research

CHECKLIST REPORT

2018

Achieving High-Value Analytics with Data Virtualization

How organizations can reduce the cost of diverse data access and increase analytics performance

By David Stodder

Sponsored by

denodo

tdwi
Transforming Data
With Intelligence™

JULY 2018

TDWI CHECKLIST REPORT

Achieving High-Value Analytics with Data Virtualization

How organizations can reduce the cost of diverse data access and increase analytics performance

By David Stodder

Transforming Data
With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

TABLE OF CONTENTS

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies. Inclusion of a vendor, product, or service in TDWI research does not constitute an endorsement by TDWI or its management. Sponsorship of a publication should not be construed as an endorsement of the sponsor organization or validation of its claims.

FOREWORD

In most industries today, increasing the power and scope of analytics is critical for successfully solving business challenges and developing innovative products and services. Decision makers in every line of business and department need analytics insights to engage effectively with customers, run operations efficiently, and evaluate risks and threats. Analytics projects tend to have a voracious appetite for data; this means that as more executives, managers, and frontline personnel work with analytics, organizations need to make careful decisions about the best way to enable personnel to find, access, and analyze their data rapidly and cost-effectively.

Analytics is also highly varied. Decision makers in most organizations begin with *descriptive analytics* to understand what has happened. They typically look at business intelligence (BI) reports of historical, aggregated data about topics such as changes in sales over time, customer spending, and inventory. More advanced decision makers are interested in *predictive insights*; they typically must rely on data scientists, statisticians, and data analysts to build models and algorithms that can analyze different data types to discover patterns and data relationships that predict what could happen.

The most advanced organizations are taking predictive insights further into *prescriptive analytics*, determining what actions they can take to produce better business outcomes and developing optimization algorithms to automate responses to predicted events or patterns.

Data virtualization helps organizations address the challenges of an expanding base of users engaging in all types of analytics, which demand access to an increasing variety of data sources. Data virtualization technology provides a virtual data layer that shields users from having to know details about the data as they access and manipulate it, such as its physical location or its format. Data virtualization is the core technology behind logical data warehouses; it enables users to connect to heterogeneous data sources without waiting for the data to be extracted and loaded into a central, physical data warehouse.

The "location transparency" achieved through data virtualization can reduce an organization's dependence on a single source of data, such as an enterprise data warehouse, Hadoop-based data lake, or cloud-based storage system. Organizations can choose to keep the data where it is and perform analytics there, often a necessity for meeting regulatory and privacy requirements.

Data virtualization is particularly helpful when users need actual or near real-time data access from multiple systems of record (such as ERP or CRM) in a complex data architecture. Users can assemble data views quickly by interacting with a universal middleware layer to meet dynamic needs rather than wait weeks or months for new data to become available in a typical enterprise data warehouse. Analytics closely tied to operational business decisions often demands such timely data access.

This Checklist Report discusses six areas that are critical to achieving high-value, business-driven analytics and the role data virtualization plays in realizing success in these areas.

☑ NUMBER ONE

APPLY DATA VIRTUALIZATION TO HIGH-VALUE ANALYTICS AND BI USE CASES

Organizations should aim data virtualization at their highest-value analytics use cases, that is, projects where faster and less-complicated data access and integration would have the biggest impact on business outcomes. Often, these are projects for which executive management is specifying the goals, time frame, and intended benefits. These goals are important in framing data access, integration, and management needs. Let's consider three common high-value use cases.

First, many organizations turn to analytics to better understand how to significantly change their business models, such as by adding new sales channels, revenue streams, or data services for partners or customers. Typically, the organization does not want to wait for an extended data warehouse development process; however, spinning up new data platforms on premises or in the cloud to support the new initiatives will create more data silos. A data virtualization solution can provide a semantic layer that enables easier integration of respective data definitions and catalogs created for the underlying data source systems. The semantic layer is helpful when new initiatives add or change data definitions and structures. It can make data integration easier and give users location transparency for BI and analytics.

A second use case is when an organization needs to analyze data from both legacy and new systems, including data created in the cloud. This situation can create confusion that slows down analytics projects. Location transparency provided by data virtualization solutions can play a critical role. Using an abstraction layer, data virtualization masks the complexities of the underlying sources. Analysts, users, and data scientists can focus on analyzing the data and developing models and algorithms regardless of where the data comes from.

A third use case is when an organization has created a data lake or hub, often built with the Hadoop Data File System (HDFS) or other NoSQL technologies, and it now wants to apply advanced analytics to the data. Organizations typically find that data proliferates in the lake or hub quickly, with fragile and chaotically changing data sets. To reap value sooner, the organization chooses not to wait to develop data warehouses or data marts fed by the lake or hub and instead relies on data scientists and developers to write specialized, ad hoc coding routines to access data in those sources. These routines are often flawed and redundant, making it difficult for the organization to derive sustained value.

Data virtualization allows organizations to think differently about such projects. Data virtualization can offer a viable alternative to setting up a data lake or hub. Leading solutions now employ massively parallel processing (MPP) technology in their own middleware layers to accelerate performance and scale as data volumes rise. Thus, rather than go to the time and expense of building the data lake and moving data into it—only to move data out again into a data warehouse or data mart for analysis—organizations could identify desired data sets in the source systems and access them using data virtualization technology.

However, if the organization has already built a data lake or hub to support big data analytics, a virtualization layer would still be helpful in enabling location independence over the hub or lake as well as other systems with relevant data. New data sets could be defined in the virtual layer so users could access and analyze data from both the big data platform and traditional systems transparently.

☑ NUMBER TWO

ENHANCE BUSINESS ANALYTICS AGILITY WITH DATA VIRTUALIZATION

With the pace of change accelerating in nearly all industries, business agility is a prized quality. Organizations need to harvest data so they can detect change, adjust processes and behavior in response, and be proactive in taking advantage of unanticipated opportunities. Data virtualization can increase an organization's ability to access and integrate data, thereby increasing agility.

Agility requires speed; advantages accrue to organizations that can respond quickly to changes in their environment. To do so, business decision makers need to access current, live data for BI visualizations and analytics; they do not have time to wait out traditional data warehousing extraction and loading processes. They may need to develop and run high-performance analytics models and algorithms against multiple data sources, including those set up dynamically on cloud platforms to meet immediate business requirements.

Data virtualization can help users interact with data quickly to achieve business agility. Solutions can create virtual views of data coming from live data sources. The underlying complexity is hidden; users don't need to understand how to code queries appropriately for each data source. Once received, the source platforms will interpret and execute queries properly. Data integration can be handled on demand by defining virtual data marts in the data virtualization layer. Users do not have to wait for transformation engines to store the results in an intermediate physical platform, which typically must be preconfigured by IT administrators. Users have the flexibility to change virtual views as needed for their BI and analytics rather than wait for IT administrators to redo ETL processes and replicate different data.

The location transparency afforded by data virtualization is useful for organizations that plan to store significant quantities of data in the cloud—not just with one cloud storage service but in a multicloud architecture. In addition to taking advantage of competition among cloud providers and ensuring platform elasticity, organizations with operations distributed worldwide often have to set up multicloud architectures for regulatory adherence so they can respect local data ownership requirements.

Modern data virtualization solutions can be hosted in the cloud; organizations do not need to manage their virtualization layer on premises. The virtualization layer can enable organizations to take advantage of the price/performance of processing the data and analytics in the cloud by using the platform's native database system. Because data virtualization can provide location transparency across cloud and on-premises data, users can view data for BI and analytics without needing to know where the data is physically located. For these reasons, organizations should evaluate data virtualization as a means of giving users the flexibility and agility to focus on the business domain rather than the complexities of the data access architecture.

NUMBER THREE

PROVIDE A COMPLETE VIEW BY INTEGRATING DATA VIRTUALIZATION WITH DATA CATALOGING

Data catalogs, glossaries, and metadata repositories have always been important, but as data volume and variety increase, these shared resources become even more vital for BI and analytics. Ideally, the data catalog is a central repository; it contains metadata—that is, descriptive information about data sets from one or more sources, how the data sets are defined, and where to find them.

Of course, organizations that have numerous databases and applications will often have multiple data catalogs, glossaries, or metadata repositories specialized for each platform or for their department or project, which can make getting an enterprise view of data and metadata difficult. Data lakes using HDFS present additional challenges because they often contain highly varied data that is not consistently documented.

Data virtualization solutions can help by providing services that hook into these catalogs and repositories to assemble a more complete view within one interface. Users can find and preview data through the data virtualization system and apply knowledge about the data coming from the data catalog, glossary, or metadata repository.

With unified views of data and metadata, BI and analytics users can search, query, and discover relevant data and content about customers, products, or other topics of interest more easily and quickly. In this way, data virtualization integrated with data cataloging can help reduce conflicts about who has the correct data. Users can feel more confident when they communicate and collaborate on BI and analytics that they are "speaking the same language."

Another important role for a data catalog, glossary, or metadata repository is to gather knowledge about data sets that users can share. This can include information about who produced the data set and its quality and relevance for certain requirements. Some combined solutions enable BI and analytics users and data stewards to tag and annotate data sets so others can determine more quickly whether the data sets are appropriate for their personalized views. Some solutions also collect usage statistics that help IT administrators and business users learn which data sets are the most and least used.

Shared resources such as data catalogs are critical to improving data quality. One of the biggest sources of delay in analyzing and visualizing data is quality issues, including misspellings, inconsistencies, and false values. With multiple sources including data lakes and cloud-based storage, data quality problems can become difficult to manage and resolve. Data virtualization can enable organizations to discover and fix data quality problems, including through use of profiling.

Finally, some data virtualization solutions can provide a virtual master reference data resource against which other sources can be compared. Organizations can use this golden record, based on the catalog, for integrating data coming from a data lake, which is often messy and full of data quality problems.

NUMBER FOUR

APPLY DATA VIRTUALIZATION TO EASE GOVERNANCE AND DATA LINEAGE TRACKING

Governance is rising in importance as data proliferates across sources and organizations collect and store potentially sensitive data, including personally identifiable information (PII) about customers, patients, partners, and other entities. Regulations such as the European Union's General Data Protection Regulation (GDPR) have raised the regulatory stakes of exposing PII, at least about EU citizens; noncompliance with GDPR could cost organizations 4 percent of their annual revenue or 20 million euros (about $24 million), whichever is higher. Organizations also need policies and practices for tracking and managing other types of sensitive data related to intellectual property, financial information, and strategic plans.

Governance can be challenging because of the volume and variety of data spread across multiple sources and the difficulty most organizations have in controlling ad hoc development and self-service data access. Data virtualization can help with governance on many fronts, but a primary tactic is using the layer to provide a single point of entry to the data sources from BI and analytics tools. Organizations can govern entry in one place—in the data virtualization layer—rather than tracking data access, replication, extraction, and transformation across numerous physical instances (such as data marts, spreadsheets, data lakes, and flat files).

Some data virtualization solutions also enable administrators to set policies for controlling access and unifying those policies with security procedures such as single sign-on. Organizations can define integrated access and security policies in the data virtualization layer. They can maintain governance and security authorizations in the layer even as user views change.

Logical views provided by data virtualization are also useful for setting up auditing and data lineage tracking procedures, which are essential to data governance. GDPR and other regulations require audits; some data virtualization solutions can track data lineage and provide an account for auditing purposes when necessary. Data lineage tracking can show the data path behind front-end reports, dashboards, visualizations, and analytics. It is important not only for governance but also for business decision making; data lineage is how users and analysts show the sources of their insights, how the data was transformed, and whether the data is fragile or stable over time.

Many organizations are extending governance beyond regulatory and security policies to stewardship of the data and ensuring that users and analysts are working with high-quality, trusted, secure data that is consistent and can be shared. Organizations can set up policies and use technologies such as data virtualization and data cataloging to support ongoing efforts to raise the quality and consistency of the data. By eliminating steps for making intermediate, replicated data stores for ETL and other processes, data virtualization can simplify data quality management and stewardship with fewer systems to track and better tracking of user activities.

✓ NUMBER FIVE

DELIVER HIGH PERFORMANCE FOR ALL BI AND ANALYTICS WORKLOADS

Analytics workloads are growing in size, variety, and complexity. Self-service business users and analysts accustomed to simpler BI reporting and descriptive analytics are pushing into predictive and operational analytics to explore what could happen and apply insights that have an immediate impact on operations. At the other end of the spectrum, data scientists are using various techniques and models to predict and prescribe what actions to take to achieve beneficial outcomes and are developing machine learning and other types of algorithms to operate at scale and with autonomy.

Preparing and provisioning data for this range of analytics can be daunting, time-consuming, and frustrating when organizations are pushing for more real-time analytics or want to move a large number of analytics projects into production faster using DevOps methodologies. Data virtualization, however, can shorten the time it takes to do so because this approach avoids replicating data and creating intermediate physical data stores (such as data marts, data warehouses, and even data lakes) which, although bypassing transformation stages, still require replication and loading of massive quantities of data.

Where creation of these physical, historical data stores is necessary, data virtualization can complement them by enabling organizations to provide users, analysts, and data scientists with a single point for more dynamic, real-time access to heterogeneous sources.

Data virtualization solutions can take advantage of several technology capabilities to achieve high performance for critical analytics workloads. Three of the most important are:

- **Query optimization.** Data virtualization solutions can apply optimization techniques automatically to accelerate queries. One method used by some solutions is partial aggregation to require fewer returned rows of data and thereby minimize network traffic. Another technique is to push down processing to the data sources to make best use of their performance strength, such as when they are built with MPP architecture.
- **In-memory grid.** Some data virtualization systems pair partial aggregation with in-memory MPP for faster post-processing. The grid can be connected to the data virtualization system through a high-speed network to be available when the system needs to post-process large volumes of data.
- **Performance monitoring.** Some data virtualization systems enable organizations to monitor and prioritize workloads to ensure the most important ones perform optimally. Organizations can use workload monitoring capabilities to spot where problems are occurring so that they can remedy them for key BI and analytics projects.

☑ NUMBER SIX

DEPLOY DATA VIRTUALIZATION TO SUPPORT FLEXIBLE ANALYTICS APPLICATION DEVELOPMENT

Data-driven organizations want to analyze data generated everywhere: on mobile and edge devices, through their own and partners' channels, and through traditional applications and data systems. This means that application development must be both more diverse and more connected; developers need to employ standard, RESTful APIs to more easily connect services and components and enable data to flow between them. Data virtualization fits with these development trends. Solutions can provide users with real-time logical views of data from multiple sources that could be implementing multiple formats, from traditional ODBC and JDBC to emerging open data services standards such as JSON.

Having the flexibility of a unified data services layer can be critical to dynamic development of applications and data services. Developers can create applications, such as for marketing departments, that let users find the right data and test analytics models across multiple sources through one interface.

Alternatively, organizations can use data virtualization layers to set up virtual sandboxes specifically for testing models and other programs so that production operational systems are not impacted. The sandboxes can be set up to offer single, virtual views of all data about customers, inventories, or other subjects. Some data virtualization solutions support use of application development life cycle tools and apply best practices for version control, governance, security, and other requirements.

Today's data management strategies must match current trends in application development toward agile methods, DevOps, and componentization using microservices. Data virtualization can provide developers and data scientists with the flexibility to shape data management around use of these methods and practices. Data virtualization layers can interface with a variety of data output formats for applications and be customized for specific visualization requirements.

Rather than offer one monolithic enterprise BI or data warehousing system that sets in stone how users will view and interact with result sets no matter how applications are configured, a data virtualization system can offer flexibility. Developers can then fit data consumption to users' preferences, including regarding their use of mobile devices.

Developers of data-driven applications need to integrate data models and identify data relationships across heterogeneous data sources. Data virtualization solutions can enable developers and users to visualize data models and relationships from within their chosen data modeling tools. Some data virtualization solutions can import predefined external data models; this is useful for organizations that want to incorporate reference models used by their enterprise BI and data warehousing systems as well as data models used by third-party packaged applications.

Organizations should evaluate data virtualization functionality to ensure that the solution enables developers and users to integrate data models. Such capabilities are critical for organizations that are seeking to break down data silos through better integration, rather than risk adding more silos as they build new applications and data services.

A FINAL WORD

This TDWI Checklist Report has discussed six areas where data virtualization can help organizations bring flexibility and agility to how they provision data and provide integrated views of data for business users, analysts, data scientists, and developers. By establishing a virtual data layer, organizations can bypass time-consuming and difficult processes for extracting, transforming, and loading data into a data warehouse, data lake, or enterprise hub.

A data virtualization layer can help organizations unify views of data and content from heterogeneous sources, on premises and in the cloud, while relieving data consumers from sorting out the technical complexities of getting data out of each source. In this way, the technology can help organizations develop and execute high-value analytics that draw on data from multiple sources and fit the demands of a variety of users and types of projects.

ABOUT OUR SPONSOR

denodo

Denodo is the leader in data virtualization providing agile, high-performance data integration and data abstraction across the broadest range of enterprise, cloud, big data, and unstructured data sources and real-time data services 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 needs for agile BI, big data analytics, Web and cloud integration, single-view applications, and enterprise data services.

Data virtualization, and the Denodo Platform in particular, marks a paradigm shift in the approach that organizations take toward accessing, integrating, and provisioning the data required to meet business requirements. Denodo Platform simplifies data access and makes critical business information available immediately to end users and consuming applications. Denodo's data virtualization innovations, including rich metadata and advanced capabilities in self-service, search, and data discovery, make the analytics journey compelling.

Denodo is well-funded, profitable, and privately held. For more information, visit www.denodo.com or call +1 877 556 2531 or +44 (0) 20 7869 8053.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

ABOUT THE AUTHOR

David Stodder is senior director of TDWI Research for business intelligence. He focuses on providing research-based insights and best practices for organizations implementing BI, analytics, data discovery, data visualization, performance management, and related technologies and methods and has been a thought leader in the field for over two decades. Previously, he headed up his own independent firm and served as vice president and research director with Ventana Research. He was the founding chief editor of *Intelligent Enterprise* where he also served as editorial director for nine years. You can reach him by email (dstodder@tdwi.org), on Twitter (@dbstodder), and on LinkedIn (linkedin.com/in/davidstodder).

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data management solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on business intelligence, data warehousing, and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence, data warehousing, and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

Transforming Data
With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org