R20
CONSULTANCY

The Business Benefits of Data Virtualization

A Whitepaper

Rick F. van der Lans
Independent Business Intelligence Analyst
R20/Consultancy

May 2019

Sponsored by

denodo

Table of Contents

1 Executive Summary 1

2 Data: The Hidden Business Asset 2

3 Modern Business Requirements for Data Delivery 3

4 Two Examples of Businesses Struggling with Data Requirements 6

5 Traditional Solutions to Unlock Data 7

6 Data Virtualization in a Nutshell 9

7 The Business Benefits of Data Virtualization 10

8 Closing Remarks 12

About the Author 13

About Denodo 13

1 Executive Summary

Data is deeply buried and almost unattainable for business users.

Using Data More Effectively – To become *data-driven* and to fulfill the *digital transformation* dream, data must be moved center stage and into the spotlight. Unfortunately, in many organizations most of the data is deeply buried and almost unattainable for business users who want to use data more effectively and more efficiently to improve and optimize business and decision-making processes. To become data-driven, organizations need to unlock and free all that data in an easy and flexible way.

Data is Deeply Buried – Most enterprise data is deeply buried and hidden inside applications and databases. Because data has been stored in a multitude of systems using different technologies and can be complex to access, it can be costly and time-consuming to locate the required data, extract it from the source systems, integrate it, and turn it into meaningful and actionable information. So, although the data exists and can be turned into a valuable business asset, it may be out of reach for business users. If business users don't or can't have access to the data to improve their business processes or decision-making processes, the data is of diminished value.

The requirements for data delivery have changed.

Modern Business Requirements for Data Delivery – The requirements for data delivery have severely changed over time. What was required twenty years ago is very different from what is required today. For example, some business users need access to real-time data. Such a requirement was unthinkable twenty years ago as it was impossible to implement. The technology wasn't ready for such requirements. Also, business users now require the immediate fulfillment of new, ad-hoc, and unplanned data requests. Again, this need did not exist some time ago. Business users could accept a delay of a few weeks.

Today, to improve and optimize business and decision-making processes, data must be:

- **Available –** Data must be available when business users need it, regardless of whether it's new or old, simple or complex, or detailed or aggregated. It should not be hidden from business users, nor should its existence be unknown.
- **Integrated –** Data from multiple systems must be combined and presented as if it comes from one system.
- **Consistent –** Data must be consistent across different reports or dashboards.
- **Correct –** Data must be in line with reality; incorrect data can lead to incorrect business decisions.
- **Timely –** Data must be available to business users at the right time, even if this means seconds after the data was produced. Data received too late, may have no value.
- **Instant –** Data must be available the instant business users ask for it, even when the data request is unique and completely ad-hoc. It should not take days before it becomes available.
- **Documented –** Data must be documented and explained using descriptive metadata; without metadata it may have no business value.
- **Trusted –** Business users must trust the data with which they make their decisions, and they need the data processing to be as transparent as possible.
- **Actionable –** Data must be shaped, filtered, and aggregated so that only relevant data is presented and is aimed at taking business actions.
- **Adaptable –** Data must be able to adapt when the business changes and when data definitions and data processing regulations change.

Popular Solutions to Unlock the Data Silos – Several different data delivery solutions are now being used to unlock data silos for business users. The most common solution is the *data warehouse*. Unfortunately, this solution was originally not designed for quick development, but to improve and guarantee the quality of the data and therefore the reports. Supporting the requirements for timely, instant, available, and adaptable data is challenging for a data warehouse. This is due to all the data copying processes taking place in a data warehouse environment.

Another solution, called *self-service BI,* enables business users to develop their own reports with dedicated tools that can access the data from almost any system. With a self-service BI solution, business users develop their reports independently from one another; or in isolation. There is practically no sharing or reusing of specifications amongst these users. The consequence is that it is hard to fulfill the requirements concerning integrated data, consistent data, and correct data.

More recently, a third solution has become popular: the *data lake*. Initially the data lake was developed to support a small group of users, the data scientists. Nowadays, the data lake is regarded more as a multi-purpose data repository that contains a large portion of the enterprise data. Data lakes are not like data warehouses in which data is cleansed, integrated, standardized, and documented. Data lakes commonly contain raw data from many different sources. It's a more experimental environment than the data warehouse. This can make it a challenge to support the requirements concerning integrated data, consistent data, trusted data, and documented data.

The Data Virtualization Solution to Unlock the Data from its Silos – Each of these three solutions has some real drawbacks with respect to data delivery. The latest solution that tries to support all the requirements for data delivery is *data virtualization*. This whitepaper describes how data virtualization can help to free and unlock the deeply buried data and to make it available to most business users. It explains how data virtualization supports the modern requirements for data integration and delivery. Most of them are described using real use cases.

2 Data: The Hidden Business Asset

Top executives want to do more with data.

Data-Driven Organizations – Popular topics in today's boardrooms are *digital transformation*, the *data economy*, and becoming more *data driven*. Regardless of how they interpret these terms, it always implies that the top executives want to do more with data. They want to use data more effectively and more efficiently to improve and optimize business and decision-making processes. Maybe data has not always been regarded as a *valuable business asset*, but due to new insights, organizations now start to see the business value of data (or at least they must).

Data is Deeply Buried – Unfortunately, most data is deeply buried. Data is hidden inside applications, such as CRM, ERP, and human resource systems; it's inside databases, data warehouses, excel spreadsheets, and files; and, it's stored on-premises and in private and public clouds. Additionally, in large companies operating internationally, data can be stored in systems distributed across the globe.

Beside the fact that data is stored in many different systems, it's also stored with different types of technologies. Data may be stored in a SQL database, a Hadoop cluster, excel files, and so on. All these systems speak their own languages and store data in unique formats. Quite regularly, if the data has been discovered by the business users, their reporting tools can't even access the data, because support for those languages and formats doesn't exist.

Data is Unattainable for Business Users – The problem is not that there is no data. On the contrary, data is available in abundance—it's everywhere. However, because data has been stored in a complex, heterogeneous, and dispersed way, it can be costly and time-consuming for business users to locate the required data, extract it from the system, integrate it, and turn it into meaningful information. So, although the data exists and can be turned into a valuable business asset, it may be out of reach for business users. In fact, data can be so unattainable that they may not even know that specific types of data that can be extremely useful to their decision making is stored in some IT system.

The consequence of unattainable data is manifold. Certain management reports cannot be developed on time, certain business risks or opportunities are not identified (or not on time), reports showing incorrect data may lead to incorrect decisions, reports for legislators are complex to audit, there can be issues with compliance, and so on.

Summary – Even if business users are able to locate the required data, they may not be able to access it, extract it from the source system, and turn it into meaningful information. In this respect, data can be compared with oil and gas. Oil and gas can also be deeply buried in the ground and very expensive and time-consuming to locate, extract, and process.

Data is potentially a valuable business asset. Unfortunately, it remains a *potentially* valuable asset if business users don't or can't have access to the data they require to improve their business processes or decision-making processes. And some data loses value if it's not used in a timely manner. In that respect, data is like water; if you don't do anything with it, it slowly vaporizes. The value of data also slowly vaporizes over time.

Data remains a potentially valuable asset if business users don't or can't have access to it.

3 Modern Business Requirements for Data Delivery

In a perfect world, the delivery of data to business users will meet each of the ten requirements; mentioned above; see Figure 1. Below, each of these requirements is explained in greater detail.

Available Data – All the data gathered over the years by an organization must be available for reporting and analytics when business users ask for it, and it must be made available in the format they require. Restrictions defined by regulations and laws may apply with respect to data usage. Within the context of such restrictions, data should not be hidden nor should its existence be unknown to business users. Unnecessarily withholding data from the business can be detrimental to its health.

Integrated Data – There was a time when all the data produced by an organization was stored centrally in one location in the data center of the organization's basement. This was a long time ago. Nowadays, data is stored in numerous systems scattered across the planet. For example, it's stored in cloud applications, in a multitude of remote databases, in local spreadsheets, and in distributed data clusters. Moreover, many different technologies are used to store the data, ranging from SQL databases via Hadoop clusters to plain flat files. Although data has been dispersed across all those systems for many practical and technical reasons, this hasn't made life easier for reporting data to business users.

More business insights can be found when data from multiple systems is integrated.

Business users require data from multiple systems to be integrated before they can work with it or present it. This is because more business insights can be found when data from multiple systems is combined. For example, it's useful to know how many products customers buy, but it's even more useful if this information is combined with how many products those customers return. Ideally, business users can integrate data without restrictions.

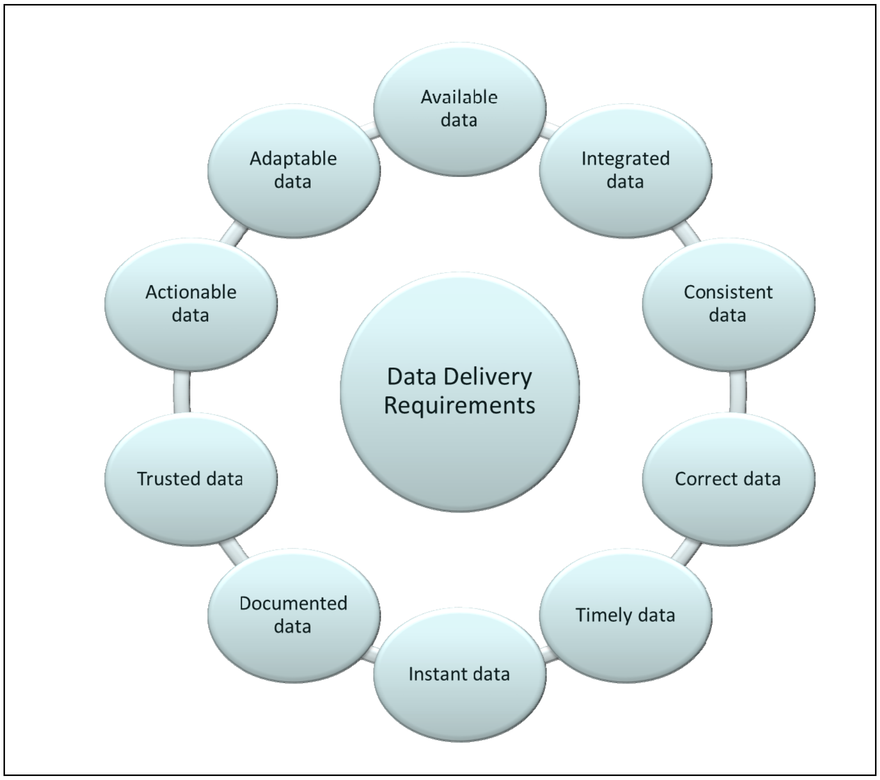

Figure 1 *The ten requirements for modern data delivery.*

Consistent Data – The consistency of data is important to the business. If two reports show different results, which one can be relied upon for decision making? Commonly, data must be transformed, filtered, aggregated, and integrated before it is presented to business users. When developing a report, such *transformation specifications* must be defined. If all reports used the same transformation specifications, data consistency would be easy to guarantee.

In many organizations different reporting and analytical tools are used for presenting data. These tools support their own languages for implementing transformation specifications. As a result, the reports are developed as isolated solutions. Developing these specifications is like inventing the wheel over and over again. Unfortunately, it is practically impossible to guarantee that two specific transformation specifications are implemented in similar ways in two different reports, so the reports may present inconsistent results.

Transparency leads to trust of the data.

Correct Data – Not all the data in IT systems is perfect. For example, names may contain typos, data elements may not match, and incorrect values may have been entered. Data is not perfect for numerous reasons. It's evident that the quality of data has an impact on the quality of decisions. Since incorrect data can lead to incorrect decisions, it's important that data is regularly evaluated for its correctness.

To deliver the highest possible quality, most of the data that comes from internal and external sources must be corrected before it can be used by business users. In traditional data warehouse architectures immense development time is spent on getting the right controls and checks in place. Many of the transformation specifications are implemented to detect and correct incorrect data.

Data received too late, may have no business value.

Timely Data – Data must be available to business users when they need it. As mentioned above, data received too late, may have lost its value. Some business users can work with data that was produced weeks before, especially if they look at historic data and do trend analysis. In this case, having access to data that is 100% up-to-date is not a necessity. But more business users need access to real-time data. This is especially true of operations management, the work force, suppliers, and customers. So, a data delivery solution should deliver timely data or data that meets the business users' requirements for timeliness.

Instant Data – Many reports are used by business users day in and day out. Other predefined reports are mailed to external legislators regularly. Both of these types of reports meet business needs that were known long beforehand. So IT can spend time on developing them.

But what if a sudden, unexpected business event occurs that requires a new report built with data that has never been requested before? It's important that IT can fulfill such ad-hoc requests for data instantly, because the urgency of the event demands it. In other words, data must be available instantly when the business needs it, even when the data request is unique and completely ad-hoc. The speed of data delivery is crucial for decision making.

Documented Data – Documentation must be available to explain what the data means and what its precise definition is. Without that, data has no business value. When business users make decisions based on data they fully need to understand and trust that data.

Without descriptions data has no business value.

When data is delivered to business users their legitimate questions are: What does this data mean? How up-to-date is the data? What is the source, internal or external? For example, if data is labeled sales, does that mean net sales or gross sales? Has the number of units returned already been subtracted from units sold or not? To answer these questions, they need access to *data definitions. Metadata* unambiguously describes and defines what the data means. Business users must know what every data element means. They need definitions and explanations of the data, information on the source of the data, and how the data was processed. Without proper metadata, data can be close to meaningless, and it is of questionable value as a business asset.

Trusted Tata – This requirement is related to the previous one, since documentation increases the trust in the correctness of the data. However, how correct is the documentation? Therefore, supplying users with information on how the data was processed can also help to increase trust. Business users must have access to specifications that show where the data comes from, how it is processed, and what kind of logic is applied to it. In other words, data processing must be as transparent as possible, because transparency leads to trust.

Actionable Data – It is easy for IT to overwhelm business users with piles of data and reports. The amount of data available for reporting is almost endless. Additionally, the same data can be presented in many different ways. The challenge is to present the right data in the right format when it is needed. This is a major design challenge. When done correctly, the data presented can lead to taking the right business actions. In other words, data must be shaped, filtered, and aggregated in such a way that it becomes actionable data. For example, showing the monthly sales data for the last twelve months may be less useful than showing in which month sales is unexpectedly severely lower than in the same month the previous year. The second option may lead to a manager investigating what has happened. The real challenge is to present the data at the right time, in the right form, and with the right level of detail that makes it actionable for business users.

It must be possible to change data delivery at the speed of the business.

Adaptable Data – Businesses change constantly. They change with respect to the products and services they offer, they change because of internal reorganizations, they have to change due to new regulations, they change to address new customer groups, and so on. All such changes can lead to new data requirements. They can lead to changes of transformation specifications, data element definitions, or reports. Adaptable data is data that can easily be adjusted when changes occur. Only then can data delivery adapt to the changing world of the business. It must be possible to change data delivery at the speed of business. Presenting data that relates to the situation before the change has no business value.

4 Two Examples of Businesses Struggling with Data Requirements

In this section two real examples are described in which business problems occurred because the IT system did not meet some of the data delivery requirements.

Utility Company – The first example is from a utility company that is just one of several comparable companies that are all part of a national organization. They all try to innovate their businesses by adhering to several operational and financial *key performance indicators* (KPIs). One of their most important KPIs is called Not-Charged-Water (NCW). This KPI shows the amount of clean drinking water pumped into the water system divided by the amount of clean drinking water charged to consumers. The lower this KPI, the better it is. It would be ideal if this KPI were equal to one.

After several years, the company discovered that there was a problem with the value used for how much water was charged. The charged amount is the sum of all the invoices paid plus the outstanding invoices. The problem was that some customers with outstanding invoices blocked their payments, because they didn't agree with their invoices. For some unknown reason, these blocked invoices were not included in the calculations. Although this was only a small percentage of the total number of invoices, it still influenced the KPI in a negative way. Business users were not aware of this and would interpret this as non-invoiced water.

They also discovered that due to technical IT problems a group of reminders were "parked" somewhere. These invoices were also not included in the calculations of the KPI.

Business users did have access to the data (in the form of KPIs), but the data was not correct, or the presented data (the KPI) was not correctly interpreted by business users. They probably didn't even know that the concepts of parked data and blocked reminders existed. This is a typical example of data delivery with problems concerning correct data, documented data, and adaptable data.

Media Archival Company – The second example relates to a company that manages the audiovisual heritage of a European country. The collected data amounts to the biggest digital media archive of Europe. Customers are TV and radio channels, private customers, media professionals, teachers and students, and scientists. This company lacked one centralized system to store customer data, so the data was scattered across many systems. Especially from an analytical standpoint, to discover how and why the media data was used and by whom, it must come together. The current situation with respect to customer data is as follows:

- Many different reports exist but they are all aimed at showing data related to one business process.
- Reports that combine data from different IT systems are scarce.
- Many IT systems that contain customer data are very hard to access or there is uncertainty about the data quality.
- Most reports don't show customer data at all.
- Customer data is stored across many different systems.
- The expertise level of business intelligence is very low within the organization
- The company does have a focus on data governance, but purely governance of data stored in the archive, and not customer data.

The required data for most customer-related reports was available, but not attainable. It was replicated, it was dispersed across many systems, the data quality was unclear, and no real metadata was available. Therefore, the media company was very restricted in analyzing the data. Discovering how and why the media data was used and by whom was close to impossible. Customer data was clearly not regarded by this company as a valuable business asset, which it could have been. At a minimum, this company has problems with the requirements concerning integrated data, consistent data, documented data, and available data.

Summary – Both examples are representative of many organizations across the world, which are struggling with exploiting their data investment and meeting their data delivery requirements.

5 Traditional Solutions to Unlock Data

The data delivery requirements described in this whitepaper are not new. Organizations have been struggling with them for years. This section describes three of the more popular solutions designed to solve the problems of unlocking data.

Solution 1: The Data Warehouse Architecture – The most common solution developed by IT to make data available for business usage, is to develop a data warehouse architecture. This architecture consists of a chain of databases, such as a staging area, a central data warehouse, and multiple data marts; see Figure 2. *Extract-Transform-Load* (ETL) programs are used to copy data from one database to another. For example, data is first copied to a staging area, next to an enterprise data warehouse, and finally to data marts. With each step, data is gradually turned into the format and shape that business users require.

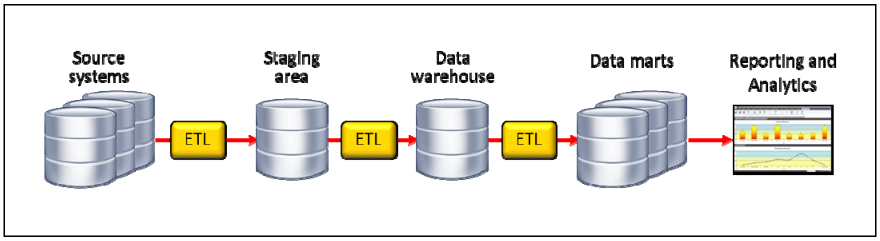

Figure 2 *The classic data warehouse architecture consists of a chain of databases in which ETL programs copy data from one database to another.*

This solution has proven its worth. The advantages are related to the fact that it's a managed environment. During the copying process, data is verified and corrected if needed. And when users access the data marts, they have access to the metadata describing exactly what all the data means. Most of the reports that are produced by these architectures are formally tested, governed, and audited.

Timely data: Delivering real time data to business users timely is very hard and sometimes just plain impossible with this architecture, because the data can't be "pumped" through the chain of databases quickly enough.

Adaptable data: Unfortunately, the data warehouse architecture is a very rigid architecture. For example, implementing changes can be a complex exercise. Also, if business users find a new data source that they want to use, it can take a month or longer before this data becomes available via this chain of databases. This architecture is not designed for quick development, but to improve and guarantee the quality of the data and therefore the reports. Making changes to the existing environment is not always easy, because many specifications are shared by multiple users. Making a change can have unwanted impacts on other reports. Therefore, supporting adaptable data is not easy.

Available data: It is very uncommon that all enterprise data can be made available for reporting through a data warehouse architecture. And if more data is needed, it may take quite some time before all the transformation specifications are developed that pump the data through the chain.

Instant data: The data warehouse architecture is commonly too rigid for instant, ad-hoc data requests. And if it is at all possible, only ad-hoc requests on data that is already available within the data marts and data warehouse can be executed.

Solution 2: Self-Service BI – The second solution is to allow users to develop their own reports and to give them tools to access the data in the original files and systems. They should also be able to combine data from multiple data sources. S*elf-service BI* doesn't have the disadvantages of the first solution. There is no need to develop an entire architecture first. If the business users receive access to the data, they can start developing their reports.

Documented data: The challenge, however, is knowing what all the data means. Do they have access to the correct metadata describing and defining the data? Without the correct metadata, working with this data is time-consuming and risky. There is a risk that users interpret the data incorrectly, and this leads to incorrect results in the reports.

Correct data: The quality of the self-service reports and dashboards may also be doubtful. The reports are rarely governable and auditable. A study by Wayne Eckerson[1] confirmed these problems. It indicated that in over 60% of the cases, self-service BI leads to report chaos. In a way, this solution is almost the opposite of the first one. The speed with which data reports can be developed is higher, but data quality is a big concern.

Integrated data: If business users develop their own reports, they are also responsible for integrating data coming from multiple systems. Developing integration specifications can be a complex exercise. It requires a deep understanding of how data is organized and stored in source systems. For the same reason, transformation specifications must be developed for many reports.

Consistent data: To deliver consistent data, who will guarantee that every set of specifications is implemented in exactly the same way by each business user? Business users are now also partly responsible for checking that the data they access is correct before it is presented.

Solution 3: The Data Lake – More recently, a third solution has become popular: the *data lake*. Unfortunately, it's not a very well-defined concept, so IT specialists have different views of what a data lake is supposed to be. For many, it's a data storage environment specifically designed for data scientists and other investigative users, and for others it's an environment in which all enterprise data is stored. A data lake can be used by all kinds of users, from business users requiring simple dashboards to data scientists who need advanced analytics.

Integrated data: In many data lakes, the data that comes from the systems is stored in its original form and, thus, is not integrated. It is left to business users and reports to integrate the data. As indicated for the self-service BI solution, this integration can be a complex exercise. For the same reason it can be a challenge to develop reports that present consistent data.

Documented data: In many cases the data stored in data lakes is undocumented and almost no metadata is available. This lack of metadata complicates the development of data reports. The question that also remains is whether the business users implement the specifications correctly and completely.

Correct data: When source systems contain incorrect data, the data must be cleansed before it's presented to business users. Most data lakes contain data copied unmodified from the source systems. This means that the correctness of the data is to a large extent dependent on the data quality of these source systems, and thus the quality of the data will be diverse. Most importantly, the correctness of external data can be questionable and there may be no way to verify its correctness.

Summary – Each current popular solution for data delivery suffers from some real practical disadvantages. Data stewards almost always have to choose between fast report development and perfect data quality. The latest solution for data delivery that supports all the data requirements is

[1] W. Eckerson; April 2013; see http://insideanalysis.com/2013/04/the-promise-of-self-service-bi/

based on *data virtualization*. The next section contains a short introduction and explains its capabilities with respect to the data delivery requirements.

6 Data Virtualization in a Nutshell

Introduction – Data virtualization is a *data abstraction layer* that sits between data sources and reporting tools; see Figure 3. It can access almost any kind of data source and can make that data available to reporting and analytical tools. The reports extract the data from the sources via a data virtualization layer that integrates, transforms, filters, and aggregates the data into the required form. In other words, all the transformation specifications required to process the data, that are commonly spread across an entire data warehouse environment, are now defined centrally within the data virtualization layer. Metadata to describe and define the data also resides within this layer.

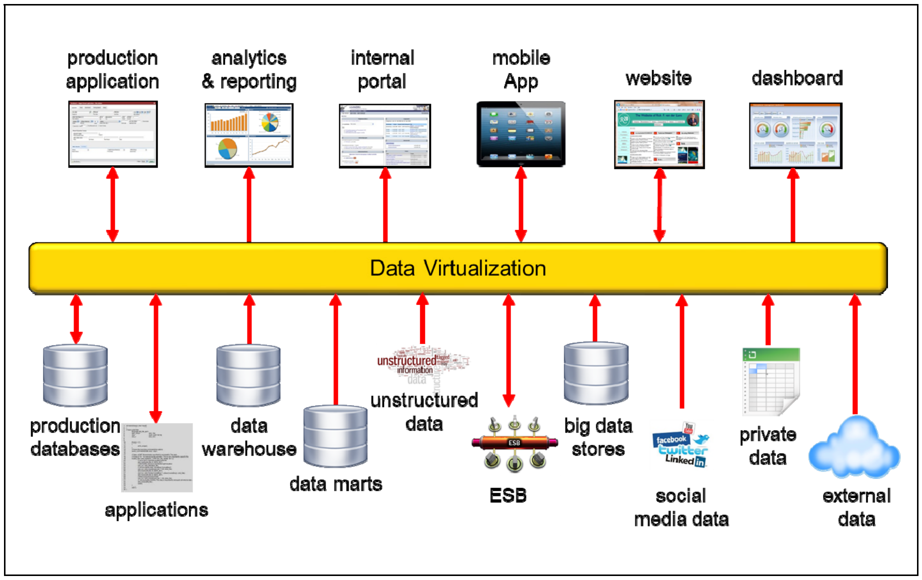

Figure 3 *Data virtualization is an abstraction layer that decouples data consumers from data stores.*

Whereas most traditional technologies, such as ETL, require data to be stored after the transformation specifications have been processed, data virtualization has been optimized to execute the transformation specifications on-demand, which offers more agility for data delivery. If needed, data virtualization supports caching, which means that data results are physically stored after the transformation specifications have been executed. The results can then be reused by multiple users and reports. In other words, storage of temporary results is optional and not mandatory allowing for more *lean integration* solutions. The term lean integration implies fast development and high adaptability, but not a lightweight solution.

Data virtualization allows for lean integration.

Supporting Data Delivery Requirements – Data virtualization supports all the modern requirements for data delivery.

Available data – Due to its lean data integration capabilities, it's easy to make data available quickly from almost any kind of source to a large group of business users using numerous reporting tools. Data virtualization unlocks data's potential very quickly.

Integrated data – Like ETL tools, data virtualization is designed to integrate data from a heterogeneous set of data sources. This integrated data can then be made available to all kinds of business users and reports. Reports share the integration specifications.

Consistent data – All types of transformation specifications can be defined in a data virtualization layer, including filters, aggregations, calculations, and projections, which can then be used by all reporting tools. Specifications are defined only once and reused as many times as possible. By reusing them, there is a high level of consistency across the reported data. Transformation specifications are centralized, rather than being scattered across different systems.

Correct data – Incorrect data can be detected and corrected if needed. This can be implemented using the transformation specifications. If required, external professional data cleansing technologies can be invoked to cleanse the data as well.

Timely data – Data virtualization delivers data in real time data, including data from operational systems, in which new data is continuously being entered. When a data virtualization server is linked to such an operational system, data becomes instantly available for reporting. As mentioned above, data must be available to business users at the right time, even if this means seconds after the data was produced. Data virtualization does not require that data from the source is first copied to another data store before it can be used for reporting, as is required by alternative solutions. In other words, a *lift and shift of data* is not needed.

Instant data – Because of the lean integration style that doesn't require data to be copied before it can be used, data can be made available for reporting in days instead of weeks. It's a perfect platform for ad-hoc data requests.

Documented data – Data virtualization enables all the available data elements to be documented, described, and defined. Tags can be defined for data elements and business users can use these tags to search for them. All of the relationships between data elements can be displayed using lineage and impact analysis diagrams.

Trusted data – Because all the transformation specifications are defined and stored centrally, it's easy to show how the data is transformed from the sources to the result. This information can be investigated by business users. This increased transparency makes the environment more trusted.

Actionable data – Whether data is actionable is not dependent on the solution. With data virtualization, data can also be shaped into actionable data for the business users. This technology supports all the required functionality.

Adaptable data – Because all the transformation specifications are stored centrally and because there is less need for physical database components, it's easier to adapt the environment to the needs of the business. If organizations using this technology are asked how they would summarize the advantages and strengths of data virtualization with one word, they will probably use terms such as agile and flexible. Both refer to adaptable data.

7 The Business Benefits of Data Virtualization

This section describes the business benefits of data virtualization through several use cases. Each use case addresses one of the data delivery requirements. These use cases are taken from different industries. The organizations requested to stay anonymous.

Consistent Data – In a national transportation and infrastructure organization, the need for ready-made reports was diminishing. Business users wanted to have the data delivered in files, which they would use to develop their own reports. Commonly, data virtualization delivers data through a SQL interface which is supported by popular reporting tools, such as Microsoft PowerBI, QlikSense, and Tableau. Additionally, data virtualization can deliver data to data consumers in many other ways. Data can be delivered as files in all kinds of formats, it can be delivered through a JSON/REST interface to support, for example, Java apps running on mobile devices, and through SQL. Making this data available through another interface is easy to implement with data virtualization. Regardless of how the data is delivered, data virtualization is responsible for cleansing, correcting, standardizing, and aggregating the data.

The organization switched to data virtualization to make it easy to deliver the data in any kind of format without having to develop separate solutions for each form of data delivery. This means that the data virtualization layer processes all forms of data delivery with the same set of transformation specifications by leading to consistent data across all the data delivery forms, whether it’s SQL, JSON, or Java.

Integrated Data – A large, well-known electronics company had decided to become more data-driven. This change was forced by the need to analyze the available data more extensively and more deeply. In addition, they wanted to build a new system in the cloud. They moved their data to several cloud-based data storage technologies, including AWS Redshift, AWS RDS, and AWS S3. Each of these technologies has its pros and cons and they support different interfaces and languages.

The data virtualization server is the heart of the new system. It stands as a doorway in front of the cloud-based systems, masking the differences across the systems. Besides making all of that data available for analytics and reporting, it makes it easy to use the data. Additionally, the data virtualization server provides data governance and data security capabilities.

The data virtualization server is responsible for integrating the data stored in all the cloud-based data sources and in the on-premises data sources. Business users see one integrated logical database. They are shielded from the distribution of data across all these systems and migration to the cloud.

Instant Data – A large data and analytics provider in the oil and gas industry brings together data from a variety of public and private sources, including highly structured data, images, and text. The data acquisition process is very complicated due to the nature of their data.

The company uses data virtualization as a single platform for all forms of data delivery. The goal was simply to serve more data customers in a shorter period of time. Before data virtualization, data delivery would easily take them one to two weeks to develop. Now, with data virtualization, this has been reduced to half-a-day. This company has experienced an impressive acceleration of data delivery to their customers. Additionally, the current solution supports more customers than before. The company is now able to deliver new data almost instantly.

Timely Data – An investment company experienced tough competition. In addition, regulations required the company to deliver data about its financial situation. Due to the rigidness of its existing data warehouse architecture, the company could not cope with these new demands for data. The required reports took too long to develop.

The decision was made to develop a new architecture based on data virtualization that would be able to deliver data on time, in the right format, and at the proper data quality level. The company refers to it as its business access layer. The key benefits are fast access to governed and secure data. Additionally, it also enables the company to blend real-time data with more stale data from the data warehouse environment. This was specifically important for the competition perspective. This new solution fully supported the company’s requirements for timely data.

Available Data – A midsize insurance company was still running its back-office systems on legacy mainframes. Because of the high costs and shrinking mainframe expertise, the company decided to modernize its back-office systems. The company chose a modern NoSQL database server to store its key transactional data, which has a highly complex structure and contains the entire history of enterprise data. The demand for more data and reporting was also increasing.

In spite of the successful migration, the company encountered a problem with the NoSQL system—the interface and language to extract data from the NoSQL system was very technical and not supported by the company's reporting and analytical tools. Therefore, the company installed a data virtualization server between the NoSQL sources and the reporting tools, to give business users easy and fast access to all of the data. This enabled the company to fully exploit modern big data technology without having to deal with its technical peculiarities.

Now all the data is available through the data virtualization for reporting.

Adaptable Data – For many years, a national bureau of statistics had operated in the same way. The organization delivered statistical results to its national government and other organizations. To do that, the bureau collected data from many sources and initiated studies to gather new data and insights.

However, the bureau realized that if it would continue working like this, commercial data marketplaces might possibly take their place, rendering them obsolete. Therefore, the bureau developed a new vision. It wanted to change its business model by becoming a large data marketplace itself. The bureau would also operate as a marketplace for other government organizations, such as additional local government organizations on the state and city level. The bureau wanted to deliver its own data plus data from other organizations, and deliver the combined data to many organizations. The bureau wanted to become the central entity for statistical data in the country.

The bureau had two important requirements: making data available fast and improving data quality. To satisfy both needs, the bureau deployed a data virtualization server, which acts as the central building block of their data delivery architecture.

8 Closing Remarks

To become data-driven and fulfill the digital transformation dream, data must be moved center stage and into the spotlight. But in many of today's organizations the data is deeply buried and almost unattainable for business users. Data virtualization can help unlock and free all that data in an easy and flexible way. It can help turn data into a valuable business asset.

Data is moving center stage to become more data-driven.

To improve and optimize business processes, decision-making processes, and the delivery of data to business users, data virtualization can help to make data available, integrated, consistent, correct, timely, instant, documented, trusted, actionable, and adaptable.

The lean and agile data integration capabilities of data virtualization help to free and unlock the data that has been hidden for so long. Data that was unattainable for most business users can now be freely accessed. It definitely assists with transforming existing data into a valuable business asset.

Data virtualization helps to free and unlock the data to become a valuable business asset.

About the Author

Rick van der Lans is a highly-respected independent analyst, consultant, author, and internationally acclaimed lecturer specializing in data warehousing, business intelligence, big data, database technology, and data virtualization. He works for R20/Consultancy (www.r20.nl), which he founded in 1987. In 2018 he was selected the sixth most influential BI analyst worldwide by onalytica.com[2].

He has presented countless seminars, webinars, and keynotes at industry-leading conferences. For many years, he has served as the chairman of the annual *European Enterprise Data and Business Intelligence Conference* in London and the annual *Data Warehousing and Business Intelligence Summit*.

Rick helps clients worldwide to design their data warehouse, big data, and business intelligence architectures and solutions and assists them with selecting the right products. He has been influential in introducing the new logical data warehouse architecture worldwide which helps organizations to develop more agile business intelligence systems. He introduced the business intelligence architecture called the *Data Delivery Platform* in 2009 in a number of articles[3] all published at B-eye-Network.com.

Over the years, Rick has written hundreds of articles and blogs for newspapers and websites and has authored many educational and popular white papers for a long list of vendors. He was the author of the first available book on SQL[4], entitled *Introduction to SQL*, which has been translated into several languages with more than 100,000 copies sold. More recently, he published his book[5] *Data Virtualization for Business Intelligence Systems*.

For more information please visit www.r20.nl, or send an email to rick@r20.nl. You can also get in touch with him via LinkedIn and Twitter (@Rick_vanderlans).

Ambassador of Kadenza: Rick works closely together with the consultants of Kadenza in many projects. Kadenza is a consultancy company specializing in business intelligence, data management, big data, data warehousing, data virtualization, and analytics. Their joint experiences and insights are shared in seminars, webinars, blogs, and white papers.

For more information please visit www.r20.nl, or send an email to rick@r20.nl. You can also get in touch with him via LinkedIn and Twitter (@Rick_vanderlans).

About Denodo

Denodo 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 by enabling faster and easier access to unified business information for agile BI, big data analytics, Web, and cloud integration, single-view applications, and enterprise data services. Denodo is well-funded, profitable, and privately held. For more information, visit www.denodo.com.

[2] Onalytica.com, *Business Intelligence – Top Influencers, Brands and Publications*, June 2018; see http://www.onalytica.com/blog/posts/business-intelligence-top-influencers-brands-publications/

[3] See http://www.b-eye-network.com/channels/5087/view/12495

[4] R.F. van der Lans, *Introduction to SQL; Mastering the Relational Database Language*, fourth edition, Addison-Wesley, 2007.

[5] R.F. van der Lans, *Data Virtualization for Business Intelligence Systems*, Morgan Kaufmann Publishers, 2012.