Brought to you by:

denodo

Data Virtualization

for dummies®

A Wiley Brand

Integrate structured and unstructured data

Create a real-time data abstraction layer

Enable information delivery and self-service

Lawrence C. Miller

Denodo Special Edition

About Denodo

Denodo is the leader in data virtualization software, providing unmatched performance; unified access to the broadest range of enterprise, big data, cloud and unstructured sources; and the most agile data services provisioning and governance at less than half the cost of traditional data integration. Denodo customers have gained significant business agility and ROI by creating a unified virtual data layer that serves strategic enterprise-wide information needs for agile BI, big data analytics, web and cloud integration, single-view applications, and enterprise data services across every major industry.

Leaders worldwide such as Amgen, Asurion, Autodesk, Vodafone, Seagate, Sumitomo Mitsui Trust Bank, Festo, and SwissRe rely on Denodo for mission-critical business processes such as business transformation to a subscription revenue model, logical data access for cloud modernization, and information self-service, as well as data governance and compliance initiatives. Find out more at **www.denodo.com.**

Data Virtualization

Denodo Special Edition

by Lawrence C. Miller

for dummies®
A Wiley Brand

Data Virtualization For Dummies®, Denodo Special Edition

Published by: **John Wiley & Sons, Ltd.,** The Atrium, Southern Gate Chichester, West Sussex, www.wiley.com

ISBN 978-1-119-55849-1 (pbk); ISBN 978-1-119-55851-4 (ebk)

Printed in Great Britain

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

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

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

Project Editor: Martin V. Minner

Associate Publisher: Katie Mohr

Editorial Manager: Rev Mengle

Business Development Representative: Frazer Hossack

Production Editor: Tamilmani Varadharaj

Denodo Review Team: Paul Moxon, Pablo Alvarez, Ravi Shankar, Lakshmi Randall, Becky Smith, Amy Flippant

Introduction

Organizations are challenged by ever-growing data volumes, as well as data types that are increasingly diverse. With the advent of big data and the proliferation of multiple information channels, organizations must store, discover, access, and share massive volumes of traditional and new data sources.

At the same time, more business opportunities can be realized only if large and diverse sources can be integrated in near-real time, if not in real time. In today's complex data landscape, it is no longer feasible to replicate data from myriad sources into a central repository because of the associated costs and delays in accessing the data. Cloud storage architectures have helped, but they still establish independent data silos that cannot be seamlessly integrated with other systems, such as traditional data warehouses.

Data virtualization is a modern approach to data integration. It transcends the limitations of traditional techniques by delivering a simplified, unified, and integrated view of trusted business data in real time or near-real time, as needed by consuming applications, processes, analytics, or business users.

About This Book

Data Virtualization For Dummies, Denodo Special Edition, consists of seven chapters that explore

- The challenges of data silos, data overload, and regulatory compliance (Chapter 1)
- What data virtualization is and how it helps businesses (Chapter 2)
- Data virtualization use cases (Chapter 3)
- How data virtualization enables big data solutions (Chapter 4)
- Data virtualization in the cloud (Chapter 5)
- How to get started with data virtualization (Chapter 6)
- Key things to know about data virtualization (Chapter 7)

Foolish Assumptions

It's been said that most assumptions have outlived their uselessness, but I assume a few things nonetheless.

Mainly, I assume that you are someone who uses or manages data in your organization, such as:

- A data warehouse manager, data engineer, or database administrator responsible for making data available to the business in an agile, cost-effective, and secure manner
- A data analyst or data scientist needing fast and reliable access to large and diverse data sets
- A business user who regularly needs to access data to make informed and timely decisions with all the best available data

Icons Used in This Book

Throughout this book, I occasionally use special icons to call attention to important information. Here's what to expect:

REMEMBER

This icon points out information you should commit to your nonvolatile memory, your gray matter, or your noggin — along with anniversaries and birthdays.

TECHNICAL STUFF

You won't find a map of the human genome here, but if you seek to attain the seventh level of NERD-vana, perk up! This icon explains the jargon beneath the jargon.

TIP

Tips are appreciated, never expected — and I sure hope you'll appreciate these tips. This icon points out useful nuggets of information.

WARNING

These alerts point out the stuff your mother warned you about (well, probably not), but they do offer practical advice to help you avoid potentially costly or frustrating mistakes.

IN THIS CHAPTER

- Eliminating siloed data in the enterprise
- Dealing with different data sources and types
- Understanding regulatory compliance requirements
- Learning the basics of data virtualization

Chapter 1
Data, Data Everywhere

In this chapter, you learn about modern data challenges including data silos, disparate data sources and types, and regulatory compliance. You also learn what data virtualization is — and what it isn't.

Unlocking Data Silos

Data silos — data sources that can't easily be shared across systems and applications — have plagued the IT and business landscape for many years. These silos exist within organizations for a variety of reasons, such as:

- Older, legacy systems have trouble communicating with more modern systems.
- On-premises systems have difficulty communicating with cloud-based systems.
- Multiple disparate storage systems have been deployed over the years as existing systems near storage capacity or performance degrades.
- Some systems work only with specific applications.

- Some systems are configured to be accessed only by specified individuals or groups.
- Companies acquire other companies with systems that are configured differently.

Data silos make it challenging for business users to access and analyze all of the available data within an organization. Data silos can lead to inaccurate results or conclusions and delayed decision making with incomplete or imperfect data. The lack of a single "source of truth" also creates doubt in the veracity of the data.

Managing the Data Swamp

Managing the deluge of digital data is challenging for every business today. In addition to the sheer volume of data, businesses must manage multiple different data types — including structured, unstructured, and semi-structured data — from multiple data sources. These different data types must often be extracted from their sources, transformed into a different format, and loaded into the consuming application (a process known as Extract, Transform, and Load, or ETL) before they can be used by the business. ETL processes (discussed in Chapter 2) are often scripted or manual processes that require IT assistance, run in scheduled batches, and are inflexible, which introduces further complexity and delay.

Navigating the Compliance Landscape

New legislation and regulations mandating data protection requirements are a constant and costly challenge for organizations in practically every industry. Regulations such as the U.S. Health Insurance Portability and Accountability Act (HIPAA), the U.S. Gramm-Leach-Bliley Act (GLBA), and Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) establish data privacy, protection, and retention requirements for certain businesses and industries.

More recently, the European Union's (EU) General Data Protection Regulation (GDPR) went into effect on May 25, 2018. All

businesses that serve EU citizens, regardless of where the business is located, are required to comply. The GDPR details how companies must protect personal information. Companies that fail to comply with the GDPR will not only be subject to hefty fines, but may also face lawsuits and additional audits. To comply with the GDPR, companies will have to demonstrate that personal data is

- Processed lawfully and fairly, and in a transparent way.
- Collected for specific, explicit, and legitimate purposes.
- Limited to what is necessary for processing.
- Kept accurate and up to date.
- Stored so that the subject is identified only when necessary.
- Processed in a secure manner so it does not fall into the wrong hands or become lost, damaged, or destroyed.
- Protected "by design." All new systems must be developed with privacy in mind.

Companies need a bird's-eye view into all of their data, as well as a way to establish security controls over the entire infrastructure from a single point. Data virtualization provides this capability, enabling companies to quickly and easily comply with data protection mandates without investing in new hardware or re-building existing systems from the ground up.

What Is Data Virtualization?

Data virtualization delivers a simplified, unified, and integrated view of trusted business data in real time or near-real time as needed by the consuming applications, processes, analytics, or business users. Data virtualization integrates data from disparate sources, locations, and formats, without replicating the data, to create a single, virtual data layer that delivers unified data services to support multiple applications and users (see Figure 1-1). The result is faster access to all data, less replication and cost, and more agility to change.

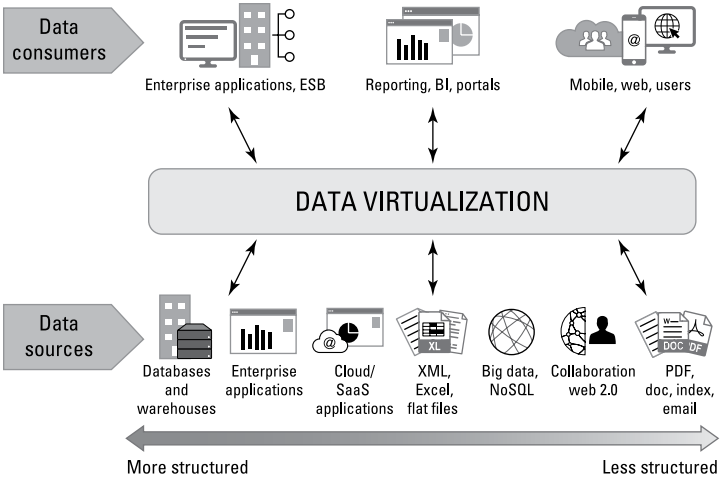


FIGURE 1-1: Data virtualization integrates data from disparate sources, locations, and formats to support multiple applications and users.

Whereas most data integration solutions move a copy of the data to a new, consolidated source, data virtualization offers a completely different approach. Rather than moving the data, data virtualization provides a view of the integrated data, leaving the source data exactly where it is. Companies do not have to pay the costs of moving and housing the data, and yet they gain the benefits of data integration.

Data virtualization performs many of the same transformation and quality functions as traditional data integration — such as ETL, data replication, data federation, Enterprise Service Bus (ESB) and others — but leverages modern technology to deliver real-time data integration at a lower cost, with more speed and agility. It can replace traditional data integration and reduce the need for replicated data marts and data warehouses in many cases.

Data virtualization is also an abstraction layer and a data services layer. In this sense, it is highly complementary to use between original and derived data sources, ETL, ESB and other middleware, applications, and devices, whether on-premises or cloud-based.

Data virtualization delivers the following key capabilities:

- **Logical abstraction and decoupling:** Disparate data sources, middleware, and consuming applications that use or expect specific platforms and interfaces, formats, schema, security protocols, query paradigms, and other idiosyncrasies can now interact easily through data virtualization.
- **Data federation on steroids:** Data federation is a subset of data virtualization, but enhanced with more intelligent real-time query optimization, caching, in-memory, and hybrid strategies that are automatically (or manually) chosen, based on source constraints, application need, and network awareness.
- **Semantic integration of structured and unstructured data:** Data virtualization is one of the few technologies that bridge the semantic understanding of unstructured and web data with the schema-based understanding of structured data to enable integration and data quality improvements.
- **Agile data services provisioning:** Data virtualization promotes the application programming interface (API) economy. Any primary, derived, integrated, or virtual data source can be made accessible in a different format or protocol than the original, with controlled access in a matter of minutes.
- **Unified data governance and fine-grained security with complete auditability:** Data virtualization enables fine-grained control over sensitive customer information stored across multiple systems by establishing a single, unified access layer across on-premises and off-premises systems. All data is made discoverable and can be easily integrated through a single virtual layer that exposes redundancy and quality issues faster. Data virtualization imposes data model governance and security from source to output data services, and consistency in integration and data quality rules. When data consumers need to access a source, they do so through the data virtualization layer, which contains the metadata for accessing each source, and returns a secure, virtualized view of the data to the consumer in real time. These views are traceable and auditable, and will be delivered only to authorized consumers.

- **Eliminates unnecessary data movement:** With a data virtualization layer in place, no data replication is required for reporting purposes, and no Extract, Transform, and Load (ETL) scripts must be rewritten. A data virtualization layer operates with a company's existing infrastructure, configured exactly as it is. The data virtualization layer merely abstracts the access functions so that users perceive the data as existing in a single virtual repository. However, if for performance reasons data must be persisted, data virtualization tools also offer simple ways to persist a data set by simply enabling some settings in the model. Replication becomes just another option, not a necessity.
- **Complete data lineage and agile business rules:** At any point in time, companies can understand and report on the full lineage of any sensitive data set, including its original source, all views, and all modifications. In addition, through the data virtualization layer, companies can establish sophisticated rules for automating compliance, such as masking data on the fly, so it cannot be viewed by users who lack the requisite credentials. Such rules can be applied quickly and effectively across diverse systems because they are applied in the data virtualization layer.
- **Secures data-at-rest and data-in-motion:** The data virtualization layer can perform role-based authentication at any level, such as guest, employee, or corporate; apply data-specific permissions including row- and column-level masking; and define schema-wide permissions and policy-based security. The virtualization layer secures data in transit via Secure Sockets Layer/Transport Layer Security (SSL/TLS) protocols and authenticates users via industry-proven protocols such as Lightweight Directory Access Protocol (LDAP), pass-through with Kerberos, Windows Single Sign-On (SSO), Open Authorization (OAuth), Simple and Protected GSS-API Negotiation Mechanism (SPNEGO) authentication, OAuth and SAML authentication, and Java Database Connectivity/Open Database Connectivity (JDBC/ODBC) Security.

- **Facilitates privacy by design:** Data virtualization is also particularly well suited to helping companies comply with the GDPR's "protected by design" requirement. By definition, a data virtualization layer does not require a source to be of a prescribed type, or to be accessed in a certain way. New sources can easily be added to the infrastructure by connecting them to the data virtualization layer, where they are immediately subject to the same security controls and auditability as any other source on the system, irrespective of the data source technology.

Data virtualization delivers abstracted and integrated information in real time from disparate sources to multiple applications and users. It is also easy to build, consume, and maintain. To build virtual data services, the user follows three simple steps (see Figure 1-2):

- **Connect and virtualize any source.** Quickly access disparate structured and unstructured data sources using included connectors. Introspect their metadata and expose as normalized source views in the data virtualization layer.
- **Combine and integrate into business data views.** Combine, integrate, transform, and cleanse source views into canonical model-driven business views of data in a graphical user interface (GUI) or through documented scripting.
- **Connect and secure data services.** Any of the virtual data views can be secured and published as SQL views or dozens of other data services formats.

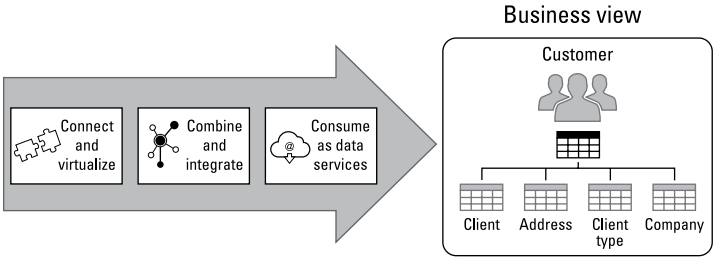


FIGURE 1-2: Building virtual data services.

WHAT DATA VIRTUALIZATION IS *NOT*

Some vendors use buzzwords for marketing other products to capitalize on the popularity of data virtualization. To help dispel any confusion, remember that data virtualization is not:

- **Data visualization:** It sounds similar, but *visualization* refers to the display of data to end-users graphically as charts, graphs, maps, reports, and so on. Data *virtualization* is middleware that provides data services to other data visualization tools and applications. Although it has some data visualization for users and developers, that is not the main use.
- **A replicated data store:** Data virtualization does not normally persist or replicate data from source systems to itself. It only stores metadata for the virtual views and integration logic. If caching is enabled, it stores some data temporarily in a cache or in-memory database. Virtual data can be persisted if desired by simply invoking it as a source using ETL. Thus, data virtualization is a powerful, but lightweight and agile, solution.
- **A logical data warehouse:** A logical data warehouse is an architectural concept and not a platform. Data virtualization is an essential technology used in creating a logical data warehouse by combining multiple data sources, data warehouses, and big data stores like Hadoop.
- **Data federation:** Data virtualization is a superset of capabilities that includes advanced data federation.
- **Virtualized data storage:** Some companies and products use the same term *data virtualization* to describe virtualized database software or storage hardware virtualization solutions. They do not provide real-time data integration and data services across disparate structured and unstructured data sources.
- **Virtualization:** When the term *virtualization* is used alone, it typically refers to hardware virtualization — servers, storage disks, networks, and so on.

IN THIS CHAPTER

- Assessing the pros and cons of ETL, ESB, and data virtualization
- Using traditional data integration techniques and data virtualization together
- Enabling business agility with data virtualization
- Empowering business users with self-service access to real-time data

Chapter 2
Introducing Data Virtualization

In this chapter, you look at traditional data integration techniques such as Extract, Transform, and Load (ETL) processes and Enterprise Service Bus (ESB) architectures, as well as how data virtualization complements these techniques, enables business agility, and makes self-service a reality for business users.

Going Beyond Traditional Data Integration

The problem with data silos (discussed in Chapter 1) is that no one can easily query all of the available data. Instead, each data silo must be queried separately, and the results then must be manually consolidated. This process is costly, time-consuming, and

inefficient. To bring the data together, companies typically use one or more of the following data integration strategies:

- **Extract, Transform, and Load (ETL)** processes, which copy the data from the different silos and move it to a central location, such as a data warehouse.
- **Enterprise Service Buses (ESBs),** which establish a communication system for applications, enabling them to share information.
- **Data virtualization,** which creates real-time, integrated views of the data in data silos and makes them available to applications, analysts, and business users.

Extract, Transform, and Load (ETL) processes

Extract, Transform, and Load (ETL) processes were the first data integration strategies, introduced as early as the 1970s.

The basic ETL process follows these steps:

1. The data is *extracted* from the source.
2. The extracted copy of the data is *transformed* into the format and structure required by its final destination.
3. The transformed copy of the data is *loaded* into its final destination, such as an operational data store, a data mart, or a data warehouse.

TIP

Some processes do the transformation in the final step and are therefore called "ELT" processes, but the basic concept is the same.

REMEMBER

Pros and cons of ETL processes include

- **Pros:**
  - ETL processes are efficient and effective at moving data in bulk.
  - The technology is well understood and supported by established vendors.

- ETL tools have features that sufficiently support bulk/batch data movement.
- Most organizations have in-house ETL capabilities.

» **Cons:**

- Moving data is not always the best approach because this results in a new repository that must maintained.
- Large organizations can have thousands of ETL processes running each night, synchronized by scripts that are difficult to modify.
- Typically, ETL processes are not collaborative; the end-users must wait until the data is ready.
- ETL processes cannot handle today's data volumes and complex data types.

Enterprise Service Bus (ESB)

ESBs, introduced in 2002, use a message bus to exchange information between applications. The message bus essentially acts as a translator between the applications, enabling the applications to communicate via the bus. An ESB decouples systems and allows them to communicate without depending on, or even knowing about, other systems on the bus (see Figure 2-1).

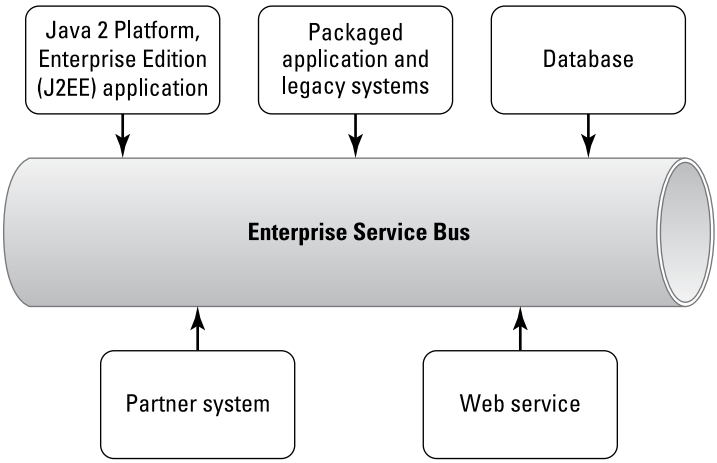


FIGURE 2-1: An ESB decouples systems and applications and enables communication between them.

ESBs form the underpinnings of service-oriented architecture (SOA), in which applications can easily share services across an organization. ESBs were born out of the need to move away from point-to-point integration which, like ETL scripts, are hard to maintain over time.

REMEMBER

Pros and cons of ESBs include

- **Pros:**
  - Applications are decoupled.
  - They can be used to orchestrate business logic using message flows.
  - ESB technology is mature and is supported by established vendors.
  - ESBs can address operational scenarios by using messages to trigger events.
- **Cons:**
  - ESBs cannot integrate application data to deliver on analytical use cases.
  - Queries are static and can only be scheduled; ESBs do not easily support ad hoc queries.
  - Database queries are restricted to one source at a time. Joins and other multiple-source functions are performed in memory, which drains resources.
  - ESBs are suitable only for operational use cases that involve small result sets.

Data virtualization

Data virtualization creates integrated views of data drawn from disparate sources, locations, and formats, without replicating the data, and delivers these views, in real time, to multiple applications and users. Data virtualization can draw from a wide variety of structured, semi-structured, and unstructured sources, and can deliver to a wide variety of consumers.

Because no replication is involved, the data virtualization layer contains no source data; it contains only the metadata required to access each of the applicable sources, as well as any global instructions that the organization may want to implement, such as security or governance controls.

Users and applications query the data virtualization layer which, in turn, gets the data from various sources (see Figure 2-2). The data virtualization layer abstracts users and applications from the complexities of access. The data virtualization layer appears as a single, unified repository to all consumers.

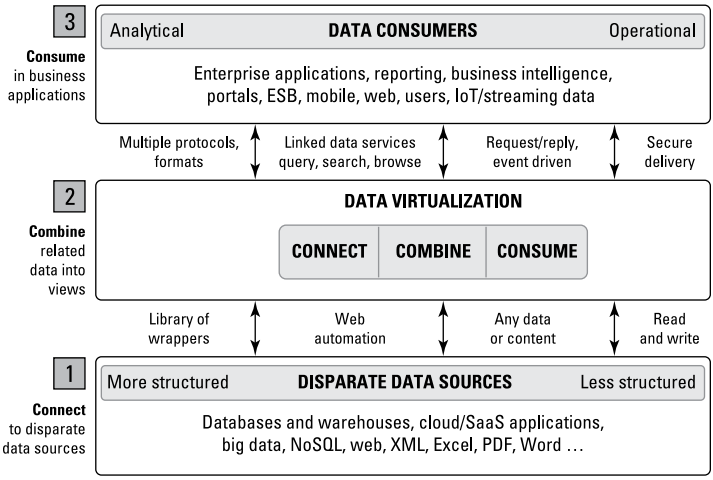


FIGURE 2-2: Data virtualization connects disparate data sources, combines related data into views, and publishes the data to applications for all data consumers.

Fundamental attributes that define the capabilities of a true data virtualization platform include the following:

- **Universal data access to any source or data type:** Data engines automatically connect, navigate, and extract data from any internal or external source and all data types including structured, unstructured, and web.
- **Unified virtual data layer:** Powerful transformations and relationships are built using an integrated modeling and execution environment to normalize, transform, improve quality, and relate data across heterogeneous source types using common metadata and semantics. An extended relational data model allows disparate data types to be represented natively in the virtual layer, thereby minimizing effort and maximizing performance.

- **Universal data publishing:** Combined information is published as reusable data services in multiple formats such as SQL query, Simple Object Access Protocol (SOAP), Representational State Transfer (REST) and Open Data Protocol (OData) web services, messaging, mobile feeds, keyword-based search, and so on. Hybrid delivery modes (such as virtual real time, cache, batch, and message-based) to consuming applications are also supported.
- **Agile high performance:** Advanced real-time dynamic optimization is supplemented by intelligent caching and scheduled batches for flexible mixed workloads. Read/write access with enterprise-class reliability and scalability — even for web and unstructured sources — is supported.
- **Unified data governance:** An enterprise-wide single entry point for data and metadata management, security, audit, logging, and monitoring is enabled through built-in tools and instrumentation, as well as integration to external data management tools.
- **Agile development of pervasive, self-service data services:** Complexity is hidden from application developers and business users. Consuming applications and data sources are decoupled, allowing data services to be easily created, extended, and used.

REMEMBER

Pros and cons of data virtualization include

- **Pros:**
  - Seamlessly federates two or more disparate data sources (makes them appear and function as one), including a mix of structured and unstructured sources.
  - Provides value-added features such as intelligent real-time query optimization, caching, in-memory processing, and custom optimization strategies based on source constraints, application needs, and data volumes.
  - Any primary, derived, integrated, or virtual data source can be made accessible, via an application programming interface (API), in a different format or protocol than the original, with controlled access, in minutes.
  - All data is accessible through a single virtual layer, which quickly exposes redundancy, consistency, and data quality issues, and enables the application of universal, end-to-end governance and security controls.

- Provides data catalog capabilities to enhance data stewardship and to support business initiatives such as a digital marketplace; governance, risk management, and compliance (GRC); and data-as-a-service.

» **Cons:**

- Lack of support for complex bulk/batch data flows, which might be required by a few use cases — for example, conditional flows with multiple targets, multi-pass loops, surrogate key management, and so on.

Table 2-1 shows which data integration strategies are best suited to various use cases.

TABLE 2-1 Data Integration Use Cases and Strategies

Use Case	Data Virtualization	ETL	ESB
Moving data into enterprise data warehouse or operational data store		x	
Migrating enterprise data warehouse to cloud	x	x	
Data unification	x		
Customer 720	x		
Real-time insights	x		x
Virtual data marts	x		
Physical data marts		x	
Agile reporting from enterprise data warehouse and other sources	x		
Logical data warehouse	x		
Data warehouse offloading	x	x	
Application synchronization		x	x
Metadata discovery and enrichment	x		
Self-service analytics	x		
ETL seeding (decouple ETL from sources)	x		
Event-driven workflows			x

Complementing ETL and ESB

Data virtualization supports a wide variety of sources and targets, which makes it an ideal data integration strategy to complement ETL processes and ESBs.

ETL processes were designed for moving data into data warehouses and similar environments, and they are particularly well suited to this task. However, ETL processes cannot easily support cloud-based sources. Data virtualization can complement ETL processes in the following ways:

- Seamlessly connecting on-premises with cloud data sources without the need to consolidate data within a single repository.
- Enabling the migration from on-premises to cloud-based systems without interrupting business continuity.
- Data warehouse offloading in which data virtualization not only helps with the offloading process, but also unifies data across the traditional data warehouse and the new repository such as Hadoop, Amazon Web Services (AWS) S3, or a cloud-based data store.
- Real-time integration of disparate data sources.
- Replacing ETL processes with data virtualization where faster access to data is needed.

Data virtualization can also complement an ESB and enhance its performance. Adding new sources to an ESB can be complex; sources like relational databases, web, or cloud-based sources, flat files, or email messages are not immediately enabled for the SOA that the ESB supports. To streamline this process, all sources that the ESB cannot handle can be unified by the data virtualization layer before being passed to the ESB. This architecture leverages the best qualities of both technologies: Data virtualization unifies disparate sources, and ESBs deliver the critical messages to support the business process.

Delivering Faster

Modern businesses require fast access to the most up-to-date and accurate data to make strategic decisions, anticipate customer needs, and stay ahead of competitors.

Physically moving, warehousing, and storing data multiple times costs money and slows down the business when changes are needed. ETL processes (discussed earlier in this chapter), many of which are often manual, introduce opportunities for costly errors and delay access to data. Replication between multiple data sources across network links adds further delay, cost, and complexity.

Data virtualization enables fast data architectures such as logical data warehouse, virtual data marts, self-service business intelligence (BI), and operational analytics. By aggregating the latest data residing in various source systems without the need to physically move the data, data virtualization helps IT rapidly deliver data for business users to use within their BI systems.

REMEMBER

The need to physically move data is one of the main culprits of data delay in traditional data architectures.

The IT landscape is also becoming more complex and distributed with additional data silos being created as big data stores and cloud solutions are adopted. Data virtualization counters this data deluge with a fast data architecture for advanced analytics and data warehouse offloading. This architecture enables IT to leverage the lower costs that big data and cloud solutions afford, while significantly improving the time to data delivery with real-time access.

Data services are increasingly critical to application development. Data virtualization enables rapid application development with a unified data services layer creating a logical data abstraction of all structured and unstructured data from the underlying sources.

TIP

Using data virtualization, IT can develop data services in less than half a day, whereas traditional data integration methods such as ETL can often take one to two weeks.

Single-view applications such as a single view of customers, products, inventories, and so on, improve the efficacy of call center agents with rapid response times, and that of sales and marketing teams with targeted campaigns. Data virtualization can underpin these single-view architectures by virtually aggregating different master data in real time, without having to replicate, centrally store, and manage the data.

To support both data-oriented and business-oriented users, data virtualization can provide an easy-to-use, yet sophisticated data modeling environment to deliver data access, manage data, and

provide data services delivery. Data virtualization helps IT teams and business users respond quickly to rapidly changing requirements in the following ways:

- An integrated development environment that meets the needs of both IT and business users with a user-friendly drag-and-drop and low-code developer studio that is geared to data-oriented teams such as data engineers, power users, and integrators, who can publish data services with just a few clicks.
- A comprehensive catalog of business views classified and tagged according to business categories for easy access allows users to browse and navigate, discover the relationships in the data sets, carry out searches over the metadata and the data itself to properly validate datasets, inspect tree view and detailed lineage information, and export query results.
- Pre-built connectors to source and target systems so that IT can quickly connect to disparate sources and ensure maximum performance.
- Flexible data delivery with one-click publication of powerful REST and OData web services.

Making Self-Service a Reality

Self-service analytics has been a long-sought panacea, promising to liberate business users to perform analytics without IT assistance and freeing IT to focus on other strategic business initiatives.

Today many desktop analysis tools enable users to slice and dice data and serve it up in a variety of full-featured reports and dashboards. However, several key challenges at the data level have prevented self-service analytics from becoming a reality:

- **Fragmented data:** Data is spread across multiple heterogeneous databases, data warehouses, cloud and big data systems, NoSQL sources, and flat files.

- **Multiple, high-maintenance data integration initiatives:** When a business user needs to query across multiple heterogeneous sources, companies often charge IT with creating ad hoc point-to-point integrations using ETL processes. If a source must be changed, these processes must be rewritten, which is costly and time consuming.
- **Data delays:** It can often take months to deliver requested data using legacy data integration processes, increasing the likelihood that the data will be inaccurate or irrelevant.
- **Poor data integrity:** Without a single "source of truth," business users may inadvertently use less authoritative sources, resulting in data of questionable quality.
- **Untraceable data lineage:** If users collect data from sources directly, they many not keep an accurate record of where the data came from, hindering the ability to determine data quality and further eroding trust in the data.

Data virtualization overcomes these challenges to make self-service a reality for business users:

- **Fragmented data is seamlessly unified.** With a data virtualization layer in place, all of the data in its various formats across myriad systems appears to users as though it sits in a single, easily accessible repository.
- **High-maintenance data integrations are replaced by a single, low-maintenance data virtualization layer.** Unlike legacy data integration technologies such as ETL scripts, data virtualization can easily accommodate changes to the source data without heavy modification.
- **Data delays are virtually eliminated because data can be accessed in real time.** Integrated views of the data, even across numerous heterogeneous sources, can be delivered to users in real time.
- **Data integrity is preserved.** Because all of the data sources are accessed through the data virtualization layer, companies can use the data virtualization layer to establish strong governance protocols and specify authoritative sources.
- **Data lineage is fully traceable.** Because all data flows through the data virtualization layer, data lineage is fully traceable from users to sources.

Data virtualization increases operational efficiencies, reduces costs and complexity, minimizes data replication, and fosters data reusability and collaboration. By enabling self-service for business users, data virtualization further enables businesses to speed decision making and time to market by reducing the reliance on limited IT resources to access and integrate data.

TIP

Data virtualization allows for replication, but only when it is necessary.

IN THIS CHAPTER

- Enabling self-service business intelligence
- Delivering a superior customer experience
- Strengthening data governance and security
- Exposing data sources as data services

Chapter 3
Exploring Data Virtualization Use Cases

This chapter explains some common data virtualization use cases and provides several real-world data virtualization success stories.

Making Agile BI a Reality

One of the most common uses of data virtualization is for agile reporting, operational business intelligence (BI), and real-time dashboards that require timely aggregation, analysis, and presentation of the most relevant data from multiple sources. Both individuals and managers must monitor performance to help make daily operational decisions in key business processes such as sales, support, manufacturing, logistics, finance, legal, and compliance.

REMEMBER

Data virtualization enables IT to be more agile in responding to this almost insatiable demand for actionable information. With data virtualization, the data remains in the source data stores. Replication, with its accompanying staging, transformation, and batch copying tools and processes, is not required. Access to the data is through virtual views that can be quickly created (and discarded, if necessary). Changes are similarly quick, making

iterative report and dashboard creation, with almost immediate involvement and feedback from the business stakeholders, a reality.

Benefits of data virtualization for agile BI projects include

- **Significantly minimizes replication:** From the data virtualization layer, the user can point to the original data sources, partially cache any data, and build "virtual data marts" that are defined in the data virtualization layer, thus avoiding the creation of new repositories (and more copies of the data).
- **Makes it easier to change views and be more responsive to business requests:** Changes are carried out in the data virtualization layer, avoiding modifying ETL scripts across the entire replication chain.
- **Enables access to real-time data for operational BI:** By enabling direct access to operational systems, data virtualization delivers data to consuming applications significantly faster than traditional data integration approaches.
- **Integrates any data type:** From semi-structured to structured and unstructured, data virtualization seamlessly integrates data from all data sources.
- **Enables self-service BI:** Users can run any report against "self-service ready" data services created and managed by IT, and mitigate the potential administration and governance nightmares with security and governance providing "self-service with guard rails."

DATA VIRTUALIZATION POWERS THE DATA REVOLUTION AT FESTO

Festo is dedicated to maximizing productivity and competitiveness for process manufacturing companies, paving the way for their digital transformation. Many aspects of Industry 4.0 are already a reality for Festo Group, as the company develops future-oriented products founded on innovative, energy-efficient technologies, intuitive human-machine collaboration, and advanced training.

Business need

To continue the innovation that has always been at the forefront for Festo, the company needs to optimize operational efficiency, automate manufacturing processes, and deliver on-demand services to its business consumers. This includes finding smarter ways to streamline how the company aggregates and analyzes data. It also underscores the need for an agile solution that would better enable Festo to monetize its customer-facing data products.

Festo also needed its business users to become self-sufficient with reporting and analysis and reduce their reliance on IT for preparing and surfacing the data they need. In addition, Festo's business teams had launched strategic projects to maximize energy efficiency, and they needed to be able to provide instant visibility on energy usage directly to the shop floor teams.

However, Festo was challenged in finding an agile and robust way to integrate the data from the existing silos, which included the data warehouse, machine data sources, and other sources, in a way that would reduce the reliance on IT by the business users while providing the quick turnaround and flexibility that the users were demanding.

Solution

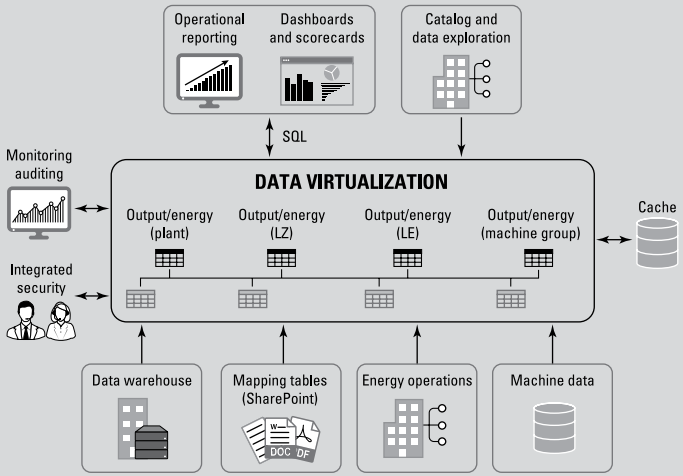
The Festo big data team developed a big data analytics framework to provide a data marketplace to better support the business (see the accompanying figure). Using the Denodo platform, this framework integrates data from numerous on-premises and cloud systems, including streaming data, machine data, and data-at-rest, and provides access to the integrated data in real time. Because the framework establishes a unified access layer, it provides consistent data access and governance across the different silos of data. As a result, business users now have easy access to all the data they need, when they need it.

To meet the demands of the business and deliver speed, flexibility, and agility, Festo implemented the Denodo platform as a key component within the big data analytics framework. The logical layer delivered by the Denodo platform provides virtual views that are tailored for business analysts, data scientists, and developers across multiple departments. "This is a win-win for us, as the business now has the flexibility they need, and they no longer have to rely on IT when they want to pull data," says Diethard Frank, IT Product Management

(continued)

(continued)

Big Data at Festo. The views incorporate data from local sources to help stakeholders meet last-mile requirements. The Denodo Platform also gains efficiencies because it removes the need to replicate data. Data remains in the source data stores, and it is accessed through business-focused virtual views.



The Festo big data analytics framework.

Benefits

The Denodo platform supports Festo's big data analytics framework by

- Delivering enhanced insight across the business without having to physically move data
- Simplifying data consumption, because data virtualization is source-agnostic and provides a single endpoint for accessing all data
- Quickly integrating new data sources and making them available to user communities in real time
- Facilitating smarter decision making via additional information-enrichment capabilities
- Increasing the speed and agility of both business and IT, because business users can now drive and maintain their own dashboards, significantly increasing customer satisfaction.

Getting a Complete View of Your Customers

Too often, information about a customer, supplier, product or project is dispersed among various systems and sources. This information must be brought together into a "holistic view" to conduct business more efficiently and trigger innovation. The most common implementation of this use case is "single customer view" data services for contact centers and customer self-service portals.

Knowledge and insight about customers leads to positive customer experiences and is the lifeblood of every business. The success of a relationship between an organization and its customers is determined by the quality of their interactions throughout the customer lifecycle. During these crucial moments, organizations can create loyalty, differentiate themselves from competitors, and increase the value of the relationship with their customers. If not managed properly, these interactions can undermine the success of that relationship. One of the major challenges facing organizations today is the need to create a single, accurate, consistent, and timely view of their customers — a view that cuts across all applications, systems, business units, and customer touch points.

REMEMBER

An organization's contact center is a key customer touch point. For many, the contact center is the primary channel for providing enriching customer service experiences. For others, the contact center is not only a part of the customer service but also a sales and marketing channel. Both types of organizations require fast access to a diverse array of information about the customer. Positive and correct interaction with customers in the contact center requires not only a good communication infrastructure but, more importantly, it requires the implementation of internal processes and information infrastructure closely aligned with the customer needs.

Unfortunately, the tools available to contact center personnel too often fall short. The reality is that for many organizations, the lack of complete, consistent, and timely data required by their customer-facing processes is a major barrier to maximizing business value in customer interactions. Fragmented data and lack of integration into business processes create obstacles that reduce the quality of customer service, cause lost business opportunities, and lead to poor customer satisfaction. These obstacles also lead to inefficiencies, low productivity, and poor morale in contact center staff.

Customer-centric data virtualization solutions include

- Unified desktops for contact centers
- Enhanced web self-service portals
- Single customer view
- Real-time performance monitoring
- Automated access to external web information

TIP

Benefits of data virtualization include

- Improved customer service at a reduced cost
- Improved staff motivation and effectiveness
- Reduced customer churn
- Enhanced competitive differentiation
- Improved cross-selling and up-selling capability

HOW GETSMARTER CREATED A SINGLE VIEW OF THE STUDENT LIFECYCLE

GetSmarter has been experiencing steep growth for quite some time because of the popularity of its university-accredited online courses, hundreds of which are available through the company's online campuses in video and audio formats. As GetSmarter's customer base and operations grew larger and more complex, so did its data repositories, which contain a variety of functional data covering marketing, finance, courses, students, and many other domains. To deliver faster deployments, GetSmarter's IT created many microservices targeting business contexts or entities such as enrollment, registration, the student portal, the user interface, payments, billing, and communications.

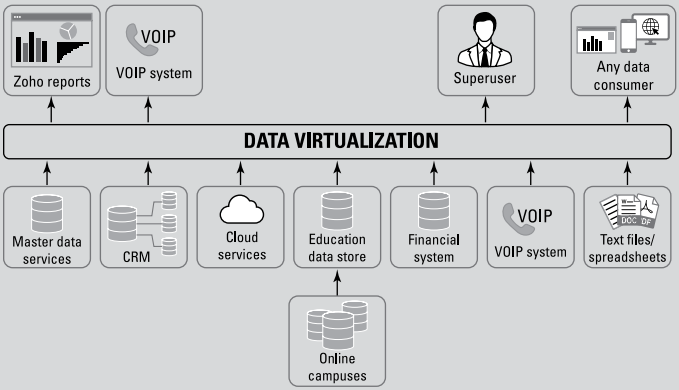
Business need

Although microservices accelerated GetSmarter's feature deployment and product announcements, the company's reporting tools now needed to talk to many databases instead of just one. GetSmarter needed a way to accommodate the many-to-many reporting tool and avoid having to manage multiple database connections. The company needed a solution that would de-clutter the microservice-based data

architecture and reduce its rigidity. While GetSmarter took the purist approach of relying on API calls to transfer data within the microservices architecture, the company still retains many legacy systems and reporting tools for marketing, finance, and online campuses, which require a simpler, more agile data access mechanism. Also, GetSmarter did not use a big bang approach in moving to the microservices architecture because the company wanted to phase out the legacy systems over time without disrupting day-to-day operations. That propelled GetSmarter to look for a smart solution for its data access layer. Finally, GetSmarter needed this data access layer to offer comprehensive security, data privacy, and data governance.

Solution

After evaluating a variety of data integration patterns and solutions, GetSmarter chose the Denodo platform, hosted on Amazon Web Services (AWS), as the unifying fabric and data access layer to integrate and abstract the company's microservices architecture, cloud and SaaS-based data sources, reporting tools, and other legacy systems (see the accompanying figure). As GetSmarter already had presence on AWS, and as the Denodo Platform is readily available on the AWS marketplace, deploying the Denodo platform was a seamless experience for GetSmarter.



How GetSmarter unified the fabric and data access layer.

Leveraging the Denodo platform's advanced data virtualization capabilities, GetSmarter virtualized the company's master data management (MDM) system, microservices, and legacy systems and created a combined virtual view, adhering to a single semantic model, to reduce

(continued)

(continued)

the business impact of reporting. The company created a single view of predefined, consistent entities such as students, courses, and pricing, which are easily recognizable by business stakeholders across the enterprise. GetSmarter leveraged the Denodo platform's data governance and security features to facilitate compliance with the European Union's General Data Protection Regulation (GDPR), the UK Data Protection Act, and South Africa's Protection of Personal Information (PoPI), in addition to ensuring that GetSmarter's own data layer was feature-rich in terms of security, risk, and compliance.

Benefits

Benefits of the Denodo platform for GetSmarter include

- GetSmarter is able to announce and launch products and services faster than ever, while keeping service quality intact, directly boosting revenue.
- With a single view of centralized information, GetSmarter business users can make faster business decisions.
- By virtualizing and combining voice over IP (VoIP) and customer relationship management (CRM) systems, GetSmarter is able to assign appropriate student success managers to students, significantly increasing customer satisfaction.

Improving Data Governance and Security

The data virtualization platform provides a unified entry point for data governance at the virtual data layer using data and metadata discovery, lineage, change impact analysis and propagation, metadata sharing, and support for mapping physical to logical and canonical data models or contract-first web services. Data virtualization also provides policy-based data services security, integrated with user/role management systems with granular read/write and row/column access control. Service levels can also be differentiated and tracked based on users and roles. Finally, a complete suite of real-time monitoring, audit, and logging capabilities ensures enterprise readiness.

REMEMBER

One critical factor in ensuring that governance, risk management, and compliance (GRC) is healthy and functioning is effective information sharing capabilities. For governance, information sharing enables the communication and understanding necessary

for senior management to work effectively through all layers of an organization. For risk management, information sharing enables organizations to gain unified views of risk across the many varieties of risk, many of which are handled by dedicated departments. For compliance, information sharing enables rapid reporting while minimizing the impact to business operations.

Unfortunately, organizations face a number of challenges with regard to information sharing, particularly around the integration of data, including

- **Disparate data sources:** Data is often fragmented across myriad internal and external data sources.
- **Different data formats:** Across the disparate source systems, data is often stored in different formats.
- **Different data standards:** Each industry has its own standards for identifying data entities. Within an organization, it is not uncommon to find data adhering to different standards.
- **Incomplete data:** Data sets cannot be easily shared if records are missing or invalid due to a reliance on data stored on different systems.
- **Unprocessed data:** Data cannot be easily shared with other systems if it contains calculations that are not part of day-to-day operations.
- **Sensitive data:** Sensitive data can be shared, but only with individuals who need to see it and have the privileges to see it. To share sensitive data, sophisticated systems must be put in place that provide granular, selective access to sensitive information.


REMEMBER

Data virtualization enables organizations to easily create aggregated, consistent views of data, such as risk data, from across the organization, and these views can be selectively shared with full adherence to an organization's data access and privacy policies.

Specific data virtualization capabilities for GRC and security include

- **Purpose-based processing:** Role-based access ensures that views can be reused for multiple purposes. Users and applications can access a single view but ensure that the data returned is applicable for the user's or application's purpose.

- **Consent-based processing:** Consent management systems can be integrated, and row-level and column-level policies can be applied in real time. Custom policies also have access to context information.
- **Data minimization:** Virtual models for data necessary for a given purpose can be created.
- **Data anonymization:** Views can provide anonymized reporting of data and/or allow access only to aggregated data.

LOGITECH ACHIEVES CLOUD MODERNIZATION

For several years, Logitech had been developing and delivering data services for analytics using on-premises systems. However, provisioning data services for business users has been reactive, time consuming, and inefficient. The company's modern product and service offerings, such as security video analysis and smart home devices, required predictive analytics, real-time data analytics, and cognitive science. To gain these capabilities and be able to offer the right service to business users at the right time, Logitech wanted to move IT operations to the cloud. Cloud technology would empower IT organizations to redefine the way data services are produced and delivered.

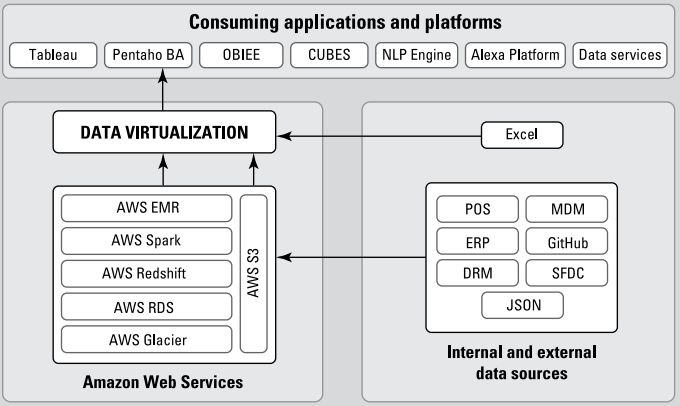
Business need

Logitech's data warehouse was not effectively meeting evolving business needs with scalability issues that impeded the company's data integration efforts. The data warehouse was also difficult to set up and tune for performance. The Logitech IT organization needed to support a growing list of data services use cases, ranging from delivering structured to unstructured data, at a velocity ranging from batch to real time, and the current infrastructure would not be able to keep pace. Logitech needed to transition IT operations to the cloud so the company could provide a more reliable, efficient, and cost-effective form of data extraction for analytics. While Logitech built its cloud-based data analytics platform, some sources of data remained on-premises. Logitech needed a solution that could seamlessly integrate all its on-premises and cloud components.

Solution

Logitech chose the Denodo Platform, hosted on Amazon AWS. In this new architecture, data from various on-premises sources and other

third-party cloud-based sources such as DRM, MDM, ERP, POS, Github, and Salesforce, are loaded into Amazon S3. Denodo also integrates data coming directly to it from on-premises Excel files, machine generated data, social media data, other Internet data.



Asurion data virtualization architecture.

After creating a single consistent data store, Denodo feeds analytics and reporting applications such as Tableau, Pentaho BA, and web services. In the Logitech infrastructure, Denodo has become the single source of truth, feeding the entire consumption layer.

Benefits

Data virtualization made Logitech's cloud journey not only possible, but possible as a live migration, with minimal impact on business operations. The Denodo platform helped Logitech hold down costs while reaching exceptional service levels. By providing the flexibility and efficiency of integrating on-premises and cloud components in real-time, the Denodo Platform has enabled business users to consume information in an easy, self-service manner. In addition, it has freed business users to become "tool agnostic" because all tools tap into the same unified view of the data. Through rapid prototyping and the optimal usage of resources, the platform has significantly reduced Logitech's operational expenses. The Denodo Platform provides a consistent data governance, security, and metadata management layer across all of Logitech's data consumers. Finally, the Denodo Platform's query optimization capabilities accelerate Logitech's data science and analytics efforts. Logitech considers this to be a critical component in the company's journey to continuous innovation.

Delivering Data Services

Data virtualization creates a unified data services layer that offers a more flexible way to architect the enterprise IT ecosystem, decoupling applications from information repositories and fostering reusability. Data services can be consumed by any application, internal users, or partners and can easily evolve by creating connections to new data sources or data combinations for new business requests.

This service-oriented approach to data has numerous benefits, including

- A consistent view of the underlying data sources, reducing the confusion caused by reports showing different results because the report developers interpreted the data differently.
- Support for multiple access protocols so that the consuming application or system can access the data service in the manner that is best for them. For example, a reporting tool can access the service using an Open Database Connectivity (ODBC) connection, and an Enterprise Service Bus (ESB) can access the same service (and same data) using a Representational State Transfer (REST) JavaScript Object Notation (JSON) web service.
- A single control point for security and governance. Being able to see who accessed the data (via the data service), when they accessed the data, and what queries they performed (as well as the result sets that were returned) is important in all industries, but this capability is particularly critical in regulated industries dealing with private or confidential data. Data virtualization provides the control point for enforcing and auditing data access through the exposed data services.

INDIANA UNIVERSITY IMPROVES STRATEGIC DECISION MAKING

As with most educational institutions, Indiana University (IU) has had a long history in business reporting and business intelligence, dating back to the days of the mainframe. Recently, the university began a new, multi-year project called the Decision Support Initiative (DSI), dedicated to helping IU improve decision making through enhanced data, models, and processes. DSI aims to improve access to data and

analytic technologies, thereby providing transparency across the IU system, leading to better-informed decision outcomes.

Business need

To empower decision makers and enhance decision making at all levels within IU, the university needed to drastically increase the availability of timely, relevant, and accurate information. Historically, data and its corresponding business logic were stored across multiple, siloed systems, making it extremely time consuming to gather and combine the relevant information decision makers needed. In some cases, data activities would fail entirely as required data elements could not be found and no common definition of sources of record were kept. Furthermore, the university's data integration toolset, primarily built around Extract, Transform, and Load (ETL) processing, required broad skillsets and scarce resources to deploy, maintain, and manage. As a result, the development time needed for information access was long — so long, in fact, that by the time data was retrieved, it was often less useful or even irrelevant for decisions.

In addition to the noted challenges of data and development timeliness, data security and privacy were also at risk within the traditional university reporting approach, as row-level access controls were integral only within the enterprise data warehouse (EDW). Other data sources and reporting environment offshoots (shadow systems) outside the EDW often lacked this same, fine-grained access control, thereby increasing the possibility of a compromise.

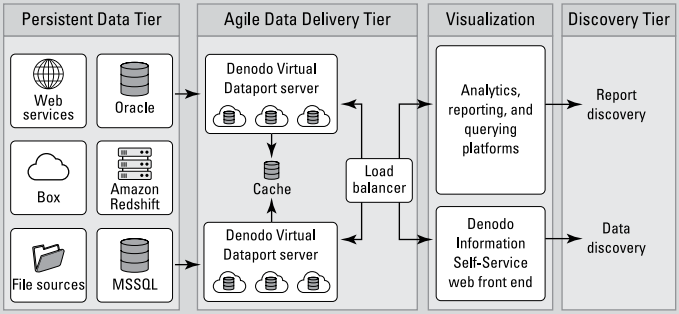
Solution

Responding to these historical challenges and preparing for the future, members of the DSI team focused on the creation of a single system of easily consumable information assets to meet the data needs of the university (see the accompanying figure). Harnessing the power of data virtualization, IU chose the Denodo platform to create a logical data warehouse (LDW). In this architecture, Denodo connects the university systems of record to data-consuming applications, providing heterogeneous data connectivity, delivery, security, and governance services. With Denodo, IU has successfully combined data sources such as Oracle, MS-SQL, Amazon Redshift, web services, and Box.com, securely serving information to consuming applications such as Tableau and Excel. Perhaps most importantly, the Denodo platform vastly improves the university's security and compliance posture, providing fine-grained access control and auditing across data sources of

(continued)

(continued)

many types. With Denodo, user management is simplified via active directory groups and Security Assertion Markup Language (SAML) integrations, providing the right access to the right users at the right time.



Indiana University's logical data warehouse.

Benefits

The Denodo platform seamlessly delivers the information required to improve outcomes while requiring fewer technical resources, improving security, and enhancing agility when compared to its traditional data warehouse forerunner. In fact, the university is leveraging the Denodo platform well beyond the LDW, satisfying data blending and access needs through use of Denodo Java extensibility in new and inventive ways. Although still in its early days, the Denodo platform shows great promise of becoming IU's enterprise platform of choice for information access and management, including self-service business intelligence. This solution provides these benefits:

- The Denodo platform has significantly improved information agility across the university. Data can now be defined and accessed almost instantaneously, no matter where it resides and with minimal effort.
- Diverse data spread across the entire enterprise can now be accessed securely with a proper authorization structure. Rules can be applied no matter when the data is accessed or where the data is stored.
- Core Business Intelligence logic is becoming centralized, thus reducing duplication of effort and enhancing development efficiency.
- IU now has a searchable data dictionary, which helps report writers find the data they need and help improve the self-service experience when using the Denodo Information Self-Service tool.

IN THIS CHAPTER

- Keeping afloat in massive data lakes
- Taking stock of data warehouses
- Analyzing big and small data in the Internet of Things

Chapter 4
Delivering Big Data Solutions That Work

In this chapter, you learn how to manage big data challenges, such as data lakes, data warehouses, and the Internet of Things (IoT), with data virtualization.

Managing Data Lakes

The expanding volume and variety of enterprise data originating from internal and external sources makes it challenging for businesses to harness their big data for actionable insights A *data lake* is a logical collection of data repositories — relational database management systems (RDBMS), enterprise data warehouses (EDWs), Hadoop, NoSQL, and so on — in which you can store data in the repository and format that is most suited for the data. In a data lake, you can process the data with different compute engines (such as Spark, SQL-on-Hadoop, or memory grids) based on the compute workload and requirements. Data virtualization creates a data discovery and access layer to hide the complexities of the "logical" data lake from the data consumers (see Figure 4-1). Data virtualization technology provides an agile and cost-effective approach to combining, governing, and managing data in data lakes, as well as to overcoming the inherent challenges presented by data lake silos.

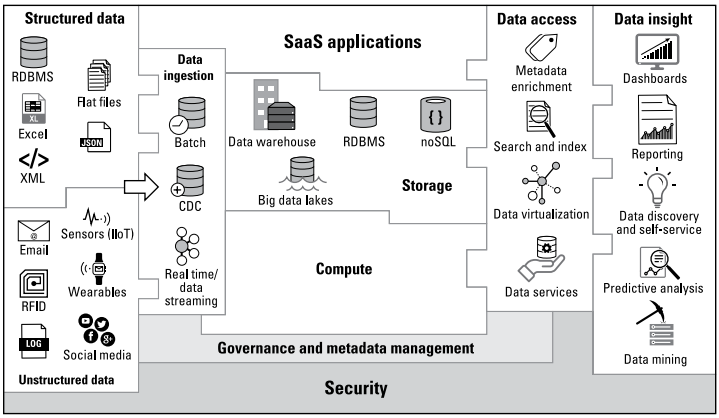


FIGURE 4-1: Data virtualization combines one or more physical data lakes with other enterprise data to create a virtual or logical data lake.

Capabilities and benefits of a single logical data lake using data virtualization include the following:

- Improves the enterprise functionality of data lakes by combining one or more data lakes with other enterprise data.
- Provides a way to access data from separate systems through an abstraction layer that makes it appear as if the data were in a single data lake.
- Improves an organization's ability to govern and extract more value from its data lakes by extending them as logical data lakes.
- Allows for easily persisting data into the lake when necessary, while preserving its lineage for easier governance. This is useful, for example, when moving data that will be used for a machine learning process.

REMEMBER

Data virtualization bridges different architectures such as data lakes, traditional data warehouses, and others. This technology enables organizations to retain existing solution investments while at the same time modernizing their data architecture to support new requirements in an agile manner.

Streamlining with a Logical Data Warehouse

Business depends on actionable intelligence, and for years this capability has been furnished by business intelligence tools that pull the data from a data warehouse. Business stakeholders have come to expect that all of the relevant data is copied to the data warehouse, even if they sometimes have to wait a day for the latest data to arrive.

With all of the recent technological advances, such as big data and cloud analytics, business analysts also expect to encounter fewer limitations in their ability to access business intelligence. Unfortunately, they are experiencing just the opposite. Newer sources, such as data from social media platforms, data from in-process transactions, or raw data about the movement of machines, isn't formatted for a traditional data warehouse. If this data were reformatted and stored in the data warehouse, the volume and cost would quickly grow. Instead, many companies are storing this data in less expensive, cloud-based storage systems like Hadoop, but the problem remains: Not all of the data is in the data warehouse at the same time, so not all of the data is available for reporting.

With a logical data warehouse, all of the data stays in place, but it appears as though it sits in a single place, so all of the data is always available for reporting. The Data Warehouse Institute (TDWI) defines the *logical data warehouse* as "a logical or virtual layer of the data warehouse architecture that integrates the physical layers of architecture under it." This logical or virtual layer sits on top of the traditional data warehouse and acts as an interpreter between the business analysts and consuming applications attempting to access a data set and the data source (see Figure 4-2).

REMEMBER

Data consumers access the data in a logical data warehouse through a layer that sits above the separate data sources, bringing them together for reporting. The logical data warehouse does not contain any actual data; it only contains the intelligence for accessing each of the sources.

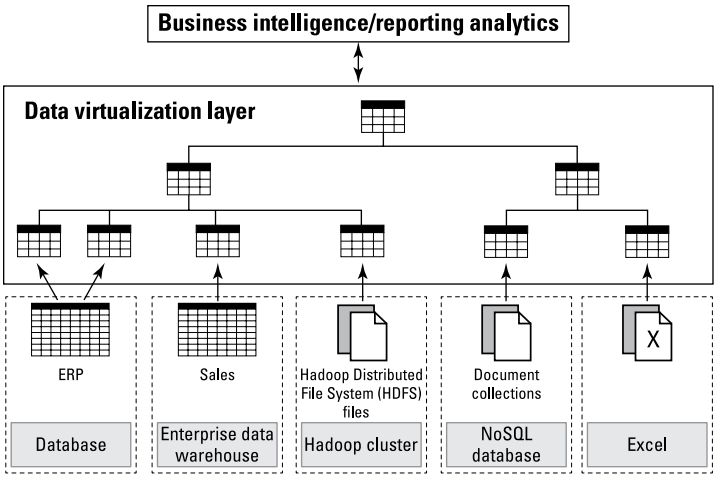


FIGURE 4-2: The data virtualization layer in a logical data warehouse.

Some common use cases for logical data warehouses include

- **Virtual data mart:** Multiple data sources are integrated to create a business-friendly data model that appears to the users and reporting tools as if the sources reside in a single place.
- **Data warehouse extended with master data management (MDM):** The MDM system stores the "golden records" for customers or products, so when business analysts pull data from the data warehouse, it is synced with the master data in real time, greatly reducing costly errors while enhancing customer service agility.
- **Data warehouse extended with cloud data:** Data from different cloud environments, such as Salesforce.com, is not always stored in the data warehouse, and this can greatly impede business intelligence. The logical data warehouse provides seamless, real-time access to virtually any data stored in the cloud, fully integrated with data stored in the data warehouse for powerful business intelligence initiatives.

- **Integration of multiple data warehouses:** Logical data warehouses can integrate two or more existing data warehouses and make them appear as one. This is a powerful capability because many companies are forced to accommodate two or more data warehouses — for example, because of a merger. Physically integrating the data warehouses, by migrating the data from the separate warehouses into a single, monolithic data warehouse, is costly and time consuming.
- **Data warehouse historical offloading:** Many companies are using cloud-based storage clusters like Hadoop as an inexpensive way to store high volumes of historical data. However, this data is then separated from data in the data warehouse for reporting purposes. In such cases, a logical data warehouse blends data from the two systems, again in real time, so you can run queries across all the data without disturbing your business processes.
- **Data warehouse extension:** Companies often want to store frequently used data on hand, in the data warehouse, but store seldom-used data in cloud-based storage. A logical data warehouse enables companies to store the data anywhere they choose, without impeding real-time business intelligence.

Business benefits of logical data warehousing include

- Gaining access to all enterprise data for business reporting in real time
- Improving upselling and cross-selling opportunities
- Enhancing the speed of operations and customer service
- Supporting data migrations without interrupting business continuity
- Maintaining independence from IT for data access

Integrating IoT Analytics

The rapid rise of the Internet of Things (IoT) has further fueled the already explosive rate of digital data growth globally. In addition to increasing the quantity of data, the IoT exponentially

increases the number of data sources and data types. These factors further contribute to the challenges of traditional data lakes and data warehouses (discussed earlier in this chapter), making data virtualization ever more pertinent.

IoT data types include structured, unstructured, semi-structured, and object data from numerous types of IoT devices such as:

- Autonomous vehicles
- Environmental sensors
- Industrial systems
- Security, surveillance, and monitoring cameras
- Smart homes
- Personal wearable technology
- Virtual reality (VR) and augmented reality (AR)

Much of this data is created, processed, and stored in the cloud. Replicating, analyzing, and accessing this data in real time is neither practical nor desirable with traditional data integration solutions. According to the last EMC-sponsored IDC Digital Universe Study, less than one percent of digital data globally is analyzed. Another challenge with IoT data is that, in addition to big data, certain types of data (sometimes referred to as "small data") have a very short useful life. This makes the need for real-time analytics to obtain value from IoT data sources ever more crucial.

Data virtualization can be used to combine streaming data from equipment sensors, stored in Hadoop, with data from a CRM system, for example, that holds customer support ticketing data and/or revenue data from sales (see Figure 4-3). This combined data can provide valuable insights to help make informed business decisions in real time, enable location-based discounts to improve customer experience, or provide sales and support information to increase retention.

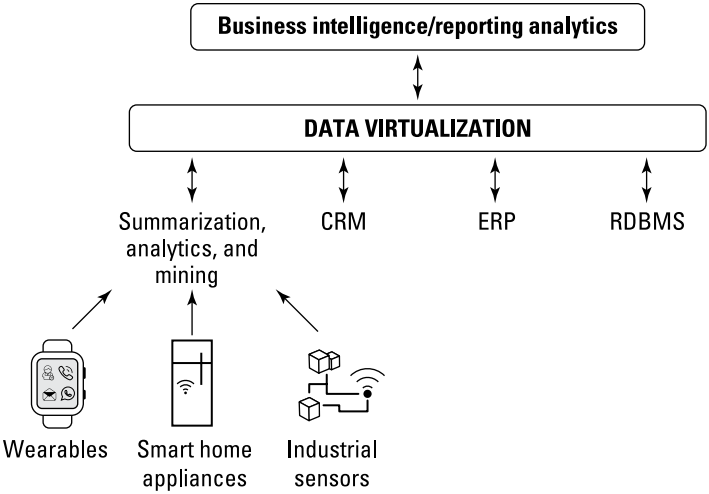


FIGURE 4-3: Combining data from IoT devices and other data sources with data virtualization.

TIP

Data virtualization solutions such as logical data lakes and logical data warehouses can enable real-time IoT analytics.

IN THIS CHAPTER

- Looking at cloud options
- Addressing application and data migration challenges
- Overcoming cloud limitations with a hybrid data hub
- Exploring cloud use cases

Chapter 5
Taking the Pain Out of Cloud Adoption

In this chapter, you take a look at how to address application and data migration challenges with data virtualization, what a hybrid data hub is, and how it benefits organizations moving to the cloud. This chapter also introduces some key use cases for data virtualization in the cloud.

Exploring the Options

Cloud technology is rapidly evolving and gaining widespread adoption. Forrester Research predicts that more than half of global enterprises will rely on at least one public cloud platform — such as Amazon Web Services (AWS), Google Cloud Platform (GCP), or Microsoft Azure — to drive their digital transformation strategies.

A recent cloud usage survey conducted by Denodo found that organizations are eager to adopt cloud-based architectures in an effort to support their digital transformation efforts, drive efficiencies, and strengthen customer satisfaction. According to the survey, the majority of those polled (76 percent) are already using the cloud, with almost half using AWS (47 percent), followed by Microsoft Azure (20 percent), and Google Cloud Platform

(13 percent). Half of respondents are implementing a virtual private cloud, with provider preferences closely aligned with those for general cloud usage. The most widely implemented used cases among survey respondents are

- **Cloud analytics:** (49 percent for AWS, 59 percent for Azure)
- **Cloud storage:** (45 percent on AWS, 29 percent on Azure)
- **Cloud data warehouse:** (40 percent on AWS, 41 percent on Azure)

Despite this rapid adoption of cloud technology, leveraging the agility and flexibility that the cloud provides can be challenging for organizations with a data architecture that consists of both on-premises and cloud data sources.

REMEMBER

Data virtualization addresses this mixed environment by creating a hybrid data fabric that spans both cloud and on-premises solutions. It also provides a unified data access and a strong security and governance layer for enterprise data. Specifically, data virtualization is being used to support

- **Cloud modernization:** Data virtualization facilitates the transition from legacy, typically monolithic applications and application suites deployed on-premises, to specialized software-as-a-service (SaaS) applications in the cloud.
- **Cloud analytics:** Data virtualization enables analytics in the cloud by facilitating the movement of data from on-premises operational systems to an analytics platform and by providing seamless access to all data.
- **Hybrid data fabric:** Data virtualization provides a hybrid data fabric by seamlessly integrating data across applications on-premises and in the cloud.

Minimizing the Impact

Many companies invest in large on-premises business suites with tightly integrated components that can be tailored to meet the needs of the business so users can be assured that they are accessing consistent, authoritative data sets.

However, such systems are extremely costly, and companies are moving to newer, less-expensive cloud-based SaaS solutions. Many companies that once relied on Oracle E-Business Suite, PeopleSoft, or Siebel are now experimenting with tools such as Salesforce, Netsuite, Workday, or Taleo.

Similarly, companies are moving to cloud-based repositories that can dynamically and cost effectively scale up or down to avoid having to purchase new storage hardware.

Unfortunately, both application and data migrations are fraught with challenges, and companies must overcome these challenges if they hope to take full advantage of cloud benefits.

Application migration challenges include

- Because they have moved from a single monolithic application to a set of individual applications, they no longer have a holistic view of the data across the various components. They can download the data from multiple applications via application programming interfaces (APIs) and then merge it, but this is time consuming and complex.
- With multiple applications, it will be difficult to maintain control over the number of licenses that a company owns. If, for example, a large number of marketing stakeholders want to access Salesforce simply to read about customer trends, it might be difficult to justify licenses for all of them.

Data migration challenges include

- During the migration, certain operations such as operational reporting, analytics, and ad hoc reporting may be disrupted because the underlying infrastructure is being moved.
- After the migration, if problems occur, it is often difficult and time-consuming to revert to the old system, which further extends the downtime.
- If a company needs to retain certain data on the on-premises system for compliance or any other purpose, providing simultaneous access to both systems is difficult.

REMEMBER

Data virtualization overcomes each of these challenges. With data virtualization, companies can seamlessly move to cloud-based apps and migrate data to cloud-based repositories, thus gaining all of the benefits that cloud solutions can provide.

Most data integration technologies first copy the data and then move the copy to a new, consolidated repository. Data virtualization, in contrast, provides real-time views of the integrated data without replication.

Because data virtualization provides access to data in real time, from a variety of systems that are normally very time consuming to integrate, such as transactional processing systems, cloud-based storage systems, and SaaS applications, it can support a wide variety of uses, such as:

- **Integrating SaaS applications, in real time:** By establishing a data virtualization layer above a company's myriad SaaS applications, companies can simulate the performance of a monolithic, on-premises business suite. The data virtualization layer can resolve inconsistencies and establish rules of authenticity. A data virtualization layer can even loop through a company's master data management (MDM) system for this purpose.
- **Reducing SaaS licensing fees:** Users access the data virtualization layer which, in turn, accesses the appropriate SaaS applications. Companies therefore can dramatically reduce the number of required licenses. Companies can also use the virtualization layer to automate read-only feeds that cut across multiple apps.
- **Enabling zero-downtime migrations to the cloud:** Because the data virtualization layer abstracts data consumers from the complexities of accessing the various data sources, users do not have to know where the data is stored, and migrations can occur even without users' knowledge. Companies can move the data according to their own schedules, without affecting day-to-day operations.
- **Enabling hybrid infrastructures:** Because the data virtualization layer abstracts data consumers from access complexities, legacy and modern systems can exist simultaneously, in a hybrid configuration, to meet compliance requirements or simply to accommodate a company's preference.

Creating a Data Gateway

Organizations are moving to cloud-based infrastructure to enable SaaS solutions and other service-oriented initiatives and forgo many of the costs associated with maintaining on-premises infrastructure. However, transitions to the cloud are not easy for the following reasons:

- **The silos remain.** Cloud storage is inherently limitless, so many organizations begin their migrations to the cloud by moving data, department by department, into single, monolithic, cloud-based data lakes for analytics purposes. However, they cannot query across the entire data assets within a data lake because the data is stored in different formats and schemata. Additionally, data lakes cannot exist in a vacuum. Enterprises also rely on SaaS applications, operational applications and other data repositories for their information needs, thus relegating a data lake to become just another data silo.
- **Data is delayed by network congestion.** Finally, because data is dispersed among many different servers connected by the network, it is, by its very nature, subject to latency, based on the amount of traffic taking up available bandwidth.

Hybrid data hubs (see Figure 5-1), enabled by data virtualization, overcome common cloud challenges. A hybrid data hub provides seamless, real-time access to on-premises, cloud, and SaaS sources, all without data replication and its associated costs and risks. Because of data virtualization, the hybrid data hub doesn't store any replicated source data; instead, it contains the metadata required to access each of the applicable sources. This technology creates unified, virtual views across all of the various heterogeneous sources, including the silos within data lakes. For SaaS sources, it abstracts each of their associated APIs so that organizations can point any of their standard reporting tools directly at the hub, which in turn gets the data directly from the SaaS application.

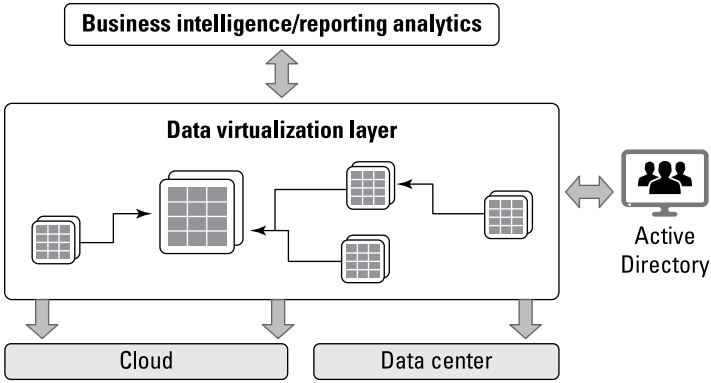


FIGURE 5-1: Hybrid data fabric architecture.

Hybrid data hubs seamlessly unify data stored in both on-premises and cloud sources. Because users are abstracted from the complexities of accessing each source separately, they suffer no impact when organizations switch sources or move the data from one to another. For this reason, hybrid data hubs enable organizations to migrate to cloud infrastructure on their own time, with zero downtime, and no impact on day-to-day operations.

Because consumers access all data through the hybrid data hub, it provides a single gateway from which to manage all security protocols, greatly simplifying the security challenge, and enabling security to be more effectively monitored and controlled.

Additionally, advanced hybrid data hub solutions can scale as needed to meet the changing needs of any organization and help IT avoid purchasing unnecessary server time.

Finally, advanced hybrid data hub solutions offer specific features for overcoming network latency such as caching with incremental queries. With this feature, a user's query across the hybrid data hub delivers three results:

- Cached data, taken at regular intervals
- Changed data, based on user defined thresholds
- A merge of both of cached and changed data

This results in a significant reduction in latency, simply by caching the static data and only reporting on the changed data.

TIP

Benefits of a hybrid data hub include

- Integrating data from silos within data lakes, from cloud and on-premises sources simultaneously, from SaaS applications and other sources, without replication, in real time
- Reducing the inherent latency of accessing data from network sources
- Enabling migrations from on-premises to cloud sources with no impact on day-to-day operations
- Tracking cloud usage by department or individual
- Managing cloud and on-premises security from a single point
- Scaling the solution only as needed, to hold down costs

Looking at Cloud Use Cases

There are many possible use cases for data virtualization in the cloud. This section introduces some of the key use cases.

Analytics in the cloud

With big data technologies such as Hadoop and Spark making advanced analytics more affordable, many organizations have started advanced analytics initiatives such as predictive analytics, recommendation engines, chatbots, and so on. These initiatives are forcing organizations to decide whether they want to continue with their existing analytics infrastructure — typically on-premises and based around a data warehouse — or whether they should move some, if not all, of the analytics processing to the cloud.

REMEMBER

The benefits of moving analytics processing to the cloud include increased agility and flexibility, as well as lower costs. Moving analytics processing to the cloud allows data to be stored in the storage or database most appropriate for the data and its uses. This is a better option than trying to force-fit all data into a predefined schema, such as a data warehouse schema.

Cloud data gateways

With more applications and data being moved to the cloud, it's becoming more difficult to find the data that users need. The disparate nature of these applications and cloud-based data stores makes integrating the data for reports and dashboards all the more difficult. Providing a data integration platform to integrate the data from these different sources and make it readily available to users accelerates the time to extract value from the data.

Cloud modernization

As part of their overall cloud strategy, many organizations are migrating from monolithic on-premises application suites to lower cost software as a service (SaaS) applications. In many cases, moving to cloud-based SaaS applications results in using multiple SaaS applications to provide the same breadth of functionality as the on-premises suite.

For example, moving from a large on-premises HR suite might result in multiple SaaS applications — for payroll/compensation, benefits, recruitment, general timesheet management, and so on. This is because the SaaS applications are typically tightly focused on one aspect of the overall function rather than trying to replicate the complete functionality of the suite. After migrating to SaaS applications, the data becomes fragmented and spread across multiple applications. Trying to get a complete and comprehensive view across the various SaaS applications becomes challenging and a point of frustration for the users.


TIP

Using a data access layer to connect to the various applications and expose a set of unified views of the application data makes it easier for users to get to the data that they need. The data access layer can be "front-ended" with an enterprise portal that acts as the user interface.

Multi-cloud integration

Many companies are moving to the cloud, but not just a single cloud. They are moving to multiple clouds, whether clouds from multiple infrastructure providers or simply multiple regions within the same cloud infrastructure. The RightScale *2018 State of the Cloud Report* found that 81 percent of enterprises have a

multi-cloud strategy consisting of an average of five clouds. The result is that data is spread across different clouds or different cloud regions, thus creating more data silos.

In most situations, the usual approach of copying the data into a centralized, consolidated repository is not an option. The volumes of data can make simply replicating the data impractical and, in the case of different cloud regions, data privacy regulations might prohibit the copying of data out of the local jurisdiction.

REMEMBER

Using data virtualization to provide access to remote data in other clouds or other cloud regions enables users to access the data without copying it. The data remains where it is stored — in the other cloud or cloud region — and only the data required by the user is retrieved on demand.

Cloud migration

Many businesses are migrating their data to the cloud to take advantage of the lower cost and dynamic scalability of cloud infrastructure. Data repositories in the cloud are no different in this respect and can dynamically scale up or down, depending upon the workload and policies defined by the user. This makes these data services very attractive and affordable to many organizations. However, migrating data from on-premises databases and data warehouses to the cloud can be a major project with many potential pitfalls.

Migrating small data sets is usually a straightforward "lift and shift" operation (that is, copy the data from the on-premises database then switch over to the cloud database). This approach obviously has an impact on the tools and applications connecting to the on-premises database, but these can be switched over to the cloud database in a controlled and managed fashion, if necessary.

Larger data sets are a different proposition. It's simply not possible to copy all of the data to the cloud data store in a timely manner. It will take long enough to copy the data that the data copied first will become out of sync with the on-premises data. This results in synchronization processes having to be created, which further complicates the whole migration project.

TIP

A less risky and more manageable approach is to migrate the data to the cloud using a phased approach. Typically, the less critical and less frequently accessed data will be migrated first. This is low risk and allows the migration team to test how the cloud data warehouse performs and scales before moving on to more critical data. This phased approach reduces the risk, but it also results in live data being located in both the on-premises data store and the cloud data warehouse while the migration is occurring.

REMEMBER

Using data virtualization as a data access layer can hide the migration from the users. They simply connect to the data virtualization platform to access the data. Whether the data is still in the on-premises data store or has been migrated to the cloud is an issue for the data virtualization platform to handle — the user doesn't need to know.

IN THIS CHAPTER

- Meeting business objectives and goals
- Focusing on quick wins
- Creating a center of excellence
- Delivering digital transformation

Chapter 6
Getting Started with Data Virtualization

In this chapter, you take the first steps toward data virtualization in your organization, from aligning your project with business objectives to driving digital transformation.

Aligning with Business Objectives

As with any project or initiative, data virtualization must be aligned with clearly defined business objectives and goals to ensure the success of the project. Specific business objectives might include

- Enabling self-service, real-time analytics for business users
- Improving customer experience with a 360-degree view of the customer
- Strengthening data governance, compliance, and security throughout the organization
- Improving data quality and confidence with a "virtual" single source of truth
- Eliminating data silos and reducing data duplication, replication, and storage, and their associated costs

Practically every modern organization has data management and integration challenges that must be addressed. Although data virtualization can help to address many of these challenges, it's important to prioritize each business and technical challenge to ensure you can focus on the most important issues for your organization. You can then align your business objectives and priorities with the most important data virtualization capabilities to focus on for your organization, such as:

- Delivering a single data access point for all data in the organization, regardless of where it resides
- Providing real-time operational data to support up-to-the-minute data requirements
- Unifying data security across the organization
- Decoupling analytics and applications from their physical data structures to minimize end-user impact during data infrastructure changes
- Enabling federated joins of data residing in disparate locations, thereby reducing the need to copy and restore data
- Creating a bridge between big data and relational data sources

Starting Small with a Clear Focus

Data virtualization doesn't have to be an all-or-nothing proposition. Remember, you don't have to "boil the ocean" with data virtualization. After building a data virtualization layer, you can begin to add data sources — based on your business priorities — as needed. Once you are comfortable that you have achieved the desired result, you can move on to the next data source until you've addressed all of your enterprise-wide data needs.

TIP

Identify a small pilot project that will quickly demonstrate the benefits that data virtualization can deliver to the wider organization.

It is also important to keep your goals in mind. Think of the ideal data architecture you would like to deploy, and work step-by-step toward that architecture. In many large companies, the final goal is a unified data delivery layer that completely abstracts any changes in the sources from the impact it would have on the users. This isn't an easy target, but with each project you can move one step closer to your ideal data architecture.

As you begin to introduce data virtualization in your organization, consider the following to ensure that you achieve the maximum benefits from data virtualization:

- Architect with an enterprise perspective from the beginning, working with your data architecture teams.
- Determine who should be responsible for the data virtualization platform. With the ability to harness data virtualization for many use cases — such as agile business intelligence (BI), deploying web services, and customer 360 — you need to involve many teams including data architects; Extract, Transform, and Load (ETL) administrators; BI teams; database administrators; and others.
- Leverage data virtualization capabilities that allow you to implement robust user security and governance, such as masking sensitive data and limiting which rows specific users can access. These capabilities help you mitigate potential security risks and governance challenges that may be created by exposing data to a wider community of users with data virtualization.
- Harness data virtualization catalog and lineage capabilities to help IT and user communities understand where specific data comes from, as well as to build trust across the user communities in the solutions they are using.

Laying the Foundation for a Center of Excellence

A successful data virtualization deployment needs proper management, similar to what a database or a data warehouse requires. With data virtualization as a key component of your data architecture, you need to have a focused team. Typical responsibilities managed by the center of excellence include

- Providing best practices and architecture guidelines that guide the usage of the technology in the company
- Supporting development by providing usage expertise and performance fine tuning

- Maintaining and managing a solid monitoring and auditing strategy that guarantees smooth operations
- Securing the system to make sure that data is exposed only to the roles and users that have the proper credentials

In short, the center of excellence ensures that the architecture is used properly and the data virtualization software is fully supported and governed.

In addition, given the position of data virtualization as middleware, the center of excellence acts as a key mediator between the owners of the data sources (such as enterprise data warehouses, customer relationship management, and others) and the business stakeholders who request access to the data.

Fueling the Digital Revolution

Digital business transformation has changed the world of data analysis and accelerated the speed at which business demands data. An enterprise data warehouse (EDW) is best suited for data and information needs that are stable, which contrasts sharply with today's dynamic and rapidly changing business environment. Physical consolidation of all relevant data into a data warehouse cannot keep up with the rapid pace of diagnosis, discovery, and decisions that characterize modern business management.

A digital marketplace that enables data users to shop for relevant, contextual data to meet their needs is replacing EDW as the primary business-facing solution for data access. A digital marketplace includes data assets from EDW and many other sources such as:

- Data warehouses
- Data lakes
- SaaS applications
- Master data management (MDM) repositories
- Operational data stores
- Enterprise resource planning (ERP) systems
- Legacy data stores
- Third-party data providers

In contrast to the outdated notion that data must be stored, processed, and managed as a technical asset, a digital marketplace accommodates a modern data paradigm wherein data is viewed as a service and a knowledge asset that is accessed, processed, analyzed, and reported by everyone who has a need for information; that is, a reusable and shareable business resource.

REMEMBER

A digital marketplace is a single storefront that facilitates access by various kinds of users to curated and relevant data in their preferred format regardless of where it is stored.

Data virtualization enables such a digital marketplace through its ability to rapidly integrate a myriad of data sources and deliver data assets to the marketplace. Data virtualization facilitates access to real-time data and provides for the addition of new data assets to the marketplace in an agile manner. With data virtualization, the digital marketplace can meet the needs and preferences of individual analysts as well as enterprise-wide use cases. Additionally, data virtualization secures and governs access to all data assets in the marketplace.

TIP

Get started with a proof of concept project using a free trial of Denodo Platform in AWS or Azure. You can also test drive data virtualization with Denodo Express, which is free to download and use. Go to `www.denodo.com/en/denodo-platform/denodo-express` to get started.

IN THIS CHAPTER

- Integrating and managing data cheaper and faster
- Enabling self-service and working with traditional data warehousing
- Securing data and saving valuable time
- Testing deployments and enabling a big data fabric

Chapter 7
Ten Things You Need to Know about Data Virtualization

Here are ten things you need to know about data virtualization:

- **It is cheaper to maintain than traditional integration tools.** Physically replicating, moving, and storing data multiple times is expensive. Data virtualization creates a virtual data layer, which eliminates the need for replication or storage costs.
- **It is a faster way to manage data.** Rather than having to wait hours or even days for your results with traditional data integration methods, data virtualization provides results in real time.
- **It maximizes performance.** Poor performance is often due to network latency (that is, the delay before a transfer of data begins). Data virtualization connects directly to the source and provides actionable insight in real time.

- **It goes far beyond data federation.** Data virtualization is a superset of data federation technology. It includes the advanced capabilities of performance optimization as well as self-service search and discovery.
- **It enables self-service business intelligence (BI).** Data virtualization can empower business users to leverage data on their own rather than always having to rely on the technical team.
- **It complements traditional data warehousing.** Data virtualization works alongside and complements traditional warehousing tools.
- **It ensures secure data governance.** Data virtualization enables a centralized point of access to all kinds of information in the enterprise, enabling security management, data governance, and performance monitoring.
- **It offers a great return on investment (ROI).** A typical data virtualization project pays back within six months of implementation. With data virtualization, businesses can achieve 50 to 80 percent time savings over traditional integration methods.
- **It is more agile than traditional methods.** Data virtualization technology includes prototyping capabilities, meaning you can test out your strategy before implementing it on an enterprise scale.
- **It gives the right context to big data fabric.** Big data fabric enabled by data virtualization integrates data, prepares it for predictive analytics, and makes it available to the consumer in real time.


TIP

Data virtualization also requires far fewer developers than traditional Extract, Transform, and Load (ETL) processes. For every four ETL developers, you only need one data virtualization developer.

denodo

SAVE TIME AND MONEY
WITH DATA VIRTUALIZATION

Data virtualization is the fastest and most cost-effective way to integrate all your data, making it easily accessible from a single abstraction layer

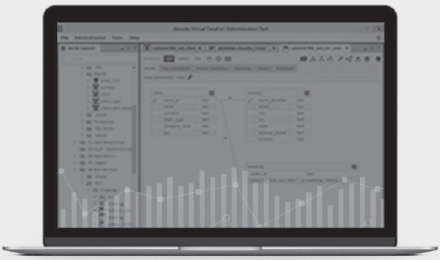
- **Agile BI**
- **Cloud Modernization**
- **Data Services**
- **Logical Data Warehouse / Lake**
- **Single View of Customer**

- Reduce integration time by **85%**
- Decrease costs by **80%**
- Realize ROI within **3 months**

Take Denodo for a Test Drive on The Cloud

Experience the benefits of data virtualization for agile BI and analytics

www.denodo.com/TestDrive



Integrate trusted business data

With the advent of big data and the proliferation of multiple information channels, organizations must store, discover, access, and share massive volumes of traditional and new data sources. Data virtualization transcends the limitations of traditional data integration techniques such as ETL by delivering a simplified, unified, and integrated view of trusted business data. This book is your guide to putting data virtualization to work in your organization.

Inside...

- Conquer siloed data in the enterprise
- Integrate data sources and types
- Cope with regulatory requirements
- Explore data virtualization use cases
- Deliver big data solutions that work
- Take the pain out of cloud adoption
- Drive digital transformation

denodo

Lawrence C. Miller has worked in information technology for more than 25 years. He has written more than 60 For Dummies books.

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

ISBN: 978-1-119-55849-1
Not For Resale

for dummies®
A Wiley Brand

Also available as an e-book

WILEY END USER LICENSE AGREEMENT

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