Data Virtualization: a Standardized Front Door to Company-Wide Data Opens the Way for (Digital) Business Success

62

Christian Kurze, Michael Schopp, and Paul Moxon

Abstract

Marc Andreessen famously said that software was "eating the world," and today, this rings particularly true, as many companies, if not most, are transforming into software-oriented companies. All major industries are going digital and are shifting their emphasis from hardware to software, from product to service, and from process to data and analytics. However, today's digitization initiatives face a number of problems. Complex IT landscapes have been built during the last 10 to 20 years, and although the architectures are in place and fulfill their current purposes they result in data silos from which data cannot easily be accessed, combined, or used for new (digital) initiatives. This chapter, supported by several real-world examples, describes how data virtualization provides access to complete, company-wide information across multiple data silos in an economically and technologically feasible way. Data virtualization helps to deliver quality data on time, while also offering enterprise features like security and auditing. Moreover, data virtualization enables companies to combine the best aspects of different technologies to build a solid, flexible, and maintainable data architecture for today and the future.

C. Kurze (✉) · M. Schopp · P. Moxon
Denodo Technologies
Munich, Germany
e-mail: ckurze@denodo.com

M. Schopp
e-mail: mschopp@denodo.com

P. Moxon
e-mail: pmoxon@denodo.com

C. Linnhoff-Popien et al. (eds.), *Digital Marketplaces Unleashed*,
https://doi.org/10.1007/978-3-662-49275-8_62

62.1 The Need to Go Digital

In recent years, power has shifted from the companies to the customer. With a few clicks, customers can check competing offers and easily switch from one vendor to another. The customer, in short, is king, and expects to be treated like one. Today, to attract and retain customers, businesses need to transform from being product-centric to being customer-centric, and strive for business continuity in all processes. Such a transformation is not easy, despite the fact that the world's information, much of which is made up of customer and machine data, is doubling every 1.5 years, and is expected to double every day by 2050. Buckminster Fuller has first stated this exponential growth in his book "Critical Path" back in 1982 [1] and proves to be right nowadays with the advent of Big Data and the Internet of Things (IoT).

Such a transformation requires businesses to be able to quickly access massive databases of customer data and immediately respond to customers' needs, even when their needs change in a few minutes, such as after a purchase or a similar compelling event. Such a transformation also requires an optimized supply chain and a completely reimagined value chain. It also requires an open dialog between management, business departments, and IT about streamlined integration across business functions, with an increased focus on project outcomes, better decision-making, efficient operations, and an orientation towards developing new products and services that meet changing customer needs.

The biggest impediment to digital transformation continues to be massive quantities of heterogeneous data, since it is located in separate silos, requiring costly, time-consuming integration, or multiple, manual steps to process a simple query across the entire data set. Nearly all initiatives driven by digitization require rapid data integration in one form or another, including the deployment of new APIs, big data and/or cloud sources, mobile solutions, SaaS offerings, and interfaces with the Internet of Things.

The model of the monolithic enterprise data warehouse (or data mega store) fades away, and data heterogeneity has come to be accepted as the new normal. Hadoop, Cloud, NoSQL and other new sources appeared rapidly during the last years. This new world of distributed and diverse data needed by many apps and users is real, and it will not go away. Such a world demands that businesses develop a fast data strategy; otherwise, businesses will simply not be able to leverage the wealth of data that is already in their hands. As Forrester [2] says, "Business stakeholders at the executive and line-of-business level need data faster to keep up with customers, competitors, and partners." Well-known numbers underpin the need for timely data availability: Fixing a product after delivery costs 10 to 30 times more than during its construction or production process. Retaining an annoyed customer may cost $100 in discounts, agent calls, and process costs, whereas fixing the problem earlier could cost as little as $5.

Data virtualization enables the use of agile, real-time, self-service data technologies that deliver data to business users in real or near-real time, to effect faster outcomes.

62.2 Modern Strategies and Architectures

Many companies begin their transformation by building data labs and analytics, and establishing data science teams within the business departments. But the biggest challenge is how to provide these teams with all the data while still maintaining compliance.

Data silos, as we mentioned above, hinder the flexible access and shared usage of data across the organization. Ideally, all analytics, reports, processes, and applications (web, mobile, desktop) should see exactly the same customer, product, and partner data.

To meet this ideal, a variety of vendors proposes technology solutions and architectures. Gartner categorizes them into three integration and semantic layer alternatives [3]:

1. *Applications and business intelligence tools as the data integration or semantic layer*: This approach delegates data integration to end-user tools and applications, but results in a duplication of effort, since it is necessary to perform integration multiple times in different tools, so changes in the back-end would require reengineering on the front-end. In addition, the primary focus of end-user tools is not to function as integration middleware, but as user-friendly applications.
2. *Enterprise data warehouses as the data integration or semantic layer*: In this scenario, infrastructure vendors provide access to data not stored in the data warehouse in a pure query federation mode, often coupled with the traditional replication of data into the data warehouse. The data warehouse, as the integration and semantic layer, remains the "center of the data universe." Although this approach appears attractive to organizations that are already heavily invested in enterprise data warehouses, it does not address the big picture. What happens when there are more than one enterprise data warehouses (often based on different technologies)? Not all data can be replicated into, or accessed via, the data warehouse, and project and storage costs would increase.
3. *Data virtualization as the data integration or semantic layer*: Moving data integration and the semantic layer into an independent data virtualization platform leverages the native capabilities of such an abstraction layer to access data across multiple heterogeneous data sources. Recommended by Gartner, this approach provides business-oriented models for the underlying data via a logical-to-physical mapping. Moreover, the virtual layer enforces common, consistent security and governance policies. Advanced data virtualization solutions provide – in addition to the commonly accepted relational access to data – native support for complex data structures commonly found in recent big data technologies, which are effective for providing timely data to any data consumers.

We argue that an optimal fast data strategy is built around data virtualization, since it establishes an intelligent abstraction layer above the heterogeneous sources, which acts as a unified data access layer across the entire enterprise. Data virtualization makes it possible to combine any kind of data, such as big data and streams from the IoT, with existing data assets like customers and products.

62.2.1 How Can Data Virtualization Transform a Business?

Let us take a closer look at the transformation to customer-centricity, effected by leveraging data virtualization.

To support cross- and up-sell initiatives, sales teams need complete, updated information about the customer as well as related information about products, channels, and warranties. Marketing, support, and executive teams all need access to the same information for their different purposes.

The typical IT architecture presents a challenge, since it is often the result of more than 20 years of development and, as we mentioned above, is often characterized by siloed data stored across many disparate systems. In these architectures, each department accesses different systems in different, manual ways and IT responds with multiple point-to-point data integrations and even more data silos for each application. As a result, business users do not get answers in time to complete their business. As Forrester [2] says, "data bottlenecks create business bottlenecks."

In the past 10 to 15 years, data warehouses have provided the solution to this problem. But recently, NoSQL, big data, and cloud technologies have challenged the data warehouse approach. In contrast with these technologies, data warehousing is too expensive, or it simply takes too much time to replicate data within the enterprise. Also, legal restrictions forbid businesses to physically store data in certain cases, though they are allowed to use the data in different combinations. However, these new technologies do not immediately solve the heterogeneity problem; they may be able to store data in any format, but a user can only access or query across data in a format that their individual application can accept.

Data virtualization eases these challenges by providing a data abstraction layer which has access to all data sources – either internal or external data, on-premise or in the cloud, structured, semi- or unstructured. As shown in Fig. 62.1, the data abstraction layer acts as a single virtual repository that integrates any data in real time or near real time from disparate data sources, whether internal or external, into coherent data services that support business transactions, analytics, predictive analytics, and other workloads and patterns [2].

Data is published to various data consumers in multiple protocols, made available for querying, searching, and browsing in request/reply or event-driven mode. More importantly, a robust data abstraction layer provides an enterprise-ready security and governance framework, which enables the secure delivery of data [4]. Data virtualization provides governed self-service for all human and machine users of data, both inside and outside of the company. With such a broad access layer, special focus has to be put on performance and scalability – horizontally and vertically – in order to ensure Service Level Agreements for data. Current software solutions offer robust answers to the questions. Furthermore, with the help of the flexible, advanced role-based access and authorization mechanisms, data isolation for privacy and legal reasons is possible and can be audited with built-in mechanisms. Common data exchange based on files or shadow IT solutions by-pass such capabilities completely.

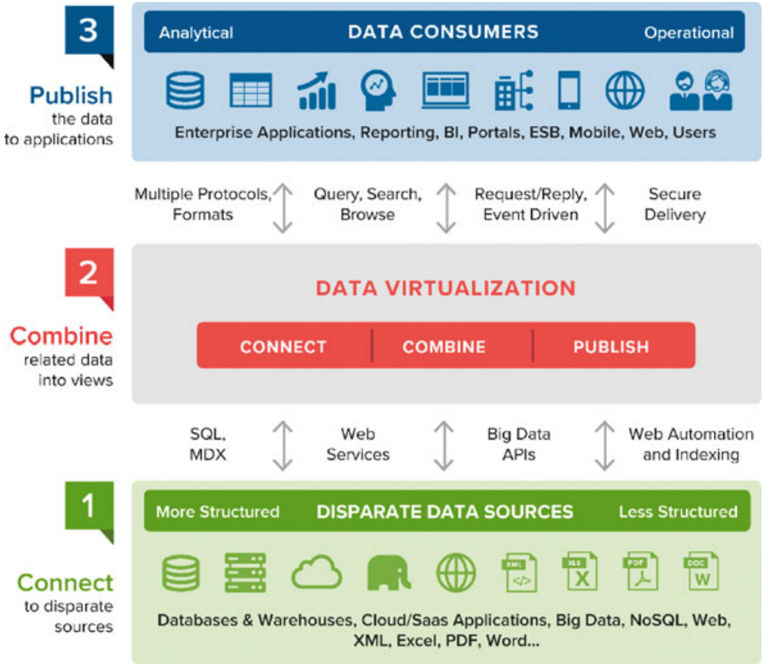

Fig. 62.1 Data Virtualization Architecture

The strategic value of data virtualization, as the standardized front door to company-wide data, is fourfold:

1. Unifying a diverse universe of data assets and helping to enforce data policies, focusing on data governance, discovery, unified modeling, security, and data auditing.
2. Maintaining efficiency in data operations with an eye towards reducing costs and shielding users from complexity, minimizing data replication, and encouraging data reusability and collaboration.
3. Enabling business agility with initiatives like enterprise data marketplaces, which focus on agile lines of business and enabling these lines to quickly launch new products, get closer to the customer, and offer data visibility and rapid data provisioning.
4. Innovating through big data and by adding new sources for enterprise use.

62.2.2 How Do Companies Transform?

Transforming a company cannot happen overnight. In addition to legacy systems and siloed data and applications, companies also often have to struggle with political and legal barriers. Usually, there is a management decision about the expected business outcomes, which should be followed by a digital initiative.

It is critical for the data virtualization architecture to enable full transparency between business and IT to encourage fluent communication between both teams. In addition, it must provide the flexibility to enable business departments to take full responsibility for new digital outcomes while enabling IT to keep control over data asset management, which includes security, governance, self-service, cloud-first strategies, API-based integration, and the delivery of new (digital) services.

In undergoing a transformation, many enterprises experience a conflict between stability and efficiency (traditional IT) on the one hand, and experimental, agile IT on the other, which is focused on time-to-market and app evolution, and is therefore more aligned with the business. Some describe this as a conflict between systems of record (like Enterprise Resource Planning, Customer Relationship Management, etc.) and systems of engagement (like data warehouses, data lakes, etc.). It is important to note that apps, reports, and humans need data from both modes of operation (traditional and experimental, or systems-of-record and systems-of-engagement) in order to focus on the customer and value creation networks. Because data virtualization can provide access to any source, it enables companies to address both modes simultaneously.

A transformation of this nature could be executed in the following way: First, complete a matrix of customer interaction channels (such as social media, website, point of sales) and critical decision points in the buying process (early stage to up-sell and cross-sell), and determine the most important customer journeys through this matrix. For example, a customer might see a viral video on YouTube about a particular product or offering. Later on, he visits the website and performs more research on the product. If he then visits a point-of-sale, he might make a purchase or his decision to make a purpose can be – at least – influenced. Alternatively, a customer might go to a garage for service and end up purchasing a new car. By recognizing and mapping out the customer interaction channels and critical decision points in the buying process, the next-best-action is clear. For example, when providing a courtesy car to a customer whose car is being serviced, should the dealer provide a larger model of the same brand (e. g. an upgraded SUV) because the customer has reached the age when she might be thinking about starting a family and, hence, might be interested in a larger vehicle? In doing so, the dealer can instill the idea of buying a new vehicle that matches the customer's life circumstances, but also (and most importantly, from the dealer's perspective) buying within the same brand.

Second, the enterprise architecture team needs to design the architecture that would best support these (and future) customer journeys. The design should reflect a combination of application and data assessment that would result in capabilities supported by specific systems and data stores. The team develops clear guidelines concerning the best use of tools and vendors, and these guidelines act as blueprints for individual projects. The architecture is constantly evaluated based on its ability to deliver the defined customer journeys. As soon as the use cases are fulfilled, the architecture matures.

In the third stage, build services: functional services (e. g. open purchase order), data services (e. g. master data), and combined services (which execute functions and read or modify data). As always, the service creation follows the modeled customer journeys.

62.3 Use Cases: Data Virtualization Driving Digital Innovation

In this section, we describe five common usage patterns of data virtualization, illustrated by real-world examples.

62.3.1 Governed Self-Service: Business Intelligence & Analytics

Agile business intelligence embraces approaches like logical data warehouses, virtual data marts, (governed) self-service, and operational business intelligence and analytics. All of these approaches rely on one or more data warehouses that are unified by data virtualization, which provides a logical data access layer to the disparate sources. Compared to a physical data warehouse, and the cost of maintaining multiple ETL (extract, transform and load) processes, such a solution can cost as much as 80% less.

One of the largest CAD software vendors leveraged data virtualization to switch from using a perpetual license model to using a subscription-based model, without disrupting BI and other business users. From a technical perspective, they built up an 800+ terabyte cloud-based data lake and combined this data with on-premise customer and financial data. Data consumers are not only reports and dashboards, but also operational and cloud-based applications. The company leveraged data virtualization because of four main benefits:

- **Availability:** Channeling end-user access through a single governance point simplifies administration.
- **Usability:** The logical data warehouse provides a single (virtual) repository, simplifying end-user access and enhancing BI.
- **Integrity:** Only the published views in the logical data warehouse are publically available. Along with data ownership, this guarantees the quality and proper licensing of the entire data set.
- **Security:** The logical access layer provides a single point for authentication, authorization, audit trail creation, and monitoring, for all enterprise-class operations.

With the new architecture in place, projects that might have required five weeks of skilled programmer time (focusing on ETL and web services development) and four weeks of testing, can now be completed in two weeks, using just one data virtualization developer and two testers. The company could also forego the need for additional hardware and software, as well as the need to maintain a heavy maintenance schedule over multiple years.

62.3.2 Big Data and Cloud Integration

Integrating big data implementations with cloud sources in real time comes into play for a wide variety of modernization initiatives, including advanced analytics, data warehouse offloading, the liberalization of big data and the cloud, SaaS integration, and hybrid analytics. The benefits of such initiatives derive from combining big data with enterprise data in real time, providing insights for informed business decisions. For example, wearables, smart home appliances, and industrial sensors often leverage up-to-date Hadoop technologies. Combining this real-time streaming data with existing enterprise data – some companies call it “small” data – for the larger context, provides real insight and value for digital businesses. The key is to make multiple data sets appear as a single data set, without replicating all data into a single repository. Even on top of a single data lake, the standardized and business-user friendly access layer has proven to be valuable.

Let us consider the example of a heavy equipment manufacturer. The company’s sensor data is captured and stored in a Hadoop cluster, but this data alone does not provide value to the company. It is the combination of this data with the parts inventory, the historical maintenance data, and the internal and external dealer data that enables the company to effectively train its predictive models, which predict potential failures. These models provide value to the company in two ways: end-users gain productivity by reducing unplanned downtime, and “pays the company back” with increased loyalty. The company increases revenue from the improved sale of service and parts, while at the same time reducing the costs of parts failure. In addition, by integrating the full supply chain of spare parts, the company benefits from the “network effect”: On a regular basis, the right parts are at the customer’s site when needed, along with a service technician with the skills to install them.

62.3.3 Broad Data Usage: Data as a Service

Data virtualization provides a way to provision data beyond the traditional methods based on SQL (Structured Query Language). Data virtualization establishes a data API and therefore serves as a single layer for all data services, published in multiple formats, such as RESTful or SOAP web services. This capability accelerates the agile development of applications by providing a unified data services layer, logical data abstraction, and linked data services. Developers no longer need to hunt down data, which can save thousands of developer hours.

Drillinginfo is a company that not only provides industry-leading oil and gas intelligence, tools and services, but also provides the most widely adopted software platform in the oil and gas industry. With the help of data virtualization, Drillinginfo reduced its new-product launch time from two weeks to less than a day. In addition, the company’s data API is made public to consumers so as to automate real-time data provisioning. Some of the dashboards are made public on the company’s website: https://diindex.drillinginfo.com/.

In the reinsurance industry, large players leverage data virtualization for 360° views of customers, contracts, deals, and risks, and these views are accessible company-wide, via RESTful web services that follow the OData standard. This data virtualization layer enables end-users to navigate through the data without a deep knowledge about the underlying schemas and the ways in which heterogeneous data sets are connected. Portals, and applications that serve internal data consumers and self-service customer portals, also access the standardized layer.

62.3.4 Operational Excellence: Single-View Applications

Data virtualization enables applications to provide a single, authoritative view across myriad disparate data sets. A single view of the customer enables call centers and portals to improve responsiveness and accelerate upselling opportunities; a single view of the product yields streamlined catalog services; a single view of the inventory speeds reconciliation efforts; and vertical-specific views enable self-service search, discovery, and exploration functionality. Combined with linked data services, navigation through business-oriented entities is a core capability that provides considerable power to business departments.

Jazztel leveraged data virtualization to enable an application to provide unified views of the customer across more than 30 data sources, including systems for provisioning, invoicing, CRM, incidents, and ERP. These views are consumed by the contact and call center, as well as the client extranet. Internal reporting draws on the same virtual entities. Client call times were reduced by 10% while solving 90% of the problems during the first call; customer retention has doubled, and the back office workload has been reduced by more than 50%.

62.3.5 Modernization, Mergers and Acquisitions, Divestments

Many businesses struggle to provide value-added services to their customers because of legacy systems that are hard to integrate. Data virtualization not only offers abstraction capabilities that ease this burden by integrating the data without replicating it, but also ameliorates mergers and acquisitions as well as divestments. It provides consumers with access to the data, regardless of the disposition of the relevant sources.

The story of AAA demonstrates the importance of decoupling data consumers from data sources in large corporate transformations. Regulatory forces mandated that AAA separated its non-profit automobile club from the profitable insurance business. Unfortunately, the organization operated highly interconnected systems and a single data center. To ease the burden of the physical migration, the whole application landscape was decoupled, via data virtualization, in horizontal and vertical layers. This step not only led to faster compliance with the regulations, but it also opened up time for the physical system migration. Also, during the migration phase, the new data center was able to communicate with the old data center in a controlled way through the virtual layer; even complex applications were changed step-by-step without interfering with the physical system migration. AAA called the initiative “changing the wheels at 70 miles per hour”.

In modernization initiatives and other corporate transformations, data virtualization can minimize the number of point-to-point connections, ease access to the data, provide view spanning across multiple systems and therefore reduces the necessary efforts and increases the potential for IT to create new strategic value.

Data virtualization can also ease the migration of whole architectures, or parts of it, into the cloud. Think of IoT cloud offerings; they are easy to set up, but they still need to integrate into the company's data backbone.

62.4 Summary

Digital transformation is becoming the new status quo. Uber and Airbnb, now household names, are also familiar examples of traditional businesses being disrupted by software based businesses. But also consider the fintech and insurtech industries. All kinds of new technologies are disrupting traditional on-premise models in automotive, retail, and other industries. 3D printing disrupts typical production and retail processes; the IoT and wearables are about to disrupt even more sectors, e. g. manufacturing, pharma, and life-sciences.

To stay ahead of these developments and transform into real, data-driven enterprises, IT and business teams need to work closely together, with IT holding responsibility for information management and provisioning, and business teams being responsible for analytics and acting on outcomes. Fact-based decision making needs to be incorporated into all processes, which requires the appropriate technologies. The key asset is data, supported and protected by effective information management.

The IT infrastructures of most companies have been in development for more than 20 years and are challenged by all the new technologies that have emerged in the last five years. Many data silos still exist, and to leverage data for digital business outcomes, companies need fast data strategies for delivering data at the speed of business. This chapter outlined an approach that uses data virtualization as a data integration layer, providing data consumers with instantaneous, unified views of the data across myriad, disparate sources. The data virtualization layer enables governed self-service across the whole enterprise, with access to the data for all groups in the enterprise, fully aligned with security and access policies.

The five use cases that we presented illustrate shifts from traditional to digital business models. As seen in these examples, data virtualization creates a path for process optimization, big data integration, and cloud analytics. Data virtualization also paves the way for enhancements to the data warehouse, business intelligence modernization, and the overall transformation of IT architectures, all while maintaining regulatory compliance.

With this kind of power, companies are able to respond immediately to customers' changing needs and have the capacity to disrupt traditional limitations across the entire business lifecycle. Manufacturers, for example, might be able to design a sports shoe, a fashion accessory, or a car body, using just a laptop, and bring the products to market in just a few days. With the power of data virtualization, companies are limited only by their imaginations.

References

1. R. Fuller Buckminster, Critical Path, 2nd Edition ed., New York: Griffin, 1982.
2. Forrester Research, "Create A Road Map For A Real-Time, Agile, Self-Service Data Platform," 2015.
3. Gartner Research, "The Big Data Warehouse Deal: The Future of Data Management Solution for Analytics," 2016.
4. R. van der Lans, Data Virtualization for Business Intelligence Systems, Watham: Morgan Kaufmann, 2012.