter. What you need is a high-performance SQL layer for interacting with the data.

Maintaining workload isolation

Many users will likely be accessing your data lake at the same time, which can consume huge amounts of resources. You don't want ad hoc data-exploration activities slowing down important analyses. That means your data lake must isolate workloads and allocate resources to the jobs that truly matter.

If you have a regular event that requires a burst of compute resources, such as the periodic training of machine learning models or a large quarterly accounting close process, ensure your data lake architecture enables workload isolation.

Make your cloud vendors give you the necessary flexibility. You should be able to configure these resources to "go to sleep" automatically after a predetermined period of inactivity. This ensures you always have resources but you don't have to pay for usage when the resources aren't needed.

Interacting with data in an object store

Your data lake should allow you to query data housed in Amazon S3, Microsoft Azure Blob Storage, or Google Cloud Storage *where it resides*. You can maintain the data lake as a single source of truth, even for multi-cloud environments. Having a single source of truth eliminates the time-consuming task of keeping multiple data repositories in sync.

TIP

Make sure your SQL abstraction layer has a simple way to connect to the data in its raw format, so that when you want to interact with it you don't have to move it from place to place. This involves pointers to external tables and, ideally, materialized views (see Figure 4-3).

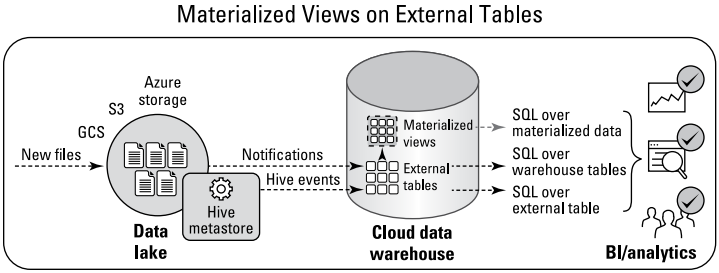

FIGURE 4-3: Materialized views can be created to significantly improve the query performance on external files.

Materialized views precalculate metadata and statistics about data in external files, thus speeding up the queries that are run on them. They let you *materialize* all of the data in your data lake, or just the part of the data lake you query most frequently. The views are automatically refreshed in the background, eliminating the need to build a data extract, transform, and load (ETL) layer or orchestration pipeline.

Here are some of the things you can achieve with external tables:

- Query data directly from Amazon S3, Microsoft Azure Blob Storage, or Google Cloud Platform and ingest it natively into your data warehouse.
- Maintain your data lake as a single source of truth, eliminating the need to copy and transfer data.
- Achieve fast analytics on data from an external source, wherever it resides.

REMEMBER

External tables store file-level metadata about the data files such as the file path, a version identifier, and partitioning information. This enables querying data stored in files in a data lake as if it were inside a single database.

Resizing compute clusters

TIP

A flexible, cloud-built data lake can automatically scale concurrency by transparently creating a new cluster and then automatically balancing the load. When the load subsides and the queries catch up, the second cluster automatically spins down. Look for a cloud provider that allows you to dynamically expand independent resources to handle sudden concurrency issues. You should

be able to specify the number of clusters you would like or let it happen automatically.

If you have a four-node cluster and you're expecting a temporary surge in data, you can easily add more compute power. For example, you might resize the cluster to 16 or 32 nodes and then, once the team is done with its analytics, you can scale the cluster back to its original configuration. To ensure uninterrupted service, make sure your cloud vendor allows you to scale the service while the cluster is running.

Creating a User-Friendly Environment with Metadata

Unfortunately, much of the data in first-generation data lakes isn't usable because it hasn't been cataloged. There isn't any metadata (data about the data), which means there are no designations or identifiers on the data. With no method for prioritizing data, it's difficult for users to target and retrieve the right information in a timely fashion.

REMEMBER

Ultimately, a data lake should democratize access to the data. People of all skill levels — including line-of-business managers, financial analysts, executives, data scientists, and data engineering specialists — should have easy and ready access to the system to perform the analytics they need.

Of course, that doesn't mean users have free reign. Security, access control, and governance are essential, as discussed in Chapter 3. Role-based access mechanisms define precisely which data people are allowed to see. Personally identifiable information (PII) can be housed within protected databases, and storage resources can be physically or virtually isolated to distinguish workloads from each other.

To uphold all pertinent compliance requirements, you should continually monitor and audit user access. The cloud-built data lake should have an intuitive, graphical user interface for IT managers and data engineers that provides global access to the data, metadata, data processing, and auditing functions.

FIVE CHARACTERISTICS OF A DATA LAKE BUILT FOR THE CLOUD

A cloud-optimized architecture will simplify your data lake. For maximum flexibility and control, make sure that your cloud data lake service has the following characteristics:

- A multi-cluster, shared-data architecture
- Independent scaling of compute and storage resources
- The ability to add users without affecting performance
- Tools to load and query data simultaneously, without degrading performance
- A robust metadata service that is fundamental to the object storage environment

For more information, please read *Cloud Data Warehousing For Dummies*.

IN THIS CHAPTER

- Moving from yesterday to today (and tomorrow)
- Maximizing scalability
- Minimizing costs
- Boosting productivity
- Simplifying the environment

Chapter 5
Assessing the Benefits of a Modern Cloud Data Lake

This chapter offers examples of modern data lakes in production settings, along with suggestions about what you can do to achieve similar results at your organization.

Getting Here from There

The first two generations of data lakes were constructed either by using open source technologies, such as Apache Hadoop, or by customizing an object store from a cloud storage provider such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform. These earlier approaches created multiple issues:

- Getting all forms of raw data into the data lake was fairly straightforward, but getting insights from that data was nearly impossible. Only a few technical elites could do so.
- IT departments had to over-plan and therefore overpay for enough compute to accommodate infrequent spikes in activity. Eventually, activity surpassed physical compute limits, causing queries and other workloads to stall.

- The technical professionals who knew how to administer and manage these installations were in short supply.
- Security and governance were an afterthought, leaving data at risk and organizations behind in compliance.

These limitations inspired modern, cloud-built solutions to collect large amounts of varying data types, in their raw forms, and store it all in a cohesive data lake that can draw from and synthesize a variety of versatile data storage repositories.

Increasing Scalability Options

On-premises data lakes are expensive because you have to purchase and maintain hardware and software infrastructure, and the hardware limits independent scalability. In the case of Hadoop, the basic data-access architecture requires you to store data on the compute nodes in the cluster. This requirement forces you to size these clusters to accommodate peak processing loads, but much of that capacity goes unused most of the time. These massive clusters create significant processing overhead, which constrains performance.

REMEMBER

Contrast that with a data lake that leverages a cloud data warehouse, or that utilizes a flexible object storage service such as Amazon S3, Microsoft Azure Blob Storage, or Google Cloud Storage. Now you have scalability, resiliency, and throughput that are much better than what you could achieve in an on-premises data center, and at a fraction of the cost. Storage and compute resources are fully independent yet logically integrated, and designed to scale automatically and independently from each other.

Modern data lakes, designed and built for the cloud, enable native data loading and analytics on structured, semi-structured, and unstructured data from one central location, with near-infinite scalability of storage and compute resources. They support concurrency for an unlimited number of users and workloads, and can easily scale up and down to handle fluctuations in usage, without adversely impacting performance. The bottom line: You pay only for what you use.

TIP

Here are some guidelines to consider to ensure scalability:

- Choose a platform that gives you easy, elastic, and independent scaling of compute and storage.

- Build your data lake to enable multiple, independently scalable compute clusters that share a single copy of the data but eliminate contention between workloads.
- Take advantage of auto-scaling when concurrency surges.

ZEETO GAINS DEEPER INSIGHTS, LIMITLESS SCALABILITY

CASE STUDY

Zeeto's real-time bidding platform surveys website visitors and uses the insights to help customers bid on highly targeted advertising. But Zeeto needed an architecture that could stream and query platform "events," which amounted to billions of rows of data each year. Zeeto also wanted to model data, test a variety of options, and iterate quickly.

Chief Technology Officer Matt Ferguson decided to deploy a serverless data ingestion service that could automatically and continuously load raw data in five-minute event streams. Previously, it took as long as five days for Zeeto's engineering team to conceive, design, and deliver reports to the management team. Now the team can produce these comprehensive reports in a few hours, and the continuous loading of streaming event data delivers near-real-time insights.

Account managers use the data to coach advertisers on their bidding strategies and click-through rates, while the CEO can access reports that update every 15 minutes, spanning all publishers and advertisers. The data pipeline eliminates the need for scripts and scheduling tools. It can also integrate changes since the last load — a technology known as *change data capture*.

Previously, as Zeeto added customers, the technology team had to purge critical business performance data because customer demand was outpacing the capacity of its rigid infrastructure. Now, Zeeto's "zero-management" infrastructure gives the company more time for value-add activities, and having unlimited scalability eliminates obstacles to growth.

According to Ferguson, "Being able to iterate and produce results in hours instead of days, and days instead of weeks, has been transformational."

Reducing Deployment and Management Costs

Traditional data lake projects were often criticized for long timelines and runaway costs. Consider what it would take to construct a large server cluster. Don't forget to add in the licensing fees for hardware and software plus the time and expertise required to set up, manage, deploy, tune, secure, and back up the environment. You're talking millions of dollars. Data lakes built on an object store were less expensive to get started, but required customer integration and tedious administration.

Moving to the cloud for your data lake will save you the significant expense of buying, maintaining, and securing an on-premises system. The cloud vendor should automatically handle all the technical chores associated with server provisioning, data governance, data protection, data security, and performance tuning. That frees up your analytics teams to focus on gaining the most value from the organization's data.

TIP

Here are some thoughts for keeping costs under control:

- Pay for usage by the second, not by the minute or month or a timeframe 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 what you require day-to-day.

REMEMBER

On-premises data lake solutions require large, up-front capital investments treated as assets and depreciated over time. Cloud solutions represent much smaller up-front investments, which are treated as an operating expense that's deducted monthly against revenues. Work with a modern cloud data lake vendor that offers usage-based pricing, which lets you pay as you go for resources actually consumed. You shouldn't agree to any multi-year licenses or service contracts, although you may get a better rate by committing to a minimum volume of usage.

Gaining Insights from All Types of Data

With all your diverse data sources and metadata located and integrated in a single system, your users can more easily obtain data-driven insights. And they can often do so without asking the technical staff for help. You'll end up with a single, integrated system for easily storing and analyzing vast amounts of data.

TIP

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

It also should enable native data loading and analytics on these mixed data formats with complete transactional integrity, and store these diverse data types in their native form, without creating new data silos.

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.
- Use pipeline tools that allow for native data loading with transactional integrity.

IAC PUBLISHING LABS: AN EFFICIENT WAY TO PROCESS JSON DATA

CASE STUDY

IAC Publishing Labs is one of the largest collections of premium publishers on the web. Its web properties include Ask.com, shopping.net, and consumersearch.com, which provide digital experiences to millions of people throughout the world.

(continued)

(continued)

As part of its work managing more than 300 million events and 50 million keywords for bidding and monetization, IAC's business intelligence (BI) team imports more than 1.5 terabytes of web log data every day, in JSON format. Previously the team stored six months' worth of this data in two on-premises systems: a 36-node massively parallel processing (MPP) warehouse and an even larger Hadoop cluster. However, the load procedures required complex preprocessing, using scripting technologies such as C, Python, and Perl to coalesce and manage the JSON data set.

IAC decided to retire its large on-premises systems in favor of a cloud-built data lake for processing data. Now, instead of loading data every five hours, the BI team can load data every 15 seconds. And instead of processing data once per day for three to four hours, the team can process data every hour, in as little as five minutes.

IAC Publishing Labs now has one current, consistent version of the truth in an AWS cloud, and its BI team can use SQL tools to directly query the JSON data from the data lake that sits on top of AWS. The team uses SQL in the data lake and processes data by spinning up data marts as required. Thirty or more analysts can concurrently query the environment, with no performance issues. The team can also load the data 24 hours a day, 365 days a year, without causing contention across time zones or impeding the work of other groups across the globe.

Thanks to this fast and flexible data lake, BI activities have shifted from a cost center to a value center.

Boosting Productivity for Business and IT

First-generation data lakes, based on Hadoop, required administrators to constantly attend to capacity planning, resource allocation, performance optimization, and other complex tasks. That's a drain on scarce IT talent. Although a cloud object store eliminates the security and hardware management overhead, this cobbled-together architecture can slow things down and require lots of manual tuning for analytic performance.

With a modern data lake built for the cloud, security, tuning, and performance optimizations are built into the managed

service as one cohesive package. That means your highly trained software engineers won't have to focus on mundane maintenance tasks. They can spend their valuable time creating innovative applications.

TIP

Offload routine technology chores to a cloud vendor so your IT pros can shift their attention away from time-consuming, low-level management tasks. Let them focus on helping the business derive insights from enterprise data.

Simplifying the Environment

The bold promise of the data lake has not wavered since the early implementations: Organizations still want a standards-based solution for high-performance querying and reporting on data at scale. However, traditional on-premises data lakes have not kept up with exploding analytic demands.

Today's data comes from a mix of sources including relational and NoSQL databases, IoT devices, and data generated by SaaS and enterprise applications. These data sources have different formats, models, and structures, which often are stored in different types of platforms.

REMEMBER

Bringing all this data together remains a challenge for legacy data warehouses and early data lake platforms, but a modern data lake can eliminate data silos and consolidate multiple types of information. That includes the object storage options commonly found in modern cloud infrastructure platforms from Amazon, Microsoft, Google, and other vendors.

Examining the benefits of object storage

Today's cloud architectures can optionally use object storage to handle various data types. You can store virtually any kind of data in these object storage mechanisms, without bothering with the extensive planning, software customization, programming, and server provisioning required in Hadoop environments. And you can use SQL and other familiar tools to explore, visualize, and analyze the data.

It can take a lot of work to create a data lake that works end-to-end, especially if you're using open source technologies such as Hadoop, or simply adding a SQL abstraction layer to an object store. The architecture can be complex, and there are several moving pieces to consider.


TIP

You don't have to integrate lots of open source software packages to obtain robust capabilities with a pre-integrated cloud data lake based on readily available and relatively inexpensive object storage mechanisms. You can sign up for the service and use it immediately.

Offering more advice


TIP

Keep these capabilities in mind as you select the optimum cloud-built solution for your data lake:

- **Fast analytics:** Look for performance optimizations and performance-boosting techniques such as data pruning, which ensures only the data necessary to complete a query is scanned.
- **Centralized data:** Having a single, integrated service minimizes complexity, reduces costs, and eliminates multiple points of failure.
- **Familiar tools:** The entire environment should be accessible via standard SQL tools and commands. In addition, support for NoSQL enables you to easily store and analyze many newer forms of data, such as IoT data generated from machines, and social media data from social networks.
- **Cohesive architecture:** All storage objects should be internal to the cloud data lake (first-class storage objects) rather than stored in an external data bucket (second-class storage objects). That will minimize data movement and maximize performance.

IN THIS CHAPTER

- Planning for a speedy deployment
- Leveraging cost-effective cloud services
- Characteristics of a successful implementation

Chapter 6
Six Steps for Planning Your Cloud Data Lake

Building a modern data lake requires technology to easily store data in raw form, provide immediate exploration of that data, refine it in a consistent and managed way, and make it easy to support a broad range of operational analytics. Follow these steps to get started:

1. **Identify the data mix.** Identify the exact data sources, types, and locations of the data you plan to load into your data lake. Next, consider how extensively the data will be used. Do you plan to share data within your ecosystem to enrich analytics? If so, how are you sharing that data now? Identify archaic data sharing methods such as FTP and email, and consider how you can replace them with a modern data-sharing architecture.
2. **Consider the repository.** A data lake uses a single repository to efficiently store all your data. The repository can support a range of use cases including data archival; data integration across a variety of data sources; ETL offloading from legacy data warehouses; and complex data processing across batch, streaming, and machine-learning workloads. Will you stage data from an existing data warehouse or data store? Will all

your data land in a cloud storage bucket such as Amazon S3, Microsoft Azure Blob, or Google Cloud Platform? If so, will you have to integrate the data lake with the storage bucket? If not, the cloud data lake can serve as your central data repository.

3. **Define the data pipeline.** As you plan what method you will use to ingest data, consider initial data loads as well as incremental updates. Do you have historical data sets you would like to migrate? If so, you will likely want to set up a one-time transfer of this historical information to the data lake. Will you continually refresh that data as new transactions occur? If so, you'll want to establish a continuous stream of data moving through the pipeline.
4. **Check pertinent use cases.** Do you want to replace or augment an existing Hadoop data lake? 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 want to work with a provider to configure a data lake using pre-integrated technologies? If you have a large investment in a traditional data lake or data warehouse solution, then you'll likely want to complement and extend that investment with newer technologies.
5. **Apply governance and security.** Building a successful data lake requires good stewardship to maintain data quality, uphold regulatory guidelines, keep data secure, and make sure the lake doesn't turn into a swamp. You need to decide who is responsible for governing and securing that data, both for your initial data loads and on a continuous basis as new data is ingested (see Chapter 3 for more details).
6. **Keep it simple.** Once your implementation is complete, you should not have to install, configure, update, or maintain hardware and software. Backups, performance tuning, security updates, and other management requirements should be part of the basic service (see Chapter 4 for more details).

Get insights fast from all your data by all your users with a cloud data lake

The concept of first-generation data lakes aimed to create a single repository for storing, integrating, and analyzing all of an organization's data. As years passed, reality set in and most data lake initiatives failed. Today, organizations still want to achieve that aim: a cloud data lake that is simple yet powerful, flexible and affordable, and provides unparalleled business value. Read this book to learn how the modern cloud data lake provides all of this and more to enable data-driven decision-making across your organization.

Inside...

- Why the cloud data lake emerged
- How to evaluate different data lakes
- How to easily enable a modern data lake with the modern data platform
- How to maximize scale and lower costs
- Why data security, governance, and sovereignty are data lake essentials
- How a data lake enables data sharing

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: 978-1-119-66624-0
Not For Resale

for dummies®
A Wiley Brand

Also available as an e-book

WILEY END USER LICENSE AGREEMENT

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