O'REILLY®
Business Guide

The Rise of Logical Data Management

An Essential Data Strategy for Transforming Your Business in the Age of AI

Christopher Gardner

denodo

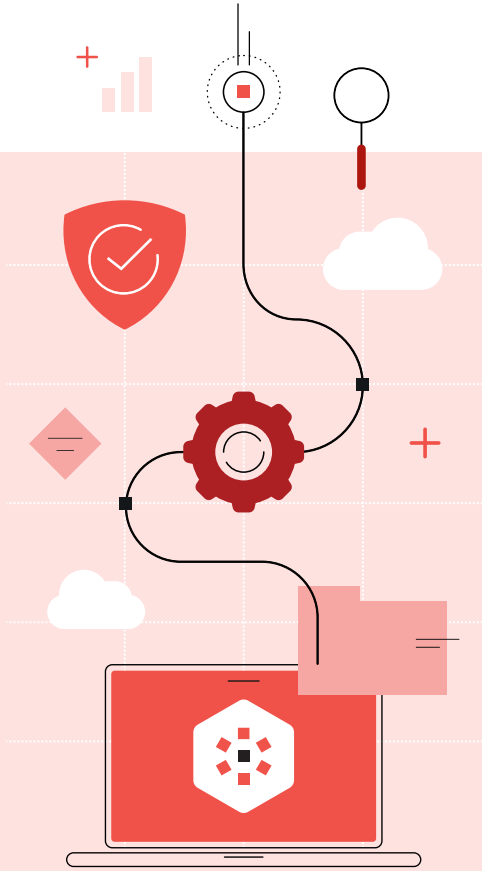
LOGICAL DATA MANAGEMENT

Deliver Data in the Language and Speed of Business, in this Age of AI and Increased Data Complexity

The Denodo Platform enables the rapid delivery of data products through a unified business data layer with centralized security and governance.

Learn why leading companies have voted us a Gartner® Peer Insights Customers' Choice four years in a row, with an average rating of 4.6 out of 5, as of 2024.

GET STARTED



Agility is only the beginning...

The Denodo Platform is smart, flexible, powerful, and affordable, to support myriad modern use cases.

Compared with traditional data management platforms, it enables:

- **65%** faster data delivery
- **67%** less data preparation
- **83%** faster time-to-value

Source: *Independent third-party analyst firm in a 2021 report on the impact of data virtualization, a key component in logical data management.*

The Rise of Logical Data Management

An Essential Data Strategy for Transforming Your Business in the Age of AI

Christopher Gardner

O'REILLY®

The Rise of Logical Data Management
by Christopher Gardner

Published by O'Reilly Media, Inc., 141 Stony Circle, Suite 195, Santa Rosa, CA 95401.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Andy Kwan
Development Editor: Gary O'Brien
Production Editor: Kristen Brown
Copyeditor: J.M. Olejarz
Proofreader: Sharon Wilkey

Cover Designer: Susan Thompson
Cover Illustrator: Ellie Volckhausen
Interior Designer: David Futato
Interior Illustrator: Kate Dullea

August 2025: First Edition

Revision History for the First Edition
2025-08-21: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781098176044* for release details.

This work is part of a collaboration between O'Reilly and Denodo. See our statement of editorial independence.

978-1-098-17601-3

[LSI]

Table of Contents

Preface

Data management has been around as long as we have had data. To the layperson it sounds easy: you have an observation, and you record it. But as anybody who has worked with data can attest, managing it is often fraught with complexities. Figuring out how to structure data so it can be found easily and quickly, as well as keeping it accurate, is hard in itself. On top of that, the same data needs to serve the disparate needs of different users.

Whether you are navigating the complexities of today's data landscapes or have largely simplified your landscape with a modern data lakehouse solution, you may encounter challenges delivering the right data, formatted in the right way and at the right time, to meet dynamic business needs. You might find that a data lakehouse, on its own, is unable to provide business users with self-service data access, or that it cannot deliver real-time, governed data to artificial intelligence (AI) applications, which is necessary for their success.

Who This Book Is For

This book is for business leaders and senior technologists who are charting a course through today's complex data ecosystems. It provides an in-depth look at logical data management, which can operate in tandem with modern data lakehouse solutions, augmenting their capabilities.

This book sheds light on the inner workings of logical data management and demonstrates its strategic value across various businesses. Whether the objectives involve enhancing AI capabilities; promoting data democratization; transforming customer experiences; or

strengthening governance, risk, and compliance initiatives, data management has proven successful for numerous companies. As you navigate through the upcoming pages, you'll find ample examples that illuminate the realm of possibilities and guide you toward achieving powerful results.

What You Will Learn

As you explore this book, you'll gain insights into logical data management approaches that will equip you with the knowledge to transform your organization's data strategy. Here's what you can expect to learn:

- The essence of logical data management and its transformative effect on data management
- How logical data management platforms can augment modern data platforms like cloud data warehouses and data lakehouses with critical capabilities such as self-service data access and AI-ready data
- Traditional data management's flaws and the imperative for change
- The foundational principles and operations of logical data management
- How logical data management works with agentic AI to deliver your company's data in real time
- How to broaden data access and fortify organizational decision-making
- Insights to innovate customer interactions
- Logical data management's role in redefining governance, risk, and compliance
- Tactics to elevate operational efficiency and agility
- Logical data management's contribution to IT infrastructure advancement
- Future developments and trends in data management

By the end of this guide, you will have the tools not just to manage data, but to use it as a driving force for innovation and progress. Let's get started.

O'Reilly Online Learning

For more than 40 years, *O'Reilly Media* has provided technology and business training, knowledge, and insight to help companies succeed.

Our unique network of experts and innovators share their knowledge and expertise through books, articles, and our online learning platform. O'Reilly's online learning platform gives you on-demand access to live training courses, in-depth learning paths, interactive coding environments, and a vast collection of text and video from O'Reilly and 200+ other publishers. For more information, visit *https://oreilly.com*.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
141 Stony Circle, Suite 195
Santa Rosa, CA 95401
800-889-8969 (in the United States or Canada)
707-827-7019 (international or local)
707-829-0104 (fax)
support@oreilly.com
https://oreilly.com/about/contact.html

We have a web page for this book, where we list errata and any additional information. You can access this page at *https://oreil.ly/riseLDM*.

For news and information about our books and courses, visit *https://oreilly.com*.

Find us on LinkedIn: *https://linkedin.com/company/oreilly-media*.

Watch us on YouTube: *https://youtube.com/oreillymedia*.

Acknowledgments

I would like to thank the tech reviewers for their valuable feedback: Bhargavi Reddy, Anant Kumar, and Aarohi Tripathi.

CHAPTER 1
The Challenges of Managing Data

In the whirlwind of the 2020s, marked by unprecedented challenges and technological leaps, the quest to master data remains more critical than ever. With the rise of transformative technologies like generative AI, the urgency for organizations to effectively manage and utilize data has intensified. Data, and lots of it, is at the heart of everything. Yet despite modern advances, distributed data continues to be a major challenge. The ability to provide actionable insights to decision makers, crucial for navigating complex scenarios and driving success, hinges on the proficiency of data management.

Enter logical data management—a strategic approach designed to meet this need head-on, guiding organizations in the art of transforming data into a language that resonates with its users swiftly and accurately. It's not just about managing data; it's about empowering data to be a decisive tool for achieving customer satisfaction, operational excellence, and innovative breakthroughs. And it's not about replacing solutions like cloud platforms, cloud data warehouses, or data lakehouses; it's about augmenting their capabilities in powerful new ways.

But before we dive into logical data management, let's review the challenges of managing data in today's landscape through the eyes of data analysts, who need to gather and translate it into actionable insights for your business. This data can come from anywhere in the organization and occasionally from outside as well. Analysts' challenge is to get access to that data, understand it, analyze it, and deliver the results to leaders at all levels of the organization. This

requires them to be skilled not only in data analytics and reporting, but also in identifying the right data sources, understanding the key questions the executive team wants answered, and working with data to provide those answers.

The Growing Distribution of Data

To be successful, you want data from multiple sources. This is not just the data your own organization creates during the normal process of business. This is also data from beyond the scope of business function:

Additional internal data
: The data within your organization comes from many distributed sources. This information is created at different times, by different teams and systems, and optimized for different purposes. Data can come from business processes, machinery, customers, and more. How do you successfully distribute these data sets, combine them, and optimize them for use at an organization-wide level without sacrificing the domain-specific demands?

Data from related business processes
: Some data comes from sources related to, but not part of, the business. They can include suppliers, marketing campaigns, and product distribution providers. These sources are not part of the main business process per se, but the information they provide is no less valuable. They can open up opportunities for saving on costs, improving logistics, and avoiding potential obstacles that may impact your business.

Data from external sources
: Some data might seem unrelated to the business but still have an impact. Information about road construction, weather, customer demographics, social media, and more all have impacts on the success of a business. For example, a natural disaster might disrupt suppliers within certain regions of the country. These external sources are a primary source for analysis by data scientists and AI.

All this data is not easily accessible. It comes in multiple formats and multiple structures. Companies need a way to consolidate it all to make it usable. The information needs to be collected, organized,

and stored for reporting purposes. It might also need to be combined with business-related data to derive its full impact. All of this takes time, storage space, processing power, and the employee resources to manage it all.

Data science has also influenced the evolution of data management. Data scientists look across many domains to extract patterns, correlations, and trends. They derive predictive and prescriptive plans that help the company make decisions looking ahead, instead of reporting on what happened in the past.

The analysis that data scientists perform takes large volumes of data from a wide variety of sources both in and outside the organization. The variety creates a challenge for the business, as these data sources are likely in different formats and structures. To make them accessible and usable for the data scientists, the data needs to be translated into a common format. It needs to be massaged and manipulated to enable it for use in analysis. This is especially true when implementing AI applications, which have stricter requirements for data access, data integrity, data cleanliness, and often require data to be delivered in close to real time.

To complicate things further, data is no longer limited to a structured format. A *structured data set* is one that can be categorized and organized into a table. Its information fits neatly into rows and columns and uses discrete data types to hold its values. This is what data warehouses are built from. The information is organized in a way that can be easily pulled, categorized, and measured.

The alternative to the structured format is unstructured data. *Unstructured data* doesn't follow a consistent format and usually requires much more work to extract information from. Examples include images, videos, audio, web pages, survey data, and more. To utilize unstructured data, businesses need a place to store it and methods to extract its values, such as natural language processing, machine learning, and AI. This need unfortunately leads to a whole new set of challenges.

Transitioning from the Traditional Data Flow

How do you translate data into decisions, given such a wide variety of sources? For years, organizations relied on extracting and loading it into a data warehouse. In this data flow, the divisions within

the company produce information using transactional systems. The data is usually structured for the purpose of running an application and not for reporting.

As a result, most organizations extract the data out of the transactional database and place it into a reporting database. Through a process called ETL (extract, transform, load), data is drawn from the transactional systems and translated into a structure that better reflects business process reporting needs. From there, the data is loaded into a reporting database, where it becomes available for analysts to develop reports with.

Data Warehouses

Data warehouses contain data from all over the organization, as shown in Figure 1-1, but the information is often not detailed enough for specific reporting needs. When this happens, a smaller subset database is developed called a *data mart*. The data mart contains a focused set of data specific to a functional section or line of business within the organization. This detailed data allows analysts to develop reports at a more granular level.

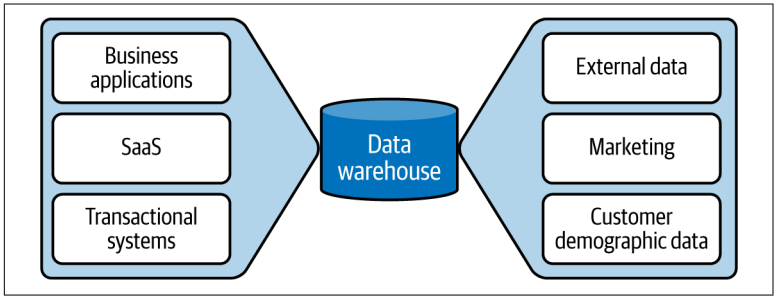


Figure 1-1. An example of how data can be collected in a data warehouse. The business collects as much information as possible and loads it into a central location, which creates a group of schemas, tables, and data for analysis.

Data warehouses and data marts provide a basis upon which analysts can build reports. Until recently, those reports were one of the few ways executive leadership could gain understanding into a company's performance. The landscape is changing, however. Data science, AI, and machine learning have transformed the way organizations look at information. They've also exposed the challenges related to traditional data management.

Data warehouses have done a great service to the business community, integrating information from online transaction processing (OLTP) databases and consolidating them into analytical data in an online analytical processing (OLAP) database. But one of the downsides of data warehouses is that they often require substantial storage and computing resources in order to scale with the needs of a large organization. If a data warehouse grows and exceeds available storage and compute capacity, the administrator has to procure more, resulting in escalating costs.

Another issue is that the queries that business analyst users write against an OLAP database are very different from the ones sent to an OLTP database. They are usually complex and expensive to run. Data warehouse administrators have to work hard to optimize the tables and make compromises by denormalizing the data, and creating indices that speed up queries but slow down writing data to the data warehouse. In addition, it's often difficult, if not impossible, to get all relevant business data into a warehouse. This means that the warehouse is not a complete solution.

We also have to consider the cumbersome and time-consuming expense of ETL jobs. As analysts and business users demand more real-time data, there is urgency to increase the frequency of loading data into the warehouse. Bottlenecks in processing time and bandwidth are encountered, and more slowdown occurs in reading and writing to the data warehouse.

Another point to consider is that with many business users having different needs, it is easy to encounter situations where the same data has to be copied and modified into many different formats for various use cases. Data marts can add complexity, processing time, and expense to serve all these end users. Conversely, with ETL it can also be difficult to reuse transformation logic across multiple pipelines, so common transformation logic ends up repeated multiple times and maintenance becomes very hard.

Finally, there are the issues of access and permissions. The ETL model creates a structure where data exists in multiple locations. In addition to the large amounts of storage space required to maintain these separate locations, security administrators also have to manage access and permissions to both as well. Since multiple domains within the business can administer their data in various ways, it's

challenging to develop a central governance strategy to manage those separate locations in addition to the central data warehouse.

The data landscape has been moving beyond the capabilities of the traditional data warehouse, as new technology and practices have been emerging in the marketplace with relation to data. Technologies such as AI and machine learning are opening avenues to new insights. The data science field is exploring how statistics can create predictive and prescriptive models for future business decisions. Data formats are also expanding from structured (in columns and rows) to unstructured (like audio, video, and images).

How can you take advantage of these new innovations? How does your current data pipeline support or prohibit utilizing these resources? Most importantly, how can your company overcome existing obstacles to take advantage of the opportunities these data sets provide?

Data Lakes

For unstructured data, many organizations turn to *data lakes*. These are storage structures that maintain information in its raw form. Data lakes can hold both structured and unstructured data, as there is no requirement to modify its format.

They also alter the way data is integrated and stored by extracting and loading it without transforming it until it's needed. This transfers the resources needed to transform the data from the extraction tool, and places it on the tool that's utilizing it for analysis, as seen in Figure 1-2. Data lakes can also be the source for an ETL process that extracts the structured data from the lake, transforms it, and loads it for use in a data warehouse.

It is probably no surprise that data lakes' greatest strength, their flexibility, is also their weakness. By making data lakes a free-for-all where any and all information can be uploaded in any format, things can become chaotic very quickly. Data lakes typically have a *schema on read* strategy, meaning that there is no enforcement for the data to follow a certain format or shape (unlike relational databases, which have strict table structures and rules). This is why if data governance is not prioritized, a data lake quickly becomes at risk of becoming a "data swamp."

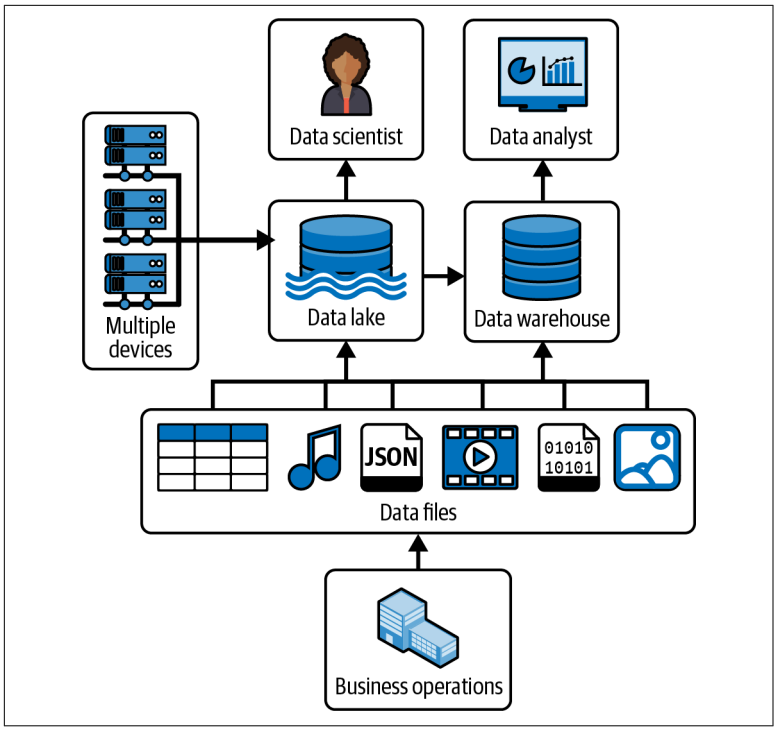


Figure 1-2. The data lake serves as a location for both structured and unstructured data. Data in the data lake works together with a data warehouse to meet reporting needs.

Data lakes also put an emphasis on ingesting data in its raw formats straight from its sources, which helps keep data timely. This has a downside of those data sets not being processed to be user-friendly. Data lakes often are the domains of power users like data scientists and data engineers, who have technical skills beyond just SQL. Data lakes also are not performant for analytical queries. This inevitably created some tension with business end users who are proficient in SQL but are not engineers or programmers, and can add processes and work that are expensive. For this reason, a data warehouse is still needed to receive cleaned and transformed data from a data lake for business end users. A strong argument can be made that this marginalizes the value of a data lake, bringing all the problems of data warehouses. I will talk about how the data lakehouse attempts to address this shortcoming later.

Pulling back, you can look at a data lake as the solution to data silos existing across an organization but still not addressing the problems of data warehouses and ease of use for analytics users. A data lake can work great for certain environments, technically proficient users who prioritize flexibility and scalability above all else. This is why data warehouses (and many alternatives like NoSQL) continue to be used. So let's talk about the next logical step, absorbing the data warehouse function into the data lake.

Data Lakehouses

In some cases, businesses may take a hybrid approach to data, storing it in a *data lakehouse*. In a lakehouse, data is stored in both a structured and unstructured format. This is illustrated in Figure 1-3.

The structured data is utilized by the business intelligence side of the company for analysis and reporting. The unstructured data is utilized by data scientists through AI and machine learning. Some organizations use natural language or other AI tools to analyze the contents of the unstructured data in order to apply a structure or categorization to its contents.

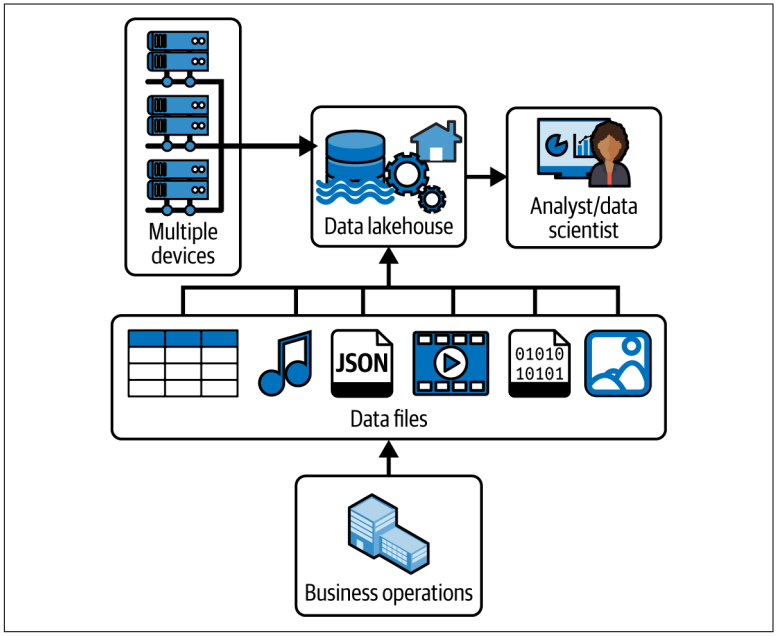


Figure 1-3. A data lakehouse serves as a hybrid version of a data lake and a data warehouse.

Open table formats such as Delta Lake and Apache Iceberg add data warehousing functions like transactionality, schema evolution, and incremental updates to data lakes, which enable faster, reliable analytics across diverse data sets. In addition, these formats are supported by several lakehouse processing engines, providing some protection against vendor lock-in.

Using data lakehouses poses several problems. First, like data warehouses, data lakes can never be an organization's sole repository. Some on-premises data may have to remain on premises, for compliance with data privacy and other regulations. Also, some organizations will want to foster and maintain a multicloud infrastructure, to leverage the best features of different cloud platforms. In addition, transferring all the data from a cloud platform into a data lakehouse can be extremely costly because of the cloud platform's (often quite high) ingress and egress charges and the additional overhead of governing and securing multiple copies of the same data.

Second, data lakehouses also still require time and effort to unify the data. Businesses develop data across multiple domains, and it's no small task to combine that data and make it available for analysis. All the challenges associated with creating, maintaining, and reusing data transformation pipelines in the data warehouse still hold in the lakehouse. Lakehouses also have the risk of creating bottlenecks that slow down provisioning data. Comparatively, a logical data layer provides a more agile alternative based on a semantic layer, which minimizes data replication, fosters reusability, and enables self-service.

Third, data lakehouses still rely on batch processes, meaning the data they provide lags behind. This makes them less desirable for AI applications that require more real-time data.

Finally, while data in open table formats can be processed by multiple lakehouse engines, users and consuming applications are still dependent on the access technologies and query dialects of their vendor. Thus vendor lock-in is a significant risk. However, transitioning to a new system may force costly changes in the consuming applications. In turn, as we will see, logical data management completely abstracts data consumers from changes in the underlying infrastructure.

The traditional approaches for data within organizations involve full centralization. Their goal is to bring all the data from the individual domains together into a single repository, such as a warehouse, lake, or lakehouse. These approaches work differently, but they benefit the organization by centralizing the data into a single location, though they also suffer from many limitations. These types of systems lock a company into a single vendor while also creating bottlenecks in the delivery of the data, and limiting access to data in real time.

While data lakehouses aim to unify structured and unstructured data for broader analytics, they fall short when it comes to powering today's AI and generative AI (GenAI) workloads. These workloads require not only high data volume, but also low-latency, contextually rich, and frequently updated data pipelines. Lakehouses rely on batch-oriented processing and delayed transformation, which impedes the real-time responsiveness and agility needed by generative models and agentic AI systems that must reason over fast-changing business conditions. Without a semantic layer to make data comprehensible and consistent across sources, and without consistent access to real-time updated data, AI initiatives are often plagued by unreliable results, high preparation overhead, and lack of trust from business stakeholders.

Similarly, organizations aiming to scale self-service and data product development often discover that lakehouses, while centralizing data, do not make it meaningfully accessible to downstream business users. Data remains locked behind technical barriers, such as unfamiliar schemas, lack of business-friendly metadata, and performance trade-offs. Teams trying to build reusable, governed data products across business domains struggle with the rigidity of lakehouse schemas and the overhead of replicating and transforming data for each new use case. This often leads to duplicated efforts, shadow IT, and slow delivery. Logical data management overcomes these constraints by enabling a consistent semantic layer and real-time integration without requiring up-front data centralization, making true self-service and scalable data products possible.

Data for Everyone

In addition to the challenges of integrating data, organizations have a broader challenge in making data accessible to a wider range of users. Data democratization is about providing access and analytic

resources to everyone at all levels of the organization, to empower them to make informed decisions. When you empower employees with data, you enable them to evaluate the processes and procedures more specific to them. This encourages them to identify inefficiencies and challenges within their own jobs and instigate change to make improvements.

To enable data democratization, you need data that is specific to each user's line of business. You need data that is easy for each user to understand, easy to access, and easy to draw information from. A data warehouse is not likely to contain such granular data, and it is also not likely to be easy to read and understand for an average employee. Data within an organization is distributed across multiple domains and in multiple formats and structures, and even with a lakehouse, some data is not easily connected to.

A typical employee requires data created by and related to their position, but they also benefit from data elsewhere in the organization. For example, a sales employee might need information about overall supply stock, competitor prices, or upcoming marketing efforts. You want to provide your employees with data they can use to make decisions on their own, and that data can come from a wide variety of places:

Data from the organization
: Employees might need access to the business processes of the organization. For example, an employee working in a sales role may want to use marketing data to derive insights on how to organize their storefront.

Data from the line of business
: The employee generates data as part of their individual task, and often this is the data most relevant to them for analysis. This data is also the most time-sensitive, as it reflects the day-to-day activities of the employee in their job.

Data from outside the organization
: Insights can be drawn from beyond the organization itself. Information on weather, suppliers, and more can directly influence business decisions at the employee level.

Employees can benefit from a wide variety of data in their jobs, but how do you make this data available to them? More importantly, how do you structure that data in a way that is easy to understand

and utilize? Ideally you want a solution that is self-service and requires as little maintenance as possible, while also being complete and detailed enough to provide the insights they need.

Conclusion

Demand for data at all levels of your organization, combined with larger, more varied data sets, requires a different approach to data management. Data warehouses lack the detail and accessible format needed for democratization. Comparatively, data lakes and data lakehouses demand additional features. Furthermore, the business data environment is complex, leading to poor-quality information that is difficult to utilize.

What options does a company have to meet today's demands? In the next chapter, we'll introduce you to logical data management and how it can radically transform a data infrastructure without having to replace any existing solutions.

CHAPTER 2

Understanding Logical Data Management

Now that we've covered some of the challenges with managing data, the imperative is clear: organizations need a strategy that empowers them to wield data with confidence, providing governed, secure, and timely information for business imperatives. Such a strategy must encompass a comprehensive framework, detailing the requisite technology, processes, personnel, and policies to sustainably manage an entity's informational assets. The strategy should bring to life the organization's vision for data acquisition, preservation, sharing, and application.

The resulting infrastructure from this strategic blueprint should refine data management at every juncture, ensuring accessibility for all involved parties. The aim is to arm the business with the data it needs to secure a competitive edge and chart a course for success. This endeavor calls for a deep dive into logical data management, a concept that will be defined and explored in the following section.

Defining Logical Data Management

Logical data management is a data management approach based on logical or virtual connections to data. This approach is applied to an organization's existing data structure to provide more uniform control over security, create a central data marketplace, and unify disparate underlying systems into a common format. It is a transformative approach to data handling, moving beyond the conventional

constraints of physical data consolidation and adopting a dynamic, virtualized strategy. This approach enhances the integration and accessibility of information, making it a more flexible and efficient resource for the organization. While ETL and data lakes move data to a central location, logical data management provides improved data quality and accessibility, as seen in Table 2-1.

Table 2-1. A high-level comparison of data management approaches

	ETL	Data lakes	Logical data management
Focus	Data integration and preparation for analytics	Storage of raw structured and unstructured data	Central, defined data elements in a format that is accessible and understandable
Purpose	Moves data from source system to central repository after cleaning and transformation	Central location for structured and unstructured data, especially for advanced analytics and machine learning	Governed data from all domains of a business, facilitating data democratization
Benefits	Clean structured data for reporting and analysis	Large storage of varying types of raw data that is flexible for exploration	Improved data quality, easier access to data, and less siloed data sources

A key technology in logical data management is *data virtualization*, which allows users to access multiple sources of data as if they are a single source. Data virtualization creates a common data access layer on top of distributed data sources, making it faster and easier to combine information from different sources. This technology improves accessibility, allowing data consumers and business decision-makers to access the information they need when they need it, expressed in the language of the business, without being restricted by physical limitations. This approach also fosters agility and consistency, as the data products delivered by the data layer are based on reusable components representing common business entities.

In essence, data virtualization is about leveraging technology to make data more flexible, integrated, and accessible, thereby empowering organizations to make more informed decisions. It's a move toward viewing data not just as a static asset, but as a dynamic resource that can drive business growth and innovation.

At its core lies the principle of abstraction, which separates the data layer from the confines of physical storage mechanisms. This

separation is crucial, as it grants businesses the agility to access data from a multitude of sources, irrespective of their location—whether they're cloud-based systems, on-premises databases, or a fusion of both. This separation also allows business data needs to evolve independently of the changes in the underlying data infrastructure provoked by cloud transition and IT modernization initiatives. The implementation of a logical data layer is akin to removing barriers within an organization, dismantling the silos that have long hindered a unified view of the information landscape, and fostering a collaborative environment where data managers and consumers can work in concert. A logical data layer can be implemented above any existing data infrastructure, including cloud data warehouses and data lakehouses.

The strategic value of logical data management is significant. It propels organizations into a future of rapid decision making, efficient operations, constant innovation, and prioritized customer satisfaction. For example, a multinational corporation can leverage this approach to integrate customer data from various touchpoints, enhancing service delivery and marketing strategies. Similarly, financial institutions can simplify governance and stay ahead of regulatory changes by centralizing policy management in the common logical data layer for all data sources and consumers. In the next section, you'll learn about more benefits that logical data management brings to the table.

The Benefits of Logical Data Management

As a leader in your organization, your days are filled with critical decisions that hinge on the accuracy and availability of data. Traditional data management systems might leave you waiting for IT to consolidate reports from various departments, a process fraught with delays and potential inaccuracies. Logical data management changes this narrative by enabling a streamlined, real-time view of your company's data landscape that is not conditioned by data infrastructure and that guarantees consistency by design.

It gives you a bird's-eye view of operations, finance, customer interactions, and market trends, all through a single pane of glass. This unified view is not just a convenience; it is a game changer. For instance, when evaluating the performance of a new product line, logical data management enables you to instantly pull together sales

figures, production costs, and customer feedback, painting a complete picture of the product's journey from conception to market reception.

Here are the key benefits of logical data management, with each one illustrated by an example so you can envision how it might apply to your own context:

Harmonization of data assets
: Enables swift location and utilization of information from across many disparate data sources, allowing, for example, a customer service representative to access a unified view of customer data from customer relationship management (CRM) systems, sales platforms, and support channels, leading to more efficient and personalized customer interactions.

Enhanced security and compliance
: Provides a common data layer to enforce uniform policies across data platforms, ensuring data protection and compliance with regulations like General Data Protection Regulation (GDPR) or Sarbanes-Oxley Act (SOX), thereby instilling trust in health-care professionals regarding the protection of patient records.

Fostering collaboration
: By exposing data in the language of the business, logical data management bridges the gap between data managers and users, such as by allowing marketing teams to work seamlessly with IT to refine customer segmentation models, thereby boosting the effectiveness of marketing campaigns.

Process improvement
: The common data layer reveals opportunities for process improvement and promotes data reuse and sharing—for example, enabling a supply chain manager on an ecommerce platform to use insights from it to optimize inventory management systems.

Comprehensive data perspectives
: Provides comprehensive data perspectives by integrating data holistically, allowing an operations manager in an energy company to merge sensor data with operational and market data for a unified predictive maintenance view, for instance.

Operational efficiency
: Boosts operational efficiency by optimizing data management tasks and automating processes, such as route optimization for logistics managers, saving time and resources. A 2021 report published by Forrester indicates that a data virtualization solution reduces time to delivery of data over traditional ETL processes by 60%–80%.

Improved data accessibility and literacy
: Improves data accessibility and literacy, fostering transparency and civic engagement, and enabling city planners to make public data more accessible to citizens.

Driving innovation
: Drives innovation by empowering the business with the data people need in formats they can use, and by feeding AI and machine learning algorithms with diverse, quality data, unlocking new possibilities and competitive advantages, including tailoring customer interactions in real time and adapting quickly to market trends.

Agnostic to underlying data sources
: The logical data layer is completely agnostic to the underlying data sources that populate it. This is beneficial when organizations upgrade, change, or otherwise modify existing data sources. When a system is modernized from on premises to cloud or updated in some other way, the logical layer can easily be adjusted with little to no disruption to business. This is much more complicated, time-consuming, and costly with traditional data management methods.

The Benefits: An Architectural View

A logical data layer reads data from a wide variety of sources and adds transformations that make them appear as if they were a single source. Logical data layers are not restricted by the confines of the physical layer formats. You can connect to diverse data types and combine disparate data sources together to create relationships that otherwise would not be possible at a physical level (see Figure 2-1).

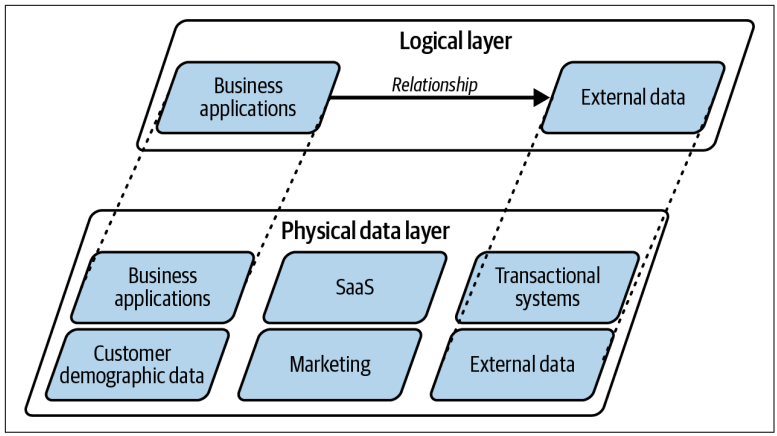


Figure 2-1. A logical data layer sits atop the physical data sources and provides a lens that translates the data. This allows disparate data sources to be combined with user-defined relationships without always having to replicate data.

The data itself still resides in its original location, meaning it does not always need to be replicated as part of the process (it *can* be, but it need not be). You can think of a logical data layer as a library, where you have access to the knowledge it contains without having to obtain your own copy of every book. The logical data layer pulls only the necessary fields from the variety of physical data sources and uses its own language to create new fields, tables, and relationships between the sources. Since the data does not need to be physically copied, the amount of resources required to maintain the layer is much smaller.

The logical data layer simplifies life for your analysts, as all the data is now available using a single model and query language. This means they no longer need to have knowledge of all the languages and technologies used in the data sources of your company. Data stored in databases, web services, software-as-a-service (SaaS) applications, data warehouses, data lakes, or any other system is accessible in the same way through the logical layer. The logical layer is in charge of automatically translating user queries to the underlying languages and technologies used by the various sources. It can also potentially offer several query languages and access technologies beyond SQL, such as REST, GraphQL, or OData to facilitate compatibility with multiple consuming applications.

Additionally, the logical data layer is not restricted to the nomenclature of the physical layer. This means your developers can rename tables and fields to common names used throughout the business. This makes the layer much easier to use and accessible to your employees who have less experience dealing with the data tables. The logical layer also provides a way for your business to connect to process-specific data sets, enabling your less technical employees to utilize data to make decisions closer to their line of work.

The data is also easily reusable. A business that uses logical data management can set up virtual views with key information related to the core functions of the business. These can be used to create domain-specific data products. For example, a business can create virtual data views for key business entities such as customers, products, or sales. These can then be used to generate reports using consistent data tailored to business domains, such as marketing, sales, or regional locations. In turn, in physical data architectures, preparing data for consumption requires creating new data copies for each use case. This complicates maintenance and can create a governance nightmare, especially with sensitive information.

The logical data layer is extremely flexible and makes it easy to add additional data sets. There is no need to copy information from one location to another when a new system is added. The logical data layer can easily be configured to point at the new source and access the needed data. This includes process-specific data sets, such as machine usage, points of sale, and CRMs.

Since there is no need to copy data, access to it can be closer to real time. For example, if your business has monitoring software that tracks production, you want that data to be as close to real time as possible. A logical data layer can sit atop the production data and expose it for use without the need to copy it to a reporting database. This ensures that the employees responsible for maintaining that production line have access to that data and can make faster decisions when issues arise. Chapter 9 covers additional benefits related to logical data management in more detail.

As new source systems arise and current systems update, the tables and fields reflected in the logical data layer do not need to change. The nomenclature and standards of the logical data layer adapt to these changes, keeping a consistent and familiar view for your users. This also reduces the need to re-create or modify reports as

the systems and underlying sources are modified—especially when transitioning your business to the cloud.

For your users, a logical data layer improves visibility into data lineage. One question that often arises with reporting is "Where did this data come from?" For traditional data management systems, this question can sometimes be challenging to answer. When the information is pulled from production systems, transformed, and loaded elsewhere, the source field and tables can get lost. Tracing a field in a report back through the reporting and ETL changes can be anywhere from challenging to impossible for end users. A logical data layer simplifies this by simply providing a lens into the data instead of copying it.

This clear view into how the data moves also facilitates better data definitions, improves data governance, and simplifies security. Instead of managing security for multiple data sources, administrators can focus on managing access to the logical data layer. This removes the need for access expertise and management in multiple systems, requiring the administrators to focus instead on the single logical data layer.

Logical data management transcends traditional data management, propelling organizations into a future-ready, data-empowered state. By understanding and leveraging its strategic benefits, you can harness data more effectively, drive innovation, and secure a competitive edge in the marketplace. The next section covers how logical data management can enhance various leadership roles.

Empowering Leadership Roles with Logical Data Management

Let's explore how logical data management can empower you to make better decisions by looking at a few more examples through the lens of various leadership roles:

IT leaders: Navigating the digital maze
: According to the 2024 Gartner CIO and Technology Executive Survey, budget constraints loom large, making innovation a delicate balance with efficiency. Consider an infrastructure leader responsible for modernizing the organization's IT infrastructure. Logical data management becomes their compass: they integrate data on infrastructure costs, performance metrics, and

resource utilization. By identifying inefficiencies and reallocating resources strategically, they optimize the technology budget. With seamless data access across legacy systems, cloud platforms, and emerging technologies, they can make data-driven decisions during digital transformation initiatives. For instance, they might enable real-time analytics for customer insights or streamline DevOps processes. Moreover, with logical data management, the infrastructure leader unifies data silos, integrates on-premises and cloud environments, and enables agility (e.g., they might orchestrate hybrid solutions, ensuring smooth migration and scalability). Even within budget constraints, they're able to facilitate efficient, innovative solutions.

Line of business leaders: Crafting exceptional customer experiences
: Business leaders hold the key to shaping exceptional customer experiences. They understand that timely data is critical for informed decision making, yet they often face a dilemma: reliance on IT teams for data access. Imagine a chief marketing officer (CMO) aiming to personalize marketing campaigns. With logical data management, the CMO gains direct access to customer insights, behavioral patterns, and preferences. By analyzing data across all relevant data sources, the CMO tailors marketing messages, predicts trends, and delivers personalized experiences. Similarly, a chief operations officer can optimize supply chains, reduce lead times, and enhance product availability by accessing integrated real-time data across warehouses, suppliers, and logistics partners. By breaking free from IT bottlenecks, business leaders can stay agile, exceed customer expectations, and drive innovation.

Data leaders: Governing the data estate
: Data leaders are the custodians of the data kingdom. Their mission is to simplify data governance while promoting a data culture that enables self-service capabilities. Logical data management can help them achieve this, ensuring data quality, integrity, and compliance with regulations. The advent of low-code and no-code connectivity solutions, coupled with a reduced reliance on ETL processes, means there is less need for many developers to be dedicated solely to data management. Instead, existing developers can focus on delivering value-added solutions and products to the business at a faster pace, without the need to constantly build complex ETL processes or support monolithic architectures.

Compliance leaders: Shielding against risks

Compliance leaders are the guardians of the organization, protecting it against evolving risks and compliance requirements. Logical data management can equip them with the agility to respond to new types of risks or compliance regulations as they arise. For data consumers, this means they can trust the data they are using. Logical data management ensures that data is accessed in a manner that respects individual privacy rights and adheres to ethical standards. This includes mechanisms for consent management, data anonymization, and secure data handling practices. Logical data management provides a framework to ensure compliance with various data protection regulations and ethical guidelines. This not only mitigates legal risks but also enhances the organization's reputation for responsible data management. Moreover, a well-governed logical data management strategy promotes transparency in data processes. It allows both data consumers and decision makers to understand how data is being used, fostering trust and ethical use of data within the organization. Additional information about risk mitigation is covered in Chapter 7.

Operations leaders: Turbocharging operational excellence

Operations leaders, the driving force of an organization, are under constant pressure to enhance efficiency, productivity, and quality in the realms of manufacturing and supply chain, financial operations, marketing operations, infrastructure operations, and more. Logical data management can enable real-time, data-driven decision making for improved operational performance. Gartner reveals that 85% of infrastructure and operations leaders aim to have more automation in the next two to three years. A Deloitte study indicates that 86% of manufacturing executives see smart factory solutions as primary competitiveness drivers in the next five years. However, a Gartner report shows only 12% of infrastructure and operations leaders exceed CIO expectations, suggesting a significant scope for improvement through effective data management.

Financial leaders: Pioneering the digital frontier

Financial leaders play a pivotal role as navigators, guiding organizations through the complexities of turning data into a valuable asset for the whole organization at a cost that is manageable and predictable. Logical data management can

serve as their compass. It optimizes costs by eliminating data redundancy and streamlining resource allocation. It ensures compliance with regulations, safeguarding sensitive financial information. By providing agile access to integrated data, logical data management empowers timely decision making, enabling strategic initiatives to thrive within budgetary constraints.

Conclusion

The future of logical data management is promising, with trends indicating a continued shift toward distributed data management. As organizations deal with data and applications that remain distributed across regional and cloud boundaries, logical data management will play a crucial role in managing this distributed data efficiently and cost-effectively.

For data consumers, logical data management acts as a bridge, seamlessly connecting disparate data sources while ensuring security and compliance. Data discovery becomes intuitive, and the power of information lies just a query away.

Logical data management offers substantial benefits, including improved decision making and data consistency. However, organizations should consider challenges such as implementation effort, initial costs, resource allocation, and balancing rigor with agility. Smaller companies may need to weigh the benefits against their investment. Legacy systems, data governance, scalability, and complexity also play a role. A thoughtful approach and ongoing evaluation are crucial for successful adoption.

In the next and following chapters, you'll see how logical data management is being put into action today to solve intricate business and technical challenges with efficiency and intelligence.

CHAPTER 3

Data Mesh and Data Fabric

The earlier chapters identified many questions about how your business deals with distributed data. How do you collect that data to make it accessible to all your employees in a single unified format and location? How do you govern the data to ensure that it is accurate while also being secure? How do you create centralized accessible data while still maintaining the granularity and detail needed for domain-specific reporting?

These questions all challenge modern business methods for data management, and two approaches have been developed to tackle these problems: data fabric and data mesh. However, as we shall see, they approach the problem in different ways. This chapter digs into both of these solutions and uncovers how logical data management is the basis for both.

What Is a Data Mesh?

A *data mesh* is a mix of technical and business practices that empower domains within a company to manage and maintain their own data (see Figure 3-1). This approach enables them to meet their reporting needs without needing to force the data into a centralized governed structure. While some aspects of the data are controlled centrally through federated data management, the remainder is left to the specific data domain to maintain.

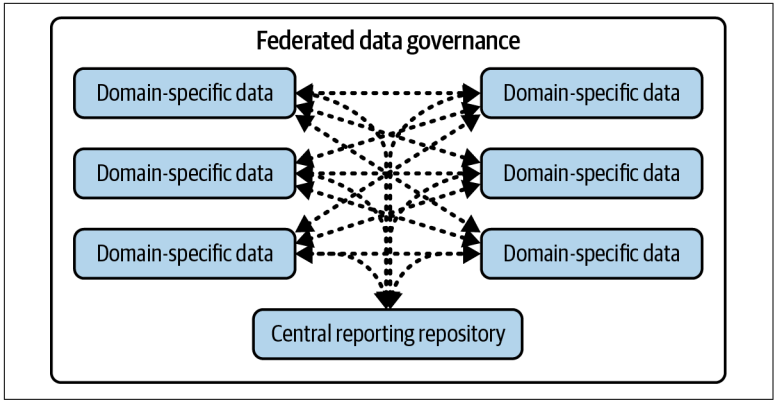


Figure 3-1. A data mesh empowers business-specific domains to manage their own data, allowing them to meet their own reporting requirements.

For example, imagine a retailer with multiple stores across the world. This retailer needs data from all domains of its organization to make effective decisions—human resources, suppliers, transportation, marketing, and more. A data mesh allows each of these domains to operate independently, generating data for its own needs. At the same time, these domains create data products—collections of data packaged for easy access and usage by the rest of the organization. These are designed with business language and logic in mind, for use by particular individuals or functions.

While the individual domains are responsible for their own data, they still share and rely on information from other parts of the organization. The main difference is that the data released by the individual domains is cleaned, organized, and validated in a way to make it easier to use and interpreted by other parts of the business. In some organizations, the data generated for external use can then be placed in a central reporting structure, such as a data lake or data warehouse.

What Are the Benefits of a Data Mesh?

A data mesh is popular for many reasons. First, the data resides with the part of the organization that knows it best. Traditional data environments rely on a central team of data engineers to collect, structure, and clean the data before providing it back to the business. While these engineers are experts in data management, they

usually lack the business understanding required to process the data efficiently according to each domain's needs. The people who best understand customer data are the customer-facing teams, the people who best understand finance data are the finance teams, HR will best understand employee data, and so forth; it is usually unrealistic to expect one central data team to have the same expertise in the way each team's data should be structured and used. This results in a continuous back-and-forth between the data engineers and the subject matter experts within the department, causing slowdowns and potential quality issues.

A data mesh solves this problem by shifting the responsibility of data management to the organizational unit where it originates. For example, a company using a data mesh allows its HR department to manage HR data. This makes sense, as the employees most familiar with the business processes related to human resources are also responsible for controlling the collection, organization, and release of their data. Their knowledge about the HR systems enables them to provide faster and more accurate data to the company. This approach also empowers them to structure their data for their own reporting needs.

A data mesh also meets unit-specific reporting needs. In a traditional environment, data is collected from all over the organization and forced into a centralized structure. This structure, usually a warehouse, follows a distinct set of rules and requirements to ensure that the information within utilizes the same structure, governance, and design. The end users need to know only one platform to access and report on business data.

While the centralized repository provides a single source for information about the business, the data within is often too broad or not organized correctly to meet the reporting requirements of the individual domains. When these issues arise, the solution is often to make shadow systems or separate reporting data sets specific to the task that needs to be analyzed. A data mesh reverses this process in some ways by ensuring that the local reporting needs are met first. The organizational unit can then selectively choose which data and granularity level is needed by the business before releasing that data centrally.

The final challenge a data mesh tries to solve is flexibility. Businesses need to respond to change quickly. The structures, processes,

policies, and procedures a company follows react to changes in the global business environment, society, and customer demands. This means that new data, new fields, new tables, and even new domains arise within the company on a continuous basis. If a company is using a centralized data management process, the data requirements associated with these changes are often complex and take large amounts of time to fully integrate into existing systems.

With a data mesh, the company can quickly adjust by integrating a new data product within an existing data area. This ensures that the data product created within that domain is specific to its reporting needs. The business experts associated with the new data take responsibility for collecting, structuring, and organizing the data for their own needs. Teams can then select the appropriate tables and fields required and ensure they are clean, complete, and accurate before exposing the data to the rest of the company. The data mesh allows the business to integrate the new data product quickly when compared to traditional alternatives.

The final benefit of a data mesh is governance at multiple levels. The subject matter experts within a business domain can set the security rules for their specific domain. They can set policies that allow them to dig into data details that are relevant to their business processes but aren't as important to the organization overall. Comparatively, global administrators can set business-wide security rules such as those required for regulatory compliance. The governance policies of the mesh fall under the global governance and can be managed in a more federated format. There is one type of database, one security model, and a centralized team dedicated to monitoring and maintaining access.

What Is a Data Fabric?

A *data fabric*, on the other hand, is an architecture of interconnected data sets. Forrester defines it as "orchestrating disparate data sources intelligently and securely in a self-service manner, leveraging multiple data platforms to provide a unified, trusted, comprehensive view to customers across the enterprise." Gartner defines it as "an emerging data management and integration design concept that supports data across the business through flexible, augmented, and sometimes automated data integration." The underlying theme of both is that technology creates a layer across disparate application

data and reorganizes it in a way that more accurately reflects the business language, needs, and demands, as illustrated in Figure 3-2. This enables end-user access with a common way to integrate, govern, and secure the data. A data fabric is configured to process the data through automation, apply data simplification, and enable self-service access.

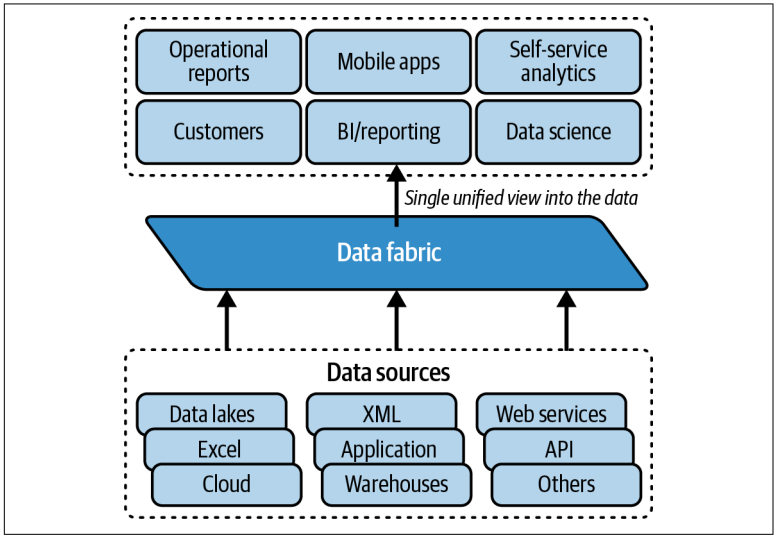


Figure 3-2. A data fabric pulls together disparate data sources and makes them accessible to analysts and end users.

Data fabrics have been used in many areas to solve data management challenges as well. Think of a higher education institution. It collects data from multiple sources and multiple systems related to recruitment, enrollment, tuition, fundraising, athletics, and possibly more. This institution needs a tool that can integrate these disparate sources and technologies into a central usable source for reporting and analysis. A data fabric allows the institution to track students from application to graduation and beyond through a central unified platform.

What Are the Benefits of a Data Fabric?

The biggest benefit of a data fabric is democratization of data. The data created by business systems is often structured to meet the needs of that system. For example, if your business utilizes a shipping tracking system, the data it generates is structured in a way

to make that shipping software run most efficiently. Unfortunately, that means that the data is also probably not designed to meet the needs of your analysts or employees. A data fabric sits atop that complex structure and simplifies it, making it easier to read, easier to understand, and simpler to use.

The second benefit of a data fabric is metadata—information about data usage and access that is collected as part of the system. This metadata provides valuable information about what data is used, when it is used, and how it is combined with other data. The system can then provide feedback to administrators on where potential bottlenecks exist in the flow of data. It can also provide recommendations as to places where automation can improve performance. More about this functionality is discussed in Chapter 9.

Data fabric implementations are also flexible with new content. When content is added or existing content changes, the fabric can be manipulated to integrate these changes without the complexity of joining together disparate data types and sources. This reduces the need for custom data integration processes to pull information in traditional data management systems, which means faster times between implementation and access to the data.

One final benefit of a data fabric is broader access to information across an organization. This approach creates a virtual layer atop the data, giving the organization the ability to utilize business-related terminology instead of complex data source acronyms and abbreviations. This allows employees at any level of your organization to integrate other aspects of the business into their domain-specific data. Using a data fabric breaks down the silos that exist within traditional data management systems. It also improves data delivery times by up to 65% over traditional ETL processes, according to Forrester.

Using Logical Data Management as a Basis for Data Mesh and Data Fabric

The concepts within a data mesh and a data fabric are not mutually exclusive. In fact, the technology of a data fabric can work collaboratively with the distributed nature of a data mesh to build a structure that meets both domain-specific needs and broader data

accessibility. This combination of technology and ideas is built upon logical data management.

How is logical data management a basis for both data fabric and data mesh ideas? The most obvious similarity between all three is that data in big organizations is diverse and distributed. It can come from machinery, customers, HR, weather, shipping, and more. In many places, these data sources function independently and provide valuable insights; however, many companies wish to combine that data with information from other parts of the organization. Disparate data types, field values, and specific terminology can make the data challenging to combine and even more challenging to use outside its source business domain.

Data Fabric

How does logical data management facilitate a data fabric specifically? Information exists in many domains across a business. Logical data management allows multiple data sources to connect and combine under a single layer. This approach combines the tables within individual sources, but also weaves together disparate data sources. The end user sees a cohesive view of all the data without needing to understand the query language, structure, and syntax of the underlying sources.

Beyond creating views of distributed data, logical data management also translates domain- and data-specific terminology. The logical layer takes the complexity of product-specific field names, table names, and logic and organizes it into a nomenclature and a structure that make sense to consumers within the business. The benefit of this is that data consumers are provided with data products containing familiar terms, making the information accessible for analysis by anyone in (or potentially outside) the organization. The logical layer provides flexibility as well by making it easier to adapt to underlying data changes without impacting the information already available and familiar to data consumers.

Logical data management also incorporates the idea of centralized control for data governance. When companies implement a logical layer, they move the governance from the individual sources to a central access point. This enables them to give the right people access to the right data in a distributed landscape. Logical data management also enables AI tools to access metadata, which it

can then use to better manage the data performance through automation. At the same time, data within individual departments is managed independently. Each department controls and secures its own data source, enabling teams to flexibly meet their own access requirements and reporting needs.

Data Mesh

What about a data mesh? Logical data management provides several base components of a data mesh. First, it provides a location to share data that is usable business-wide without replicating or altering domain-specific sources. The detailed data within the domain is still utilized, managed, and owned by those with expertise. Since each domain is responsible for its own data, teams can manage access, definitions, and structure independently.

Next, any data that is relevant to the rest of the organization is translated up through the virtual layer, making the data accessible and understandable. The domain determines how data from its domain can be translated into usable resources for the central business. The domain owners create a data product that uses business terms and common language that is then provided to the rest of the business through the logical data layer.

The data products provided by the source business domains are then provided to users within a self-service interface. Data consumers no longer require access to multiple domain-specific sources. Additionally, the information is presented in a way that works for the business without affecting the underlying data. This eliminates the need for fluency in multiple coding languages. Data consumers are provided a platform where they can access multiple areas of data labeled using business terms, and data sets from multiple domains are woven together through the virtual layer.

Finally, logical data management allows organizations to develop a central governance policy across the virtualized data in addition to those specific to each domain. Experts within each domain are tasked with providing the right fields and tables to integrate and make available to the rest of the organization. While each domain administrator sets governance policies on their specific data, a central governance group sets standards and governance for the company as a whole.

In short, logical data management implements a self-service platform for business users to access data across the organization. Logical data management centralizes data governance and simplifies enabling data governance policies. While data is labeled with business nomenclature, domain-specific data is still available for strategic analysis. Additionally, the logical layer allows these specific domains to identify, rename, validate, and provide area-specific data to make it more accessible to the rest of the organization.

Case Studies

The best way to understand how a logical data layer improves data democratization within your organization is to review examples of using a data fabric and data mesh within other businesses.

Finance

In 2017, a leading financial institution was facing data challenges. Each department had its own architecture and data policies. Data was siloed and replicated, maintained through separate ETL tools. As a result, governance over data within the organization was lacking, and data lineage was difficult to track. The company needed a way to modernize its data management.

The company reached out to vendors for proofs of concept. The goal was to find a solution that would integrate with the current landscape, consolidate business logic and data access, and establish a faster delivery of data to consumers. The company settled on a *logical data warehouse*, a logical architecture placed on enterprise applications and existing data stores to provide a centralized location for data access without the replication of the data itself.

While this solution solved the problem of decentralized data repositories, it unfortunately created additional challenges. The first obstacle was finding developers who could translate domain-specific data sources into the new warehouse. Many of the data sources originated in software-specific tools, requiring knowledge of each vendor's structure, terminology, and language. Additionally, the developers needed to regularly meet with domain experts to understand the data and determine which data should be extracted to the warehouse and how to make that data make sense from a business language standpoint.

To complicate things further, many of the domain experts created views in intermediate systems to make it easier for the logical data warehouse to read and ingest. This broke the lineage of the data, increasing the difficulty of tracing the data back to its source. Additionally, some views were complex and created performance issues when they were pulled into the new warehouse.

These challenges led to a logical data mesh, which takes aspects of a logical data warehouse and blends it with the concepts of a data mesh. The ownership of the data and the content that is pulled is controlled by the domains themselves. The experts within the domain build the base data mesh view that is then shared with the logical data warehouse developer. This ensures that the most pertinent and accurate data is shared within the organization.

The logical data warehouse developers integrate the data mesh view into the logical data warehouse, making it available to developers in other domains. These other developers integrate that data with their own and publish a separate data product. Both domain developers work with data contracts, which identify to the software engineers data dependencies within their specific domain while ensuring interoperability and reusability across the data products. This helps the company avoid potential pitfalls from updates that may alter schemas or cause potential cascading issues.

The end result for the institution connected 16 systems, and over 85% of employees utilized the solution. They now call over 13,000 queries daily, using self-service tools such as Tableau and R. The solution contains 28 virtual databases, and 21 are specific to data-mesh-enabled domains. The solution also improved response times, using caching when needed. It unified the business under a central platform, providing data to everyone at all levels and tying together what was once a very distributed data system.

Aeronautics

Another organization also benefits from utilizing a data mesh solution. This company deals in aeronautics. When it started investigating data solutions, it struggled with diverse data sources, organizational change, data replication, and visibility into data lineage. The company was looking for a solution that would allow users at all levels of the organization to access data. The organization wanted clear governance and reusable data, and it needed the data to

be trusted and validated by business experts but still easy to access for users lacking expertise.

These parameters underlie the structure of a data mesh. A data mesh allows domain-specific experts to collect, validate, and deliver data to the central developers, who structure and blend it with data from other areas of the business. To implement the mesh, the company utilized a logical data management platform. This allowed it to maintain data in the source locations while providing data products for use across the company. The restructuring took the data out of the language of the tools that generated it and into the language of the business. It clarified the path that data travels from where it is collected to where it needs to be used. The data was then notated with business-appropriate labels and tags.

The result was a structure that encouraged self-service. Users moved away from replicated local data sets and started using a freely explorable central source for reporting needs. The responsibility for populating the central data fell to the domain experts, who managed and maintained virtual database layers, distributing the work and organizing it based on expertise. They controlled access through tags and global policies to ensure that the data is accessible to the appropriate users downstream.

The end users benefited from trusted data, as the domain-specific virtual databases are actively managed. This means the data is cleaned and curated for end use. The company now benefits from a flexible system that can adapt to new technologies and changes in the business operations. New domains can be added quickly and easily to the existing structure in much the same way the existing areas are. More importantly, analytics users can contribute to the domain-specific virtual databases with their own reusable products.

The company currently has around two hundred users across the organization. There are two domain-specific virtual databases, with another four arriving soon. In addition, 38 workspace-level virtual databases are in use. Their mesh consists of over seventy data sources and currently employs nearly seven hundred views!

The solution this company implemented is only one step in its data management journey. The company continues to add new products, bringing in new data sources and domains. It is continuing to mature its processes and determine how to federate diverse data into a central

usable resource. Eventually it hopes to incorporate machine learning and AI to help with data curation and insight generation.

Conclusion

The benefits of a data fabric and a data mesh, when supported by a logical approach to data management, are easy to identify. These solutions combine the business expertise of domains with the simplicity of business language to make easy-to-use data sources for democratized reporting. They also make the lineage of the data easy to follow while providing a central location for governance and security.

In the next chapter, you'll get an overview of how logical data management works. You'll also see in more detail how data integration, management, and delivery of data through virtual views can benefit your own business.

CHAPTER 4

How Logical Data Management Works: An Overview

You've seen how businesses rely on different strategies to deliver data for analysis. Some use a data fabric to connect disparate data sources. Others rely on a decentralized data architecture, called a data mesh, to utilize the knowledge of their subject matter experts to deliver domain-specific data to the organization. Many use a mixture of a data fabric and a data mesh.

Both of these techniques can be supported by logical data management, but how does that management work, specifically? How does your company integrate its many systems under a system that facilitates integration, management, and delivery without replicating the huge volumes of data that are required to make it run?

Data Virtualization

The core underlying concept and technology behind logical data management is *data virtualization*, a data integration technique that allows your end users to access data from any location within your organization without needing to know where it came from, where it is located, or how it is formatted. This technique also provides a view that allows users to access and analyze information without needing to know the technical aspects of the source. More importantly, data virtualization does not *require* the replication of data. It instead provides a lens into existing data, making it easier to access, understand, and utilize.

You can think of data virtualization as being similar to a computer's operating system. The underlying logic of the chips, circuits, processor, and more are extremely complex and impossible to fully grasp without specific knowledge of computer engineering and programming. Data virtualization utilizes those complex components to provide an accessible interface to a complex set of underlying data.

Data virtualization utilizes a metadata-only middle layer to separate the data from the underlying source. This concept is then applied across all the domains within the organization, creating a way to access multiple data sources from a single interface, such as a dashboard or application. Your users no longer need to understand the complexity or technology of the original data source. Instead, all data from across the organization can be viewed from a single place (see Figure 4-1).

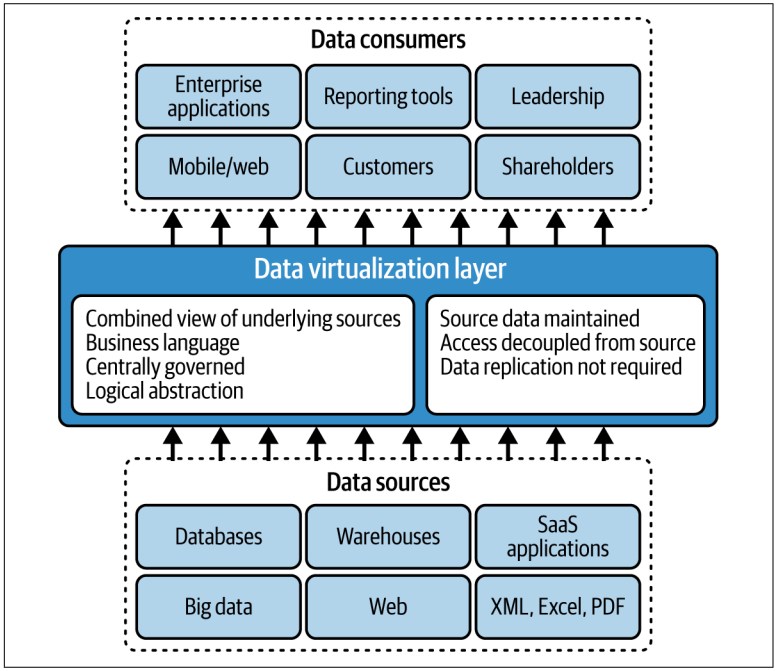


Figure 4-1. The data virtualization layer sits atop your company's data sources, creating a view for business data consumers. The complexities of the source data are abstracted through the layer into a data product that is easier to understand and access.

To understand how data virtualization works, you need to understand how data is pulled from your operational systems. The applications your organization uses to do business generate data; however, it comes in a wide variety of formats and structures and utilizes multiple methods of security. Your analysts, as a result, need to be technically fluent in multiple programming languages and methods for pulling that data. They need to be able to work in various databases, security, and reporting tools in addition to having knowledge of your organization's business processes.

Data virtualization removes the complexity of directly accessing diverse, underlying systems. Instead, data virtualization connects to these systems on behalf of the user. This initial connection relies on knowledge (metadata) about the data source, and data virtualization abstracts users from all the technical details behind the scenes. Data virtualization handles the technical details of connecting to these data sources, generating consolidated views, and ensuring the security of the underlying data.

Data virtualization also transforms each original source's views of the data into a single, virtual view, making the data easier to access and understand. Instead of multiple sources, languages, and types of security, users see a single representation of the underlying sources. In short, multiple systems in your business become a single, homogeneous view that can be easily structured for—and labeled by—different functions of your business.

The simplified version of data is also an important benefit of virtualization. Often, the type and structure of domain-specific data is complex and challenging to users outside that domain. Rather than training end users to understand that domain's query language and complex structure, the virtual layer provides a business language lens into the data that is easier to understand and use. The structure and design of this lens can be monitored and controlled by the subject matter experts to ensure data quality and integrity.

Data virtualization can be deployed to simplify data views for just one or two data sources, or it can be deployed as a layer across most if not all of an organization's data assets. When data virtualization is deployed as an enterprise-wide layer, it becomes a *logical layer*, or *logical data layer*, which forms the heart of a logical data management platform.

Logical data layers, enabled by data virtualization, also improve data governance and security. Since the users in your organization connect directly to the logical layer, your security team can focus on securing this one layer, rather than the many data sources across the entire data landscape. Permissions are granted at a single layer, and access is controlled at a single point. The original source data is still accessible to users with domain expertise, but a simplified, translated, aggregated, and secure version can be provided to the rest of the organization.

Data virtualization software can connect to a plethora of data sources. Within many sources, there are multiple ways to connect to and pull that data into the virtualized view. This means that your technical experts have flexibility in how the data is integrated, including on-the-fly multisource federation and point-to-point replication. The flexibility decreases the time needed to get from source data to delivered product and enables your administrators to develop an optimized strategy for each data source.

Additionally, the need for data replication is reduced. The logical layer sits upon existing data sets, and new, derived data products can be defined on top without worrying about the original location and format. Replication becomes an option, not an obligation. This is much faster than waiting for an ETL process to replicate the data into a warehouse before reporting on it. Sometimes replication processes are necessary, but often they delay timely access to information or prevent the system from accessing the freshest data. In Chapter 6, we will discuss these ideas in more detail.

On the other side of the data flow, many types of data virtualization solutions provide the data in multiple formats. This is extremely beneficial since it means that the underlying data can be accessed in a format that is familiar and easy to use for each end user. Many vendors provide query options, such as SQL, while also providing APIs or web services, such as OData, REST, GraphQL, JSON, and more. Each user within your organization can access the data in the way they are most comfortable and familiar with.

The semantic modeling built into data virtualization allows for terminology and data structures specific to the organizational domain. Each area within the business can create and consume the parts of the data set that best suit its needs. Since the layer is entirely managed by the organization, the schemas, tables, and field names

can represent business terms rather than the technology that creates and stores the data. This approach makes data accessible to all users across all domains and minimizes confusion about the data and its structure.

Inventory is a great example of how an accessible data source could be beneficial to multiple teams. Finance may want information about resources on hand to calculate budget. Marketing may want to know about inventory surplus to support an upcoming promotion. Your warehouse manager and shipping teams may be interested in what supplies are available and what might need to be ordered soon. These and many more examples are just some of the ways an accessible source of data benefits multiple areas of an organization.

Security Policies and Administrative Concerns

With any data system, the primary concerns are privacy, security, and quality. Logical data management overcomes many of these obstacles through the centralization of governance. The logical layer sits upon existing systems and translates the data through a business lens, making it understandable and accessible to the end users. To ensure that everything translates accurately and efficiently, administrators need to work with the business domains to verify that the logical layer provides accurate data to the appropriate consumers.

In your business, you likely have multiple domains that generate that data. You likely have information being captured from human resources, finance, suppliers, marketing, and more. Each department utilizes different applications to maintain its processes, and likewise, each has its own data structure, technology, and security. Those responsible for securing and providing access to those data sets likely have to be familiar with different types of databases, different types of security, and multiple layers of access.

With multiple types of operational software and databases comes varying types of security capabilities. The permission options and levels within each can be vastly different. Data that is restricted in one database may be accessible with the same permission levels in another. Other systems may not have a data access control policy at all. It becomes increasingly challenging for your security team to consolidate the access rights, access levels, and permissions across the wide variety of underlying systems.

To complicate things further, some of the systems may not integrate with your company's single sign-on software. This means multiple login processes, multiple accounts, and multiple passwords for your end users to track. At the same time, your security administrators need to develop separate systems for tracking and managing employee arrivals and departures to ensure that data access is added or removed appropriately.

In addition, many systems that provide data to your organization are designed for generating analytics, but those account for only a portion of the data your company needs. Some production systems generate data but are difficult to access because of service level agreements. Others may not support the physical load of both business processes and data reporting. Though the systems are challenging to access and secure, the data within them is still valuable to your organization.

Logical data management improves the management of all your data through unified views of disparate data. Rather than focusing on multiple source locations for data security and management, your administrators can focus on securing at the logical layer. Access and permissions are controlled through the logical layer rather than the individual source systems. In addition, the security of the logical layer can often easily be integrated with your company's single sign-on service.

A centralized security process for data also enables simplified auditing. Your security administrators can check and validate access to all data sets at once. When new employees arrive that need access, they can be granted permissions to all appropriate systems at once. Likewise, when an employee leaves or retires, administrators can quickly identify all existing access grants and remove them. This simplification of security lends itself to role-based grants as well, meaning that users within a specific role in the organization all have access to the same data. Users can be given a role rather than individually being granted permission to each separate data source.

Improving Performance, Working with Caching, and Other Special Cases

As we discuss logical data management and data virtualization, you may wonder how this additional logical layer might affect query performance. It seems like having an additional step between the consumer and the data source would delay access to data. In reality, logical data management platforms do not impede performance; rather, they offer a variety of additional optimization options, so you can choose the best approach for each use case.

In many cases, the data that is needed for a particular query is already in a data warehouse or lakehouse, systems with plenty of muscle to process data at scale. A data virtualization engine in a logical data platform is able to utilize this power. It can leverage the underlying system to run the query, passing through the consumer's queries directly to the underlying system data source. The power and performance of the execution is determined by the capabilities of the underlying systems.

However, it is not always appropriate to utilize only the underlying data warehouse or lakehouse to process the queries. Sometimes the queries require additional tools to join, manipulate, or aggregate the data. Let's use a few examples to explain how a logical data platform can tackle more complex scenarios:

- Even in powerful systems like a data lakehouse, some queries take too long for the patience of an avid analyst. Furthermore, executing the query every time can be expensive from an economic point of view (think pay-per-use cloud data warehouses). For these situations, logical platforms offer several options, like *caching* or *aggregate-aware acceleration,* which can transparently replicate all or part of the data in the logical platform without needing to manually create and maintain replication pipelines. Acceleration factors in these situations can improve the performance by a factor of 10 or more.
- In the modern business landscape, you may need to look beyond the data available in the data lakehouse and include other pieces of information. Data virtualization can provide a closer-to-real-time view of data within your organization while also opening the doors to otherwise inaccessible sources of data. Queries could come from a variety of sources and be distributed

across multiple data sources. Imagine a report that combines Salesforce information with your data warehouse to make a sales forecast that is based on both the historical data in the data warehouse and real-time data available only in Salesforce. Data virtualization engines have specific capabilities to manage multisource execution (also called *data federation*) in an optimized way, including advanced federation techniques and massively parallel processing. The options for caching and selective replication also make sense when data freshness is not critical.

- A logical layer may also sit atop operational production systems, allowing your business access to this data with guardrails. Depending on your use case, your business can allow access to the production data sources while simultaneously limiting the number, size, and frequency of queries being placed against them. This brings operational data to your employees and pushes data-driven insights and decisions closer to real time.

As you can see, logical data management platforms increase your options in terms of query execution beyond those provided by conventional physical architectures. For each data product exposed in a logical platform, for example, you can do the following:

- Leverage the processing power of the data source, without any data replication
- Combine multiple data sources in a single query
- Use caching and other acceleration techniques to partially replicate certain data sets for performance and/or cost reasons
- Choose to materialize its data in a specific system, such as a data lakehouse

Modern logical data management platforms optimize performance through several methods that are explained in detail in Chapter 6.

Conclusion

How does your business approach a logical data management implementation? There are multiple ways. One method is to start with a targeted data set. Choose a domain within your organization and use it to compare the benefits against your existing data management approach. Alternatively, apply logical data management at an enterprise level. Because it is a logical data layer that sits upon

your existing systems, you can apply it without changing existing data management processes. You can continue to collect data and report on it with your existing management processes while also experimenting with a logical data layer.

Data virtualization is the cornerstone of logical data management. Using it, your company can get the benefits of a data fabric, a data mesh, or both, to create a unified data platform. This not only embraces domain-specific subject matter expertise but also creates a data resource that is connected and accessible to your employees. Data virtualization provides centralized governance and security while also translating complex data structures into easy-to-use business language. The next few chapters will dig into the heart of logical data management to uncover how it translates your business language, adapts to growth and change, and provides a central location for governance and security.

CHAPTER 5
The Semantics of Business

Successful businesses make decisions based on data. In the past, these key decisions fell to company leadership and executive officers. They relied on dashboards and reports designed to give them insight into how the business was performing. Today, however, many organizations are realizing that decision making at the top level is not enough. They need to empower employees to make good choices at all levels of the organization. As a result, there is a push toward data democratization.

Data democratization is the process of enabling data access throughout an organization to empower employees to make informed decisions. This is a very powerful and beneficial idea, but first, your company needs to answer some challenging questions:

- How does your company get the right data to the right people?
- How can that data be made accessible and understandable to all its consumers?
- How does your company ensure that the data being used is accurate and reliable?
- How can your company deliver data in a timely manner to promote decision making that is closer to real time?

The major obstacle underlying all of these questions is disparate, diverse, complex, non-user-friendly data. In addition, these data sources often lack quality, documentation, and insight into their lineage. The data sources your employees rely on to make decisions

can be from anywhere. They can be housed in different formats, different structures, and different security models. Based on its source, data is also likely stored in a language that employees of the business are not familiar with. A data democratization initiative struggles to deal with these obstacles. It faces challenges in overcoming the multiple formats and semantics as well as the variety of security methods and policies associated with each disparate data source.

In this chapter, you'll see how a logical layer utilizes semantic models to address these obstacles.

What Is a Semantic Model?

The previous chapter discussed data virtualization, which provides view-like access to underlying complex data systems. Unlike a traditional view, it can traverse multiple physical data sources and formats, creating a holistic view. A *semantic model* describes the objects in a database and their relationships to one another. This model provides a view of the data from the perspective of the business, making the data easier to understand and use. In short, it is the translation between business systems and data consumers.

The simplest way to understand a semantic model is through an example. Imagine you are running a small community college. Your school has information about all sorts of things from finance, to HR, to students and classes. Any student in your school has information recorded about them. The admissions office tracks information about their application and admission, whereas the finance team may track information about their tuition payments.

Each of these parts of the organization sees the student in different ways and tracks different information; however, only one student is being tracked. The relationship between these data sets is valuable, as all this information is required to progress the student from application to degree. At the same time, each area has specialized knowledge about the student that is specific to their function. A semantic model is a formal representation of the relationships between these separate domains along with the attributes that each tracks in its own data set.

Data virtualization exposes data in formats that are easily recognized and understood by the data consumers. The data underlying these virtual databases does not necessarily come from the same source

and may contain additional information that helps to classify and organize the data within. Multiple views can be created depending on the target audience. If you consider the higher education example from earlier, you can see that student financial information is relevant to the financial management team, which tracks incomes and expenses for the college. At the same time, that data is also relevant to the student, who needs information on how much tuition is owed and when (Figure 5-1).

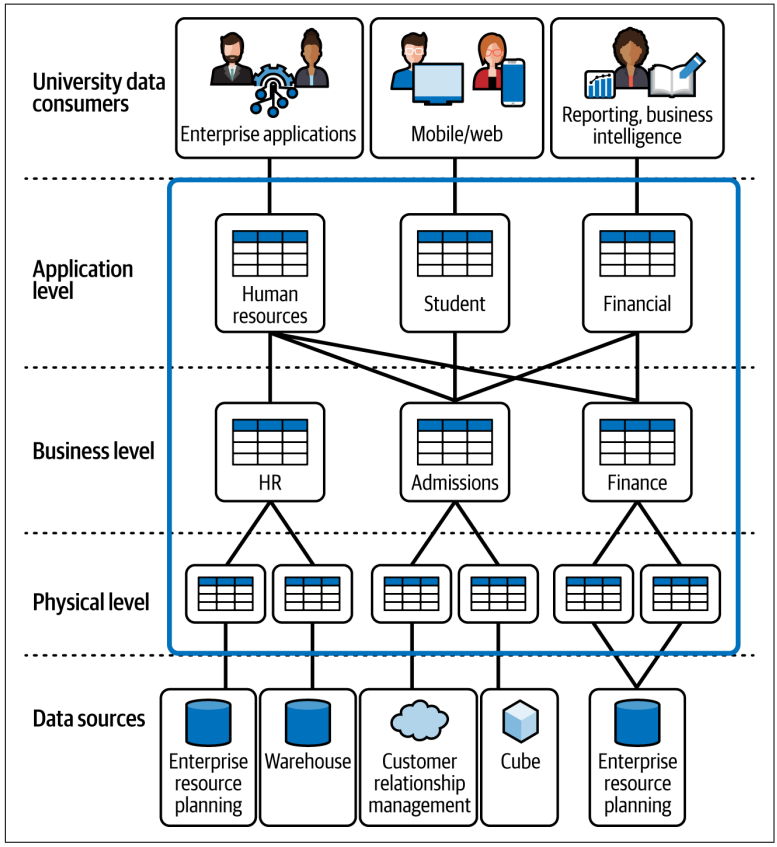


Figure 5-1. Data virtualization is used to provide a semantic layer, which is a virtual application of a semantic model, to assist end users in understanding and using data from complex operational systems. The resulting data product can be configured to meet the needs of specific consumer applications.

These data views are the semantic models of the data. Semantic models are the metadata that define meaning and relationships

among pieces of data, while a semantic layer applies those models to the actual underlying data. The semantic layer translates operational data into something that is accessible and understandable to the end data consumer. The logical layer is usually structured in a hierarchy, which enables reusability. In other words, a data view is created that provides the key fields relevant to the organization. The view is then repeatedly used to create data products, which are designed to meet the needs of the consumer. These data views typically include additional semantic information, such as descriptions of the view, descriptions of the fields, and other tags.

The Power of Semantic Models in Data Management

Semantic models enable organizations to organize, manage, and access data. But what benefits do they provide, exactly? How does a semantic model improve your company's access to data? Let's look at a few benefits:

Consistency
: The semantic model is a business-friendly definition of data from multiple complex underlying data sources. This means that even though the data is pulled from multiple sources, the model can align terms and definitions. In this way, consistent governance can be applied that creates a common definition and term for an object even though it may be labeled or used in other ways within the various domains.

  Imagine a distribution center that is responsible for fulfilling online orders for products. The warehouse team may have values for the field Order Status that relate to identifying and collecting the product, such as "allocated" or "in queue." The shipping team, on the other hand, may have values such as "packaged" or "shipped." A semantic layer combines the data from both areas and sets a standard to each. This gives a holistic view of the order status regardless of its progress from received to fulfilled.

Maintenance
: Another advantage of a semantic layer is maintenance. Since the data is translated through the semantic model, adjustments are made centrally. Changes to definitions and values happen in the model rather than requiring administrators to track down the field through multiple data sources and domains.

Development
: The semantic model provides clear labels in business language that is easy to understand for not only the data consumers, but also the developers. This makes it faster and less complex to develop. Developers can focus on achieving the objectives of the business instead of reinventing data structures.

  Likewise, the semantic model reduces the chances for data quality problems. The data is available in a central representation rather than multiple separate underlying data sources. This allows for better maintainability, improved explainability, and reduced complexity, and decreases the chance of implementing poor-quality data.

Reusability versus unification
: The unified approach to data does have one drawback. If a data quality issue is introduced in the mapping of the source data, that error can propagate across the organization—everywhere it is used. This is of course addressed by the fact that it can be repaired quickly in one central location rather than multiple underlying data pulls.

  This challenge is something that businesses need to balance and be aware of. The semantic model is like a foundation upon which a structure is built. A single flaw can bring down a building. Conversely, it is oftentimes much easier to build upon that base than it is to build multiple foundations for separate independent structures. Your business needs to decide which approach to take when building the data structure, and this is where a logical data model can help.

Adaptability and Business Context: Speaking the Language of Business

Business processes are usually built on software that stores data designed for the optimization of that process. For example, your company probably uses a human capital management tool to manage and maintain your workforce. As such, the tool generates and delivers data in a format that works for the tracking of human resources. Unfortunately, the names, design, and structure of this data rarely support the reporting process.

The structure of source data is often designed with the product in mind, leading to strangely named fields and complex structures that are unintelligible to a data consumer. Instead, your company should use terms that it is familiar with and that are well-defined. It's much easier to report on Employee Years of Service than it is to report on EMPL_YR_SVC_CD. The semantic layer makes this translation by creating a logical view of the data structured for business purposes and labeled with business-related terms. It does this through three aspects:

Simplified data landscape
: The semantic layer works by translating user queries against the semantic layer into physical requests against the underlying data. Additionally, the language of the semantic layer is separate from that of the underlying base data, meaning that this abstraction also applies when building the semantic models. This makes the design process easier to facilitate, requiring less knowledge in the base data language.

Customization
: The semantic model is easily customized for the way it is to be used. A higher education institution might utilize a semantic model to track faculty, students, and staff, with the flexibility to adapt the software to meet their needs. Each division of the institution may have different methodologies and processes that it uses to keep the school running, and each requires different subsets and structures of the underlying data. At the same time, the data still needs to be standardized to meet regulatory reporting requirements.

Enhanced data accessibility
: Semantic models take the language of the source data and make it accessible to your users. Most data consumers are not fluent in data coding languages. They may lack the proper tools and knowledge to run a SQL query against a data source, for instance. Translating the data into a common language makes it much easier for users to perform self-service analytics, as the data is user-friendly, logically structured, and easy to understand.

In summary, semantic models provide a layer of translation between the source data and the language of the business. The model makes accessing data easier, simplifies complex models, and aligns the data with business processes.

Enabling Logical Semantic Models Through Data Virtualization

Using a semantic model has multiple benefits, but how does your business enable one? Data virtualization can provide many capabilities that enable semantic models. Understanding these capabilities is beneficial for enabling a backend for one:

Portability
: Data systems evolve. Sometimes your company needs to change vendors, hardware, or make changes to a more cost-effective platform. If your semantic models are tied to a specific technology, you lack this power to adapt. Data virtualization is not tied to specific technology. It is able to evolve to integrate with multiple data systems and infrastructures, allowing you to move data from source to source without interrupting your business processes and data flow.

Unification
: Semantic models tie your business together. As a result, they should not be restricted to certain types of systems. Your company may have data in source systems, warehouses, lakehouses, or more. Data virtualization enables this connectivity. It can reflect data from multiple sources and types to provide it as a unified data view. This provides a single location to look for data and is invaluable to self-service, which will be discussed in detail later in this chapter.

Streamlined development
: Data virtualization products typically provide a graphical interface that allows nontechnical access to data. This enables subject matter experts to interact with, modify, and change data to meet shifting business needs. Changes in the model also need to be as automated as possible. AI can be used to identify changes in source data, do impact analysis, or track data lineage to adjust the model on the fly. Development should be based on existing models as much as possible, and reusability is key. Data virtualization tools provide these capabilities.

  Data virtualization also provides modeling from a top-down perspective. You can generate your semantic model by using an external application and then import it into the data virtualization tool as interface views. Developers can then implement

these interface views and ensure that they are compliant with the model.

In addition to technical adaptation, the semantic model should be able to adjust to business changes as well, integrating with tools such as Git, Jenkins, or Azure DevOps. This ensures that the models created evolve with the business data sources and customer integrations simultaneously in choreographed migration events to minimize disruption.

Security

The most crucial aspect of data management is security, and it is important to review the security of your data at two separate levels.

Security during development is vital to allowing your teams to collaborate while allowing only a select few to create or edit objects. Independence between domains should be established, especially those that reuse one another's semantic models. It's also vital to maintain that separation as it delineates a separation of duties. The developers who set up access control and performance tuning should not be the same as those who translate the business terms into the data descriptions.

Security during execution refers to who can access the data, and is usually based on roles and session attributes. Semantic models enable access restrictions, data masking, and row-level security through role-based and attribute-based access control. This helps your organization ensure that the correct people are accessing the correct data and mitigates security risk. The upcoming chapters elaborate further on existing security capabilities and how to use them effectively.

Observability

Access to the data is only part of the security process. In addition to controlling who has access, organizations also need to see how that data is being used and where. Data virtualization provides insights on usage and usage patterns. This provides not only an audit of where data is being accessed, but also information on improving data accessibility. For example, if multiple users are repeatedly accessing a certain set of data, the virtualization tool can identify it as popular for others seeking insights.

Extensibility
: Semantic models can be built to simplify the search and exploration of data through virtualization. Using a solid solution enables data consumers to identify data quality and popularity through labeled levels such as gold, silver, or bronze. A solution also helps track data lineage, identify the data owner, and provide a platform to reach out with data questions. This ensures that the consumers of the data are using the right data in the correct way and provides a feedback loop for the business domains to ensure they meet business needs.

Flexibility in execution strategy
: The semantic model is independent of the execution strategy used to obtain the required data. Different use cases or data products can use different strategies, and these strategies can change for performance or cost reasons without affecting data consumers. For instance, a semantic model designed for self-service analytics may implement aggregation-aware techniques, whereas a specific data view that is frequently used may implement caching to ensure very fast performance. Yet another view may access the data source directly to ensure that the data is fresh. The goal of the semantic model is to meet the needs of the data consumers and abstract them from the changes in these execution methods.

Multiple delivery methods
: This chapter has focused heavily on how data is interpreted into a semantic model, but equally important is how it can be consumed. Your end users need a platform that provides data in a format that is easy to use and understandable to the end user. Data virtualization is valuable for this reason. People can use SQL if they are familiar with it, or there are options to pull data through many data visualization tools, such as Tableau or Microsoft Power BI. Applications can access the data in multiple ways as well. Data virtualization can provide data in REST, OData, or GraphQL API format.

  Open data formats such as Iceberg and Delta also provide options to share data across teams, reducing replication and increasing reusability.

Other Technologies That Enable Semantic Models

Semantic models can be implemented through a variety of other technologies. It's important to understand where these models exist and what their limitations are:

Reporting tools
: Many reporting tools allow users to restructure and relabel data to meet reporting needs. This is incredibly convenient when developing end-user reports; however, the structure and labels created within the tool are usually limited to the tool itself. This reduces reusability and negates some of the flexibility of a traditional semantic model.

Multidimensional engines
: Many companies elect to restructure and relabel data into a multidimensional storage solution, also known as a *cube*. The benefits of a cube are that it allows for advanced modeling, security, and documentation. The data can be structured or aggregated for easier reporting or improved speed. Unfortunately, like the reporting tools, the usage is limited. Most cubes are limited to a specific use case, making them difficult to reuse for other purposes.

SQL transformation tools
: Some SQL tools, such as dbt, exist to translate otherwise complex data. In short, a developer writes SQL code against an existing table or tables to generate more usable data in another table. This method of translation is often combined with continuous integration and continuous development practices. A major obstacle with this approach is security during development: since all teams share the same codebase, independence between domains is very hard to achieve. In addition, the approach is limited to a single data source, and the SQL code needs to be written using the specific dialect of the underlying system, reducing the code's portability.

Defining a Data Marketplace

As mentioned in the earlier sections, semantic models are designed to translate complex data systems into usable, business-defined tools. They make data accessible to a wide variety of data consumers while utilizing a framework that is governed and secure. The

semantic model should be well-documented, easy to explore, and well-defined. The data should be reliable and accessible through a wide variety of methods to enable trustworthy and timely data reporting at all levels of an organization. All these aspects come together to generate a data marketplace or a data catalog. A *data marketplace* is a complex and full-featured interface placed on top of file systems to make it easier to discover and access data. A *data catalog* is a simpler version of a data marketplace.

A data marketplace is an ecommerce-like experience for all things data within an organization. The various domains of an organization generate data that is integrated, transformed, translated, and made available in simple business terms. Users can then explore, search, and select the data that most meets their reporting requirements. Once they find the data they need, they should be able to easily extract it and utilize it in a tool of their choice, such as Tableau, Microsoft Excel, or another reporting application.

To build an effective data marketplace, businesses need to treat the data like a commodity in a store. The store will be successful if the products within are easy to search through, easy to find, and easy to "purchase." The quality of the product, in this case the data, is valuable to the consumer, and maintaining that quality is a top priority for the business. To accomplish these goals and ensure that your data consumers are happy, your business needs to focus on these key components of a data marketplace:

Discover and query
: Your data products need to be discoverable. The interface of the data marketplace needs to be integrated with advanced search capabilities. It should be integrated with natural language processing, for example, to ensure that the terms used by the consumer are tied to the data products within. Additionally, the data marketplace needs to be able to categorize data products and include features that facilitate navigation, such as autocomplete, product suggestions, and search filters.

Collaboration and sharing
: It's not beneficial to an organization to limit data discovery and insights to those with reporting knowledge. Data needs to be shared with others through insights, annotations, and feedback questions. This empowers users to collaborate to discover new insights or request additional data access to analyze data

they were previously unaware of. At the same time, the system should be integrated with methods for users to request that access. There should also be methods for sending data inquiries to the stewards and for the stewards to provide responses.

Security and governance
: The data marketplace is a store for users to access data, but not everyone should have access to all the data. Restricting and monitoring data usage is critical to ensuring that the right data is being used by the correct people. There should be audit trails and data lineage tracking in addition to roles and attribute reviews when it comes to monitoring access and integrity.

Integration with existing tools
: Data is no good if it is not accessible in a variety of ways. Your marketplace should support multiple methods of data access through reporting tools and platforms. You need to be able to meet the needs of your users regardless of whether they pull the data through a complex data query, visualize it in a reporting tool, or dump it into an Excel file.

A data marketplace can revolutionize the way a company does business. It is the primary way to drive a platform of data democratization. It enables your employees to utilize data to make informed decisions regardless of their level of experience with the data or their role in the organization. The data marketplace should be easy to access, easy to utilize, and easy to maintain. Combined with a well-organized semantic model, a data marketplace empowers your employees to use data, improving data literacy, decision making, and collaboration throughout the organization.

Conclusion

A semantic model can greatly improve the way your organization interacts with data, and is empowered by data virtualization. The model simplifies data access through abstraction, separating your data consumers from the complexity of business process systems. It allows your data and domain experts to collaborate to create a platform that provides data in business-oriented structures and terms. A semantic model also integrates perfectly with a data marketplace, which creates an interface through which your employees can explore and extract data for their own use and decision making.

CHAPTER 6

Scalability and Performance

Your business needs data from multiple sources inside and outside the organization to enable effective decision making. It needs to access data in multiple formats and through multiple security mechanisms. Logical data management serves as a conduit that connects these separate sources together under a single layer, creating an accessible, easy-to-navigate, and easy-to-understand data marketplace for anyone within your organization to utilize. But what happens when data grows or changes? How can data virtualization help your company adapt and mature its data environment without sacrificing the speed and depth of the data itself?

It may seem like data virtualization is another step between the source data and the user; however, when correctly configured, the difference in speed is minimal, and data virtualization can even speed up slow data sources by using caching and query acceleration techniques. There are many opportunities to optimize performance, and this chapter digs deeper into them. We will explore the multiple data virtualization approaches and how each works to optimize performance. We'll also look at other tools and techniques that work collaboratively or as part of the data virtualization software to further improve data speeds.

Approaches to Data Virtualization

A data virtualization tool needs to connect to multiple sources at once to provide a holistic view of your company's data. The data needs to be translated into business language and combined with other sources, often resulting in extremely large, complex, and time-intensive queries. As a result, many data virtualization vendors rely on the underlying data sources to handle the data processing. Instead of the data virtualization tool running the query and returning the results, the tool pushes the processing down to the underlying data source system. Even then, many queries rely on multiple sources at once, or depend on data sources with limited or no existing processing power (such as static files). To understand how data virtualization tools handle these situations, let's first examine the two approaches they use to implement the virtual layer:

Specialized data virtualization engines

This type of implementation relies on an external component that manages the connections to all the underlying data sources. It analyzes incoming query requests and determines which data source contains the requested data. When multiple tables from different sources are requested, this type of data virtualization will take control of the data federation and optimize the query based on its purpose (Figure 6-1).

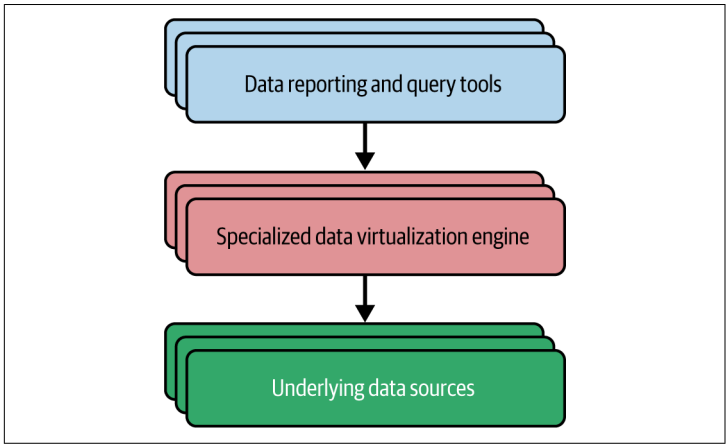


Figure 6-1. A specialized virtualization tool handles all the transactions between the querying tool and the requested data.

Data engines with virtualization extensions

This type of implementation utilizes virtualization as an extension of an existing data source. Many data lake engines that use massively parallel processing (MPP), such as Apache Spark and Trino, employ this type of architecture to manage internal queries and to connect to external data sources. When data is requested in this type of implementation, the worker nodes are responsible for querying the external sources and directing the result into the MPP engine (Figure 6-2). This engine is a system of processing nodes that can run multiple tasks simultaneously and independently.

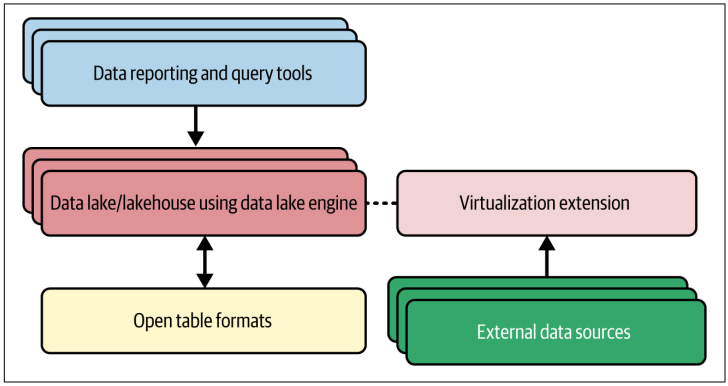


Figure 6-2. A data engine with a virtualization extension pulls data from external sources to process utilizing its own internal format.

The next few sections explore how each of these approaches impacts query performance in different ways.

Specialized Data Virtualization Engines

The major difference between data virtualization and traditional data management is movement. In traditional management, the data is copied from the source systems and placed in a central repository, such as a data warehouse, for reporting and analysis. Data virtualization engines do not need to move the data. They only hold metadata of objects such as tables, views, and stored procedures. When data is required, the virtualization engine instead relies on the underlying data source to perform the queries.

The metadata stored by the virtualization engine includes details about the underlying data sources, such as the data type, the vendor, and the version. The metadata can also contain information about the data statistics, including cost estimation, configuration, pagination in APIs, and more. This metadata is important to query speed, as the virtualization engine uses it to determine the optimal methods for pulling data from that source. Unlike a data engine with extensions, some specialized data virtualization engines also connect to a wider range of data sources, including databases, data warehouses, data lakes, APIs, SaaS applications, and multidimensional databases. Most data engines with virtualization extensions are restricted to relational databases and NoSQL.

The structure of a specialized data virtualization engine allows it to improve performance through continued analysis. Virtualization can optimize the query in two ways. The first option is rule based, which is simply a set of predefined rules that it can use to optimize the query. The alternative is cost-based optimization, which uses statistics from the data source to determine the optimal result. Cost-based optimization utilizes the metadata from the data sources to determine which form of optimization to apply when returning the query (Figure 6-3).

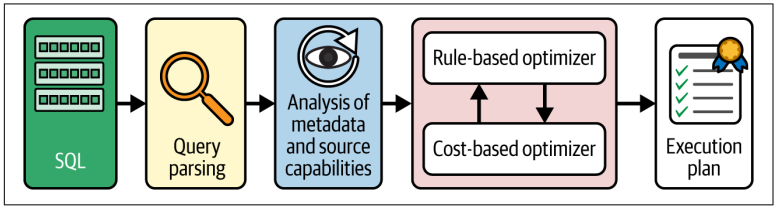


Figure 6-3. When a user initiates a query, the virtualization engine parses the query and analyzes the metadata to determine the most optimal method for returning the data.

Optimization can happen in a couple of ways, depending on the location of the requested data:

Queried data comes from a single system
: In this method, all the data resides within a single underlying source. The virtualization engine can determine whether it is better to leverage the power of the underlying system to process the query or to run the query itself. In many cases, the most efficient method is to let the underlying source execute the query;

however, some sources lack the power to do that execution (Excel, JSON, etc.). In these cases, the virtualization engine will bring in the data and perform any additional operations. This is called *postprocessing*.

Queried data comes from multiple systems
: It's regularly necessary to pull data from multiple systems to provide the information requested by the user. In these cases, the virtualization engine needs to not only look at the underlying sources for optimization, but also determine the best methods for operations, such as joins, blends, transformations, and aggregations. In this method, the cost-based optimizer plays a significant role. It utilizes partial data sets, data source statistics, and other data source details to determine the cost of execution. The optimizer identifies multiple methods for performing the query and then selects the most optimal of the options.

The data virtualization engine can combine the two methods, pushing down some of the operations to the underlying sources while performing other operations postprocessing. When this method is used, the engine creates execution branches to pull the data from the underlying data sets and combine them. The optimizer may iterate through multiple branches by using smaller data pulls and information about the sources to optimize the query. The final result is then usually compiled by the virtualization engine to return the final result to the user.

The easiest way to understand this is to look at an example. Imagine you're in charge of a large medical facility. You have data about patients and visits that needs to be combined for analysis. The patient data set is in Oracle and contains 3 million rows, but the visit data resides in Amazon Redshift and is significantly larger, with 200 million rows. Your virtualization engine will need to combine the data sets and aggregate by patient. The optimization engine explores multiple options for combining and aggregating these data sets (Figure 6-4).

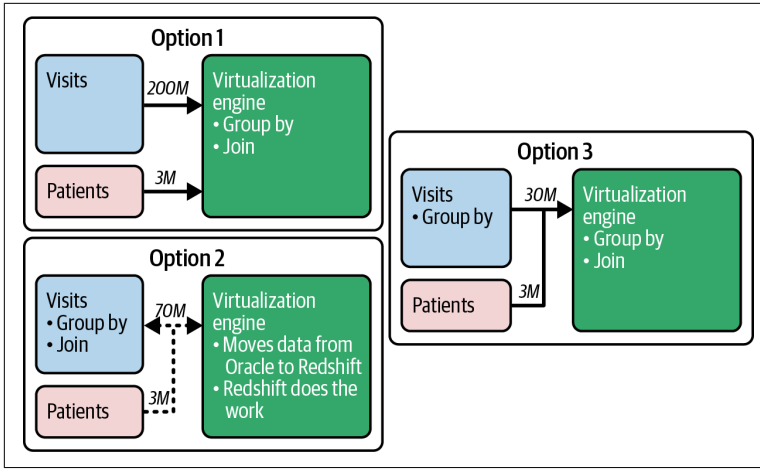


Figure 6-4. Potential methods for pulling, joining, and aggregating patient and visit data.

Option 1

The entirety of the data is pulled from both tables. The virtualization engine then groups the data and joins the tables together to provide the result. With this optimization plan, the virtualization engine does all the work of joining and aggregating the data. The bottleneck is the transfer of all that data over the network, significantly slowing the process.

Option 2

The virtualization engine takes multiple steps. First, the relatively smaller patient data set is extracted from the Oracle table and placed into a temporary table in Redshift with the visit data. The virtualization then pushes down the grouping and joins to Redshift. The Redshift application is able to quickly aggregate and join the data, and returns a relatively small amount of data back to the virtualization engine to then return to the requestor.

Option 3

The data is still delivered to the virtualization engine by both underlying data sources; however, the virtualization engine pushes the visit data aggregation to Redshift. It is grouped by patient before being sent over the network to the virtualization engine. This means that a significantly smaller amount of data is transferred over the network, and the data virtualization engine can then join and group the data before returning it to the requester.

Each of these options accomplishes the task of joining visit data to patient data and aggregating the results. Option 1 is much slower than the other two options specifically because of the amount of data that needs to be transferred over the network. The other two options are significantly faster, and the data virtualization engine will compare the two approaches before choosing the optimal solution.

The difference between all these options in execution time could be milliseconds or could be hours, depending on how much optimization can be achieved. The role of the virtualization engine is to determine which will provide the fastest results. Because of HIPAA, health-care data can't always be copied to different locations because of security and privacy concerns related to patient information. This factor should also be taken into consideration when deciding between these options.

Data Engines with Data Virtualization Extensions

An alternative to the specialized virtualization engine is a normal data engine with a virtualization extension. This approach is most commonly used by data lake providers. While originally created for the Apache Hadoop system, it quickly evolved to adapt to cloud and other object storage formats. Data lake engines decouple the processing from the storage of the data. They rely on MPP to connect and process many actions at once. Because the processing is separated from the storage, it can incorporate connectors to other systems, such as external relational databases or NoSQL platforms.

The coordinator node takes the query from the client and parses it by using the information in the metastore. The node uses its optimizer to develop an execution plan, which splits the initial client request into multiple smaller requests that can then be run independently and simultaneously on one or more worker nodes. This parallel processing reduces query time by having multiple tasks happen concurrently. Each worker is responsible for its own portion of the data pull and processing, and the coordinator directs and tracks that responsibility. The nodes, illustrated in Figure 6-5, are listed here with their responsibilities:

Metastore
: This is a database containing the metadata information for the objects being requested. The database contains information such as columns, data types, partitions, and performance statistics.

Worker nodes
: These are the powerhouses of the data lake. They complete the collection of the data from the source(s) and process the data. Each node works independently through the direction of the coordinator node to pull data, process it, and return results based on the query and direction of the coordinator node. Workers are able to communicate with each other as well.

Coordinator node
: The coordinator accepts the initial request from the client application. It then uses statistics and metadata from the metastore to parse the query, develop the execution plan, and instruct the worker nodes. It is also responsible for returning the results of the query back to the client.

Object storage
: Data sets are stored here. This is a distributed file system that's often in the cloud, and the files are stored in formats specialized for analytics, such as Parquet, Optimized Row Columnar, Delta, or Iceberg.

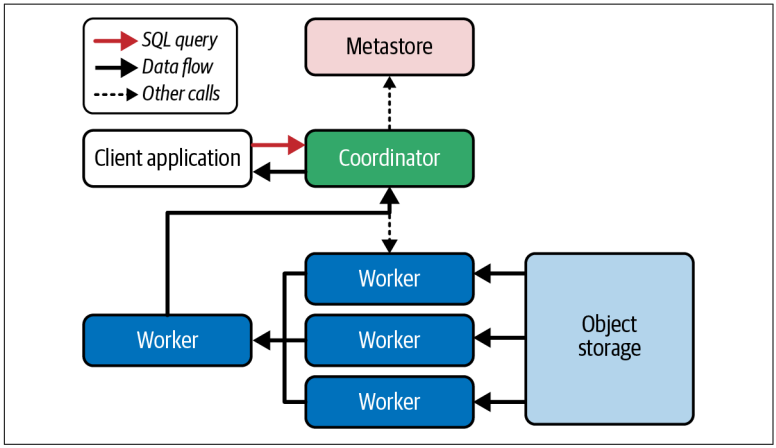


Figure 6-5. In this simple data lake pipeline, data objects are stored in the object storage repository and usually take a format specialized for analytics.

With external data sources, the transactions work similarly. Because the virtualization is an extension of a regular data engine, it is able to utilize the power of the data lake MPP engine and its parts. Unlike data within the lake, the data is not in a specialized format and instead resides outside the lake. Each worker node connects to one external data source and pulls the data in for processing.

This approach can be positive or negative. The workers can independently connect to separate data sources and pull in data at the same time; however, the resulting pulls are very similar to Option 1 in the specialized data virtualization engine. If the data files are large, the process becomes bottlenecked by the bandwidth of the network. Additionally, data lake optimization engines are designed to work with data within the lake; they do not offer the degree of federation that a specialized data virtualization engine can. The data lake can split a query against Parquet files between multiple workers, but the external data sources are restricted to one worker each.

Hybrid Data Architectures

Some vendors have started integrating both types of data virtualization architectures into a single approach, and it is not difficult to see why. The data lake coordinator node behaves similarly to a specialized data virtualization tool. The node directs underlying workers similarly to the way a specialized data virtualization tool pushes down actions to the underlying data sources. In fact, some data lake vendors have started advertising improved push-down capability as part of their enterprise releases.

Unfortunately, the push-down capabilities of data lakes are often limited. While they can perform push-down on simple predicates, such as some filters in a WHERE clause, they often lack the SQL dialect translation necessary to integrate with many external sources. They also lack advanced optimization techniques—such as Options 2 and 3, discussed previously—for specialized data virtualization engines. This means they often need to bring the majority of the data through the network at query time, which is typically unacceptable for large data volumes. Additionally, connections to external sources (e.g., databases outside the data lake) are channeled through query-based interfaces (using technologies like Arrow Flight or JDBC). This means the data lake engine cannot read

the data in parallel using several workers, as it can do when reading from its own data storage.

Conversely, many specialized data virtualization vendors will advertise their use of MPP engines. This processing power enables them to take better advantage of integration with data lakes while also using parallel processing earlier on in multisource queries. Denodo, for example, embeds an open source data lake engine as a built-in component. It allows the tool to access data lakes but also provides power for use in federated scenarios where MPP can be employed.

Other Optimization Techniques

Logical data management platforms utilize other methods to improve performance as well. Each method has a benefit and a place for use. Caching is best suited for frequently used raw data, whereas aggregate awareness benefits more frequently used summarized data. Some data virtualization systems, such as Denodo, utilize AI to recommend an option. Some examples of optimization include:

Caching
: Data replication can be used to create performance improvements and to protect operational data sources from complex queries. In the context of logical data management and data virtualization, *caching* refers to replicating all or part of the data of a certain data view in the logical data layer. Advanced logical data platforms allow you to create and automatically refresh these copies either incrementally or in full without having to create any code or pipeline. When all or part of the data used for a query is cached, the execution engine will use the cached copy instead of accessing the original data source.

Aggregate-aware acceleration
: A smaller data set will always perform faster than a larger one. Aggregate-aware acceleration is the idea that intermediate aggregation can be the basis for calculating an overall result.

  Advanced data virtualization platforms utilize data summaries, which are precomputed intermediate aggregations that can be used as a starting point to compute many queries that follow a similar pattern. The concept is similar to materialized views in databases like Oracle.

For instance, if you precompute the total sales by customer and day in a data summary, the optimizer will be able to transparently use the summary as an intermediate step to answer queries such as total sales of VIP customers by week or total sales of customers from Arizona by month in the last year. With large data volumes, this can be orders of magnitude faster than computing the queries from scratch.

Data transfer and serialization/deserialization
: Transferring data between systems creates another opportunity for optimization, especially in slow networks. Systems can be configured to optimize data compression and properly manage data chunks to improve performance. Serialization and deserialization is the process of converting the data into a format that is easily transferable, transferring it, and then reverting the format on the other side.

  Unfortunately, this process of serializing and deserializing in itself can be a bottleneck. Some new protocols have entered the data tool marketplace that are making an effort to create a standard, limiting the amount of serialization and deserialization. One example of this effort is Arrow Flight, which is a protocol used to decrease the amount of unnecessary serialization in tools such as BigQuery, Denodo, Snowflake, and some Python libraries.

Conclusion

Multiple principles and techniques are associated with a logical layer that can be used to improve performance. In fact, when properly implemented, a data virtualization engine can have minimal overhead and negligible impacts on performance over direct queries against the data source. Implemented correctly, it is possible to access data across multiple sources without needing additional replication.

The right optimization strategies and configurations enable your business to utilize a virtual layer on your data that is easy to use and understand while also being optimized and responsive. Additional techniques, such as caching and aggregate-aware acceleration, can even improve the performance of the original data sources.

CHAPTER 7
Data Governance and Security

Data is the lifeblood of your organization. Without it, you lack the ability to make informed decisions, improve the efficiency of workflows, or avoid potential losses and obstacles. But data is influential only if it is accurate, timely, and accessible to the people who need it. To enable this, you need processes in place to maintain high-quality standardized data, enforce well-defined data security policies, and protect customer and employee privacy. In this chapter, we will take a look at how logical data management provides your business with the tools to govern, secure, and deliver that data. You will understand how various teams within your organization will utilize these tools to not only ensure the security and accuracy of the data, but also continuously monitor the flow of data to avoid interruptions or mistakes.

This chapter is divided into three key components. First, you'll dig deep into how traditional data governance is defined and applied. You'll see how access, authentication, and authorization work together to ensure that the right data is available only to the right people. You'll also see how logical data management implements these ideas to keep your data secure.

From there, you'll explore how different groups in your organization work together to ensure that your data is not only secured, but also accurate, reliable, and accessible. You'll do this through an example of a federated governance model. It will step you through the process of data access authorization from request to eventual approval (or denial).

You'll get a glimpse into how all these systems of governance are maintained and monitored. Decisions are made based on the data provided, so users who access the data need to be sure that it is accurate, reliable, and complete. The operation and observation of your data ecosystem is key to keeping it running smoothly and efficiently. This monitoring is especially vital as your business works toward a self-service initiative. Finally, you'll see how continuous integration/continuous delivery (CI/CD) work together as a pipeline to migrate data from requirements gathering to release.

Security in Data Governance

If you research *data governance* online, you will discover a wide variety of definitions and ideas. Some sites list three major components, while others reference four pillars, and still others describe the five C's. In fact, *data governance* is an umbrella term used to describe a set of concepts and components that revolve around securing access to, and maintaining the quality of, data from its origin through when it is finally disposed of or archived—the entire data life cycle. There are many steps in this life cycle, and many of them revolve around access.

Governance has multiple parts that make it function, which is why it is sometimes hard to define. Normally, it is thought of as the guidelines for your business with regards to the way data is managed, and it covers details such as privacy, security, data quality, and delivery. A well-designed governance plan will ensure that your company has processes and procedures developed to handle data delivery and also that it meets regulatory requirements, has protection from breaches, meets business standards, and operates as efficiently as possible.

A major aspect of governance is data security. Data security controls who can access the data and what they can do with that access. This is also known as *authentication* and *authorization*. Recently, access control policies have been added to authorization, which further evolves how control is managed. All these components play a key part in the way data is secured in logical data management applications. Let's define a few key terms:

Authentication
: Authentication is access to data, and the most basic form of authentication is a username and password. Traditional data platforms have relied on the global systems within the organization for authentication. Examples of these include Active Directory and single sign-on technologies such as Kerberos. As data has shifted onto the internet, authentication methods have shifted toward identity providers, which are services that store and manage user identities online. These services utilize protocols such as OAuth, OpenID, and SAML to provide a more granular approach to single sign-on. Logical data management vendors integrate into a company's authentication platform to embed themselves as part of the data ecosystem.

Authorization
: Whereas authentication gives access to reach the data, authorization controls what the user can do with the data after they have access. The main rights associated with data include levels such as read, write, or delete. Additionally, there are usually levels related to administrative roles that control access to the data and role management and system configuration. Many systems also provide more fine-grained access, allowing the business to map roles to responsibilities, which separates the stewardship and management of the data. This means that different teams can have different roles related to their responsibilities with the data.

Authorization with security policies
: Authorizing user capabilities table by table can be tedious, time-consuming, and prone to error. As a result, access control policies have increased in popularity. These are rules utilizing business terminology that dictate how data can be seen or used within a table. They use words or tags to define what access a role or roles have to the table. For example, data with a #ssn tag might have an access policy that requires users to have a certain role before they can access it. Data access policies can go beyond that, though. They can be used to obfuscate, mask, or encrypt data as well. Users who do not have the appropriate tag would see **** instead of a Social Security number, for example.

The benefit of data access policies is that they can be centrally managed. Rather than going from table to table to change authorizations, administrators can instead change the policy centrally. The change then percolates down through all the tables and fields where the tag is applied. This makes data access not only easier to change, but also easier to manage and audit. Centrally managed policies also reduces the number of opportunities for errors.

Logical data management platforms rely heavily on authorization with data access policies, as it is extremely easy to define and enforce them centrally. The policies can be applied on both operational and information data sources together, which enables consistency throughout the data life cycle.

Access to source data

As mentioned in earlier chapters, logical data platforms provide a layer between the source data and the end consumers. Therefore, two levels of authentication are involved in getting data from the source to the data consumers. The logical layer is accessed by consumers through the processes we just mentioned, but how does the logical layer authenticate to the source data? There are two approaches to this challenge—a service account or pass-through credentials.

A service account is a generic account that is created to access the source data and expose it in the logical layer. Security in this configuration is applied to data through the logical layer. Comparatively, the pass-through method takes the credentials used to access the logical layer and passes them through to the underlying source. This can be complex to configure, but can also be beneficial if the organization already has established privileges in place for the source data through single sign-on techniques.

The strategy a company uses depends on many factors. Does the company already have established rights to source data systems? Does the source data system work with single sign-on? As a result, many companies adopt a hybrid approach to security with logical data management. Security still exists at the source data, but policies and rules are applied at the logical layer to create consistency with data access regardless of the source data capabilities.

Other Data Governance Capabilities

Data governance goes well beyond security. In fact, many separate parts work together to ensure that the data is delivered accurately to those who need it to make decisions. Many of these components are relevant to a logical data strategy, data virtualization, data catalogs, and marketplaces.

Data lakehouses provide many of these governance components in a very basic format. The idea of curated zones is inherited from data lake usage patterns. Logical layers go beyond this basic support, enabling all these features as part of the data ecosystem. This helps organizations build an environment of data self-service, providing a guided, curated approach to data access rather than forcing users to stumble through raw data sets. Components of data governance include the following:

Data documentation
: Documentation is the best way to illustrate to end users the meanings behind the fields and tables available for use. Data stewards or subject matter experts are relied on heavily to ensure that the data they share is used correctly and accurately.

Data lineage
: It is important to understand what a data field is and what it contains. Part of this understanding is achieved by tracing the data back to its source. Data lineage tracks where the data is sourced as well as how it has been changed or manipulated prior to delivery to end users.

Data quality and data profiling metrics
: Data will inevitably have problems with missing values or malformed data. Data quality and data profiling metrics are ways to measure that accuracy and completeness. This helps the end users understand how high a data set's quality is for reporting purposes. It can take many forms. Here are some examples:

  Medallion (gold, silver, bronze)
  : Most commonly used in the context of data lakes, in which raw data is often copied to the lake's storage and then goes through one or more transformations before being deemed ready for business use. *Bronze* refers to raw data that comes from original data sources without any modification and that may have quality issues. *Silver* refers to data that has

been cleansed and deemed accurate. *Gold* is additionally transformed to be meaningful to a business user. For example, an executive report should be based on a gold-standard quality of data, while the data analyst producing the report should be able to trust that its underlying data is "silver," meaning it is clean and accurate.

Numeric
: Numeric metrics may indicate a ranking or exact number of fields that have missing or incorrectly formatted values. An example is a numeric value on customer data to identify the number of rows with incorrectly formed zip codes.

Percentage
: A percentage may indicate to the end user what portion of the rows of data is complete for certain fields. An example is marketing data that is missing some email addresses.

  The values should be a clear indicator to the end users of the data's quality, or a measure of the number or percentage of rows that are lacking data. This will help the user make better decisions about which data to use to accomplish reporting tasks.

Classifications
: Classification can help identify data by using business-labeled domains or categories. This helps the end user understand the type of data available and where it came from in terms that are familiar to them. Categorization makes it easier to quickly find the data relevant to a query, versus searching with keywords or scanning table by table.

Endorsements and warnings
: This capability goes together with the data quality and profile metrics. Administrators of data sets can identify how they should be utilized to answer specific queries. They can indicate whether the data set is complete and whether it contains the values necessary to answer specific questions. The capability is an excellent way to communicate across domains to indicate whether a data set is good or lacking. Warnings can also be added to provide a great way for an administrator to indicate to the rest of a business when ETL processes fail or data becomes out-of-date because of other issues.

Data Governance at Work: An Example

The concepts of data governance make sense when listed out, but how do they apply to your business? What does it look like when you manage your business by using a federated governance model? Let's take a look at a hypothetical example to see how the components of data governance work together to get the right data to the right people.

Imagine you work for a large retail organization with hundreds of stores, multiple domains, and, most importantly, large quantities of data. The domains within your business all govern their own data, and they treat it like a data product. They ensure that it is curated, documented, and classified according to the needs of the business. They are also very wary of how their data is being used and are sure to secure it, requiring the right authentication and authorization to use it.

In this example, you are a marketing director for your region, and it is your responsibility to create a marketing campaign targeting young-to-middle-aged individuals with your company's latest athletic gear. To ensure you reach the right customers, you need data. So, you dig into your company's data marketplace, researching the data sets to identify those that might provide the customer list and demographics you require. As you search the marketplace, you look through lineage, endorsements, data quality metrics, descriptions, and more. After a bit of searching, you land on a customer data set that seems to meet your needs. Unfortunately, it's managed by a different department, plus there are strict access controls because of privacy regulations, and so you do not have access to the data.

Since you are browsing data sets in the marketplace, you can simply request access as you browse. The governance and security settings on your company's data kick in, and the manager of that data set is notified that you would like access. This person will evaluate your request, your requirements, and ask any questions they may have about your needs. As you are a marketing director for a specific region, the security manager may restrict you to accessing information for customers within your region, and redact sensitive and personally identifying information. This is easy for them to do, as the data is well curated, tagged, and contains a column designated for the region.

Once you have access to the data, you use the well-labeled fields and documentation to identify an age field. This will be perfect for identifying customers within the age range you desire! Unfortunately, when you pull the data, you notice that the results for that field are all asterisks. The documentation on the data indicates that the age field was redacted because of global IT policies, as it could be used to directly identify customers in the data set (Figure 7-1).

Instead, it is recommended that you use the customer age range field. It seems like you are not the first person to look for this information. When you pull the age range field, you see it is divided into age ranges of 10 years, which will still work great for your marketing needs while maintaining privacy!

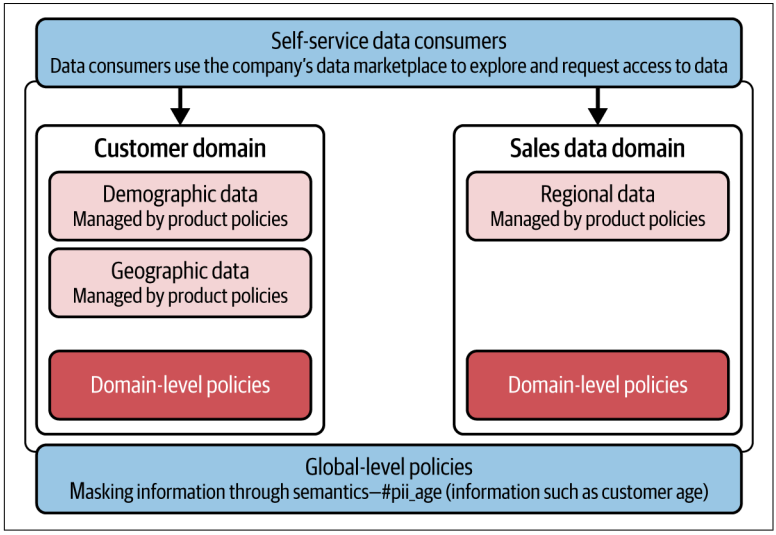


Figure 7-1. A sample federated governance model. The consumers can easily browse and request data. The access and authorization that is granted or denied to that consumer is managed through product, domain-level, and global-level data policies.

Data Governance in Real Time: Operating Your Platform

You've seen how security and governance work hand in hand, but there are more technical aspects to making data flow smoothly and securely through your business. Once you have a strong, well-established data environment, you need safeguards to ensure it stays

that way. A strong monitoring practice is relevant to many roles and is an intrinsic part of modern data strategies. You need insight into not only who accesses the data, but also from where and how. You also need to track data performance by using metrics about CPU, memory, and network traffic. Logical data management is one way to enable this monitoring, to govern and secure disparate data sources in real time.

Traditionally, these tasks have been relegated to the technology and operations teams, as they need to understand the systems and be able to react quickly to potential issues. In more recent years, the responsibility of monitoring data within an organization has shifted somewhat. Data stewards, data security teams, and others have started stepping into the role of data monitors for several reasons. Let's look at these roles:

Data stewards
: Data stewards ensure that data is accurate, secure, and compliant. They understand where it comes from, how it's created, and how it should be used. They are able to identify which data is heavily used, which is popular, and which is not. This can help the data domains understand which of their data sets need more thorough documentation and which could potentially be deprioritized. More recently, data stewards have also taken on the role of data product owners in those organizations that have adopted data mesh and data product principles.

Data security team
: Metadata provides a wealth of information when properly monitored. Data security teams can watch the usage of data sets to identify patterns and search for potential data breaches. They can also track requirements and report to regulatory agencies with regards to the enforcement of regulation compliance.

Financial operations (FinOps)
: It may seem odd to see this team listed in an area that is traditionally technical; however, financial teams have a growing responsibility when it comes to data. As the traditional on-premises data storage is slowly replaced by on-cloud solutions, the financial team is responsible for tracking pay-per-use ecosystems. It is often their responsibility to track data consumption, anticipate data growth, and manage potential steps to keep costs at bay.

Where does logical data management fit into the world of monitoring and observing your data? It is the middle layer between the source data and the consumer, which means it has both the monitoring and the management tools necessary to easily track usage across the data ecosystem and respond to potential issues. Logical data management offers insight into the data demands and reporting needs of the organization. It can use metadata to identify the users, reports, or data sets that might potentially increase costs. Logical layers can dynamically restrict access when certain criteria are met, shifting processing and workflows to less expensive alternatives without the consumer being aware.

Development Operations and CI/CD

The final component of governance and security is development operations. This is the development of software or data products and the path they follow from creation to production to release. The agility of the data teams is heavily dependent on the agility of the methods used by the data engineers. To build a successful logical data platform, a company needs to adopt a development plan that is both extremely agile and managed by the practices of CI/CD.

CI/CD is a set of practices that ensure that development of new data products is both efficient and speedy. Instead of one large project that takes many years to complete, coding is done in smaller chunks and produces smaller, more frequent releases. CI/CD is comprised of three key fundamentals (see Figure 7-2):

Continuous integration
: Data engineers do not work in a silo, but rather build on one another's work. To be successful at this, companies need to adopt practices similar to those in coding software. They need a data platform that enables version control. This platform needs to support branch-based development and features that enable cross-team collaboration. This enables data engineers to work collaboratively and simultaneously on one or multiple projects. Most modern logical data platforms support these capabilities and integrate with other tools in the ecosystem.

Continuous delivery
: Any developer will tell you that the key part of any code is testing before release, and data engineers are no different. The development of new data products should include development

testing as well as end-user testing. If possible, automated testing can be employed to improve turnaround time on certain developments.

Continuous deployment
: Once a data product is developed and tested, it is ready for release to production. Part of the development operation is this release process. Migrations often span multiple systems and include multiple sources, lakehouses, reporting tools, and logical models. All these systems need to provide endpoints for migration and testing. CI/CD frameworks such as Jenkins provide an organized method to manage the release processes.

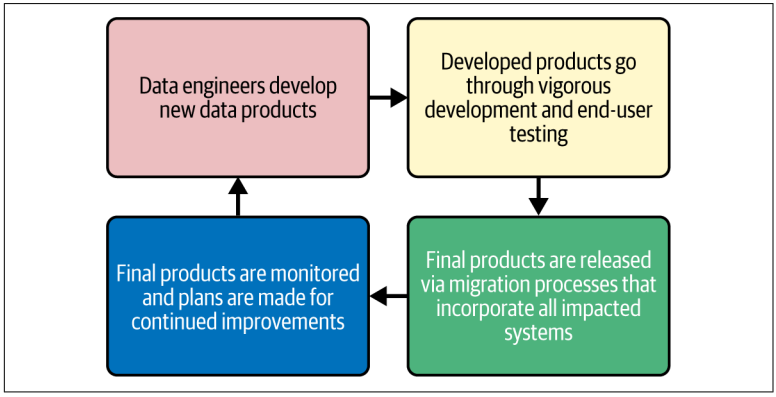


Figure 7-2. A simple model illustrating the planning, development, testing, and deployment of new data products.

Modern companies need to consider their development processes as part of their logical data environment adoption. These processes need to be agile and follow CI/CD requirements to ensure that data products are developed as quickly and accurately as possible. This will enable the company's continued data improvements while limiting the number of IT bottlenecks.

Conclusion

When you consider a logical data management platform for your business, you need to keep security and governance in the forefront. You need to ensure that the data you create is accessible to those who need it and restricted from those who do not. At the same time, you need a tool that will ensure you can easily track who has access

to your data and what level of access they have. A logical data management platform should enable multiple domains in your business access to participate in the management and flow of data. Finally, you need to remember that an agile process will ensure that new data products are produced at a steady pace and that improvements can quickly and easily be made.

CHAPTER 8

Logical Data Management and AI

The business world evolved with the advent of artificial intelligence, and it has likely impacted your company as well. Generative AI (GenAI) is a subset of AI that focuses on creating content such as text, images, and code. The arrival of large language models (LLMs) and GenAI reflect the growth of the technology. Businesses are using AI to improve innovation and integrate automation in their production processes. They are using it in diverse ways and through unique and different use cases. But how do logical data management and AI work together to improve your business data environment? In this chapter, we will explore the symbiotic relationship between AI and logical data management.

Logical Data Management as a Support for AI

AI and data are closely related. The technology relies heavily on data to fuel its engine. It relies on vast repositories of data that it uses to derive patterns, relationships, and outliers. The data that builds this knowledge base is known as *training data*, and the quality of AI insights is dependent on the quality of the training data fed into the system. The larger the volume and higher the quality of data that is fed into the training data set, the more accurate the predictions and decisions that AI produces based on this data. The more diverse and comprehensive the training data, the better AI will be at predicting new, unseen data and results.

A logical data layer creates a platform that delivers the trustworthy data on which LLMs depend. This ensures that the data powering the AI engine is reliable and useful. Likewise, a logical layer consolidates the large variety of data sources and domains within your organization under a single structure, simplifying access and configuration. This section will dive into two of the many areas where logical data management provides the foundation needed for powerful AI implementation.

LLMs contain encyclopedic information about all historical events, all literature ever written, and more. Unfortunately, they have a weakness: while they contain this vast wealth of knowledge, they have little to no information about your organization specifically. They do not know your service or production processes, customers, employees, or products and services without additional integration and effort. This raises a question: how do you inject that business-specific information into a GenAI application? How do you supply the information it needs about your organization to enable it to provide valuable insights and support specific to you?

Logical Data Layer to Power Retrieval-Augmented Generation Frameworks

In 2020, Meta released a paper discussing a technique called *retrieval-augmented generation* (*RAG*), which sought to address this challenge in a safe and effective way. RAG enhances LLMs by integrating with an organization's internal data sources, such as CRMs, product databases, and more. It routes the query through a vector database to identify additional sources of information related to the query. It then takes these results and shares them along with the original question back to the LLM. A vector database is simply a specialized database that stores data with numerical representations to enable similarity searches and faster data retrieval.

The easiest way to understand how this works is through an example. Imagine a child in school whose teacher has asked them to write a report on Greek mythology. The student reviews the assignment and then approaches their parents for help. The parents know a bit about Greek mythology, but not in enough detail to help the child properly compose the report, as seen in Figure 8-1.

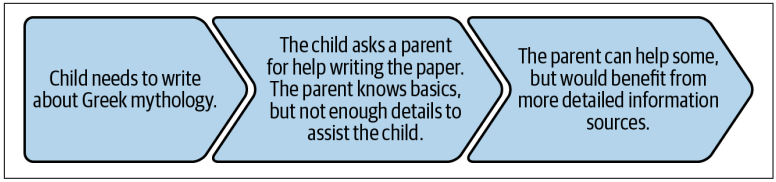


Figure 8-1. In this example, the parent is the LLM. They have knowledge that can help the student write the paper, but would benefit from another, more detailed and accurate source of information.

A RAG framework takes the information search one step further. RAG is similar to the parent looking for resources on Greek mythology on their bookshelves. They utilize these books to find the best, most detailed information for the report and add context to help the student with the assignment. With these books, they sit down with their child and help them write the paper. This additional information source would provide the student with more information to write a much more thorough and detailed report about the topic (Figure 8-2).

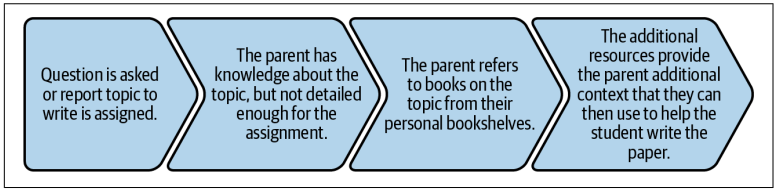


Figure 8-2. A RAG framework uses vector database sources external to the LLM to improve accuracy and context before delivering a query to the LLM itself.

With the power of AI and RAG, your company can enable conversational business intelligence (BI), the ability to ask questions in natural language and get responses in a similar format, creating a direct conversation with your data. For example, you can ask, "How many laptops do we have in stock in the warehouse?" The LLM can augment the response (with RAG) by using information from the organization's inventory data to provide a natural response, just as if you were speaking to a colleague in the hall. You can then dig deeper into that information to produce charts on historical levels of laptops in the warehouse, sales history, or supplier information. This capability is the doorway to true data democratization. Anyone can ask a question and get a result without needing knowledge of the

data structure, a query language, or an analytics tool. They simply ask a question and get a result.

Unfortunately, implementing such a model is no small task. As companies try to embrace conversational BI, they've landed on two approaches. Each has its benefits and drawbacks:

Data replication in vector databases
: This technique is the easiest to implement. It enables contextual searches associated with RAG processes by indexing content within a vector database. This works great for images and PDFs, whose content can be included in your responses. However, it has significant drawbacks when applied to structured data. For example, it lacks the fine-grained security control often found in traditional databases. Additionally, common analytical questions that require joining or aggregating data points across very large data sets are better addressed using SQL queries than using this technique.

Text-to-SQL translation (query RAG)
: This method relies on a different idea, which is translating your natural language questions into SQL queries. For instance, "How many books do we have on Greek mythology?" can be translated into `SELECT count(*) from books WHERE topic = 'Greek mythology'`. This overcomes some of the shortcomings of the previous approach but also introduces challenges. Often, this approach can be applied to only a single data source. However, many questions you ask may rely on operational data, the application, or third-party systems. The reality is that most often this information will be in multiple systems, making this approach hard to implement.

Can logical data management help your organization achieve that data democratization and conversational BI? Fortunately, the answer is yes! Logical data management can bring some of the benefits of both approaches while providing a solid foundation to address their unique challenges.

On one hand, logical data management supports RAG by organizing, cataloging, tagging, and vetting data from all your sources. Logical data management is sort of similar to a card catalog or a book database with details about each book in the library. It provides rich metadata about these books to ensure that the search returns results pertinent to your research and not things tangentially related.

Within your business, you are utilizing business-specific data to augment your query.

On the other hand, it enables strong text-to-SQL functionalities, but unlike other systems, it can make it work across your entire data landscape. You're not tied anymore to a single data source—you can ask questions about any system, or even ask questions that require going across multiple sources.

In the following section, we will look at these benefits in more detail.

The Role of Logical Data Management in Conversational BI

It should come as no surprise that a logical data layer is ideal for developing the data management foundation that AI applications depend on. A semantic layer provides context against structured data views. This enables AI applications to find the appropriate data view and ensures that the LLMs can access all the data they need regardless of the underlying source. This allows the organization to develop a single, secure location for implementing text-to-SQL integration. A logical data management solution benefits the implementation of AI in numerous other ways as well:

Single point of access
: LLMs need to query all the data within an organization to provide a holistic representation. A logical data layer provides that single point of access, enabling text-to-SQL pipelines against business systems, warehouses, data marts, and more.

Semantic layer
: A logical data layer translates operational terminology into business-related syntax. This means that LLMs based on a logical layer have access to business context and knowledge through business definitions, categories, tags, sample values, and more.

Reduced replication
: The logical layer rests above the underlying data source, reducing the need to replicate data and providing closer-to-real-time data access.

More detailed access and permissions
: The security provided by utilizing logical data management is far more granular and easier to manage than traditional methods, ensuring that the correct users have the correct level of access to specific data.

Vendor-neutral
: Logical data layers are able to translate the underlying data sets into a common format. This means that LLMs and the power of conversational BI can be applied to any format of underlying data without the need for migration or translation of the data into another source.

Enterprise-wide data governance
: Logical data management connects multiple data sets and data domains under a single logical layer. This logical layer ensures that the data is well governed, cleaned, and validated, which is vital when establishing AI-ready data.

Faster access to data
: Logical data management is able to sit atop multiple sources, including those producing data in real time, whereas centralized approaches require copying the data into a central location. This means that the data fed into AI from a logical source is more up-to-date and closer to real time.

This list is by no means exhaustive. It simply illustrates many of the benefits provided by a logical data layer in preparation for enabling an AI application. A logical data layer serves as a layer that makes your data AI-ready. The combination of an LLM and a logical data layer is a productive first step in developing an AI implementation and enabling natural language processing within your organization.

AI is only as good as the data fed into it. Bad data results in bad responses and strange results. If it is customer-facing AI, this may lead to a bad perception of the company or even lawsuits. Logical data management as a tool prevents this through translated, validated, and well-structured data within your organization.

Working Toward a Multiagent Future

Researchers and developers are continuously working to improve AI. They are researching ways to make the AI systems more robust, scalable, faster, and efficient. One way they are achieving this is through multiagent systems: collections of AI systems working together to solve more challenging problems autonomously. Each agent can provide expertise in a particular task, and communicate with other agents so that together they can achieve more complex, multistep goals.

Logical data management can play a large role as these multiagent applications become more available. It is invaluable as access to data in a diverse, complex ecosystem becomes more challenging. This is even more evident when requirements like security and governance are added to the mix.

To facilitate this integration, new protocols like Model Context Protocol (MCP) give AI agents a simple, standardized way to plug into tools, data, and services. A logical layer, accessible via MCP, can become a fundamental part of a modern, AI-ready data architecture.

This multiagent approach also has a huge potential as an extension of conversational BI. Today, many organizations struggle to get clear answers to open-ended or interpretive questions like "Why are sales dipping in one region but not another?" or "What's driving customer churn despite high satisfaction scores?" Traditional GenAI tools can retrieve facts, but they fall short when the question calls for deeper reasoning, cross-functional analysis, or correlations that aren't immediately obvious. A multiagent AI-system, on the other hand, can give insights into the reasons behind the dip and what marketing campaigns are recommended to improve sales. These kinds of systems create a research plan that can leverage multiple agents and data queries to address the problem.

This new class of AI capabilities is designed to simulate the work of a skilled data analyst. AI understands the intent behind the question, analyzes context, reasons across data sources, and synthesizes insights that are explainable and supported by evidence. It can go beyond retrieval to explain why things happen; handle nuanced, multipart questions with contextual understanding; and deliver traceable, insight-rich answers ideal for real decision making. This dramatically reduces time-to-insight, while expanding the possibilities for data-driven discovery.

Using AI to Build, Support, and Sustain Logical Data Layers

Thus far, we've explored the power of logical data management in supporting an AI infrastructure. On the flip side, AI can also support logical data management. In this section, we'll explore how AI can boost the capabilities of logical data management as well.

Many people are familiar with the tools available within GenAI. It is built on an LLM, which means that it is based on a spoken language. The terminology and syntax employed against the AI tool utilizes the natural language humans speak every day, making it accessible to nearly everyone. As a result, users employ the LLMs to not only summarize text, but also to manipulate it, summarize it, and create it. The tool can help write a paper or an email, or aggregate notes from a business meeting. It can also be used to provide more natural language tables and field names for building a logical data layer.

AI is also very useful in generating and updating code. Code such as SQL, Python, C, Java, and more are simply extensions of everyday language. As a result, the LLM can understand not only the syntax of the code, but also what it is designed to do, how it is designed to do it, and how it can be streamlined to perform more efficiently. These capabilities simplify complex problems for developers and coders, making their work more efficient. The code capabilities within AI can also be used to automate mundane or tedious tasks, freeing up time for more valuable work. This can also be applied when developing a logical layer atop a wide variety of underlying domain data sources.

GenAI as a Data Management Productivity Tool

There are many ways AI can assist with supporting a logical data layer. Let's dig in further on these capabilities to understand how AI supports logical data management:

Data stewardship / data product ownership
: The tools of GenAI also assist with monitoring and managing the data. The tools can be used to generate descriptions and labels for data sets, tables, and fields. They do this through context, utilizing information from the underlying query, the lineage, data profiling, and more. They can also dynamically tag sensitive data to assist with implementing security policies.

Logical data management simplifies this for GenAI by centralizing the access, but conversely, GenAI can analyze the structure and configuration of that security to ensure that it meets the standards and requirements of the organization. GenAI can also be used for data validation and quality assurance.

Data engineering
: The people responsible for the design, maintenance, and infrastructure of the data processes also benefit from GenAI. For example, it is helpful in defining data models by identifying optimal joins, improving SQL queries, or finding syntax errors within code.

  GenAI can even suggest simplified table and field names and help write documentation. For example, PROD_ID_CD might become Product ID. This translation, together with extended documentation that brings in additional context and details, makes it easier for end users to work with and understand the components of the data without needing to search for definitions or analyze through trial and error.

  GenAI can also be used to generate synthetic data in a logical layer. AI uses existing data combined with statistical patterns and processes to create imaginary data, which it then uses to populate a hypothetical future version of the environment or potential test cases. This synthetic data enables engineers to stress-test the data environment long before it reaches a potentially unstable state. It also enables testing in an environment without risking data privacy, security, and confidentiality.

This list of uses for GenAI is just a small sampling. It simply illustrates how GenAI can be applied to a logical data management system to bring additional benefits, functionality, and agility. Within logical data management, its power goes further. GenAI is no longer restricted to specific systems, but instead it can be applied to all data assets within. This helps enable better stewardship, a more streamlined data pipeline, and improved data democratization across the organization.

GenAI as a Data Transformation Tool

It is important to remember that LLMs can process not only text, but also images, video, audio, and other unstructured data. This means that you can query information in these unstructured assets

by using a query language like SQL. How is this beneficial, and why would you want to do it? Let's look at an example.

Imagine that you have customer reviews for your products stored in a database. Some of those reviews are bad, and you wish to understand what is causing the main pain points in order to improve customer satisfaction. Using traditional data management approaches, building these kinds of pipelines usually involves engaging external tools or APIs and machine learning models in multiple steps, with a data scientist creating custom code in a language like Python. With the power of SQL and the LLM as a transformation tool, you can simply run a query to do that in a single shot! You can use the LLM to extract its sentiment, and to focus on the ones with negative reviews. You can also ask the LLM to summarize the top three causes of that negative impact. Imagine doing that in simple SQL sentences like this one:

```
SELECT reviewername,
summarize_ai(group_concat(';', review))
FROM
    (SELECT reviewername, review
    FROM online_reviews
    WHERE sentiment_ai(review)
    GROUP BY reviewername)
```

Since multimodal LLMs can process documents and images, you can even apply this idea to extract information from a PDF (e.g., "What's the name of the seller in this contract?") or to get the email address of all clients who have uploaded a blurry image, so you can ask them to upload a new one.

Logical data management can utilize the classification and summarizations generated by AI tools. This can include both structured and unstructured data in a wide variety of formats. GenAI facilitates and automates the translation and preparation tasks required to integrate unstructured data into a logical data layer. It can also facilitate data governance and security.

Conclusion

AI and logical data management work hand in hand. Logical data management provides a reliable source of information AI can use to augment responses for queries against LLMs. It also creates a single layer for access to all of your company's data without needing to integrate multiple data types, data structures, and security models.

Conversely, AI can support the creation and management of your logical data environment. AI can be used to analyze unstructured data sources to enable integration with a logical data layer. It can also be used to assist with writing code, monitoring data use and access, and facilitating friendlier data access for democratization. This illustrates how AI and logical data management are mutually beneficial.

CHAPTER 9

Realizing the Benefits of Logical Data Management

Thus far, you've seen the benefits of logical data management from a technical viewpoint. You've seen how it overlays and connects diverse data sets, reduces data replication, and provides a central location for managing data within your organization. You've seen its relevance in data mesh, data fabric, data governance, and GenAI. But what does logical data management mean to your company in terms of improving the experience of your customers, employees, and other stakeholders? How can this approach reduce costs and increase revenue? This chapter digs into the benefits of logical data management from a business perspective to help you understand more clearly the benefits it provides to your organization.

To begin, let's examine the benefits at a high level. There are three major ways your organization can benefit from a logical data management approach: 1) more insights and better decisions through data democratization, 2) better customer experience, reduced risks, and lower costs through improved business operations, and 3) IT infrastructure optimization:

Data democratization
: We've already explored how logical data management can make data accessible to employees at all levels of the organization, but what benefits does that provide? Democratization enables all employees to utilize data to make better and more timely decisions. This approach empowers them to improve their own

productivity, reduce waste, and deliver more value to the business. Simply giving employees access to data is not enough, though. To make data democratization work, the data provided needs to be trusted and understandable by business users. Data must be validated, easy to access and use, and delivered on a timely basis, and all of this done at scale across a large organization with potentially many data users. The major benefit of data democratization is the ability for more people in the organization to generate their own insights and make effective decisions based on the data. Companies that prioritize this score better in almost every business metric.

Business operations

One of the biggest challenges of running a business is keeping the various business functions informed with data. The front office deals with customers and needs information to provide the best experience. Users in the front office are interested in marketing campaign success, customer loyalty, and external influences on purchasing patterns with the goal of optimizing revenue. At the same time, governance, risk, and compliance teams (also known as "the middle office") are concerned with measuring and managing risks to the company. They need information on strategic threats to the market, compliance and regulation, managing costs and financial losses, and business operations. They need data to predict, prepare for, and respond to challenges the business might face during normal operations. Finally, the back office is focused on making the business more efficient. They need data from the entire course of the business to ensure that the goods or services are provided in the most streamlined way possible. An IDC European data analytics, automation, and AI survey found that businesses saw significantly improved business outcomes when utilizing data-driven decision making, use of data platforms, and business automation.

IT infrastructure optimization

Data requires IT resources. It needs storage, processing power, a network, and, most importantly, highly skilled people. It is no wonder that IT infrastructure needs have grown over the years with the increased demand for data. With this demand comes increased loads, more risks, and higher costs, to the point that data management has become a significant portion of

IT costs. Logical data management plays an important role in simplifying this infrastructure, reducing costs, and minimizing risks. It provides a way to unify data from multiple domains, reduce replication costs, provision access, and provide insight into data consumption.

The application of logical data management is pivotal in all these areas, but to truly see its potential, we need to dig into the details.

Data Democratization and Self-Service

The data in your organization is valuable only if it is accurate and accessible. Data self-service enables your employees to run queries independently, develop reports, and generate insights without the delay of relying on a central reporting team. This fosters a culture of data-driven decision making. Each employee has the tools and access required to generate insights on their job and role within the organization. This enables accelerated decision making, innovation, and increased productivity. It also frees reporting and IT resources to tackle more demanding and strategic tasks.

While the benefits of data self-service are numerous, the journey toward it is not without challenges. First, many organizations rely heavily on IT to provide data access and delivery, resulting in increased delays and inefficiencies. Second, the data within an organization is often fragmented, diverse, and distributed, requiring multiple steps and challenges to be overcome with regard to database type, security, and access. Finally, an inherent delay in traditional data management processes creates a lag between data generation and actionable insights. This reduces a business's flexibility and adaptability in reacting to change.

Data self-service can be enabled in multiple ways:

Self-service analytics
: Many organizations enable and encourage its professionals to generate their own insights and make better decisions through data, with minimal IT support. This is called *self-service analytics*. And with the advent of GenAI, conversational BI is a form of self-service analytics that enables all users to become analysts, using only their natural language. Organizations are creating policies to encourage self-service through direct access to the data. However, many users and AI tools may struggle to find,

access, and subsequently utilize the right data, especially if it is distributed across multiple systems they may not be familiar with, or accessible only through technical interfaces they won't have the skills to use. The risk of data leaks and privacy violations increases as more users and AI tools leverage potentially sensitive data. Logical data management solves all these challenges by providing a unified business-friendly interface while also providing the required controls and security.

At the same time, logical data management enables organizations to see the value of adopting a self-service culture through quantitative measures. It allows them to compare business process KPIs against data usage statistics such as the number of users querying the data, the number of queries performed, and the percentage of the enterprise data being utilized. These metrics make it easy to demonstrate which users and what decisions and actions were supported by self-service data, and thus measure its value to the business.

Distributed data product ownership

The employees directly involved in creating the data are often the most experienced in using it. For example, HR employees best understand how to use and report on employee data, and marketing employees best understand marketing campaign data. As a result, the various teams within the business must become more invested in their own data and its quality as it is made available to the rest of the organization. Businesses must encourage and empower those domain experts to take ownership of that data. This was covered in more detail in Chapter 3, where we examined the application of a data mesh.

In this context, the data products' owners are embedded within business teams and are responsible for the entire life cycle of the data, and ideally should be able to do so without IT having to do much of the work. Logical data management provides many attributes that can enable this form of self-service, including data virtualization, a semantic layer, a federated governance model, and valuable insights to data usage.

A newer impact of data products is the ability to measure the value of data and access to data. This is called the *data return on investment* (*data ROI*). In this metric, companies are looking to measure the financial impact of data access and usage in

quantifiable ways. The individual domains are responsible for these metrics related to the creation of data and its storage compared to that data's benefit to the organization. Logical data management can support the measurement of data ROI through the creation, utilization, and monitoring of data products' success metrics.

Intelligent/AI application development
: Application teams also want ownership of the data that powers their applications. Many of these applications are becoming, or are being replaced by, AI agents, with autonomous actions being taken based on events or conditions. The previous chapter discusses AI agents, or agentic AI, in more detail; however, some business benefits are worth mentioning here:

  - Reduced time spent on AI application development related to data integration, preparation, and management. This increase in speed improves deployment times of the AI-powered application and its subsequent revisions, meaning the application is available sooner to provide value to the business through improved revenue, decreased costs, reduced risks, and more.
  - AI-powered applications produce more errors when the data is incorrect, outdated, or missing. That is why an organization needs to build its AI application on fast, reliable data. Incorrect values returned through an AI application may result in adverse consequences, such as customers getting incorrect chatbot responses, incorrect amounts of supply being ordered, a patient being provided the incorrect medication, and other outcomes that can harm the business and its stakeholders. A logical data layer can provide the fast, accurate data that AI tools depend on for accurate responses.

Business Operations

Organizations are looking for ways to operate their business processes more efficiently from end to end. A retailer, for example, wants to ensure that customers have a positive experience at each touchpoint with the business, including mobile apps, websites, and physical stores. A manufacturer wants to optimize its production process by reducing the costs of its supply, and optimizing

production and product delivery. There are many examples of how businesses can benefit from incorporating data in their decision-making processes, but there are also significant challenges:

Unified data platform
: Data within a business is produced at multiple points and often through disparate systems. To see the full picture of an organization, businesses need to connect to and pull data from each. Traditional data management systems sometimes lack the necessary resources to connect to certain data sources because of the database type, the size of the data, or regulatory requirements. Logical data management systems get around these obstacles by not requiring the data to be copied to a central location ahead of time.

Data recency
: Data is more valuable when it's fresh. For example, when your customers are engaging with your website, you want to provide them with options that match their current needs. This means not only pulling their order history, but also tracking their most recent browsing patterns and interactions with you. What are they searching for, and what have they bought in the past? What is something that they may want to buy now as they browse through your site? The same is true for data security and monitoring. You want your security and compliance officers to know about potential threats as quickly as possible to mitigate risk. The same is also true with supply chain or distribution disruptions. You want your operations staff to know about such issues as close to real time as possible, to act quickly and minimize adverse impacts on the business.

  All these demands depend on fast, reliable data. Traditional data management systems rely on a data pipeline that copies data from the source system into a reporting repository. Keeping this data close to real time is difficult and expensive. Logical data management, on the other hand, can layer upon existing source systems, providing a window on near-real-time data as your organization goes about its business practices, without requiring expensive, always-on data replication. This means your business is agile and can adapt to changing conditions while also minimizing the costs of doing so.

Semantics
: Your company is likely built on multiple systems producing data about business processes. This data is presented in a format that best suits the needs of the process, meaning it is probably not in a syntax that is easily understood by those outside the specific domain. Likewise, multiple systems can produce data on the same topic. For example, customer data may reside in both sales and help desk applications, security alerts may come from multiple monitoring tools, and a supplier data set might provide many of the same data points as an inventory data set. But which one is more reliable and authoritative? Finally, if the right data set doesn't yet exist, how long will it take to build a data pipeline to create one?

  When your employees are looking for data, they need clearly defined data sets and metadata that can assist them in finding information. They need a single place to go to find the data they are looking for, and they need the data in that repository to be in a format that is easily understandable and in the context of the business. A logical data layer can help provide that business context through *semantics*, and as discussed in Chapter 5, create a central marketplace that any employee can visit to connect to and analyze data.

These business practices describe how to consolidate data, improve its speed, and make it more accessible; however, how does this help with core business processes? How do these technical improvements increase customer satisfaction, improve business processes, and reduce potential risks? Let's look at how logical data management might impact these key aspects of your business.

Personalized Customer Experience

Logical data management is invaluable when it comes to combining disparate customer data sources, because of its ability to sit upon source systems and provide data in near real time. This enables your business to be extremely agile in many ways:

New and improved customer experiences
: Many things impact your company's ability to sell goods and services. The market can change very quickly because of new innovations and changing economic conditions. Competitors can release updated products or shift costs to challenge your

position in the marketplace. Your organization needs to be able to quickly make changes to your customer experience to respond to these challenges. The time it takes to respond to market changes and adapt your customer experience is known as *time-to-value*, and it is measured by the acceleration of positive revenue impacts.

Logical data management provides your business with the flexibility and data accessibility to reduce your time-to-value. Since the data is available through a central point, has access to multiple underlying sources, and is in a language that is familiar to the business, it takes only a fraction of the time to update your customer-facing experience. This means you can adapt quickly and retain customers while also avoiding challenges that might result in customer churn. Better customer experiences lead to improved satisfaction and more opportunities to increase upsell.

A great example of how this works was the Covid-19 pandemic. Businesses that were able to quickly understand how their customers were behaving in response to lockdowns and other pandemic impacts were able to pivot their operations and satisfy their customers much more quickly. Comparatively, those businesses that needed additional time, resources, and effort to gather the right data and make necessary customer-facing changes struggled to keep customers happy and spending money, or lost them entirely to more nimble competitors.

Personalization

When a customer interacts with your business, you want them to find what they are looking for so you can make a sale and bring in revenue. That's why it's important to have near-real-time access to customer data, demographics, purchase history, and more. When a customer is on your website, you want to draw their attention to things they are likely to purchase. You will not be successful if your website is advertising athletic apparel to a user searching for formal business attire!

You need sentiment, preferences, and demographics from places like social media and elsewhere to develop a personalized experience for your customer. Logical data management supports this by making disparate data sources like social media and customer demographics accessible for GenAI in real time. The

data you track about your customer can work together with GenAI to provide an experience that is tailored specifically to your customer, but only if your data is accurate and recent. Logical data management is able to pull from source systems with improved speed and accuracy. This means that when a customer visits your website, it will provide relevant opportunities based on who the customer is and what they are doing at that point in time.

Personalization of the customer experience needs to extend to the entire customer base, which means the data pipeline it relies on needs to be flexible and adaptive enough to grow as the company's customer base grows. Any improvement in customer experience improves engagement, and it is measurable through things like satisfaction surveys, retention, or simply how much your customers spend.

Measuring and demonstrating business metrics in real time
: Logical data management can source data from the underlying systems in near real time. This means your business has access to data that is optimally fresh and relevant to dashboards and reports tracking your company's key performance indicators. It also means that these reports can stay accurate and up-to-date without waiting for the delays inherent in a traditional data management system. It means your company is more nimble and able to provide better customer service, grab opportunities, and dodge potential obstacles.

Real-Time Operational Intelligence

The benefits of logical data management reach beyond the front office. Back-office operations can also benefit from fast, accessible data. The primary difference is that the back office is focused on improving business processes and efficiency, versus improving customer experience. Real-time data is invaluable as market changes can impact what the company is selling as well as the resources it needs to produce that product or service. There are several places where the back office can benefit from real-time, actionable data:

Optimization of supply chain and distribution
: You need real-time insight into your business's supplies. You need to be able to not only track the resources available to you, but also react to adverse or sudden disruptions, such as natural

disasters or geopolitical issues. This helps you avoid issues like running out of stock when a supplier is no longer operational, and positions you to quickly adapt when issues arise. This, in turn, reduces gaps in your ability to serve customers, reduces operating costs, and increases revenue. Often, tracking is done through real-time dashboards powered through automation and built on real-time unified views of the data.

Manufacturing productivity and automation
: Manufacturers need insights into the production line to monitor processes in real time. Robotics and control systems operate in complex environments with little or no human interaction. This means the machines must be fed data to ensure that they function as optimally as possible. The availability of real-time data is paramount to avoiding issues with productivity. Data collection and availability also need to be flexible, as manufacturing processes change and improve. At the same time, real-time data insights help ensure that the amount of product produced reflects market demand. All these factors come into play in multiple industries. For example:

  - Consumer products: food and household products
  - Complex durable products: automotive, aviation, and medical products
  - Energy production: oil and gas from wells, through refining and into distribution
  - Pharmaceuticals: medicine and other chemical products

  Each of these industries and many others require repeatable, scalable processes to create the product they provide. As part of this production, they need insight into the data around the entire value chain, from market demand all the way back to supply chain and manufacturing processes, to operate as efficiently as possible. Logical data management enables a unified real-time view of this entire value chain, regardless of the different types and locations of applications involved.

Internet of Things and predictive maintenance
: The Internet of Things has become mainstream. Computers and sensors are everywhere, taking readings and collecting data. In manufacturing, they are especially prevalent. Sensors on machinery track everything from production values, to

temperature, to maintenance processes. This data is extremely valuable as it can be used to predict and prevent unexpected interruptions in the process. When a machine overheats, a truck breaks down, or a part wears out, the manufacturing process can grind to a halt. These events are costly to the company and need to be limited in time and frequency. This is where data comes in. Information collected about the manufacturing process and the machines that run it needs to be analyzed for potential issues. Some data can even be fed back into the machines performing the work as a method of automating preventative maintenance. If a machine gets too hot or a product starts to show signs of a defect, the sensors on the machine can detect an issue and pause production to limit the amount of waste and downtime. The data can also be used to alert employees when a situation needs to be addressed.

Unfortunately, such devices are highly diverse in terms of types, data formats, and locations on the network. Moreover, not all outage situations are equal in terms of the downstream impacts on the business. Using traditional data management techniques to integrate and identify the potential business impacts of outage situations, separate signal from noise, and ensure the organization is proactively acting on the most critical situations, all in real time, is very difficult. Logical data management makes this easier by delivering a unified real-time view of such data.

Holistic Risk Management and Regulatory Reporting

Governance, risk, and compliance are all items that need to be addressed by your company. Those responsible for managing these tasks need insight into the business's risks and compliance issues through data, and this is something that logical data management can provide. Like other operational processes, the risk management team needs access to unified real-time views of data to ensure that risk is minimized, compliance is met, and threats are quickly identified and addressed:

Facilitating real-time risk analytics
: Real-time unified views of the data allow organizations to monitor for, identify, and address potential threats. These views into the data help identify compliance breaches and fraudulent activities such as money laundering, privacy compliance, or external cyber threats. Logical data management ensures that

this data is readily available and eliminates the delay of copying data inherent in traditional data management processes. Thus it enables faster response to, or even prevention of, adverse events before they cause substantial damage to the organization or its stakeholders.

Meeting regulatory deadlines
: The last things any company wants to face are fines or penalties associated with failed compliance audits. In addition, they want to avoid the brand and customer loyalty impacts that can result from missing deadlines, as this may be disclosed publicly by the relevant regulatory agency. This is especially true during mergers and acquisitions, when the compliance scope may expand suddenly as the acquired organization's data landscapes are incorporated and subject to the acquirer's regulatory requirements. Logical data management can offset this by providing universal access across multiple disparate underlying data sets from across both organizations, and deliver unified real-time views without having to copy or migrate data from multiple sources.

On-demand compliance posture management
: A single logical layer for accessing underlying data sources also reduces policy and security management demands associated with a diverse and disparate data environment. Instead of sourcing compliance data from across a large variety of data sources and monitoring tools, the logical layer provides a central point of compliance information, enabling a company to quickly and easily deliver real-time compliance status to compliance policy owners. Known in the industry as *compliance posture management*, the ability to continuously understand, in real time, whether an organization is in compliance—and if not, why not—depends on unified real-time views of all data for measuring and monitoring compliance.

  Compliance posture management is important for any regulatory domain in which real-time compliance status is important. Examples include the following, all of which may have disparate monitoring and logging tools generating disparate data that a logical layer can easily unify and deliver in real time:

Data security posture management
: Monitoring for unauthorized data usage, access control violations, data leaks and breaches, and unauthorized copying that would violate GDPR and similar privacy regulations, or be indicative of a potential data security threat.

Environmental, social, and governance (ESG) posture management
: Monitoring consumption of energy and other resources across the organization as well as third parties (e.g., logistics companies transporting goods on a retailer's behalf, also known as Scope 3), in order to ensure ongoing compliance with sustainability regulations such as the European Union's Corporate Sustainability Reporting Directive (CSRD).

Financial risk posture management
: This can take several forms, including live dashboards depicting the likelihood of fraud or money laundering, understanding capital markets risk, monitoring receivables, and so forth, across multiple lines of business.

Implementing real-time controls
: With near-real-time data comes real-time notifications. Organizations can define thresholds that trigger automatic actions or alerts when fraud or potential breaches are detected.

Implementing global controls
: In addition to managing and reporting on access to a company's data, logical data management also enables changes and enforcement of global access controls in real time. An administrator can change a policy on a global level to quickly lock down and enforce security in just one place, the logical data layer, which then applies across the organization's entire data environment.

IT Infrastructure Optimization

IT is traditionally charged with managing and maintaining an organization's data management processes; however, as businesses grow and technologies change, IT is struggling to keep up with the demands. Data is becoming more voluminous, diverse, and complex. At the same time, the demand for data and the variety of its users means that the rigid methodology of traditional systems

is falling behind. Businesses can no longer wait on the extensive delay associated with updating data extractions, changing schemas, and moving data through traditional data pipelines. They need large amounts of data that is reliable and accessible in as close to real time as possible.

Logical data management reduces the burden on IT by making data easier to manage, faster to integrate, and easier for end users to access and utilize. Logical data management also makes it easier to monitor costs associated with data-related workloads, including storage, compute, and network traffic. This is measurable in many areas, such as the following:

Cloud operations
: Logical data management enables a company to track real-time information on data to optimize cloud spending, tie expenditure to business outcomes, and enable flexibility in cloud resource management. It can also track costs of data comparatively between pre-logical data management and post-implementation, as well as before versus after migrating to the cloud.

Agility and zero-downtime migrations
: Your company is always innovating, and logical data management enables you to adopt these innovations and make changes to underlying infrastructure without extensive changes to how the business is consuming its data. For example, as many organizations migrate from on-premises to cloud infrastructure, logical data management provides a single view of data that the business could use on an ongoing basis and without disruption, even while data is being migrated to new technologies behind the scenes. Additionally, logical data management enables the workload to be processed in the system that runs it most optimally with regards to performance, functionality, and cost, without sacrificing access, security, and governance. Currently, many organizations are implementing GenAI by introducing new AI technologies into their business processes, and logical data management similarly can shield downstream data users from such changes. Companies can measure the impact on business infrastructure by comparing migration and adoption times of new technologies compared to the pre-logical data management environment.

Distributed queries

: Logical data management improves distributed query performance. Logical data management platforms are able to measure the performance attributes of various data sources and optimize distributed queries accordingly. They can also track cost metrics such as storage, compute, and network traffic over time, and also compared to pre-logical data-management baselines. This improved performance means faster access to data and at better cost for time-critical applications.

Conclusion

Logical data management can impact your business in many technical ways, but this chapter reveals some of the benefits from a business operations perspective. A logical data layer provides opportunities to improve business operations performance, serves as a basis for designing a self-service data culture, and reduces the dependencies on valuable IT resources. All these impacts result in improved efficiency, increased revenue, and improved customer satisfaction for your business.

CHAPTER 10
The Future of Logical Data Management

As time progresses, your company will continue to need data; however, your future data needs will come with additional challenges. Larger and more complex data sources, advancing technologies, and changes in business needs are just a few of the reasons traditional data management processes will no longer suffice. Your company needs a data management strategy that is both flexible and resilient to handle the challenges of future data demands. This chapter examines how logical data management provides that foundation. As discussed throughout this book, logical data management provides the following benefits:

A seamless data ecosystem
: Your business is a complex tapestry of diverse, interwoven parts. To gain insights on its performance overall, you need to see how all these parts work together as a cohesive whole. The future of logical data management will utilize advances in data virtualization, semantic layers, and real-time processing to remove these divisions and grant you insights with greater speed and accuracy. Data mesh, data fabric, and any future technology will work together to empower data domains to take ownership of data while enabling centralized governance and data consistency.

AI-driven integration and insights

It's no secret that AI is drastically altering the business landscape, but what future impacts might AI have on data and logical data management? Sophisticated and more complex AI models will vastly improve query performance, automate the creation of data marketplaces, and enhance data governance frameworks. Large language models already integrate with logical data platforms, allowing for more seamless natural language queries. This will facilitate data democratization and bring data insights to all levels of your organization. As a result, employees, customers, and shareholders will be able to ask questions and get answers about any facet of the business and receive insights from reliable and understandable data sources.

Ethical and responsible data management

Logical data management prioritizes a framework designed around security, compliance, and privacy. This is used to ensure that data is harnessed effectively and responsibly. It enables organizations to balance innovation with accountability, building trust with customers and shareholders. As data grows more important, the data management needs will grow as well, and logical data management is designed to adapt to these changes.

Infinite scalability and adaptability

The capabilities of data lakehouses and cloud data warehouses are extended by a logical data platform. The future of logical data management will extend and scale these capabilities well beyond the current limitations. It will enable technologies such as cloud-native architectures, edge computing, semantic layers, and query acceleration. This will help your company to be flexible and adapt to changing market and business conditions, scale your operations with less effort, and provide rapid analytic insight even under larger loads.

Empowering a data-driven society

The underlying power of logical data management lies in making data accessible and understandable to your organization, your shareholders, and your customers. It breaks down the barriers between data types and domains, provides a natural business language semantic layer, and ensures accurate, reliable data while simultaneously integrating security, privacy, and compliance. Logical data management will break down

technical barriers and bridge the gap between business and IT, aligning technology with strategic business goals.

Agility and faster time to data
: Logical data management places a virtual layer on top of existing data resources, regardless of location, database format, or data structure. This means that the virtual layer can sit upon your systems, and is flexible enough to adapt as the business changes or new data sets become available. Data virtualization uses techniques like query push-down, caching, and aggregate-aware acceleration to ensure that your business has access to its data as quickly as possible. This enables you to make faster data-driven decisions for your business.

Logical data management is a vision of how your business can interact with its data in an increasingly complex and connected world. Its future will not only enhance data capabilities, but impact how organizations operate, innovate, and thrive.

About the Author

Christopher Gardner is a business intelligence analyst and lead Tableau developer for the University of Michigan. He has over 24 years of experience with the university and has served as a data analyst and Tableau expert for the campus since 2013. Within his role for the university and through O'Reilly boot camps and classes, he has taught thousands of users how to develop data visualizations and dashboards within Tableau. In addition, Christopher is a tech editor and writer for O'Reilly, participating in various Tableau and data-related articles and books. He is a Tableau Certified Data Analyst and has maintained an equivalent of that since 2016. He holds a degree in Actuarial Mathematics from the University of Michigan.