LEARNING MADE EASY

2nd Snowflake Special Edition

Data Sharing

for dummies®
A Wiley Brand

What is secure enterprise data sharing?

How to avoid the hassles of moving shared data

How to monetize your shared data

Brought to you by:

snowflake™

Lawrence C. Miller
David Baum

About Snowflake

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

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

Data Sharing

2nd Snowflake Special Edition

by Lawrence C. Miller and David Baum

for dummies®
A Wiley Brand

Data Sharing For Dummies®, 2nd Snowflake Special Edition

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

ISBN 978-1-119-66641-7 (pbk); ISBN 978-1-119-66645-5 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

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

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

Development Editor: Nicole Sholly

Project Editor: Martin V. Minner

Executive Editor: Steve Hayes

Editorial Manager: Rev Mengle

Business Development Representative: Karen Hattan

Production Editor: Siddique Shaik

Snowflake Contributors Team: *Vincent Morello, Daniel Kuperman, Leslie Steere, Michael Nixon, Matthew Glickman*

Table of Contents

Introduction

Organizations inside an enterprise acquire crucial insight by analyzing data they share with each other. For example, finance teams need sales data to forecast future financial performance. Product management teams require marketing data to determine future products and services. Executive management needs up-to-the minute dashboards, fueled by data from many parts of the enterprise, to make timely, data-driven business decisions.

Outside an organization, retailers share sales data with their vendors to manage inventory and supply chains. Software-as-a-service (SaaS) providers share the data they collect with their customers to provide them with deeper insights into their business. Healthcare providers securely share patient data with vendors that provide ancillary products and with other business partners that analyze that data to help improve patient services. The list goes on. According to Forrester Research, more than half of global data and analytics decision makers report their firm is expanding its ability to source external data.

Data has become more than something to collect and analyze. It's an asset you can easily and securely make available inside and outside your organization to streamline operations, swiftly deliver more-personalized customer experiences, and open up new market opportunities. As a data provider, you can also securely monetize your data and create self-service relationships between your organization and an endless number of data consumers. In fact, 47 percent of global data and analytics decision makers report their organizations currently commercialize data, according to Forrester.

About This Book

Welcome to *Data Sharing For Dummies,* 2nd Snowflake Special Edition, where you explore how modern data sharing enables any organization to share and receive live data, within minutes, in a governed and secure way — with almost none of the risk, cost, headache, and delay that have plagued traditional data

sharing methods. Modern data sharing allows an organization to easily and quickly forge one-to-one, one-to-many, and many-to-many relationships to share data in new and imaginative ways and reduce time to insight to a level never before possible.

Icons Used in This Book

In this book, you'll occasionally see special icons calling attention to important information. Here's what to expect:


CASE STUDY

The case studies provide best practices from organizations that have successfully used modern data sharing methods.


REMEMBER

This icon points out information you should commit to your non-volatile memory — your gray matter.


TECHNICAL STUFF

This icon explains the jargon beneath the jargon.


TIP

This icon points out useful nuggets of information and helpful advice.


WARNING

These alerts offer practical advice to help you avoid potentially costly or frustrating mistakes.

Beyond the Book

At the end of this book, if you're thinking, "Where can I learn more?" just go to `www.snowflake.com` to find out what Snowflake offers, obtain details about modern data sharing, view webinars, get the scoop on upcoming events, and access documentation and other support. You can contact Snowflake or even try its technology for free.

IN THIS CHAPTER

- Defining data sharing
- Recognizing the importance of data sharing
- Exploring data sharing examples
- Understanding how organizations share data
- Taking advantage of data sharing opportunities

Chapter 1
Getting Up to Speed on Data Sharing Basics

Every day, organizations everywhere use data to track business results, make decisions, engage customers, define and create products, forecast trends, and more. Data is also a resource used and consumed between organizations, internal and external to one another, to collaborate on business plans, mutual initiatives, or joint opportunities.

There's no limit to how enterprises can engage and collaborate with data. However, data does not magically appear on your doorstep, figuratively speaking. It is generated at a place of origin and then distributed across the organization and analyzed to gain insights.

In this chapter, you learn about data sharing — what it is, why it matters, how and why organizations share data, and what business opportunities data sharing can create.

What Is Data Sharing?

Data can originate from the many software applications an enterprise uses to run its business, from the constant activity of visitors engaging a website, from an Internet of Things (IoT) device attached to the refrigerator in your home, or from a sensor built into something as sophisticated as the jet engine of an airliner. There are potentially endless data-creating scenarios in the modern world. Market intelligence firm IDC estimates the world's total digital data created will increase to 180 zettabytes by 2025 (one zettabyte is equal to about 1 trillion gigabytes). Unfortunately, traditional data sharing methods require moving data, which is riddled with problems. Going forward it will be impractical, if not impossible, to share vast amounts of data in meaningful ways.

Furthermore, many enterprises have come to realize they could enhance their business operations if they had access to data outside their organizations. Enterprises also recognize it is not easy to access data they don't generate themselves. Thus, *data sharing* is the act of providing access to data between business units inside the same organization, or between organizations external to each other. The organization that shares its data is called a *data provider.* The organization that wants to use shared data is called the *data consumer.* Any organization can be a data provider, data consumer, or both.

Figure 1-1 shows how organizations have traditionally shared data — by making a copy of the shared data and sending it to their data consumers. The data consumers then download the data to analyze or combine that data with their existing data for deeper insights into who their customers are, how efficiently their business operates, and into which new industries their business is heading.

But this process is slow, cumbersome, and costly and only allows for moving limited amounts of shared data. Figure 1-2 shows how modern data sharing happens without moving data. Instead, a data provider makes available live, read-only copies of data to its data consumers via modern cloud data sharing. In essence, data doesn't have to move.

REMEMBER

Modern data sharing is a feature found in modern data platforms.

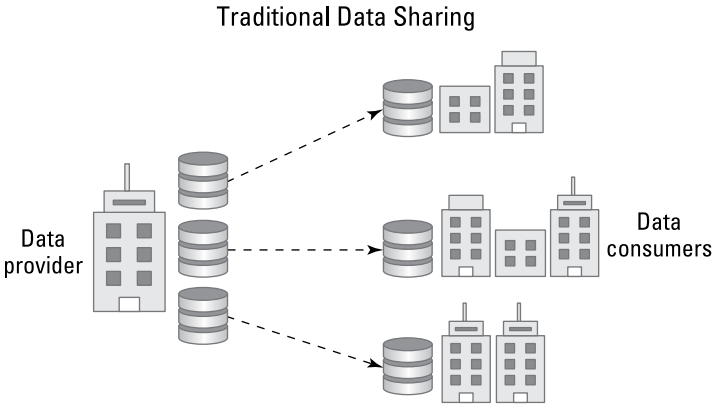

FIGURE 1-1: Traditional data sharing requires duplicating and moving data from a data provider to data consumers.

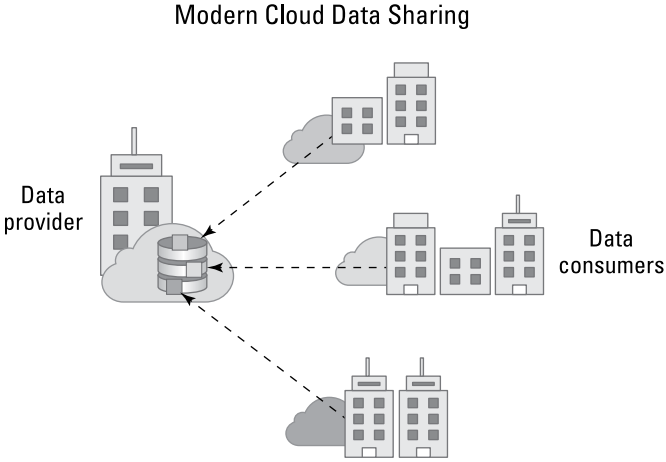

FIGURE 1-2: Modern cloud data sharing enables fast, live, secure, and governed data sharing without moving data.

CASE STUDY

PORTAL FOR JOB SEEKERS IMPROVES 300 PERCENT

Snagajob's mobile sourcing and hiring tools connect 75 million registered hourly workers to Snagajob's business subscribers, which represent 300,000 employer locations.

Snagajob had a need to share its data from its data platform with an external marketing analytics firm. The firm then used the data to reach out to Snagajob's business clients to execute targeted re-engagement campaigns on behalf of Snagajob. This relationship allowed Snagajob to avoid in-house labor costs associated with this marketing function.

To share its data, Snagajob routinely had to implement time-consuming steps that included:

- Identifying the database elements to be shared
- Extracting the data set with a client tool
- Compressing and encrypting the data in order to email it
- Emailing the file to the marketing partner

After the external marketing firm received the file, it would execute the same time-consuming steps in reverse, with the additional steps of building a database table to ingest the data and then importing the data into its target database.

Snagajob turned to a modern, built-for-the-cloud data platform that could operate as an extremely scalable data warehouse but also be its platform for data sharing. All Snagajob had to do was create a "share" that enabled its marketing partner to receive live, secure, and instantaneous access to the tables and views Snagajob shared. The marketing partner could then execute the email campaign with data provided that was always accurate and up to date, because the partner always accesses a live, read-only version of Snagajob's latest data. All of this took place in a matter of minutes, without any data movement.

Performance, reliability, and agility were dramatically increased, allowing Snagajob to reduce the implementation time for sharing data from several days to just a few hours. Snagajob saved 300 percent on costs and was able to operate much more quickly, which improved its industry competitiveness.

Data Sharing Examples

Here are just a few of the new business opportunities modern cloud data sharing makes possible:

- **Data sharing to eliminate data silos:** Develop a single source of truth for all your internal data and share it among thousands of data consumers across hundreds of business units within a single organization.
- **Data sharing for business efficiencies:** Share live data with your business partners to optimize costs, streamline operations, and provide superior customer service.
- **Data sharing as a product:** Provide live and direct access to slices of your data as a monetized service so your data consumers can enrich their own existing data.
- **Data sharing as a product differentiator:** Software-as-a-service (SaaS) providers can offer direct access to the petabytes of data generated from their business-to-business (B2B) subscribers' activity. Those subscribers can then perform deeper analysis on more of their data — analysis previously unavailable to them.

Capitalizing on these opportunities requires data sharing capabilities with uncommon speed, power, security, governance, and simplicity. These capabilities are not available with traditional data sharing methods.

How Organizations Share Data

Traditional approaches require laborious efforts to stitch together a patchwork of disparate tasks to share and move data. These processes are costly, create manual overhead, and limit how much data an organization can actually share:

- **Email:** A data file is emailed from provider to consumer.
- **File Transfer Protocol (FTP):** Data files are shared and downloaded between two computers or via the Internet.
- **Extract, transfer, load (ETL) software:** ETL software extracts data from the provider's database, transforms the data, and then loads it into the consumer's database.

- **Online file sharing services:** These are similar to FTP, but sharing and downloading data files takes place via Internet file transfer only.
- **Cloud storage:** The provider stores data in the cloud and provides the consumer with credentials for accessing it.
- **Application programming interfaces (APIs):** An API is used to initiate and manage the data transfer.

But imagine the possibilities of having on-demand access to ready-to-use, live data so you can make immediate use of that data inside a secure, governed environment.

Exploring Data Sharing Possibilities

With modern data sharing, the possibilities are practically endless. Here are a few compelling opportunities

- **Improve the customer experience:** Delivering targeted business or retail offerings with personalized marketing campaigns in a highly competitive, digital market requires a deeper understanding of your customers, competitors, and industry trends. The primary path for gaining this understanding involves acquiring data you don't already have in order to reveal what you don't already know.
- **Streamline your business:** Easily sharing data across the multitude of business units that comprise your organization, and with your business partners, creates a single source of truth. This could save billions of dollars by reconciling even the most minor data inconsistences.
- **Create new business assets from data:** Some data within a data provider will be just as valuable to thousands of external, non-competing data consumers. All of this can happen through an effortless, self-serving business model thanks to the simplicity of sharing any part of a modern cloud data platform that offers modern data sharing.

IN THIS CHAPTER

- Looking at today's data sharing methods
- Recognizing the cost of delays
- Looking at computing complexity
- Addressing business pain points

Chapter 2
Understanding Traditional Data Sharing Challenges

In this chapter, you learn about the many limitations of traditional data sharing methods and technologies.

Addressing a Multi-Faceted Problem

If you anticipate sharing data with tens, hundreds, or even thousands of internal or external data consumers — each of which has unique data sharing requirements — how can you easily support this challenge? How do you support growth without constantly building more storage clusters, managing complex software, and suffering through prolonged latencies and performance penalties — all without creating inconsistencies by sharing stale copies of data? Simply put, the traditional data warehouse platforms of today were not built to support the constant need to share data in real time.

To understand the magnitude of the challenges associated with traditional data sharing, consider the pros and cons of common

approaches (shown in Table 2-1) that a company would encounter, for example, when sharing data with a third-party service provider or another external organization such as a business partner.

TABLE 2-1 Pros and Cons of Traditional Data Sharing Approaches

Data Sharing Approach	Pros	Cons
Email	• Pervasive and ubiquitous • Infrastructure in place • Easy to compose an email and attach a file	• Not conducive for large data sets from relational databases (can't scale) • Limited size of attached files (less than 25 MB) requires large data sets to be deconstructed and zipped • Limited network bandwidth, which results in slow data transmission • Not secure, requiring custom encryption • Mirror effort required on recipients' end (receive, decrypt, reconstruct data, and so on)
File transfer protocol (FTP) — see Figure 2-1	• Well-known and long-established protocol • Availability of a wide range of FTP client software and services	• Schema changes require a great deal of lead time • Must acquire FTP client software, server, and/or service • FTP account admin setup and overhead required • Large data sets must be deconstructed and broken down in size to facilitate faster data transfers • Not natively secure; requires custom encryption scripting or secure service • Mirror effort required on recipients' end (receive, decrypt, reconstruct data) • Efforts must be repeated with each new update to a shared data set

Data Sharing Approach	Pros	Cons
Extract, transfer, load (ETL) software	• Large availability of well-established ETL software solutions • Purpose-built to extract data from a database or data source and transform the data for loading into a target database • Good for bulk and complex data movement and transformations	• Latency emerges when data changes • Expensive software, costing up to tens of thousands of dollars • Complex, requiring specialized skills to integrate and deploy • Can take months to implement • Change management and schema evolution can be difficult
Online file sharing services	• High availability of services • Generally easy to use	• Better suited for sharing of flat files, not relational database objects • Data is not ready to use (ready to analyze) • Risk associated with data inconsistencies when the original copy of the data changes
Cloud storage	• Numerous services available from large cloud storage providers	• Less-than-optimal performance when querying directly from cloud storage • Change management and schema evolution difficult, requiring a separate metadata management process • Risk exposure when data changes • Complete SQL data manipulation language (DML) semantics (for example, UPDATE, INSERT) may not be supported

(continued)

TABLE 2-1 *(continued)*

Data Sharing Approach	Pros	Cons
Application programming interfaces (APIs)	• Numerous APIs available • Wide variety of use cases • Programmatic implementation relieves some manual effort	• Data movement is required, creating risks for failed transfers • APIs process data in small amounts, creating bottlenecks for large data volumes • Performance is directly affected by available bandwidth, requiring high costs for higher bandwidths

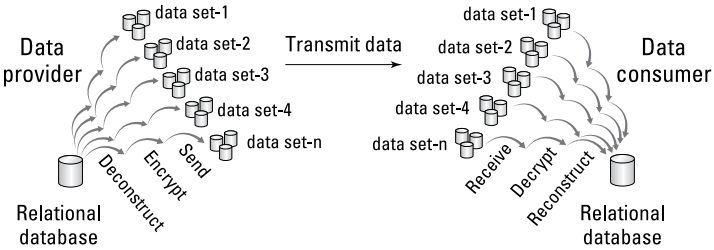

FIGURE 2-1: Multiple steps of a typical legacy FTP-based data sharing workflow.

Dealing with Conventional Data Sharing: Time-to-Value Delays

Conventional data sharing methods can create other challenges that cause more delays and require more assistance from your IT teams, including:

- **Handling increased data size:** The shared data set is often much larger than originally scoped, which creates problems with the data extraction process. You'll likely need a scripting language to automate the breakdown and extraction process, which may require additional IT assistance. The reverse process must also occur for data consumers.
- **Decrypting sensitive data:** If the data set includes sensitive information, the output files will likely need to be encrypted, masked, or redacted, which may require additional IT assistance. If the data set is encrypted, encryption keys must be securely

shared between the parties via a separate process, and the data consumer must decrypt the shared data.

- **Changing file formats and schema:** It may be necessary to change the file format multiple times if additional database attributes must be shared. When table attributes change on the data provider's end, a corresponding change must also occur on the data consumer's end.

The accumulation of all these steps results in slow and painful processes for both data providers and consumers. All of this must happen *before* any attempt to analyze and develop insights from the data, which creates time-to-value delays.

Usually, the delays and difficulty don't end with just the data transfer effort. For example:

- **Sharing data in real time:** More IT assistance is needed if the data set is shared in a more real-time fashion, rather than being sent only once per night.
- **Cleaning data:** The import process has problems, and the data isn't as clean as anticipated. For example, the data extraction may contain special characters that should have been disregarded. This means the data provider must build more-sophisticated data extraction processes, resulting in more IT assistance, costs, and delays.

SPEEDING ACCESS TO DATA FROM BILLIONS OF DEVICES

CASE STUDY

Localytics is a Boston-based company that provides a mobile engagement platform used in more than 37,000 apps on more than 2.7 billion mobile devices worldwide. Localytics gives hundreds of the world's top brands insights about their mobile users and the tools to engage with those users.

Localytics uses modern data sharing to provide its customers with access to Localytics' data without exporting that data, solving one of the biggest data challenges marketers face. Previously, users had to connect different sources of customer data from customer relationship management (CRM) systems, business intelligence (BI) tools, mobile

(continued)

(continued)

analytics, and other sources. Data was often exported and copied into other platforms, manipulated, and further analyzed. The process produced multiple copies of the data living in many places, thus increasing costs and complexity and producing inconsistent results.

ETL eliminated

Localytics removed the burden of cumbersome ETL efforts to make data directly accessible through modern data sharing, creating a much more efficient and reliable way to manage and understand customer data. Specifically, Localytics employs secure, permissions-based access to enable customers to work with session, event, and profile data from Localytics and run their own queries and custom reports against that data. Customers can also use popular BI tools to analyze their data.

Data latency reduced from three hours to three minutes

Localytics stores all its data in a modern cloud data platform, augmented with modern data sharing. Instant sharing of live data with Localytics' customers eliminated a previous data latency of three hours. With modern data sharing, real-time data is ready to query in about three minutes. Customers don't need to expend effort to use the data, which is filtered through live, secure, governed, and permissions-based sharing.

Localytics customers immediately saw the potential to eliminate their ETL processes, removing that burden from their overworked data teams and saving their organizations time and effort.

To protect against failures during the file transfer process, on either the extraction and/or import side, both the data provider and data consumer must incorporate special software code or scripts to monitor the transfer and automatically restart the process in the event of failure. This means greater effort and longer delays to develop insights and derive value from data.

Computing Complexity Challenges

Traditional options for sharing data also require scaling complex computing platforms to share even small slices of data. Complexity adds burdens and requires extra resources, including infrastructure costs — internally and externally.

The goal should be effortless sharing of limitless amounts of data with internal and external organizations, including your business partners, for collaboration and business planning. If your business model is focused on monetizing your data, you'll want the same level of effortless sharing to distribute data to as many data consumers as possible, with individualized, self-service access and security as needed.

WARNING

If you think cloud storage is the answer, think again. Sharing data using a basic cloud storage service is inefficient. It won't provide the ability for you or your data consumers to query the data in a high-performance manner or ensure data consistency. A Hadoop computing platform is not the answer either because of its inherent complexities and complications.

Sharing Data the Conventional Way: Business Pain Points

Cumbersome and complex data sharing methods combined with costly and inflexible computing platforms produce headaches for organizations that need to collaborate on data. In addition, the processing overhead required to extract data from a traditional data platform and transfer that data to other organizations delays the value shared data provides. Additionally, every time data changes, data extraction and transfer processes must be repeated because shared data is always a static version and becomes stale immediately.

The inability to extract insight from data quickly is an inhibitor to maximizing the commercial value from data. Data consumers encounter delays in developing insights, which can lead to dissatisfaction with the data provider. Furthermore, because common methods for sharing data can't incorporate changes immediately, data consumers risk executing analytics on incomplete data. This can lead to less accurate analytics or faulty conclusions for business decisions.

Data sharing within an organization

Data sharing scenarios within an organization include:

- Sales groups share data with finance groups to track sales and revenue to forecast an organization's performance.

- Marketing teams monitor and analyze customer data to predict behavior and align demand generation programs.
- Different subsidiaries of an organization share data with each other to better align their go-to-market plans and gain more understanding of the separate areas of the business.

When functional groups within an organization cannot share data effectively, data silos result, and business collaboration suffers. Each group will maintain its own data warehouse or *data mart* — a copy of some of the data from the corporate data warehouse. Data silo and data mart sprawl ensue and create unnecessary burdens for IT and data platform teams.

Data sharing business-to-business (B2B)

Some examples of B2B data sharing scenarios include:

- A hotel booking website shares reservation patterns and trends with hotel properties to develop promotional and pricing programs.
- A grocery chain provides store sales data to suppliers to ensure shelves are adequately stocked to meet demand.
- Retailers share in-store sales data to fashion merchandising so the hottest trends are always available.

Whether sharing data to external organizations or receiving shared data from them, if organizations cannot collaborate on data they are less efficient and run the risk of operating at a higher cost and lower productivity.

Monetizing data

An example of monetizing shared data is a data service company that gathers mobile phone location information and usage data and then shares the information with advertising agencies and marketing groups so they can execute highly targeted campaigns to specific consumers.

Look for more details about data monetization in Chapter 3.

IN THIS CHAPTER

- Tracing the history of data sharing in business
- Looking more deeply at data sharing scenarios and the value to the business

Chapter 3
Recognizing the Business Value of Sharing Data

In this chapter, you learn how data sharing methods have evolved in business, why data sharing is critical to any business, how businesses share data internally and externally, and how the cloud and software-as-a-service (SaaS) change the data sharing model.

Looking Back at the Early Days of Data Sharing

Understanding the business value of data sharing today requires a historical perspective. Not long ago, it was considered the norm for organizations to host and support multiple business applications within their own data centers. There would be an application for finance, another for marketing, others for sales, human resources, operations, and so on. Just ten years ago, large companies would host and run hundreds of business applications from their own data centers.

Each of these applications would also have an associated database. These databases were not optimized for analytics and did not share data between applications. In order to analyze this data, each business unit in charge of a database would have to extract, transform, and load (ETL) the data into its own *data mart,* which is a smaller, stand-alone version of a data warehouse. Then, to develop business intelligence across an organization or to execute analytics against companywide data, data would have to be sent through the ETL process from the individual data marts into a central data warehouse. This data would then be prepared for analytics. The entire process was slow and cumbersome. Data formats varied across the applications, requiring further modeling and transformation into a new data warehouse. But at least you had access to the data because it was in your own data center.

The bottom line is that no company survives without some level of internal data sharing.

Assessing the Business Value of Data Sharing for Organizations

Data sharing across and beyond an organization consists of four basic workflows:

- **Across lines of business (LOBs):** Sharing data between business units within the same organization
- **Between organizations:** Outbound data sharing to another, separate organization to benefit your business
- **Between organizations:** Receiving inbound data shared from another organization to benefit your business
- **Monetizing data:** Sharing live data as a service so data consumers can enrich their own, existing data

Sharing data across LOBs

Within the same organization, business units depend on email, spreadsheets, shared network drives, application programming interfaces (APIs), and other methods for communicating and for sharing data. Sharing data across an organization enables and fosters increased levels of business intelligence and drives timely and informed business decisions.

Within an organization, however, data is often locked in silos. Mergers or acquisitions, firewall restrictions, or other business or technology barriers often restrict an organization from easily sharing data across its business units. These physical or logical separations of infrastructure can prevent two or more business units from accessing all available data within an organization to deliver all-inclusive, data-driven insights. These data silos emerge when an organization relies on a traditional, on-premises data warehouse or a traditional data warehouse ported to the cloud.

Sharing between organizations: Outbound

External data sharing takes place all the time. A vendor-supplier relationship, a partner relationship, a developer-producer relationship — or any number of other business relationships — all require two or more organizations to collaborate with data to drive business. In Figure 3-1, the primary organization is sharing data, outbound, to the partner organization.

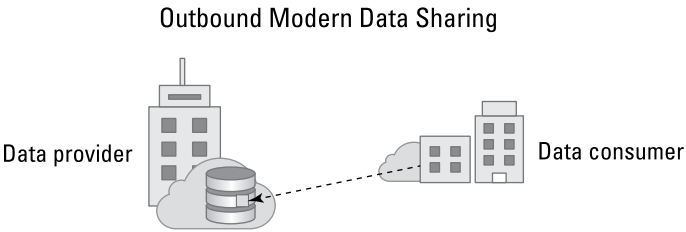

FIGURE 3-1: An organization, acting as the data provider, shares data with its supplier, the data consumer.

For example, in a vendor-supplier relationship with data sharing, a supplier knows in advance when to replenish the stock of a particular item. Well-managed inventory also prevents overstocking, minimizing the need to significantly reduce prices.

Sharing between organizations: Inbound

Increasingly, organizations engage outside service companies. These contracted companies can specialize in logistics, shipping, marketing services, or sales operations, just to name a few. For example, a large retailer would collect massive amounts of demographic data about its target customers. The retailer would then share this data as a data provider to a data analytics company.

STREAMLINING INTERNAL OPERATIONS WITH DATA SHARING

CASE STUDY

Rakuten Rewards is a global conglomerate that since 1997 has helped shape the way people shop online, offering cash-back deals and shopping rewards on the world's largest selection of products and services. Each company division or subsidiary has specific legal requirements and permissions related to which data can be shared, creating a complicated data sharing infrastructure within the company.

Not only was the internal process of sharing data cumbersome, but it also prevented business units from accessing each other's data sets for making more-informed business decisions.

Rakuten chose a modern, cloud-built data platform that enables its business units and subsidiaries to easily receive governed and shared data within minutes and combine it with their own data sources for even deeper insights.

As Rakuten Vice President of Analytics Mark Stange-Tregear recounted, the resulting transformation of internal data sharing "not only allowed us to easily share data with groups outside our legal umbrella but also allowed other groups with specific expertise or niche analytics skills to work on our data sets without us having to hire outside resources, making us a more nimble organization."

From there, the analytics company would analyze the data for the retailer. It would then provide the analysis back to the retailer in the form of an inbound data share, as shown in Figure 3-2.

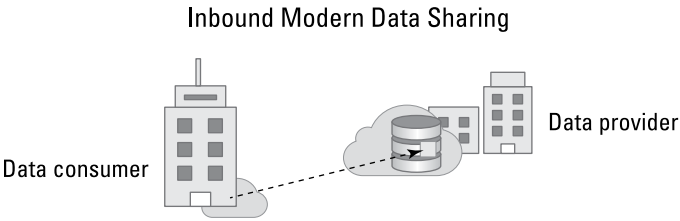

FIGURE 3-2: The organization is the data consumer, accessing the data from its outside data analytics vendor, which is the data provider.

In other scenarios, the organization contracts a service provider to perform a function the organization chooses not to perform in-house. In turn, the service provider generates data as a result of that service — data that belongs to the organization, which is the service provider's customer. With inbound data sharing between organizations, the data generated by the service provider is shared with its customer. The customer then executes additional analytics to develop deeper insights and value from additional data generated outside its data center but within its business ecosystem.

Monetizing data

Data can also take on more significance today than just day-to-day collaboration. Data is a business asset — a currency. As such, data can offer different types of value depending on the organization that wants to consume that data. Thus, as with any asset, data has value. To monetize the value of its data, a provider can sell data to consumers that can then use the data to advance their own business objectives (see Figure 3-3).

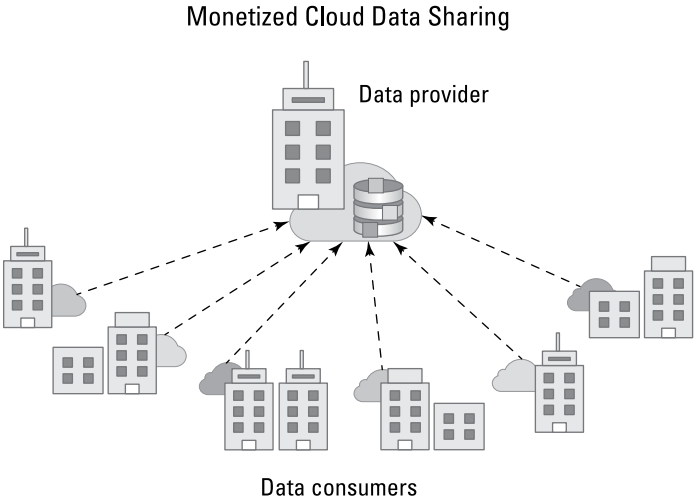

FIGURE 3-3: An enterprise data provider creates new revenue opportunities by sharing data with other organizations acting as data consumers.

Data consumers can use shared data without having to capture and collect it themselves. They can benefit directly from analyzing that data or combining it with other data to enhance its value.

But harnessing the value of data — either for consumption, mass collaboration, or value-added business opportunities — requires an easy method for enabling data access without actually moving the data. Traditional data sharing methods are too costly, risky, and labor intensive. Monetizing data requires the ability to easily and affordably create a secure, self-service business model to share data between data providers and data consumers.

SHARING DATA TO HELP SERVE 90 MILLION GAMERS

CASE STUDY

PlayFab provides back-end services and data logistics for its game studio customers that serve 75 million monthly average gamers and 15 million daily average gamers. PlayFab shares player data for each game with each respective game studio. PlayFab also anonymizes aggregated data across all studios and shares those results with all its game studio customers. On the receiving end of the shared data, game studios optimize games based on the shared player data from PlayFab.

When gathering data, and sharing it with game studios, PlayFab regularly encountered challenges such as capturing the right data the first time, getting data from client devices and moving it into the data pipeline, managing constantly changing data schemas driven by new game data events, and reducing soaring costs due to moving and transferring high volumes of game data. To solve these challenges, PlayFab adopted a modern, cloud-built data platform.

With modern data sharing, PlayFab sets up secure, governed, and live views of the data with each game studio under a straightforward, self-service business model. This method avoids ETL entirely. Secure and governed views guarantee each game studio's data is truly isolated and game designers get direct access to a direct feed of live game data, without any custom import required.

IN THIS CHAPTER

- Exploring a modern data sharing architecture
- Allowing access to database tables and views for data consumers
- Looking at a modern data sharing example
- Protecting sensitive data with secure views

Chapter 4
Enabling Live Data Sharing

In this chapter, you learn how a modern, built-for-the-cloud data platform architecture helps data providers and data consumers overcome traditional data sharing challenges. You also learn how to use real-time data sharing from inside a modern cloud data platform environment to quickly and easily enable secure and governed views of live data for your data consumers.

Exploring Modern Data Sharing

As discussed in Chapter 2, data sharing is a multi-faceted challenge. But traditional methods of data sharing fundamentally address only one part of the challenge — providing data consumers with access to a provider's data. Although traditional data warehouses and *data lakes* (repositories that store massive amounts of raw data until needed for analysis) were designed to make data usable, their underlying architectures are not capable of modern data sharing — that is, providing data access to data consumers without having to move the provider's data.

Traditional data sharing is slow, and it reduces an organization's ability to execute quickly. In addition, a lack of security

and governance, among other things, means traditional data warehouses and data lake architectures cannot support unlimited concurrent access by data consumers or real-time data changes by data providers without cumbersome unloading and transferring of data, as shown in Figure 4-1. This puts data consumers at risk of operating on stale (static) data.

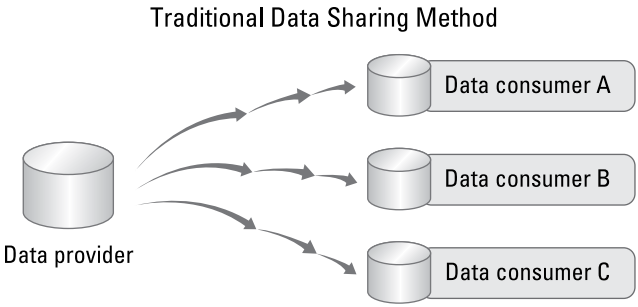

FIGURE 4-1: Traditional data sharing requires cumbersome, multi-step processes from both the data provider and its consumers.

The lack of a comprehensive solution creates a struggle for data providers and consumers to easily share data and ensure data consistency. It also limits the ability to monetize data.

Winning with modern data sharing

With modern data sharing inside a modern, cloud-built data platform, in a matter of minutes you (as a data provider) can enable live access to any of your data for any number of data consumers, inside or outside your organization. You can share data across internal business units, with business partners across your ecosystem, and with external organizations to easily support richer analytics, data-driven initiatives, new business models, and new revenue streams (see Figure 4-2).

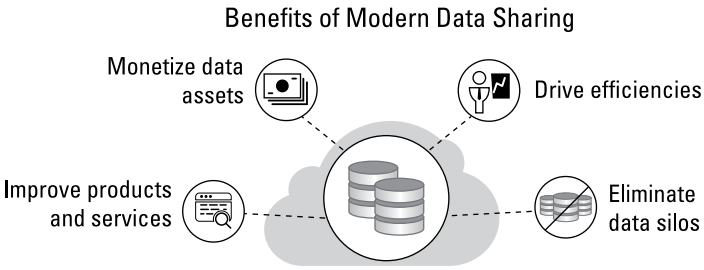

FIGURE 4-2: Providers can improve business with modern cloud data sharing.

With modern data sharing, ready-to-use data is immediately available in real time. In a modern cloud data platform architecture, query speeds on shared data are exponentially faster and fortified with limitless storage and compute resources. Modern data sharing extends the architecture and functionality of the modern cloud data platform to share data. Enterprises can grant read-only access to their live, ready-to-use data (structured and semi-structured) in a secure and governed environment. Data consumers can then choose to combine (JOIN) data from other organizations to augment and deepen their data analytics.

REMEMBER

Only the scalability, elasticity, and flexibility of a multi-tenant cloud data platform makes it possible to store data from diverse sources and share that data among a large number of data consumers without contention or competition for resources.

Making data sharing easy

Enterprises can realize the business benefits of modern cloud data sharing thanks to a cloud-built data platform, which

- **Eliminates movement and copying of data:** Modern data sharing offers direct, real-time access to live data in a secure, managed, and controlled environment.
- **Provides ready-to-use data:** Data consumers get the full capabilities of a data warehouse, allowing them to query and analyze shared data as soon as they're given access to it. They can combine shared data with their own data. Security, governance, data schema, and metadata are all provided within the modern data platform.
- **Protects personally identifiable information (PII) and complies with industry requirements:** Data providers can easily and securely share data while ensuring no PII or sensitive data is compromised by using advanced sharing functions, while still enabling the data consumer to make full use of the data. In the same way, data consumers who are interested in sharing data for enrichment by a data provider can do so without exposing PII.
- **Enables data sharing without added costs:** Modern data sharing eliminates the duplicative costs of building the infrastructure needed to store shared data, since data consumers view the shared data directly from the data provider without having to copy or move data.

» **Enables data sharing with unlimited data providers and consumers:** A modern cloud data platform can serve an unlimited number of data providers and consumers, with full transactional integrity and data consistency.

Modern cloud data sharing eliminates the delays, cost, and friction of existing methods, which provide only primitive mechanisms for data publishing, access, and control. Modern data sharing is built on three key architectural innovations, which we discuss in the next few sections.

Decoupling of storage, compute, and services

The separation of storage and compute resources is a fundamental part of a modern data sharing architecture, as shown in Figure 4-3. All data is stored, in optimized form and without any loss of data fidelity, in the cloud. A single copy of the data stored in a modern cloud data platform — a single source of truth — can be accessed concurrently by any number of independent compute clusters, enabling an organization to perform any number of internal workloads, such as analytics.

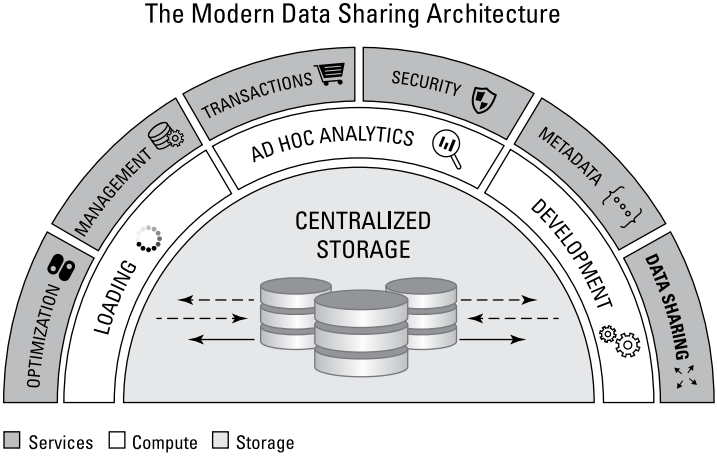

FIGURE 4-3: A modern data sharing architecture built for the cloud with storage, compute, and services completely separate but logically integrated.

Decoupling of storage and compute is also critical for sharing data. It enables data consumers to directly access shared data, using their own data platform compute power. But data consumers don't pay for storage costs (because the shared data doesn't

move), and the data provider doesn't pay for any of the compute resources that a data consumer uses to analyze shared data.

Managing multi-tenant transactions

Making shared data usable requires access to data and coordination across all data consumers to ensure consistency, security, and performance. The services layer is a key part of a modern data sharing architecture. Global metadata, transactions, and security are all managed from here, making the services layer the control tower that tracks, logs, and directs access to data for every database element and object contained within the data platform, as shown in Figure 4-4.

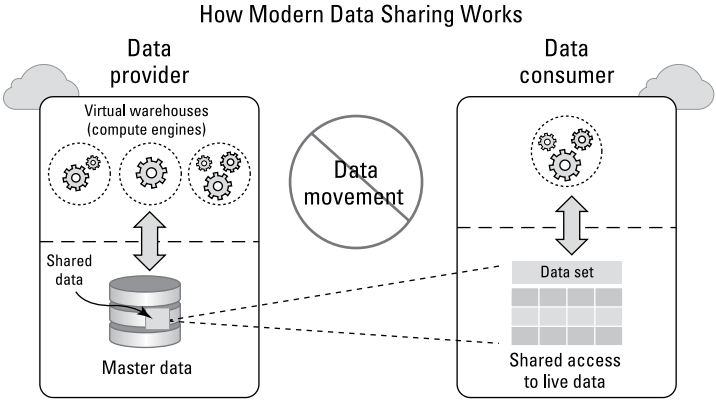

FIGURE 4-4: Metadata in a modern data sharing architecture enables live data access between a data provider and data consumer, without moving data.

Additionally, the services layer provides transactional consistency across all data providers and consumers, ensuring all data users see a consistent view of live and up-to-data data. A data provider can update shared data in real time. Likewise, after transactions are committed, all data consumers can simultaneously view the data provider's updates and immediately query the shared data — all with transactional, ACID-based consistency.

ACID is a consistency model that defines a set of properties to ensure transactions in a relational database are valid, even in the event of multi-statement transactions and processing errors, as well as power failures and crashes. The properties of ACID are:

- **Atomicity ("all or nothing"):** Every operation in a transaction must succeed for the transaction to be completed.

If even a single operation fails, the entire transaction is rolled back and the database state is left unchanged.

- **Consistency:** The completion of any transaction brings the database from one valid state to another valid state.
- **Isolation:** Concurrent transactions do not contend for access to the data and are run as if each transaction executed sequentially.
- **Durability:** After a transaction is committed, it remains committed.

Data providers also benefit by having to share only the specific data the consumer needs, which makes the process easier and more secure.

Gaining unlimited concurrency

With modern data sharing, shared data can be accessed by large numbers of concurrent data consumers, as shown in Figure 4-5. In contrast, the architecture of traditional data warehouses forces all users to compete for resources, creating a struggle to deliver optimum performance and consistency. Automatic scaling of concurrency takes simultaneous query processing even further in modern data sharing by automating the scaling of additional compute engines without manual intervention.

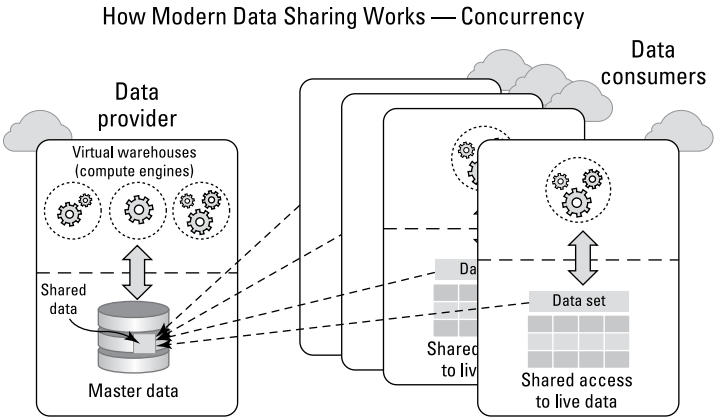

FIGURE 4-5: Unlimited concurrency with a modern data sharing architecture.

Using Modern Cloud Data Sharing

Modern cloud data sharing allows access to database tables and views for any user of a modern cloud data platform. When a data provider shares data with a data consumer, the database object and view are all from within the data provider's environment.

In addition, a modern cloud data platform gives data providers granular control of access to database tables and secure views via *shares.* Data consumers can query a provider's database only if granted access privileges. Once the data provider creates a share, the data consumer can then query the data.

Instant access, without data copying or movement, is made possible because all database objects are maintained and updated only in a modern cloud data platform, and orchestrated by its global metadata management services.

Learning How It All Works

As a data provider, the first step to sharing data is to specify what database tables and views to share with specific data consumers. This is done via a data share object, effectively an "empty shell" that houses the references to the actual database and the shared database objects. Data shares are first-class objects in a modern cloud platform environment for which it provides a set of data definition language (DDL) commands for creating and managing shares. Commands include `create share`, `alter share`, `drop share`, and others. Access commands include `grant` and `revoke` privileges. Once a share is created, the data provider grants access to the specific database and database objects it shares. The SQL semantics are as follows:

1. **Create the share.** The following example creates an empty share named `sales_s`:

```
create share sales_s;
```

2. **Add privileges for objects in the share.** Grant usage on the primary object before granting usage on any objects within the primary object. For example, grant usage on a database before granting usage on any schemas contained within the database. Complete all grants for the data share before

adding the data sharing consumer(s). The following example grants privileges for the `sales_db` database, the `aggregates_eula` schema, and the `aggregate_1` table to the data-share object:

```
grant usage on database sales_db to
   share sales_s;
grant usage on schema sales_
   db.aggregates_eula to share sales_s;
grant select on table sales_
   db.aggregates_eula.aggregate_1 to
   share sales_s;
```

3. **Confirm the contents of the share:**

```
show grants to share sales_s;
```

4. **Grant access to the share for the intended data consumer(s).** The following example makes the `sales_s` share available to data consumers A and B:

```
alter share sales_s add
   accounts=data_consumerA, data_
   consumerB;
```

`data_consumerA` and `data_consumerB` now can see their individual shared data and can create their databases from the shared data as necessary.

Controlling Access to Shared Data by Using Secure Views

What if you have sensitive data in your database? With modern data sharing, you are not limited to sharing entire databases or entire database tables. If portions of a table are subject to strict security and confidentiality policies, sharing the entire table exposes the sensitive data. With a command utility called *secure view*, you can control access to shared data and avoid security breaches, as shown in Figure 4-6.

For example, for online retailers to plan inventory levels, they need to share merchandise and sales data with their distributors. However, the table within the database that contains the sales data also contains sensitive customer ID information, which must be blocked and protected.

How Modern Data Sharing Works with Individual Secure Views

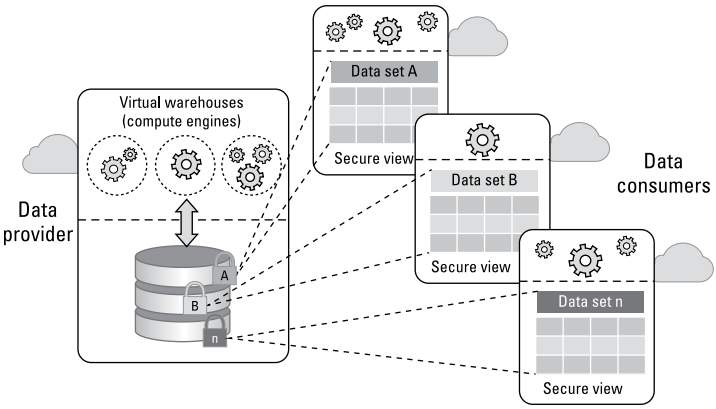

FIGURE 4-6: Secure views with modern data sharing allow data providers to protect access to sensitive data.

To demonstrate how to accomplish this using a secure view, follow the SQL semantics from the previous section to create a data share object, `sales_s` for database `sales_db`. For this example, assume `sales_db` and `sales_s` already exist. The schema for `sales_db` is named `public`, within which, table `unitsales` is constructed as `unitsales (customerid, sku, date, qty)` and is populated with data.

When you need to plan new inventory with your distributor, you want to provide access to the unit sales data, but not the customer data, which is sensitive. Therefore, a secure view named `distributor_sales_data` is created from the `unitsales` table, just for the distributor.

With modern data sharing, the steps are accomplished as follows:

1. **Create the secure view.** Assuming the database and schema are already created and populated with data, the next step is to create a secure view on the `unitsales` table:

```
create secure view sales_db.
   public.distributor_sales_data as
   select sku, date, qty
   from sales_db.public.unitsales;
```

This logic creates a secure view named, but without the sensitive `customerID` data. The data included in the view are `sku`, `date`, and `qty`.

2. **In the sales_s share container, add privileges for the secure view:**

```
grant usage on database sales_db to
   share sales_s;
grant usage on schema sales_db.public
   to share sales_s;
grant select on view sales_db.public.distributor_
   sales_data to
   share sales_s;
```

This logic enables the share (container), `sales_s` to have privileges for the `distributor_sales_data` secure view.

3. **Confirm the contents of the share:**

```
desc share sales_s;
```

The user is provided a readout confirmation of the share, `sale_s`.

4. **Grant the distributor access to the share** (it is assumed the distributor is also a modern cloud-built data platform user):

```
alter share sales_s add
   accounts=<distributor_name>;
```

5. **To see your share:**

```
show shares;
```

As this demonstrates, data providers can easily share data while controlling data consumers' access to data with a secure view. Sensitive data is protected, and data consumers gain access to non-sensitive data for their own analytics, without the need to copy or move data.

IN THIS CHAPTER

- Collaborating on truly shared data
- Leveraging low-cost compute power
- Using analytics to drive faster business decisions
- Taking advantage of data commercialization opportunities

Chapter 5
Assessing the Impact of Modern Data Sharing

In this chapter, you learn how modern data sharing enables real-time collaboration. You also learn which technologies and trends have enabled a modern data sharing architecture, as well as how modern data sharing can enable organizations to quickly create new business assets from data.

Modern Approaches to Data Collaboration

Traditionally, data sharing has meant sharing copies of data (see Figure 5-1). This creates a myriad of challenges, including the following:

- Multiple, static, and out-of-sync versions of the same data exist in different environments, across multiple data silos.
- No practical single source of truth or governance exists for data in the organization.

- Critical business decisions are made based on outdated, incomplete, or inaccurate data.
- Electronic discovery costs escalate when multiple sources of data within and outside an organization must be identified, searched, and produced for litigation support.
- The potential number of data breaches and accidental data loss/disclosure risks multiply, along with their associated costs, such as breach notifications, credit monitoring services, damage to an organization's brand, customer churn, litigation, forensic analysis, and recovery.

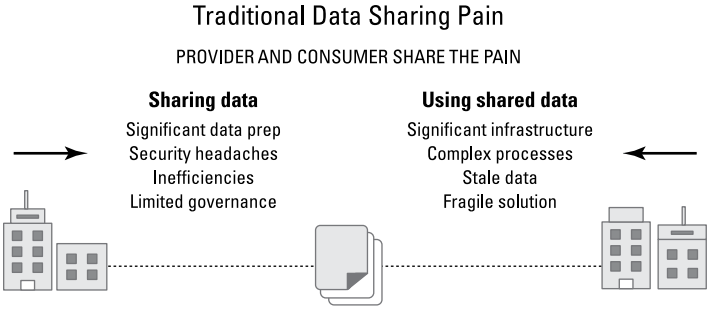

FIGURE 5-1: The difficulties of traditional data sharing methods.

Modern data sharing enables fast, cost-effective, and secure collaboration inside and outside the organization by providing data consumers with real-time access to a single copy of the same data. With modern data sharing, data providers can share data using the same modern cloud data platform they use to run their organizations.

The Democratization of Computing

Enterprise-grade computing power is everywhere, and it's getting even more powerful, faster, and less expensive every day. The cloud has further accelerated this trend, making limitless

computing resources easily available at massive scale to any-sized organization.

With this much power, business users perform many analytical computing tasks rather than submitting them as queued job requests for IT administrators to perform. Business analysts have the power to run advanced analytics against large data sets. Data scientists can perform predictive analytics and develop machine learning algorithms that serve as the basis for artificial intelligence. Business executives can execute swiftly with up-to-the-minute analytics dashboards; product management executives achieve faster time to market for new products and services. In short, massive computing power is now within everyone's reach.

Data-Driven Business Decisions

To ensure the most complete view of customers, today's organizations require a multi-channel approach to gathering data from various channels (such as websites, mobile devices, point-of-sale terminals, call centers, and so on). Traditional data warehouses not only fail when sharing data, but also perform poorly when needing to uncover meaningful relationships between different forms of structured and semi-structured data.

Semi-structured data, such as JSON data, does not conform to the standards of traditionally structured data and includes data generated from newer data sources, such as social media sites, clickstreams, mobile devices, and Internet of Things (IoT) devices.

But collecting and analyzing all this data from multiple channels is exactly what successful organizations must do. Outcomes from insightful analytics can help them better target new products and services. So, it's understandable those that want to provide data as a service, or as a value-added business asset, are just as interested in delivering access to data quickly and easily, so other, non-competing organizations can benefit.

CASE STUDY

REDUCING COSTS AND STREAMLINING OPERATIONS

Heap Analytics provides web and mobile analytics for companies across a number of industries. To share raw data with customers, the company would drop files into Amazon S3 buckets, or host files with a cloud-based vendor for customers who weren't sophisticated enough to handle all the extract, transfer, load (ETL) activities required to then analyze the data. Heap's costs and efforts included:

Engineering work: Five to ten hours of work per week debugging pipelines, performing cluster maintenance, handling unique customer setups, and resolving resource contention

Customer service work: Five or more hours of solutions time per week getting proper access; scheduling cluster operations; and creating, maintaining, and resizing hosted clusters with a traditional cloud data warehouse

By leveraging a modern cloud-built data platform with modern, secure data sharing capabilities, Heap eliminated the engineering work and reduced customer service work significantly. It also saw gains in other areas. For example, Heap can now share data to its customers, who can access it immediately and see value sooner. The modern solution decreased Heap's support burden by requiring less back and forth on permissions and security approvals, no custom setups and less infrastructural variance, and painless cluster maintenance, with no resource contention.

The Commercialization of Data

By removing traditional barriers of on-premises and cloud data warehouses, modern cloud data sharing introduces at least four new economic opportunities that enable enterprises to share data as packaged and monetized assets, quickly and securely, powering a true data economy:

- **Data monetization:** Many companies produce and sell data, some of which started nearly 100 years ago. Now, with modern data sharing, any organization can turn its data,

regardless of its size, into a business asset by charging for access to slices of its data repository. This low-cost, zero-headache solution enables data companies to immediately meet data consumers' urgent demands for fresh data with up-to-the-minute accuracy, which maintains the highest value for data. Organizations have greater power and platform capabilities to support, move forward, and implement data monetization strategies.

- **Data sharing with business partners:** Sharing data directly with business partners is not new. But the effortless sharing of live data is groundbreaking. Modern cloud data sharing enables the following:
  - Organizations can share data instantly with business partners in their ecosystem — for example, supply chain, distribution, marketing, third-party sales, and so on.
  - Data consumers can query against live data to mutually benefit both the data consumer and data provider, no matter which organization owns the data and no matter which one takes on the role of either the data provider or data consumer.
  - Business results can be delivered faster from faster execution of analytics based on internal and external data sources accessed immediately, without overhead, and at a fraction of the cost of traditional data sharing methods.
- **Breaking down enterprise data silos:** Modern data sharing eliminates data silos strewn across an organization and beyond by allowing you to store structured and semi-structured data in a modern cloud data platform. You can share seamlessly, without downloading or replication. Your systems that were previously siloed can be tightly integrated, without manual integration or the need for data pipelines.
- **Near-zero management to reach more data consumers:** Only a modern cloud data platform eliminates the traditional and time-consuming methods needed to manage a legacy data warehouse. Performance is built into the modern data platform, so there's no infrastructure to tweak, no knobs to turn, and no tuning required. With near-zero management required, organizations can pursue more, far-reaching data sharing strategies to target a larger base of data consumers across their business units, with business partners inside their ecosystem, and with other organizations as part of the data economy.

THOUSANDS OF DATA CONSUMERS — ONE SOURCE OF TRUTH

CASE STUDY

PlaceIQ aggregates, collects, and anonymizes data from thousands of applications on mobile devices. It then makes the data available for companies that want to target and reach out to mobile consumers based on their location and behavior. PlaceIQ customers may include marketing companies, advertising agencies, and product producers, to name a few.

One data source — thousands of self-service users: PlaceIQ customers execute highly targeted campaigns based on the individual usage and geographic movement of mobile device users. To do so, each PlaceIQ customer may require a unique slice of the data to qualify potential prospects. The cost and effort of traditional methods made it impossible to effortlessly deliver the data within a self-service business model.

Effective business models — zero-data management: A modern cloud data platform, built for data sharing, resolved the challenges faced by PlaceIQ to scale and deliver thousands of individual data consumer subscriptions with governed, secure slices of the data. PlaceIQ now uses far fewer resources to manage its single source of truth, while enabling its data consumers to self-serve their own data subscription. With modern data sharing, PlaceIQ focuses on developing new use cases for its data that are differentiated and priced accordingly for every customer.

Simplicity for PlaceIQ data consumers: PlaceIQ customers also benefit by having a simple environment to assess shared data before merging it with their own data sets. In doing so, PlaceIQ customers do not have to sort through large amounts of irrelevant data to find the information they desire. Plus, all the data from the modern data platform meets common security and certification requirements. PlaceIQ customers can blend data sets without compliance issues.

IN THIS CHAPTER

- Creating data exchanges to extend data to third parties
- Setting up real-time data sharing
- Overcoming challenges with traditional data marketplaces

Chapter 6
Using Exchanges to Share Data and Monetize Insights

Exchanging data within a controlled ecosystem can yield rich analytics and deep insights, leading to more-informed decision-making. In this chapter, you learn how your organization can participate in a data exchange to leverage third-party data and create new data assets.

Examining the History of Data Marketplaces and Exchanges

A data marketplace, also known as a data exchange, allows participants to discover useful data sets. Traditional data exchanges are accessible to participants through a portal or app store-like environment. Many data providers offer a basic data set that is free. They then charge for additional data sets and any data services they may provide.

Thousands of online marketplaces today link buyers and sellers. Typically, the data vendor handles data transformation, preparation, and loading, while the marketplace is in charge of discovery, collaboration, licensing, and auditing. It's a huge opportunity: In its September 2018 report, "Value of Data: The Dawn of the Data Marketplace," Accenture estimates the data-as-a-service (DaaS) market is poised to reach $10.4 billion by 2021.

In contrast, a modern data exchange is built on the architecture of modern data sharing, enabling one-to-one, one-to-many, and now many-to-many data sharing relationships between data providers and data consumers. It eliminates the cost and effort associated with traditional data transfer, which plague most data marketplaces.

Most organizations already share their data, or plan to do so, via data sharing and their cloud-built data platform. There is an immense and rapidly expanding marketplace for monetizing data. In addition to earning additional revenue by sharing slices of your data to data consumers, it's an opportunity to forge new alliances and glean new insights by combining your data with data and data services from other providers. We look more closely at data monetization in Chapter 7.

WHY USE A DATA EXCHANGE?

Use an exchange to provide and consume shared data and data services to:

- **Create market opportunities:** Tap new revenue streams by charging data consumers for secure and governed access to read-only versions of your data sets.
- **Improve security and data access:** Minimize the risk and cost associated with copying data by opening secure, real-time connections to read-only data sets.
- **Improve the customer experience:** Give customers a fast, secure, and cost-effective way to access data and data services, increasing consumption and improving satisfaction for your data subscribers.

Moving into the Modern Age with Real-Time Data Sharing

Most companies derive value from a data exchange fairly quickly as they leverage the deeper insights from additional data and data services. Those data consumers acquire data from third-party providers in several ways, including via an API, using a self-service interface for discovery and analysis, buying raw data via download or file transfer, and using an app that reveals trends and insights. After that, two common questions arise:

- How do I scale my data-sharing efforts to include external parties, such as customers and partners?
- How do I find relevant data I can easily source from third parties without bogging down my analytics operations?

The answer to both questions hinges on architecture. Modern data exchanges don't require data movement; extract, transform, load (ETL) technology; or constant updates to share data with consumers. There is no need to transfer data through File Transfer Protocol (FTP) or to develop and maintain APIs. And since data is shared rather than copied, no additional cloud storage is required. *Data providers* can easily publish data and make it available for instant discovery, query, and enrichment by *data consumers*, as shown in Figure 6-1.

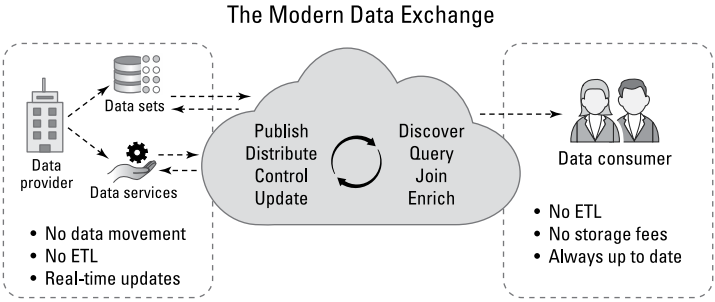

FIGURE 6-1: An architecture for an efficient, real-time data exchange.

Establishing the Right Architecture for Exchanging Data

Developing and maintaining APIs levies a heavy toll on data providers and data consumers. The providers must configure and manage the APIs, while consumers have to establish custom links to all marketplace vendors with which they want to do business. Instead, the ideal data exchange provides the following mechanisms for both providers and consumers:

- Facilitate real-time access to live data sets without APIs or ETL processes.
- Upload structured and semi-structured data sets without having to move or transform the data.
- Create different views of the data and share these views securely while protecting personally identifiable information (PII).
- Provide data in a ready-to-use format to data consumers.

Some cloud vendors offer data exchange options to their customers by leveraging the inherent capabilities of their multi-tenant SaaS architectures. To ensure good performance in these exchanges, leading cloud vendors should offer a multi-cluster, shared data architecture that separates compute and storage resources. This allows data providers to incur almost no cost to share data and allows data consumers to pay only for the compute resources, not storage, to analyze the shared data.

To maximize convenience, an exchange should be free to join and should allow participants to instantly publish data sets, ensuring other authorized users of the ecosystem can always access up-to-date data. The data storage layer should reside in a scalable, pay-as-you-go cloud layer, such as Amazon Simple Storage Service (S3), Microsoft Azure, or Google Cloud Platform (GCP). The data platform provider wraps these storage assets in a services layer to ensure everything is secure, properly maintained, tuned, and optimized for self-service access. This type of architecture makes it possible to efficiently exchange data from one centralized system, with dynamic elasticity. Data providers can share an unlimited amount of data, yet recipients pay only for the data and resources they actually use.

CREATING THE IDEAL DATA MARKETPLACE

CASE STUDY

Weather data is an important and vital asset for business intelligence across industries. In the U.S. alone, weather accounts for more than $600 billion a year in lost revenue. Weather strategies help companies forecast sales, mitigate risk, adjust transportation routes, and confidently make decisions.

The Weather Source team has 75 years of meteorology and climatology expertise. The company provides weather data from the year 2000 to present and forecasts of up to 15 days. Each of Weather Source's hyper-local solutions can be tailored to the points of interest most relevant to a customer's business.

Weather Source joined a data exchange to expand access to its data delivery services. The data exchange, built on top of a secure data sharing platform, offers an easy way for businesses to access data.

With the exchange, Weather Source is now able to reduce costs and effort involved in publishing its data sets. Data transformation, loading, and reconstruction are no longer required. New objects become immediately available to all consumers, providing real-time data access across the entire ecosystem. This more efficient operation offers customers better access to Weather Source data in a SQL-friendly format that's always up to date.

TIP

If you want to ensure long-term flexibility and market expansion for the data exchange you join, work with a cloud data platform provider that can replicate data among multiple geographic regions as well as store data in many different clouds — including AWS, Microsoft Azure, and Google Cloud — all as part of one seamless exchange. That way, you won't be locked into one type of object store or one particular cloud vendor as your data exchange evolves.

Some cloud data platform providers allow their customers to create *private data exchanges* to share and monetize data and data services within a self-contained ecosystem. Data platform customers can leverage the exchange to set up new lines of business by making their data and data services available to their end customers and partners. Consumers can easily access these data exchanges, paying only for data and data services they use, as per the terms set by the provider.

GETTING HELP FROM YOUR DATA EXCHANGE VENDOR

TIP

A modern data exchange should provide many unique benefits, including:

- Access to thousands of data sets across industries and functions
- An easy way to search for and discover data for deeper analytics
- Access to live data sets without manual intervention
- Full access to granular and historical data
- No need to build and maintain ETL processes
- No added data storage costs because data isn't moved or copied
- Access to value-add data services to enhance your data analytics

REMEMBER

Older exchanges depend on FTP downloads, ETL transformations, and API connections, to name a few. Modern exchanges use streaming data services or SQL queries to extract information, without the need to transform or copy data.

IN THIS CHAPTER

- Discovering data monetization opportunities
- Evaluating governance and distribution models
- Pricing your data as a service

Chapter 7
Monetizing Your Data

Data monetization is a process by which a data provider charges data consumers a fee to gain access to the provider's data or data services, so a data consumer can enrich its existing data sets to benefit its business and its customers. Some providers offer access to the data itself while others offer services, such as data modeling, data enrichment, and data analytics.

Nielsen is a pioneer in this space, and its service is so pervasive it has become a household name. For more than 90 years, the company has collected, analyzed, and sold consumer data to media companies, advertising firms, retail organizations, and many other industries. Nielsen's proven business model has inspired numerous niche data sharing opportunities. For example, in the financial services industry, some companies collect stock market data, package it, and sell it to brokers and hedge funds. In the gaming industry, PlayFab is well-known for providing gaming and analytics services for its live gaming studio customers (consumers), powering more than 2,500 games that serve more than 1 billion gamers. PlayFab provides its data consumers real-time analytics, dashboards, and custom reports about their players' activity, so studios can optimize their gamers' experience and drive more business.

Examining Industry Opportunities

Organizations can monetize data in countless ways, as shown in Figure 7-1. For example, a telecommunications company can sell location data to help retailers target consumers with ads. Consumer packaged goods companies can share purchasing data with online advertisers — or directly with customers. A logistics company might sell data about transportation patterns and shipping activity as an indication of economic trends. Transportation agencies can collect data from tollbooths and bridges to optimize navigation apps and minimize traffic congestion. From tracking engine performance to monitoring human biometrics, data sharing reveals previously unobtainable insights so organizations can spot trends, make predictions, and take corrective action.

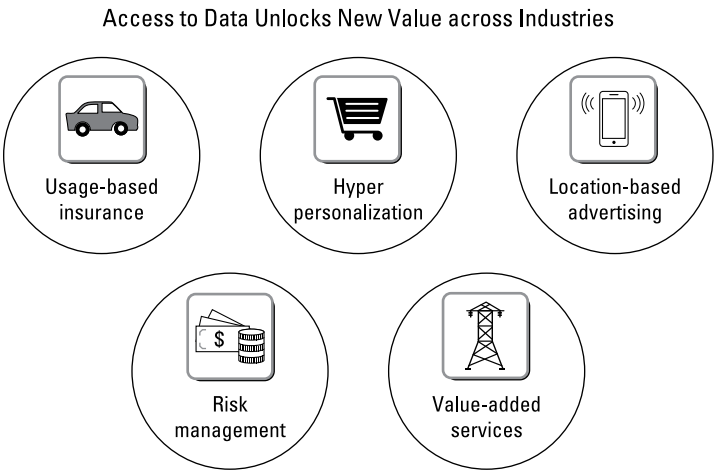

FIGURE 7-1: Data sharing unlocks valuable opportunities in every industry.

Acquiring data from other organizations likewise can enhance internal operations. For example, customer service departments may want external data about callers to segment prospects and knowledgeably route opportunities. Marketing professionals look to external data sources to target their messages and tailor their campaigns. Sales teams collect and aggregate third-party data about the organizations they sell to, so they can properly qualify leads. Risk analysts gather data from cybersecurity experts to help identify network intruders and circumvent fraud.

Getting in Front of Data Monetization Trends

Modern data sharing is growing every year as the volume of data that organizations collect expands exponentially. A Forrester Research survey of global data and analytics decision makers revealed that more than 75 percent of respondents are expanding their use of external data and nearly 50 percent of these companies already commercialize their data.

From Internet traffic to weather data, social media trends to purchasing patterns, your data may be valuable to third parties. Whether or not you are monetizing your data, chances are that your competitors are. In IDC's "2019 Predictions for Digital Transformation," the research firm predicted that 80 percent of enterprises will create data management and monetization capabilities by 2020, and that by 2023, 95 percent of entities will have incorporated new digital KPI sets.

To determine where the opportunities lie, ask yourself these questions:

- Are your customers requesting access to specific data they don't have access to today?
- Are they in search of more data, better data, better insights, or new insights?
- Are customers complaining about how you currently share data? Is fulfilling their requests a burden to your staff?

TIP

Start by identifying the area of value for your customers. This will help you determine what type of data and access to give them, and how you might extend your current business practices to maximize future data sharing opportunities.

As you consider a data monetization strategy, consider these options:

- Selling access to raw data
- Charging for specific data services (such as data enrichment, data modeling, and so on)
- Offering data analytics for a fee

Observing Good Data Governance

For all types of data sharing scenarios, you must observe all government and industry regulations regarding the security and privacy of consumer data, as shown in Figure 7-2. Regardless what data you plan to share, setting conditions and limitations is important. Appoint a data governance champion to pay attention to these legislative mandates; set internal policies and procedures; respond to customer requests; and control access to data using secure views, secure functions, and secure joins — as described in the next chapter.

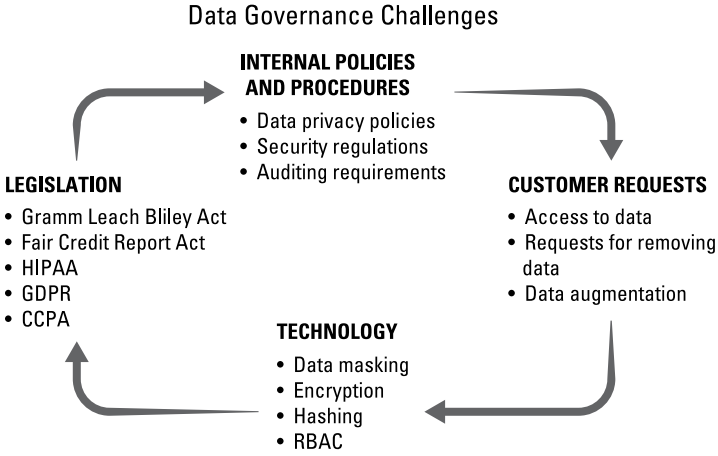

FIGURE 7-2: Make sure your data sharing practice complies with all pertinent regulations and doesn't compromise customer privacy.

Evaluating Distribution Channels

As discussed elsewhere in this book, sharing data requires a secure, flexible business model that simplifies the exchange of data between data providers and data consumers. There are several potential distribution channels for selling data and data services and building market momentum:

- **Direct to customers:** The advantage of this approach is that you're selling to a well-known audience, with the chance to deepen those relationships. The disadvantages are that it's a limited market, and you have to construct a dedicated sharing platform.

- **Through a data broker:** The advantage is ease of adoption because the broker supplies a built-in clientele. The disadvantages include lack of control over the data and the inability to forge direct relationships with users.
- **Via a data exchange**. As explained in Chapter 6, a data exchange is a scalable platform that supports flexible, bidirectional data sharing. This is a great way to build your brand and establish new types of direct relationships with customers. The only possible disadvantage involves adoption: If you're starting with an early-stage ecosystem, it may take time for the service to deliver on its full potential. In this case, look for a data exchange that will help you generate brand awareness and scale as part of its value proposition, versus a traditional data marketplace where your data sits there waiting for someone to find it.

Pricing Your Data Sharing Service

There are a number of ways you can bill customers and partners for your data. For data that is streamed or continuously updated, monthly subscription fees are common. However, if your data set is relatively static, then most of the value is derived at the initial time of consumption, with less value over time. In these instances, it may be more cost-effective to ask customers to pay a one-time fee rather than charge them for a subscription.

You can also consider offering tiered pricing according to different levels of service:

- **Free:** Select access to a limited data set
- **Standard:** Full access to the data set
- **Premium:** Full access to the data set plus mobile alerts, basic insights, custom dashboards, and role-based reports
- **Elite:** Premium plus predictive analytics and decision-support tools

DATA SHARING SERVICE OPENS NEW REVENUE STREAMS

CASE STUDY

Maintaining product quality in the fast-moving consumer goods (FMCG) industry requires rapid decisions related to inventory, warehouse management, shipping, and delivery. By helping grocery suppliers visualize trends in their data and share insights with retailers, Atheon Analytics keeps goods moving to the right place at the right time. Atheon's SKUtrak service is the leading flow-of-goods tracker for FMCG suppliers throughout the United Kingdom.

Fresh groceries are a byproduct of fresh data, and Atheon's new data sharing service, based on its cloud-built data platform allows the company to easily share data to SKUtrak customers, so they always see the latest insights but without complex data-copy or data-moving procedures. "We provide data directly from retailers and share it with suppliers and customers to make sure their products get to the supermarkets where they need to be," Atheon's head of service delivery Rose Ahearne said. "The data platform helped with scalability and created new product opportunities."

Having a cloud-built analytics platform to securely share real-time data is now an essential part of Atheon's business. Through visualizing data in SKUtrak, retailers and their suppliers can explore regular patterns and anomalies to identify opportunities for improvement across the entire grocery supply chain. This results in increased availability, reduced waste, and better-met shopper needs. "Data sharing offers a great new revenue stream," Ahearne added. "It has become part of our product portfolio. Developers and consultants aren't waiting around for data, and productivity has gone through the roof."

IN THIS CHAPTER

- Complying with government and industry regulations
- Keeping sensitive data secure
- Selectively sharing and controlling access to a data set

Chapter 8
Governing Your Data

Many organizations need to securely link their own data with data generated by their partners, suppliers, customers, and industry peers, but they have concerns about protecting personally identifiable information (PII), protected health information (PHI), competitive data, and other forms of fine-grained data.

It's a pressing issue: Today's businesses 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.

For the GDPR, as an example, there are three key areas to consider:

- Principles related to how personal data is processed and handled
- The rights of individuals (*data subjects*)
- Accountability, including being able to document and demonstrate data compliance

GDPR requires organizations to rein in the runaway complexity that multiple data sources create and to establish a modern data governance strategy on a global scale. Not only do most organizations keep their data in multiple locations (disparate data warehouses, data marts, servers, and clouds), but they also suffer from *hidden layers of data*. This phrase describes the countless copies of data that exist within organizations as a result of siloed teams and departments sharing information.

Copies of data are rampant when organizations rely on older technologies, such as traditional on-premises data warehouses, which require data to be exported and copied in order to be shared. And if distributing data sets via cloud storage, File Transfer Protocol (FTP), or APIs is hard to track, what about employees who manually share files and spreadsheets via thumb drives or email? Without the right technology, fulfilling compliance requirements is difficult, because when part or all of a data set is copied, you have to apply your efforts to multiple data sets, possibly in multiple locations.

Fortunately, fulfilling data privacy and protection requirements is significantly simplified when you have the right technologies in place for sharing that data. Instead of dealing with any number of data sets in any number of locations, you can establish a single copy of your data, governed by advanced technologies for accessing and sharing it.

REMEMBER

You can ease your compliance efforts by keeping a single "source of truth" of historical and live data in a single location and granting on-demand access to governed slices of that copy of data.

To determine what type of data security you need, start by identifying the data sharing scenario that most closely matches your use case. Throughout this chapter, we explain five such scenarios.

Sharing the Same Data with One or Multiple Consumers

In this situation, the need is simple: You have a table or view you'd like to share with one or many data consumers (see Figure 8-1). For instance, if a retail chain wanted to share its entire database with 50 franchises, it could use this method. The process is

relatively straightforward. First, the retailer would need to create a *share*, a database object that grants permissions and data access to the data consumers.

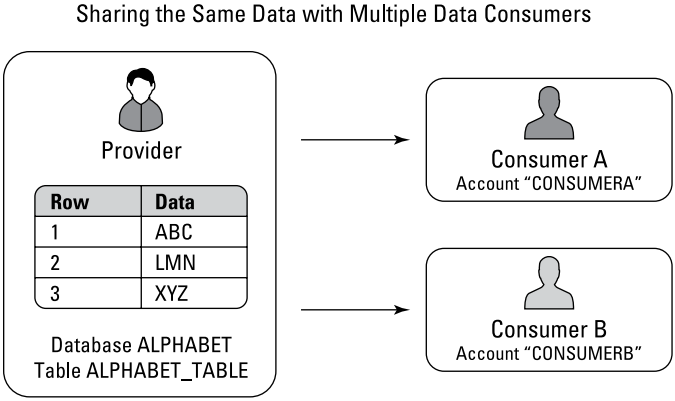

FIGURE 8-1: Using permissions to share the same data with multiple consumers.

To complete the process, the retailer would add permissions for all related database objects it wants to share, and then add the 50 franchise accounts to that share. This could be automated via a scheduled script or through a workflow or provisioning system using Python or SQL.

Sharing Different Subsets of Data with Multiple Consumers

This scenario is more sophisticated because it involves sharing different subsets of data with multiple consumers. For instance, a car manufacturer might want to share data with its dealers about production plans. Some dealers may be confined to a single state, while others span many states. Using this method, the manufacturer could share the pertinent data with each dealer, regardless of dealer location.

For example, Figure 8-2 shows data in one table broken up into two groups. Group 1 contains row 1, which is visible only to Consumer A. Group 2 contains rows 2 and 3, which are visible to all consumers.

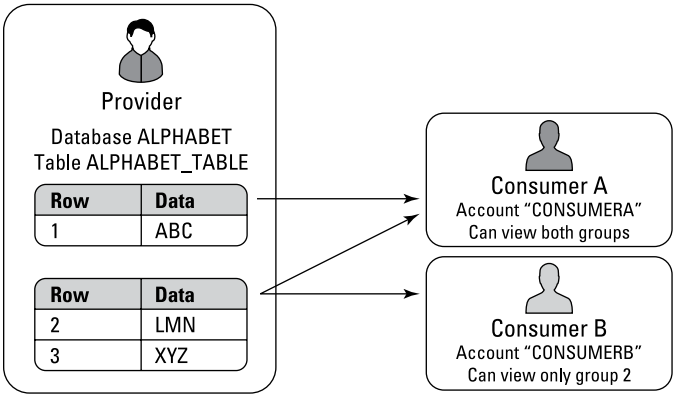

FIGURE 8-2: How to share the different subsets of data with multiple consumers.

Facilitating this kind of sharing requires a couple of additional steps. First, the data provider needs to create an *entitlements table* to match the dealer account names to the set of objects each dealer needs to access. Second, the manufacturer must create a *secure view* that controls the data each dealer can access when that dealer runs queries.

Using Secure Views to Provide Predefined Slices of Data

As explained in Chapter 4, if any portion of a table is subject to security and confidentiality policies, sharing the entire table could expose sensitive data. *Secure views* allow you to control access to the data and avoid potential security breaches.

Secure views are an effective way to enforce table, column, row, and even cell-level security when sharing data between organizations external to each other.

Data providers can each create secure views of their data and share access to those views with other users of the SaaS cloud data platform, even if these users are in other organizations. This is also a useful method for organizations that want to grant database

access directly to multiple end customers. Secure views allow these customers to see only their specific rows of data from each table, but not the rows that contain data about other customers.

When should you consider using secure views? The following are some guidelines:

- To allow data consumers to access data without compromising security
- When PII and PHI are involved
- To restrict data access when multiple consumers need access to the same database

Answering Questions While Maintaining Data Privacy

Despite the power and flexibility of secure views, nothing prevents users from running `SELECT *` queries and viewing or exporting all the data made visible to them. To mitigate this type of access, you can create *secure shared user-defined functions (UDFs)*: small pieces of SQL or JavaScript code that securely operate against raw data. Secure shared UDFs allow the owner of the data to limit the specific types of questions and analyses that can be asked against their data. It also lets a user ask specific questions of detailed data without giving that user the ability to directly view or export the raw data.

For example, imagine a retailer that wants to allow its suppliers to see which items are commonly sold together with theirs, a common practice known as *market basket analysis.* To enable this type of analysis while preventing users from seeing the raw data, a SQL statement can be wrapped in a secure UDF, with an input parameter to specify the item number that is being selected for market basket analysis.

Secure shared UDFs allow data owners to provide data consumers with functions that can run across detailed data while preventing the other parties from viewing or exporting the raw data.

Data providers can grant rights to multiple consumers to use their secure shared UDFs, even if those users are in different accounts or different organizations. Secure UDFs have become a great way to guarantee data security and privacy, giving data providers and consumers the following benefits:

- Allows consumers to run queries without viewing the source data or underlying logic
- Limits the questions consumers can ask — in other words, the types of queries they can run
- Gives providers the ability to provide valuable data to consumers without worrying about disclosing confidential information or PII

Using Secure Joins to Combine Data between Companies

When sharing data with parties outside your organization, a data provider may find itself sharing data among two or more organizations that have common customers. This situation comes up frequently between healthcare payers and providers, which each maintain their own records about patients. How do you ensure that when sharing data, each healthcare company can see data about common patients only? How do you offer bidirectional access to the data without revealing PII, as governed by HIPAA guidelines? Can you exchange data about patients without actually sharing or transferring the names or any other PII or PHI of these patients?

Secure joins allow you to jointly analyze only the records that are in both of the data sets, and to prevent each data consumer from seeing records about non-overlapping patients. For more details regarding secure joins, see "Secure Joins: How to Join Data Between Companies While Respecting PII," at `www.snowflake.com/blog/secure-joins-how-to-join-data-between-companies-while-respecting-pii`.

REMEMBER

A comprehensive data sharing service upholds security and complies with government regulations by using *secure views, secure shared user defined functions (UDFs)*, and *secure joins*.

IN THIS CHAPTER

- Evaluating opportunities and defining your organization's role
- Confirming your data sharing capabilities
- Covering all the bases
- Running a proof of concept and executing for success

Chapter 9
Six Steps to Advance Your Business with Modern Data Sharing

Now that you understand the enormous potential of modern data sharing and the challenges of traditional data sharing methods, it's time to consider the possible impact and benefits of modern data sharing. This chapter outlines six key steps to help you and your organization get started with modern data sharing to advance your business:

1. **Uncover data sharing barriers and opportunities.** The goal in this step is to gather a snapshot of data sharing requirements within your organization — both now and for the near-term future. You need a firm handle on data flows and work processes already in place to share data.

   When you have this information, focus on identifying the data that has the potential to produce the most value. Ultimately, the objective is to uncover laborious data sharing and data transferring processes robbing you of productivity and resources within your IT, data warehouse, and business analytics groups. These barriers create delays in producing business value from data. Identify the barriers and set the

stage for easier and faster execution of your data sharing business plans. Here's what to look for:

- **Data:** What types of data must the data warehouse contain? At what rate is new data created? How often will data move into the warehouse? Will you, now or in the future, want to also have easy access to a data exchange via your data platform?
- **Data flows:** Identify the data sharing already taking place within different groups of your organization and with external entities. Map out these data flows to understand who generates the data, who consumes the data, and how that is being done.
- **Work processes:** Which teams currently manage the work processes for data sharing or data transfers? Identify the tools being used, and whether they are extract, transform, and load (ETL) tools or data replication tools. This, combined with the mapping of data flows, will help you determine the current costs associated with data sharing.
- **Future time-to-market and time-to-insight objectives:** Know your data sharing needs in the present and near-term future, without regard to your current tools and available expertise.

2. **Define your role for each use case.** Identify each current and future data sharing scenario and whether your business unit or organization will be a data provider, data consumer, or both:
   - Identify the organizations, internal and external, that must be brought on board. All data sharing relationships have at least two stakeholders — the data provider and the data consumer(s).
   - Identify and engage your data providers and consumers, and confirm everyone has access to the capabilities of modern data sharing across multiple geographic regions and cloud infrastructure providers.
3. **Confirm your data platform solution can easily and cost-effectively enable modern data sharing.** Look for the following capabilities of modern data sharing:
   - *Data does not move:* Modern data sharing enables data to be shared without any data movement, ETL, or file transfer. This is the lion's share of the cumbersome work

required to share data. Eliminating data movement puts you on a better path for limitless data sharing.

- *Real-time updates:* As a data provider updates its data, every consumer should see the change as soon as it is committed. This should happen without any extra work by the data provider or consumer. Data consumers want this level of integrity, which increases the value of your data.
- *Individual secure views and secure functions:* Modern data sharing is not about simply creating access to your entire data warehouse for all your data consumers. It's about having the granular control to easily provide the necessary view of the data as required for each data consumer, while protecting sensitive data and complying with strict industry regulations, such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA). Modern data sharing enables one-to-one, one-to-many, and many-to-many data sharing relationships. The tool you choose should give you this flexibility.
- *Create Your Own Private Data Exchange:* Easily create your own private data exchange where you control who can publish, share, and access the data.

4. **Implement a proof of concept (PoC).** After investigating data sharing options, viewing demos, asking questions, and meeting with vendor teams, you should execute a PoC as soon as possible. A PoC is a process of testing a solution to determine how well it serves your needs.

   In addition, consider what else you can do above and beyond what you do today with data. If you had a modern, cloud-built data platform that enabled modern data sharing, what additional business value could the system deliver? You may want to monetize some of your data. How will you accomplish this? When setting up your PoC, list all current and future requirements and success criteria.

   For example, if your primary complaint about your current data sharing methods is that queries take too long to run, don't focus solely on that issue. Your PoC should validate assumptions about all or most high-value requirements, including ease of migrating your data to the new solution, loading new structured and semi-structured data, running queries, and handling multiple workloads.

5. **Create an outreach plan.** Outline the steps necessary to engage your data consumers. Because modern data sharing can facilitate one-to-one, one-to-many, and many-to-many data sharing relationships, you must communicate the value of your data and the benefits data consumers can expect to receive.

6. **Execute — demonstrate time-to-market and time-to-value improvements to your business stakeholders.** After completing your PoC, demonstrate the benefits of modern data sharing to your stakeholders. Estimate time savings and related cost savings for your organization as the data provider and/or the data consumer. Demonstrate the improvements in productivity your organization will gain and, if applicable, forecast the revenue potential for monetizing your data. You should be able to develop a complete picture of the ROI potential for modern data sharing and a modern cloud-built data platform. You will then be well on your way to taking data sharing to new levels of capabilities and opportunities for your organization.

Gain deeper insights into new business opportunities with modern data sharing

Organizations that share live, governed, and secure data across their business units, and within their business ecosystems, better serve their customers and more quickly reveal new market opportunities. However, traditional data sharing methods have produced only limited results thanks to cumbersome, costly, and risky processes. In addition, they don't enable you to safely monetize your data via a modern data exchange. This book reveals how you can easily and securely share data via the modern, cloud-built data platform.

Inside...

- Why data sharing emerged 100 years ago
- Why early methods limit what's possible
- How a cloud-built architecture enables live, secure, and governed data sharing
- How to govern and monetize your data
- Real-world data sharing case studies
- Steps to start sharing live data

Lawrence C. Miller has worked in information technology for more than 25 years. He has written more than 70 For Dummies books. **David Baum** is a freelance business writer specializing in science and technology.

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

ISBN: 978-1-119-66641-7
Not For Resale

for dummies®
A Wiley Brand

Also available as an e-book

WILEY END USER LICENSE AGREEMENT

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