Modernizing Data Architectures for a Digital Age Using Data Virtualization

A Whitepaper

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

October 2019

Sponsored by

Table of Contents

1 Executive Summary

Introduction – Countless organizations have started *data architecture modernization* projects. These projects are initiated because they understand the increasing business value of data. By using it more effectively, more widely, and more deeply they can improve and optimize business processes and decision-making processes. It is regarded a requirement in order to stay competitive, cost-effective, and efficient. In essence, organizations want and need to become more data-driven in order to effectively participate in the emerging digital economy.

Current data architectures are not sufficient and need to be modernized.

They understand the potential business value of data, but equally they understand that the architecture of the existing data delivery systems, such as the data warehouse (and to some extent their data lakes), is not sufficient. These systems that rely on physical data movement and redundant data storage do not always have the right performance, scalability, and functionality. In order to future-proof data architectures for the next evolution of analytics their current systems need to be modernized.

New Business Requirements – A wide range of reasons exists why organizations need to modernize, such as enabling self-service reporting and dashboards; ensuring advanced analytics are equipped to work with (near) real-time data instead of yesterday's data; combining internal with external data coming from one of the many public, commercial, or social data sources to enrich analytical insights; deploying AI/machine learning/data science to discover patterns or trends in the data that may improve decisions, automate, or optimize business processes; and deploying IoT technology to monitor machines and business processes in much more detail to improve efficiency and reduce risk.

Additionally, other types of requirements have impact on BI systems as well, such as the impact of new regulations on data protection and privacy. Not everyone is allowed to see all the personally identifiable information (PII) anymore, some data needs to anonymized, and 'the right to be forgotten' needs to be implemented somehow. More business users and regulators require that the entire 'factory' that delivers them data becomes more transparent, which implies more up-to-date data catalogs and metadata.

Modernization and not Replacement – Modernization is normally not a simple matter of deploying more computer power; for example, by replacing one tool with another, by replacing the current database server with a faster one; or, by migrating data to the cloud. There is no quick fix. Nor is replacing the entire architecture with a new one a viable alternative. This is too risk prone. For modernization to be effective, it must be seamless. The current business operations cannot falter because of this exercise. Modernization of a data architecture is not a simple replacement of one architecture with a new one, but involves the improvement of existing modules and removal of weak modules.

Modernization of data architectures must be a seamless process.

This Whitepaper – This whitepaper describes how *data virtualization* can help provide a seamless evolution to the capabilities of an existing data architecture. With data virtualization data architectures can be modernized without disturbing the existing analytical workload. Basically, data virtualization can extend an existing data architecture to more quickly unlock and exploit all the existing data, to present more low latency data, and to support new forms of data usage, such as data science, without the need for mass replacement.

Data virtualization can help to modernize data architectures without mass replacement.

The data lake, self-service BI, cloud technology, and data catalog are often mentioned in modernization projects. The whitepaper also describes how data virtualization can help to simplify inclusion of such concepts in a new and future-proof data architecture.

2 The New World of Data Delivery

Data Has Become a Key Business Asset – It cannot have escaped anyone's attention that the role of data within organizations has been elevated. Especially top executives understand the potential business value of data. For a long time, storing and processing data was seen as a necessity. Data needed to be stored to support key business processes, such as order-to-bill, procure-to-pay, and plan-to-inventory. But so much more can be done with data. New technologies and tools allow us to be more creative with data. When used the right way, it can be used to strengthen the competitive situation, increase market share, optimize and streamline business processes, minimize industrial waste, optimize transport of goods, deploy resources more efficiently, automatically verify dubious financial transactions, and so on.

It has taken some time, but new initiatives, such as *digital transformation*, becoming an *insights-driven organization*, and the exploitation of AI and machine learning, have convinced top executives that data is a key business asset. Using data for competitive advantage has become a recurring topic at boards of directors meetings.

Raising the Bar for Data Delivery Systems – This growing importance of data raises the bar for the analytics systems that support the business and decision-making processes. Organizations want to do more with data, they want to exploit the data they have more effectively, efficiently, and widely. Examples of 'doing more' with data today means:

New requirements have raised the bar for analytics systems.

- Enabling self-service reporting, dashboards, and for analytics to work with (near) real-time data. Showing yesterday's data is not adequate anymore. Offering service interfaces with which external parties can request data without human intervention.

- Combining internal with external data coming from one of the many public, commercial or social data sources to enrich the analytical capabilities.

- Accelerating AI/machine learning initiatives where data scientists need to discover patterns or trends in the data that were unknown before. This can lead to predictive or descriptive models that help to improve decisions or optimize business processes.

- Simplifying deployment of IoT technology to generate more data on the machines and business processes. In other words, creating more detailed data to possibly get more insights.

- Offering edge analytics in which real-time data is analyzed continuously and near the place where the data is produced by the sensor or business process (the edge).

Other types of requirements have an impact on analytics systems as well, such as the impact of new regulations, for example GDPR, on data delivery, and the requirement by business users and regulators that the entire 'factory' that delivers them data is more transparent. Transparency implies that an organization is able to show how the data moves from start to end, from the data source to the reports, and how the data is filtered, cleansed, transformed, integrated, and aggregated by this process.

Upgrading Analytics Systems – Moving data into the spotlights of the organization has an enormous impact on the existing data and analytics systems, such as BI systems, reporting systems, data entry systems, and websites. The amount of data stored is increased, much more detailed data is requested, more reports are processed, more users are given access, more forms of data usage are deployed, the accuracy of data is improved, and so on. Additionally, because data becomes involved in so many decision processes, the quality of those decisions is directly dependent on the quality of the data. Incorrect or stale data leads to incorrect decisions.

Organizations have come to depend increasingly on the availability and quality of data.

In other words, because the bar for data usage is raised, organizations have come to depend increasingly on the availability and quality of data, and that impacts the analytics systems. However, implementing all the above requirements can disrupt existing systems, because it overreaches their technical limits. This leads to:

- poor performance
- increased time to market new reports
- system instabilities
- missed business opportunities
- isolated report development in the business departments

The key reason for this disruption is that the existing data architectures, integration tools, applications, and processes were initially not selected and designed for this new, comprehensive workload. They have been stretched beyond their limits.

The existing systems have been stretched beyond their limits.

Modernizing the Data Architecture – The solution is normally not a simple matter of deploying more computer power; for example, by replacing one tool by another, by replacing the current database server by a faster one; or by migrating to the cloud. There is no quick fix. Quite often, the entire architecture responsible for data delivery must be re-evaluated and modernized. Serious changes need to be made. Data flows may have to be redesigned, extra data stores may have to be developed, sourcing the entire system to a scalable cloud platform may be the solution, other techniques for designing data structures may have to be used, and so on.

But whatever needs to be done to support the new workload, replacing the entire architecture by a new one is almost never an option, because it is considered too risk prone. If modernization takes place, it must be seamless. The current workload may not falter because of this exercise. Modernization of the data architecture is not a simple replacement of one architecture by a new one. Modernization involves the improvement of existing modules and the removal of weak modules.

Modernization involves the improvement of existing modules and removing of weak modules.

3 Challenges of Designing New Data Architectures

The previous section mentioned several business reasons for the modernization of data architectures. This section describes some of the technical challenges resulting from those business reasons that demand development of new architectures for data delivery systems.

Real-time Data – Business users increasingly need access to near real-time data. Especially operational decision makers, the workforce itself, and online customers can't work with yesterday's data.

Additionally, for many automated decisions *real-time data* (or zero-latency data) is a necessity. Data must be analyzed right after it has been produced. The requirement for real-time data applies to, for example, numerous industrial environments, gaming systems, and financial systems.

Real-time data is a necessity for many automated decisions.

Technically, this means that data must be moved to the data consumers after production as quickly as possible. Most existing systems were not designed for real-time data usage and don't include the right technologies. The way they are designed now by copying the data several times from one database to another using several ETL processes, won't work. It creates too much data latency and potentially introduces errors in the process. A new data architecture must be able to support functionality to stream data to data consumers or to replicate data in real-time.

Preserving Data History – In the old days disk storage was expensive. Engineers would do everything to design systems in such a way that the amount of data stored was minimal. One of the prevalent tricks used to minimize data storage was limiting the data history. Most changes to data were not kept. Old values and records were simply deleted and replaced by new ones. Sometimes no history was preserved at all or only a limited history of data was kept.

Removing data saves disk space, but it also limits analytical capabilities, especially the ones in which processes need to be analyzed historically. A requirement for the organization may be to store a complete history of most of the data. This can have a serious consequence for the tools and technologies required in a new data architecture and for how we keep track of all that data.

Organizations may require a complete history of data.

Big Data – Related to the previous topic is one that is still quite popular in the IT industry, namely *big data*. Note that this term is starting to fade, because large data volumes are becoming the norm for the amount of data organizations have to manage.

Big data means different things to different organizations. For example, it can mean that organizations want to analyze all the texts (e.g. contracts and social media messages), voice messages (e.g. conversation between air controllers and pilots and of call center operators), images (e.g. car damages due to accidents), and videos (e.g. from security cameras and cameras at airports and retail stores). Big data can also mean that more business processes are monitored in detail resulting in massive amounts of detailed data indicating the progress and success of those processes. Again, this will impact the technology and tools to be used. Big data can also refer to streaming data that might need to be analyzed right after it has been produced.

Undoubtedly, whatever big data means for an organization, the new, increased amount of data stored, processed, and analyzed will have a serious impact on the technology used in the new data architecture.

Big data has a serious impact on the technology used.

Bi-modal Support – Most existing analytics systems are developed to support traditional forms of reporting, dashboards, and analysis. These forms of analytics must be governable, auditable, and formally tested. Lately, more and more business departments want to develop their own reports using *self-service analytics tools*. These self-service reports may be exploratory in nature and have a short life span. Some reports are even developed for one-time usage only to solve a problem.

The term *bi-modal* refers to implementing these more agile, experimental solutions in combination with the more stable, traditional, and formal forms of reporting. This combining of self-service reports and formal reports normally leads to a clash of cultures and technologies. It is a challenge to develop a single data architecture that supports bi-modal satisfactorily. It is like designing a car that must be

Implementing bi-modal in a new data architecture is a challenge.

superfast and yet can carry a heavy load. Those two requirements contradict.

Cloud Platform Independency and Interoperability – Moving applications and systems to the cloud is not unusual anymore. Many cloud platforms exist, including those from Amazon, Google, and Microsoft. They all have their strong and weak points. A new data architecture should be as cloud platform independent as possible in order to allow an organization the flexibility to switch between technologies that best meet their requirements.

Cloud platform independency and interoperability are key requirements.

Additionally, organizations may already use multiple cloud platforms. In this case, the new data architecture must support *cloud platform interoperability*. For example, it must be transparent for reports requiring data from two systems running on two different cloud platforms. Cloud platform independency and cloud platform interoperability are key requirements for modern data architectures.

Data Science – Data science enables organizations to create analytical models to support them with better and/or faster decisions. Techniques such as statistics, deep learning, machine learning and AI are deployed to create these analytical models. First, a new data architecture should help data scientists to easily develop new models in an experimental and investigative way. Second, studies show that data scientists spend commonly 80% of their time on data preparation work and only 20% on the real analytical work. A new data architecture must assist in shortening the time it takes to acquire and prepare data.

New data architectures must shorten the data preparation time.

Data Streaming – The need for real-time data can result in the requirement to process streaming data, such as IoT data. Data is streamed (or pushed) directly from the source to feed real-time streaming applications that transform or react to new information. New data architectures need to accommodate streaming data with systems that can support asynchronous publish and subscribe scenarios similar to a message queue or enterprise messaging system. This may have to happen without any form of temporary data storage. Many questions need to be answered, such as how is instant data cleansing implemented, how fast can analytical models be deployed, can the infrastructure process the data streaming workload, does the streamed data be stored as well and can the database technology keep up with the data ingestion rate produced by the streams, and so on. A new data architecture has to support this streaming of data.

New data architectures must support data streaming.

Service-Oriented Interfaces – Not all data consumers are humans. Increasingly, data needs to be made available through decoupled, stateless service interfaces. For example, services must be developed that are externalized allowing other parties, such as suppliers and agents, to build applications that access them. In this case, the data consumer is not a human but an application. Similarly, the platforms on which the new data architecture will operate may also be service-oriented, deploying microservices and containers.

Data needs to be made available through service interfaces.

A consequence of this type of data usage is that it can lead to an enormous and highly unpredictable query workload. The technologies used in the new data architecture must be powerful and scalable enough to support it.

These are some of the challenges facing organizations when they design their new data architectures and when selecting new technologies, tools, and techniques.

4 Data Virtualization as Catalyst for Modern Data Architectures

Existing data delivery systems can be modernized and be made future-proof in many different ways and through many different technologies and tools. A technology that can help with an almost seamless modernization of a data architecture is *data virtualization*. A common, somewhat descriptive definition[1] of data virtualization is: *Data virtualization offers data consumers a unified, abstracted, and encapsulated view for querying and manipulating data stored in a heterogeneous set of data stores.*

The common risk of modernization is that it leads to the existing reporting system being temporarily offline. Data virtualization enables data architectures to be modernized without disturbing the existing reporting and analytics workload. This section lists some of the features of data virtualization that can ease data architecture modernization.

Through query pushdown the full query power of a database platform is used.

Improved Performance Through Query Pushdown – When data virtualization servers receive a request for data, they can determine how much of the processing is executed by the data virtualization server itself and how much by the underlying data sources. Pushing query processing to the data source is called *query pushdown*; see Figure 1. The advantage of this approach is that the full query power of that database platform is used. This is especially valuable for new platforms, such as Hadoop and some of the analytical SQL systems, because they offer fast parallel query processing capabilities. Query pushdown capabilities can speed up performance by exploiting that query power.

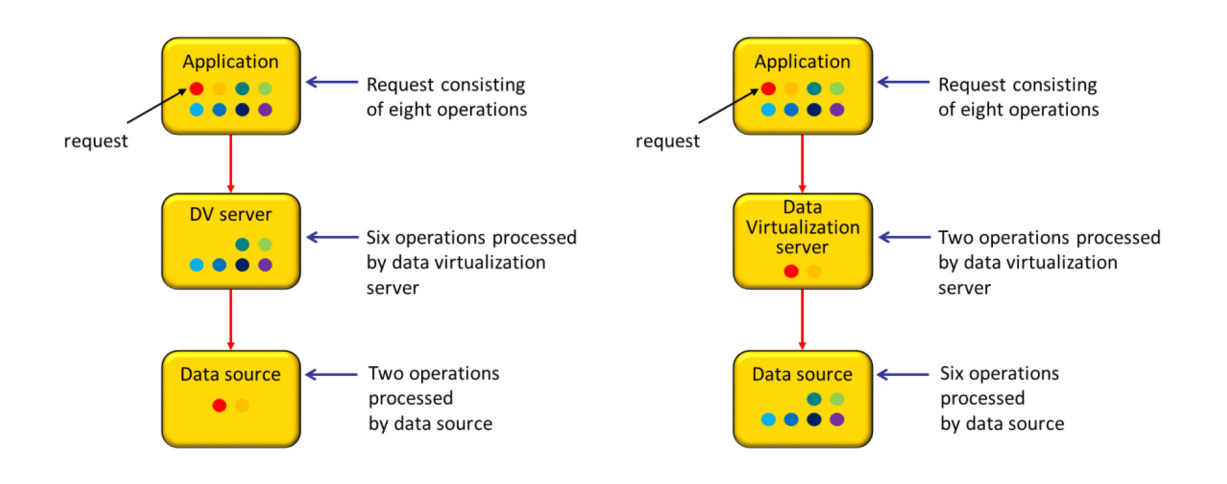
Application
request
Request consisting of eight operations
DV server
Six operations processed by data virtualization server
Data source
Two operations processed by data source
Application
request
Request consisting of eight operations
Data Virtualization server
Two operations processed by data virtualization server
Data source
Six operations processed by data source

Figure 1 *Using query pushdown to exploit the full query power of the underlying data source.*

Data virtualization supports a centralized security layer.

Improved Data Security Through Centralization – When data is dispersed across several different database platforms, those responsible for data security have to deal with different security systems. This means that for every user different types of specifications for authorization, authentication, and anonymization need to be defined to indicate what this user is allowed to do with the data. Potentially, this is a security nightmare. How are all those specifications kept in synch? Take as example an employee who resigned. He has to be removed from the user list from every system that he had access to.

[1] R.F. van der Lans, *Data Virtualization: Selected Writings*, Lulu.com, September 2019; see http://www.r20.nl/DataVirtualizationBook.htm

With data virtualization one centralized security layer of specification can be defined and applied across all data sources. The ability to specify who is allowed to see which part of the data needs to be defined only once.

Improved Data Anonymization Through Filtering – New regulations and laws are forcing organizations to implement anonymization rules regarding *personally identifiable information* (PII). Some PII needs to be masked, other PII must be garbled in such way that it can't be traced back to a specific person. Data virtualization servers support filtering and transformation features to implement many of those anonymization rules. This can be done without impacting the original databases. As with data security rules, these anonymization rules are defined centrally and can operate on data that is virtualized from any kind of database platform.

Data virtualization supports features for data anonymization.

Note that the stored data remains unchanged, unmasked, and ungarbled. Evidently, if the need exists to query the data directly and bypass the data virtualization server, then the anonymization rules must be implemented in the database platform itself.

Improved Development Speed Through Centralized Specifications – To transform data coming from a variety of data sources into meaningful data for business users, it needs to be processed, which means it needs to be filtered, cleansed, integrated, aggregated, calculated, and so on. In traditional analytics systems, these *transformation specifications* are implemented across the entire system, in the reports, in the ETL programs, inside the databases themselves, and as snippets of code inside handwritten programs. This is detrimental to productivity and maintenance.

Data virtualization servers allow most of the transformation specifications to be defined centrally in a *repository*; see Figure 2. Specifications are defined once and can be reused many times and thus improving development speed and easing maintenance Additionally, built-in lineage and impact analysis capabilities allow developers (and business users) to study the specifications and see how they are related.

Transformation specifications can be defined centrally with a data virtualization server.

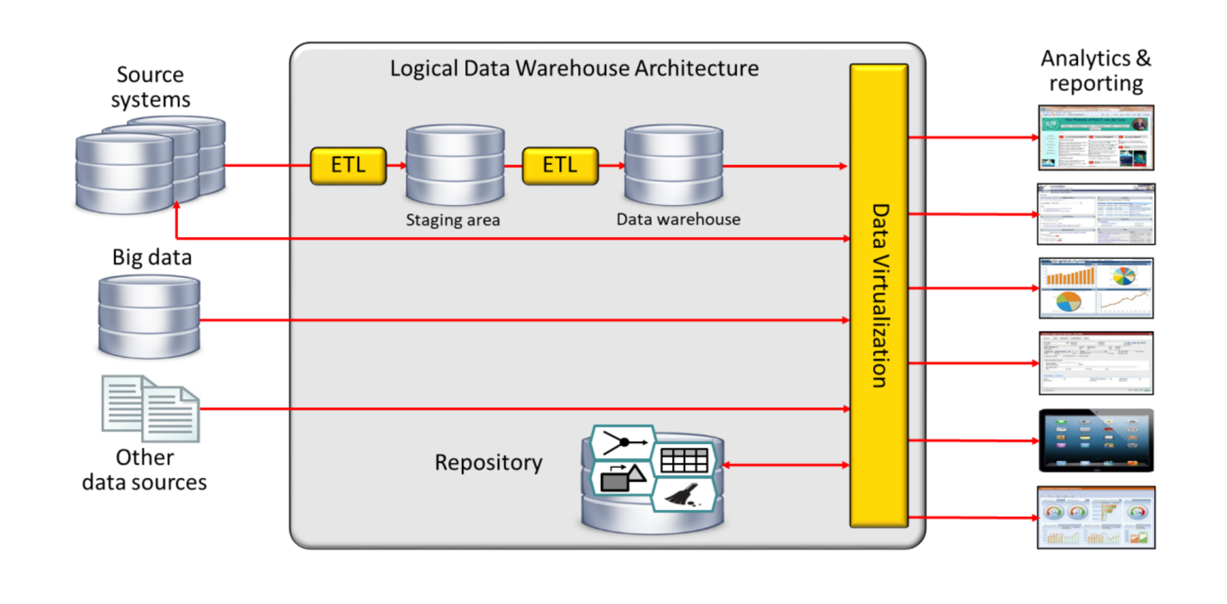

Figure 2 *Data virtualization servers store all the transformation specifications in a central repository.*

Easy Database Migration Through Database Independency – Data virtualization servers can extract data from all kinds of database servers, including most SQL products, Hadoop, spreadsheets, JSON and XML sources, and applications. Reporting tools can use a data virtualization server to access any of those database servers using one and the same language. Whether they access data stored in Hadoop, Oracle, or SnowflakeDB, all these systems can be queried using the same statements. In other words, data virtualization makes the reports database technology independent.

This feature can be used when the need exists to migrate data to a modern database platform; see Figure 3. In Phase 1 the data virtualization server layer is implemented in between the reports and the existing database. In Phase 2, the data is copied to the new database, and in Phase 3 the data virtualization server is redirected to access the new database. This can all be done without making no or minor changes to the reports. Note that the migration of data can be done table by table, or all the data can be migrated in one big step.

Data virtualization makes reports database technology independent and therefore easier to migrate.

Database independency is a valuable feature to move gradually and seamlessly to a new data architecture.

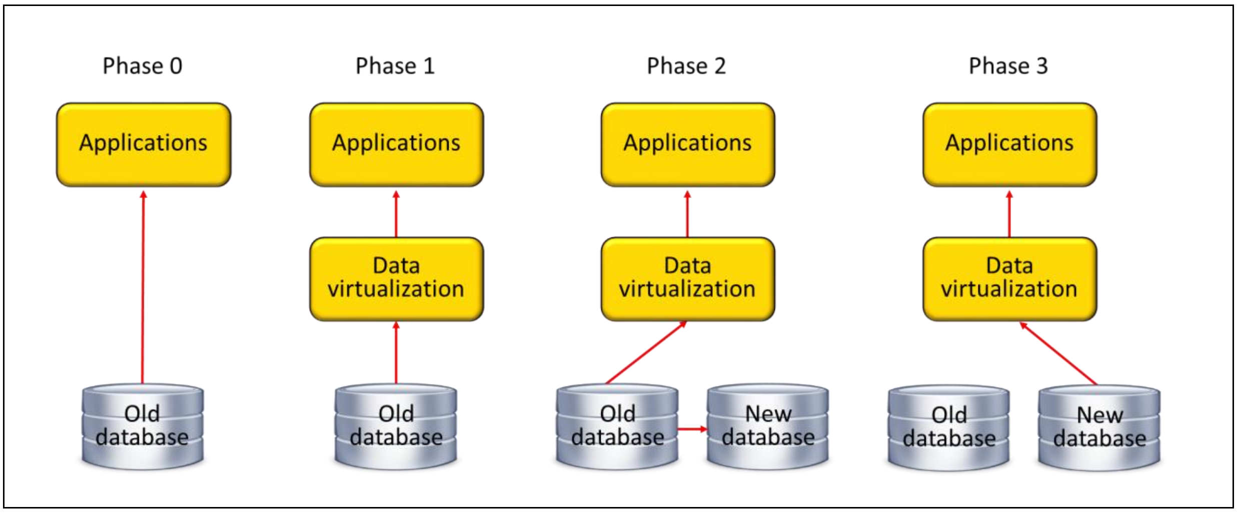

Figure 3 *Migrating to a modern database platform.*

Decouple through Data Abstraction – Data virtualization can be used to provide a *virtual data abstraction layer* implemented atop an entire data architecture, decoupling the reporting from the technologies used to store and process the data; see Figure 4. With data virtualization acting as an abstraction layer, changes can be made to the data storage layer without impacting the reports. For example, different technologies can be used to copy the data from one database to another, database layers may be skipped to speed up data delivery to the reports, different database design techniques can be used to keep track of more data history, other data warehouse solutions can be linked to this architecture but still presenting it as one integrated architecture, and so on. In other words, with data virtualization, the data storage layer can be modernized without impacting the reports.

Data virtualization decouples reports from the existing data architecture.

Note that this wrapping approach can also help to integrate new reporting systems coming from an acquisition of another organization.

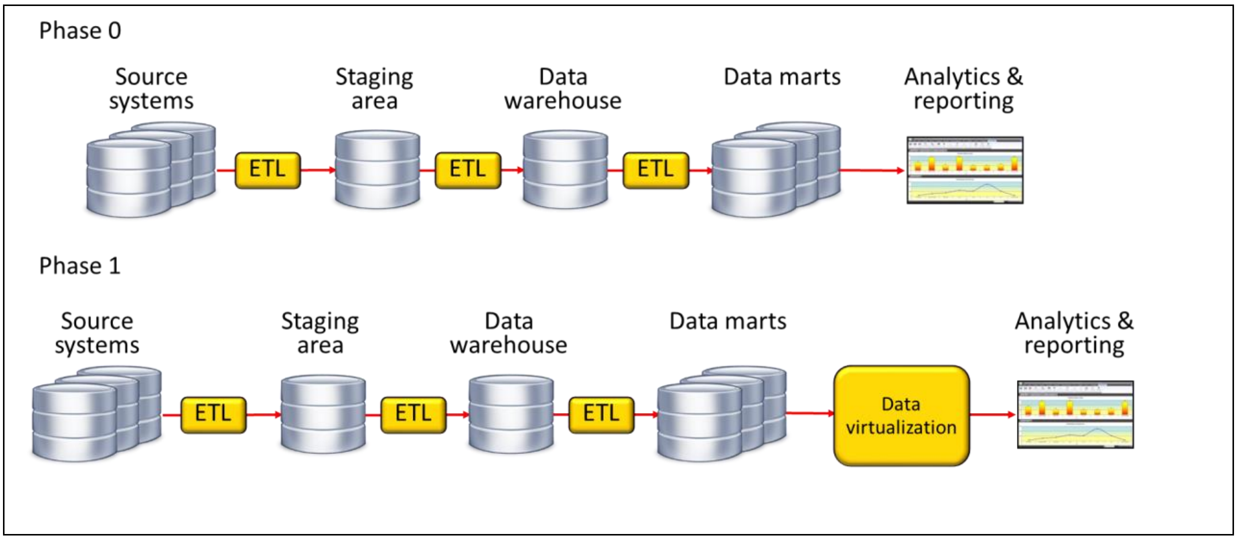

Figure 4 *Wrapping an architecture with data virtualization.*

Improved Data Transparency Through the Data Catalog – Increasingly, organizations are turning to data catalogs as a means to store valuable data assets for fast retrieval. Within data virtualization, a data catalog can be used to store the metadata or virtual views that have been defined, described, and documented. Additionally, these can be tagged and enriched with additional descriptions to make them more understandable and more easily searchable. Developers, business users, and data scientists can gain practical benefits from a data virtualization catalog, using it to search for data elements they need for their reports or algorithms, and is especially helpful for more ad-hoc investigative environments.

Data virtualization servers support a central and searchable data catalog.

Easy Cloud Adoption Through Abstraction – Modernization of data architectures may involve the adoption of a *cloud platform*, such as those from Amazon, Google, or Microsoft. Technically, this can mean several things. First, it can mean that a database is migrated to the cloud. This migration of a database can be hidden for all the reports and applications through a data virtualization server. In this case, the reports and data virtualization server still run on premises and the latter takes care of all the access to the cloud-based database; see Figure 5. This migration of the data doesn't need not to be executed in one step, but can be done gradually. The data virtualization server handles the queries that join data stored in the new cloud based platform and data still remaining in the on-premises database. Especially if databases are large, such a gradual or phased data migration is preferred to guarantee 100% data availability.

Seamless and gradual migration to cloud platforms using data virtualization.

Second, if an organization wants to exploit the fast analytical database technology available in the cloud, a data virtualization server can hide the fact that the data is not only moved to the cloud but also to another database platform. The data virtualization server hides the fact that the data is accessible through a slightly different SQL dialect and uses query pushdown to exploit the full power of the new database platform.

Third, the organization has started to use proprietary cloud applications, such as Salesforce.com and NetSuite. The need to integrate data produced in these cloud applications must be accessible for reporting and integration with other databases. Data virtualization servers can access cloud applications as if they are databases.

Fourth, if the entire architecture needs to be migrated to the cloud, the virtual data abstraction layer capabilities enable this step by step.

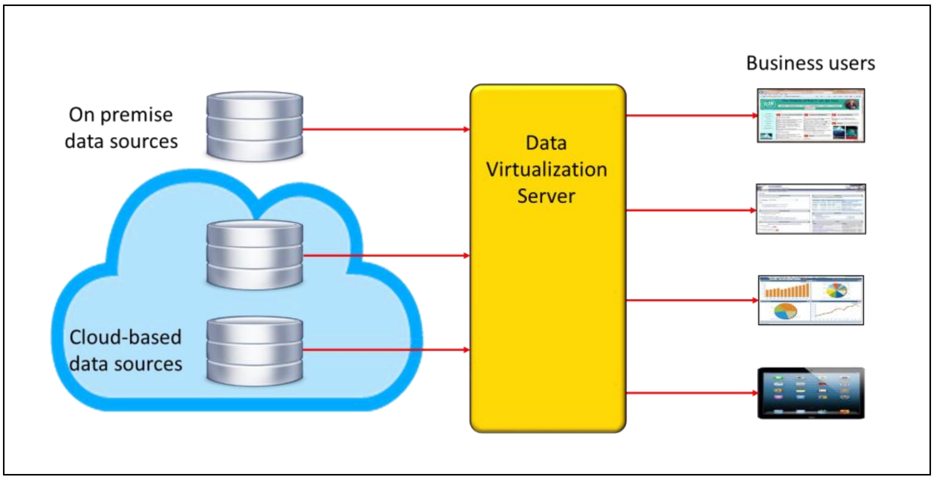

Figure 5 *Hiding the migration of a database to a cloud platform with data virtualization.*

> *Data virtualization can deliver real-time data to reports by accessing transactional systems.*

Real-time Data Delivery Through On-Demand Access – Data virtualization servers can access any kind of database, including the databases belonging to transactional systems. Data virtualization servers can help reports to show real-time data stored in those transactional databases; see Figure 6. The challenge of developing such reports is that they do not interfere with the running transactions, nor should the reports be slow because of this. Data virtualization servers support features to cater for this, such as advanced query optimization, the ability to force the data virtualization server to execute the precise queries you want, and caching transactional data.

Also, the query pushdown capabilities allow a data virtualization server to use the full power of the underlying database platform. And if the user accesses large amounts of data, data virtualization uses the parallel query processing capabilities to speed up performance. Caching can also help to minimize interference, by caching some of the data that barely ever changes, such as tables with reference data.

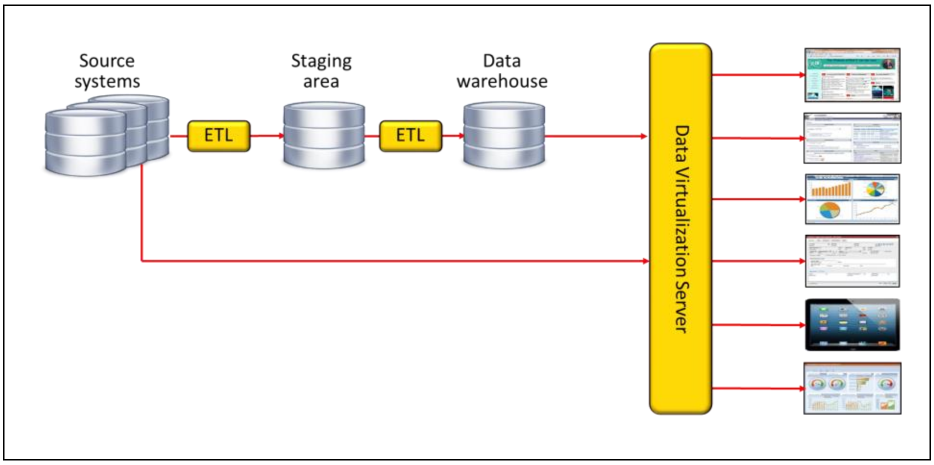

Figure 6 *Data virtualization can present real-time data by accessing transaction systems directly.*

To summarize, data virtualization helps to modernize existing data architectures seamlessly and gradually. There is no need for a complete replacement of the entire architecture, which would be highly risk prone.

5 Other Solutions for Data Architecture Modernization

Other concepts and technologies are often proposed to help modernize a data architecture. This section describes for each of them their respective drawbacks and benefits. Additionally, this section explains how data virtualization can help to simplify adoption of such concepts in a new and future-proof data architecture.

The data lake is the focal point for every form of data usage.

Data Lake – A form of modernization is by copying all (or most) of the original data to a central data lake. In this case the data lake becomes the focal point for every form of data access and data delivery. This can be done roughly in two ways. First, some organizations include the data lake in between the source systems and the data warehouse; see Figure 7. In this way the data lake acts as, what used to be called, an *operational data store*. Such a data lake is used by business users, such as data scientists, directly and by the ETL processes that copy the data from the lake to the data warehouse. Replication or streaming technology is used to keep the data lake up to date. From a complete data history perspective and integration of data, this is an interesting alternative. However, it scores poorly on making real-time data available.

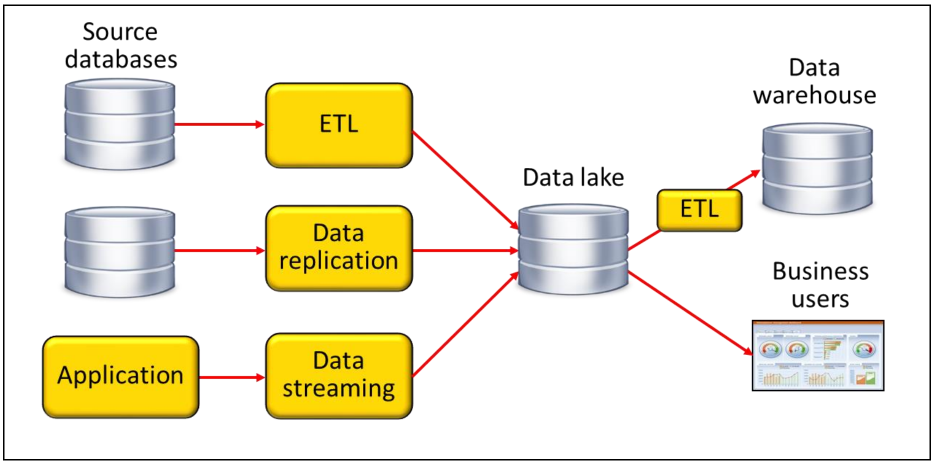

Figure 7 *Including a data lake in an existing data architecture.*

Second, the data lake is introduced as a central repository for storing all the enterprise data and external data. Here, the data lake replaces the existing data warehouse or similar systems. Implementing such a data lake has an enormous impact on a large part of the data architecture. Reports may have to be changed and ETL processes as well.

To make this work, all the data needs to be moved to the data lake. The solution would be to link the data lake into the existing data architecture with data virtualization. The introduction of a data lake in an existing data architecture can lead to migration of reports currently operating on a data warehouse, a data mart or some other database. This may not be a simple one-to-one migration. Data in a data lake is quite often structured differently than in one of the other databases mentioned. This can mean that the code accessing the data must be redeveloped. With a data virtualization server this changed data structure can be hidden for the reports. This simplifies the introduction of the data lake within an existing data architecture.

Cloud platforms unburden organizations and offer limitless scalability.

Cloud Platforms – Cloud platforms have proven their value to organizations. Evidently, different organizations experience different business benefits. One dominant business value is the *unburdening* that cloud platforms deliver. Organizations don't need to manage and monitor their hardware environments anymore. No reason to equip special data center rooms with hardware and raised floors, deal with fast communication lines, and control the temperature and adhere to the right service level agreements. A second business benefit is the almost unlimited availability of processing and storage capacity. This offers an organization easy and limitless scalability; additional compute or storage power can be available within seconds when needed.

Cloud platforms play an important role in modernizing data architectures. Some organizations move their existing software components to the cloud. Others try to exploit the cloud platforms optimally by also adopting *native technologies* developed and optimized specifically for specific cloud platforms, such as Google BigQuery and Amazon Athena. If reports are developed straight on these systems a certain lock-in develops. The dependency on a specific cloud platform increases when such native cloud technologies are deployed. In other words, these technologies make it harder for customers to migrate again in the future. As indicated in this whitepaper, the dependency of these optimized but proprietary technologies can be minimized by using data virtualization, while retaining all the performance, scalability, and flexibility benefits.

Many organizations have already reached the stage in which they use multiple cloud platforms. For example, they run SalesForce.com on one cloud, some of their data is stored in Hadoop on AWS, and they have installed SAP on Azure. Regardless of the rest of the data architecture, a solution is needed to integrate data coming from these multiple cloud platforms. Technology-wise this is a *hybrid-cloud query*. Data virtualization can handle this type of query and can completely hide the fact that data is stored on different cloud platforms.

Self-Service Analytics – Sometimes the new requirements of business users are not implemented by modernizing the architecture, but by giving report development functionality in the hands of business users. Self-service analytics is a good example of this. Reports in the old days were always developed by IT specialists. Nowadays, some of the products are sufficiently intuitive for business users to develop the reports themselves. In other words, if modernization means that the functionality and the presentation forms or reporting and analytics need to be improved, this is definitely an interesting route.

Self-service analytics is about giving report development functionality in the hands of business users.

In fact, some of these tools now even support integration functionality. Quite easily, business users can define how to integrate data from multiple sources. This is sometimes referred to as *citizen integration*. Terms as data wrangling and data preparation refer to similar functionality. All these terms mean that more integration and preparation capabilities are available to business users. Note that the limiting factor is still that they must access databases developed and owned by IT. So, they are still confined by that world.

Self-service analytics has much to offer to an organization. The big drawback is that many specifications for integration, aggregation, cleansing, and filtering, are developed in an isolated fashion. This possibly leads to inconsistent and incorrect report results, it decreases productivity of the business users, it impairs maintenance, and the quality of the specifications developed by the business users is hard to test and rarely ever auditable.

Data virtualization can minimize this risk by allowing many of these specifications to be defined centrally. In this case, the specifications can be shared and reused by all the self-service reports. They can even be shared across different tools. This solves the problems described.

Leveraging Metadata – In many modernization projects the focal point is improvement of maintaining and delivering metadata. Organization-wide a growing need for data transparency, data governance, and searchable metadata exists. The focus in such projects is not only on developing a 'better' data architecture, but also the ability to describe and search metadata, especially business metadata.

A few alternative solutions exist. Many of those solutions allow one to build up a standalone data catalog in which all the metadata is stored, defined, documented, and linked. Business users and developers can browse this catalog and search for specific terms and see lineage. Unfortunately, with these standalone solutions most of the metadata is not always automatically updated when parts of the systems are deleted or changed. To update them, a manual or partly automated action has to take place. The risk is that this is omitted and the data catalog becomes slowly out of date and eventually obsolete.

Data virtualization servers, such as the Denodo Platform, support a built-in data catalog. When a data object is defined, it can be described with technical and business metadata. Lineage specifications are automatically kept up to date. Deleting a data element automatically changes the catalog. In other words, the catalog is more integrated with the operational systems than the standalone solutions and therefore always up to date.

The Denodo Platform is one of the few data virtualization with which metadata elements can be linked to reports. Evidently, this data catalog won't contain all the metadata.

The Logical Data Warehouse – The most obvious approach of modernization is the one where the existing architecture is extended or some components are replaced. For example, with the *logical data warehouse architecture* the existing classic data warehouse architecture can be wrapped; see Figure 8. In this new architecture all the data consumers access the data via the data virtualization server. Some of the data accessed may come from the existing data warehouse, some may come from new big data sources, or from any other data source. The data virtualization server hides where the data comes from. This permits a gradual introduction of data virtualization and a gradual migration to a completely new architecture. For example, when all the data consumers access all the data via the data virtualization layer, it becomes easier to make changes to the data warehouse. The data structure can be changed or the way how the data warehouse is loaded without impacting the reports.

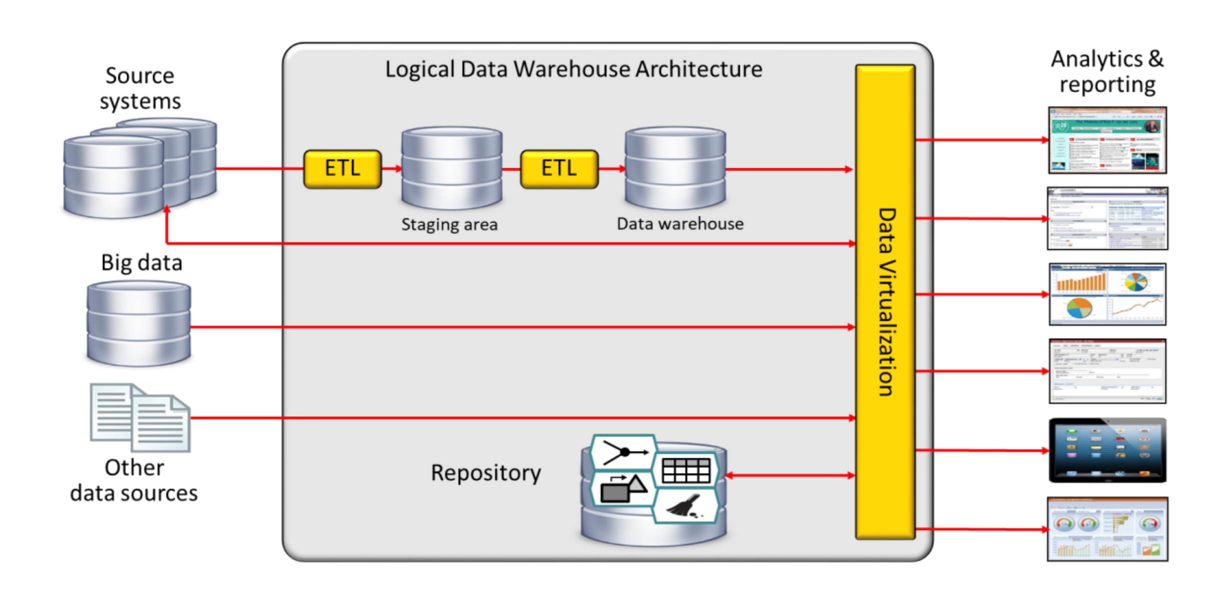

Figure 8 *High-level overview of the logical data warehouse architecture.*

6 Closing Remarks

A full rip and replace approach of data architectures is not an option.

It is naive to think that an existing data architecture can be switched of on one day and the new one switched on the next day. For numerous reasons a full rip and replace approach is in most situations not realistic. For example, if the architecture supports 24x7 data access, when should the switch be made? Or, a switch may involve closing the old system and copying the data to the new system before it can be made operational, and especially if it concerns big data, just the copying process itself may take too long.

The ideal migration is a seamless and gradual one, one barely noticed by existing data consumers. The ideal migration is an evolutionary migration from the old to the new data architecture. Step by step, small modules of the existing data architectures are replaced by new ones. Due to a gradual migration, a quick modernization makes it possible for an organization to compete in the current digital age and fulfill its digital transformation dream.

The many wrapping and abstraction features of data virtualization servers offer such a gradual and seamless migration. New modules can easily be added to the existing architecture, old ones can be replaced. Also, the adoption of cloud platforms becomes easier and is less risk prone when a data virtualization is able to hide it. It also makes the new data architecture less dependent on one cloud platform and its native technology.

Other technologies and approaches used for modernization, such as data lakes, self-service BI, cloud platforms, and data catalogs, can also be relevant. If so, data virtualization can work with all of them. In fact, most often data virtualization can amplify their benefits.

Combining data virtualization with data lakes, self-service BI, cloud platforms, and data catalogs has a synergetic effect.

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.

He is the author of several books on computing, including his new *Data Virtualization: Selected Writings*[4] and *Data Virtualization for Business Intelligence Systems*[5]. Some of these books are available in different languages. Books such as the popular *Introduction to SQL* is available in English, Dutch, Italian, Chinese, and German and is sold worldwide. Over the years, he has authored 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[6], entitled *Introduction to SQL*, which has been translated into several languages with more than 100,000 copies sold.

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 Axians Business Analytics Laren (formerly Kadenza): This consultancy company specializes in business intelligence, data management, big data, data warehousing, data virtualization, and analytics. In this part-time role, Rick works closely together with the consultants in many projects. Their joint experiences and insights are shared in seminars, webinars, blogs, and whitepapers.

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, *Data Virtualization: Selected Writings*, Lulu.com, September 2019; see http://www.r20.nl/DataVirtualizationBook.htm

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

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