LEARNING MADE EASY

Snowflake
Special Edition

Cloud Data Platforms

for dummies®
A Wiley Brand

What is a cloud data platform?

How to connect all data, insights, and users

How to choose a cloud data platform

Brought to you by:

snowflake

David Baum

About Snowflake

Snowflake started with a clear vision: Make modern data analytics effective, affordable, and accessible to all data users. The Snowflake Cloud Data Platform shatters the barriers that prevent organizations from unleashing the true value from their data. Thousands of customers deploy Snowflake to advance their businesses beyond what was once possible by deriving all the insights from all their data by all their business users. Snowflake equips organizations with a single, integrated platform that offers the data warehouse built for the cloud; instant, secure, and governed access to their entire network of data; and a core architecture to enable many other types of data workloads, including a single platform for developing modern data applications. Snowflake: Data without limits.

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

Cloud Data Platforms

Snowflake Special Edition

by David Baum

for dummies®
A Wiley Brand

Cloud Data Platforms For Dummies®, Snowflake Special Edition

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

ISBN 978-1-119-71389-0 (pbk); ISBN 978-1-119-71384-5 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. Some of the people who helped bring this book to market include:

Development Editor: Nicole Sholly

Project Editor: Martin V. Minner

Editorial Manager: Rev Mengle

Executive Editor: Steve Hayes

Business Development Representative: Karen Hattan

Production Editor: Mohammed Zafar Ali

Snowflake Contributors Team: Vincent Morello, Clarke Patterson, Leslie Steere

Table of Contents

Introduction

Data is fundamental to creating efficient business operations, discovering new revenue opportunities, and delivering exceptional customer experiences. But yesterday's tools for acquiring, storing, and sharing data have not kept up with today's burgeoning demands. Data is growing in volume and diversity, taxing the limits of the applications and workloads that depend on it, from simple reporting and data visualization apps to advanced machine learning algorithms. A cloud data platform is not a disparate set of tools. It's a unified, pervasive set of services that houses all your data and puts it to work, securely and consistently. It enables the modern data warehouse for analytics, high-performance access to data lakes for data exploration, data engineering for ingestion and transformation, data science for creating machine learning models, data application development and operation, and data exchanges for securely sharing data among authorized users.

About This Book

This book explains how you can establish a cloud data platform that minimizes the complexity of securely loading, transforming, integrating, analyzing, and sharing data associated with traditional on-premises and cloud solutions. In addition, you'll learn how to compare platform solutions and how those solutions support many types of applications, workloads, and clouds. You'll also get insights into how a cloud data platform can be the hub of a more insightful and responsive enterprise and a foundation for transforming outdated business processes and accelerating the uptake of digital business models. Read on to learn how you can establish a single cloud data platform for all types of data, without incurring the excessive cost and complexity inherent in traditional data management and analytics solutions. Discover how your IT team and other business units can step away from tedious data administration and focus on delivering great experiences with data. This book shows you how to:

- Supply business users with simple but powerful data analytics they need, when they need it.

- Make quick, accurate, and consistent business decisions based on simple and advanced analytic approaches.
- Eliminate the expense of buying, maintaining, and securing on-premises data management and analytics solutions.
- Easily deliver performance benefits and security measures that require almost no setup or maintenance.
- Efficiently share and monetize your data without making copies of the data or moving it from place to place.
- Empower your data scientists, data engineers, and data application developers to extract value from your data in ways not possible with your existing solutions.

Icons Used in This Book

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

TIP

Advice guiding you on better ways to use a cloud data platform in your organization.

REMEMBER

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

CASE STUDY

Case studies about organizations using cloud data platforms to improve their data management and analytics activities.

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 Snowflake Cloud Data Platform, obtain details about plans and pricing, view webinars, access detailed documentation, or get in touch with a member of the Snowflake team.

IN THIS CHAPTER

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

Chapter 1
Getting Up to Speed with Cloud Data Platforms

Over the last four decades, the software industry has produced a wide variety of solutions for storing and analyzing data. The most prominent of these solutions have included data warehouses and, more recently, data lakes. These solutions made it possible to work with both traditional forms of data and newer data types generated from websites, mobile devices, Internet of Things (IoT) devices, and other more recent technologies. But traditional data warehouses and data lakes were too complex, costly, and limiting to use. They did catapult data and analytics from the enterprise back office into the executive suite, but they weren't architected for the differences and dynamics of today's data. Nor could they take full advantage of the cloud's native capabilities, such as instantly scaling nearly unlimited data storage and compute resources.

Over the past ten years, vendors that have offered on-premises data warehouses and data lakes have migrated their solutions to the cloud. For the most part, these first-generation cloud data solutions provide better price and performance than their on-premises cousins. But they weren't built from the ground up for the cloud, so they can't take full advantage of the resources and elastic capabilities of the cloud.

The industry has learned from the benefits and drawbacks of these solutions and carried that knowledge forward. Each solution was a stepping stone and solved an important problem. Yet, the world's insatiable appetite for accessing and analyzing more data in more ways has always remained at least one step ahead.

Traditional enterprise data warehouses (EDWs) did a decent job of analyzing an organization's traditional, or structured, data to inform business decisions. As the desire for analytics steadily increased, these solutions couldn't handle the volume, velocity, and variety of data and analytic workloads prevailing in the modern enterprise. These traditional EDWs also offer little help to developers and software engineers creating new types of cloud-based data applications.

Forward-looking organizations seek a single, powerful, and extensible *data platform* that can handle legacy needs in conjunction with a raft of new and pressing requirements, such as contending with a broad array of data types, supporting new types of analytics, enabling robust and automated data pipelines, securely sharing data, developing data applications, and scaling effortlessly to accommodate ever-increasing analytic workloads. They need a data platform that guarantees scale, performance, and concurrency — a platform combining the best components of on-premises EDWs, cloud data warehouses, modern data lakes, and secure ways to share and consume shared data from within a single, cohesive solution. They need a cloud data platform.

UNDERSTANDING DATA TYPES

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

Defining the Cloud Data Platform

A modern cloud data platform must deliver the best of legacy data warehouses and data lakes, the true benefits of the cloud, and much more (see Figure 1-1). Based on the lessons of the past, such a platform should be built on four fundamentals:

- A single platform that supports many different workloads
- Secure, governed access to all data and all types of data
- Instant and near-infinite performance and scalability
- A zero-maintenance, cost-effective cloud service that's instantly available and extremely easy to use

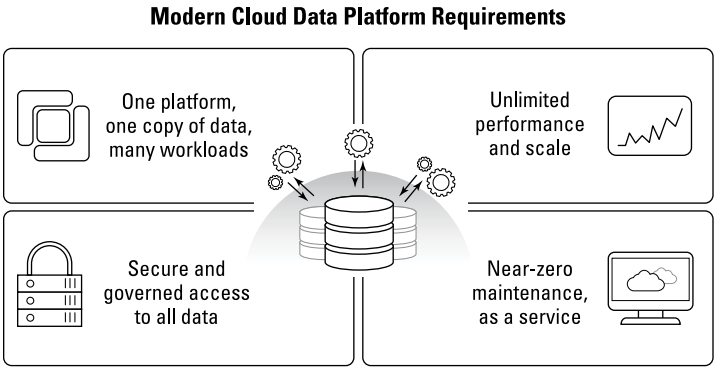

FIGURE 1-1: The fundamental elements of a modern cloud data platform.

Understanding the Multi-Cluster Shared Data Architecture

With a traditional data platform, fixed compute and storage resources limit *concurrency* — the ability to perform many tasks simultaneously and/or to allow many users to access the same data and resources. With a multi-cluster, shared data architecture, compute and storage resources are separate and can be scaled independently to leverage all the resources of the cloud. This allows multiple users to query the same data without degrading performance, even as other workloads are taking place simultaneously, such as ingesting data via a data engineering pipeline or training a machine learning model.

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

- **Storage:** A single place for all structured or semi-structured data
- **Compute:** Independent compute resources dedicated to each individual workload to eradicate contention for resources
- **Services:** A common services layer that handles infrastructure, security, metadata, and query optimization

Ideally, this architecture should be *cloud agnostic*, providing a consistent layer of services to each cloud region and all cloud infrastructure regardless of which cloud platforms are used (see Figure 1-2).

Built on versatile binary large object (BLOB) storage, the *storage layer* holds your data, tables, and query results. This scalable repository should handle both structured and semi-structured data and should span multiple regions within a single cloud and across competing clouds.

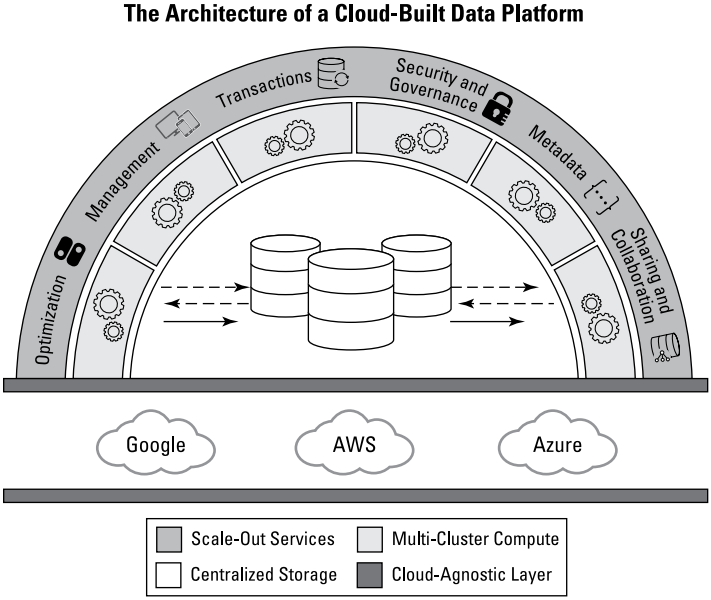

FIGURE 1-2: A modern cloud data platform should seamlessly operate across multiple clouds and enable many types of modern data workloads.

The *compute layer* can process enormous quantities of data with maximum speed and efficiency. You should be able to specify the number of dedicated clusters you want to use for each workload or have the option to let the service scale automatically.

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

Why You Need a Cloud Data Platform

Whatever industry or market you operate in, learning how to easily and securely utilize your data in a multitude of ways will determine how you run your business and how you address current and future market opportunities.

Having a greater volume and variety of data at your fingertips, and stored and managed in a consistent way, opens the door to these opportunities. To take advantage, you need a data platform to easily store and organize data regardless of format; provide convenient access to that data; and improve the speed at which you can analyze and share data across your organization and within your ecosystem of customers, suppliers, business partners, and beyond.

A cloud data platform also helps you take advantage of three important technology trends:

- **The rise of the cloud:** Traditional data center infrastructure is sized for a known set of data management tasks. The cloud offers unlimited capacity for storing and processing data. This opens the door to an unprecedented number of concurrent high-performance workloads within a centralized platform.
- **The explosion of data:** Data will continue to grow in both size and variety, driven, in part, by the proliferation of Internet, mobile, social, and IoT technologies that produce immense quantities of raw but vital data. This growth happens every second of every day. And within the new data lies valuable insights for organizations with the technology, resources, and commitment to tap its potential.

- **The diversification of analytics:** On some level, every business user is an analyst. When, where, and how these users perform analytics is changing quickly, however, as data and analytics are democratized across the enterprise. As the appetite for data continues to grow, analytics will become central to more and more business processes, from historical quarterly reporting to forward-looking predictive and prescriptive analytics.

UNIFYING DATA AND WORKLOADS

CASE STUDY

Devon Energy, an independent oil and natural gas company, relies on historic and predictive analytics to operate its assets at maximum capacity. The company depends on a diverse set of structured and semi-structured data sources to make these determinations. It tried to create an on-premises data platform and associated data warehouse to leverage these diverse data sets, with disappointing results: One didn't scale well; one was too complex to build and maintain; and the third, a data lake based on Hadoop, was difficult for users to access. Frustrated, Devon chose a cloud data platform strategy to unify its data and extend analytics throughout its enterprise.

This platform, known as the Devon Data Hub, combines a data lake and a data warehouse into one solution, ensuring governed access to a growing volume of data ingested from more than 30 data sources. With more than 1,000 users issuing approximately 4 million queries per month, the data hub has two layers to support different classes of users: The enterprise data layer is created and governed by the IT staff; the community data layer is maintained by business users, with minimal governance to reduce the barriers to entry. This second group of relatively nontechnical users can access data without involving the IT team or having to extract and join it using third-party tools. Reports that took days now run in minutes, and data loading processes that took more than half a day now run in half an hour.

Structured and semi-structured data now lives in the cloud data platform, yet all workloads are managed by a common set of cloud services governing how data is accessed, shared, and secured. Devon pays only for the cloud resources it uses and can effortlessly scale each workload in response to business needs, without having to overprovision capacity or worry about resource-contention issues.

IN THIS CHAPTER

- Understanding the problem with traditional data management approaches
- Forming a new vision for data platforms
- Looking at the evolution and limitations of data lakes
- Reviewing the advantages of a cloud-built data platform

Chapter 2
Considering the Exponential Growth and Diversity of Data

Reviewing a little recent history is essential to understanding today's pressing needs for a cloud data platform. Not so long ago, organizations dealt mostly with data entered manually into large software solutions, including customer relationship management (CRM), enterprise resource planning (ERP), supply chain management (SCM), and many other types of business applications. Sounds big, but the amount of data was small and fairly predictable. Most of it was maintained in on-premises data centers.

Fast-forward to the present. Today's businesses are awash in data, not only from these enterprise applications but also from mobile phones, websites, point-of-sale terminals, and billions upon billions of interconnected devices that track the world around us. In fact, some modern businesses *are* their data (ride-sharing services, for example). A large amount of this data is gathered by sensors and ingested by software applications that reveal moment-to-moment conditions, from the surge of energy through an electric grid to traffic conditions on a busy highway.

Legacy on-premises and first-generation cloud data platforms can't keep up with the relentless creation, acquisition, storage, analysis, and sharing of these diverse data sets. Much of the data is semi-structured, which means it doesn't fit neatly into the traditional data warehouse that emerged 40 years ago. In addition, some data types, such as images and audio files, are wholly unstructured, and must be maintained as binary large objects (BLOBs) within an object-based storage system.

Examining the Problems with Traditional Data Management Approaches

In response, many organizations have established unique solutions for each type of data and each type of workload, such as a data warehouse for operational reporting, data marts for departmental reporting and analytics, and data lakes for data exploration on data they don't yet know the value of. They also have implemented specialized extract, transform, and load (ETL) tools to rationalize different types of data into common formats, and set up data pipelines to orchestrate the exchange of data among databases and computing platforms. Each of these "point solutions" requires independent hardware resources and specialized software, with each connection point creating implementation and maintenance issues.

But what if those organizations had a complete data platform that could adequately address the performance, scalability, and concurrency required to deal with these diverse data requirements? Consider traditional analytics. On-premises enterprise data warehouses (EDWs) work best with well-defined, structured data. Some newer data warehouses have demonstrated the ability to work with both structured and semi-structured data. However, as explained in ensuing chapters, these newer solutions don't have the architectural flexibility to simultaneously work with both types of data as well as to support the multitude of other workloads that need to run at the same time, such as a data engineering service that ingests streaming data.

Cloud-based data warehouses can help by offering some elasticity and better price/performance than their on-premises alternatives, but many of them are limited in their ability to handle all types of data for all types of users. These limitations have motivated the

formation of data lakes, which are designed to hold huge quantities of raw, semi-structured data in their native formats.

Chronicling the rise of data lakes

Initially, the core technology used to create data lakes was based on the Apache Hadoop ecosystem, an open source software framework that distributes data storage and processing among commodity hardware. The ecosystem includes tools for ingesting data, processing it, and storing it in its native form.

Unfortunately, most Hadoop-based data lake projects did not yield much value, partially as a result of their complex distributed architectures, which required custom coding for data transformation and integration. As Forrester analyst Noel Yuhanna notes in his report "Big Data Fabric 2.0 Drives Data Democratization" (May 9, 2019), "simply putting lots of diverse data into Hadoop or a data lake won't magically create meaningful insights without further integration, transformation, enrichment, and orchestration."

The data lake achieved what the data warehouse could not: a way to store all your data, regardless of structure, in a single repository. However, business users found it difficult to integrate and analyze this vast pool of data, and many organizations had a hard time finding, recruiting, and retaining the highly specialized data professionals needed to keep the data lake viable. As a result, most of today's Hadoop-based data lakes can't effectively organize all of an organization's data, which originates from dozens or even hundreds of data streams and data silos that must be loaded at different frequencies, such as once per day, once per hour, or via a continuous data stream.

Finally, the Hadoop architecture has rudimentary controls for determining who can see the data and was not designed to comply with important industry standards governing the security and privacy of that data, including the Health Insurance Portability and Accountability Act (HIPAA), Payment Card Industry Data Security Standard (PCI DSS), and General Data Protection Regulation (GDPR). Without adequate data security controls to resolve vulnerabilities, and data governance procedures to orchestrate changes, data lakes can quickly become *data swamps* — unorganized pools of data that are difficult to use, understand, and share with business users. The greater the quantity and variety of data, the more significant this problem becomes with a data lake.

TIP

Whether you're dealing with data from weblogs, Internet of Things (IoT) data from equipment sensors, or social media data, the volume and complexity of these semi-structured data sources can make it difficult to obtain insights from a conventional data lake.

Leveraging the cloud

Keeping all your data on site, either in Hadoop or in another type of data lake or data warehouse, quickly gets expensive. It's difficult to know in advance how much compute and storage you might need to analyze these huge data sets. The tendency is to over-provision: to pay for more resources than you need most of the time to handle occasional spikes in usage. Even when the storage infrastructure is sized to handle peak loads, once a system gets popular those workloads invariably increase and may exceed those limits. This leads to resource contention, where multiple workloads compete for a finite set of computing and storage resources. It also leads to queueing workloads, a problem exacerbated by the formation of specialized data marts in which a portion of a data warehouse is sequestered and made available for departmental analytics, such as for marketing, finance, or sales. No matter how much capacity planning you do up front, it can be difficult to predict the future, and you may not have the capital budget to continually increase the compute and storage capabilities required to support these workloads.

Some cloud vendors have introduced alternative data lake solutions, such as Amazon Simple Storage Service (S3), Microsoft Azure Blob, and Google Cloud Storage. These highly elastic cloud storage solutions enable you to store unlimited amounts of data in their native formats. Organizations can leverage these general-purpose object storage environments to create their own data lakes from scratch.

TECHNICAL STUFF

Object storage systems manage data as *objects* that include the data itself, metadata that describes what the data is used for, and a unique identifier. These systems allow organizations to store massive amounts of structured, semi-structured, and unstructured data.

However, while these cloud data lake solutions freed customers from having to manage the hardware stack as they did with Hadoop, customers still had 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, establishing effective data governance, and many other essential activities — all before they get down to the important work of creating useful analytics.

The progression from on-premises data warehouses and data lakes to cloud-based alternatives, culminating with a modern cloud data platform, is illustrated in Figure 2-1. Only a modern cloud data platform empowers all types of users to manage all types of data with exceptional performance, using the industry-standard SQL language that powers the world's most popular analytics and data visualization tools.

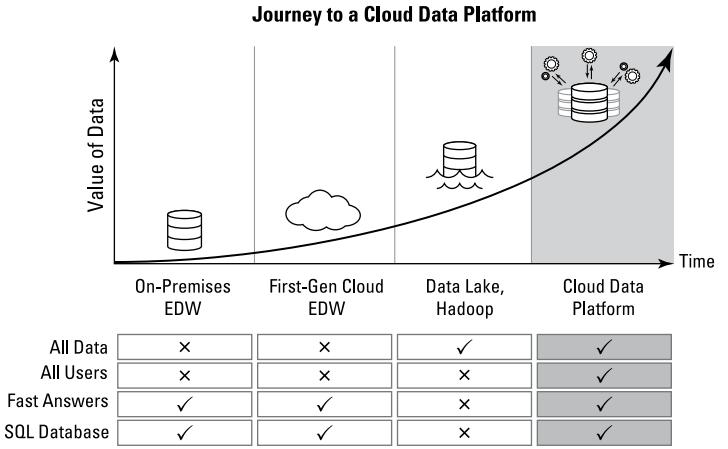

FIGURE 2-1: A cloud data platform combines the best of enterprise data warehouses, modern data lakes, and cloud capabilities to handle all data and workloads.

Cloud-Washed Versus Cloud-Built

Not all cloud solutions have the same pedigree. Many began their lives as on-premises solutions and were later ported to the cloud — sometimes called *cloud-washed* solutions. By contrast, *cloud-built* platforms have been designed first and foremost for the cloud.

To identify a solution built on a cloud-optimized architecture, look for the following characteristics:

- Centralized storage for all data
- Independent scaling of compute and storage resources
- Near-unlimited concurrency without competing for resources
- Ability to load and query data simultaneously without degrading performance
- Availability across clouds and regions with built-in replication to enhance business continuity and simplify expansion
- A robust metadata service that applies across the entire system
- The capacity to securely share and discover shared data

The importance of each of these attributes is explained in succeeding chapters.

THE ADVANTAGES OF THE CLOUD

With usage-based pricing and near-infinite scalability, cloud solutions are a natural place for storing and integrating large quantities of data. The properties of the cloud that make it particularly well-suited for large data sets and intensive analytics workloads include the following:

- **Unlimited resources:** Cloud infrastructure delivers nearly unlimited resources, on demand, within seconds. You can pay by the second for what you use, making it possible to instantly support any scale of users and workloads without compromising performance.
- **Rapid ramp-up:** When you choose a cloud-built solution, you avoid costly up-front capital investments in hardware, software, and other infrastructure, not to mention the ongoing operational costs of maintaining, updating, and securing on-premises systems. This advantage allows you to focus on securely analyzing, sharing, and even monetizing your data.
- **Natural integration point:** With the rise of new types of data exchanges, marketplaces, and other online data services, a great deal of the data organizations need comes from external sources. Predicting how much data you will use in advance is nearly impossible. Bringing it together in the cloud is easier and cheaper than trying to accommodate the data in-house.

Exploring the Advantages of a Unified Cloud Data Platform

Today's organizations want an easier way to cost-effectively load, transform, integrate, and analyze unlimited amounts of structured and semi-structured data, in their native formats, in a versatile data platform. They want to simplify and democratize the exploration of that data, automate routine data management activities, and support a broad range of data and analytics workloads. And they want to collect, store, and analyze their data in one place, so they can easily obtain all types of insights from all their data.

A properly architected cloud data platform brings this all together with a common set of *services* that streamline how that data is used. As discussed in subsequent chapters, the platform also enables you to consolidate diverse analytic activities, orchestrate the secure sharing and exchange of data, and create modern data analytics applications and populate them with data.

By creating a single place for all types of data and all types of data workloads, a cloud data platform will dramatically simplify your infrastructure, without incurring the costs inherent in traditional architectures. For example, by centralizing data, you reduce the number of stages the data needs to move through before it becomes actionable, which eliminates the need for complex data pipeline tools. By reducing the wait time for data, users can obtain the data and insights they need, when they need them, so they can immediately spot business opportunities and address pressing issues.

Chapter 3 explains how a modern cloud data platform can enable a long-term strategy for maximizing your data assets by streamlining essential workloads, such as analytics, data sharing, data ingestion, and data science.

CASE STUDY

GAMING THE SYSTEM

With more than 300 million registered players globally, Electronic Arts (EA) is a leader in digital interactive entertainment. EA offices, or studios, located throughout the world create immersive gaming experiences. This geographic diversification has bred tremendous creative diversity. It has also complicated human resources (HR) and finance processes because each studio franchise observes unique contractual obligations, legal scenarios, and cultural norms.

EA needed a cloud data platform to enforce corporate-wide consistency. Previously, each franchise collected and analyzed data with its own set of business intelligence (BI) tools, so the company couldn't possibly gain a 360-degree view of revenues. In some European markets, it took 11 days to consolidate financial results from the previous week. Now, EA's cloud data platform consolidates multiple types of data and governs it according to a common set of rules, eliminating duplication and misinterpretation. It's easier to roll up data to the corporate level, while management has visibility into the performance of each franchise. Furthermore, queries that took 30 seconds on EA's previous legacy platform now take 3 seconds, and reports that took a month to create are available in minutes.

Since the company started distributing its games via digital downloads rather than through physical media, occasional spikes in demand have placed crushing pressure on EA's content distribution system. IT professionals had to refresh the data once per week. With escalating volume, the old platform couldn't handle the load. Now EA's new cloud data platform enables the company to add compute and storage capacity on demand to accommodate seasonal market fluctuations, and EA doesn't have to pay for extra capacity during slower times of the year. The new platform also simplifies maintenance. "As soon as we dropped in the data . . . the [cloud data platform] started performing without tuning, indexing, or tablespace management," reports EA's chief BI architect, Vlad Valeyev.

IN THIS CHAPTER

- Taking stock of your data and analytic needs
- Discovering the value of a single platform for all data
- Understanding the importance of secure and governed data
- Demonstrating the power of a zero-maintenance platform

Chapter 3
Selecting a Modern Cloud Data Platform

Organizations outgrow their existing data platforms for a variety of reasons. In many instances, limitations surface in response to competitive threats that require an organization to acquire new types of data and experiment with new types of data workloads. For example, a data scientist may set out to create a predictive analytics model that helps the sales team mitigate customer churn. Machine learning is a data-intensive activity, and the new model requires a huge amount of new data to train machine learning models. The success of this sales initiative depends on the ability to store and process a large volume of new data about customer behavior.

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

Meanwhile, the sales team wants to study social media posts to discern trends, issues, and attitudes within the customer base. This data arrives as JavaScript Object Notation (JSON), in a

semi-structured format. Sales analysts want to visualize this data in conjunction with some enterprise resource planning (ERP) transactions stored in a relational database, including historic data about sales, service, and purchase history.

Finally, another division within the organization wants to display these purchase patterns as data points on a digital map. This requires new data from a geographic information system, along with unstructured image data to depict products and sales locations. Traditional data platforms can't keep up with the new data engineering, data science, data sharing, and other abilities organizations need in order to acquire and harness this new data.

Reviewing Your Analytic Needs

These are the kinds of business scenarios that can cause an organization to look for a more modern and versatile data platform (see Figure 3-1). Consider your own needs. You may have a data platform or data management system that works well for a certain type of data, but you want to take on new business projects that require the analysis, and sharing, of new data types. Or perhaps you want to rethink your data acquisition strategy — to engineer better methods for ingesting and loading data into your platform. As you gather more data and the value of that data grows, you may want to monetize that data via a cloud data exchange to turn your data into a strategic business asset.

Even without adding new types of workloads, legacy on-premises or cloud-washed data platforms may make it impossible to handle an escalating data load. Data may be coming in more quickly and in larger quantities than before, and it might be too expensive to scale the system within the confines of your current architecture.

Whether you're motivated by business reasons or technology reasons to upgrade to a new platform, all of these scenarios reveal four fundamental principles — four "must haves" that guide your selection of an optimal cloud data platform:

- A single platform for many workloads . . .
- That enables secure, governed access to all data . . .
- And delivers unlimited performance and scale . . .
- Offered as a service via a usage-based, zero-maintenance platform

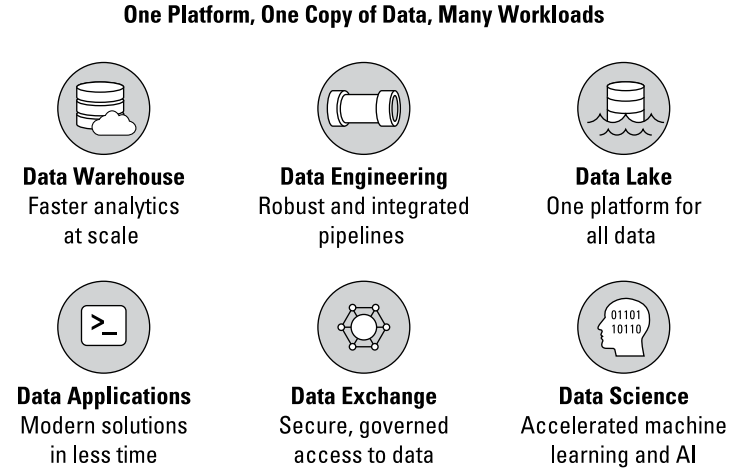

FIGURE 3-1: A cloud data platform should be powerful, flexible, and extensible to handle the most important data workloads.

A Single Platform for All Your Data Workloads

To maximize the performance of all data workloads, you need a highly elastic cloud platform. The platform should dynamically bring together the optimal set of resources for each workload and each usage scenario, including the precise balance of compute power and storage capacity. All you should have to worry about is loading, querying, and working with your data.

A complete cloud data platform can scale to any number of concurrent workloads. Data engineers, data scientists, and business analysts can all interact with the same single source of data without resource contention or data delays. These platforms are also extensible, enabling you to connect other software solutions for transforming and analyzing data.

The platform should automate everything from how data is stored and processed to how transactions are managed, data is secured, and metadata is managed. One platform, governed by one set of

services, will support the needs of analysts, data scientists, data engineers, and application developers creating new data products for your internal stakeholders or your external customers. And it will enable you to maximize your data resources and more easily deliver products and services to solve pressing business needs.


TIP

When properly architected, a cloud data platform can offer virtually unlimited scale for every workload, making it possible for thousands of users to analyze and share data concurrently, with no bottlenecks.

Securing and Governing Data

A cloud data platform should focus on simplifying how you protect and govern your data, to thwart breaches while complying with all industry and regional data regulations. Data security depends on three essential factors:

- **Confidentiality:** Preventing unauthorized access to data
- **Integrity:** Ensuring data is not modified or corrupted and is properly governed
- **Collaboration**: Providing ways for disparate teams to collaborate across shared sets of curated data

A cloud data platform should also have provisions for *securing* your data while enforcing comprehensive data governance. This ensures that no matter what type of workload, application, or procedure is invoked, the data is seen only by authorized users, and that all users obtain consistent results. For example, finance teams need sales data to forecast financial performance. Product managers require marketing data to develop new products and services. Executives need data from all parts of the enterprise to make timely, data-driven business decisions. Enterprise-wide data governance helps these teams work in concert and share consistent business results. (See Chapter 8 for more details on data security and governance.)

Data is an asset. To maximize its use, you must be able to securely make it available, across and outside your organization, for a number of benefits. You must be able to *share* that data, so it doesn't have to be copied or moved, as discussed in Chapter 6.

Unleashing a Zero-Maintenance Platform

Traditional data platforms are derived from legacy environments that work on closed networks with local data. As analytic applications and other data-driven workloads gain popularity, these legacy platforms slow down under the weight of too many concurrent users and burgeoning volumes of data. When your cloud data platform is architected first and foremost for the cloud, it can automatically provision limitless amounts of compute power to support any number of users and workloads, without affecting performance.

Easy access to data, and the ability to share it, is a top priority for most organizations. According to the Forrester Research report "Design Data Governance for the Data Economy," 80 percent of organizations want to expand their use of external data. A more recent Forrester report, "The Insights Professional's Guide to External Data Sourcing," finds 47 percent of organizations currently commercialize their data while 76 percent have launched, or plan to launch, initiatives for improving their ability to source external data.

All of these organizations depend on data, but none of them wants to own tedious database maintenance, system management, and IT administration.

A modern data platform built for the cloud from the ground up should enable you to develop and use whatever data applications you need without having to provision infrastructure or manage a complex software environment. It should be taken care of as part of a zero-maintenance platform that automates security and manages each workload to guarantee top performance. Zero maintenance means no infrastructure to manage, and no knobs to turn, so you can focus on the data and not on managing the platform.

Your platform should also offer per-second billing, which enables each user and workgroup to pay only for the precise storage and compute resources used in per-second increments, so you never have to pay for idle capacity.

A robust cloud data platform should make it easy to store and analyze data, and should also be the centerpiece for ingesting

new data, for developing data applications, and for building data exchanges that allow you to easily and securely share governed data. With a flexible architecture, delivered as a service, that's both powerful and easy to use, the potential is practically limitless.

DEMOCRATIZING DATA FOR ANALYSIS

CASE STUDY

When U.K. supermarket giant Sainsbury's set out to make analytics more cost-effective and accessible to its employees, the first step was to consolidate, cleanse, and reorganize all of its data assets into a cloud-built environment. In addition to being the second-largest general merchandise and clothing business in the U.K., Sainsbury's owns a bank and hundreds of retail stores. The organization has thousands of employees and millions of customers and performs billions of transactions each year.

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

These three systems now publish raw data directly to the cloud data platform. It initially lands in a data lake and then flows through a curation layer to populate a dashboard that streams data to the digital trading teams. Data scientists and machine learning engineers can access granular raw data straight from the data lake; store managers can obtain standard reports via a self-service dashboard. The cloud data platform separates storage and compute resources, improving performance and eliminating resource contention for thousands of users.

"One of the core tenets of our strategy is to know our customers better than anybody else, so we can offer them fair prices and great quality," Sainsbury's Chief Data and Analytics Officer Helen Hunter said. "This involves bringing together disparate data assets and democratizing data for the good of every user."

IN THIS CHAPTER

- Assessing time to value for your cloud data platform
- Confronting the limitations of on-premises solutions
- Dissecting the differences between "platforms" and "toolkits"

Chapter 4
Why Your Data Platform Belongs in the Cloud

Chapters 1–3 show the value of having a single but integrated cloud data platform, powered by a comprehensive services layer and architected to take full advantage of the boundless resources of the cloud. In this chapter, you learn the key considerations for a cloud data platform and why this solution provides more value at much less cost compared to implementing a traditional, on-premises data platform. First, note that a modern cloud data platform is both *evolutionary* and *revolutionary*.

- **Evolutionary:** The computer industry has spawned a variety of sophisticated on-premises data platforms. Each new solution builds on the capabilities and breakthroughs of its predecessors. Many of these data management and analytic solutions still exist: According to IDC, 75 percent of analytic workloads are still on premises; only about 25 percent have moved to the cloud. But the analyst firm predicts that split will shift from 75/25 to 50/50 by 2023.
- **Revolutionary:** Most of these popular legacy systems were conceived and developed long before the cloud became a viable alternative. They were architected to leverage a set of finite resources and to work within the confines of a corporate data center. This imposes inherent

limitations on how fundamental storage and compute resources are provisioned and used. Cloud-built platforms were designed for just the opposite — to utilize nearly unlimited storage and compute resources.

Assessing Time to Value

Organizations are gradually moving more and more of their analytic systems from on-premises data centers to the cloud. Several good reasons account for this migration. First, time to value: Deploying a conventional data platform in your data center can take a year or more. A lot can change in that time — from small reassessments of the business requirements to major reversals that may cause stakeholders to question the purpose or value of the project. A long implementation cycle exposes the endeavor to staff attrition, economic downturns, revenue shortfalls, and the perennial delays that stem from "scope creep" in response to all of these developments.

Calculating Hard Costs

Second are the upfront costs you'll incur to implement an on-premises solution. Hardware costs typically include computer servers, storage devices, data center space to house the hardware, a high-speed network to access the data, and redundant power supplies. You need industrial-strength HVAC systems and other physical infrastructure necessary for keeping the system up and running. If your data platform hosts mission-critical data and workloads, you may need to configure and maintain a comparable disaster recovery site. You'll also have to spend hundreds of thousands of dollars in software licensing fees, and possibly additional fees for each named user of the platform. Annual support contracts typically comprise 20 percent of the original license cost.

You'll need to consider the cost of extra hardware and software to house *data marts* (subsets of a data warehouse oriented to meet the demands of specific groups of users or lines of business), and the cost and effort of increasing the storage and compute capacities of these configurations once they reach their peak.

Sizing the System

An on-premises deployment requires skilled IT professionals to size the system and configure it for expected workloads. You typically have to size the system for peak usage, even if that need occurs only one day a month, one week every quarter, or one month every year. More often than not you're paying for something you're not actually using. Or, you could fail to size the system large enough, so it fails under a peak load and leaves you with unhappy customers or unfulfilled orders. The configuration generally includes technology professionals to determine the following:

- Number and speed of central processing units (CPUs)
- Amount of memory
- Number, type, and size of disks for storing data
- Input/output (I/O) *bandwidth* (a measure of how much data can be transferred at a given time)

With a cloud data platform, teams spend time working with data, not managing infrastructure. Storage and compute resources can scale independently, elastically, and infinitely to meet the shifting demands of each workload, as well as to accommodate peak usage periods without reducing the performance of many concurrent workloads. This is an immense advantage that becomes clearer in subsequent chapters.

The cloud, by its very nature, gives you flexible capacity, ensuring you'll always have what you need, when you need it. Not all cloud data platforms are the same, however, so you'll want to make sure you have the correct architecture to take full advantage of what the cloud offers.


REMEMBER

A cloud data platform can deliver value much more quickly, and at a more attractive price point, than an on-premises solution. Exorbitant CapEx costs associated with developing and implementing an on-premises system are replaced by the affordable OpEx, usage-based pricing of a cloud solution.

Hiring Skilled Personnel

Throughout the lifetime of your data and analytics projects, you'll need to retain skilled IT personnel to configure, deploy, and maintain your on-premises solutions. For example, a traditional data lake may require IT professionals who understand the Hadoop ecosystem, including tools such as Java and Spark. These skilled specialists will likely spend their time on low-value activities, such as managing infrastructure, rather than creating new data applications that add value to the business. This can create potential bottlenecks when issues arise, because you may not have enough staff to handle development, maintenance, and troubleshooting. Furthermore, it means you, not the data platform vendor, are ultimately responsible for keeping the system performant and continuously up and running.

This responsibility extends to securing your data, and the stakes are high. If you opt for an on-premises data platform, you're solely responsible for managing sensitive data, which involves constant attention to firewall protection, security protocols, data encryption, user roles, and access privileges, as well as responding in real time to cybersecurity threats. Effective data security is complex and costly to implement, especially in terms of human resources. Poorly implemented security measures expose you to even more costs if your data platform is breached.

Protecting and Recovering Data

All information systems are vulnerable to data loss from equipment failure, power outages, theft, vandalism, and natural disasters. If those systems are on-premises, it's your responsibility to keep them online. You must back up those systems regularly and store backups at a remote location. You'll want a backup power supply to prevent data loss and ensure your data is always available to your users. If a disaster or other mishap occurs, you'll need procedures in place to recover data within the expected time window, using the most recent backups. If your data is mission-critical, you may want to maintain a disaster recovery site in another location, along with the infrastructure, licenses, and personnel to ensure there are no gaps in service.

A functionally robust cloud service includes these backup and recovery capabilities as an inherent part of the basic service. It is

architected for geographic redundancy. Your data will be stored off premises and automatically backed up to multiple locations to protect your intellectual capital and to enable you to recover quickly if problems arise. The best data platforms support multiple cloud providers in case you need to easily and quickly migrate data to another cloud for business or technology reasons.

Knowing the Differences Between Platforms and Toolkits

Your data platform vendor may claim to have a complete "platform," but look at its origins. Today's enterprise software vendors rose to prominence on the strength of their legacy applications, databases, and tools, which were created largely to run in on-premises data centers. In step with the rest of the software industry, these vendors have made a wholesale effort to move their legacy solutions to the cloud. But those on-premises architectures were designed to leverage a finite set of data center resources, not the boundless resources of the cloud. Some of their fundamental limitations cannot be easily rectified.

Furthermore, many software vendors established dominance partly by acquiring "point" solutions that handle specific tasks and integrating them into a toolkit of cloud offerings. Thus, their cloud "platforms" include a disparate collection of business applications, database management systems, application development tools, data integration tools, data preparation tools, security and identity management utilities, and business intelligence and analytics options, each with unique query and display technologies. Yes, a platform-as-a-service (PaaS) layer loosely connects these solutions, but it's far from seamless, and in some cases the burden is on you to integrate the components.

The same issues plague the generic offerings from cloud storage and compute providers. These vendors began by offering infrastructure-as-a-service (IaaS) solutions — fundamental hardware, such as servers and storage devices — as cloud services. Their customers have gradually moved software applications and databases into these cloud environments, happy to let the vendor take charge of the hardware layer. Over time, these cloud vendors added a PaaS layer, which includes operating systems, hypervisors (computer software or hardware that creates and runs virtual

machines — basically, virtual environments that function as virtual computer systems), application servers, database management systems, and other basic software infrastructure. This laid the groundwork for a software-as-a-service (SaaS) layer as well, which includes various types of software applications, tools, and utilities, some supplied by the primary cloud vendor and others supplied by technology partners.

It's up to you to select the components you want to use and to ensure they are properly integrated, secured, governed, and maintained, and work properly with your data. A complete cloud data platform should handle most of these tasks for you.

ONE PLATFORM FOR ALL WORKLOADS

CASE STUDY

Yamaha, the world's largest manufacturer of musical instruments, uses its cloud data platform to ingest and analyze large amounts of transactional and relationship data. Hourly ingestion of enterprise resource planning (ERP) and customer relationship management (CRM) data into the platform yields timely sales insights about dealer order volume, customer credit limits, inventory availability, and sales pipeline performance.

The platform makes it easy to share data and run multiple workloads throughout the organization. Customer success teams use ERP data to monitor product support metrics, measure customer satisfaction, and decide when to ramp up hiring. Marketing and finance teams rely on data from the platform to inform pricing and budgetary decisions. Having a multi-cluster shared data architecture has eliminated resource contention, enabled fresher data imports, and expedited data-visualization activities. There are numerous other benefits as well, including the ability to instantly scale for any number of users, near-zero maintenance and per-second pricing, streamlined ingestion of large data sets, separate storage and compute resources that eliminate reporting delays, and native SQL support for data exploration.

IN THIS CHAPTER

- Recognizing how today's workers use data
- Establishing secure, democratized access and collaboration
- Realizing what's possible with a cloud data platform
- Envisioning a future-proof architecture: any workload, any cloud

Chapter 5
Comparing Cloud Data Platforms

Not all data platforms have the same pedigrees. Many began their lives as on-premises solutions or toolkits and were later ported to the cloud. As opposed to these *cloud-washed* solutions, *cloud-built* platforms have been designed first and foremost for the cloud. This chapter introduces the technologies that differentiate these two strategies.

Acknowledging How Today's Workers Use Data

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

For example, managers, supervisors, and line-of-business (LOB) workers generally want data delivered within the context of the business processes they use daily. They want to visualize data through intuitive charts and graphs. They want to monitor business conditions by displaying data through portals, menus, and

operational dashboards — ideally through easy-to-use apps on computers, tablets, and phones.

Professional analysts are better equipped to deal with data in its raw form. Many have been trained to create business intelligence apps, to load data into spreadsheets, and to create pivot tables and generate custom reports. They're comfortable creating data models, joining tables, and imposing a sensible structure on a data set. They're familiar with using SQL to create and issue queries.

Data scientists leverage massive data sets to build and train machine learning models. They manipulate data from one or more central locations and run procedures to consolidate, cleanse, and import data into a wide range of data science tools. They create predictive analytics for the business community.

Data engineers know how to collect data from various locations, load it into data repositories, and move it from place to place. They use data pipelines and extract, transform, and load (ETL) tools to populate databases, in real time or in batch mode, and to refresh those databases at periodic intervals.

IT departments typically employ software engineers, web developers, database administrators, and DevOps (a combination of development and operations) professionals to develop and deploy data-driven applications. These professionals collect data and apply it to unique business problems.

They all want access to that data fast: *low-latency data access*. Your data platform must be optimized to provide near-real-time access to an ever-growing collection of data.

Democratizing data access and collaboration

As they carry out their respective activities, an overriding imperative drives all these types of users: the need for *unlimited access to data*. Business professionals, data analysts, data engineers, data scientists, and application developers need to confidently work with the same single source of truth that ensures consistent outcomes.

The right cloud data platform makes this experience possible. You can leverage all your data simultaneously without having to import or export data from one system to another. This is a sharp contrast from legacy data platforms, which are typically

optimized for a single type of data, forcing you to create silos for different data and workloads. Now you can "open the aperture" to all your data without introducing additional complexity.

REMEMBER

Your cloud data platform should democratize access to your data so people of all skill levels — LOB managers, financial analysts, executives, data scientists, and data engineers — can readily access the data and perform the data and analytic workloads critical to their roles.

Maintaining data where it resides

Many data platforms masquerade as cloud solutions, when in fact they have merely been "lifted and shifted" from a legacy, on-premises heritage. This diverse software-solutions landscape invariably spawns diverse data sets, such as:

- A data warehouse populated with structured data from enterprise software systems in one geographic region
- An on-premises data lake, constructed for newer, semi-structured data types in another region
- An assortment of local databases, data marts, and operational data stores established in the cloud and on premises and created to solve unique departmental needs

Furthermore, each public cloud provider has different levels of regional presence, and data sovereignty requirements may require you to keep data processing operations within the regions you serve, leading to even greater data diversity.

With a cloud-built platform, everything is designed from the outset to take advantage of the cloud. All parts of the solution fit together — including the central repository where data is stored. This enables you to store all of your business data once and in a single solution. All data is accessible to all workloads, leveraging a common fabric of cloud services for security, identity management, transaction management, and other functions.

Whether your data is stored in the cloud data platform itself or in an external repository, such as an object store from one of the cloud infrastructure providers, all users should have a single interface enabling them to view and manage that data. A cloud data platform should allow you to access data in these external tables just as easily as you can access it from the main platform — and with exceptional performance.

ASKING THE RIGHT QUESTIONS

Cloud-built means created from the start to take advantage of the cloud, with each cloud platform component designed to complement the others. To ensure you obtain superior, cloud-built capabilities, ask the vendor these questions:

- Does the platform completely separate but logically integrate storage and compute resources to scale them independently, maximizing performance and minimizing cost?
- Does it gracefully handle a near-infinite number of simultaneous workloads (concurrency) without degrading performance or forcing users to contend for a finite set of resources?
- Will it support diverse data types and integrate data from multiple locations?
- Will you incur downtime to perform maintenance and upgrades?
- Can the platform do all of this automatically without the complexity, expense, and effort of manually tuning and securing the system?

Realizing What's Possible with a Cloud Data Platform

A flexible cloud data platform allows organizations to use their traditional business intelligence (BI) tools and newer, more advanced technologies devoted to artificial intelligence, machine learning, data science, and other forward-looking data analytic activities (see Figure 5-1). It combines data warehouses, subject-specific data marts, and data lakes into a single source of truth that powers multiple types of workloads, including the development and operation of data applications.

A cloud data platform should simplify the process of storing, transforming, integrating, managing, and analyzing all types of data. It should also streamline how diverse teams *share* data, so they can collaborate on a common data set without having to maintain multiple copies of data or move it from place to place. Consistent data governance makes it easier to enforce data-access restrictions dictating who can see what data. Having these

controls in place improves data security and reduces risk, so all members of an organization can work in concert to boost revenue, improve efficiency, and reveal new and disruptive opportunities.

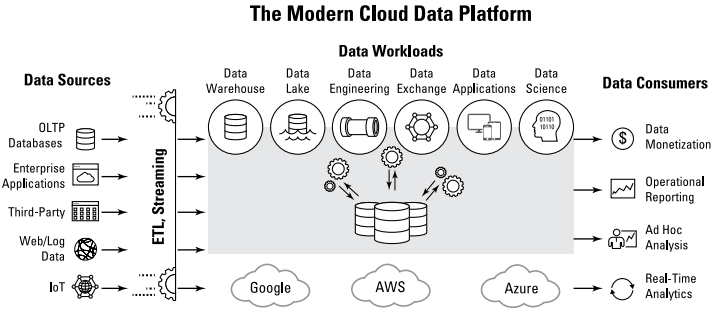

FIGURE 5-1: A cloud data platform should handle any data source and data workload, and serve data consumers of all levels and needs.

As explained elsewhere in this book, a modern cloud data platform is built on a *multi-cluster, shared data architecture* that offers the scale, flexibility, security, and ease of use large and emerging organizations require. End-to-end platform services automate everything from data storage and processing to transaction management, security, governance, and metadata (data about the data) management — simplifying collaboration and enforcing data quality. Delivered as a service and with consistent functionality across multiple regions and clouds, the data platform must enable instant and near-infinite elasticity.

A centralized data repository allows an organization's business units and its business partners and customers to securely share governed data without having to copy that data or move it from place to place. This versatile architecture simplifies data sharing and minimizes governance and compliance issues that arise from managing multiple copies of the same data set.

Enabling Global Flexibility: Any Workload, Any Cloud

As described in Chapter 1, a modern cloud data platform should compile queries and coordinate database transactions across multiple regions and clouds. Using a technology called *global state*

management, it can also maintain transactional integrity for all data, in any cloud, anywhere in the world. This type of "all or nothing" database integrity means that if a database operation is interrupted in midstream, for any reason, it is "rolled back" without any change to the database, ensuring that partial transactions will not affect your data's overall integrity.

In addition, a common metadata layer enforces *consistency*, so all users obtain consistent results, and all workloads enable consistent outcomes, no matter how many queries and transactions are conducted.

These unique capabilities allow the platform to securely govern data stored in multiple clouds and across multiple regions, worldwide. This type of multi-cloud, cross-cloud platform has tremendous advantages for sharing data, enabling business continuity, and complying with unique industry-specific and region-specific data sovereignty requirements, as Chapter 7 reveals.

IN THIS CHAPTER

- Establishing an efficient data sharing architecture
- Controlling access to sensitive data with secure views and shares
- Ensuring users have a consistent, up-to-date view of the data
- Supporting data governance initiatives

Chapter 6
Enabling Secure and Governed Access to Data

Many business initiatives require users to share data. Traditional data sharing methods require organizations to constantly copy and transfer that data from place to place. Users may like having local copies of the data, but system administrators quickly lose control: The more copies of data, and the more users, the more difficult it is to secure that data and govern it over its lifecycle. Data becomes stale the moment it's copied. An organization's ability to securely share it across its enterprise, its business ecosystem, and beyond is almost impossible with traditional data sharing methods.

Some organizations attempt to automate data-synchronization operations by setting up programmatic links among data sources, such as with File Transfer Protocol (FTP) or application programming interfaces (APIs). In other instances, users simply email data back and forth, or rely on file sharing utilities to share content. However, because the data quickly becomes dated and must be continually refreshed, it requires constant copying, movement, and management. In addition, file sharing utilities rarely uphold corporate, industry, and region-specific data security and regulatory policies.

Establishing a Secure Data Sharing Architecture

Imagine the possibilities if you could provide on-demand access to ready-to-use, live data inside a secure, governed environment. What if you could easily share data among multiple business units and seamlessly exchange data with your business partners, with no copying or synchronization, and everybody leveraging the same single source of truth? A cloud data platform should allow you to seamlessly and securely share data internally among departments and subsidiaries and externally with partners, suppliers, vendors, and even customers.

By sharing data within a modern cloud data platform, you can enable *live access* to any subset of your data for any number of data consumers, inside and outside your organization, to support analytics endeavors and other data-driven initiatives. All database objects are centrally maintained and updated by the data platform, in conjunction with end-to-end security, governance, and metadata management services. You don't have to link applications, set up complex procedures, or use FTP to keep data current. And because data is shared rather than copied, no additional storage is required (see Figure 6-1).

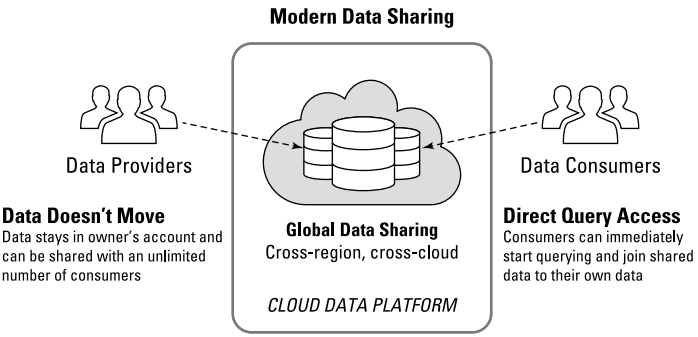

FIGURE 6-1: A cloud data platform streamlines data sharing between data providers and data consumers, even across multiple regions and clouds.

A modern cloud data platform allows you to easily share subsets or "slices" of your data in a secure and governed way. Rather than physically transferring data to internal or external consumers, you can authorize those consumers with read-only access to a governed portion of a live data set, accessible via SQL. That means

shared data can be accessed by large numbers of concurrent consumers, without competing for resources. Performance is exceptional due to the cloud's limitless storage and compute resources.

Securing sensitive data

What if you have sensitive data in your database? If portions of a database table are subject to strict security and confidentiality policies, sharing the entire table would expose that sensitive data. Choose a modern cloud data platform that allows data providers to control access to individual database tables and *secure views*. (A view is the result of a query that displays some or all of the fields in a table.)

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


TIP

A cloud data platform facilitates data sharing by enabling authorized members of a cloud ecosystem to access live, read-only versions of the data. If you don't have to keep track of data in multiple places, controlling what the data includes and how it should be updated becomes easy. It's also easy to monitor who interacts with your data.

Guaranteeing transactional consistency

As explained in Chapter 1, the services layer is the linchpin of a modern cloud data platform. It manages metadata, transactions, security, and other operations. It performs these activities locally or globally across multiple regions and clouds as it tracks, logs, and directs access to data for every database element and object contained within the platform.

Within the context of data sharing, the services layer enforces *transactional consistency* to ensure all users see a consistent, up-to-date view of the data at all times. If the data provider updates the data set or commits new transactions, all data consumers can simultaneously view these updates and immediately query the shared data, with ACID-based consistency.

ACID (atomicity, consistency, isolation, and durability) is an industry-standard consistency model that ensures transactions are valid even in the event of processing errors, power failures, and system crashes.

Finding New Ways to Support Data Governance Initiatives

Because a cloud data platform allows organizations to easily share their data and consume shared data across their business units and with partners, it also offers new ways for governing data across the enterprise. As it applies to master data in particular, a cloud data platform can serve as a new tool for defining, evolving, and controlling core data attributes across business units or with partners.

A complete cloud data platform also makes it easy to exchange and monetize data through a data marketplace or *exchange* that links data providers with data consumers. It leverages the principles of modern data sharing to enable one-to-one, one-to-many, and many-to-many data sharing relationships. Data stewards use exchanges to publish governed sets of data for consistent consumption by a wide range of parties. As changes are made to these governed data sets, they are immediately available for consumption, greatly simplifying the standardization and distribution of live, governed data across the enterprise.

IN THIS CHAPTER

- Ensuring worldwide business continuity
- Implementing the right clouds for the right locales
- Complying with data sovereignty regulations
- Using a single code base for all clouds to simplify administration

Chapter 7
Maximizing Options with a Multi-Cloud Strategy

Your cloud data platform should allow you to easily move data workloads among multiple clouds and multiple regions within each cloud, so you can locate data where it makes sense and mix and match clouds as you see fit. Having this type of deployment flexibility assists with geographic expansion, improves business continuity, and allows you to use different cloud services in different regions — without vendor lock-in. You can securely move data anywhere in the world while selecting the cloud storage options that best meet your needs. Here are a few key terms to be aware of:

REMEMBER

- **Multi-cloud** means you can use several different clouds, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform.
- **Cross-cloud** means you can access data from all of these clouds simultaneously, as well as seamlessly migrate data workloads from one cloud to another.
- **Data replication** is the process of storing data in more than one location to ensure it's protected and always available in the event of a regional outage.

Enabling Business Continuity

A complete and modern cloud data platform automatically replicates databases and keeps them synchronized across regions and clouds, boosting availability, automating the *failover* of workloads (switching to a redundant or standby computer), and guaranteeing instant access and recovery for databases of any size (see Figure 7-1).

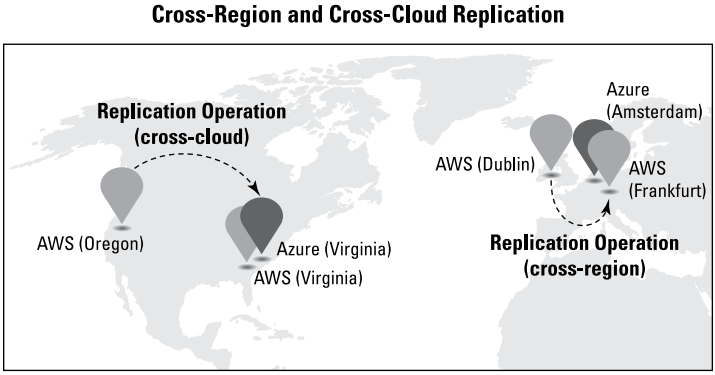

FIGURE 7-1: A modern cloud data platform automatically replicates databases and keeps them synchronized across regions and clouds.

Cross-cloud data replication is especially important for business continuity and disaster recovery. If an outage takes place, you can instantly resume data and analytic activities without incurring downtime. For example, if a disaster occurs in a particular part of the world, or if you experience problems with one of your cloud providers, you can immediately access live data replicated to a different region or cloud service. Your data platform should automatically synchronize data across cloud platforms and regions.

Enlisting the Right Cloud for the Right Requirements

Each department and division within your organization may have unique requirements. Rather than demand all business units use the same provider, a multi-cloud strategy allows each unit to use the cloud that works best for that unit. This is a strategic

advantage for global companies, because not all cloud providers operate in all regions. It's also useful if you acquire a company that has standardized on a cloud different from the one you're using.

Some industries, such as healthcare and financial services, uphold data sovereignty requirements that mandate you keep data within the same region as the people who use that data. Cultural and linguistic considerations may also motivate organizations to store data within the regions they serve. If your platform supports this type of seamless data portability, you'll have an easier time complying with regulations. The right data platform gives you the flexibility to select the cloud provider with the best offering and strongest regional presence.

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


TIP

Each public cloud provider has different levels of regional presence. A cross-cloud data platform should enable the free and secure movement of data anywhere in the world, while also allowing you to select cloud storage vendors that meet the needs of each application and business unit.

Minimizing Administrative Chores with a Single Code Base

When working with multiple clouds, how do you ensure the same security configurations and administrative techniques apply to all of your cloud providers? Will you have to resolve differences in audit trails and event logs? Will your cybersecurity experts have to deal with different rule sets, or work with multiple key management systems to encrypt data? A unified code base spanning all cloud platforms simplifies these operations. You won't need to hire people with unique skill sets or maintain familiarity with multiple clouds.

With a single code base, accessible via a cloud service, the software is always on and instantly updated to include the latest functionality. The cloud provider is responsible not only for keeping the cloud data platform online, but also for upgrading and maintaining the system. Rather than having to endure the large disruptive software upgrades that plagued yesterday's on-premises environments, most cloud providers update their software continuously. Small, incremental updates minimize the need to retrain your staff or sustain major changes to your operation.

Sharing Data More Broadly

As discussed in Chapter 6, data replication is the fundamental technology that allows you to share data when that data spans multiple regions and clouds. It also extends your reach when you want to monetize your data or bring business partners into your own data exchange.

Look for a cloud data platform that is *cloud agnostic,* so you can configure data sharing services to meet each set of requirements, no matter which clouds your partners use.

Having a single source of truth eliminates the time-consuming task of keeping multiple data repositories in sync. And no matter where your data resides, the user experience will be exactly the same, even as you uphold geo-residency requirements and comply with data sovereignty mandates. For example, analysts can query data housed in AWS, Microsoft Azure, or Google Cloud Platform, using the same procedures for all of them.

IN THIS CHAPTER

- Benefitting from high availability, reliability, and resiliency
- Keeping data safe with end-to-end security and governance
- Integrating third-party tools easily and securely

Chapter 8
Leveraging a Secure, Resilient, Always-On Architecture

A cloud data platform is not a disparate set of tools or services. Instead, it should be one integrated platform that enables many workloads, including data warehouses for analytics, data lakes for data exploration, data engineering for ingestion and transformation of data, data science for creating AI and machine learning models, data application development and operation, and data exchanges for easily and securely sharing data among authorized users. Each of these workloads has unique attributes, but all depend on the universal principles of availability, reliability, extensibility, durability, security, and governance. That's what this chapter is about.

Benefitting from a Reliable, Always Up-to-Date Platform

A cloud data platform should be always on. As discussed in Chapter 4, this is a primary requirement of a cloud solution. Whatever SaaS provider you choose, that vendor should provide a service-level agreement (SLA) that stipulates uptime requirements of a minimum of 99.995 percent, which equates to about 28 minutes of downtime per year.

A cloud data platform should be available, reliable, dependable, and constantly and seamlessly updated with new functionality. Today's cloud solutions are updated continuously, so you always have the latest functionality, and you never have to endure a lengthy upgrade. The typical best practice for a cloud provider is to add new features every two weeks, without disrupting customer data and workloads. Ideally, guardrails should be in place if customers want to "opt out" of a particular update until they are ready.

Protecting Data with a Secure, Governed Platform

Protecting your data and complying with industry and regional data protection regulations must be fundamental to the architecture, implementation, and operation of a cloud data platform service. All aspects of the service must be centered on protecting your data — not just backing it up routinely and replicating it to multiple geographic locations, as described previously. Your cloud data platform vendor must demonstrate that it upholds a multilayered security strategy that includes access control, data storage, and physical infrastructure in conjunction with comprehensive monitoring and alerts and verifiable cybersecurity practices (see Figure 8-1).

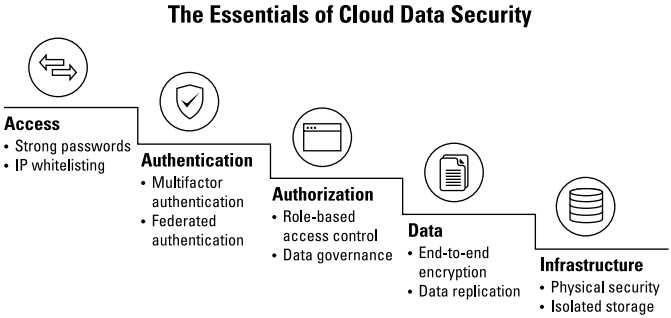

FIGURE 8-1: Security in five stages, from access to infrastructure.

Enforcing access, authentication, and authorization policies

A cloud data platform should always *authorize* users, *authenticate* their credentials, and grant users *access* only to the data they're authorized to see. *Role-based access control* applies these restrictions based on each user's role and position in the company. For example, finance personnel can view data from the general ledger, HR professionals can view data on salaries and benefits, and so on. These controls can be applied to all database objects including the tables where data is stored; the schemas that describe the database structure; and any virtual extensions to the database, such as views.

However, even with strong access and authorization procedures in place, data breaches can result from *social engineering attacks*, in which authorized users unwittingly reveal sensitive information to a hacker, such as a username and password. To help prevent this occurrence, your cloud data platform should provide *multifactor authentication*, which requires secondary verification, such as a one-time security code sent to a user's mobile phone.

Single sign-on (SSO) procedures simplify password management by making it easier for people to log into multiple applications directly from other sanctioned applications. A modern cloud data platform should employ *federated authentication* to centralize these identity management and access control procedures for all applications that use the data platform, making it easier for your team to manage user credentials and database privileges.

Encrypting data

Encrypting data involves applying an encryption algorithm to translate readable text into *ciphertext*, which contains a form of the original *plaintext* that is unreadable by a human or computer without the proper cipher to decrypt it. *Decryption*, the inverse of *encryption*, is the process of turning ciphertext into readable plaintext. This is fundamental to security, and your cloud data platform should ensure it happens, automatically, all the time, and without impacting the performance of your data-dependent workloads. Data should be encrypted *in transit* and *at rest*, which means from the time it leaves your premises, through the Internet or another network connection, and into the platform. It should be encrypted when it's stored on disk, when it's moved into a staging location, when it's placed within a database object, and when it's cached within a data repository. Query results should also be encrypted.

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

Establishing a strong cybersecurity posture

Equipment failures, network breaches, and maintenance mishaps can result in data loss and introduce inconsistencies into your data. Your cloud data platform vendor should have procedures for safeguarding against accidental or intentional destruction. The vendor should also employ round-the-clock monitoring of your data and infrastructure, both by humans and by machines. Security "events," generated by cybersecurity monitoring systems that watch over the network, should be automatically logged in a tamper-resistant security information and event management system (SEIM). Automatic alerts should be sent to security personnel when suspicious activity is detected.

As an added protection, file-integrity monitoring tools ensure critical system files aren't tampered with, and IP address *whitelists* enable you to restrict access to only trusted users and networks.

(A whitelist specifies sanctioned email addresses and domain names.)

Software patches and security updates must be installed on all pertinent software components as soon as those updates are available. Your cloud data platform vendor should also deploy periodic security testing (also known as *penetration* testing) by an independent security firm to proactively check for vulnerabilities.

Physical security measures in the cloud data center should include biometric access controls, armed guards, and video surveillance cameras to ensure no one gains unauthorized access. All physical and virtual machines must be further controlled with rigorous software procedures for auditing, monitoring, and alerting.

Ensuring data protection, retention, and redundancy

All database objects (data structures such as views and indexes used to store and reference data) should be centrally maintained and updated by the data platform, as described in Chapter 6. This is a fundamental principle that makes all your other security practices more effective. For example, it's one of the things that makes secure data sharing possible: Rather than physically transferring data to internal or external consumers, and losing control of that data, you can keep it in the platform and authorize data consumers with the appropriate levels of access. Chapter 6 also discusses how you can limit exposure of the underlying database tables by setting up *secure view*s.

Data replication, a concept introduced in Chapter 7, is a technology that automatically maintains your data in multiple locations to ensure it's protected and always available in the event of a regional outage. In case of a mishap, this enables you to instantly restore previous versions of your data in a table or database within a specified retention period, as governed by your SLA with your cloud provider. A complete data-retention strategy should go beyond duplicating data within the same cloud region or zone. It should replicate that data among multiple availability zones for geographic redundancy.

At first glance, this may appear to negate the centralized "all data in one place" strategy advocated throughout this book. It doesn't. That's the beauty of data replication within the context of a cloud data platform: Although your data may be maintained in multiple availability zones, the platform applies end-to-end security, governance, and metadata management services. It's all synchronized by the platform, with centralized command and control.

Requiring tenant isolation

If your cloud data platform vendor uses a multi-tenant environment, in which many customers share the same physical infrastructure, make sure each customer has a virtual data platform isolated from all other data platforms. For storage, this isolation should extend down to the virtual machine (VM) layer, governed by independent directories and unique encryption keys. Some vendors also offer dedicated virtual private networks (VPNs) and bridges from a customer's systems into the cloud data platform to keep all communications and data transmissions separate and distinct.

Maintaining governance and compliance

Among other things, data governance ensures corporate data is properly accessed and used and that day-to-day data management practices comply with all pertinent regulatory requirements. Governance policies establish rules and procedures to control the ownership and accessibility of your data. The types of information that commonly fall under these guidelines include credit card information, Social Security numbers, dates of birth, IP network information, and geolocation coordinates.

Compliance isn't about just robust cybersecurity practices. It's also about ensuring your data platform provider has the required security procedures in place, and can prove it. Your cloud data platform vendor must demonstrate it adequately monitors and responds to threats and security incidents and has sufficient incident response procedures in place.


TIP

Industry-standard attestation reports verify cloud vendors use appropriate security controls. For example, the PCI DSS attestation of compliance reveals whether your vendor properly stores and processes credit card information. Make sure your vendors

provide a copy of the entire report for each pertinent standard, not just the cover letters.

In addition to industry-standard technology certifications such as ISO/IEC 27001 and SOC 1/SOC 2 Type II, verify your cloud provider complies with all applicable government and industry regulations. Depending on your business, this might include PCI, the Health Insurance Portability and Accountability Act (HIPAA) and Health Information Trust Alliance (HITRUST), and FedRAMP certifications. Cloud vendors should also supply evidence that third-party software vendors they work with are compliant and that they perform regular security audits.

Extending Access to Third-Party Solutions

Finally, a cloud data platform should make it easy to incorporate and integrate third-party solutions. It should anchor a thriving ecosystem of technology partners to give you options for what you do with your data. To uphold strong security, this integration can be achieved with two types of controls for network security: private connectivity, and via a known range of IP addresses. Accounts in the cloud data platform can be made accessible via private IP addresses, just like internal applications running on the network. The data never traverses the public Internet, which significantly reduces exposure to common security threats.

For example, automated machine learning (AutoML) tools from third-party vendors enable data scientists to select algorithms, tune parameters, and iteratively create and assess machine learning models. The results can be consumed through dashboards, reports, and predictive analytics via connections to other ecosystem partners. Both the raw data and the machine learning results reside in the data platform for easy access.

Security, reliability, and availability are essential attributes of a cloud data platform. They should be "baked in" to the architecture, and inherent in every workload that depends on that platform to make your organization more profitable, responsive, and productive.

GOVERNED AND SECURE DATA SHARING

CASE STUDY

Greenhouse is a marketing agency that leverages data to help clients make digital advertising decisions. With data at the core of its creative, media, and consulting services, the company collects and analyzes billions of interactions from client websites, apps, and advertising campaigns.

Seeking to let clients query these raw data sets, Greenhouse's data science team began exploring modern data sharing solutions. With the onset of the General Data Protection Regulation (GDPR), a legal framework that sets guidelines for the protection of personal data in European Union member states, Greenhouse needed to provide customers with access to their own data and comply with the pertinent regulatory requirements. Previously, sharing this raw data was a costly and tedious process that involved configuring cloud environments, connecting data buckets, copying massive data sets, spinning up compute clusters, and troubleshooting issues.

Today, Greenhouse ingests each client's interaction data into a cloud data platform to create a single source of truth. Clients can view, query, and analyze their own data through a web interface. Continuous data protection, end-to-end encryption, and SOC 2 Type II certification help Greenhouse comply with GDPR regulations.

Greenhouse's data scientists use the data platform to analyze large data sets, collaborate with clients, and share rich interaction insights. Clients can query the interaction data to anticipate behavioral trends, track key performance indicators (KPIs), and maximize the return on their advertising spending — without relying on Greenhouse staff to answer every question. The cloud data platform enables immediate account-to-account sharing of data and simplifies the joining of raw interactions with data from other sources, such as a client's customer relationship management (CRM) system. This makes it easier to collaborate and discover value in the data.

IN THIS CHAPTER

- Empowering users with data
- Eliminating low-level tasks for high-level data professionals

Chapter 9
Maximizing the Efforts of Your Workforce

Organizations in every industry depend on their data to improve efficiencies, reduce costs, and discover new business opportunities. All of today's data-intensive activities are more efficient when they're driven by a unified data platform. All are more cost-effective when you have one usage-based service that allows independent scaling of workloads. All are easier to plan for, track, and execute when the platform uses metering technology to monitor which departments use which resources.

A cloud data platform empowers workers with varying skill levels to maximize their use of data (as explained in Chapter 5). It democratizes access to everyone. Perhaps most importantly, the platform maximizes the efficiency of highly skilled data professionals, so they can spend less time on routine tasks and more time on creating and managing the data-driven workloads that drive the business forward.

Putting Skilled Data Professionals to Work

Software developers, data architects, data scientists, and database administrators make up an elite, much-sought-after workforce that's pivotal to innovation. Unfortunately, with traditional

data platforms and tools, these highly paid workers are often far from productive. In a 2017 article titled "The Cognitive Coder," *InfoWorld* called it the 80/20 data science dilemma: Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data, the magazine reported.

Without a cloud data platform, this problem persists today. With a modern cloud data platform, however, highly paid technology professionals can spend much more of their time being productive with their data, and less time on routine administration and maintenance. Skilled tech workers can focus on accelerating analytics, building data-intensive applications, and creating predictive models to predict churn, identify valuable customers, optimize offers, and more.

CASE STUDY

USING TECH RESOURCES WISELY

Micron Technology manufactures memory and storage systems for a broad range of applications. Previously, the technology company managed a legacy on-premises data warehouse, supported by a large team of IT professionals and database administrators (DBAs) dedicated to supporting servers and databases, scheduling and managing backups, and performing hardware upgrades. According to Jeff Shelman, Micron's director of technical solutions, smart manufacturing, and artificial intelligence, "Management of our on-premises data warehouse was eating up our most valuable IT resources."

In four months, Micron migrated hundreds of applications, approximately 5,000 tables, and 500 terabytes of data into a cloud data platform. The separation of storage from compute resources immediately simplified workload management. The IT team was able to shift 50 percent of its DBAs away from data warehouse support and over to more strategic projects. In addition, the number of downtime events dropped by more than 95 percent.

With a more stable data warehousing platform, Micron's business users can focus on leveraging data to improve the manufacturing process, where a fraction of a percentage of process improvement can translate to millions of dollars in cost savings.

IN THIS CHAPTER

- Moving data warehouses into the fast lane
- Creating more-versatile data lakes
- Streamlining data engineering tasks
- Sharing data
- Developing new data applications
- Fostering the work of data scientists

Chapter 10
How Do You Use a Cloud Data Platform?

A cloud data platform is about maximizing the value of your data. You do this by bringing together modern technologies for storing, sharing, and analyzing that data; for ingesting new types of data; for building new data applications; and for delivering cutting-edge data science projects. This chapter discusses how a modern cloud data platform can enable, automate, and improve these important workloads.

Creating a Modern Data Warehouse

A cloud data platform enables a modern data warehouse with a single source of data for any scale of data, workloads, and users. You can load data in its native forms and use SQL to query both semi-structured and structured data within a single unified system. This allows you to focus on using your data to derive useful

insights rather than on maintaining complex hardware and software. With a modern data warehouse, you can

- Query both structured and semi-structured data using SQL
- Provide consistent service levels to an unlimited number of concurrent data warehouse users and workloads
- Seamlessly and securely share governed, read-only slices of your data across business units within your organization, and with partners and customers — all without copying or moving data
- Use your existing integration tools to ingest data into the data warehouse, and use your favorite business intelligence (BI), analytics, and machine learning solutions to extend its value

Building or Augmenting Data Lakes

You can build or augment an existing data lake by using a cloud data platform as your central repository. Or, you can land your data in a cloud storage solution, such as Amazon S3, Azure Data Lake Storage, or Google Cloud Storage, and use the platform to accelerate data transformations and analytics in your existing data lake. A multi-cluster, shared data architecture will yield dramatically better performance than traditional alternatives. Having a standard SQL interface means you can more efficiently discover value hidden within the data lake and quickly deliver data-driven insights to all your business users.

Streamlining Data Engineering

Traditional data pipelines require complex ETL procedures to extract, transform, and load data. Furthermore, these legacy solutions can't handle and process all types of data, such as structured data from enterprise applications, machine-generated data from IoT systems, streaming data from social media feeds, JSON event data, and weblog data from Internet and mobile apps. A modern cloud data platform allows you to easily and efficiently ingest all these data types using standard SQL (see Figure 10-1). It can support a range of popular data ingestion styles, including batch integration and streaming integration with Apache Kafka.

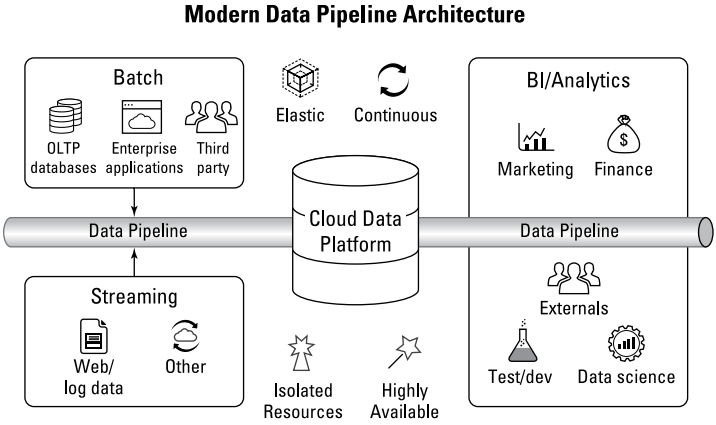

FIGURE 10-1: A modern data pipeline should support any form and velocity of data, for every type of analysis.

Exchanging Data Precisely, Easily, and Securely

Sharing data shouldn't involve copying and moving data. A cloud data platform enables organizations to easily share slices of their data, and receive shared data, in a secure and governed way — without requiring constant data movement or updates to keep data current. Authorized members of a cloud ecosystem can tap into live, read-only versions of the data. Rather than physically transferring data to internal or external consumers, the platform enables read-only access to a governed portion of the live data set via SQL.

This type of advanced data sharing encourages collaboration by making it easier to broadly share your data across business units, with an ecosystem of business partners, and with other external data constituents. It also allows you to monetize your data to create new revenue-generating services.

The cloud data platform you choose should also extend modern data sharing technology to a much bigger scale in the form of a *data exchange,* where customers can acquire third-party data, combine it with their own data, and gain new insights. Because data isn't copied or moved in these scenarios, you eliminate the cost, headache, and delays associated with traditional data exchanges and marketplaces, which deliver only subsets or "slices" of data that must be continually refreshed.

Developing New Data Applications

Software developers need cloud-native development tools to accelerate the process of developing and deploying new applications. A cloud data platform provides unlimited compute and storage resources for development, iteration, testing, and quality assurance (QA) activities. It eliminates the need to build infrastructure and automatically handles provisioning, availability, tuning, data protection, and other operations across multiple clouds. Ideally, the platform should include connections to popular languages, tools, and utilities — including Python, Node.js, Go, .NET, Java, and SQL — and allow users to query structured *and* semi-structured data with standard SQL.

Advancing Data Science Initiatives

Data scientists require massive amounts of data to build and train machine learning models. Finding, retrieving, consolidating, cleaning, and loading training data takes up as much as 80 percent of a data scientist's time. After the data has been properly prepared, machine learning models must be trained and periodically retrained, which requires fresh data to be processed through the cycle.

A modern cloud data platform can satisfy the entire data lifecycle of machine learning, artificial intelligence, and predictive application development. It consolidates data in one central location, and from there it can be imported easily into a wide range of data science notebook and AutoML tools. It also natively supports Scala, R, Java, Python, and other languages. These capabilities empower data scientists to more easily develop and deliver new applications, as well as to train the associated machine learning models.

IN THIS CHAPTER

- Figuring out what you need
- Identifying the data and workloads you want to migrate
- Assessing success criteria
- Comparing solutions and options
- Determining total cost of ownership and the return on your investment

Chapter 11
Five Steps to Getting Started with a Cloud Data Platform

This chapter guides you through five key steps to choosing a cloud data platform for your organization. The process begins with evaluating your data needs and concludes with testing your top choice. By the end, you'll have a plan to help you choose your solution with confidence.

Step 1: Evaluate Your Needs

Consider the nature of your data, the skills and tools already in place, your usage needs, your future plans, and how a data platform can take your business further than you imagined.

Fit with existing skills, tools, and processes: What tools and skills from your team must apply to your cloud data platform? What business processes will it impact, such as order-to-cash and hire-to-retire? Finally, which departments will a cloud data

platform impact, such as marketing, sales, finance, HR, supply chain, transportation, and manufacturing?

Usage: Which users and applications will access the cloud data platform? What types of queries will you run, and by how many users? How much data will users need to access, and how quickly? Which workloads will you run, and how will they vary over time?

Data sharing: Do you plan to securely share data across your organization and with customers and/or partners? Will you allow these consumers to access raw data, or will you enrich that data by adding data analytics services to your data exchange? Will you look for data monetization opportunities?

Global access: Will you store data in a public object store, such as Amazon S3, Azure Blob Storage, or Google Cloud Storage — or in a combination of all three? Do you have specific functional, regional, or data sovereignty requirements? Do you need a cross-cloud architecture to maximize deployment options, bolster disaster recovery, or to ensure global business continuity?

Resources: How much do you want to invest to monitor and manage availability, performance, and security? Do you want to pay professionals who have experience with database management, data engineering, and application development, along with a DevOps team to streamline these activities? Or do you need a data platform that removes those burdens?

Step 2: Migrate or Start Fresh

Every cloud data platform project starts with assessing how much of your existing environment should migrate to the new system.

Is this a brand-new project? If so, it often makes sense to design the project to take full advantage of the capabilities of a cloud data platform rather than carry forward an existing implementation that may have architectural constraints, as discussed in Chapters 4 and 5.

Which parts of your current data platform and dependent systems cause the most problems? A well-planned migration could focus on moving the most problematic workloads to the cloud data platform first. Or, you may want to migrate easier, more straightforward workloads to get quick wins.

What constraints in your current systems will a cloud data platform eliminate? The right cloud data platform should eliminate the need for tools and processes designed to work around resource constraints, the disruptive effort required to add capacity, and having to optimize costs.

Will your existing applications work with the new platform? Business intelligence, data visualization, and other analytic tools should be easy to adapt to the new architecture.

How are your data and analytics requirements likely to change in the future? A cloud data platform is designed to evolve to accommodate new data, technologies, and capabilities as they emerge. This evolution will likely reveal new opportunities to capitalize on advanced capabilities, such as IoT, machine learning, and AI.

Step 3: Establish Success Criteria

How will you measure the success of the new cloud data platform initiative? Identify the most important business and technical requirements, with a focus on performance, concurrency, simplicity, and total cost of ownership (TCO).

If your new cloud data platform has capabilities not available in your previous system, and those capabilities are relevant to evaluating the business and technical success of your new solution, be sure to include them. As you establish your new solution's success criteria, determine how to gauge success, including which criteria are quantifiable and how to measure them, and which criteria are qualitative and how to assess them.

Step 4: Evaluate Solutions

Make sure your choice meets these architectural criteria:

- Natively integrates structured and semi-structured data, stores it all in one place, and avoids creating data silos
- Streamlines the data pipeline so new data is available for analysis in the shortest possible amount of time
- Dedicates isolated resources to each workload

- Shares data without having to copy or move live data, and easily connects data providers and consumers
- Replicates databases and keeps them synchronized across accounts, cloud platforms, and regions to improve business continuity and streamline expansion
- Scales compute and storage independently and automatically, and scales concurrency without slowing performance

Step 5: Calculate TCO and ROI

If you choose a cloud data platform based on price, consider the TCO for a conventional data platform: the cost of licensing, typically based on the number of users; hardware (servers, storage devices, networking); data center (office space, electricity, administration, maintenance, and ongoing management); data security (password protection and encryption); solutions to ensure availability and resiliency; support for scaling and concurrency; and creation of development and staging environments.

For some solutions, you might need to consider additional costs, such as building and managing multiple data marts, having multiple copies of data in different data marts, training people, and having multiple systems (for example, SQL and NoSQL) to handle diverse data.

Calculating the costs of cloud data platform options is usually easier, but it varies according to the vendor's services. Assuming you outsource everything to the vendor by choosing a data-platform-as-a-service offering, you can calculate the TCO based on the expected usage fees. If you opt to use an external object store from one of the big cloud vendors, you need to add the costs of that vendor's services as well.

Organizations typically calculate the return on investment (ROI) over the expected lifetime of the data platform, which is commonly one to three years. A key caveat: People often assume a cloud system runs 24/7 and at high capacity, overlooking the savings possible when it is scaled up and down dynamically in response to changing demand and only charges by the second.

Unite all your data, insights, and users with a cloud data platform

In just six years, the number of software-as-a-service workloads globally has ballooned to more than 400 million. The benefits of cloud over on-premises computing are clear, but an age-old problem persists. All of those workloads, and an organization's vast landscape of internal solutions, generate individual silos of data. They are nearly impossible to unite inside a single solution, preventing organizations from obtaining the deepest data-driven insights possible. This book reveals how organizations of any size can easily and securely store, integrate, analyze, and share their data, across multiple clouds and regions, for a business without barriers.

Inside...

- Why the cloud data platform emerged
- How to select a cloud data platform
- Why a modern cloud architecture matters
- What data workloads you can tackle
- What business benefits will emerge
- How to get started with a modern cloud data platform
- Real-world case studies

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

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

ISBN: 978-1-119-71389-0
Not For Resale

for dummies®
A Wiley Brand

Also available as an e-book

WILEY END USER LICENSE AGREEMENT

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