denodo

Whitepaper

DATA VIRTUALIZATION:

The Modern Data Integration Solution

Contents

Abstract	3
Introduction	4
Traditional Integration Techniques	5
Data Virtualization	6
The Five Tiers of Data Virtualization Products	8
Summary: 10 Things to Know about Data Virtualization	9
Denodo Platform: The Modern Data Virtualization Platform	10

Abstract

Organizations continue to struggle with integrating data quickly enough to support the needs of business stakeholders, who need integrated data faster and faster with each passing day. Traditional data integration technologies have not been able to solve the fundamental problem, as they deliver data in scheduled batches, and cannot support many of today's rich and complex data types. Data virtualization is a modern data integration approach that is already meeting today's data integration challenges, providing the foundation for data integration in the future. This paper covers the fundamental challenge, explains why traditional solutions fall short, and introduces data virtualization as the core solution.

Introduction

We have been living through an unprecedented explosion in the volume, variety, and velocity of incoming data. This is not new, but emerging technologies, such as the cloud and big data systems, which have brought large volumes of disparate data, have only compounded the problem. Different sources of data are still stored in functional silos, separate from other sources of data. Today, even data lakes contain multiple data silos.

Business stakeholders need immediate information for real-time decisions, but this is challenging when the information they need is scattered across multiple sources. Similarly, initiatives like cloud first, app modernization, and big data analytics cannot move forward until data from key sources is brought together as a unified source. Unfortunately, traditional data integration techniques have proven to be resource intensive, time consuming, and costly.

Traditional Integration Techniques

Most data integration approaches involve an extract, transform, and load (ETL) process, or a closely related process. ETL processes were introduced as early as the 1970s, and though they have matured and grown diverse over the years, they follow three basic steps that correspond with their name:

1. First, the data is extracted from the sources.
2. Next, the extracted data is transformed into the format and structure required by its final destination.
3. Finally, the transformed data is loaded into its final destination, be it an operational data store, a data mart, or a data warehouse.

ETL processes are not one-size-fits-all solutions. Each is carefully scripted and tested to accommodate the unique requirements of each individual source and the final target system.

Some processes do the transformation in the final step, and are therefore called "ELT processes," but the basic concept is the same: Once the scripts are written and the processes are tested, they copy large amounts of data from one or more sources and replicate it in a single, consolidated system, via a series of scheduled batch processes, applying all of the necessary transformations along the way.

ETL processes offer a number of distinct advantages, which is why they are still being used today:

- They are extremely 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 or batch data movement.
- Most organizations have ETL competencies in-house.

In recent years, however, as the data landscape has grown more complex, and as the need to gain actionable intelligence from consolidated data has become more acute, organizations are learning that ETL processes also have certain disadvantages:

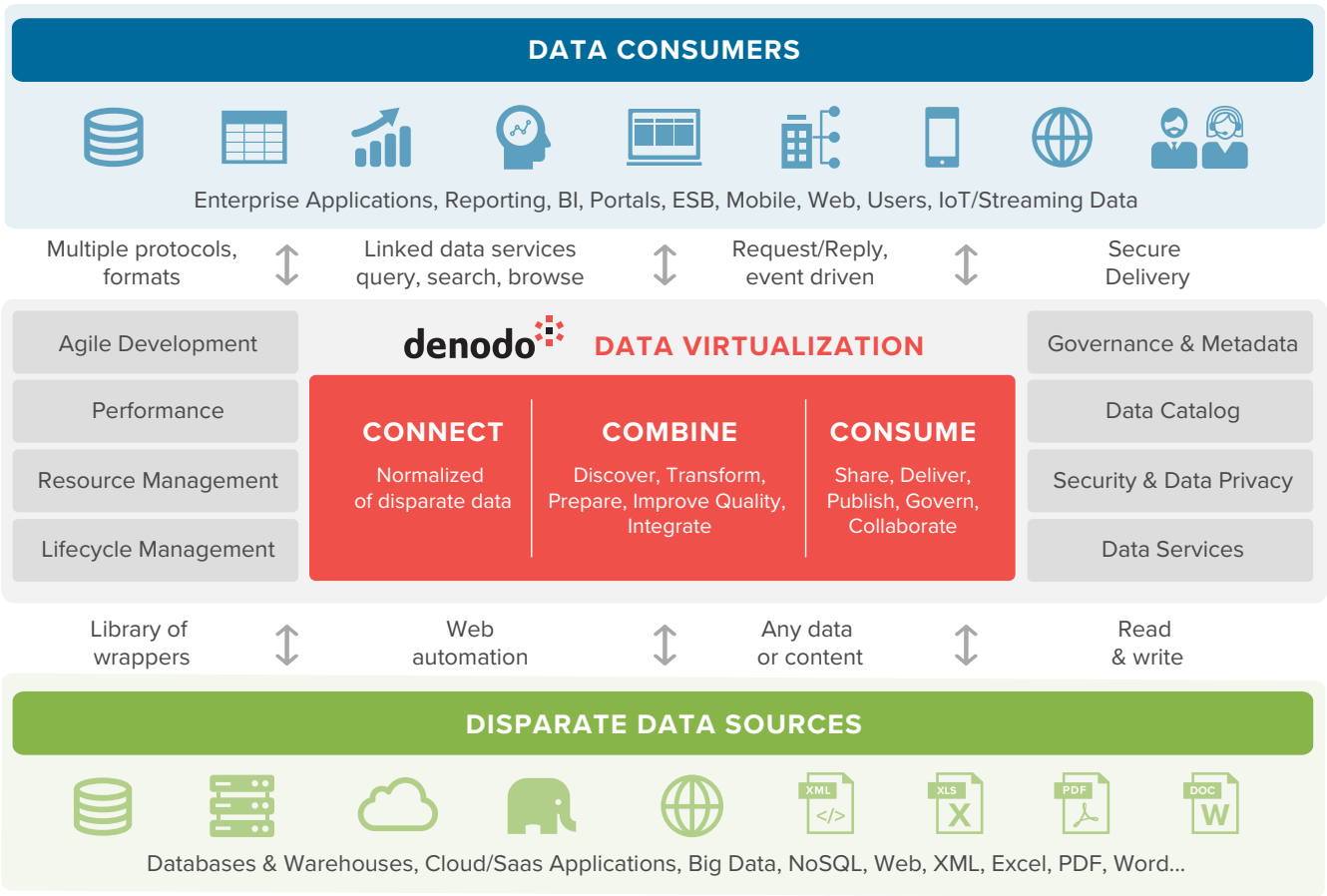
- Moving data is not always the best approach, as this results in a new repository that needs to be maintained, which is both resource-intensive and costly.
- Large organizations can have thousands of ETL processes running each night, synchronized by scripts that are difficult to modify if needed.
- Because ETL processes deliver data in scheduled batches, the end user needs to wait until the data has been delivered. Depending on the configuration and schedule, batches can be delivered quickly, but never on an instantaneous, as-needed basis. For this reason, many ETL processes are set up for overnight delivery.
- ETL processes cannot handle today's data volumes and complex data types, such as minute-by-minute transactional data or the fluctuating readings from machine sensors.

Data Virtualization

Data virtualization is a data integration strategy that uses a completely different approach: Rather than physically moving the data to a new, consolidated location, data virtualization provides a real-time view of the consolidated data, leaving the source data exactly where it is.

Sophisticated data virtualization solutions go a step further: They establish an enterprise data-access layer, providing universal access to all of an organization's critical data sources. When business users need to access data, they query the data virtualization layer, which in turn gets the data from the applicable data sources. Because the data virtualization layer takes care of the data-access component, it abstracts these users from complexities such as where the data is stored or what format it is in. Depending on how a data virtualization layer is implemented, it can enable business users to simply ask questions and receive answers, letting the data virtualization layer handle the underlying complexity.

In most cases, these seamless, "self-service" scenarios will not involve business users querying the data virtualization layer directly; instead, they would most likely interact with an application, a Web portal, or another type of user-centered interface that will, in turn, get the required data from the data virtualization layer. The essential architecture is that the data virtualization layer sits between all data sources on one hand, and all data consumers on the other, be they individuals or applications, as shown in the diagram below:



It is important to note that because data virtualization does not replicate any data, the data virtualization layer itself contains no data; instead, it merely contains the metadata required to access the various sources. This architecture has many advantages beyond the fact that the data virtualization layer is “light” and therefore easy to implement. For one, it means that enterprise-wide access controls can easily be applied at the data virtualization layer, rather than at each and every source system. It also provides a central location to which developers can connect APIs for accommodating different sources, from the most structured the least structured.

Data virtualization is therefore a modern data integration strategy. It performs many of the same transformation and quality-control functions as traditional data integration solutions, but it provides real-time data integration at a lower cost, with faster speeds and increased agility. It can either replace traditional data integration processes and their associated data marts and data warehouses, or simply augment them, extending their capabilities.

As an abstraction layer and as a data services layer, data virtualization can be easily leveraged between original and derived data sources, ETL processes, enterprise service busses (ESBs), and other middleware, applications, and devices, whether on-premises or cloud-based, to provide flexibility between layers of information and business technology.

Clearly, data virtualization offers distinct advantages over traditional, replication-based data integration approaches:

- It can seamlessly federate two or more disparate data sources (making them appear and function as one), including a mix of structured and unstructured sources.
- It can support value-added features such as intelligent real-time query optimization, caching, in-memory processing, and custom optimization strategies based on source constraints, application needs, or network awareness.
- Via an API, 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 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.

Data virtualization has but one disadvantage: Unlike ETL processes, it doesn’t support bulk or batch data movement, which might be required by a few use cases. However, as noted above, data virtualization can be easily implemented alongside ETL processes.

The Five Tiers of Data Virtualization Products, from "Feature" to "Enterprise Platform"

As data virtualization solutions gain in popularity, some of their features are being included in other products, or as an add-on module or feature. It is therefore important to distinguish between an add-on, or a built-in data virtualization product, and a full-fledged enterprise data virtualization platform, which is capable of establishing an enterprise data-access layer as described above.

THE FIVE TIERS OF DATA VIRTUALIZATION PRODUCTS ARE:

Data blending features. These are often included as part of a business intelligence (BI) tool. Data blending combines multiple sources to feed the BI tool, but the output is only available for that tool alone and cannot be accessed from any other external application for consumption.

Data services modules. These are often offered for an additional cost by data integration suite or data warehouse vendors. They provide robust data modeling and transformation, but their query optimization, caching, virtual security layers, support for unstructured sources, and overall performance, tends to be weak. This is because they are often designed to prototype ETL processes or master data management (MDM) tools.

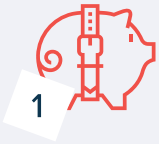
"SQLification" products. This is an emerging category, particularly among big data and Hadoop vendors. They virtualize underlying big data technologies and enable them to be combined with relational data sources and flat files so they can be queried using standard SQL. This can be effective for the big data stack, but not beyond.

Cloud data services. These are often deployed in the cloud and have pre-packaged integrations to SaaS and cloud applications, cloud databases, and few desktop and on-premises tools like Microsoft Excel. Unlike a true data virtualization product, however, with tiered views and delegatable query execution, these products expose normalized APIs across cloud sources for easy data exchange in projects of medium volume. Projects involving big data analytics, major enterprise systems, mainframes, large databases, flat files, and unstructured data are beyond the scope of such services.

Data virtualization platform. These are built from the ground up to provide data virtualization capabilities for the enterprise in a many-to-many fashion through a unified virtual data layer. Data virtualization platforms are designed for agility and speed across a wide range of use cases, agnostic to sources and consumers, and they both out-perform and collaborate with other middleware solutions.

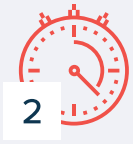
Summary: 10 Things to Know about Data Virtualization

This brief paper covered the benefits of a data virtualization platform, which can either augment an existing traditional data integration solution or completely replace it. To summarize, here are 10 things to know about data integration:



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 waiting hours or days for results, data virtualization provides results in real time.

It complements traditional data warehousing.

Data virtualization can be implemented right alongside existing data warehouse solutions.

It maximizes performance.

Performance is often hampered by the delay before a data transfer begins. Data virtualization connects directly to the source and provides actionable insight in real time.

It enables self-service BI.

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 enables secure data governance.

Data virtualization establishes a centralized access point for all kinds of information and metadata in the enterprise, enabling security management, data governance, and performance monitoring.

It goes far beyond data federation.

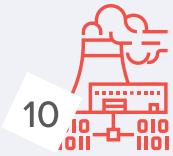
Data virtualization is a superset of the ten-year-old data federation technology. Unlike data federation, it offers performance optimization as well as self-service search and discovery.

It offers significant ROI.

A typical data virtualization project pays back in less than six months of implementation. With data virtualization, business can achieve 50% to 80% time savings over traditional integration methods.

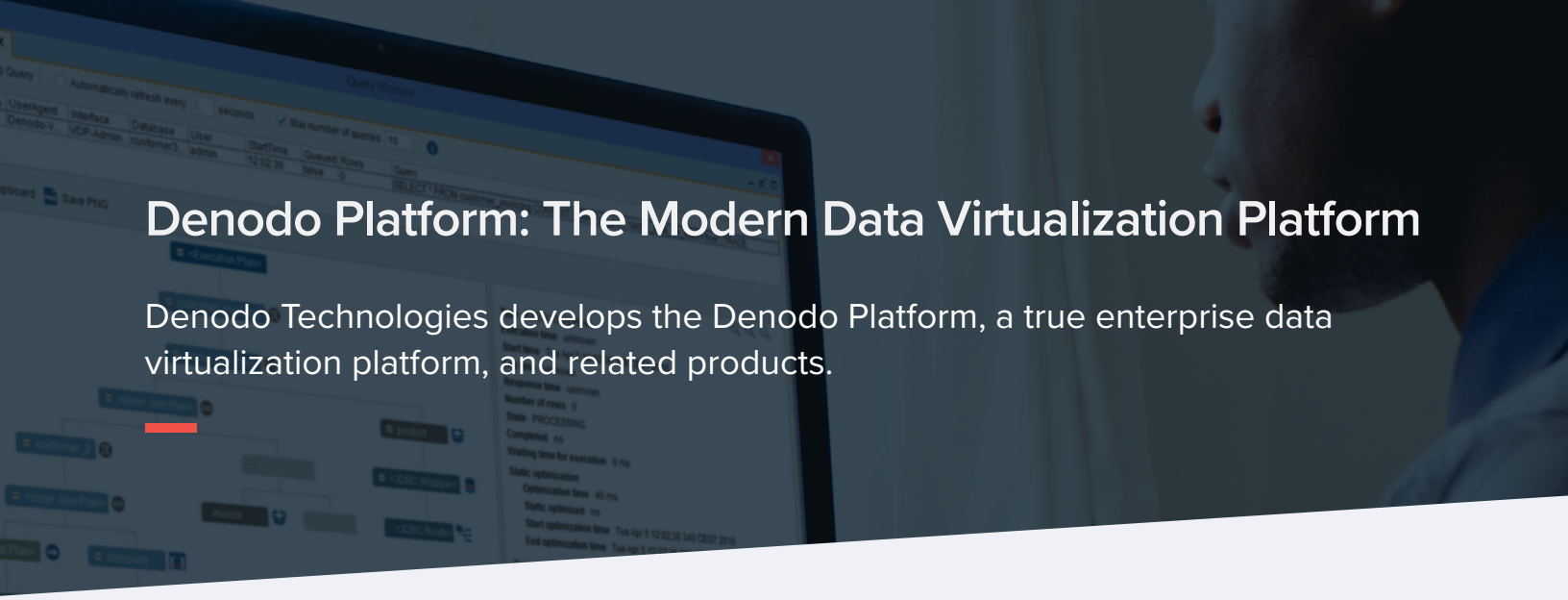
It is more agile than traditional methods.

Data virtualization enables seamless prototyping, or the ability to test out a strategy before implementing it on an enterprise scale.



It provides 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 consumers in real time.



Denodo Platform: The Modern Data Virtualization Platform

Denodo Technologies develops the Denodo Platform, a true enterprise data virtualization platform, and related products.

THE DENODO PLATFORM GOES BEYOND EVERY OTHER DATA VIRTUALIZATION SOLUTION, OFFERING:

A Dynamic Data Catalog, providing seamless access to data via a searchable, contextualized interface.

The Dynamic Query Optimizer, which intelligently chooses the optimal query strategy for each execution, for faster access to data.

In-memory parallel processing, which further accelerates data access, now to unparalleled speeds.

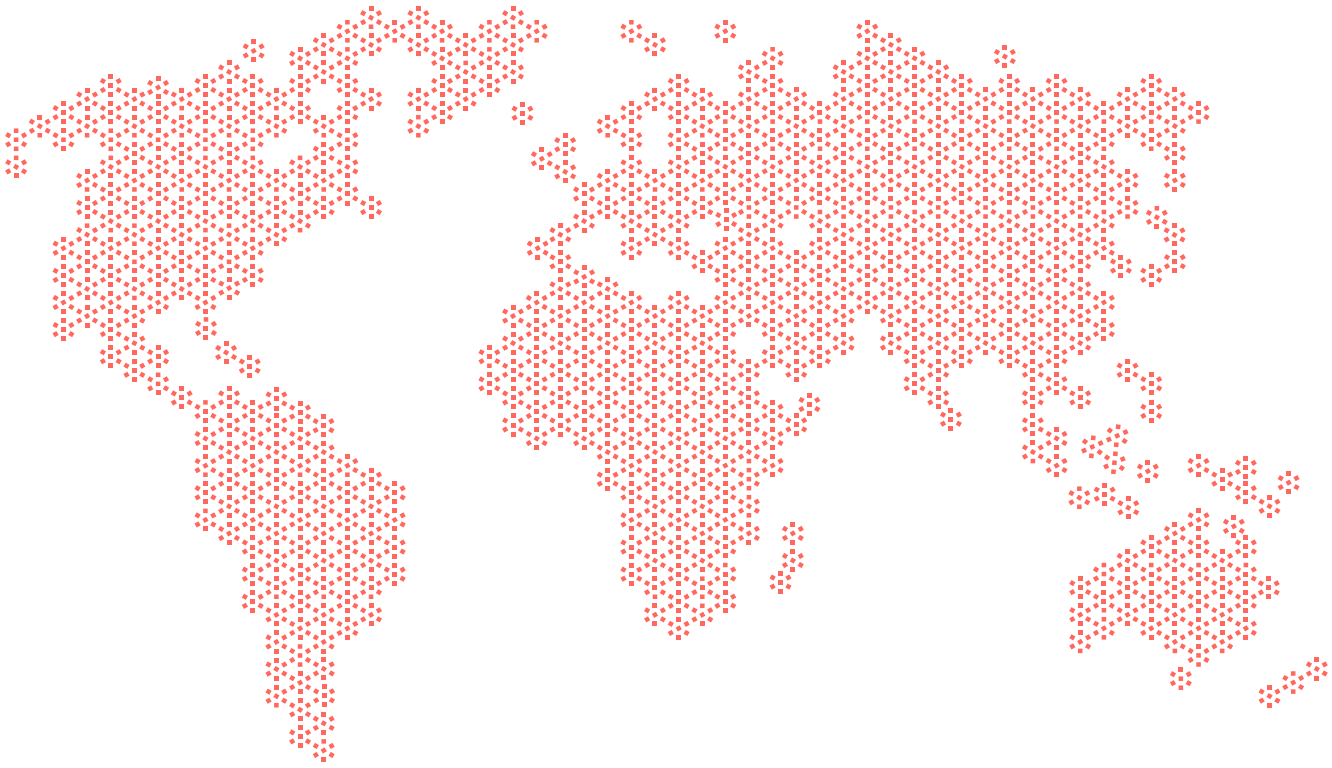
A completely redesigned interface, targeting the special needs of business and IT stakeholders.

A suite of automated lifecycle management features, so users can spend less time managing data, and more time leveraging data to make decisions.

A modern data services layer that supports OAuth 2.0, SAML, OpenAPI, OData 4, and other cloud standards for easy interoperability with current cloud systems.

Seamless security and governance, by providing secure, selective access to an organization's entire data holdings via a single point of control and administration.

Availability on leading cloud marketplaces such as Amazon Web Services (AWS) and Microsoft Azure, as well as on Docker.



denodo

Denodo Technologies is the leader in data virtualization providing agile, high performance data integration, data abstraction, and real-time data services across the broadest range of enterprise, cloud, big data, and unstructured data sources at half the cost of traditional approaches. Denodo's customers across every major industry have gained significant business agility and ROI.

Visit www.denodo.com | Email info@denodo.com | Discover community.denodo.com