LEARNING MADE EASY

Snowflake Special Edition

The Data Cloud

for dummies®
A Wiley Brand

Unite your siloed data

Securely access and share governed data globally

Execute your diverse analytic workloads

Brought to you by

snowflake®

David Baum

About Snowflake

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

The Data Cloud

Snowflake Special Edition

by David Baum

for dummies®
A Wiley Brand

The Data Cloud For Dummies®, Snowflake Special Edition

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

ISBN 978-1-119-81061-2 (pbk); ISBN 978-1-119-81062-9 (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. Some of the people who helped bring this book to market include the following:

Development Editor: Brian Walls

Project Manager: Martin V. Minner

Senior Managing Editor: Rev Mengle

Acquisitions Editor: Ashley Coffey

Business Development Representative: William Hull

Production Editor: Mohammed Zafar Ali

Snowflake Contributors Team: Vincent Morello, Elise Bergeron, Kent Graziano, Tim Fletcher, Clarke Patterson, Ganesh Subramanian, Christina Jimenez, Leslie Steere

Table of Contents

INTRODUCTION 1
Introducing Snowflake's Data Cloud 1
Icons Used in This Book 2
Beyond the Book 2

CHAPTER 1: **Sizing Up Challenges and Opportunities with Data** 3
Contending with an Immense Volume and Variety of Data 4
Succumbing to Silos — and More Silos 4
Resolving Problems with Fragmented Data 5
Attending to Data Governance 6
Embracing the Cloud 7
Breaking Down the Silos 8
Understanding the Impact of the Data Cloud 9
Sharing Data via a Cloud Network 10
Looking Ahead 12

CHAPTER 2: **Understanding the Value and Capabilities of the Data Cloud** 13
Understanding What You Can Do in the Data Cloud 13
Accessing your data 14
Governing your data 14
Making your data actionable 14
Identifying the Data Cloud's Unique Attributes 15
Standardizing on one data cloud 16
Supporting all data 16
Powering all workloads 16
Increasing Data Sharing's Potential 17
Introducing the Snowflake Platform 18
Cloud-built scale and performance 18
Exceptional economic value 18
Inherent ease of use 19
Multi-cloud and cross-cloud flexibility 19
Baked-in security 20
Unique collaboration options 20

CHAPTER 3: **Collaborating in the Data Cloud** 21
Understanding the Recursion Rate of Data 22
Introducing a Modern Way to Share Data 22
Looking Beyond the Four Walls of Your Organization 24
Tapping into Snowflake Data Marketplace 25
Differentiating Snowflake Data Marketplace 27

CHAPTER 4: **Deploying the Data Cloud Across Industries** 29
Yielding Greater Value in Financial Services 29
Delivering Better Healthcare Outcomes 32
Powering the Retail Supply Chain 34
Delivering Superior Media and Entertainment Services 38
Offering Better Public Sector Services 41

CHAPTER 5: **Drilling into Snowflake's Platform** 43
Starting with the Right Architecture 43
Building on the Lessons of History 44
Improving performance, lowering costs 45
Understanding why the right architecture matters 46
Establishing One Multi-Region, Multi-Cloud Service 47
Enjoying an Easy-to-Use Platform 48
Predicting and monitoring usage 48
Easing access to all types of data 49
Enforcing Strong Security and Governance 50

CHAPTER 6: **Running All Your Workloads** 53
Deploying Data Warehouses and Data Lakes 53
Enhancing Core Workloads 54
Executing Other Critical Workloads in the Data Cloud 55
Engineering data pipelines 55
Simplifying data science 56
Creating data applications 56
Sharing Data Without Limits 58

CHAPTER 7: **Six Steps to Getting Started with the Data Cloud** 59

Introduction

Innovators build and transform their businesses with data. To do so, they seek out technology that will enable them to easily and securely unify, integrate, analyze, and share that data — within their ecosystems and with other organizations keen to do the same.

Unfortunately, much of this data is born and remains in silos. Whether on premises or in the cloud, in software applications or from customer touchpoints, data becomes fragmented across departments, data centers, and public clouds. Supply chain operations, point-of-sale transactions, data security apps, and numerous other business processes create unique data stores. This data is difficult to access, govern, and mobilize in service of your business.

Introducing Snowflake's Data Cloud

The Data Cloud is a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance. Inside the Data Cloud, organizations can unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Wherever data or users live, the Data Cloud delivers a unified and seamless experience — even when data and workloads span multiple public clouds. The Data Cloud enables discovering, managing, and sharing data among business units, suppliers, other business partners, and customers. It also provides live access to data and data services from more than 125 partners in Snowflake Data Marketplace. The opportunities across industries are nearly endless:

- **Retailers** use the Data Cloud to easily centralize and share live data with consumer packaged goods (CPG) companies, supply chains, and other partners, allowing them to optimize pricing, accelerate inventory turns, and increase profits.
- **Financial services companies** digitize and automate processes, reduce fraud and risk exposure, and securely access second- and third-party data and combine it with their own data to deliver high-value customer services.
- **Healthcare providers** use the Data Cloud to securely share live health data internally and with partners to provide quality patient outcomes, reduce costs, and shorten time to market.

- **Media firms** centralize subscriber data and share it across brands, advertisers, ad platforms, and enrichment providers to increase subscriber lifetime value, ad revenue, and return on investment (ROI).
- **Manufacturers** use the Data Cloud to synchronize supply chain activities across their business ecosystems, increase plant productivity, and improve production quality while forging data-driven partner networks.
- **Public sector organizations** modernize IT and collaborate by securely sharing data across agencies, governments, and partners for new insights and improved citizen services.

Icons Used in This Book

Throughout this book, the following icons highlight tips, important points to remember, real-life examples, and more:


TIP

Advice guiding you on how to use the Data Cloud in your organization.


REMEMBER

Concepts worth remembering as you immerse yourself in understanding the Data Cloud.


CASE STUDY

Case studies about organizations using the Data Cloud to unify, share, and mobilize their data.


TECHNICAL STUFF

The jargon beneath the jargon, explained.

Beyond the Book

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

IN THIS CHAPTER

- Understanding data diversity
- Resolving problems with siloed data
- Unifying information in the Data Cloud
- Understanding the potential of a data ecosystem

Chapter 1
Sizing Up Challenges and Opportunities with Data

Nearly every business interaction generates data — whether via social media, mobile communications, Internet of Things (IoT) devices, ecommerce transactions, or many types of digital services. Multiply those interactions by a growing number of connected people, devices, and interaction points, and the scale is overwhelming — and multiplying every day. As a result, the business world has a greater need to store and manage data than ever before.

The amount of data created, captured, copied, and consumed in the world from 2010 to 2020 increased from 1.2 trillion gigabytes to 59 trillion gigabytes, according to a recent Forbes report. IDC estimates the amount of data created in the three years from 2020 to 2023 will be more than the data created during the previous 30 years. Partly in response to this huge influx of new data, the global cloud computing market size is expected to more than double to $832 billion by 2025, according to MarketsandMarkets research.

Yet an Accenture survey of 750 senior business and IT professionals revealed that just 37 percent of the respondents indicated they fully achieved the outcomes they expected from leveraging the cloud. Only 29 percent were completely confident their cloud

migrations would deliver the intended value within the expected timeframe. As this chapter shows, a new paradigm is needed to address these challenges and opportunities: the Data Cloud.

Contending with an Immense Volume and Variety of Data

If the immense quantity of data presents business challenges, so, too, does its variety. According to IDC's "2020 Global DataSphere" report, approximately 80 percent of all new data created is *semi-structured* (such as weblog, IoT, and mobile device data) or *unstructured* (audio, video, PDF, and other types of rich media content). Traditional databases, the backbone of enterprise computing, were not designed to centralize, integrate, analyze, and share this quantity or diversity of information, either on premises or in the cloud.

These shortcomings have left many business leaders wondering how they can participate in the new data economy — the global supply, demand, and consumption of data. The need to achieve rapid time to value has become more acute, even as the total volume and wide variety of useful data has grown. Unfortunately, most data management and analytics solutions rely on a small fraction of available data and look backward rather than forward. These systems remain important, but today's organizations need easy access to technology, data, and a global network to rapidly deliver predictive and prescriptive insights that drive operations forward, best predict and serve customers, and reveal new market opportunities.

One thing is certain: Having a data-centric operation is no longer optional. The opportunities of centralizing your data, and accessing data and data services in a standardized and seamless way from thousands of other organizations, is rapidly becoming today's way of achieving success on a scale not possible before.

Succumbing to Silos — and More Silos

Historically, technology limitations led IT teams to sequester data behind software and network perimeters. Each new data management endeavor created another data silo. Initially, these

repositories took the form of data warehouses, supplemented by data marts, data lakes, and — more recently — a plethora of new types of databases designed for machine learning and data science endeavors. As a result, the world's data is fragmented by data type, workload, geographic region, and clouds. Many organizations struggle to reconcile these differences, and data living in many silos is incredibly tough to combine and analyze.

Some use cases require data from multiple sources, such as when semi-structured data in a Hadoop data lake must be combined with relational data in one or more data warehouses. To analyze data from these diverse sources, IT professionals may have to build special-purpose data warehouses that require complex data-ingestion procedures powered by expensive, proprietary hardware. Data engineers create custom-coded procedures to extract subsets of the data from each source and merge it into yet another silo as a means of integrating these data sets.

These brittle interfaces must be continuously updated to accommodate new data sources and destinations. For example, data science apps typically require that data be maintained in a different form (or model) from business intelligence apps. Reconciling these differences and keeping all the data in sync requires extensive programming.

REMEMBER

Most advanced analytics applications and machine learning models leverage unique, individualized data sets because analyzing data across disparate sources is so difficult.

Resolving Problems with Fragmented Data

Imagine if everyone in your organization could access one common repository of consistent, governed data. Think about how easy it would be to combine all types of data without importing or exporting it from one system to another. What if executives, managers, and operational workers could leverage a single source of truth across your entire organization? How much more productive would your organization be if your software developers, data scientists, data engineers, and business analysts could spend more time analyzing data and less time preparing data for analysis?

These "what ifs" have hung like an albatross around the software industry's neck for nearly four decades, mainly because of how corporate information systems store and provide access to data. On premises or in the cloud, each production application maintains data in unique places and formats:

- Marketing data resides in a marketing automation system.
- Sales data is housed in a customer relationship management (CRM) solution.
- Finance data is stored in an enterprise resource planning (ERP) system.
- Inventory data is kept in a warehouse management solution.

Extracting production data for analysis creates another set of silos: data warehouses for operational reporting, data marts for departmental analytics, and data lakes for data mining and exploration. These data management systems require specialized extract, transform, and load (ETL) tools to load the data and prepare it for analysis.

Many organizations employ highly paid software engineers to set up data pipelines that orchestrate data exchanges among databases and computing platforms. They purchase special-purpose integration tools to rationalize the differences among data types and create new destination databases for reporting, analytics, and data science. The procedures for accessing, combining, and merging data are complex, expensive to maintain, and difficult to scale.

Attending to Data Governance

Even if you follow best practices for storing your data in a data warehouse or data lake, perennial challenges with security and governance are complicated by data privacy regulations that get more rigorous every year. For example, companies that do business in the European Union must adhere to exacting data lineage and traceability requirements to comply with General Data Protection Regulation (GDPR) requirements. Similar regulations have come into effect in California with the California Consumer Privacy Act (CCPA). Industry-specific mandates, such as the Health Insurance Portability and Accountability Act (HIPAA) in healthcare, the Payment Card Industry Data Security Standard (PCI DSS)

in ecommerce, and the Sarbanes-Oxley Act (SOX) in finance, further complicate security and governance.

REMEMBER

Each time you replicate data, you need to apply government and industry mandates against any new silo. The more data silos you have, the more complicated compliance becomes because you must follow the data trail in all its incarnations and instances. The best way to integrate these silos is to unify your data in a single repository that also provides seamless and performant access to external data.

TIP

Without proper governance, data silos stand in the way of corporate compliance by making it more difficult to trace the data's lineage, catalog the data, and apply security rules. Combining your data into a centralized repository simplifies these tasks.

Embracing the Cloud

To store and share significant amounts of data, many organizations place their data in public cloud repositories, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). These ubiquitous services have opened the floodgates for storing, sharing, and monetizing data. According to IDC's September 2020 "Quarterly Cloud IT Infrastructure Tracker," enterprise cloud spending, both public and private, increased 34.4 percent from 2019 to 2020, while non-cloud IT spending declined by 8 percent.

As IDC and other researchers point out, although cloud adoption trends have increased steadily in recent years, cloud computing and storage practices picked up additional momentum in 2020 in response to new work habits formed during the COVID-19 pandemic. Supporting work-from-home employees pushed data storage and data management activities into overdrive. Applications that facilitate remote communication and data sharing moved to the forefront as remote work became the norm.

Initially, these public cloud services became part of the solution to the growing number of data silos. However, having abundant capacity doesn't necessarily solve the data access problem. All that capacity can lead to more confusion and greater disparity due to various public clouds, inherent incompatibilities among clouds, and different subscription models. Thus, the cloud has engendered a new set of data silos, mainly because you can't easily

integrate data and workloads among popular cloud offerings. Although spinning up a cloud database now is easy, that database can quickly become its own silo, just as it did in the legacy on-premises world.

According to a December 2020 "Cloud Migration Forecast" report from Deloitte, 97 percent of IT managers plan to distribute workloads across two or more clouds to maximize resilience, meet regulatory and compliance requirements, and leverage best-of-breed services from various providers.

Easy availability of these public cloud resources adds to the data-diversity problem because individuals can use a credit card to spin up new cloud instances, often outside of an IT department's auspices. These departmental databases and information systems are not always properly deployed, backed up, secured, or integrated, and they may not comply with well-defined IT policies governing the proper dissemination, protection, and use of data.

Hand in hand with the rise of public cloud computing is a vast and growing collection of software-as-a-service (SaaS) applications, each with its data formats and repositories. Today, many businesses struggle to reconcile investments in this immense "application cloud," which can include hundreds or even thousands of unique apps and data stores at large firms.

REMEMBER

Public cloud services make it easier to store and access data, but they also have led to a new set of data silos due to the inherent incompatibilities among the vendor's clouds. Data assets can't be easily shared or moved among them, leading to more replication of data sets and more data management headaches.

Breaking Down the Silos

Everybody wants to combine data to form a complete, 360-degree picture of their constituents — including customers, prospects, citizens, patients, or any other pertinent group they serve. For example, marketing professionals commonly record customer interactions through all available touchpoints, drawing data from emails, phone calls, chats, website visits, social media posts, point-of-sale transactions, and customer service interactions to gain a complete picture of each customer and prospect.

In some cases, the marketing team has collected second- and third-party data.

Achieving these highly prized 360-degree views requires pulling all this data into one place. For example, to determine which ad campaigns generate the best leads, marketers might combine CRM data, call center data, and marketing campaign data to understand each customer's unique "journey." Businesses in other sectors face similar challenges, giving rise to such terms as "citizen-360" in government and "patient-360" in healthcare.

In all instances, the difficulty amassing and combining dissimilar types of data presents roadblocks to obtaining complete insights. For example, a healthcare provider might need to combine structured data in a medical records system with unstructured handwritten doctor's notes and semi-structured image data, such as X-rays and MRIs. In the legal industry, a mass of documents pertaining to lawsuits must be maintained in a form that allows broad, free-form search capabilities.

REMEMBER

Having data in a common repository simplifies segmenting customers and discerning trends, including whom customers contact, which channels they use, which offers interest them, and which content works best for each product, service, and campaign.

Whether in healthcare, law, marketing, finance, or any other domain, business professionals, data analysts, data engineers, data scientists, and application developers need to confidently access a single source of truth so their reporting, analytics, and data science endeavors yield consistent outcomes. The Data Cloud makes this experience possible. You can leverage all your data simultaneously, even when it resides in multiple clouds, without having to import or export data from one system to another. This architecture is a sharp contrast from how data applications were created in the past: optimized for a specific workload and a single type of data.

Understanding the Impact of the Data Cloud

The Data Cloud is a global data network that spans multiple public clouds. No matter which cloud services you use, the Data Cloud allows your entire organization to access, analyze, and share that

data in a secure, seamless, and governed manner, regardless of location.

Inside the Data Cloud, you can unite siloed data, easily share governed data, and execute various data-driven workloads. Because all data is consolidated in one place, you can eliminate the silos and the associated administrative procedures that go along with maintaining your data.

REMEMBER

As part of the Data Cloud, your data gains value through association with other data in the Data Cloud ecosystem. For example, the Data Cloud facilitates the process of easily sharing governed data with partners while leveraging a cohesive set of data services.

The Data Cloud meets the needs of businesses in the digital economy by resolving some of the software industry's most pressing tasks:

- Supporting a wide range of data-driven workloads
- Connecting a diverse set of data sources with a broad set of data consumers
- Taking advantage of cost-effective public cloud services

Sharing Data via a Cloud Network

In the digital economy, enterprises everywhere need to share data. For example, retailers commonly share sales data with vendors to manage inventory and supply chains, and telecommunications providers share subscriber engagement information with application developers to produce better customer experiences.

The Data Cloud facilitates these interactions by allowing the exchange of data and data services among data consumers, data providers, and data services providers in a seamless, transparent manner. Some companies, for example, offer data services to help you mobilize your data and derive maximum value from it. These *data services providers* can help you activate, govern, and fully understand your data's potential, driving business value. *Data integration providers* make getting your data into the Data Cloud easier. *Data science and analytics partners* help you derive value from your data, and *systems integrators* can help you securely share and monetize your data.

Within the Data Cloud, data providers can establish *one-to-one* sharing relationships, such as when a retailer shares data with a consumer packaged goods (CPG) company, or *one-to-many* relationships, such as when a weather bureau shares weather forecast data with subscribers. As described in Chapter 3, the Data Cloud also provides access to hundreds of commercial data listings via Snowflake Data Marketplace. The marketplace is a component of the Data Cloud and offers a rich supply of data sets and data services.

MITIGATING A GLOBAL PANDEMIC WITH SHARED COVID-19 DATA

CASE STUDY

To streamline access to COVID-19 health and patient data, a coalition of leading healthcare and technology companies joined forces to establish one of the largest healthcare data repositories ever created. Dubbed the COVID-19 Research Database, this cross-industry consortium, housed in the Data Cloud, is shared with thousands of academic, scientific, and medical researchers. The repository contains billions of anonymized records, including patient claims data, electronic health records (EHRs), and U.S. mortality data.

Having current data on medical claims, pharmacy claims, laboratory data, demographic data, and many other data points enables researchers and public health officials to better understand and combat the deadly coronavirus. Authorized users at dozens of medical schools and hundreds of public health institutions leverage this immense cloud data set to study how state closure policies affect non-COVID-19 healthcare services, how to balance economic and health concerns when considering reopening businesses, the impacts of COVID-19 on preventive care, the costs of testing in various regions, and many other studies. Researchers can also determine the demographic factors and preexisting conditions of people requiring a ventilator and examine the impact of quarantines and social-distancing measures.

Secure data-sharing technology permits these researchers to collaborate without having to copy or move the data. Authorized users can directly share data both within their organizations and between organizations. They can also take advantage of a growing data marketplace to access live, ready-to-query data from data providers and service providers.

TIP

The Data Cloud is based on a new global data-sharing architecture, making all users part of this federated landscape. Authorized users can access any data made available to them in the Data Cloud without having to copy or move the data to another location. The data is accessed in place, subject to a universal data governance model.

Looking Ahead

The increased generation and consumption of all kinds of data has fueled the continued growth of the global datasphere. Currently, most organizations can access only a small portion of the data available to them, both inside and outside their businesses. A new paradigm is needed, based on purpose-built technologies for storing, processing, and managing data. The Data Cloud brings the right technology to bear, enabling thousands of organizations to mobilize data in the service of their businesses. Chapter 2 explains the Data Cloud's construction and how the Data Cloud can benefit your organization.

REMEMBER

The Data Cloud is a global network where thousands of organizations can easily and securely access their own data and the growing number of data sets and data services with near-unlimited scale and concurrency. Inside the Data Cloud, you can unite siloed data, easily discover and securely share governed data, and execute a diverse set of analytic workloads. The Data Cloud enables a seamless experience across multiple public clouds.

IN THIS CHAPTER

- Revealing what you can do in the Data Cloud
- Identifying the unique attributes of the Data Cloud
- Reviewing the primary Data Cloud workloads
- Exploring modern data sharing
- Introducing the underlying technology platform

Chapter 2
Understanding the Value and Capabilities of the Data Cloud

The Data Cloud empowers business leaders, data engineers, data scientists, business analysts, and operational workers to mobilize data in a multitude of ways to best serve their business and their customers. This chapter introduces the Data Cloud's primary capabilities and highlights the unique attributes of this global data network.

Understanding What You Can Do in the Data Cloud

Snowflake created the Data Cloud so that any organization could break down its data silos (both on premises and in public clouds), execute diverse analytic workloads, and easily and securely share

governed data across subsidiaries, ecosystems, and geographies. The major benefits fall into three categories: *access, governance,* and *action.*

Accessing your data

By unifying your structured, semi-structured, and certain types of unstructured data (functionality Snowflake expects to make available in a future release), the Data Cloud eliminates the need for separate data platforms, data warehouses, data lakes, and data marts. The Data Cloud also simplifies securely sharing data throughout your organization and with your external ecosystem of partners, suppliers, customers, and other stakeholders. You can further enrich your analytics by accessing data sets, data services, and data applicatons via Snowflake Data Marketplace — which is part of the Data Cloud — without needing to copy or move data to share it within your ecosystem.

Governing your data

The Data Cloud allows you to eliminate data silos, so you can govern all the data that matters to your business. Create and enforce flexible data governance and security policies and apply them across various data workloads. Snowflake's support for important data privacy regulations minimizes the risk of a security or compliance breach, streamlining data governance efforts without hindering innovation. You can extend data access and data security policies throughout your ecosystem of partners and to data accessed from third-party providers. Additionally, you can utilize third-party data governance services via Snowflake Data Marketplace.

Making your data actionable

The Data Cloud has been architected to unleash near-unlimited power, scale, and concurrency for all your data workloads — including data engineering, data lakes, data warehousing, data science, data applications, and data sharing. In the Data Cloud, you can analyze your data, develop new data products, run machine learning models to make informed business decisions, drive innovation with fresh data-driven insights, and monetize data for new revenue streams (see Figure 2-1).

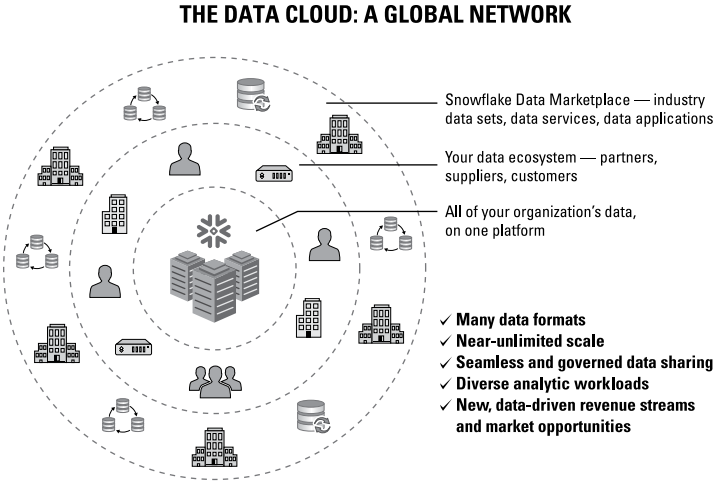

FIGURE 2-1: The Data Cloud delivers breakthroughs in technology, seamless access to data without having to move or copy it, and a multitude of business benefits.

Identifying the Data Cloud's Unique Attributes

The Data Cloud consists of two parts — platform and data. Snowflake's platform is the engine that powers the Data Cloud. The data comes from customers and other data providers that unify, access, and ultimately decide to share data via the platform.

Organizations of all sizes and across many industries centralize their data in the Data Cloud to efficiently execute diverse analytic workloads. The Data Cloud lets them run multiple workloads simultaneously and against the same data sets, such as when data engineering workloads must run concurrently with data warehouse and data science workloads.

Thanks to Snowflake's unique architecture, which separates storage and compute, the number of supported workloads is practically limitless.

Furthermore, the Data Cloud supports sharing data and services in a controlled manner without the need to copy files or create data silos. For example, you can use the Data Cloud to unify data across your supply chain, simplify interactions with business

partners, and envision entirely new revenue streams by monetizing your governed data and sharing it with other organizations in your industry.

Standardizing on one data cloud

In some ways, the Data Cloud is like the World Wide Web. The web was created to provide an "information superhighway" that allows people to share information and link documents. Likewise, anyone can plug into the Data Cloud to access, govern, and activate their data and connect to a larger data ecosystem, including accessing third-party data and data services in Snowflake Data Marketplace.

The Data Cloud is composed of multiple Snowflake regions spanning the three major public clouds on four continents, and enabling easy collaboration of data and workloads among regions. All Data Cloud workloads run consistently, and all users have a single cohesive experience, regardless of which region or cloud provider they use.

Supporting all data

In 2014, Snowflake transformed the world of data management and analytics by introducing a single platform with first-class support for both structured and semi-structured data. Snowflake expects to add support for certain types of unstructured data in a future release, including audio, video, text, imaging data, PDFs, and many other types of files. The Data Cloud can accommodate both business-generated and machine-generated data in their native formats. These broad data storage and data access capabilities enable the Data Cloud to fulfill its lofty vision: a world with unlimited access to governed data, so every organization can tackle the challenges and opportunities of today and reveal the possibilities of tomorrow.

Powering all workloads

The Data Cloud makes possible mobilizing all data in service of your business without the cost, complexity, and overhead of managing multiple database management systems, data engineering tools, and cloud vendors. However, the true value of the Data Cloud isn't just the ease with which it can store, manage, and secure your data. It also streamlines *access* to that data.

With data solutions, putting data "in" is the easy part. Getting insights out of data can be complicated, such as when you attempt to run analytics and data science workloads simultaneously. The Data Cloud, driven by the powerful Snowflake platform, which manages data from source to consumption, excels in supporting a wide range of workloads. Figure 2-2 introduces the primary Data Cloud workloads, explained in greater detail in Chapter 6.

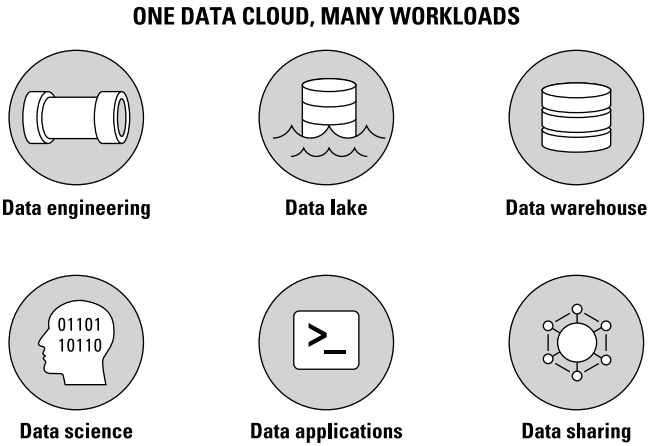

FIGURE 2-2: The Data Cloud seamlessly supports the world's most popular data management and analytic workloads with no resource contention.

Increasing Data Sharing's Potential

As introduced in Chapter 1, the Data Cloud enables a global network where thousands of organizations can quickly mobilize and share their data. *Data providers* can easily share governed data with *data consumers* without having to move or copy data. Consumers can discover, access, and join shared data with their own data to create new insights.

REMEMBER

Sharing isn't limited to data. Data service providers can also share unique solutions and business logic via their data services. For example, a data science service provider might offer pretrained machine learning models that customers can run against their data without providing access to the entire data set.

SUPERIOR DATA SHARING

The Data Cloud streamlines the process of sharing and managing data across a broad, global ecosystem in the following ways:

- Data and data services can be shared without the need to copy data.
- Authorized users can obtain live, read-only access to a current data set that is always up to date.
- Data consumers can acquire access to shared data and combine it with their data to generate new insights.
- Snowflake Data Marketplace provides rich data sets and data services from third-party providers.

Introducing the Snowflake Platform

Since its inception, Snowflake has reimagined many aspects of data management, data operations, and data analytics to leverage the potential of cloud computing. Snowflake's platform powers the Data Cloud. This section highlights the platform's primary attributes. Chapter 5 showcases the driving technologies of the Snowflake platform.

Cloud-built scale and performance

The Snowflake platform has been architected for cloud-scale computing in terms of data volume, computational performance, and concurrent workload execution. This unique cloud-built architecture enables organizations in the Data Cloud to spin up new workloads without scale or concurrency limitations and dynamically provision these workloads with as little or as much compute and storage capacity as they need. There are no hidden capacity limits, and there is no fear of resource contention with other data workloads.

Exceptional economic value

Snowflake's cost-effective utility model ensures organizations in the Data Cloud pay only for the capacity they use, in one-second increments. Capacity is provisioned automatically, based on user-defined thresholds, allowing customers to focus on consumption

while the platform manages the resources they need. Because Snowflake contracts with the public cloud vendors for immense capacity at a considerable scale, customers receive these cloud services at exceptional economic value via Snowflake.

Inherent ease of use

Snowflake's platform is dramatically more straightforward to use than previous generations of cloud data platforms and database-as-a-service offerings. Because the platform doesn't require expert knowledge to run, your database administrators can move beyond manual tuning and configuration chores and focus their expertise on creating and refining data models, extracting new insights, and deriving business value from a wide range of data.

Multi-cloud and cross-cloud flexibility

The Data Cloud spans the world's most popular cloud services: Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). You no longer have to standardize on one of them. You can leverage multiple regions and clouds according to your business needs and seamlessly share data from one region or cloud to another. For example, a large multinational organization with a multi-cloud strategy can use the Data Cloud to operate seamlessly across multiple cloud provider regions to serve its worldwide customer base and enhance its disaster-recovery strategy. Similarly, a data provider can leverage Snowflake's global infrastructure to share data with other organizations in many geographies and across multiple clouds.

REMEMBER

Because Snowflake completely abstracts the underlying public cloud platform, users and administrators don't need any cloud-specific expertise. They never have to work with these third-party cloud services directly, nor do they receive a separate bill for public cloud usage.

REMEMBER

Snowflake designed the platform to be cloud-agnostic: It runs interchangeably on AWS, Microsoft Azure, and GCP. One unified code base spans all three public cloud services, which means customers can seamlessly move data and workloads among them yet interact with one cohesive platform interface. Being cloud-agnostic is a decisive advantage when serving a worldwide user base or formulating a global disaster recovery strategy.

Baked-in security

A strong data governance and compliance framework is critical to extracting the analytical value of data. The Data Cloud allows you to store all your data while providing fast, governed, and secure access to that data. Industry-leading cybersecurity practices extend across all data and all workloads. Multilevel security includes physical security to protect all facilities where data is stored, data encryption and associated key management services, role-based access controls, object-level permissions, and robust database security. Furthermore, the Data Cloud adheres to regulatory and data privacy policies to ensure the correct handling of sensitive data.

Although new privacy requirements are still emerging in many areas, the Snowflake platform is continually evolving to help customers comply with national and industry-specific regulations and to avoid costly violations of those regulations.

Unique collaboration options

Snowflake's unique architecture provides the near-unlimited scale, concurrency, and performance for organizations to collaborate with data across their enterprises and ecosystems. Whether creating access to data sets or data services, between organizations or with traditional data providers or SaaS vendors, the Data Cloud raises the bar of what is possible. You can unify data across your supply chain to accelerate time to market. You can also establish entirely new revenue streams and collaborate with business partners in new and productive ways. Snowflake customers can gain rich insights, build great products, and deliver extraordinary services by reaching beyond their data.

IN THIS CHAPTER

- Replacing antiquated data sharing methods
- Sharing data within and beyond your organization
- Tapping into Snowflake Data Marketplace
- Getting in step with the data economy

Chapter 3
Collaborating in the Data Cloud

One of the primary obstacles to building a robust data economy is that data is difficult to share, hindering collaboration. Data owners across organizations often have data siloed in different data platforms, database models, and clouds, complicating the exchange and dissemination of their data. Raw data often exists in disparate formats (structured, semi-structured, and unstructured), inhibiting easy integration and access. Furthermore, data governance practices and data usage regulations vary from location to location and company to company, complicating compliance with the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and various industry data privacy regulations.

The Data Cloud eliminates these obstacles by leveraging two unique Snowflake capabilities:

- Snowflake Secure Data Sharing enables customers to securely share data in real time, both within and beyond their organizations, but without having to move the data.
- Snowflake Data Marketplace allows data consumers to similarly access and query third-party data sets and data services to reveal previously unimagined business insights.

Understanding the Recursion Rate of Data

A significant factor impacting today's businesses is the ratio of *unique data* (new data created and captured) to *replicated data* (existing data copied and consumed). According to IDC's 2020 Global DataSphere report, this ratio stands at 1:9 and is edging toward 1:10. In other words, data is ten times more often copied and consumed than created. This phenomenon is also known as the *recursion rate* of data — the rate at which the same data is processed again and again.

These statistics are particularly revealing when you consider the costly, cumbersome, and risky ways most data is shared. The most popular methods include setting up programmatic links among data sources, such as establishing a File Transfer Protocol (FTP) site or creating application programming interfaces (APIs) — both of which require writing and maintaining code. In other instances, users simply email copies of data back and forth or rely on file-sharing utilities to share content. Neither of these simplistic methods upholds data security and regulatory policies.

REMEMBER

Copied and shared data becomes stale almost immediately and must continually be refreshed — a primary contributor to the high recursion rate of data. Typically, you don't just move the data once. You have to keep the data set current, which requires either periodic bulk uploads to replace the previous data set or discrete transactions to record the changes.

Introducing a Modern Way to Share Data

The Data Cloud employs Snowflake's Secure Data Sharing technology, which replaces these traditional data sharing methods. Instead of archaic, maintenance-intensive, manual software procedures, Secure Data Sharing allows two or more organizations to share and consume live data safely in the Data Cloud simply by setting access permissions. Any data provider can easily grant access to the data it wants to share with its intended data consumers without having to copy, move, or manually update that data.

All database objects are centrally maintained and updated in the Data Cloud. End-to-end security, governance, and metadata management services are systematically applied, even when queries and transactions span multiple public clouds. You don't have to link applications, set up file-sharing procedures, or frequently upload new data to keep it current. It's also less expensive because data is shared rather than copied, so no additional storage is required.

Snowflake customers can acquire and share data by merely granting governed access to specific data sets in their Snowflake accounts — even when the recipient accounts are in different clouds or geographic regions of the same cloud. Providers designate data for sharing and grant permissions to it. Recipients can then query that data in place without needing physical custody of that data.

Thanks to the extensive governance and security within the Data Cloud, data providers can safely share data with partners while maintaining data control within their Snowflake accounts. They can create one-to-one, one-to-many, and many-to-many data sharing relationships. Data providers retain security and control but can opt to share slices of their data with other members of the Data Cloud. This approach's advantages are clear: data is always up to date, with near-zero latency, and any updates to shared data sets are instantly available to consumers (see Figure 3-1).

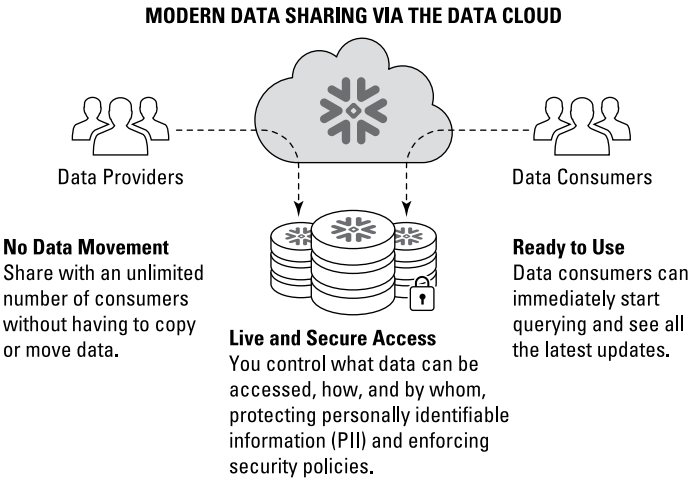

FIGURE 3-1: The Data Cloud streamlines data sharing between data providers and data consumers, even across multiple regions and clouds.

BETTER DATA SHARING

Why should you use the Data Cloud to share data?

- **Simplify the interface:** Eliminate email, multiple spreadsheets, shared network drives, APIs, and other traditional methods for copying and moving data.
- **Reduce latency**: Do away with stale data, which often yields outdated or incomplete insights.
- **Extend your horizons:** Discover, access, and share data throughout your business and far beyond, with potentially thousands of organizations.

REMEMBER

Snowflake Secure Data Sharing technology allows multiple organizations to easily share and consume live, governed data. No additional maintenance or file transfers are required, unlike with traditional data-sharing methods. You can share data and receive shared data with many different entities within your organization, across your business ecosystem, and with your customers and partners.

Looking Beyond the Four Walls of Your Organization

For more and more of today's businesses, collaboration extends beyond the boundaries of a single organization. Within the Data Cloud, you can collaborate freely with other members of your business ecosystem — for example, with your customers, your suppliers, and your partners. The Data Cloud's sharing potential creates a *network effect:* the more organizations that join the Data Cloud, the more data can be exchanged with other Snowflake customers, partners, and data providers, enhancing the value of the Data Cloud for all users.

For example, a software company that offers travel management services can use the Data Cloud to share credit card purchases, expense data, and budget information to help clients analyze corporate travel expenditures. A medical research firm can use the

Data Cloud to share COVID-19 data with local, state, and federal agencies. Financial services providers can use the Data Cloud to publish data about market trends and exchange rates, helping stock brokers and investment companies monitor market movements. And just about any company that wants to connect with suppliers, partners, and remote business units can use the Data Cloud to collaborate (see Figure 3-2).

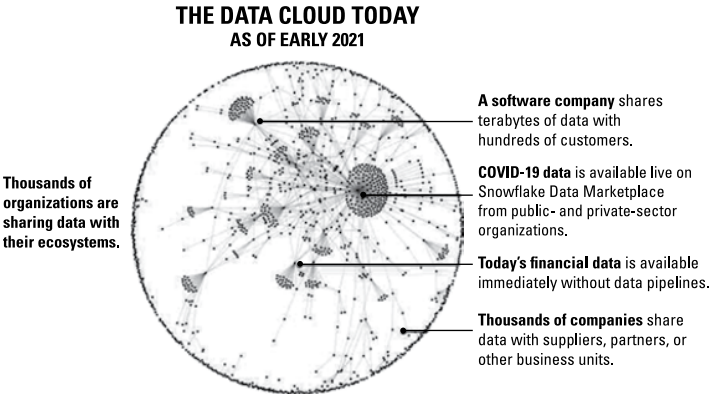

FIGURE 3-2: The Data Cloud streamlines the exchange of data among all types of companies and within all types of workgroups.

Tapping into Snowflake Data Marketplace

To further these aims and encourage the expansion of new data opportunities, the Data Cloud includes Snowflake Data Marketplace, where hundreds of data sets and data services are available from more than 125 commercial providers, including FactSet, Starschema, Weather Source, S&P Global, Wunderman Thompson, Epsilon, Tapad, AccuWeather, Airlines Reporting Corporation (ARC), and many others.

Additionally, any Snowflake customer can list its governed data sets and monetize those data sets by sharing them with other Snowflake customers via Snowflake Data Marketplace. Data consumers can discover productive ways to utilize and merge third-party data with their own data, and data providers and data services providers can instantly reach new audiences and grow their revenue.

Unlike with other data marketplaces, Snowflake's unique Data Cloud architecture empowers customers to utilize live, ready-to-query, third-party data sets — and data services — without needing to copy files or move data (see Figure 3-3).

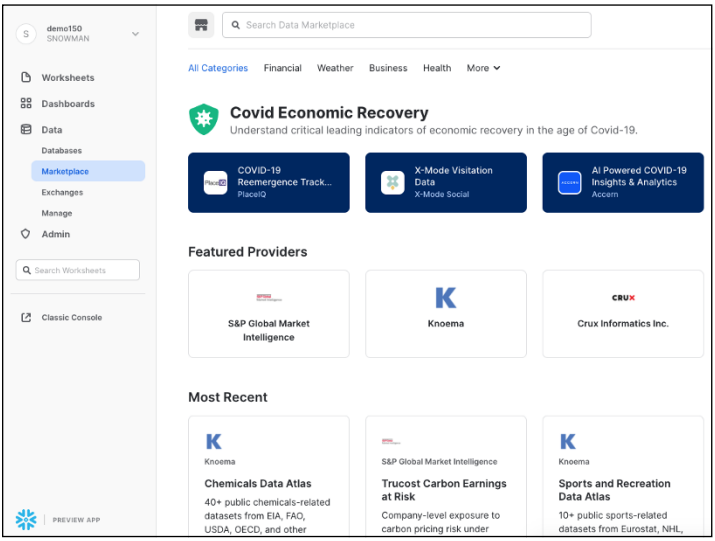

FIGURE 3-3: Snowflake Data Marketplace includes data and services from more than 125 data partners across more than a dozen categories.

Snowflake includes capabilities for discovering, managing, monitoring, and sharing this data in productive ways. Authorized data consumers can obtain secure access to the data they are permitted to see. For example:

- **ARC** offers access to the world's most comprehensive source of global air travel intelligence, spanning flights from more than 4,600 airports in 234 countries and territories.
- **S&P Global** provides prescored information to help businesses identify the credit worthiness of potential business partners, allowing them to gain a more complete picture of their credit risk exposure.
- **Heap** uses Snowflake Data Marketplace to share user, session, and event-level data from websites and applications.

Snowflake customers can also take advantage of useful *data services* in the marketplace, such as detecting fraudulent transactions, assessing risk scores, identifying cybersecurity threats in security logs, and enriching customer data for marketing purposes. For example:

- **LiveRamp** helps marketers create more-complete customer profiles by pulling in data points from more than 100 offline data sources.
- **Quantifind** performs financial risk assessments to compute risk levels for people and organizations.
- **Hunters** is an autonomous cybersecurity solution that uses artificial intelligence to detect attacks that bypass existing security solutions.
- **Seismic** shares marketing and sales-enablement data as a feature of its analytic reporting data set, along with a data model for governing users, content, and actions.


TIP

Data Cloud customers can use Snowflake Data Marketplace to access useful data sets and data services, sometimes establishing new revenue streams by monetizing their data. Demographic, marketing, weather, and COVID-19 data sets are all examples of data globally accessible on Snowflake Data Marketplace.

Differentiating Snowflake Data Marketplace

Other data marketplaces require data consumers to build a data pipeline, ingest the data as a distinct file, and store it on their systems — either on premises or in a public cloud. Data consumers must go through this process whenever the data provider updates the data set. Complex data engineering may be required to create custom interfaces, either with programmatic links or via manual file transfers.

Snowflake's Secure Data Sharing technology allows users of Snowflake Data Marketplace — and data services providers — to query live data sets with a couple of clicks. Instead of having to

ingest files, the information is automatically accessible from each consumer's Snowflake account. No additional storage, maintenance, or file transfers are required, and consumers can instantly see changes and updates as data providers update their data sets.

DISTRIBUTING REAL-TIME DATA VIA SNOWFLAKE DATA MARKETPLACE

CASE STUDY

FactSet provides financial data, market data, and analytics to tens of thousands of investment professionals. These clients integrate the data with their applications, web portals, and statistical packages to make crucial decisions.

Previously, FactSet used API calls, FTP, and Secure FTP methods to transfer data, which was time-consuming and required significant compute resources. In some instances, ingesting very large data sets took weeks. FactSet sought a more efficient way to distribute its content to clients.

Now, FactSet deploys its data sets once in the Data Cloud, making them instantly available to a nearly unlimited number of clients via Snowflake Data Marketplace. Investment professionals obtain fast, near real-time access to content that's managed by FactSet, including access to more than 20 proprietary data feeds and dozens of third-party data feeds, without incurring storage costs. They can query structured and semi-structured data using SQL, join it to their data, and analyze the merged data sets with their own tools — without utilizing cumbersome extract, transform, and load (ETL) procedures.

File-sharing processes that used to take hours, or even days if the team had to provision cloud resources, are now nearly instantaneous. Some FactSet clients use Snowflake Data Marketplace to map disjointed data sets together, such as joining FactSet queries with data from other content providers. FactSet plans to add more data sets to Snowflake Data Marketplace, such as economic and research data from the U.S. Department of the Treasury's Treasury International Capital (TIC) System and banking benchmark data from Cornerstone Advisors, a FactSet partner.

IN THIS CHAPTER

- Yielding greater value in financial services
- Generating positive outcomes in healthcare and life sciences
- Powering the retail supply chain with fresh insights
- Delivering superior media and entertainment experiences
- Offering exceptional government services to citizens

Chapter 4
Deploying the Data Cloud Across Industries

The Data Cloud allows organizations across many different industries to maximize their data's value and delivers timely insights to improve efficiencies, boost revenues, reduce costs, and deliver better experiences to customers, patients, and citizens. This chapter describes the Data Cloud's potential in several industries.

Yielding Greater Value in Financial Services

Today's financial services companies are experiencing massive change, driven, in part, by a new breed of nimble financial technology (FinTech) companies adept at using data, analytics, and artificial intelligence to deliver digital financial services for consumers and businesses. PwC revealed in a report titled "Financial Services Technology 2020 and Beyond: Embracing Disruption"

that more than three-fourths of today's financial services companies invest in technology to put their customers at the center of routine interactions and to enable hyper-personalized experiences.

The Data Cloud simplifies the process of capturing, storing, analyzing, and sharing data, as well as developing new data applications that power high-value services. The Data Cloud allows financial services firms to centralize data, mine it for insights, and merge it with second-party data from vendors or financial institution partners. They also mobilize third-party data from such sources as Snowflake Data Marketplace to achieve broader customer and investor views. For example, banks gather customer data from multiple touchpoints within their institutional, commercial, and retail segments. Having a shared source of governed data stored in the Data Cloud simplifies identifying best practices for developing new products and services that meet their customers' unique needs. Another example is investment companies and asset management firms using fresh data to power differentiated investment strategies and enable their advisory teams to make quick data-driven decisions.

Many financial services firms utilize third-party data sets to better understand the economic and societal trends that affect their businesses, along with data services that assist with critical regulatory and governance tasks, such as fraud detection, anti-money laundering, risk management, and credit assessments.

REMEMBER

Analyzing information is much simpler when firms can store all their data in a centralized, globally available, consistently governed platform.

Insurance companies use the Data Cloud to access third-party data sets from partner companies and publicly available data sets on Snowflake Data Marketplace. A property and casualty (P&C) insurer that wants to minimize claims leakage, for example, can use the Quantifind data services to investigate potentially fraudulent auto repair claims. Available in Snowflake Data Marketplace, this service uses external data sources, such as sanctions and blacklists, in conjunction with predictive risk-typology models that indicate risk levels. This data helps claims adjusters establish fraud risk scores for each individual that submits a claim.

The Data Cloud gives insurance adjusters immediate access to this third-party data, facilitating collaboration across the insurer's business ecosystem. For example, insurance executives might want to know how financial metrics such as *total annual claims paid* compare with similar metrics from other P&C insurance companies in their region. Here again, Snowflake Data Marketplace contains the insights they need: S&P Global, a Snowflake partner, supplies comprehensive statutory financial data from the National Association of Insurance Commissioners.

Snowflake's Secure Data Sharing technology makes these multiple sources of data instantly available, without the need to transfer data or set up custom APIs, making it easier to mobilize data for business intelligence or data science endeavors. For example, the P&C insurer might want to combine the Quantifind data and the S&P Global data into an executive dashboard, along with information from the internal customer relationship management (CRM) system about each auto body repair shop.

HIGH-VALUE CUSTOMER AND INVESTOR EXPERIENCES

Financial services companies use the Data Cloud to centralize their data for deeper insights into their institution, securely access second- and third-party data for broader customer and investor views, reduce fraud and risk exposure, and digitize and automate processes to focus on delivering high-value services.

- **Financial institutions:** Unify and securely share your customer data, removing silos across all departments to create an enterprisewide 360-degree view of each customer.
- **Financial services partners:** Combine customer-behavior and financial data with additional sources from your FinTech and other partners to create broader customer profiles.
- **Financial data providers:** Increase your offerings by sharing relevant customer and investment data so institutions can create deeper customer relationships.
- **The customer experience:** Financial services companies can deliver timely and personalized customer, investor, and policyholder experiences and relevant interactions at every touchpoint.

Having a universal data repository simplifies data sharing across business units, subsidiaries, and with data partners while deploying controls that enable compliance with financial regulations, such as the Sarbanes-Oxley Act and Dodd-Frank Wall Street Reform and Consumer Protection Act, as well as many other international regulations.

Delivering Better Healthcare Outcomes

Healthcare and life sciences are distinct industries with similar attributes. Healthcare focuses on delivering health services for patients, whereas life sciences focus on developing and commercializing medicines and therapies. Due to the critical nature of this work, these are highly regulated industries. In the U.S., the most notable health information security requirement affecting these industries is the Health Insurance Portability and Accountability Act (HIPAA), which governs the use of protected health information (PHI) and personally identifiable information (PII). In the European Union, health data must be secured per General Data Protection Regulation (GDPR) requirements.

The Data Cloud helps ensure the security of sensitive health data while providing seamless, near real-time, governed access to other health systems, payer networks, research partners, health analytics providers, and more. Snowflake is HIPAA-compliant and HITRUST-certified, meaning that all patient data is integrated quickly, securely, and in support of these rigorous data governance standards.

The rapid rise of patient health data generated from electronic health record (EHR) systems and connected health devices paves the way for new approaches to care delivery, clinical diagnostics, medical innovation, and regulatory decision-making. Immense data volumes flow between healthcare providers, payers, life sciences companies, and medical research institutions. The Data Cloud powers secure data management capabilities while powering both routine administrative reporting and advanced analytics and data science initiatives.

By leveraging Snowflake Secure Data Sharing capabilities, healthcare organizations can achieve true interoperability by providing secure, managed access to live data while safeguarding protected health information and maintaining compliance with data security and privacy regulations. Additionally, Snowflake Data Marketplace includes several healthcare-specific data sets and data services, from physician credentialing and physician-based market share data to public health data.

REMEMBER

Complying with industry regulations governing PHI and PII security presents ongoing challenges for many healthcare and life sciences companies. However, tremendous opportunities exist for organizations that can reimagine how to store, manage, and govern clinical and administrative data.

A SINGLE SOURCE OF TRUTH FOR EVERY PATIENT JOURNEY

Healthcare and life sciences organizations use the Data Cloud to securely share live health data internally and with their ecosystem of partners to provide quality patient outcomes, drive growth, shorten time to market, and reduce costs while deploying controls that enable compliance with HIPAA and other data-governance regulations.

- **Healthcare providers:** Create a single source of patient data, including electronic health records along with clinical, administrative, operational, and Internet of Medical Things (IoMT) data.
- **Payers:** Share your claims, billing, and other data with providers, and share formulary and diagnostics data with pharmacy and lab partners.
- **Life sciences companies:** Share your clinical, sales, marketing, and other data across business units and externally with research collaborators.
- **Research institutions:** Aggregate and share health-outcomes data collected from providers, life sciences companies, and other partners.

CASE STUDY

SHARED DATA FOR PATIENT CARE

Komodo Health believes that smarter, more innovative use of data and analytics is essential for reducing disease burden. To that end, the company has applied artificial intelligence and other advanced data science techniques to its Healthcare Map, which tracks the unique journeys of more than 320 million patients, augmented by analytic algorithms and clinical expertise.

Previously, Komodo had difficulty linking data across thousands of siloed sources. Komodo's IT team struggled to build access controls, maintain quality control processes, and scale the ingestion of terabytes of daily data. Now, by using the Data Cloud, Komodo Health can leverage more than 150 payer data sets in conjunction with more than 65 billion clinical and pharmacy encounters.

The Data Cloud enables Komodo customers to use frameworks, such as R, and programming languages, such as Python, to directly query these data assets. Komodo also offers data services that allow customers to run custom analytics on top of their data and develop data assets they can leverage to glean their own insights.

Powering the Retail Supply Chain

In the wake of the COVID-19 pandemic, retailers and consumer packaged goods (CPG) companies are under immense pressure to meet consumer needs via new delivery channels while ensuring speed, convenience, and quality. Companies across the industry are reevaluating how they use data to forecast customer demand, manage supply chains, and conduct business efficiently. Retailers can no longer rely on intuition and guesswork to decide which products to stock, which promotions to run, which pricing models to employ, and a host of other operational issues.

A mutually beneficial relationship exists between retailers and their CPG partners, rooted in the exchange of data. Each entity has siloed information that is valuable to the other. For example, retailers have granular point-of-sale (POS) data that CPG firms can't readily access. CPG companies want to analyze that data and combine it with other sources to make category and brand management recommendations. They need to know precisely

what customers are buying to produce the right goods in the right quantities.

To satisfy these needs, both retailers and CPG companies are forging connected supply chains, where intelligent demand forecasts drive automated replenishment strategies to ensure the items customers want are always available — in the right place, at the right times, and in the right quantities.

A complete source of live, governed data is critical to gleaning these consumer-level and product-level insights. Unfortunately, transactional data is often siloed within POS applications, which complicates integrating the data with other data sources, such as marketing applications, supply chain management applications, inventory control systems, and CRM systems.

TIP

Whether trying to learn more about customers, manage inventory levels more efficiently, or forecast demand for popular items, POS data is the starting point for analysis.

The Data Cloud enables retailers and CPG companies to easily share POS and other data while building mutually beneficial relationships and forging extended partner networks. Shared data is the key to synchronizing activities such as moving goods from suppliers through distribution centers and out to retail stores and ecommerce depots.

Retailers and CPG firms also depend on the Data Cloud to conduct their analyses, often using third-party data available in Snowflake Data Marketplace. For example, Numerator offers consumer purchase information from more than 1 million U.S. households, informing go-to-market strategies. Data provider Adverity offers advertising data from Google, Facebook, and Snapchat to help companies better understand customer preferences and purchasing trends. In some cases, these data services present opportunities for retailers and CPG companies to create new revenue streams by monetizing their data (see Figure 4-1).

Of course, strict consumer privacy laws, such as GDPR and the California Consumer Privacy Act (CCPA), and cybersecurity threats drive the need for relentless data governance and advanced security practices. All retailers desire to get to know their customers and anticipate their needs, but failing to secure PII data can result in significant fines for companies that are noncompliant with prevailing data privacy laws.

SECURELY SHARED DATA FOR THE RETAIL SUPPLY CHAIN

The Data Cloud enables retailers to centralize their data and securely share live data with supply chain partners while seeking to optimize pricing and inventory strategies, increase margins, and ensure consumer privacy.

- **Retailers:** Centralize siloed data across your supply chain, inventory, point of sale, CRM, customer loyalty, and marketing analytics systems.
- **CPG organizations:** Access transaction data from your retailers, combine it with other data, analyze it, and provide back brand-management insights.
- **Suppliers:** Access omnichannel sales data from retailers to predict product quantities and minimize out-of-inventory or overstock events.
- **The customer experience:** Deliver seamless and personalized experiences to customers by enabling access to data across retail and CPG businesses.

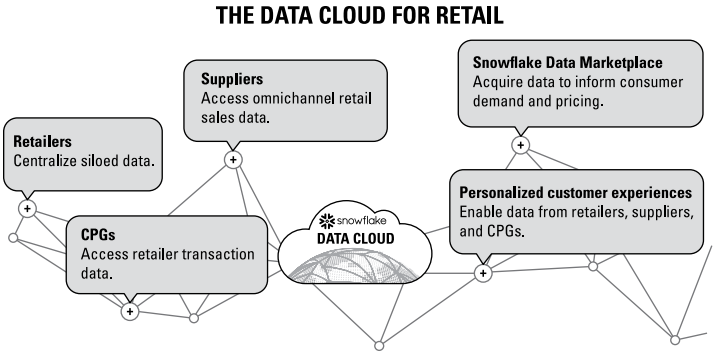

FIGURE 4-1: The Data Cloud enables all members of the supply chain to share data efficiently and invest in infrastructure that unlocks value from data.

NATIONAL RETAILER EXPANDS ITS DATA HORIZONS

CASE STUDY

At Office Depot, data comes in from many sources: websites, mobile apps, retail outlets, warehouses, supply chains, Internet of Things (IoT) devices on vehicles, and more. Business analysts use this data to measure financial performance, refine the company's marketing strategies, and enrich customer-360 profiles that indicate market trends.

In the past, having multiple data warehouses required these analysts to work with different views of data for different purposes, such as one type of repository for web and mobile data and another type for IoT data. Limited data storage resources required them to choose which data they wanted to keep and for how long, constraining their ability to conduct long-range historical analyses. Today, Office Depot uses the Data Cloud to unify all the structured and semi-structured data it needs. Having near-limitless storage and compute capacity allows managers to run detailed analytics much more quickly than before. For example, the programs that analyze weblog data to identify customer behavior on the company's ecommerce website used to take hours to run. Now, these programs run in minutes. Furthermore, Office Depot can maintain its historical data indefinitely, even if the value is not immediately apparent, rather than deleting the data due to limited storage capacity.

Office Depot also uses Snowflake's Secure Data Sharing technology to maintain data internally in one central location but share it across business units by creating near-instant access. The business team can easily share data externally, too, and is looking to monetize governed data sets within Snowflake Data Marketplace as new opportunities arise.

REMEMBER

The Data Cloud allows organizations to deploy controls that enable compliance with increasingly rigorous data governance requirements.

Delivering Superior Media and Entertainment Services

The media industry is in the throes of change as traditional business models evolve to accommodate digital content and online services, such as the video-on-demand (VOD) services reshaping today's television markets. Entrenched broadcasters and cable TV companies are losing ground to Netflix, Hulu, and dozens of other content providers, motivating many long-time subscribers to "cut the cord" on traditional cable packages, satellite services, and bundled TV content. Incumbent media providers are rapidly developing VOD platforms, crowding the market with hundreds of viewing options. Television networks, cable TV providers, streaming content providers, production studios, and the reigning tech giants Apple, Google, and Amazon are vying for subscribers in this crowded sector.

Other types of media and entertainment companies face similar hurdles in the race to embrace digital services. In addition to publishers and telecommunications firms, this vibrant industry includes:

- **Ad tech companies** that help marketing agencies and brands roll out and analyze their digital advertising efforts
- **Ad agencies, data measurement firms, and data enrichment companies** that offer creative marketing services, such as helping media companies create and enrich customer profiles
- **Art venues and sports businesses** that offer events and entertainment products, live and online
- An immense **video gaming industry** that offers packaged, networked, and cloud-based gaming options

One thing hasn't changed: all media companies need to acquire, retain, and engage audiences to earn their business via subscriptions, incremental purchases, and advertising revenue streams. Data holds the key to making the right content decisions, acquiring customers, preventing churn, and optimizing viewer experiences.

REMEMBER

Digital business models yield a steady stream of data that can help target customers, deliver personalized content recommendations, and measure the impact of marketing and outreach campaigns.

Unfortunately, many advertisers and brands are restricted in accessing the data they need and aren't sure how to merge it with other sources. Entrenched media and entertainment companies are saddled with legacy technology in which customer data is spread across silos. This fragmented data landscape complicates achieving crucial business goals and regulatory compliance — a costly and risky prospect as data privacy laws toughen.

Many media and entertainment companies understand the imperative of deploying modern data sharing to make up for these deficiencies. The Data Cloud brings together consumer and subscriber data across platforms and channels, along with purchase information. Data sharing is the starting point for unifying consumers' identities, gaining awareness of all their online activities, and ultimately revealing each subscriber's customer lifetime value (CLV).

Once these complete customer views are created, media and entertainment companies can use the Data Cloud to democratize access to these insights, enabling teams to make more-informed decisions on content production and acquisition, distribution, advertising, and product experience. Furthermore, media firms can share deeper insights about their audiences with brand advertisers, enabling advertisers to better predict the return on ad spend and increase the value of the ad inventory.

For example, advertisers collect granular customer data across many sources to create better ads and measure attribution from each ad campaign. The Data Cloud enables media and entertainment companies to cost-effectively store all their data in a globally available platform and to share governed data across business units, with advertisers, and among data enrichment partners.

REMEMBER

Within the Data Cloud, Snowflake data *cleanrooms* provide a secure way to share first-party data among media companies, advertisers, and other partners, without moving data or exposing PII. Sharing data via Snowflake enables businesses to manage and control data privacy as they capture and exchange valuable consumer data.

Additionally, Snowflake Data Marketplace includes relevant data sets and data services that simplify acquiring identity graphs, demographics data, and audience data, in conjunction with enrichment services from commercial data providers (see Figure 4-2).

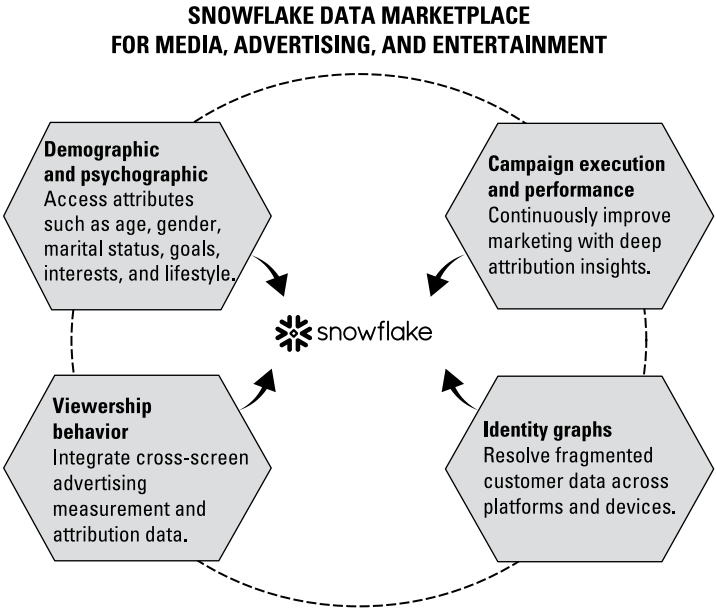

FIGURE 4-2: Snowflake Data Marketplace is the network for sourcing third-party data and data services for many industries, including media and advertising.

SUPERIOR EXPERIENCES FOR MEDIA AND ENTERTAINMENT

Media firms use the Data Cloud to securely share governed data across brands, advertisers, ad platforms, and enrichment providers. As a result, these organizations make data-driven decisions to increase subscriber lifetime value and increase return on advertising spend.

- **Media firms:** Unify your subscriber behavior data across platforms and channels with purchase information from brands to create 360-degree customer views.

- **Brands:** Unify your customer touchpoints across media channels and advertising platforms to optimize campaigns that improve conversions and sales.
- **Advertisers:** Connect your advertising spend with media consumption and purchase behavior to demonstrate marketing programs' return on investment.
- **Data providers:** Increase the scale of your digital and online offerings by seamlessly connecting your data with your customers' data via the Data Cloud.

Offering Better Public Sector Services

Accelerating scientific research by empowering agencywide collective analysis, strengthening data collaboration between federal and state agencies, and serving the public interest by improving live, secure data distribution to citizens and external consumers are just a few examples of how the Data Cloud can modernize the flow of data across and beyond government lines.

The push to embrace cloud-based technologies already has transformed IT infrastructures at every level of government. Federal, state, and local agencies have made significant strides in modernizing how data is collected, stored, and analyzed, all in service of their mission and in fulfillment of strategic IT mandates, such as annual strategic action plans from the U.S. Office of Management and Budget. Furthermore, the opportunity or need has never been greater for government agencies to mobilize data to improve citizen services, streamline administrative and operational inefficiencies, further research and innovation, and support interagency collaboration.

Through seamless and secure data sharing, the Data Cloud enables government entities to effectively eliminate data silos within agencies, across departments, among different government levels, and across clouds. Just as importantly, the Data Cloud ensures secure and governed access to all citizen data.

REMEMBER

Snowflake has done the hard work of ensuring the Data Cloud meets the privacy and security regulations required of government organizations and monitors compliance with those regulations, including FedRAMP, SOC, FISMA, NIST, and FIPS standards, on an ongoing basis.

Finding modern alternatives to the traditional ways of sharing data within government has become more pressing with the emergence of COVID-19. The state of California uses the Data Cloud to power a coordinated public health response by providing seamless, statewide data access to county officials, various state agencies, and residents to put the freshest data in the hands of those who need it most (see the case study in Chapter 1 for details). The state's shared data set includes case data from the California Department of Public Health, as well as hospital and lab data. Additionally, California leverages Snowflake Data Marketplace to provide the public with updated information about statewide cases, testing, demographics, homeless impact, and more.

Of course, the public sector must continually find ways to boost efficiencies and improve services for citizens and businesses while keeping a lid on costs. Public sector organizations can create more-effective and economical information systems by placing their data in the Data Cloud and using a growing set of third-party data services in Snowflake Data Marketplace. By leveraging the Data Cloud to address the urgent need for secure, unfettered access to all types of data, government entities have a simple, cost-effective solution for securely sharing essential information.

BETTER CITIZEN SERVICES VIA AGENCY COLLABORATION

Public sector organizations use the Data Cloud to improve collaboration with live, revocable access to critical data while upholding agency and departmental governance policies.

- **Federal agencies:** Securely access and share your operational and administrative data with agencies, state governments, and citizens.
- **State/local government:** Improve your data management and governance while receiving and sharing live, secure data with agencies, citizens, and partners.
- **Ecosystem partners:** Governments can reduce risk and improve data-sharing efficiency between public- and private-sector partners to strengthen cross-sector relationships.
- **Citizens:** Provide citizens with more effective, efficient government services powered by secure, near real-time data access.

IN THIS CHAPTER

- Starting with the right architecture
- Putting the Data Cloud in a historical context
- Unleashing near-limitless scale and performance
- Extending data and workloads across clouds
- Enforcing strong security and governance

Chapter 5
Drilling into Snowflake's Platform

Snowflake's founders set out to disrupt the previous 30 years of computing. They considered every aspect of data management and data operations in their quest to create a platform capable of near-unlimited power, scale, and world-class data governance and security. The result was a brand new, cloud-built platform, which powers the Data Cloud. This chapter examines the platform in detail.

Starting with the Right Architecture

Snowflake's platform was architected first and foremost for the cloud, unconstrained by any legacy technology heritage. Building on the growing availability, acceptance, and utility of public cloud services, Snowflake can be deployed across multiple public clouds and regions, yet is accessible via a single interface. It enables a number of diverse analytic workloads and can accommodate near-unlimited amounts and types of data, with minimal latency.

A UNIQUE, CLOUD-BUILT ARCHITECTURE

Four technology breakthroughs contribute to the platform's unique capabilities:

- A **multi-cluster shared data architecture** that powers near-unlimited scale, concurrency, and efficiency
- A **unified code base that works across multiple public clouds** as if they were one
- **Baked-in data governance and security and** features that ensure exceptional data protection
- Secure **data sharing technology** that allows any number of organizations to share and receive live data with each other — instantly and without having to move or copy data

This unique architecture allows many types of users to freely collaborate via data, both within an organization and with business partners, customers, the rapidly growing Snowflake Data Marketplace, and the thousands of organizations that comprise the Data Cloud.

Building on the Lessons of History

Traditional data platforms were architected to leverage a set of finite computing resources, often within the confines of an on-premises data center. Careful capacity planning is required to size each new data warehouse, data lake, data mart, or other workload. Because organizations don't always know in advance how popular these data environments will become, they have to overprovision — to deploy more hardware and software resources than they will initially need. Complicating matters, many data-driven workloads are characterized by occasional bursts of activity, such as when accountants close the books at the end of each month or when retailers have to predict and analyze the results of peak shopping periods each year. Sizing these environments to accommodate peak loads is wasteful, because they need all that capacity for only a small fraction of the time.

As analytic applications, data science applications, data engineering pipelines, and many other types of data applications have grown in popularity and importance, these legacy platforms have bowed under the strain.

Traditional data platforms don't scale well, and having a fixed set of compute and storage resources limits *concurrency* — the degree to which users can simultaneously access the same data and computing resources. Stymied by a linear architecture, they can't run multiple workloads in parallel, leading to long wait times for resources and data-driven insights.

With the rise of public cloud services, businesses suddenly could provision limitless amounts of compute and storage capacity. Theoretically, this allowed their traditional data environments to support a larger number of users and workloads. However, traditional database management systems and analytics platforms were not architected to take advantage of all this power and capacity. Many of them were simply "lifted and shifted" to the cloud, where they continued to operate under the limitations of their legacy heritage. These information systems have been architected to work with a finite set of resources and to utilize a single type of data — a data warehouse for structured data, a data lake for semi-structured data types, and a wide variety of local databases, data marts, and operational data stores — some in the cloud and others on premises — each created to solve a unique set of departmental needs.

Improving performance, lowering costs

Snowflake's platform was designed from the outset to take full advantage of the cloud's unique attributes. Compute and storage resources can be scaled independently to accommodate the specific needs of each application. A near-unlimited number of users can run concurrent workloads without degrading performance, such as ingesting data via a data engineering pipeline while simultaneously training a machine learning model to use that same data.

All parts of Snowflake's platform fit together — including the central repository where data is stored, the independent and near-limitless compute resources, and a layer of cloud services for security, identity management, transaction management, and many other functions.

Understanding why the right architecture matters

Why are these technical breakthroughs so important, and what do they bring to your business? As a Snowflake customer, you can spin up new diverse workloads without limitations, provisioning as little or as much computing power and storage capacity as you need down to the second. You can scale storage and compute resources independently yet be charged only for the exact resources you consume. Each workload receives a dedicated set of compute clusters, even if these workloads access the same underlying data (see Figure 5-1).

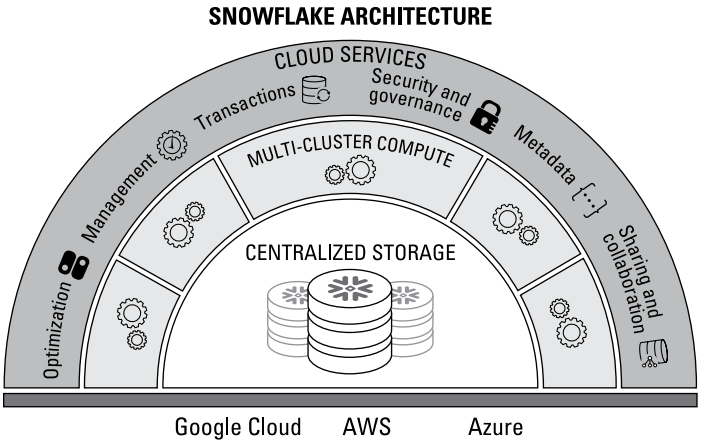

FIGURE 5-1: By separating storage from compute resources, Snowflake's platform permits multiple departments — and even multiple organizations — to run nearly unlimited concurrent workloads against the same data.

REMEMBER

Snowflake's platform provides virtually unlimited scalability and elasticity because there is virtually no limit to the number of active clusters you can dynamically deploy as each workload scales. You always have the resources you need, and you never have to pay for idle capacity.

REMEMBER

The platform is built on a patented multi-cluster, shared data architecture to enable near-limitless concurrency without impacting performance. Compute resources can scale on the fly or automatically to provide users with consistent performance, regardless of the number of queries issued at any given moment.

UNLEASHING PERFORMANCE AND SCALE

Snowflake's platform includes the following performance-enhancement services:

- **Micro-partitioning and clustering** features dynamically prune data during query execution. That means Snowflake scans only the subset of the data that is relevant to your query.
- **Materialized views** (database objects that contain the results of a query) improve performance of common and repeated query patterns, accelerating resource-intensive queries. As data is added or modified in the referenced base table, Snowflake updates materialized views automatically and transparently in the background.
- A **query acceleration** service automatically identifies parts of the query that could benefit from additional compute resources and prioritization. Query acceleration is particularly useful for data science and other scan-intensive workloads.
- A **search optimization** service further accelerates queries. For example, searches that include pattern matching within strings unlock additional use cases, such as analytics on log data, boosting the performance of cybersecurity analytics.

Establishing One Multi-Region, Multi-Cloud Service

The Data Cloud spans the major public clouds and many of their regions across four continents (see Figure 5-2). Snowflake regions are built on top of the infrastructure provided by the major cloud providers: Amazon Web Services, Microsoft Azure, and Google Cloud Platform.

A shared metadata layer defines a cohesive set of network services to orchestrate the easy transfer of data and workloads among regions and clouds. All Data Cloud analytic workloads run consistently, and all users have a cohesive experience, regardless of which region or cloud provider is used. Snowflake maintains transactional integrity for all data in any cloud, anywhere in the world. If a database operation is interrupted during execution, the data will be "rolled back" to its previous state, without any change to the database.

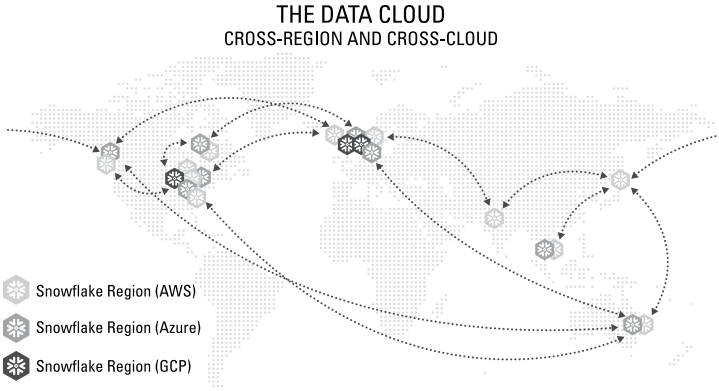

FIGURE 5-2: The Data Cloud spans four continents and the infrastructure of the three main cloud providers.

Enjoying an Easy-to-Use Platform

As discussed throughout this book, the Data Cloud allows you to develop and deploy whatever data applications you need without provisioning infrastructure or managing a complex software environment. For the most part, the platform manages itself. There is no need for tuning or optimizing performance. Database administrators (DBAs) don't need to manage indexes or partition data. The platform automatically manages each workload to maximize performance. Your DBAs can focus on deriving value from the data rather than on managing the platform.

The platform is the central point for many processes such as ingesting data, developing data applications, and securely sharing governed data throughout your extended ecosystem and beyond. Per-second billing enables each user and workgroup to pay only for the precise storage and compute resources used.

Predicting and monitoring usage

Built-in resource monitoring and management features provide complete transparency into usage and billing, enabling granular chargeback and showback capabilities tied to individual budgets. Snowflake helps customers monitor, manage, and *predict* overall usage in two important ways:

- Monitoring controls watch the platform to ensure it's consuming credits as expected
- Management controls make consumption predictable

Customers not only maintain control but can employ safeguards to eliminate runaway usage. For example, *auto suspend* and *auto resume* features automatically start and stop resource accounting when Snowflake isn't processing data. Customers can also set specific time-out periods for each type of workload, such as *immediately upon completion* for extract, transform, and load (ETL) jobs or *10 minutes* for query activities that require a warm cache.

Easing access to all types of data

The Data Cloud allows users to store, manage, and query structured and semi-structured data types using standard SQL. Snowflake expects to support certain types of unstructured data, such as audio, video, and PDFs, in a future release.

Whether you store data in Snowflake or an external repository, such as an object store from one of the cloud infrastructure providers, all users have a cohesive interface to query, view, and manage the data. Users can access data in external tables just as easily as they access it from the main platform. A common metadata layer ensures that all users obtain consistent results, and all workloads deliver consistent outcomes.

By establishing a common repository for many types of data and data workloads, the Data Cloud simplifies data management tasks and eliminates silos. Having centralized data reduces the number of stages the data needs to move through before users can act on it. Increasing the time value of data empowers users to act sooner and infuse analytics into operational business processes.

REMEMBER

Snowflake's near-zero maintenance platform enables any organization to operate across different public clouds and regions as a single cloud while satisfying industry and regional data privacy requirements.

Enforcing Strong Security and Governance

Having a unified code base spanning multiple cloud platforms is also important for enforcing strong data security. Snowflake applies the same security configurations and the same administrative techniques across all public cloud providers. Cybersecurity professionals don't have to resolve differences in audit trails and event logs or work with multiple management systems to encrypt data. Encryption is the default — data is encrypted while in transit and at rest, subject to customer-managed encryption keys.

REMEMBER

Snowflake's cybersecurity team conducts comprehensive monitoring, analyzes a constant stream of alerts, and enforces industry-leading cybersecurity practices to thwart threats and troubleshoot issues.

Role-based access controls ensure users can access only data they are explicitly permitted to see. These controls are applied to all database objects, including tables, schemas, and any virtual extensions, such as external tables — even when multiple types of data, many different regions, and more than one type of public cloud is involved. You won't need to hire people with unique skill sets for each public cloud, because Snowflake masks the differences via a universal layer of software services that spans all of them.

Policy-based security controls, such as dynamic data masking and row-access policy capabilities, redact values based on the permissions granted to each user querying the data or restrict the data's visibility depending on a user's role.

Secure views can further restrict access, such as to mask salary and Social Security fields or protect customer personally identifiable information (PII). A new tagging framework allows users to attach metadata to objects and enforce policies based on these tags.

Finally, the platform makes it easy to set up *row-level access policies* that define rules for accessing data. Snowflake's query engine dynamically applies the corresponding rules to expose only the rows each user can see. Implementing these access policies by using standard SQL syntax makes it easy for database administrators to create and update these policies.

A CONTINUOUS DATA PIPELINE FOR MACHINE LEARNING AND ANALYTICS

CASE STUDY

Convoy is a digital freight network that uses machine learning and automation software to connect shippers with carriers, efficiently moving millions of truckloads of cargo each year. Convoy's data solution relies on its data warehouse and an experimentation service that enables data scientists to analyze how new product features affect critical business metrics. It is also the foundation for Convoy's machine learning platform, which provides better matches, pricing, and service to customers.

But Convoy's existing solution was inefficient. Analysts relied on data that was bulk loaded via Apache Airflow every few hours. Not only did that prevent analysts from accessing current data, but storage costs were too high because engineers had to persist data in production databases for analytics. Convoy needed a more robust data platform that could run multiple workloads concurrently — data ingestion, machine learning, analytics, and data science — without causing contention. The solution had to accommodate JSON and VARIANT data types and support Convoy's established broad ecosystem of tools for building data applications, engineering data pipelines, creating business intelligence apps, and authorizing.

Convoy chose Snowflake to reduce data latency via micro-batch ingestion services, scale storage and compute resources independently to maximize workload performance, and support semi-structured data in native formats. Now, instead of running Apache Airflow jobs every couple of hours, Convoy runs thousands of micro-transactions in just a few seconds, using Snowpipe and the Snowflake Kafka connector to ingest data from Amazon S3 into Snowflake's Data Cloud. This continuous data pipeline reduces storage requirements because engineers no longer need to persist data in production databases. Snowflake Streams and Tasks loads new data into the Data Cloud within minutes of transactions occurring. Performance for all workloads is exceptional: Convoy has seen a 20x reduction in data integration time and 10x faster query responses with the Snowflake solution.

REMEMBER

Putting your data in the Data Cloud simplifies data governance by making it easier for business users to comply with industry-specific and region-specific data sovereignty requirements, protecting sensitive corporate data and personally identifiable information (PII). Snowflake unifies governance and security practices across all data and all workloads and helps you implement pertinent controls, even when accessing data stored in external tables. Universally applied data-access controls help you monitor and audit usage across your extended ecosystem. You can apply consistent policies to many workloads and all pertinent Snowflake accounts, even when your data resides among different clouds.

IN THIS CHAPTER

- Maximizing the value of data warehouses and data lakes
- Simplifying data science initiatives
- Engineering robust data pipelines
- Creating modern data applications
- Sharing data without limits

Chapter 6
Running All Your Workloads

Snowflake's platform has been architected to support your most diverse and critical analytic data workloads, driven by end-to-end data pipelines that move governed data from source to consumption. This chapter examines the most common workloads customers deploy in the Data Cloud: data warehouses, data lakes, data engineering, data science, data applications, and data sharing.

Deploying Data Warehouses and Data Lakes

Data warehouses were designed mainly to store and analyze structured data from relational databases. Data lakes emerged to store large amounts of raw, semi-structured data in its native forms. The early data lakes were built using on-premises Hadoop databases. More recently, organizations started using cloud-based object stores, such as Amazon Simple Storage Service (S3), Microsoft Azure Blob, and Google Cloud Storage, to construct data lakes.

These modern cloud data lake solutions free IT professionals from having to manage the hardware stack, but those professionals still have to create, integrate, and manage the software environment. This involves setting up procedures to transform data, along with establishing policies and procedures for identifying users, encrypting data, enforcing effective data governance, and many other time-consuming activities — all before they get down to the important work of creating useful analytics.

In contrast, the Data Cloud empowers all types of users to manage data with exceptional performance, using the industry-standard SQL language that powers the world's most popular analytics and data visualization tools. It combines the best of enterprise data warehouses, modern data lakes, and cloud capabilities to handle all data and workloads. Structured, semi-structured, and certain types of unstructured data (functionality that Snowflake expects to make available in a future release) can be staged in its raw form — either in the Data Cloud itself or in external tables within an object storage service.

Enhancing Core Workloads

The data warehouse was the first workload for which organizations used Snowflake to modernize their data analytics. As they continued with Snowflake, they deployed additional workloads in the Data Cloud. Snowflake continually enhances its capabilities for these core workloads. For example, the Data Cloud now supports *geospatial data* via the GEOGRAPHY data type, which uses a spherical-earth coordinate system to store such features as points, lines, and polygons on the earth's surface. You can load data from several supported formats, and Snowflake stores an optimized representation for fast queries.

Snowflake also routinely adds new features to improve data-access performance, such as the *search optimization* and *query acceleration* capabilities mentioned in Chapter 5.

The Data Cloud also includes new capabilities for maximizing the value of data lakes. This is especially important as more and more customers discover the value of utilizing semi-structured data from weblogs, Internet of Things (IoT) devices, social media networks, and other sources — much of which traditionally has been

stored in data lakes. Now, customers can store this data within the Data Cloud itself or keep the data in object stores and access it via external tables. They can leverage all the power, security, and governance the Data Cloud provides across the entire pool of data.

Other new features maximize control for customers pursuing hybrid cloud strategies. For example, you can use the Data Cloud to transform data and then push that data out of the Data Cloud via Snowflake's data export feature and into flat files or external tables. Snowflake maximizes flexibility by enabling these cyclical loading and unloading processes, giving customers the flexibility to blend many storage options within one coherent Data Cloud strategy.

Executing Other Critical Workloads in the Data Cloud

In addition to using the Data Cloud as a foundation for data warehouses and data lakes, many customers depend on Snowflake to execute data engineering, data science, and data application workloads. The reason is that the Data Cloud does much more than just store your data. It is also a proven solution for *processing* that data: running complex extract, load, and transform (ELT) operations within data pipelines; performing feature engineering within data science workloads; and handling the heavy compute requirements of today's data applications. Once your data is in the Data Cloud, processing the data there makes sense.

Engineering data pipelines

Data engineering involves collecting, preparing, transforming, and delivering data for analysis. This often happens via data pipelines that automate the transfer of data from place to place and transform that data into specific formats for certain types of analysis. The Data Cloud simplifies the process of making data "production ready" by putting it into a usable form, including manipulating data formats, scaling data systems, and enforcing data quality and security. The Data Cloud also helps data engineers create data pipelines to streamline data ingestion and transformation tasks.

TIP

Managing batch and micro-batch uploads and accommodating streaming data via Snowflake's Snowpipe service allows users to act on new data right away. For example, as soon as an item is scanned at a retail point-of-sale location, the data is available to inform restocking decisions. As soon as a bank suspects a fraudulent credit card transaction, the cardholder is queried to authorize or deny the charge.

The Data Cloud also has utilities for connecting event streams via the Snowflake Kafka connector. Developers can use SQL, Java, Scala, and Python to write and execute code within the Data Cloud using familiar concepts, such as data frames, to load data into columns and rows. Data engineers can build sophisticated data-ingestion pipelines that orchestrate multiple streams and tasks, all within the Data Cloud. Snowflake's Snowpark service, available in testing environments as of November 2020, allows data engineers to develop data pipelines using popular programming languages and then process the data either inside or outside the Data Cloud. Snowflake expects to make the service available for general consumption in a future release.

Simplifying data science

The Data Cloud streamlines the entire data science lifecycle of machine learning, artificial intelligence, and predictive application development. This empowers data scientists to acquire the data they need to develop new applications quickly and train the associated machine learning (ML) models. The Data Cloud can run complex ML algorithms internally and integrate with external ML services via Snowflake's external function framework, such as to invoke a risk-scoring model written in Java, Python, or Scala, and pull in the result set. Data scientists and data engineers can run these procedures externally or run them directly in the Data Cloud to simplify the overall process, making the Data Cloud ideal for data preparation and feature engineering.

Creating data applications

Historically, more than 30 percent of customers use Snowflake as their back end for data applications due to virtually unlimited scalability and a cost-effective, pay-as-you-go model. The Data Cloud offers near-limitless compute capacity for developing applications and executing resource-intensive workloads. This is an increasingly important capability as mobile app developers and software-as-a-service (SaaS) providers launch businesses focused on capturing, processing, and analyzing data.

MORE OPTIONS FOR DATA SCIENTISTS

Data scientists choose the Data Cloud for two primary reasons:

Access to all data: Data scientists require massive amounts of data to build and train ML models. The Data Cloud lets them store virtually all relevant data in one central location while maintaining the highest level of governance and simplifying the process of importing data into a wide range of data science notebooks and AutoML tools.

Data programmability: Snowflake integrates with external ML services so you can run advanced algorithms directly inside the Data Cloud, with support for Java, Python, Scala, and other programming frameworks. This gives data scientists and data engineers the flexibility to program in the languages of their choice.

Traditional software companies can use the Data Cloud to bring existing apps into the modern age. Many built their offerings on traditional data stacks, including legacy on-premises and cloud-washed data warehouses created before the cloud existed. Those solutions lack the cloud-native attributes that make the Data Cloud so powerful. Many traditional data apps struggle to process immense volumes of data, such as from application logs, websites, mobile devices, and IoT sensors, which frequently arrive in schemaless and semi-structured formats.


TECHNICAL STUFF

The Data Cloud provides nearly unlimited compute and storage resources for development, iteration, testing, and quality assurance (QA) of new data apps. It also supports the cloud-native development tools today's software developers use, including Python, Node.js, Go, .NET, Java, and SQL.

The Data Cloud also provides data recovery features to ensure deployments go smoothly. For example, to ensure all software releases and data are backed up properly, Snowflake provides continuous data protection services, along with built-in features that eliminate the need for traditional backup scripts and processes.

In addition to accommodating prominent application development tools and programming languages, the Data Cloud is designed to leverage DevOps principles, such as continuous integration (CI) and continuous deployment (CD), and enable developers to query structured, semi-structured, and certain types of

unstructured data (functionality that Snowflake expects to make available in a future release) with standard SQL.

By moving compute- and data-intensive workloads to the Data Cloud, data engineers, data scientists, and data application developers can dramatically simplify their IT environments. From building ML models to processing complex ETL transformations, Snowflake's multi-cluster, shared data architecture offers a robust processing engine for all these needs — and it's getting more capable every year.

Sharing Data Without Limits

As described throughout this book, the Data Cloud offers unique capabilities for organizations that wish to collaborate and share data. It facilitates internal data sharing and also external data sharing with other Data Cloud customers, either as part of a business partnership between customers or by accessing publicly available data and data services from Snowflake Data Marketplace. An entire data set or a selected subset can be shared with other Snowflake accounts. Diverse teams can collaborate without maintaining multiple copies of data or moving data from place to place. Consistent data governance enforces data-access restrictions dictating who can see what data, empowering all organization members to work in concert. Universally applied security and governance controls simplify compliance mandates and reduce cybersecurity risks.


TIP

One-to-one, one-to-many, and many-to-many data-sharing relationships can be established within a Snowflake region without copying or moving the data by granting live, read-only access to selected database objects. Additionally, users can share data and objects across regions and clouds by leveraging Snowflake's global replication technology, which transparently keeps the data in sync in the target accounts.

Sharing data within a broad or public ecosystem is also easier with the Data Cloud. Instead of the "one size fits all" approach offered by other data marketplaces, Snowflake allows data providers to structure unique data sets for each client. Snowflake's platform offers a level of governance that simply isn't possible with traditional file subscription services. Secure, row-level access can be used to specify which data is shared. Data providers can reach data consumers across clouds and regional boundaries just as easily as reaching data consumers across the hall.

IN THIS CHAPTER

- Creating Snowflake accounts
- Moving data into the Data Cloud
- Attending to security and governance
- Expanding your geographic footprint
- Sharing and monetizing data

Chapter 7
Six Steps to Getting Started with the Data Cloud

The Data Cloud fulfills Snowflake's vision for a data-connected world in which organizations can easily and securely access, unify, integrate, analyze, and share data. If you are a Snowflake customer, you are already a part of this global network. If you are a new customer, follow these steps to get started:

1. **Create an account.** Visit `www.snowflake.com` to create your first Snowflake account in any region of your choosing — most likely based on your geographic preference and favorite cloud provider. If other members of your organization create additional accounts, Snowflake automatically and securely connects them, so you can replicate data among accounts and collaborate easily across geographic regions. The command to create a new Snowflake account is executed directly from the initial account and is available to any user with an admin role.
2. **Load and integrate data.** Identify the data sources you plan to load into the Data Cloud. Will you stage data from an existing data warehouse or data store? Will you load data

continuously, or copy it in batch mode? Consider initial data loads and incremental updates. For example, you may set up a one-time bulk transfer for a historical data set. If you want to refresh that data set as new transactions occur, establish a data pipeline to manage a continuous stream of data. If you plan to load data from existing applications and databases, peruse Snowflake Data Marketplace for commercial data providers offering solutions to push data from those silos into the Data Cloud. You will discover connectors and adapters for common data types.

3. **Apply governance and security.** Building a data-driven business requires good stewardship to maintain data quality, uphold regulatory guidelines, and keep data secure. Decide who is responsible for governing and securing your data, both for your initial data loads and continuously as new data is ingested (see Chapter 2 for more details).

4. **Expand to other regions.** Your organization can expand its global presence in the Data Cloud by easily accessing additional Snowflake regions from your original account. You don't need to create additional Snowflake accounts for each new region. Accessing Snowflake from multiple regions may be valuable for satisfying GDPR data residency constraints or implementing a multi-cloud strategy.

5. **Share your data.** Do you plan to share data within your ecosystem? If so, how are you sharing data now? Do you plan to share data only within your organization or also with customers and partners? Identify outdated data sharing methods, such as FTP and email, and consider how you can replace them with Snowflake's Secure Data Sharing technology.

6. **Join Snowflake Data Marketplace.** Do you plan to utilize third-party data and data services from Snowflake Data Marketplace, or monetize your data and data services by offering them to other Snowflake customers? After creating an account and loading your data, follow these steps to participate in Snowflake Data Marketplace:

   - Identify *shares* — collections of objects to which you will grant consumer access.
   - Create one or more marketplace listings, along with your provider profile.
   - Enrich your own data by securely accessing ready-to-query data sets from more than 125 third-party data providers and data service providers.

Mobilize your data to best serve your customers and your business

IDC estimates the amount of data created from 2020 to 2023 will be more than the data created during the previous 30 years. Yet an Accenture survey revealed that just 37 percent of senior business and IT professionals indicated they fully achieved the outcomes they expected from leveraging the cloud. Enter Snowflake's Data Cloud — a global network where thousands of organizations unite their siloed data, easily discover and securely share governed data, and execute diverse analytic workloads. Read on to learn how your business can reach the next frontier.

Inside...

- The problems and opportunities with data
- What's possible in the Data Cloud
- Sharing data locally and globally
- Snowflake Data Marketplace
- The Data Cloud for industries
- The engine that drives the Data Cloud
- Running diverse analytic workloads
- Real-world Data Cloud case studies

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

Cover image: Courtesy of Snowflake

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

ISBN: 978-1-119-81061-2
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

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