LEARNING MADE EASY

2nd Snowflake Special Edition

Cloud Data Platforms

for dummies®
A Wiley Brand

Why a cloud data platform is crucial

How it accelerates data sharing and collaboration

How to choose a modern cloud data platform

Brought to you by:

snowflake®

David Baum

About Snowflake

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

Cloud Data Platforms

2nd Snowflake Special Edition

by David Baum

for dummies®
A Wiley Brand

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

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

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

ISBN 978-1-119-87548-2 (pbk); ISBN 978-1-119-87549-9 (ebk)

Publisher's Acknowledgments

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

Development Editor: Brian Walls
Project Manager: Jen Bingham
Acquisitions Editor: Ashley Coffey
Editorial Manager: Rev Mengle
Business Development Representative: Molly Daugherty
Content Refinement Specialist: Tamilmani Varadharaj
Snowflake Contributors Team: Vincent Morello, Alex Gutow, Kent Graziano, and Leslie Steere

Table of Contents

Introduction

Data analysts, data scientists, data engineers, and data application developers influence critical functions throughout the enterprise: sales, finance, supply chain, and much more. But they often work in isolation and must contend with trying to access a vast landscape of data silos.

This disparity stymies what's possible for any organization that wants to serve its customers and advance its business with data-driven insights and decisions. According to a 2021 study by Forrester Consulting titled "Unveiling Data Challenges Afflicting Businesses Around the World," 71 percent of businesses are gathering data faster than they can analyze and use it; 66 percent say they constantly need more data than their current capabilities can provide; and 84 percent claim to have significant problems with non-optimized data systems, partly due to high storage costs, outdated IT infrastructure, and manual or slow data management processes.

A well-architected and easy-to-use cloud data platform can resolve these problems by providing a single source for all your data. The platform should enable instant and near-infinite scale and concurrency of data workloads, including data pipelines, business intelligence, predictive analytics, and machine learning. As a result, users should enjoy a seamless experience, even when their data spans multiple clouds and regions. The platform should also streamline developing and delivering data applications — opening new revenue streams and creating new business models. Finally, a modern cloud data platform should give you the capability to share and monetize data across a broad business ecosystem instantly and securely.

About This Book

This book explains how to establish a modern cloud data platform that handles many types of data without incurring the excessive cost and complexity inherent in traditional data management solutions. Read on to learn how to:

- Standardize on a fully managed, usage-based data platform that supports multiple data types.
- Empower your data professionals to extract value from data in ways not possible before.
- Take advantage of baked-in data security, governance, and resiliency that spans regions and clouds.
- Efficiently access, share, and monetize data without copying or manually moving data from one environment to another.
- Implement new and changing architectural patterns such as a data mesh or a hybrid data warehouse/data lake with a single, flexible platform.

Icons Used in This Book

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

TIP

Guidance on better ways to use a cloud data platform in your organization.

REMEMBER

Concepts worth remembering as you immerse yourself in understanding cloud data platforms.

CASE STUDY

Case studies about organizations using cloud data platforms to transform how they understand their customers and their businesses.

Beyond the Book

If you like what you read in this book, visit `www.snowflake.com` to access a free trial of Snowflake's Data Cloud, obtain details about plans and pricing, view webinars, access detailed documentation, or get in touch with a member of the Snowflake team.

IN THIS CHAPTER

- Tracking the cloud data platform's history
- Defining the cloud data platform
- Introducing basic architectural tenets
- Understanding the need for and benefits of a cloud data platform

Chapter 1
Getting Up to Speed with Cloud Data Platforms

Over the last four decades, the software industry has produced various solutions for storing, processing, and analyzing data. These solutions made it possible to work with traditional forms of data and newer data types generated from websites, mobile devices, Internet of Things (IoT) devices, and data generated from other more recent technologies. Some of the new solutions were designed to democratize access to data for the business community, which has gradually moved data and analytics from the enterprise back office to frontline workers and the executive suite.

The business world has learned how to put some of this data to work in productive new ways, but many on-premises and legacy cloud platforms weren't architected for the variety and dynamics of today's data. Nor can those systems help you solve modern operational needs, such as providing a single experience across major clouds and securely sharing data globally.

Many software vendors have simply migrated their on-premises solutions to the cloud. For the most part, these first-generation cloud solutions provided better price and performance than their

on-premises cousins. However, because they weren't built from the ground up for the cloud, they struggled to take full advantage of the cloud's near-unlimited scalability and performance.

The industry has learned from the benefits and drawbacks of these solutions and carried that knowledge forward. Each solution was a stepping stone and solved an important problem. Yet, transforming those stepping stones into complete end-to-end offerings that enable organizations to deliver real value from their data continues to be a challenge.

REMEMBER

Forward-looking organizations now seek a powerful, interoperable, and fully managed *cloud data platform* that guarantees scale, performance, and concurrency — a platform that simultaneously supports analytics, data science, data engineering, and data application development, along with secure ways to share and consume shared data globally from within a single, cohesive solution.

Why You Need a Cloud Data Platform

Whatever industry or market you operate in, learning how to use your data easily and securely in a multitude of ways will determine how you run your business and how you address current and future market opportunities. A modern cloud data platform should easily enable you to marshal a single copy of your data for everybody to use simultaneously, and deliver near-unlimited bandwidth for analyzing data, sharing data, building data applications, and pursuing data science initiatives. Additionally, a modern cloud data platform should make your business users more efficient and help your IT team step away from tedious data administration, so everybody can focus on delivering great experiences with your data.

The most advanced cloud data platforms should enable instant and near-infinite elasticity, delivered as a service with consistent functionality across multiple regions and clouds. And it should allow your organization's business units and its business partners and customers to share governed data securely without having to copy the data. This versatile architecture should simplify near-instant data sharing within and between organizations directly or via a data marketplace, and minimize governance and compliance

issues by allowing everyone to rally around a single, sanctioned copy of the data.

Defining the Requirements of Modern Cloud Data Platforms

Whether architecting data pipelines, creating data science models, sharing data locally or globally, or performing many other data-intensive tasks, a modern cloud data platform must support the many ways your organization uses data. It must deliver a superset of capabilities to replace outdated systems, such as legacy data warehouses and siloed data lakes, and supply a versatile foundation for developing new data applications, building and deploying machine learning models, driving powerful insights, and simplifying the creation of complex data pipelines. Furthermore, the platform must facilitate advanced data sharing relationships and allow you to easily access commercial data sets and data services within today's expanding data marketplaces.

REMEMBER

Most importantly, your cloud data platform must take full advantage of the true benefits of the cloud, with an architecture based on three key elements (see Figure 1-1).

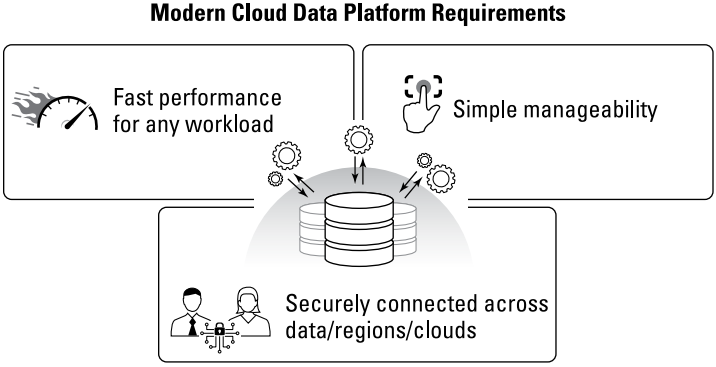


FIGURE 1-1: The fundamental elements of a modern cloud data platform.

Introducing the Architecture of Cloud Data Platforms

To best satisfy the requirements of a modern cloud data platform, the platform should be built on a modern *multi-cluster, shared data architecture,* in which compute, storage, and services are separate and can be scaled independently to leverage all the resources of the cloud (see the "Essential Architecture" sidebar). This architecture allows a near-limitless number of users to 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 properly architected cloud data platform offers the scale, flexibility, security, and ease of use that large and emerging organizations require. End-to-end platform services should automate everything from data storage and processing to transaction management, security, governance, and *metadata* (data about the data) management — simplifying collaboration and enforcing data quality.


TIP

Ideally, this architecture should be *cross-cloud,* providing a consistent layer of services across regions of a single public cloud provider and between major cloud providers (see Figure 1-2).

ESSENTIAL ARCHITECTURE

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

- **Storage:** A single place for structured, semi-structured, and unstructured data types
- **Compute:** Independent compute clusters dedicated to each workload to eradicate contention for resources
- **Services:** A common services layer that provides a unified experience by enforcing consistent security, propagating metadata, optimizing queries, and performing other essential data management tasks

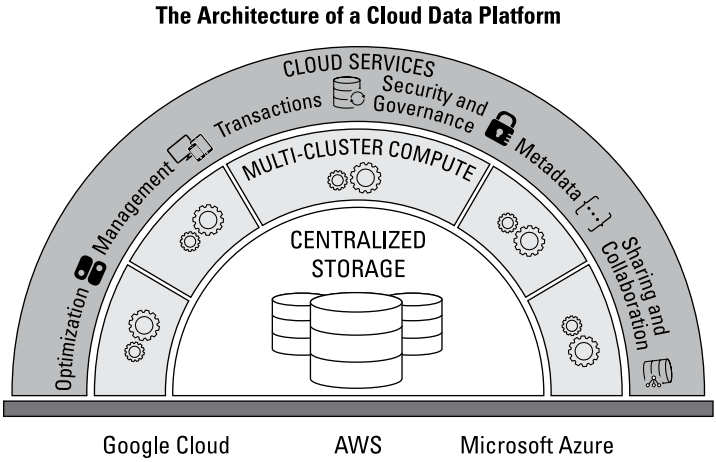


FIGURE 1-2: A modern cloud data platform should seamlessly operate across multiple clouds and apply a consistent set of data management services to many types of modern data workloads.

Built on versatile binary large object (BLOB) storage, 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* should process enormous quantities of data with maximum speed and efficiency. You should be able to easily specify the number of dedicated clusters you want to use for each workload and have the option to let the service scale automatically.

The *services layer* should coordinate transactions across all workloads and enable data loading and querying activities to happen concurrently. When each workload has its own dedicated compute resources, simultaneous operations can run in tandem, yet each operation can perform as needed.

Staying in Front of Important Trends

A cloud data platform should help you take advantage of several important technology trends that have arisen as organizations learn to leverage their data fully:

- **Advanced prescriptive and predictive analytics:** Whereas traditional analytic systems are reactive and backward-looking, predictive and prescriptive systems understand the present state or peer into the future. They recommend a specific course of action by considering dynamically shifting variables, such as moment-to-moment sales during a retail promotion or campaign. Once data scientists identify the correct algorithms and train the machine learning models, the systems predict outcomes and prescribe a course of action on their own — and they get smarter over time.
- **The opportunity to create new data applications:** A cloud data platform should make data application development more accessible not just for traditional technology companies but also for any company that sees the opportunity to offer data-driven products and services to its customers.
- **Support for modern data patterns and paradigms:** The ability to leverage new architectural frameworks beyond data lakes and data warehouses, such as a hybrid lake-warehouse or *data mesh* — a decentralized method of data management that assigns responsibility for data to the business teams that are closest to that data. Rather than one monolithic system under the auspices of a centralized IT department, a data mesh extends ownership to business experts from throughout the organization. Each business team leverages its domain knowledge to create data pipelines, catalog data, uphold data privacy mandates, and ensure data quality.
- **Easy, pervasive, and secure data sharing:** A cloud data platform should enable organizations to establish one-to-one, one-to-many, and many-to-many relationships to share and exchange data in new and imaginative ways. Secure, governed access to a single source of data not only makes internal teams more efficient but also facilitates collaboration among business partners, customers, and other constituents.
- **The rise of global data networks:** In every industry, immense data-sharing networks, exchanges, and marketplaces have emerged, propelling a growing data economy and motivating business leaders to examine new data sharing possibilities. A cloud data platform should enable these networks with almost none of the cost, complex procurement cycles, and delays that have plagued traditional exchanges and other types of data sharing.

IN THIS CHAPTER

- Understanding the problems with traditional data management approaches
- Forming a new vision for data platforms
- Acknowledging the limitations of data warehouses and data lakes
- Reviewing the advantages of a unified cloud data platform

Chapter 2
Leveraging the Exponential Growth and Diversity of Data

First-generation cloud data platforms can't keep up with the nonstop creation, acquisition, storage, analysis, and sharing of today's diverse data sets. Much of the data is semi-structured or unstructured, which means it doesn't fit neatly into the traditional data warehouse, which first emerged more than 40 years ago. Additionally, some data types, such as images and audio files, are wholly unstructured and must be maintained as binary large objects (BLOBs) within an object-based storage system that doesn't conform to traditional data management practices.

UNDERSTANDING DATA TYPES

Most data can be grouped into three basic categories:

- Structured data (customer names, dates, addresses, order history, product information, and so forth). This data type is generally maintained in a neat, predictable, and orderly form, such as tables in a relational database or the rows and columns in a spreadsheet.
- Semi-structured data (web data stored as JavaScript Object Notation [JSON] files; comma-separated value [.CSV] files; tab-delimited text files; and data stored in a markup language, such as Extensible Markup Language [XML]). These data types don't conform to traditional structured data standards but contain tags or other types of markup that identify individual, 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.

Examining the Impact of Data Silos

The value of a properly architected cloud data platform can be summed up in one word: simplicity. Many organizations have established unique solutions for each type of data and each type of workload: a data lake to explore potentially valuable raw and semi-structured data as a prelude to data science initiatives, a data warehouse for SQL-based operational reporting, or an object storage system to manage unstructured video and image data.

They have also implemented specialized extract, transform, and load (ETL) tools to rationalize different types of data into common formats and set up data pipelines to orchestrate data movement among databases and computing platforms. As a result, each type of data lands in a unique system, designed and modeled for particular needs.

Multiple disconnected silos can quickly become a maintenance and governance nightmare as users attempt to copy, move, transform, and combine data to accommodate unique requirements.

Furthermore, many legacy systems don't have the architectural flexibility to simultaneously work with structured, semi-structured, and unstructured data and support the multitude of other workloads needed to derive value, such as data engineering pipelines and machine learning models.

These limitations motivated the formation of data lakes 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 highly 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 originates from dozens or even hundreds 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.

REMEMBER

Whether data from weblogs, Internet of Things (IoT) data from equipment sensors, or social media data, the volume and complexity of these semi-structured and unstructured data sources can make obtaining insights from a conventional data warehouse or data lake difficult. A modern cloud data platform can resolve these limitations by storing all the data within a single, easy-to-manage system with features that far supersede the legacy paradigms and technologies (see Figure 2-1).

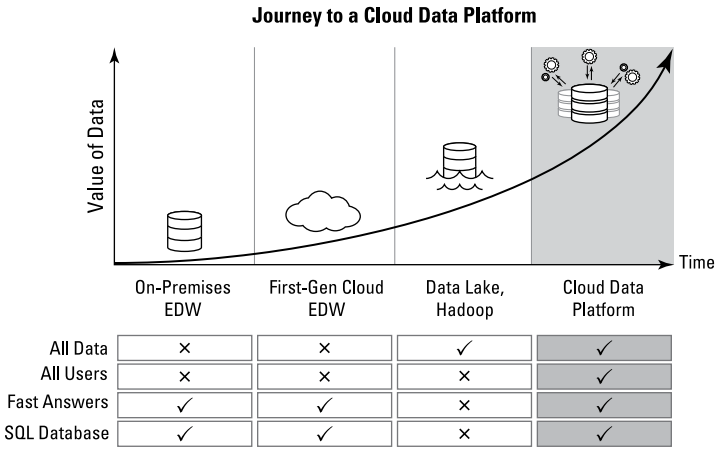


FIGURE 2-1: A cloud data platform combines the best of enterprise data warehouses, modern data lakes, object storage systems, and cloud capabilities to handle many types of data and workloads.

Understanding Problems with "Stitched Together" Platforms

Clearly, the cloud is a boon to data-intensive projects. But not all cloud data platforms have the same pedigree. Some are built on a cohesive architecture that takes full advantage of modern cloud infrastructure and features inherent integration among all platform services. Others represent "ecosystems" — dozens or even hundreds of "best of breed" services that weren't initially designed to work together.

For example, some cloud ecosystems allow you to select from hundreds of services for acquiring, storing, processing, and analyzing unique types of data. However, each service uses a different engine with its own access requirements, maintenance procedures, and learning curve. It's up to you to figure out how to make them all work together. If you don't, you will quickly find yourself confronting some of the same data silo and data access challenges you encountered in the on-premises world.

Consider a marketing team that wants to analyze customer buying behavior by geographic location and then feed the results to a data science team to create customized purchase recommendations. Each team will have to use different tools and services for each type of operation, such as feature engineering, data visualization, and ad hoc analytics. First, the data engineering team might create a data pipeline that gathers web interaction data and turns raw latitude and longitude coordinates into ZIP codes. They may use a specific tool to prepare data and load the data into a repository. After that, the marketing team might use a business intelligence service to submit queries and visualize the results via dashboards, allowing the team to associate certain types of behavior with certain users and regions. Finally, the data science team may use a complementary machine learning service to build and train a model that predicts user behavior and offers special discounts.

Each unique activity requires a unique set of tools and may require copying, extracting, or moving the data. The customer must figure out how to stitch it all together because these systems don't naturally integrate.

SOARING TO NEW HEIGHTS WITH A CLOUD DATA PLATFORM

CASE STUDY

The spirit of innovation drives nearly every aspect of the business for JetBlue, a leading airline carrier based in the U.S. In that spirit, JetBlue's data scientist and machine learning engineers use a cloud data platform, because it gives them a one-stop-shop for all their data needs. Airlines run on razor-thin margins. The data science team uses the cloud data platform to discover cost efficiencies, develop great customer experiences, and promote competitive fares, all of which boosts revenue. Data is available 24/7, which helps JetBlue maintain business continuity throughout the organization. Dynamic data masking allows the airline to control access to data based on roles. Near-real time reporting enables analysts to build dashboards that allow the operations team to make decisions as situations occur.

The data science team plans to use the cloud data platform to build better fuel prediction models. By combining internal data with external sources, such as air traffic control and weather, they can develop reports and run analyses that were not possible with their traditional data management solution.

JetBlue also uses the cloud data platform to share data with external partners. In two minutes and with only a few clicks, the data engineering team can create a secure data sharing infrastructure that formerly would have taken months of planning and weeks of development.

As JetBlue expands beyond its domestic roots, analysts can use the knowledge they have gained to craft unique experiences for new customers in new locales. As Ben Singleton, director of data science and analytics at JetBlue, said, "We like to say that we're a customer service company that just happens to fly planes. Now it almost seems as though we're also a technology company that happens to fly planes. The cloud data platform is a key part of making that happen."

Reviewing the Advantages of a Unified Data Platform

Today's organizations want an easier way to cost-effectively load, transform, integrate, and analyze unlimited amounts of structured, semi-structured, and unstructured data, in their native formats, in a versatile data platform. They want to simplify and

democratize access to that data, automate routine data management activities, efficiently govern the data, and support a broad range of data processing and analytics workloads. And they want to do all this *in one place,* so they can easily obtain and share all types of insights from all their data.

A cloud data platform will dramatically simplify your infrastructure by creating a single place for many types of data and data workloads. For example, centralizing data reduces the number of stages the data needs to move through before it becomes actionable, eliminating the need for complex data pipeline tools. Reducing the wait time for data makes it possible for users to obtain the data and insights they need, when they need them, so they can immediately spot business opportunities and address pressing issues.

REMEMBER

A cloud data platform should simplify the storage, transformation, integration, management, security, and analysis of all types of data. It should also streamline how diverse teams *share* data to collaborate on a common data set without maintaining multiple data copies or moving it from place to place. Consistent data governance makes it easier to enforce data-access restrictions, dictating who can see what data. Having these controls in place improves data security and reduces risk, so all members of an organization can work in concert to boost revenue, improve efficiency, and reveal new and disruptive opportunities.

IN THIS CHAPTER

- Taking stock of your data and analytic needs
- Staying current with the latest functionality
- Distinguishing between "cloud washed" and "cloud built" data platforms
- Reviewing the advantages of a fully managed service
- Unleashing a zero-maintenance platform

Chapter 3
Selecting a Modern, Easy-to-Use Platform

Organizations outgrow their existing data platforms for a variety of reasons. In many instances, limitations surface in response to competitive threats that require the business to acquire new types of data and experiment with new data workloads. For example, a data science team may set out to create a predictive analytics model that helps the sales team mitigate customer churn. The success of this sales initiative depends on the capability to access and iterate over the right data that best describes customer behavior.

One new venture leads to another. In this case, based on what the sales team learns about customer churn, the ecommerce team may realize it needs to simplify how customers navigate one of the company's key websites. To do this properly, analysts must look closely at the website traffic — to capture and analyze clickstream data. This brings in another massive influx of raw, semi-structured data.

Meanwhile, the support team wants to study social media posts to discern trends, issues, and attitudes within the customer base.

This data arrives as JavaScript Object Notation (JSON) in a semi-structured format. Analysts want to visualize the analysis of this data in conjunction with audio transcripts of customer support calls and some enterprise resource planning (ERP) transactions stored in a relational database, including historical data about sales, service, and purchase history.

Finally, another division wants to display these purchase patterns as data points on a digital map. This requires new data from a geographic information system. Traditional data platforms can't keep up with the latest data engineering, data science, data sharing, and other capabilities organizations need to acquire and harness this new data.

Reviewing Your Data Needs

Business scenarios like these can cause an organization to look for a more modern and versatile data platform (see Figure 3-1). Consider your own needs. You may have a data platform or data management system that works well for a certain type of data, but you want to take on new business projects that require the analysis, visualization, modeling, or sharing of new data types. Or perhaps you want to rethink your data acquisition strategy — to engineer better methods for acquiring data into your platform.

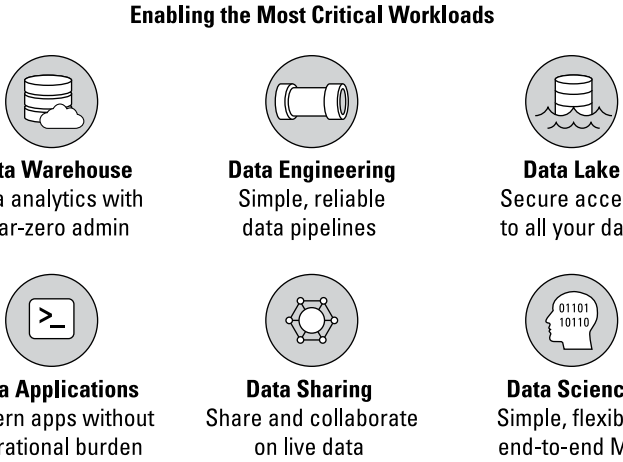


FIGURE 3-1: A modern cloud data platform should be powerful, flexible, and extensible to handle your most important data workloads.

TIP

While you gather additional data and the value of that data grows, you may want to monetize that data via a data marketplace to turn it into a strategic business asset. A modern cloud data platform should provide seamless access to a cloud data marketplace.

Leveraging External Data

Organizations with complex IT environments and a diverse data landscape can use a cloud data platform to leverage their data without importing or exporting data from external repositories. Data from various locations can be governed by a common set of services for security, identity management, transaction management, and other functions. These universal attributes pertain to data stored in the platform itself and data stored in external tables, such as an object store from one of the public cloud providers.

What are the advantages of this approach? First, all users have a single interface for viewing and managing that data. Second, in addition to the primary data store, the platform allows you to access, manage, and use data in *external tables* (read-only tables that reside in external repositories and can be used for query and join operations) just as easily as you can access it from the main platform — and with exceptional performance. Finally, you can leave data in an existing database or object store yet apply universal controls. This allows you to simplify your data environment by standardizing on a single cohesive system.

Distinguishing Between Cloud-Washed and Cloud-Built

Not all data platforms have the same pedigrees. Many began their lives as on-premises solutions or toolkits and were later ported to the cloud. As opposed to these *cloud-washed* solutions, *cloud-built* platforms have been designed first and foremost for the cloud. *Cloud-built* means created from the start to take advantage of the cloud, with each cloud platform component designed to complement the others.

TIP

To ensure you obtain superior, cloud-built capabilities, ask your cloud data platform vendor these questions:

- Does the platform completely separate but logically integrate storage and compute resources and services and scale them independently, maximizing performance and minimizing cost?
- Does it easily handle a near-infinite number of simultaneous workloads (concurrency) without degrading performance or forcing users to contend for a finite set of resources?
- Does the platform permit one-to-one, one-to-many, and many-to-many data sharing relationships without requiring people to copy or move the data?
- Does it ensure a seamless experience across regions and clouds?
- Does it facilitate collaboration by data engineers, data analysts, data scientists, and other authorized users across a single, governed data set?
- Can the platform perform all this automatically without the complexity, expense, and effort of manually tuning and securing the system?

Insisting on a Fully Managed Cloud Service

All organizations depend on data, but none wants to be bogged down with tedious database maintenance, system management, and IT administration tasks. In response, a rapidly growing industry of software vendors has emerged, offering partially or wholly managed cloud applications and other cloud solutions.

However, not all cloud services are created equal. Most cloud vendors claim to offer "managed services," but you must dig a little deeper to discover how much automation they actually provide. Ideally, all aspects of managing, updating, securing, governing, and administering your data platform should be transparent to the business community and require no extra effort by your IT

professionals. Furthermore, this level of automation should be holistic across clouds, regions, and teams, as Chapter 7 describes.

When it comes to software updates, you should always have the latest functionality, and you should never have to endure a lengthy, manual upgrade process. You, the customer, should not have to plan for updates, experience downtime, or modify your installation in any way. In the background, the cloud data platform provider should take care of all administrative tasks related to storage, encryption, table structure, query optimization, and metadata management in order to eliminate manual tasks.

By contrast, if you layer your database and other software services on infrastructure from one of the public cloud providers, you're responsible for integrating, managing, and updating all the components.

TIP

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

- Do you have to set up and manage data replication, 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 you to keep the software platform up to date?
- Does the service automatically encrypt all your data at rest and in motion with industry-standard encryption, or do you have to set up and apply encryption to the data manually? Does the encryption hinder query performance?
- Does the service scale up and out instantaneously and elastically and then release extra compute or storage resources when they are no longer in use? Or, do you 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 you have to set up change data capture (CDC) procedures to keep multiple databases in sync, or does the vendor handle that for you?

- Do you need to partition data, tune SQL queries, and optimize performance, or does the platform handle this automatically?


TIP

The best cloud data platforms are fully managed services: You click a button, and a database appears. After that, all management, administration, scaling, tuning, and data security should happen automatically in the background.

Ensuring Ease of Use for the Business

In addition to automating these common IT management tasks, a modern cloud data platform should be easy for all people to use, from business analysts to application developers to data scientists. All users should be able to focus on maximizing the potential of their data rather than managing the data platform.

With some cloud data platforms, IT is responsible for provisioning new resources and managing them. 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.

Monitoring the Costs of Cloud Usage

To ensure you don't pay for more capacity than you need, your cloud data platform should also offer usage-based pricing in conjunction with built-in resource monitoring and management features that provide complete transparency into usage and billing, with granular chargeback capabilities tied to individual budgets. Integrated usage tracking by time or by accumulated use allows you to administer cost allocations and chargebacks easily.

Finally, the platform should employ safeguards to eliminate runaway usage. For example, *auto suspend* and *auto resume* features automatically start and stop resource accounting when the platform isn't processing data. You should also be able to set specific time-out periods for each type of workload.

IN THIS CHAPTER

- Recognizing how today's workers use data
- Democratizing data access and collaboration
- Supporting new architectural paradigms and access patterns

Chapter 4
Accommodating Users, Workloads, and Access Patterns

Today, nearly every worker consumes data on some level. Everybody is a *data consumer,* but each person has different data requirements.

For example, *managers, supervisors, and line-of-business (LOB) workers* generally want data delivered within the context of the business processes they use daily, and in a form they can readily understand. They want to visualize data via intuitive charts and graphs, ideally displayed via easy-to-use apps on computers, tablets, and phones.

Analysts are better equipped to sort, summarize, and manipulate data. Many have been trained to use business intelligence apps, load data into spreadsheets, create pivot tables, and generate custom reports. They're comfortable creating data models, joining tables, and imposing a sensible structure on a data set. They're familiar with using SQL to create and issue queries.

Data scientists leverage massive data sets to build, train, and deploy machine learning models. They consolidate, cleanse, and transform data to fuel their models. To deliver new value and unlock new business opportunities, they create predictive and prescriptive analytics.

Data engineers build data pipelines and use various tools to populate databases in real time or batch mode and refresh those databases at periodic intervals. They are also responsible for cleansing data to eliminate duplications, correct inaccuracies, and resolve inconsistencies, often by incorporating input from analysts and LOB managers. Finally, data engineers handle data transformation projects, such as converting data from one format or structure into another format or structure.

Software developers and DevOps professionals develop and deploy data-driven applications for internal use and to create products for external customers. These technology professionals collect data and apply it to unique business problems. They also collect, analyze, and maintain the data the applications generate.

Data architects are tasked with delivering the right tools and infrastructure to make all these teams productive while helping to establish and enforce data security and data governance needs.

Democratizing Data Access and Collaboration

All of these workers want to access relevant data as soon as it is needed — to obtain the right data at the right time. To make this possible, a cloud data platform must be optimized to provide near-real time access to an ever-growing collection of diverse data. Business professionals, data analysts, data engineers, data scientists, and application developers need to confidently work with the same single source of data to ensure consistent outcomes, and collaborating on this unified data set should be easy.

REMEMBER

As organizations enable this level of collaboration, they need to find ways to eliminate duplicate efforts. The right cloud data platform makes this experience possible by alleviating disconnected data silos and discouraging data copying. Users should be able to leverage the data simultaneously without importing or exporting that data from one system to another.

This is a sharp contrast from legacy data platforms, which are restricted by a linear data processing architecture. These older platforms are limited in the scale and number of multiple workloads they can run in parallel, leading to long wait times or failed jobs for resources and data-driven insights. Furthermore, because they're typically optimized for a particular type of user or workload, organizations often end up with unique data silos for each unique situation.

Figure 4-1 shows that a flexible cloud data platform accommodates all these users and workloads. It supports advanced analytics and machine learning along with traditional business intelligence (BI) and data visualization. It offers all the capabilities that organizations derive from data warehouses and data lakes. It also facilitates modern data sharing relationships and empowers developers to create and maintain data applications.

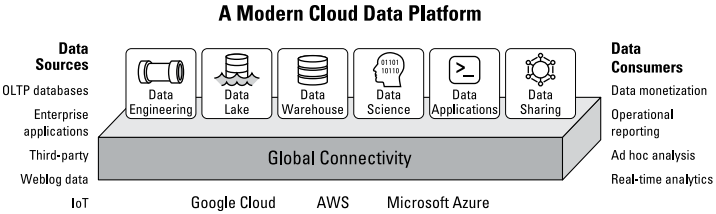


FIGURE 4-1: A cloud data platform should handle any data source and data workload and serve data consumers of all levels and needs.

Supporting New Architectural Patterns

New types of data often necessitate new architectural patterns, some of which you can't foresee in advance. A modern cloud data platform enables these architectural patterns to change and evolve according to your business needs. For example, a traditional data warehouse may evolve into a hybrid pattern that combines the best attributes of data warehouses and data lakes. Domain-specific data marts might evolve into a more manageable, better-governed *data mesh*. You need a cloud data platform to facilitate all these patterns based on some key architectural principles described below.

Empowering data teams with a data mesh

A data mesh is a design pattern for organizing data and helping domain teams gain access to that data. The basic premise is to divide large, monolithic data architectures into smaller functional domains, each managed by a dedicated team. The teams closest to the data are responsible for developing and managing the data products they use and that serve the business, including building and maintaining the data pipelines, implementing governance policies, and extending access to others who can benefit from that access.

This new architectural paradigm arose to remedy the limitations, delays, and expertise required of traditional data warehouses and data lakes, which tend to combine lots of data from lots of departments into a monolithic system managed by a central team.

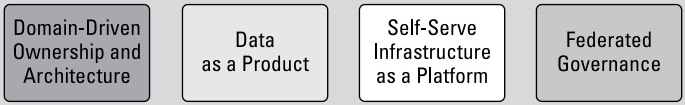
DATA MESH ARCHITECTURAL PRINCIPLES

Four primary principles underlie today's emerging data mesh architectures that help users gain the most value from their data (see the accompanying figure, which shows the four core principles of a data mesh architecture):

- **Principle 1: Domain-centric ownership and architecture.** A data mesh shifts the responsibility of data ownership into the hands of specialized teams. Domain teams control all aspects of the data as well as create and share analytics with other teams. From ensuring they have the right sources to building and maintaining data pipelines to enforcing data quality, the people who best know the data take charge of putting it to work.
- **Principle 2: Data as a product.** Domain teams aren't just responsible for the data; they are also responsible for developing and maintaining useful data products. For example, a supply chain team might create an inventory data product that a marketing team can tap into to develop new discount campaigns. Likewise, a finance team can design and share revenue products with data science teams.

- **Principle 3: Self-service infrastructure as a platform.** A data mesh eliminates complex technologies and the need for niche skills. The right cloud data platform supports a consistent set of tools and capabilities that allow domain teams to build, serve, and utilize data products without getting bogged down managing hardware and software or scaling infrastructure.
- **Principle 4: Federated governance.** Strong access controls and data protections are implemented by each domain team, mitigating risks while enforcing data privacy and compliance as new products are developed for sharing data. These governance policies should be centrally managed and interoperable across the business.

Data Mesh Architecture



Enhancing new paradigms with a cloud data platform

Even when you follow these modern design principles, a data mesh runs the risk of turning into domain-specific silos. A cloud data platform allows data teams to leverage relevant data when they need it without creating new silos or increasing operational complexity. The entire organization can securely share a single copy of data that all authorized users can discover and access immediately. People throughout the enterprise can easily access and query the data without having to move or copy it. All data is live and instantly accessible, and all updates are automatically propagated to other teams.

The best cloud data platforms can connect domain teams across regions and clouds, as Chapter 7 discusses. Each domain team can operate locally, running on its preferred cloud or region. Whether the teams work in SQL, Java, Scala, or Python — or utilize a mix of languages and techniques — the cloud data platform should easily support them. They can share data and data products as easily with a domain team on the other side of the world as they can with a team in the same office. And the organization can replicate data between regions and between multiple public clouds to operate without disruption, ensuring business continuity, allowing for regional data sovereignty differences, and upholding regulatory protections.

TIP

When anchored by a modern cloud data platform, a data mesh can incorporate many types of data (structured, semi-structured, and unstructured) and file formats, and support access to external data for comprehensive coverage of the data landscape. IT teams don't need to worry about provisioning, maintenance, upgrades, or downtime. Domain teams operate as distinct units and can scale their data products to other teams, requiring no infrastructure expertise or database tuning.

BETTER DATA ENGINEERING YIELDS BETTER INSIGHTS

CASE STUDY

Vimeo is a software-as-a-service (SaaS) company that provides professional-quality video for more than 200 million users. Vimeo ingests and analyzes large amounts of customer, marketing, and product-usage data to surface data-driven insights that support customer acquisition and upsell initiatives. Unfortunately, data engineering challenges and time-consuming system maintenance diverted attention from higher-impact initiatives.

Realizing the need for a new data environment, Vimeo subscribed to a cloud data platform to ingest fresh data directly from Vimeo's production databases. A Kafka connector simplifies the process of ingesting billions of streaming video events per day. These new data pipelines have reduced latency for reports that aggregate data from Salesforce, Amplitude, Google Analytics, and content delivery network (CDN) vendors.

The cloud data platform also increases Vimeo's ability to make data-driven decisions. Ingesting enriched data from no-code sales and marketing platform Openprise provides valuable insights about enterprise-level customers. Integrating with customer data platforms Singular and Simon Data enables a data enrichment process that helps marketers refine Vimeo's customer acquisition models. Best of all, Vimeo's data platform can support new data-driven initiatives.

Overcoming data engineering challenges has freed Vimeo's technical staff to focus on innovation and helped the company to reimagine its extract, transform, and load (ETL) practices. The cloud data platform features a multi-cluster shared data architecture that scales instantly to handle more data, users, and workloads. A near-zero maintenance infrastructure has improved uptime, reduced system administration, and eliminated concerns about stability. As a result, Vimeo can now run more queries, ingest more data, and create more business opportunities.

IN THIS CHAPTER

- Broadening analytics initiatives
- Creating more versatile data lakes
- Streamlining data engineering tasks
- Sharing and collaborating with data
- Developing new data applications
- Fostering the work of data scientists

Chapter 5
Using a Cloud Data Platform to Support Diverse Data Workloads

A cloud data platform should maximize the value of your data. It should bring together modern technologies for storing, sharing, and analyzing that data; creating modern data pipelines; building new data applications; and delivering cutting-edge data science and predictive analytics projects. A modern cloud data platform can power, scale, automate, and improve these important workloads.

Extending Beyond Data Warehouses and Data Lakes

A cloud data platform should establish a single source of data for a virtually limitless scaling of workloads and users. You should be able to use ANSI SQL to manipulate all data, including support for joins across data types and databases, as well as use modern programming languages, such as Java, Scala, and Python. The data platform should offer a superset of the best capabilities of

data warehouses, data lakes, and more. In addition, a cloud data platform should:

- **Simplify management,** eliminating administrative chores such as tuning queries, installing security patches, scaling workloads, and replicating data
- **Maximize data options,** allowing users to access near-limitless amounts of structured, unstructured, and semi-structured data (including JSON, XML, and AVRO) to build data applications, launch data science initiatives, and extract timely insights
- **Power all users and workloads,** enabling many concurrent users and multiple applications to simultaneously access the data without degrading performance
- **Minimize usage costs,** separately scaling storage and compute resources to facilitate instant, cost-efficient scalability and allowing users to pay only for what they use in per-second increments


TIP

These attributes make a cloud data platform an ideal architecture on which to deploy the best of a data lake and data warehouse in one solution. You can tap into the massive scale necessary to bring all data together without compromising on performance. Additionally, you can use the platform to augment and connect data siloed in other systems to accelerate data transformations and analytics. Having flexible access via SQL and other popular languages makes building data pipelines, running exploratory analytics, training machine learning (ML) models, and performing other data-intensive tasks easy for many types of users working across shared data.

Organizations with traditional data lakes can extend these assets by using a cloud data platform as the single source of data. Having a multi-cluster, shared data architecture yields dramatically better performance than traditional alternatives. Finally, when anchored by a cloud data platform, data can be more carefully governed, which Chapter 8 discusses.

Streamlining Data Engineering

Traditional data pipelines are often developed using legacy extract, transform, and load (ETL) procedures that may slow down or even fail as data volumes spike. They are often too rigid to accommodate evolving needs and dependencies, such as modifications to the data model; data cleansing requests from downstream users; or new data types, such as machine-generated data from Internet

of Things (IoT) systems, streaming data from social media feeds, JSON event data, and weblog data from Internet and mobile apps.

To accommodate newer forms of data and enable more timely analytics, modern data engineering workloads rely on the superior processing capabilities of a cloud data platform. With older data platforms, transformation jobs contend for resources with other workloads running on the same infrastructure. Modern cloud data platforms move the transformation process to the cloud, enabling superior scalability and elasticity. This allows data engineers to create data pipelines that extract and load raw data and transform it later, once the requirements are understood — a strategy known as ELT rather than ETL.

Thanks to the versatility of these modern data transformation workflows, data engineers can create stable, scalable data pipelines to incorporate all types of data, accommodate emerging business requirements, and use popular languages and tools. When based on a fully managed cloud service, these modern data pipelines automate many of the management, maintenance, and provisioning challenges of traditional pipeline infrastructure.

Sharing Data Easily and Securely

A cloud data platform should enable organizations to share slices of their data easily and receive shared data in a secure and governed way — without requiring constant data movement or manual updates to keep data current. Authorized members of a cloud ecosystem should be able to tap into live versions of the data. Rather than physically transferring data to internal or external consumers, the platform should enable instant access to governed portions of live data sets.

REMEMBER

This type of advanced data sharing encourages collaboration by making it easier to broadly share data across business units, with an ecosystem of business partners, and with other external organizations.

A cloud data platform also allows you to more easily monetize your data to create new revenue-generating products and services. Because data isn't copied or moved in these scenarios, you eliminate the cost, headache, and delays associated with traditional data exchanges and marketplaces, which deliver only stale subsets or "slices" of data that must be continually refreshed. Consult *Data Sharing For Dummies* for additional details.

Developing Data Applications

Today, nearly every company sees the value of leveraging data to develop new insights and share them with customers and partners, opening up new revenue streams and powering new lines of business. A cloud data platform masks DevOps complexity, so you can focus on creating innovative data applications. For example, a cloud data platform eliminates the need to build infrastructure and automatically handles provisioning, availability, tuning, data protection, and other operations. Developers can instantly spin up dedicated compute resources to support a near-unlimited number of concurrent users and workloads without requiring a dedicated engineering team to prepare the data. Operations and quality-assurance (QA) professionals can utilize DevOps workflows to:

- Create instant sandboxes with zero-copy cloning and isolated computing resources
- Access historical data or roll back to a previous version without manual backups
- Improve operational efficiency with built-in high availability, data durability, and disaster recovery utilities

Advancing Data Science

Data scientists need data to build and train ML models and predictive applications. The better the data, the better the outcomes. Finding, retrieving, consolidating, cleaning, and loading training data takes up an inordinate amount of a data scientist's time.

A modern cloud data platform should satisfy the entire data lifecycle of ML, artificial intelligence, data visualization, predictive/prescriptive analytics, and application development. It should consolidate data in one central location for easy development and flexible accessibility via a wide range of data science notebooks and AutoML tools. It should also natively support the most popular languages, including SQL, Java, and others. These capabilities enable data scientists to develop and deploy new models with less time spent on data preparation.

IN THIS CHAPTER

- Establishing a robust and efficient data sharing architecture
- Leveraging a data marketplace
- Controlling access to sensitive data while maximizing its usefulness
- Guaranteeing transactional integrity

Chapter 6
Sharing and Collaborating with Your Data

According to a 2020 Forrester Research report titled "The Insights Professional's Guide to External Data Sourcing," 47 percent of organizations currently commercialize their data, while 76 percent have launched, or plan to launch, initiatives for improving their external data sourcing. A cloud data platform should revolutionize these endeavors by easily enabling modern and secure data sharing without requiring organizations to move or copy the data.

This is in stark contrast to traditional data sharing approaches, in which *data providers* simply copy and send some or all of a primary data set to *data consumers*. Within these traditional data sharing scenarios, data is often copied via File Transfer Protocol (FTP) or an application programming interface (API) that links the two systems. In some instances, special data pipelines move the data via extract, transform, and load (ETL) procedures that extract data from the provider's database, transform it into a format suitable for consumption, and then load it into the consumer's database.

Newer data sharing methods use cloud storage services to stage data to a central location that authorized consumers can access.

However, within all these scenarios, disparities arise between the primary data set owned by the data provider and the secondary data set used by data consumers, requiring constant update procedures to keep the two versions in sync. Traditional data sharing methods are slow, cumbersome, costly, and create secondary data silos that quickly become dated or "stale," as Figure 6-1 illustrates.

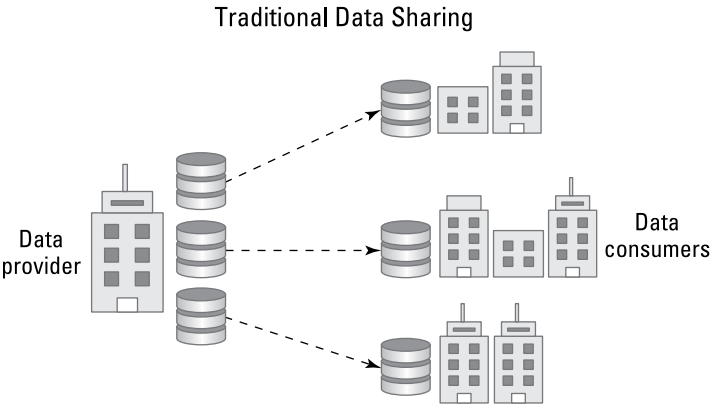


FIGURE 6-1: Traditional data sharing methods hinder organizations from extracting fresh insights from data that is always live and up to date.

Instead, a data provider can use a cloud data platform to provide data consumers access to live, read-only data that doesn't move via modern cloud data sharing. The data can be shared across cloud providers and regions without using ETL or other traditional procedures. The data is updated automatically — decreasing management overhead for both the data provider and data consumer.

When sharing live, read-only data, a data consumer can easily access and integrate the shared data set without changing the data provider's original version. When the provider updates the data set, the data consumer's read-only version is updated almost simultaneously (see Figure 6-2).

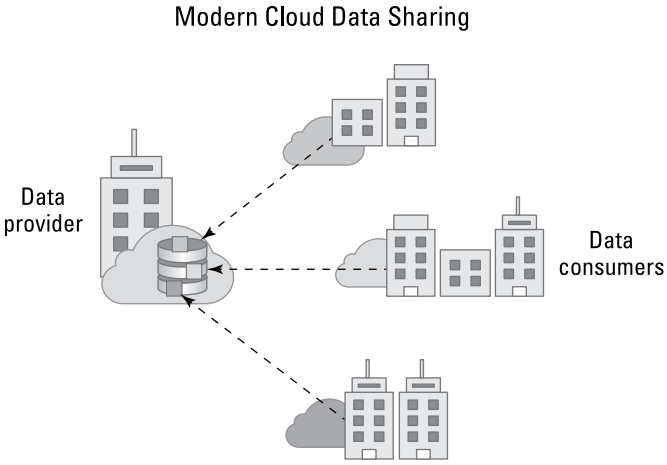


FIGURE 6-2: A modern cloud data platform enables live, governed data to be shared across clouds and regions without needing to move files across environments or create unnecessary copies.

Establishing a Robust Data Sharing Architecture

A modern cloud data platform should allow you to provide on-demand access to ready-to-use, live data inside a secure, governed environment. This will enable you to share data easily among multiple business units across your organization and seamlessly exchange data with your business partners, customers, and other entities within your business ecosystem. This all happens without copying or moving data, and with everybody leveraging the same single copy of data.

REMEMBER

Sharing data internally among departments and subsidiaries should be just as easy as sharing it externally with partners, suppliers, vendors, and even customers. With a modern cloud data platform, all database objects are centrally maintained and updated in conjunction with end-to-end security, governance, and metadata management services. As a result, you don't have to link applications, set up complex procedures, or use FTP to keep data current. And because data is shared rather than copied, no additional storage is required (see Figure 6-3).

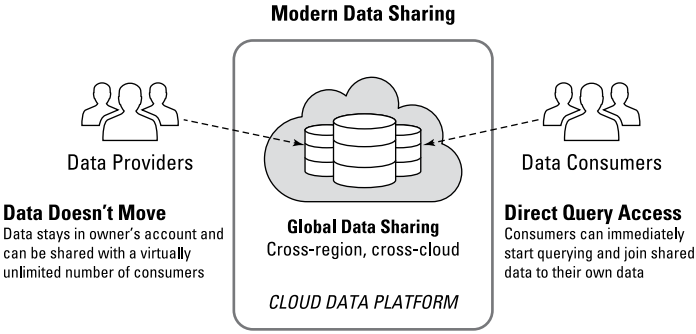


FIGURE 6-3: A cloud data platform streamlines data sharing between data providers and data consumers, even across multiple regions and clouds.

This modern data sharing architecture lets you share subsets or "slices" of your data. It also allows you to share business logic, such as user-defined functions (UDFs) written in multiple procedural languages. As with sharing data and metadata, these shared functions uphold previously defined governance and security controls.

Rather than physically transferring data, models, and functions to internal or external consumers, you can authorize those consumers with read-only access to a governed portion of a live data set, accessible via SQL and a variety of other languages or analytic tools. As a result, a near-infinite number of concurrent consumers can access shared data and logic without competing for resources. Additionally, performance is exceptional due to the cloud's near-limitless storage and compute resources.


TIP

A cloud data platform should facilitate modern data sharing by enabling authorized members of a cloud ecosystem to access live, read-only versions of the data. If you don't have to track data in multiple places, controlling what the data includes and updating the data becomes easy. So does monitoring who interacts with it.

Leveraging a Data Marketplace

Modern data sharing technology also sets the stage for collaborating and monetizing data via *data marketplaces* — online communities that facilitate the purchase and sale of data and data services. It's a burgeoning opportunity: Thousands of online marketplaces today link buyers and sellers. Typically, the data

provider handles data transformation, preparation, copying, and loading, while the marketplace oversees discovery, collaboration, licensing, and auditing. These are onerous tasks for the data provider, requiring complex data pipelines and constant update procedures that often leave the consumer with stale data. With a modern cloud data platform that replaces those manual marketplace tasks, data providers can share and monetize their data much more easily.

Some data providers share data. Others also share *data services* that put that data to work. For example, an organization might supplement its internal customer data with third-party data to better understand the age and income of groups that have purchased from its website. The same organization might subscribe to a data service that cross-references online purchase behavior with additional third-party demographic data, enabling a more personalized understanding of each group or segment.


TIP

A cloud data platform should make it easy to join a data marketplace or establish your own data exchange that enables an ecosystem of your business partners, for example, to share data and data services collaboratively. The platform should also include user-friendly search and discovery tools to make it easy for users to identify pertinent marketplace services and easy for data providers to promote their services. Thus, each marketplace participant can easily offer and acquire new data sets for exploration, analysis, and other tasks, and use a wide range of data services to add value to that data.

For example, a financial services company can examine ecommerce data sets to identify fraudulent transactions. 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. A logistics company might sell data about transportation patterns and shipping activity as an indication of economic trends.

A cloud data platform streamlines the entire process of acquiring, merging, enriching, and sharing data and data services. It's not just a network of organizations but also a network of content that includes data, logic, and applications. These capabilities are especially helpful to data scientists, who often acquire third-party data sets to expand their analyses and enrich their predictive models.

REMEMBER

With modern data sharing technology, a data provider can easily grant access to the data it wants to share with its intended marketplace consumers without managing cumbersome data pipelines. End-to-end security, multiparty governance, and metadata management services are systematically applied, even when the data consumers span multiple clouds. With updates made automatically, you don't have to link applications, set up file-sharing procedures, or frequently upload new data to keep data current.

Sharing Sensitive Data

If portions of a database table are subject to strict security and confidentiality policies, sharing the entire table would expose that sensitive data. Therefore, to minimize the chance for unauthorized access, your cloud data platform should offer flexible options for securing data.

For example, large companies may store aggregated advertising data in a *clean room* to help businesses better understand their advertising data. Data clean rooms have stringent privacy controls that don't allow businesses to view or pull any customer-level data and yet still infer trends from the aggregate data, such as how many customers purchased particular products after viewing an ad for those products.

Sometimes you need to fully mask or "lock down" the data; at other times, you need to anonymize certain fields, rows, or columns to allow people to analyze the data without seeing the sensitive elements. A cloud data platform allows organizations to use this aggregated data and combine it with their own without exposing sensitive user-level data. Choose a modern cloud data platform that allows data providers to easily control access to individual database tables with granular protections, such as secure views, which Chapter 8 discusses.

IN THIS CHAPTER

- Ensuring worldwide business continuity
- Implementing the right clouds for the right locales
- Establishing global data replication to ensure data protection and availability
- Complying with data sovereignty regulations
- Simplifying administration by using a single code base for multiple clouds

Chapter 7
Maximizing Availability and Business Continuity with a Cross-Cloud Strategy

Large organizations commonly rely on multiple on-premises data repositories while also storing data in one or more public clouds. This diverse software-solutions landscape invariably spawns diverse data sets, such as data warehouses populated with data from enterprise applications, data lakes for exploratory analysis, and a wide assortment of local databases, data marts, and operational data stores for local and departmental needs.

Organizations have been working for years to eliminate these silos — first arising from countless on-premises systems and now compounded by a plethora of cloud-based applications. As these organizations expand, they often become dependent on new sets of silos in various regions and across different clouds, making it difficult to use all data fully.

Furthermore, each public cloud provider has different levels of regional presence, and data sovereignty requirements may require organizations to keep data processing operations within the regions they serve, leading to even more silos. Each department and division within your organization may have unique requirements. Rather than demand that all business units use the same cloud provider, a multi-cloud strategy allows each unit to use the cloud that works best for that unit.

This is a strategic advantage for global companies because not all cloud providers offer the same services or operate in the geographic regions where your data and users reside. It's also useful if you acquire or merge with a company that has standardized on a cloud different from the one you're using, because it enables teams from various business units to collaborate without first undergoing a lengthy migration to a single, standard cloud.


TIP

Your cloud data platform should allow you to easily operate data workloads among multiple clouds and multiple regions within each cloud, so you can locate data where it makes sense and mix and match clouds as you see fit.

This type of deployment flexibility assists with geographic expansion, streamlines business development, improves availability, and allows you to use different cloud services in different regions — without wrestling with the unique nuances of administering each cloud. The cloud data platform should deploy the same code base that spans them all to deliver a consistent, unified experience regardless of region or cloud. This also enables you to host data seamlessly and securely while selecting the cloud options that best meet your needs.

Minimizing Administrative Chores with a Single Code Base

When working with multiple clouds, how do you ensure 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? What about tuning and scaling techniques on different clouds? Will your IT security experts have to deal with varying sets of rules on each cloud or work with multiple key management

systems to encrypt data? Will data engineers have to create unique pipelines? Will data scientists encounter obstacles when building machine learning models from multiple data sets?

Your cloud data platform should provide a unified experience across multiple cloud providers to ensure data management consistency and to simplify administrative operations. Abstracting the differences among clouds means you won't need to hire people with unique skill sets or maintain familiarity with each public cloud. Here are a few terms to be aware of:

- **Multi-cloud** means your platform operates on several clouds — Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform — and their regions.
- **Cross-cloud** means you can instantly and consistently access data from all these clouds and their regions and replicate and share data seamlessly between them.
- **Data replication** is the process of replicating data in more than one region or cloud. This can be to make data available to other parts of the business, ensure your business remains operational in the event of a failure or outage, or meet regulatory compliance requirements.

The platform experience should be the same no matter where your data resides, even as you uphold geo-residency requirements and comply with data sovereignty mandates. For example, analysts can query data housed in Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform, using the same procedures.

REMEMBER

The best cloud data platforms enforce data access, security, and governance policies that "follow the data," not only across regions but also across clouds. A shared metadata layer can define a cohesive set of network services to orchestrate data management and uphold all data protections and controls. Therefore, all users obtain consistent results, and all workloads enable consistent outcomes.

This is a boon to administrators because they don't have to learn each cloud provider's distinct data access and data governance policies. And because the data doesn't have to be moved among systems, administrators achieve stronger data security and privacy levels, with better end-to-end visibility for compliance. No matter where that data lives, no matter where it's being accessed,

administrators can easily control how the information is protected and ensure that all data-access constraints are consistently enforced.

This same logic pertains to data security and governance: There is no need to set up different policies for each cloud because the cloud data platform spans them all. Data stewards can set up all necessary roles, permissions, masks, and controls — irrespective of which clouds their teams use. All constraints, controls, and views will be consistently enforced.

These same cross-cloud administrative capabilities also simplify software development and maintenance activities for DevOps and DataOps teams. For example, rather than creating three versions of a software application for the three major clouds, a software-as-a-service (SaaS) provider can create one data app that runs on all of them.

Replicating Data to Improve Business Continuity

Business continuity involves planning for disruptions, whether those disruptions stem from an extreme weather event, a cyberattack, or an internal mishap. All businesses that depend on data need a business continuity strategy to mitigate these situations and a recovery plan to make sure the business can resume normal operations. An essential part of these continuity plans involves *data replication,* the process of automatically copying data to multiple locations to ensure it's protected and always available and automatically up to date in the event of a regional outage.

Ideally, data should be replicated to more than one region or cloud to ensure all parts of the business have the data they need, that the business can quickly resume normal operations, and that the enterprise complies with corporate compliance requirements governing the security and sustainability of data throughout its lifecycle.

Cross-cloud data replication is essential for business continuity and point-in-time consistency between primary and secondary sites. In case of a mishap, it enables you to instantly *failover* — switch to an up-to-date copy of the data in another region or

cloud — or restore previous versions of a table or database within a specified retention period. This strategy ensures that your business won't be disrupted, and you'll minimize data loss.

REMEMBER

Most importantly, data replication procedures should be *completely automated* to remove the risk of manual errors or delays. Furthermore, a complete data-retention and business continuity strategy should go beyond duplicating data within the same cloud region or zone. Instead, it should replicate that data among multiple availability zones for geographic redundancy and even between multiple clouds.

At first glance, this may appear to negate the centralized "single source of data all in one place" strategy advocated throughout this book. It doesn't. That's the beauty of automatic data replication within the context of a cloud data platform: Although your data is mastered, managed, and governed in only one location, by domain-centric data stewards and teams, it can be maintained in multiple availability zones or clouds for disaster recovery purposes. A modern cloud data platform needs to support this and apply end-to-end security, governance, and metadata management services. It's all synchronized by the platform, with centralized command and control based on these fundamental principles.

REMEMBER

A complete and modern cloud data platform should automatically replicate databases and keep them synchronized across regions and clouds, boosting availability, automating the failover of workloads, and guaranteeing instant access and recovery for databases of any size (see Figure 7-1).

Cross-Region and Cross-Cloud Replication

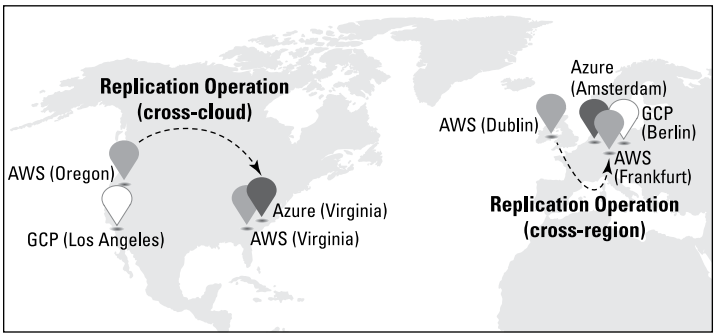


FIGURE 7-1: A modern cloud data platform automatically replicates databases and keeps them synchronized across regions and clouds.

Reacting Quickly to New Regulations

Cross-cloud deployment has become increasingly pertinent as data privacy regulations become more restrictive in Europe and elsewhere. In some instances, sudden and sweeping changes to data privacy laws may force you to reconsider which cloud provider you use.

Similarly, you may encounter a competitive situation that motivates you to move your operations off a particular cloud. Amazon, Google, and Microsoft all have extensive and diverse operations. How will you react if you have standardized on one of these cloud platforms, and the vendor later launches a new business or acquires a new company that presents a competitive threat? Rather than support that cloud vendor's business, you may opt to shift your data and workloads to a different cloud provider.

REMEMBER

A modern cloud data platform should make it easy to move your operations between clouds. It should securely govern data stored in multiple clouds and across various regions worldwide.

Accommodating Shifting Data Sovereignty Requirements

Some industries, such as healthcare and financial services, uphold data sovereignty requirements that mandate you keep data in the same locale in which it is generated. If your platform supports seamless data portability, you'll have an easier time complying with regulations. A modern cloud data platform should give you the flexibility to select the cloud provider with the best offering and strongest regional presence. It should also allow you to replicate portions of a data set and enable fine-grained control over where those portions are located so that you can put precisely the right data into the hands of the teams that need it.

Your data platform should enable these capabilities across regions and clouds without reducing the performance of operations against your primary data. In other words, analysts, data scientists, software developers, and other knowledge workers should be able to do their jobs without worrying about where they are or where their data is coming from, confident that it's always available when they need it.

TIP

Each public cloud provider has different levels of regional presence. A cross-cloud data platform should enable secure access to data globally while upholding regional data privacy laws. This allows you to select cloud providers that meet the needs of each application, business unit, and competitive scenario.

Delivering Federated Governance

Putting your data in a modern cloud data platform simplifies data governance by making it easier for business users to comply with industry-specific and region-specific requirements yet maintain the freedom to locate their data in whichever public cloud has the strongest local presence. With such "federated" governance, domain owners of data can easily define and apply fine-grained policies, all in keeping with centrally managed governance processes — even as data is shared among clouds, regions, and workloads. Domain teams can discover and query the same data, and their resulting views of the data set can change based on their role and the data sensitivity, drastically simplifying governance while allowing teams to obtain value from a single copy of data.

REMEMBER

Cloud agnostic doesn't simply mean storing your data and operating your workloads in whatever cloud you choose. It also means standardizing on a single cloud data platform built on a single code base that operates seamlessly across all the clouds your organization relies upon.

ENABLING A MULTI-CLOUD STRATEGY

CASE STUDY

Founded in 1851, financial services company Western Union enables customers to pay bills, send money, and pick up cash at more than 550,000 agent locations worldwide. To ensure exceptional experiences for more than 250 million customers across retail and digital channels, Western Union ingests and analyzes large amounts of transactional data.

(continued)

(continued)

Unfortunately, its legacy data architecture, which consisted of multiple on-premises data warehouses, made it difficult to gain a comprehensive view of each customer. Developing visualizations, provisioning users, ensuring 24/7 uptime, and performing maintenance across a wide variety of systems was operationally burdensome and diverted attention from analytics. Copying large amounts of data, as many as five times due to different ingestion processes, created dissimilarities in the data and questions about mismatched data sets. Increased demand for analytics led to resource contention, despite costly and time-consuming hardware upgrades.

Seeking to consolidate these disparate systems, Western Union implemented a cloud data platform that powers a unified, seamless experience across multiple public clouds, including Amazon Web Services and Google Cloud Platform. As a result, various divisions within the company can pick the best cloud for each use case and replicate data between clouds without additional pipelines.

Today, integrating data from Western Union's sales, marketing, and service clouds provides actionable insights to help frontline staff elevate advertising performance and customer loyalty. The cloud data platform has allowed the company to consolidate more than 30 data stores and utilize a wide variety of business intelligence tools. For example, C-level executives rely on Tableau dashboards to monitor Western Union's transaction volume and value. Data science teams use Amazon SageMaker to create machine learning models and then move the results into self-service dashboards that display metrics about model performance.

Having a single system has reduced costs by more than 50 percent. Eight data operations teams — which previously existed to keep data flowing and provide system maintenance — have been combined, reduced, or redeployed to more valuable projects. The multi-cluster shared data architecture scales instantly to handle Western Union's data, users, and workloads without resource contention. In the future, Western Union plans to use the cloud data platform and an associated data marketplace to share insights with approximately 37,000 B2B clients.

IN THIS CHAPTER

- Outlining the essential elements of cloud security and data protection
- Enforcing comprehensive yet nonintrusive data security and governance policies
- Centrally authorizing and authenticating users

Chapter 8
Leveraging a Secure and Governed Data Platform

Protecting your data and complying with industry and regional regulations is fundamental to a cloud data platform's architecture, implementation, and operation. All aspects of the service must center on maintaining security, protecting sensitive information, and complying with industry mandates.

Introducing Key Principles

The three major aspects of good governance are knowing your data, protecting your data, and unlocking data across teams and workloads — and with external data consumers (see Figure 8-1).

Centralizing control

Good governance is much easier to achieve when all database objects (data structures such as tables and views used to store and reference data) are centrally maintained and updated by the data platform. The data platform should apply fine-grained governance across all the different objects, not just the database, and those governance policies should be always replicated with the data.

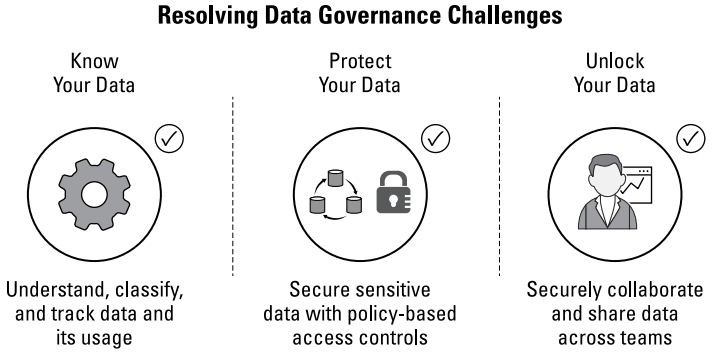


FIGURE 8-1: Comprehensive data governance is based on these three fundamental principles.

This fundamental principle makes all your other security practices more effective. For example, it's one of the things that makes secure data sharing and collaboration possible: By creating access to read-only views of a data provider's data, you can maintain a single source of the data and authorize data consumers with the appropriate levels of access.

To maximize data availability while minimizing risk, the right cloud data platform should allow you to create flexible data-access policies backed by centrally enforced protections, controls, and audit procedures. Common methods include:

- **Interaction controls,** such as secure views, secure joins, and secure user-defined functions (UDFs), are dynamically applied as people interact with the data.
- **Traceability features** allow data owners to track data where it lives to ensure protections are continually applied and allow for data deletion where appropriate (such as the "right to be forgotten").

Enforcing access policies

A cloud data platform should always *authorize* users, *authenticate* their credentials, and grant users *access* only to the data they're authorized to see. *Role-based access control* applies these restrictions based on each user's role and function in the organization. For example, finance personnel can view data from the general ledger, HR professionals can view data on salaries and benefits, and so on. These controls can be applied to all database objects,

including the tables where data is stored, the schemas that describe the database structure, and any virtual extensions to the database, such as views.


REMEMBER

Role-based access policies should be centrally managed and universally enforced. It should be easy for data owners to grant permissions and then later update or rescind them as the data set evolves and when people take on new responsibilities or move to new positions. Whether a person is accessing data from a different region or cloud, that person's permissions must remain the same. The permissions should "follow the data." With a modern cloud data platform, these security constraints should be built in and easy to set up and scale without placing an additional burden on your database administrators or IT team.

Protecting sensitive data

Once data resides within a cloud data platform, you can control access to that data in several ways, including:

- **Secure views**, such as allowing customers to see only specific rows of data from a table and not to see rows that pertain to other customers, allow organizations to control access to data and avoid potential security breaches.
- **Secure joins** can establish discrete linkages (to people, devices, cookies, or other identifiers) without exchanging or making visible any personally identifiable information (PII).
- **Secure UDFs** allow data consumers to link, join, and analyze fine-grained data while preventing other parties from viewing or exporting the raw data.

For example, by sharing only certain views, a data provider can limit the degree of exposure to the underlying tables. Data consumers can query specific databases, tables, and views only if granted access privileges. A consumer may have permission to query the view but be denied access to the rest of the table. By creating these secure views, the data provider can control access to a shared data set and avoid security breaches.


REMEMBER

Data access policies should not change the data in the underlying table: They should be dynamically applied when the table is queried. For example, a national sales database can be set up with row-level access restrictions so sales reps can see only the account information for their regions.

Without these flexible governance policies, data stewards would have to copy regional sales information into separate tables to share data with the pertinent sales regions — one table for the southwest region, another table for the northwest region, and so on. However, changes in the base table need to be copied and merged to all the regional tables, requiring constant administration. A cloud data platform simplifies this scenario by allowing a sales manager to maintain data in one base table, to which secure views and other access policies are applied dynamically at query time.

Another common method is to mask or anonymize part of the data set, such as revealing only the ZIP code fields from an address table. This would be a good way to allow data scientists to access the data they need to make regional predictions without exposing the people or households involved. This same logic can be applied to Social Security numbers, salary information, credit card information, and any other type of data that should be protected from unauthorized users. Centralized security policies can allow you to unlock more value from your data while maintaining control and minimizing risks.

Ideally, these data access policies should "follow the data" between clouds and regions, which Chapter 7 discusses, and defined roles should be applied and enforced across the entire organization for easy management.

Complying with regulations

Centralizing data governance also makes complying with data privacy mandates easier. Many organizations have concerns about the proper use of PII, protected health information (PHI), competitive data, and other types of sensitive information. In some cases, they must adhere to strict regulations governing the security and privacy of consumer data, such as the European Union's General Data Protection Regulation (GDPR), the United States' Health Insurance Portability and Accountability Act of 1996 (HIPAA), and the California Consumer Privacy Act (CCPA). These regulations must be observed throughout the entire lifecycle of your data — from creation and storage, to usage and sharing, to archiving and deletion.

REMEMBER

A cloud data platform should help you comply with all pertinent industry regulations and provide security and compliance reports upon request. Your cloud data platform vendor must demonstrate that it adequately monitors and responds to threats and security incidents and has sufficient incident response procedures

in place. Industry-standard attestation reports verify that cloud vendors use appropriate security controls. Check with your data platform provider to ensure the reports you need are available. Common certifications include:

- **OC 1 Type I:** An independent auditor's attestation of the *financial controls* the provider has had in place during the report's coverage period
- **SOC 2 Type II:** An independent auditor's attestation of the *security controls* the provider has had in place during the report's coverage period
- **PCI-DSS:** A set of data security standards from the payment card industry to which merchants must adhere when dealing with consumer data
- **HITRUST/HIPAA:** A comprehensive set of baseline security and privacy controls for organizations that deal with healthcare data
- **ISO/IEC 27001:** A set of standards for establishing, implementing, maintaining, and improving information security
- **FedRAMP Moderate:** A standardized approach to enforcing data security for federal agencies within the U.S. government
- **GxP:** Data integrity requirements for life sciences organizations that produce regulated medical products

Encrypting data

Encrypting data involves applying an encryption algorithm to translate readable text into *ciphertext*, which contains a form of the original *plaintext* that is unreadable by a human or computer without the proper cipher to decrypt it.

Decryption, the inverse of *encryption*, is the process of turning ciphertext into readable plaintext. This is fundamental to security, and your cloud data platform should ensure it happens by default, automatically, all the time, and without impacting the performance of your data-dependent workloads.

TIP

Data should be encrypted *in transit* and *at rest,* which means from the time it leaves your premises, through the Internet or another network connection, and into the platform. It should be encrypted when it's stored on disk, when it's moved into a staging location, when it's placed within a database object, and when it's cached within a data repository. Query results should also be encrypted.

The cloud data platform vendor should also protect the decryption keys that decode your data from ciphertext back to plaintext. The best cloud vendors deploy AES 256-bit encryption with a hierarchical key model. This method encrypts the encryption keys and instigates key rotation that limits the time any single key can be used, further strengthening security.

Sharing centralized data

Breaking down data silos across your ecosystem and centralizing data in a common repository makes governance easier. It also facilitates the secure and easy *sharing* of data. Chapter 7 discusses how a modern cloud data platform enables data providers to control access to data and establish secure views of that data. It allows you to share data safely and securely by easily setting up centralized policies and permission structures based on department, region, role, and other considerations. Each data team can define and manage the policies that are appropriate for their data.

A cloud data platform should also anchor a thriving ecosystem of technology partners to give you options for what you do with your data. To uphold strong security, this integration can be achieved with two types of controls for network security: private connectivity and via a known range of IP addresses. The data never traverses the public Internet, which significantly reduces exposure to common security threats. For example, automated machine learning (AutoML) tools from third-party vendors can be consumed through dashboards, reports, and predictive analytics via connections to other ecosystem partners. The data, compute processing, and machine learning results reside in the data platform for easy access.

IN THIS CHAPTER

- Maximizing performance for all types of usage
- Understanding data processing engines
- Identifying limitations with commodity cloud providers
- Establishing a cohesive set of cloud services

Chapter 9
Achieving Optimal Performance in the Cloud

In the cloud, rapid data processing means less resource consumption and lower costs. Virtually unlimited cloud resources make it easy to scale vertically and horizontally, to bring in new teams that can all run more types of operations on your data in parallel without contention. However, you need a flexible data platform to properly leverage all that compute power, provision the right amount of resources, and easily process all types of data for many kinds of workloads. Without these essential ingredients, you can't easily put the data to work for your business.

Maximizing Performance for All Data Processing Activities

A modern cloud data platform should accelerate everything from storing and processing data to handling transactions, securing data, and managing metadata. One platform, governed by one set of services, should support the needs of analysts, data scientists,

data engineers, and also application developers creating new data products for your internal stakeholders or external customers. Additionally, all users should interact with the same data without contending for resources or experiencing data processing delays.

This versatility is made possible only by a data processing engine that works exceptionally well for a wide range of workloads. As a result, your organization can standardize on one universal, flexible, and open data platform optimized for many data management and data analytic activities, rather than having to acquire, learn, and apply a unique data processing system to each task and then stitch them together (see Figure 9-1).

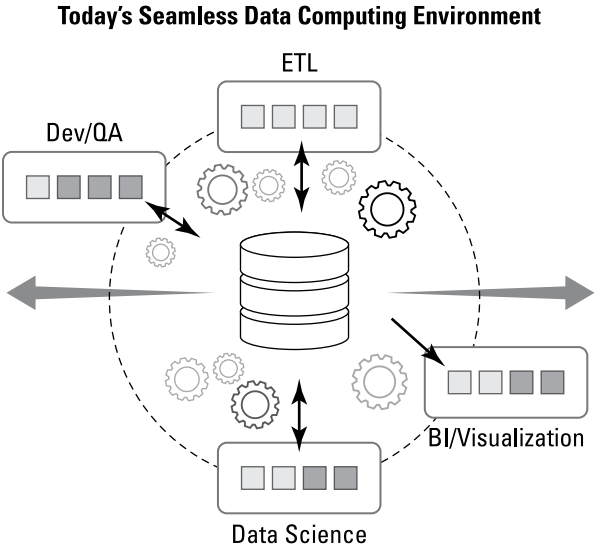


FIGURE 9-1: A modern cloud data platform should deliver the power, speed, seamlessness, and versatility of running a near-unlimited number of concurrent and interconnected data workloads, at practically any data scale.

For example, a data engineering team can process data through high-volume data pipelines while a data scientist team conducts exploratory analysis, and a business unit queries an immense data set. All these activities should happen at the same time without performance issues. They should also happen seamlessly between each other. As Chapter 1 describes, only a multi-cluster shared data architecture can enable all teams to experience great

performance and scale these separate processes at will, without resource contention. The platform should automatically manage each workload to maximize throughput and ensure consistent results, making it possible for thousands of users to simultaneously analyze and share the same single copy of data with no bottlenecks.

REMEMBER

A cloud data platform architected first and foremost for the cloud can automatically provision nearly limitless amounts of compute power to support virtually any number of users and workloads without affecting performance.

Understanding Data Integration and Performance Issues

The world is awash with data, giving rise to many data processing tools and strategies. Typically, each data workload requires a unique data processing engine — purpose-built and tuned for each workload. For example, you might need one type of engine for loading and transforming data via pipelines, another for processing analytic queries, a third for training machine learning models, and so forth. Each engine may be tied to its own data repository and may require different programming languages, such as SQL or Scala. Each must be properly tuned to maximize performance for that workload's unique attributes.

Because these data processing engines support very disparate workloads, they have very different features and functions that make it difficult, if not impossible, to easily stitch them all together to deliver a cohesive data experience for an enterprise. As a result, organizations commonly end up with a separate environment for each data workload, all operating in isolation. Each one requires unique skills, tools, services, and overhead — from data engineers developing data pipelines, to business analysts running reports, to data scientists developing predictive models, to application developers creating and maintaining data apps.

These discrete technologies may vie for a finite set of compute resources in traditional on-premises systems and many cloud

services. If one team is running a heavy data preparation job while another is crunching end-of-the-month financial reports, both teams may experience resource contention and thus poor performance or failed jobs. Adding more resources can be a lengthy process, requiring new capital expenditures, complex implementation cycles, and ongoing system maintenance.

Identifying limitations with cloud providers

The big cloud providers all offer customers near-limitless amounts of compute and storage capacity. These vendors have amassed thriving ecosystems of data processing tools and utilities, some developed internally and others by third parties. In addition to utilizing the raw data storage and compute infrastructure services from these cloud platforms, customers can choose from a wide array of add-on services for everything from preparing data to processing queries to building and training machine learning models.

These cloud platform ecosystems allow you to select from hundreds of services for accessing, preparing, and processing data. However, many of these services use unique data processing engines with their own access requirements, maintenance procedures, and learning curves. It's up to you to figure out how to make them work together. If you don't, you will quickly find yourself confronting the same data-silo and data-access challenges you encountered in the on-premises world: disparate services, each with their own data pipelines, development tools, and management utilities.

The critical issue is this: Can the cloud vendor and its associated ecosystem of add-on services fulfill all your data management and analytic needs cohesively without forcing you to master unique languages, development techniques, and management tools? What services are layered on top of the basic cloud infrastructure to handle data engineering, business analytics, data science, and other tasks? How easy is it to integrate and use data for these various activities?

In many cases, the burden is on you to figure out how to perform each business task, integrate data, and synthesize the results back into the platform. Training your team to work synergistically is no small task, especially for an organization that seeks to maximize the accessibility and usability of its data.

TIP

Having a cloud data platform that spans multiple clouds allows you to keep your data closer to the processing of that data, which minimizes data latency and maximizes performance. For example, if a new data app becomes popular in Japan, you probably don't want application data stored in America or Western Europe. Not just software companies, but any multinational firm with diverse geographic requirements, can benefit from having regional and cross-cloud flexibility. See Chapter 7 for additional details.

Reviewing limitations of point solutions

Other questions should naturally arise as you evaluate the various products within these respective cloud ecosystems. Do they perform well in concert? Do they all share a common interface? Do they complement each other as a cohesive set of services, or does it seem more like a bunch of independently developed capabilities? If you cobble together a bunch of products within a cloud vendor's ecosystem, it may be hard to make it all work together well enough to achieve your performance goals. For example, if multiple users use the same service to access data, can the cloud provider minimize resource contention among multiple teams?

REMEMBER

Some cloud data platform vendors claim that they can run all types of workloads against one common data repository, but the data processing engine isn't accessible to all users, and it doesn't dedicate resources to each workload. That means each team must lobby IT to request resources or fight for compute time with other groups. Performance degrades as contention increases. Be sure to ask vendors whether all concurrent workloads execute simultaneously without impacting the performance of other workloads and services. If they can't, end users may be forced to manage their resources and learn a specialized set of skills to use the performance engine.

SUPPORTING DATA ENGINEERING, BUSINESS INTELLIGENCE, AND PREDICTIVE ANALYTICS

CASE STUDY

When U.K. supermarket giant Sainsbury's set out to make analytics more cost-effective and accessible to its employees, the first step was to consolidate functionally siloed data from multiple operating companies into a modern cloud data platform. In addition to being the second-largest general merchandise and clothing business in the U.K., Sainsbury's owns a bank and hundreds of grocery stores. The organization has thousands of employees and millions of customers and performs billions of transactions each year.

To populate its data platform, Sainsbury's data engineers combined data from three large enterprise data stores: supply chain analytics from its food business, customer loyalty analytics from its nationwide loyalty program, and data from a traditional enterprise data warehouse. The cloud data platform made it relatively easy to rationalize these data silos into a common format and democratize the data among three groups of employees: data scientists, professional analysts, and employees who want to know more about their customers but lack the technical skills to do the analytics.

These three systems now publish raw data directly to the cloud data platform, which populates a dashboard that streams data to the digital trading teams. Data that was formerly difficult to access from Sainsbury's retail stores and other consumer-facing channels is now readily available. Store managers can obtain standard reports via cloud-based dashboards that offer visibility into customer needs and preferences. In addition, data scientists and machine learning engineers are creating new data sets by accessing raw data from a data lake, which resides within the cloud data platform. The cloud-built platform separates storage and compute resources, improving performance and eliminating resource contention for thousands of users. Queries that used to take six hours in a legacy data warehouse now run in three seconds in the cloud data platform.

"Being connected to our customers, knowing and serving them better, is a core pillar of our strategy," Sainsbury's Chief Data and Analytics Officer Helen Hunter said. "This involves bringing together disparate data assets and democratizing data for the good of every user."

IN THIS CHAPTER

- Considering your overall requirements
- Identifying the data and workloads you want to migrate
- Comparing solutions and options
- Determining total cost of ownership and the return on your investment
- Assessing success criteria

Chapter 10
Five Steps for Getting Started with a Cloud Data Platform

This chapter guides you through five key steps to choosing a cloud data platform 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 data platform can take your business in new directions. Remember, a cloud data platform isn't a disparate set of tools or services. It's 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 models, data application development and operation, and data sharing for easily and securely sharing data among authorized users.

These workloads have unique attributes, but all depend on the universal principles of availability, reliability, extensibility, durability, security, governance, and ease of use. Keep these essential workloads in mind as you ask yourself these questions:

- **Existing tools and processes:** Are there entrenched tools, work habits, and business practices you want to accommodate with your cloud data platform? What business processes will it impact, and which departments will benefit?
- **Usage:** Which users and applications will access or leverage the cloud data platform? What types of queries will you run, and by how many users? How much data will users need to access, and how quickly? Which workloads will you run, and how will they vary over time?
- **Data sharing:** Do you plan to share data within your organization and with customers and/or external partners? Will you enrich that data by adding data analytics services? Will you look for data monetization opportunities?
- **Global access:** Do you have specific functional, regional, or data sovereignty/ requirements? Do you need a cross-cloud architecture to maximize deployment options, bolster disaster recovery, or ensure global business continuity?
- **Resources:** What staff do you have in place, and to what extent can you apply those resources to these new data-driven projects, workloads, and access patterns?

Step 2: Migrate or Start Fresh

Assess how much of your existing environment you wish to carry forward:

- **Is this a brand-new project?** If so, consider how you can take full advantage of the capabilities of a cloud data platform rather than pursuing an outdated approach or strategy.
- **Which applications and workloads should you prioritize?** Consider migrating easy, straightforward workloads to the new cloud data platform first. This will allow you to obtain quick wins and solid validation from the user community before you attempt to tackle more difficult initiatives.

- **Will your existing applications work with the new platform?** Business intelligence solutions, data visualization tools, data science libraries, and other software development tools should easily adapt to the new architecture.
- **How are your requirements likely to change in the future**? As you ponder emerging data-driven projects and future application initiatives, make sure you are positioned to accommodate new data, technologies, and capabilities such as Internet of Things (IoT), machine learning, and artificial intelligence.

Step 3: Evaluate Solutions

As this book describes, your cloud data platform must take full advantage of the true benefits of the cloud, with an architecture based on three foundational pillars:

- Convenient access to data via a near-zero-maintenance environment
- Exceptional performance for concurrent data-usage activities
- Easy and secure analysis and sharing of data, both across the organization and within a broad ecosystem

In that vein, make sure your choice meets these architectural criteria:

- Natively integrates structured, semi-structured, and unstructured data and avoids creating data silos
- Includes integrated policy-based data governance controls that follow the data
- Shares live data without having to copy or move that data
- Replicates databases and keeps them synchronized across regions and clouds to improve business continuity and streamline expansion
- Scales compute and storage capacity independently and automatically, and scales concurrency instantly and near-infinitely without slowing performance

Step 4: Calculate TCO and ROI

If you choose a cloud data platform that accommodates all types of data and that has been designed first and foremost for the cloud, you should be able to pay for actual usage in per-second increments and minimize additional costs, such as maintaining multiple systems and training people to handle diverse data.

If you outsource everything to the vendor by choosing a data-platform-as-a-service offering, you can calculate the total cost of ownership (TCO) based on the expected usage fees. If you opt to use an external object store from one of the big cloud vendors, you also need to add the costs of that vendor's services.

Calculate the return on investment (ROI) over the expected lifetime of the data platform options you're considering. Don't overlook the savings possible with features such as scaling up and down dynamically in response to changing demand.

Consider the potential revenue impact of monetizing your data. A cloud data platform helps you maximize the value of your data — and not just the data you have within your own four walls but also external third-party data available via data marketplaces.

As a data provider, you can offer governed slices of your data to potentially thousands of data consumers to create new revenue streams. You can also combine your data with marketplace data to create valuable products and services.

Step 5: Establish Success Criteria

How will you measure the success of the new cloud data platform initiative? Identify the most important business and technical requirements, focusing on performance, concurrency, simplicity, and TCO.

For example, does the new data platform make your organization more productive? Does it simplify access to key workloads, break down data silos, and boost collaboration? Bringing your data together brings your teams together. Calculate the impact of standardizing on one centralized system versus struggling with a patchwork of tools, apps, and data sets. Focus on measurable, quantifiable criteria and qualitative enhancements.

Unite siloed data, execute critical data workloads, and share data securely

You're generating more data of multiple types than the hodgepodge of systems you deploy can handle. And you have data silos, numbering in the dozens, hundreds, or even thousands across your business. Even worse, exponentially more data exists outside your organization — data that if you could access it easily would provide insights and business opportunities you've never imagined before. Enter the modern cloud data platform. This book reveals how any organization can easily store, integrate, analyze, share, and acquire data across clouds and regions, locally and globally, for untold possibilities. Read on.

Inside...

- How to select a cloud data platform
- Seamlessly access your varied data
- Rely on a modern architecture to access near-unlimited compute and storage
- Accommodate your users and workloads
- Access modern data marketplaces and create new data products and applications
- Support new architectural frameworks
- Experience real-world case studies

David Baum (david@dbaumcomm.com) is a freelance business writer specializing in science and technology.

Cover Image: © ktsdesign/Shutterstock

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

ISBN: 978-1-119-87548-2
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

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