LEARNING MADE EASY

3rd Snowflake Special Edition

Cloud Data Warehousing

for dummies®
A Wiley Brand

Unlock your data's potential

Manage data easily, securely, and efficiently

Understand modern data platforms

Brought to you by

snowflake®

David Baum

About Snowflake

Snowflake enables every organization to mobilize their data with Snowflake's Data Cloud. Customers use the Data Cloud to unite siloed data, discover and securely share data, power data applications, and execute diverse AI/ML and analytic workloads. Wherever data or users live, Snowflake delivers a single data experience that spans multiple clouds and geographies. Thousands of customers across many industries, including 639 of the 2023 Forbes Global 2000 (G2K) as of July 31, 2023, use Snowflake Data Cloud to power their businesses. Learn more at **snowflake.com.**

Cloud Data Warehousing

3rd Snowflake Special Edition

by David Baum

for dummies®
A Wiley Brand

Cloud Data Warehousing For Dummies®, 3rd Snowflake Special Edition

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

ISBN 978-1-394-21162-3 (pbk); ISBN 978-1-394-21163-0 (ebk)

Publisher's Acknowledgments

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

Development Editor: Nicole Sholly
Project Manager: Jen Bingham
Acquisitions Editor: Traci Martin
Editorial Manager: Rev Mengle
Sales Manager Molly Daugherty
Content Refinement Specialist: Tamilmani Varadharaj

Table of Contents

Introduction

Data is infiltrating all types of business processes and reshaping the way companies operate. Regardless of your industry or market, the ability to manage data easily, securely, and efficiently has become vital for success.

For instance, in the realm of marketing, data is animating customer segmentation and targeted advertising, allowing businesses to craft personalized marketing campaigns based on the moment-to-moment activities of consumers. In transportation, real-time data enables travelers to optimize routes, and that same data can be aggregated to reduce traffic congestion and improve roadway efficiency. These examples highlight the immense potential of data and the transformative impact it will continue to have for years to come.

Forward-thinking organizations rely on powerful, easy-to-use, and out-of-the-box cloud data warehouses to put their data to work. The best cloud data warehouses are built on a *cloud data platform* — a unified, global solution not only for data warehousing but also for data lakes, data engineering, AI/ML, and data application development. By concurrently powering these and other workloads, a cloud data platform enables everyone in the organization to deliver valuable experiences with their data.

Delivered as an affordable, usage-based service, a cloud data platform can help your business users become more efficient and allows your IT team to break free from mundane data administration tasks. It provides consistent functionality across multiple regions and clouds with instant and near-infinite scalability. Multiple business units can securely share governed data without the complications of duplicating or copying data, as well as extend access to partners, customers, and other constituents — either directly or through a data marketplace.

About This Book

Welcome to the third edition of *Cloud Data Warehousing For Dummies* where you discover how your organization can tap into and transform the power of massive amounts of data into valuable business intelligence.

In this book, you learn how to create an innovative, cost-effective, and versatile cloud data platform that powers not only your data warehouse but also many other data workloads. Additionally, you learn how to extend an existing data warehouse to take advantage of the latest cloud technologies.

Icons Used in This Book

Throughout this book, the following icons highlight tips, important points to remember, and more.


TIP

Tips alert you to easier ways of performing a task or better ways to use cloud data warehousing in your organization.


REMEMBER

This icon highlights concepts worth remembering as you immerse yourself in the understanding and application of cloud data warehousing.


TECHNICAL STUFF

The jargon beneath the jargon, explained.


CASE STUDY

The case studies in this book reveal how organizations applied cloud data warehousing to save money and significantly improve the speed and performance of their data analytics.

Beyond the Book

If you like what you read in this book, visit `www.snowflake.com`, where you can find out more about the company's cloud data platform offering, sign up for a free Snowflake trial account.

IN THIS CHAPTER

- Understanding data warehouses, data lakes, and cloud data platforms
- Diving into the modern cloud data warehouse's history
- Exploring trends in data and analytics
- Keeping up with the shifting demands of data

Chapter 1
Introducing Cloud Data Warehousing

A traditional data warehouse required purchasing, installing, and configuring the necessary hardware, software, and infrastructure to store and analyze data. *Cloud data warehousing* emerged as an efficient, cost-effective way for organizations to scale analytics without those upfront costs. And, when a cloud data warehouse lives on a well-architected, modern cloud data platform, it not only enables organizations to accelerate analytics but also broadens data management capabilities to include other architectures, like a data lake, and can securely and efficiently run other workloads. To help you understand data warehouses, data lakes, and the modern cloud data platform, this chapter defines each, and briefly shows how the modern cloud data platform came into being. The chapter wraps up with a quick look at trends in data processing and how those trends require the ability to shift and meet new data demands.

Defining the Data Warehouse

Initially, *data warehouses* were simply relational databases that stored and queried large volumes of structured data. Today, cloud-built and hybrid cloud data warehouses can also incorporate semi-structured data, such as JavaScript Object Notation (JSON) weblogs, and unstructured data, such as images and audio conversations. This has allowed modern data warehouses to expand beyond mere analytic repositories for internal business operations and include a burgeoning volume of data from mobile apps, online games, Internet of Things (IoT) devices, social media networks, generative AI systems, and many other sources.

A *data warehouse* is a computer system dedicated to storing and analyzing data to reveal trends, patterns, and correlations that provide information and insight. Traditionally, organizations have used data warehouses to capture and integrate data collected from internal sources (usually transactional databases), including marketing, sales, production, finance, and more. However, unlike transactional databases, data warehouses are designed for analytical work. These software environments serve as federated merged repositories, collecting and aggregating data from various operational systems for analysis and generating business insights.

Defining Data Lakes

Data lakes arose to supplement traditional data warehouses because the relational model can't accommodate the current 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, such as when data scientists conduct data exploration for feature engineering and developing predictive models.

Data warehouses and data lakes are both widely used to store big data but aren't interchangeable. A *data lake* is a vast pool of raw data that is stored in a highly flexible format for future use. A *data*

warehouse is a repository of filtered data that has been preprocessed for a specific purpose. We explore these differences further in Chapter 2.

Understanding the Cloud Data Platform

A *cloud data platform* is a single, unified network that enables data analysts, data scientists, data engineers, and more to connect their data, applications, and services that are most critical for their business. It allows for workloads like data warehousing, data lake, data engineering, collaboration, AI/ ML, application development, and more. It makes it easy to share data with a diverse group of users without requiring the technology team to copy that data or establish a new data silo. It upholds centralized data security, data governance, and regulatory compliance policies to ensure that people obtain complete, consistent, and accurate data when they issue queries and generate reports — without violating data privacy mandates. It also can accommodate new architecture patterns such as a data mesh, and integrate open table formats such as Apache Iceberg tables (for more on this, see Chapter 2).

Consumption-based pricing allows each user and workgroup to allocate costs to specific accounts and cost centers with constant visibility into the compute and storage resources they use. Best of all, a modern cloud data platform operates seamlessly across multiple public clouds via one consistent interface, maximizing flexibility and avoiding the restrictions of a single cloud provider.

REMEMBER

Cloud data warehousing, which can live as a workload on a modern cloud data platform, emerged from the convergence of three major trends: 1) changes in data sources, volume, and variety; 2) increased demand for data access and analytics; and 3) technology improvements that significantly increased the efficiency of data storage, access, and analytics.

MARRIOT SIMPLIFIES ITS DATA PLATFORM AND ACHIEVES LOWER TOTAL COST OF OWNERSHIP

CASE STUDY

Marriot, a Snowflake customer, comprises 32 global brands across 139 countries, with 8,300 hotels offering 15 million hotel rooms, and 100,000 home and villa properties.

Prior to using a unified, single cloud data platform, Marriott used a mix of legacy database technologies that made their stack complex, costly due to expensive upgrades, and difficult to operate. Data engineers spent 20 percent of their time on infrastructure issues such as tuning Spark jobs.

Simplifying its data platform on Snowflake has enabled Marriott to achieve transparency and control of its data, faster speed to market, improved collaboration and data sharing, a better user experience, and lower TCO.

With Snowflake, Marriott has seen a dramatic improvement in performance and cost savings in comparison to Spark and Hive-based workloads. Many users from Marriott have commented on their improved experience with Snowflake, mentioning queries that used to take five hours or time out on Netezza that now take one hour on Snowflake. Data that previously took 48 hours to one week in Hadoop is now available nearly instantly in Snowflake.

Tracking the Emergence of Modern Cloud Data Warehousing

Traditional data platforms are designed to leverage a set of finite computing resources, often within the confines of an on-premises data center. Careful capacity planning is required to size each new data warehouse, data lake, *data mart* (a subset of a data warehouse that focuses on specific data for a particular purpose), or other data-driven workload. Because organizations don't always know how popular these workloads will become, they have to overprovision them — deploying more hardware and software resources than they expect to initially need.

As analytic applications, data science applications, data engineering pipelines, and many other types of data applications have grown in popularity and importance, many of these legacy data warehouse platforms have bowed under the strain. Restricted by a linear architecture, they can't run multiple workloads in parallel, leading to long wait times for computing resources and the data-driven insights they impart. Many users complain of slow, inefficient queries, scalability issues, and rising licensing costs as analytic workloads grow.

Complicating matters, many data-driven workloads are characterized by occasional bursts of activity, such as when the finance team closes the books at the end of the month or when data scientists train ML models. Sizing a data warehouse to accommodate peak loads is wasteful because the system needs all that capacity for only a small fraction of the time.

These issues stem, in part, from antiquated design principles. Older data warehouses use a "shared nothing" architecture that tightly couples storage, compute, and database resources. This type of architecture makes it difficult to elastically scale the database to respond to the escalating needs of many concurrent users and workgroups, as well as to accommodate occasional bursts in query activity.

The steady rise of public cloud services has empowered businesses to provision nearly limitless amounts of compute and storage capacity. Theoretically, this has allowed traditional data environments to support a larger number of users and workloads. In practice, however, older data warehouse systems were not structured to take advantage of all this power and capacity. While some of these data environments have been "lifted and shifted" to the cloud, they have continued to operate under the architectural limitations of their legacy, on-premises heritage.

In many cases, these information systems have been architected to work with a finite set of resources and to use a single type of data, which has led to data platform sprawl — a data warehouse for structured data, a data lake for semi-structured data types, and a wide variety of local databases and data marts, some in the cloud and others on-premises with each created to solve a unique set of departmental needs. This sprawl forces IT administrators to contend with the problem of *data silos*, which involves reconciling dissimilar architectures and different types of data stored in many different places.

Traditional data platforms don't scale well, and having a fixed set of compute and storage resources limits *concurrency* (the degree to which users can simultaneously access the same data and computing resources). Today, thanks to the nearly infinite resources available in the cloud, businesses can easily scale compute resources to handle an escalating volume of activity.

Looking at Data Processing Trends

Historically, businesses collected data in a well-defined, highly structured format at a reasonably predictable rate and volume. Even as the speed of older technologies advanced, data access and usage were carefully controlled and limited, given the scarcity of computing resources, to ensure acceptable performance for every user.

But now, the business world is experiencing a data deluge, with data arising from sources too numerous and varied to list. The velocity and volume of this data can quickly overwhelm a conventional data warehouse. In some cases, this can cause analytics applications to hang or even crash due to an overload of users and the workloads they attempt to run.

Adapting to Data Demands

It may be difficult to predict the amount of computing resources needed to analyze large and growing data sets, especially when an increasing share of this data originates outside your data center. This makes a cloud data platform the natural location for storing and integrating warehouse data.

A modern cloud data platform also enables *elasticity* to scale all your analytic workloads. Organizations and workgroups can acquire computing power for short periods, making projects easier to execute and allowing even small businesses to reap the benefits of a powerful data warehouse.

To take full advantage of cloud resources, a new architecture is required that separates but logically integrates storage, compute, and data warehouse services (such as metadata and user management). Chapter 3 explains that because each component

is separate, they can be expanded and contracted independently, enabling data warehouses to be more responsive and adaptable.

Adapting to the exponential increase of data also requires a fresh perspective (see Figure 1-1). The conversation must shift from how big an organization's data warehouse should be to whether it can scale cost effectively, without friction, and in the magnitude necessary to handle massive volumes and varieties of data, arriving at increasing velocity.

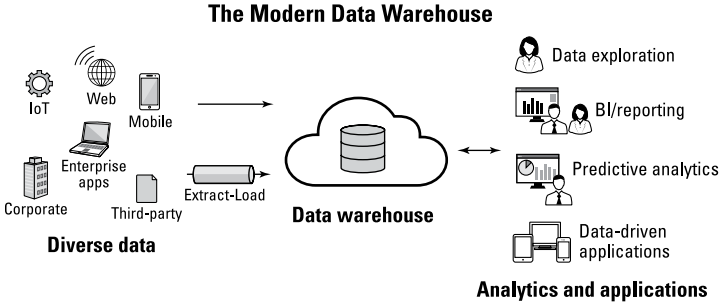

FIGURE 1-1: The modern data warehouse must support many types of data, analytic use cases, and applications.

SUMMING UP THE CHALLENGES OF DATA MANAGEMENT

The modern cloud data warehouse arose in response to several evolving data trends, all of which put a strain on legacy architectures:

- **Variety:** Data sources are numerous and varied, resulting in more-diverse data structures that must coexist in a single location to enable exhaustive and affordable analysis.
- **Resource contention:** When data storage and computation are physically tied together, analytics problems typically arise if either resource starts to run low.
- **Velocity:** Loading data in batches at specific intervals is still common, but many organizations require continuous data loading (micro batching) and streaming data (instant loading).

(continued)

(continued)

- **Elasticity:** Scaling up a conventional data warehouse to meet today's increasing storage and workload demands, when possible, is expensive, painful, and slow.
- **Diversity:** Proprietary data platforms are often complex, requiring specialized skills and lots of tuning and configuration. This worsens with the growing number of data sources, users, and queries.
- **Collaboration:** Sharing data usually requires building data pipelines and copying data around, which takes time and resources and often results in delays and negative downstream impacts.

CASE STUDY

OVERCOMING SCALABILITY ISSUES

Autodesk software solves challenges in architecture, engineering, construction, product design, manufacturing, media, and entertainment. Autodesk's customer 360 Analytics Data Platform (ADP) supports a variety of BI, data science, and customer-facing use cases.

Autodesk's data lake architecture was operationally burdensome to support and cost-prohibitive to scale. Data ingestion workloads relied on large amounts of homegrown code that led to frequent troubleshooting and unreliable data. Data-access-control limitations presented data governance challenges.

Performance issues inhibited Autodesk's product teams and business users from accessing timely insights. Lack of trust in ADP caused teams to consider building their own data environments. Near-zero maintenance reduced administrative work and freed up technical staff to focus on increasing analytics.

Adding native SQL support and an extensive network of connectors, drivers, and programming languages simplified data ingestion and transformation.

Autodesk's reimagined data architecture allows the data platform team to support even more self-service analytics use cases and gain the following benefits:

- Significantly reduced administration overhead (by 3x)
- 10x faster data ingestion and transformation
- Increased self-service access to analytics powered ML workloads

IN THIS CHAPTER

- Supporting many languages
- Working with many data formats
- Organizing data files with open table formats
- Utilizing new architectural patterns
- Simplifying data management with a data mesh
- Taking a modern approach to data lakes

Chapter 2
Standardizing on a Versatile Data Platform

Regardless of your industry or market, the capability to harness your data easily and securely in a multitude of ways has become paramount for success. A modern cloud data platform empowers you to consolidate your data, providing unlimited bandwidth for data analysis, data sharing, data engineering, application development, and data science initiatives. As a result, your business users become more efficient and your IT team can break free from mundane data administration tasks, allowing everyone to focus on delivering valuable experiences.

Each role has unique data requirements — from developers to data architects to operational workers. As a result, a cloud data warehouse must live on a cloud data platform that can work with numerous programming languages, be compatible with prevailing architectural patterns, and integrate smoothly with a wide variety of data formats.

Supporting Many Languages

SQL, Python, Scala, Java, JavaScript — developers interact with many languages to access data and build data applications, including non-coding languages, natural languages, and conversational interfaces, such as generative AI tools that use programming languages behind the scenes.

A cloud data warehouse should live on a cloud data platform that works seamlessly with these languages. In addition, business analysts should be able to use ANSI SQL to manipulate all data, including support for joins across data types and databases.

Flexible access via SQL and other popular languages makes it easier to build data pipelines, run exploratory analytics, train ML models, and perform other data-intensive tasks. This is the starting point for enabling a broad set of business intelligence (BI), reporting, and analytic use cases.

Working with Many Data Formats

Traditional data warehouses are optimized for storing relational data in predefined tables. However, today's data warehouses must accommodate many other data types and file formats, including raw and streaming data from weblogs, equipment sensors, social media networks, and other sources that don't conform to a rigid tabular structure. Web data may be stored as JSON files. Spreadsheets may occupy comma-separated value (CSV) formats or tab-delimited text files. And data interchanged among multiple applications may be defined in extensible markup language (XML), complete with tags and other coding that identify distinct entities within the data.

A cloud data platform should natively support popular semi-structured data formats, including the following:

- **JSON,** a lightweight, plain-text, data-interchange format based on a subset of the JavaScript Programming Language. JSON data can be produced by any application.
- **Apache Avro,** an open-source data serialization and Remote Procedure Call (RPC) framework originally developed for use with Apache Hadoop. Avro utilizes schemas defined in JSON

to produce serialized data in a compact binary format. The serialized data can be sent to any destination (that is, application or program) where it can be easily deserialized because the schema is included in the data.

- **Apache ORC (Optimized Row Columnar),** a columnar format used to speed up Apache Hive queries. ORC was designed for efficient compression in Hadoop and improved performance of Hive for reading, writing, and processing data.
- **Apache Parquet,** a compressed, efficient columnar data representation designed for projects in the Hadoop ecosystem. This file format supports complex nested data structures and uses Dremel record shredding and assembly algorithms.
- **XML,** a markup language that defines a set of rules for encoding documents. XML was originally based on standard generalized markup language (SGML), another markup language developed for standardizing the structure and elements that comprise a document.

THE THREE BASIC DATA TYPES

Most data can be grouped into three basic categories:

- **Structured data** (customer names, dates, addresses, order history, product information, and so forth) is generally maintained in a neat, predictable, and orderly form, such as the tables in a relational database or the rows and columns in a spreadsheet.
- **Semi-structured data** (web data, spreadsheet data, XML data) doesn't conform to traditional structured data standards but contains tags or other types of markups that identify distinct entities within the data.
- **Unstructured data** (audio, video, images, PDFs, and other documents) doesn't conform to a predefined data model or is not organized in a predefined manner. Unstructured information may contain textual information, such as dates, numbers, and facts that are not logically organized into the fields of a database or semantically tagged document.

REMEMBER

A complete cloud data platform can store diverse types of data in their native formats without creating data silos or imposing unique schemas to access 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.

Utilizing Open Table Formats

In addition to standardizing on a cloud data platform that supports JSON, Avro, Parquet, and XML file formats, make sure it works with your desired table format, whether proprietary or open source. Apache Iceberg is a widely popular open table format with a large ecosystem of contributors, vendors, and users, ensuring you don't lock your data into any single vendor.

Iceberg adds a SQL-like table structure to the unstructured and semi-structured data stored in files and documents. You can store Iceberg metadata and data files in your object storage and query them in-place. This allows computing engines, such as Spark, Trino, PrestoDB, Apache Flink, Hive, and Snowflake, to easily manage and inspect the data.

Open table data formats have tremendous momentum from the commercial and open-source communities. Will your data platform support them if needed?

TECHNICAL STUFF

Even when most of your data is maintained in a centralized data warehouse repository, it's still possible to accommodate data in *external tables* (read-only tables that can be used for query and join operations) and *materialized views* (database objects that contain the precomputed results of a query). This architecture enables seamless, high-performance analytics and governance, even when the data arises from more than one location.

Supporting New Architectural Patterns

One reason technology projects fail is because the stakeholders fail to look ahead. Don't just look at your current state; consider how your business may evolve in the future.

Historically, companies have invested in special-purpose technologies and data platforms, and it's a huge effort to migrate

them to more open and versatile formats. Such migrations can become a massive undertaking, sort of like trying to copy a lifetime's worth of family movies from an analog VHS format to a digital format like MP4.

With new types of data, you may encounter new architectural patterns that you didn't predict. For instance, you may want a data warehouse to be transformed into a hybrid pattern that merges the strengths of data warehouses and data lakes. Additionally, domain-specific data marts could evolve into a more streamlined and regulated data mesh.

A modern data platform supporting the data warehouse workload must be able to accommodate these patterns and easily adapt to your evolving business needs, as shown in Figure 2-1.

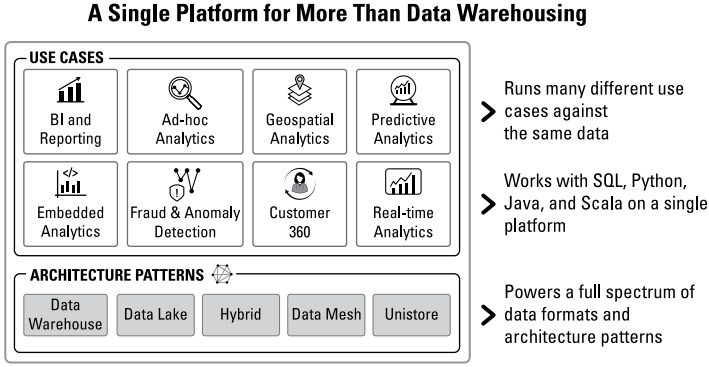

FIGURE 2-1: A versatile data platform powers a full spectrum of use cases, whether data is stored inside a data warehouse or in external tables.

REMEMBER

By making rigid demands about how to structure your data, you may unwittingly determine how to structure your business. The right data platform will allow you to do new things in familiar ways through a familiar interface. This maximizes flexibility as your business evolves.

Improving Control with a Data Mesh

A *data mesh* simplifies the process of managing massive data architectures by breaking them down into smaller functional domains, each overseen by a dedicated team. These domain teams are responsible for crucial tasks, such as building and maintaining data pipelines, implementing governance policies, upholding

data privacy mandates, and ensuring data quality. Rather than creating silos, a data mesh breaks them down — it distributes data responsibilities across different teams or domains while maintaining data discoverability and accessibility.

This architectural pattern confirms that the teams working with the data have in-depth knowledge and expertise, fostering greater ownership and accountability as each data set aligns with the overall needs of the business. By distributing data responsibilities across the organization, a data mesh fosters a culture of data democratization and encourages cross-functional collaboration.

When anchored by a modern cloud data platform, a data mesh can incorporate many types of data and file formats and accommodate external data sources, different workloads, and multiple clouds.

Moving Beyond Data Lakes

Data lakes are designed to store huge quantities of raw data in their native formats in a single repository. However, business users often find accessing and securing this vast pool of data difficult, and many organizations have a hard time finding, recruiting, and retaining the specialized IT experts needed to access the data and prepare it for downstream analytics and data science use cases. Additionally, most of today's data lakes can't effectively organize all of an organization's data, which may originate from dozens of data streams and data silos that must be loaded at different frequencies, such as once per day, once per hour, or via a continuous data stream.

In response, hybrid platforms have emerged that combine the best attributes of data warehouses and data lakes into a single platform. 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.

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 nearly 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.

IN THIS CHAPTER

- Defining essential architectural attributes
- Enabling data workloads across regions and clouds
- Organizing your data for out-of-the-box analytics

Chapter 3
Architecting a Cloud Data Platform That Just Works

Creating an effective cloud data warehouse isn't just a matter of repurposing yesterday's on-premises technologies or moving existing analytic applications and databases from your data center to a cloud vendor's infrastructure. Properly leveraging the power and scale of the cloud requires a new mindset, a new set of management principles, and new cloud-built capabilities.

Outlining the Primary Architectural Components

To best satisfy the requirements of diverse and ever-escalating data workloads, a modern cloud data platform should be built on a *multi-cluster, shared data architecture,* in which separate compute, storage, and services can be scaled independently to leverage all the resources of the cloud.

A modern cloud data warehouse includes a central persisted data repository that is accessible from all compute nodes. Like a shared-nothing architecture, it processes queries using MPP (massively parallel processing) compute clusters.

This architecture allows maximum scalability, because each node in the cluster stores a portion of the entire data set locally. A near-limitless number of users can query the same data concurrently without degrading performance, even while other workloads are executing simultaneously, such as running a batch processing pipeline, training a machine learning model, or exploring data with ad hoc queries. A multi-cluster, shared data architecture includes four layers that are logically integrated yet scale independently from one another:

- The *storage layer* holds your data, tables, and query results. This scalable repository should handle structured, semi-structured, and unstructured data and span multiple regions within a single cloud and across major public clouds.
- The *compute layer* processes enormous quantities of data with maximum speed and efficiency. You can easily specify the number of dedicated clusters you want to use for each workload (thus eliminating contention for resources) and have the option to let the service scale automatically.
- The *services layer* coordinates transactions across all workloads and enables concurrent data loading and querying activities, enforcing security, propagating metadata, optimizing queries, and performing other important data management tasks. When each workload has its own dedicated compute resources, operations can run simultaneously and perform as needed.
- The *cross cloud and global layer* globally connects data and applications across regions and clouds, securely, through a single, consistent experience, and is described further below.

Spanning Multiple Regions and Clouds

Many companies store data in multiple clouds and regions, necessitating a cohesive cross-cloud strategy that can attain business continuity, resilience, and collaboration no matter where data is located. A recent survey, part of Snowflake's Data Trends Report,

examined data usage patterns at 7,800 organizations — all Snowflake customers. According to the survey, the number of organizations operating across the three leading public cloud providers (Amazon Web Services, Microsoft Azure, and Google Cloud) grew 207% during the 12 months ending January 2023.

These companies need data warehouses that can store and manage data consistently across many different geographic regions and clouds. However, when working with multiple cloud providers, how do you ensure that the same security configurations, administrative techniques, analytics practices, and data pipelines apply to all your cloud providers? For example, will you have to resolve differences in audit trails and event logs or apply unique tuning and scaling techniques on each cloud? Will your security experts have to deal with varying sets of rules or work with multiple key management systems to encrypt data? Will data engineers have to create unique pipelines?

A cross-cloud data platform enables data administrators to apply consistent policies to all data in all areas. This makes it easier to keep up with changing regulations, apply regional locality controls, and take advantage of whichever public cloud services best match your evolving business strategy.

Once you have this type of technology layer in place, it quickly becomes a competitive advantage, allowing you to achieve results faster, comply with data governance procedures more easily, and maintain uninterrupted operations through seamless data replication (see Figure 3-1).

A cross-cloud data warehouse provides a consistent layer of services across regions of a single public cloud provider and between major cloud providers, with the following emphases:

- **Continuity:** The data warehouse must offer inherent resiliency to eliminate disruptions, comply with changing regulations, and simplify data migrations among different vendor clouds.
- **Governance:** Your data warehouse should offer flexible policies, tags, and lineage capabilities that follow the data, ensuring consistent enforcement across users, workloads, clouds, and regions.

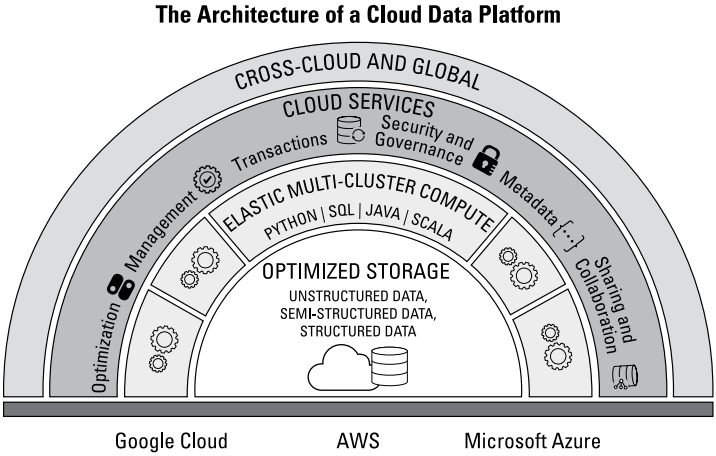

FIGURE 3-1: A modern cloud data platform should seamlessly operate across multiple clouds and apply a consistent set of data management services to many types of data workloads.

» **Collaboration:** A cloud data warehouse should allow workers to instantly discover, access, and share data, services, and applications across clouds and regions, without requiring complex integration technology such as file transfer protocol (FTP) or extract, transform, and load (ETL) procedures.

Consolidating Data for Out-of-the-Box Analytics

One of the fundamental principles of this book is to encourage all stakeholders in your organization — including line-of-business managers, data analysts, data engineers, data scientists, application developers, and frontline workers — to actively leverage the same single source of data. This ensures consistent outcomes and accelerates time to insight by reducing the time spent wrangling data.

In practice, rallying the enterprise around a single source of truth is rarely a seamless process, mainly due to how corporate information systems have been designed and implemented over the last several decades. Whether on-premises or in the cloud, each

production application creates its own data silo. For example, marketing data resides in a marketing automation system, sales data in a customer relationship management (CRM) system, finance data in an enterprise resource planning (ERP) system, and inventory data in a warehouse management system, among others.

These disparities are carried over to the analytic databases derived from these production systems. Operational reporting may be the province of a data warehouse, while departmental analytics relies on data marts and data mining, or exploration requires a data lake. Sharing data among these systems may need specialized data pipelines powered by complex ETL procedures. The situation has become even more complex with the rise of thousands of software-as-a-service (SaaS) tools and mobile apps, each with its own unique sources of data.

Achieving operational efficiency

All cloud data warehouse vendors offer some degree of automation, but it's crucial to delve deeper to determine the level of automation they truly provide. Ideally, your data warehouse platform should be seamlessly managed, updated, secured, governed, and administered without requiring extra effort from your IT team.

When it comes to software updates, you should automatically receive the latest functionality without enduring a lengthy, manual upgrade process. You shouldn't have to worry about planning for updates, experiencing downtime, or making modifications to your installation. The cloud data warehouse provider handles all administrative tasks related to storage, encryption, table structure, query optimization, and metadata management behind the scenes, effectively eliminating the need for manual administration.

To determine how much work will be necessary, ask your cloud data warehouse vendor these questions:

- Do we have to optimize resource usage or manually scale the system, such as requesting an additional cluster when more compute power is required?
- Does the provider automatically apply software updates, such as security patches, as soon as those updates are available? Or does it merely manage the underlying infrastructure and require us to keep the software platform up to date?

- Does the service automatically encrypt all our data at rest and in motion with industry-standard encryption, or do we have to set up and apply encryption to the data manually? Does the encryption system hinder query performance?
- Does the service scale up and out instantaneously and elastically and then release extra compute or storage resources when they're no longer in use? Or do we have to handle these tasks manually?
- Does the cloud provider automatically replicate your data to ensure business continuity across regions? After cross-regional replication is established, do we have to set up change data capture (CDC) procedures to keep multiple databases in sync, or does the vendor handle that for us?
- Do we need to partition data, tune SQL queries, and optimize performance, or does the platform handle this automatically?

Provisioning and managing resources

Your cloud data warehouse should allow you to right-size the computing infrastructure to match the resource needs of each workload. For example, if you're running a data pipeline with low compute requirements, you can match a small cluster to that workload rather than incur the cost of an overprovisioned cluster. If you need to test new machine learning modules or run advanced analytics, you can utilize a large cluster.

The best cloud data platforms have an elastic performance engine that permits variable concurrency without resource contention, tuning, or the need to manage the system. The data platform supports any number of users, quantity of jobs, or volume data with reliable multi-cluster resource isolation. This gives you fine-grained scalability for each workload while minimizing usage costs.

REMEMBER

With some cloud data platforms, IT is responsible for provisioning and managing new resources. In other platforms, all the infrastructure is provisioned and managed behind the scenes; you simply run your queries or processing jobs and the cloud data platform does the rest, abstracting technical complexities and automating system management activities in the background.

IN THIS CHAPTER

- Ensuring value through consumption-based pricing
- Using columnar storage to maximize efficiency
- Looking at the right metrics to keep costs down
- Improving performance and total cost of ownership (TCO) by fine-tuning compute resources

Chapter 4
Achieving Exceptional Price and Performance

Fast analytical performance is crucial for data-informed decision-making. However, the more data you ingest and process in your data warehouse, the more cloud resources you consume, which can have a direct impact on costs.

There are three essential aspects to cost optimization in a cloud data warehouse:

- **Visibility:** Users can fully understand their spending and attribute it accurately to designated cost centers.
- **Control:** Administrators can set limits and take corrective actions to govern resource use.
- **Optimization:** Companies can identify inefficient spending and reallocate funds for more impact.

This chapter dives into these aspects and describes how to achieve cutting-edge performance while simultaneously monitoring data warehouse costs and optimizing resource use.

Utilizing Consumption-Based Pricing

Make sure that the pricing model for your cloud data warehouse matches the value you obtain from it. Paying for a set amount of storage and computing power, commonly known as *subscription-based pricing*, can incur significant yearly costs and typically requires regular management. To ensure that you don't pay for more capacity than you need, your cloud data platform should offer usage-based pricing.


TIP

Usage-based pricing allows you to choose how data users at your organization consume resources. Some cloud data platforms allow you to pay for usage per second with a one-minute minimum, increasing control over costs.

Maximizing Efficiency with Columnar Storage

Data uploaded into the data warehouse should be reorganized into a compressed columnar format. Because columnar databases use less memory to output data, more data can be stored, speeding up queries.

Examine the terms of your usage agreement: Expect to pay only for storage you use, not for excess or reserved storage capacity. You also shouldn't pay to clone databases within your data warehouse for development and testing activities. You want to be able to reference — not copy — your data multiple times and therefore not have to pay extra for storage. Chapter 6 covers data sharing and collaboration in detail.


REMEMBER

Compute resources are more expensive than storage resources, so your data warehouse service should allow you to scale each resource independently and make it easy to spin up exactly the compute resources you need under a usage-based pricing model. The vendor should bill you only for the resources you use — down to the second — and automatically suspend compute resources when you stop using them. It's useful to receive those charges in an all-inclusive bill with no hidden costs or fees.

Calculating and Controlling Costs

As enterprises migrate IT workloads to the cloud, they're transitioning from a world of scarcity to a world of abundance marked by nearly limitless data storage resources and nonstop data processing capacity. It's important to control costs and rein in excessive consumption.

The cost of using a cloud data warehouse is typically based on three interrelated metrics: data transfer volume, data storage consumption, and compute resources. A cloud data platform separates these three services to give administrators complete control over data warehouse usage.

Your data platform must make it easy to track the consumption of all cloud services. This includes built-in resource monitoring and management features that provide transparency into usage and billing, ideally with granular chargeback capabilities to tie usage to individual budgets, departments, and workgroups.

Data warehouse administrators can set guardrails to ensure that no individual or workgroup spends more than expected. For example, they can set time-out periods for each type of workload along with *auto suspend* and *auto resume* features to automatically start and stop resource accounting when the platform isn't processing data.

They may also set limits at a granular level, such as determining how long a query can run before it's terminated, which helps to avoid unexpected costs associated with runaway queries.

Optimizing Performance and TCO

Fine-tuning the compute resources provided by a cloud data warehouse can improve the performance of a query or set of queries. Administrators can resize the environment whenever necessary, even while running production workloads in tandem. They can also start or stop the entire data warehouse at any time to optimize overall price and performance.

Look for a cloud data warehouse solution that automatically optimizes performance and eliminates administrative effort to incorporate new resources. Whether it's search optimization (SO) capabilities, more efficient storage compression techniques, or reduced compilation time for SQL queries, you shouldn't have to do anything to gain access to new features or the latest capabilities.

That's the beauty of subscribing to cloud services from a reputable data platform provider: New functionality appears instantly, without tedious upgrade cycles. Regularly released platform optimizations and software updates continuously improve performance, often while simultaneously lowering costs.

AUTOMATION DRIVES INNOVATION

CASE STUDY

Veradigm, a Snowflake customer, is a technology company that delivers care and financial solutions to healthcare providers. To provide stakeholders with actionable data and insights, the company ingests and analyzes large amounts of data on electronic health records, disease registry data, and claims data.

Unfortunately, with Veradigm's legacy data warehouse environment, onboarding new data sources took up to nine months. Furthermore, data processing limitations made it difficult to join tables that contained medication, laboratory, and other healthcare data.

Realizing the need for a more modern data environment, Veradigm subscribed to a cloud data platform with a multi-cluster shared data architecture. The platform automatically scales storage and compute resources, eliminating performance issues, lowering costs, and offering more granular control. For example, one group at Veradigm reduced its resource consumption from $40,000 per month to less than $4,000 per month, even though team members were processing twice as much data.

With a fully managed infrastructure and near-zero maintenance, Veradigm's cloud data platform has enabled the company to support additional data use cases such as a data lake without increasing headcount and easily meets its service level agreements (SLAs) for each workload. All data resides in one multipurpose repository, which is much simpler than wrangling multiple disparate data sets.

IN THIS CHAPTER

- Securing data through encryption, user login controls, access controls, and more
- Applying governance policies to protect data and maintain the quality of your data

Chapter 5
Bolstering Data Security and Governance

In recent years, there has been a spike in the proliferation of data generated and collected by organizations. With data from third-party sources becoming more common — such as data from SaaS apps, popular application clouds, data marketplaces, data exchanges, and more — data security, data privacy, data governance, and regulatory compliance have become much more complicated. Organizations need to understand the source of common threats and take a hard look at who might be trying to misuse, breach, or attack their database management systems. For example, trade secrets may be valuable to industry competitors, while energy grid information is a target for political saboteurs. Understanding these realities is the starting point for setting up comprehensive security, governance, and compliance policies that can be consistently enforced across your entire data estate.

Exploring the Fundamentals of Database Security

Securing your data and complying with pertinent regulations is fundamental to the architecture, implementation, and operation of a cloud data warehouse service. All aspects of the service must be centered on protecting your data as part of a multilayered strategy that considers both current and evolving security threats. Your security strategy should address external interfaces, access control, data storage, and physical infrastructure in conjunction with comprehensive network monitoring, alerts, and verifiable cybersecurity practices.

Eliminating security silos

Some organizations enforce security and governance policies by creating unique data silos and then limiting access to each silo based on account, region, role, and other variables. This approach complicates data governance. Rather than creating unique data silos with unique data protection policies, establish universal, application-level controls that apply to one centralized repository.

TIP

Just as it is important to eliminate data silos, a good security strategy seeks to eliminate identity silos as well.

Encrypting data by default

Encrypting data means applying an encryption algorithm to translate the clear text into cipher text. All warehouse data should be encrypted by default using the latest security standards and best practices. Encrypt data from the time it leaves your premises, through the internet, and into the warehouse: when it's stored on disk, moved into a staging location, placed within a database object, and cached within a virtual data warehouse. Query results should also be encrypted.

The vendor must protect the decryption keys that decode your data. The best service providers employ AES 256-bit encryption with a hierarchical key model. This method encrypts the encryption keys and instigates key rotation that limits the time during which any single key can be used.

Data encryption and key management must be always on and entirely transparent. Having the option to supply your own encryption keys is important so that you can disconnect the cloud provider from your data if necessary.

Verifying vendor participation

Some cloud data warehouse vendors automate only rudimentary security capabilities, leaving many aspects of data encryption, access control, and security monitoring to the customer. Other vendors handle these tasks for you. Before standardizing on a cloud data platform for your data warehouse deployment, ask the vendor these questions:

- Does the service enforce essential security attributes by default, such as encryption, threat detection, and incident response?
- Does it follow Center for Internet Security (CIS) Benchmarks for configuring IT systems, software, networks, and cloud infrastructure?
- Are security controls global, comprehensive, and easy to configure?
- Does the vendor subscribe to a shared responsibility model, and is it clear who's responsible for which aspects of security?
- Can we bring our own identity and establish SSO (single sign-on)?
- Can our data administrators set granular access controls (such as column- and row-level restrictions), along with role-based access to database tables?
- Is security applied not only to the central data repository but to external tables as well?
- Does the vendor regularly perform compliance audits and have the necessary security attestations to show?

Patching, updates, and network monitoring

Software patches and security updates must be installed on all pertinent software components as soon as those updates are available. The vendor should deploy periodic security testing

(also known as penetration testing) by an independent security firm to proactively check for vulnerabilities.

As an added protection, file integrity monitoring (FIM) tools ensure that critical system files aren't tampered with, and IP address allowed lists enable you to restrict access to the data warehouse to only trusted networks.

Security "events," generated by cybersecurity monitoring systems that watch over the network, need to be automatically logged in a tamper-resistant security information and event management (SIEM) system. Automatic alerts should be sent to security personnel when suspicious activity is detected.

Ensuring data protection, retention, and redundancy

In case of a mishap, you should be able to instantly restore or query previous versions of your data in a table or database within a specified retention period, as governed by your service-level agreement (SLA) with the cloud data warehouse provider. A complete data-retention strategy goes beyond duplicating data within the same cloud region or zone; it replicates that data among multiple availability zones for geographic redundancy. Optionally, automatic failover to these other zones can ensure continuous business operations.

Securing marketplace data

A growing number of organizations leverage a data warehouse to develop data applications not only for internal use but also for external use via a data marketplace. Sharing data through marketplace apps necessitates another level of security. Data providers must be able to guard, monitor, and review application submissions to vet potential users.

In some cases, data providers create data clean rooms that enforce designated governance policies. These sanitized data sets can be confidently shared with partners and other external constituents without exposing sensitive information.

Controlling user logins

For maximum convenience and security, a cloud data warehouse will allow you to apply your chosen SSO and identity access

management (IAM) procedures. The data warehouse should also permit you to apply multifactor authentication (MFA) at the account level. This permits you to require some or all users to pass through a secondary level of verification such as entering a one-time security code sent to the user's mobile phone.

SSO procedures and federated authentication make it easier for people to log in to the data warehouse service directly from other sanctioned applications. *Federated authentication* centralizes identity management and access control procedures, making it easier for data warehouse stakeholders to manage user access privileges.

Applying access controls

To protect sensitive data, a cloud data warehouse service must authorize users, authenticate credentials, and grant people access only to the data they're authorized to see. *Role-based access control (RBAC)* policies need to be applied to all database objects, including tables, schemas, and virtual extensions to the data warehouse.

Ideally, data administrators can apply granular access controls down to the rows and columns of database tables. For example, this type of control could be used to permit users to see basic employee data but not Social Security numbers, salaries, and other sensitive information.

Governing How People View, Access, and Interact with Your Data

Governance policies establish rules and procedures to control the ownership and accessibility of your data. Applying global, universal data governance policies allows you to scale your data estate with confidence.

For example, interaction controls, like secure views, secure joins, and secure user-defined functions (UDFs), are applied as people interact with the data:

- **Secure views** give data custodians control over data access, preventing security breaches. For instance, customers can view specific rows of data from a table that excludes rows pertaining to other customers.

- **Secure joins** establish linkages without revealing personally identifiable information (PII). It allows discreet connections to people, devices, cookies, or other identifiers.
- **Secure UDFs** let users analyze fine-grained data while protecting raw data from being viewed or exported by other parties.

Protecting your data

Organizations concerned about safeguarding sensitive data can control access at a more granular level. Common data protection methods include the following:

- **Row access policies** allow users to see only information relevant to them. For example, sales reps may only access customer data for their own accounts while regional managers can access all customer data within their regions.
- **Dynamic data masking** selectively conceals data during queries. This technology allows you to store PII and still perform robust analytics on the data without exposing it to unauthorized users.
- **External tokenization** transforms data into an unrecognizable string of characters with no meaningful value in case of a system breach. The data can be dynamically detokenized at query runtime.

Classifying and identifying data

Classification and identification policies help you avoid data privacy leaks and compliance breaches by tracking the types of data in use, its lineage, and how it changes. For example, you can use *object tagging* to control access to confidential and sensitive information such as salary amounts and Social Security numbers.

Traceability tools let users track data wherever it resides, ensuring continuous protections and enabling data deletion when necessary (including the "right to be forgotten").

Data lineage tools, whether embedded in the cloud data platform or provided as additional services, help you understand how data flows through your data-processing systems. This knowledge

assists compliance officers in tracing the usage of sensitive data, including its sources, destinations, and any transformations along the way.

Demanding attestations and compliance certifications

Compliance isn't just about robust cybersecurity practices. It's also about ensuring that your data warehouse provider can prove it has the required security procedures in place. Industry-standard attestation reports that verify cloud vendors use appropriate security controls. For example, a cloud data warehouse vendor needs to demonstrate that it adequately monitors and responds to threats and security incidents and has established sufficient incident response procedures.

In addition to industry-standard technology certifications, such as ISO/IEC 27001 and SOC 1/SOC 2 Type II, you'll want to verify that your data warehouse provider complies with all applicable government and industry regulations. Depending on your business, this could include the following:

- Payment Card Industry Data Security Standards (PCI-DSS)
- GxP data integrity requirements
- HIPAA/HITRUST privacy controls
- ISO/IEC 27001 security management provisions
- International Traffic in Arms Regulations (ITAR)
- FedRAMP certifications

TIP

Ask your providers to supply complete attestation reports for each pertinent standard.

Monitoring data quality

Data governance requires rigorous oversight to maintain the quality of the data your company uses internally and shares with external constituents. Bad data can lead to missed or poor business decisions, loss of revenue, and increased costs. *Data stewards* — charged with overseeing data quality — must be empowered to proactively uncover anomalies in the data, such as when data is corrupt, inaccurate, or not being refreshed often enough to be

relevant. The best data platforms include out-of-the-box system metrics for the most common types of data quality issues, and make it easy to define, measure, and monitor data quality via integrated, cloud-native facilities (see Figure 5-1).

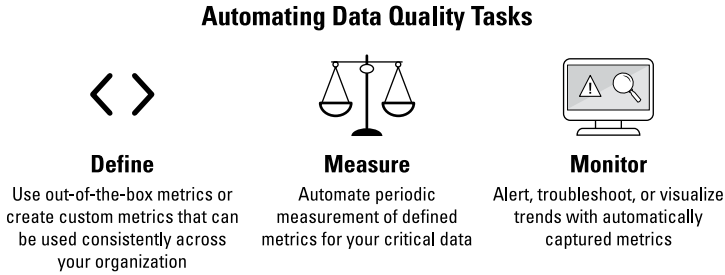

FIGURE 5-1: A complete cloud data platform empowers data stewards to enforce data quality via cloud-native management facilities.


TIP

Establishing comprehensive security and governance policies is not only about reducing risk but also about increasing productivity. If your data platform lacks an integrated set of applications for data custodians, data stewards, compliance officers, and other experts, you'll have to cobble together these capabilities from third-party tools. At best, this scattered approach will make it difficult to enforce organization-wide policies. At worst, it will introduce delays — or even cause users to mistrust the data, leading to poor decision-making, a lack of a data-driven culture, and inefficiency.

As you provide access to your users, pay attention to these tenets of data governance:

- **Know your data:** Classify data, tag sensitive data, and audit data usage
- **Protect your data:** Secure sensitive and regulated data with granular access policies
- **Connect your ecosystem:** Seamlessly extend your data governance policies as you share data, internally and externally, across regions and clouds.

IN THIS CHAPTER

- Recognizing and overcoming technology limitations
- Sharing data without copying or duplication
- Extending security and governance policies to shared data
- Monetizing data and data services via a data marketplace

Chapter 6
Enabling Data Sharing

Data sharing is the act of providing access to data — both within an enterprise and between enterprises. The organization that makes its data available, or shares its data, is a *data provider.* The organization that wants to use the shared data is a *data consumer.* Any organization can be a data provider, a data consumer, or both.

There's an abundance of potential value to unlock from the world's burgeoning data sources. Until recently, however, no technology existed for sharing data without a significant amount of risk, cost, headache, and delay.

Confronting Technical Challenges

Traditional data-sharing methods, such as File Transfer Protocol (FTP), application programming interfaces (APIs), and email, require you to make a copy of the shared data and send it to your data consumers. These cumbersome, costly, and risky methods produce static data that quickly becomes dated and must be refreshed with more current versions, requiring constant data movement and management via data pipelines, and causing a loss

of data version control. These complexities, coupled with database inconsistencies, authenticity headaches, and the difficulty of sharing large volumes of data add up to frustrating, expensive, and time-consuming data exchange processes.

TIP

Look for a cloud data platform that allows you to accomplish the following:

- Share data easily and securely across clouds, companies, teams, departments, functions, and business units
- Easily set up security and governance with built-in permissions and roles for ease of administration
- Share data, views, and dashboards to permit collaborative decision-making through a single, consistent user interface
- Deliver direct access to live, ready-to-query data across clouds and regions with on-demand fulfillment and no programmatic APIs, FTP transfers, or ETL procedures
- Safely share highly sensitive or regulated data without exposing it to unauthorized users by applying privacy-enhancing technologies and cross-cloud data clean rooms

Sharing without Copying

A cloud data platform is ideal for a data-sharing service because it enables authorized members of a cloud ecosystem to tap into live, read-only versions of the data. Organizations can easily share and receive slices of data in a secure and governed way. This method doesn't require data movement, extract, transform, load (ETL) technology, or constant updates to keep data current. There's no need to transfer data via FTP or to configure APIs to link applications. Because data is shared rather than copied, no additional cloud storage is required. With this superior architecture, data providers can easily and securely publish data for instant discovery, query, and enrichment by data consumers, as shown in Figure 6-1.

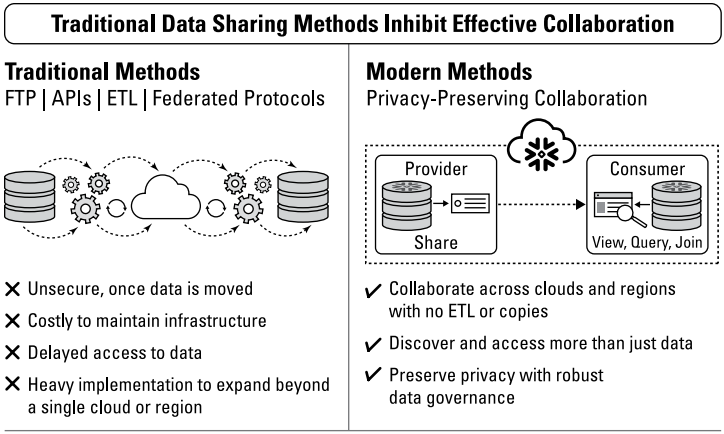

FIGURE 6-1: Identifying the attributes of modern data sharing.

Protecting Sensitive Data

In some cases, portions of a data warehouse are subject to strict security and confidentiality policies. Before you can share these parts of the data set, you may need to mask or anonymize certain fields, rows, or columns. This allows people to analyze the data without seeing the sensitive data elements.

Choose a cloud data platform that allows data providers to easily control access to individual database tables with granular protections policies and privacy-enhancing technologies. All the pertinent data security and governance capabilities should apply to your data-sharing architecture (for more on this, see Chapter 5). For example, controlling who can view and analyze sensitive or regulated data should be easy. Furthermore, you need to be able to share tables without exposing designated elements, either through privacy-enhancing technologies, such as aggregation and projection constraints, or data clean rooms.

Monetizing Your Data

Modern data-sharing technology sets the stage for collaborating and monetizing data via *marketplaces* — online communities that facilitate the purchase and sale of data and data services. For example, a telecommunications company can sell location data

to help retailers target consumers with ads. Consumer packaged goods companies can share purchasing data with online advertisers or directly with customers.

In addition to monetizing data, a marketplace allows you to monetize business logic, such as user-defined functions (UDFs), as well as applications.

If sharing data and applications through a marketplace is important to you, opt for a cloud data platform that has a thriving marketplace associated with it. Some platforms make it easy to discover third-party data, data services, and applications from hundreds, or even thousands, of providers, and can market and deliver your data products and services (see Figure 6-2).

Marketplace customers can use cloud credits and budgets to purchase data and data services. Such platforms may also offer built-in facilities to meter application usage and handle the associated billing. These capabilities allow data providers to focus on supplying value-added data services rather than getting caught up in administrative chores.

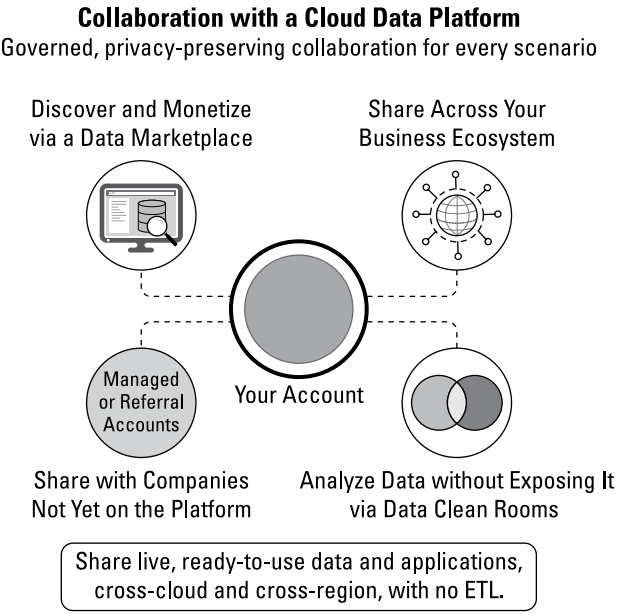

FIGURE 6-2: A cloud data platform enables you to securely leverage your data warehouse to share and collaborate with your data, for every scenario.

IN THIS CHAPTER

- Accommodating geospatial analytics
- Optimizing search activities
- Exploring the benefits of ML-powered functions
- Developing AI applications
- Understanding the importance of automation

Chapter 7
Advancing Analytics

Business intelligence (BI) is no longer merely the domain of executives, professional analysts, and data scientists. An effective cloud data platform that supports data warehouse workloads establishes not only a common repository for all types of data and analytics but also empowers diverse teams to collaborate and easily manage data. Popular analytic methods include the following:

- **Ad hoc analytics** allow business users to answer discrete business questions iteratively, such as tracking monthly sales or reviewing on-hand product inventory. Dynamic elasticity and dedicated resources for each workload power these queries without slowing down other workloads.
- **Event-driven analytics** constantly incorporate new data to update reports and dashboards so managers can monitor the business in real time or near-real time. Ingesting and processing streaming data requires an elastic data warehouse to handle variations and spikes in data flow.
- **Embedded analytics** operate as separate and distinct business processes within applications. The cloud facilitates data transfers from cloud-based applications to a cloud data warehouse where inherent scalability and elasticity can better support fluctuations in users and workloads.

The data warehouse workload in your cloud data platform should support a broad ecosystem of third-party BI solutions, as well as offer native tools for specific types of analysis. Some of the primary capabilities are summarized below.

Considering Geospatial Analytics

Most companies use geospatial data due to its capability to offer insights into location-based trends and patterns. For example, retailers collect geospatial data about sales territories, store locations, and customer addresses to design better supply chains. Healthcare companies collect geospatial data to track the penetration of viruses and diseases. Telecommunications firms use it to monitor subscriber usage and optimize their communications networks. Logistics companies collect it to plan routes and optimize shipping activities. In some cases, this data is stored as simple numeric coordinates. In other cases, it resides in specialty data types such as spherical (geography) or flat surface (geometry).

Collecting and analyzing spatial data involves new methods of data integration, analysis, governance, and interpretation. Traditional data warehouse systems can't handle location data at scale because they have limited processing power, lack robust spatial analysis capabilities, and are difficult to integrate with geographic information systems (GIS).


TIP

Select a data platform that can store and process any type of spatial vector object and perform complex geospatial transformations, such as converting geographic coordinates to street addresses. The processing engine must be able to handle location data at scale and seamlessly integrate with leading GIS tools.

Optimizing Search Functions

Search optimization features can significantly improve the performance of certain types of queries on tables such as the following:

- Queries that use selected geospatial functions with geography values
- Selective point lookup queries on tables

- Substring and regular expression searches
- Queries on fields in columns that use certain types of predicates

Your cloud data platform should offer optimized search capabilities that allow analysts to efficiently explore and query large volumes of data for point lookup queries, log analytics, star joins, substring searches, and geospatial searches. These capabilities are especially useful for needle in the haystack searches (such as a customer lookup) along with cybersecurity and log search use cases (such as when an analyst seeks to find the logs for a particular IP address).

Arming Data Analysts with ML

Many data analysts want to take advantage of the benefits of machine learning (ML) but are daunted by the complexity of ML frameworks. In response, some cloud data platform vendors have created SQL functions that use ML to detect patterns in data. When backed by a robust data processing engine, these ML functions make it easy to scale from one to millions of dimension-value combinations. In addition, data engineers can integrate calls to these functions into their data pipelines just as any other SQL function. Some examples of SQL functions with ML under-the-hood include the following:

- **Forecasting functions** allow data scientists to construct accurate time series forecasts with automated handling of seasonality, scaling, and other variables.
- **Anomaly detection functions** empower analysts to find outlier events that should be investigated for suspicious activity, along with unlikely situations that should be excluded from future analysis.

Developing AI Applications

You may start out using a cloud data platform for a traditional warehousing workload. As your volume of data grows, as your data analysts advance, and as you hire data scientists to join your team, you can start using the cloud data platform to store and

process artificial intelligence (AI)/ML workflows, train predictive models, and then put those models into production.

ML algorithms learn from data; the more data you provide, the more capable they become. A cloud data platform gives you one place to instantly access all relevant data for AI and ML workflows without complex data pipelines. It enables data science teams to store and process nearly limitless volumes of data at a progressively lower cost via powerful arrays of computers that can be scaled up and down at will. It unifies data security and data governance activities, fosters collaboration, and provides elastic scalability for data science and related analytic endeavors.

The most advanced cloud data platforms allow developers to deploy containerized data apps on accelerated computing infrastructure such as leading graphical processing units (GPUs), expanding the processing power that can be applied to these resource-intensive workloads. One popular application for these advanced processing scenarios is the ability to natively run large language models (LLMs) within the platform and access them through an associated marketplace. This arrangement allows cloud data platform customers to utilize these applications within their own accounts.


TIP

Although your data platform should be able to securely deploy and process non-SQL code — including Python, Java, and Scala — SQL remains the industry standard for querying data. As such, your cloud data platform's data warehousing workload should include innovative SQL tools for data management, data transformation, data integration, visualization, BI, and all types of analysis.

Automating Development, Deployment, and Monetization

As AI becomes a more important aspect of many of today's software development projects, a cloud data platform gives advanced data analysts — and data scientists — native tools to facilitate ML application development such as turning data and ML models into interactive applications. These platforms should work readily with popular open-source frameworks, tools, and languages, and include native libraries and functions that automate the data science life cycle. Some platforms even have out-of-the-box capabilities to turn Python scripts into web apps with no front-end development required.

Chapter 8
Four Steps for Getting Started with Cloud Data Warehousing

This chapter guides you through four key steps to choosing a cloud data warehouse for your organization.

Step 1: Evaluate Your Needs

Consider the nature of your data, the skills and tools already in place, your usage needs, your plans, and how a cloud data platform can take your business in new directions. Think beyond data warehousing (storing and analyzing data). Ideally, you want one integrated platform that enables many workloads, including data warehouses for analytics; data lakes for data exploration; data engineering for data ingestion and transformation; data science for developing predictive applications and machine learning (ML) models; data application development and operation; and data sharing for easily and securely sharing data among authorized users.

Step 2: Migrate or Start Fresh

Assess how much of your existing environment should migrate to the new data platform and what should be built from scratch. Defining strategy and goals, taking account of budget and resources to migrate, and understanding your data volume can help you make this decision. To better understand which approach is best for your organization, talk to the professional services team of the data platform you're considering. Your BI solutions, data visualization tools, data science libraries, and other software development tools must easily adapt to the new architecture.

Step 3: Calculate TCO

Select a vendor that allows you to pay for actual usage in per-second increments. Consumption-based pricing eliminates software license fees, reduces infrastructure costs, and minimizes maintenance so you can reallocate technology resources to higher-value business priorities. Plus, when it comes to minimizing TCO, don't overlook the value of productivity.

TIP

Don't overlook the savings possible with features such as scaling up and down dynamically in response to changing demand.

Step 4: Set Up a Proof of Concept

Request a POC from a prospective vendor with the general understanding that if the solution performs satisfactorily, you'll subscribe to the service.

A proof of concept (POC) tests a solution to determine how well it serves your needs and meets your success criteria. Request a POC from a prospective vendor with the general understanding that if the solution performs satisfactorily, you'll subscribe to the service. Obtaining first-hand experience via a POC will set you up for success with future data warehouse endeavors.

Transform data into valuable business intelligence

Forward-thinking organizations rely on powerful, easy-to-use, and fully-managed cloud data warehouses to put their data to work. This book shows you how to unlock your data potential on an innovative, cost-effective, and versatile cloud data platform that powers not only your data warehouse, but also many other data workloads. Additionally, you'll learn how to extend an existing data warehouse to take advantage of the latest cloud technologies.

Inside...

- Keep up with the shifting demands of data
- Five steps for getting started with cloud data warehousing
- Share data without copying or deleting
- Apply governance policies to protect data
- Enable data workloads across regions and clouds
- Real-world 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: 978-1-394-21162-3
Not For Resale

WILEY END USER LICENSE AGREEMENT

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