Data Virtualization for Business Intelligence Systems

Revolutionizing Data Integration for Data Warehouses

MORGAN KAUFMANN

Rick F. van der Lans

13.5 The Future of Data Virtualization According to James Markarian, CTO of Informatica Corporation

Exponential growth in traditional transactional data plus new information from social media, call detail records, sensors and devices, and geo-location systems - has made it imperative that businesses take the lead in harnessing data to derive new insights for competitive advantage. To turn these new insights into business opportunities, both business and IT executives are being challenged to rethink their information management practices, break down organizational and data silos, and improve business/IT collaboration.

While many companies know that data is key to driving competitive advantage, they also know that they may not be as effective as they could be at transforming it into a useful asset. The data is there, but it's fragmented across systems - it's in the cloud, it's on the desktop, it's on mobile devices, it's in a variety of data sources. In fact, it's everywhere and it's duplicated again, and again. It's often not available when it's needed and when it is available, it might be full of errors, or does not deliver the value desired to improve operations and increase revenue. And, as data grows, the cost to manage it only increases.

This is particularly true in the case of business intelligence, the top priority for the CIO. According to Forrester Research, Inc.: "As data volumes and information complexity continue to skyrocket, traditional business intelligence tools try hard to keep up with ever increasing and changing demands. But it's an uphill battle - and business intelligence tools and applications do not always keep up the right level of pace and advancement. As a result, the rift between business requirements for on-demand information and real-time decisions, and business intelligence applications and IT support staff's ability to support them continues to grow."

According to the report:

- 66% of BI requirements change on between a daily and monthly basis
- 71% of the respondents said they have to ask data analysts to create custom reports for them
- 36% of custom report requests require a custom cube or data mart to answer the request
- 77% of respondents cited that it takes between days and months to get their BI requests fulfilled

It is clear – in order to turn data into new insights and thus opportunities, businesses need to pay greater attention to data integration to provide more actionable, trustworthy, and timely information in their BI projects and initiatives. They need to put data integration and business user-driven self-service to work for uncovering new insights that the business needs and can trust. Doing this will drive growth, reduce costs, and deliver innovations through spotting patterns and trends while meeting compliance and risk mandates more effectively. But how do you accomplish this?

How to Maximize Return on Data with Data Virtualization

Here are some critical factors that organizations must consider in order to maximize their return on data:

- They must be more business-focused and think from a perspective of an end-to-end business use of information and associated processes.
- They must employ self-service and business-IT collaboration best practices to enable the business to own their data, while IT retains control and governance.
- They must employ agile data integration techniques that the *cut wait and waste* in the process while complementing traditional approaches to BI.
- They need to define user-driven service-level agreements (SLAs) around data latency, accuracy, consistency, availability and uptime, and understand system implications.

- They must use data to complement their business strategy and technical infrastructure, instead of reinventing the entire wheel.

Of late, data virtualization has evolved as an agile data integration concept to enable more agile BI. However, traditional BI approaches including data integration, data warehousing, and other complex data processing initiatives are not going away anytime soon. This is because data will continue to be heterogeneous, dirty, and semantically different across systems. To succeed, data virtualization must co-exist, reuse and complement existing infrastructure and investments made to solve these problems rather than be a Band-Aid for a small subset of special use cases. It must also involve the business user early and often to ensure that the data is trustworthy.

Beyond Looking Under the Hood

The trick is to gain competitive advantage by accelerating the delivery of critical data and reports, and be able to trust and consume them instantly. But, data virtualization must be done right to support the critical success factors. Very often data virtualization borrows heavily from its data federation legacy. The primary use case data federation does well is to access and merge already cleaned and conformed data in real-time, leaving the heavy-lifting for other processing logic to make this possible. So, the time advantage gained is lost as one realizes the federated data had to be prepped for federation. As a result, the ROI simply disappears.

So, do go beyond looking under the hood and ask a few hard questions. To what extent does the solution support data transformation? Is it nominal, limited to what you can do programmatically through SQL or XQuery? Is there any data profiling or will you require staging and further processing? Is it profiling of both logic and sources, just sources, or neither? Is data cleansing and conforming simplistic, hand-coded, or non-existent? How about reuse? Can you quickly and easily reuse virtual views for any use case, including batch? To do data virtualization right, it requires a deep and thorough understanding of very complex problems that exist in the data integration domain.

So what's our perspective? Simply put, data virtualization must take a page from the world of virtual machines. Data virtualization must do the heavy lifting of accessing, profiling, cleansing, transforming and delivering federated data to and from any application, on-demand. It must handle all the underlying data complexity in order to provide conformed and trusted data, re-using the logic for either batch or real-time operation, whether through SQL, SOA, REST, JSON, or new acronyms yet to be specified. Data virtualization must be built ground-up on the *cut the wait and waste* best practices discussed in the book on *Lean Integration* by Schmidt and Lyle.

By starting with a logical data model, giving *business* and *IT* role-based visibility into the data early in the process, enabling data profiling on federated data to show and resolve the data integrity issues, applying advanced transformations including data quality in real-time to federated data, and completely and instantly reusing the integration logic or virtual views for batch or real-time, you can *cut the wait and waste* throughout the process. By leveraging optimizations, parallelism, pipelining, identity resolution and other complex transformational capabilities you can only find in a mature data integration platform, data virtualization can enable more agile business intelligence.

Finally, with enterprises generating huge volumes of data, the types of data changing enormously, and need for shorter data processing speeds, data virtualization can maximize the return on data. You can ensure that immediate action is taken on new insights derived from both *big data* and existing data. You can combine on-premise data with data in the cloud, on-demand. With the world becoming more mobile, you can provide access to disparate data by provisioning it to any device. Done right, data virtualization can give you the agile data integration foundation you need to embrace what we call secular megatrends - *social*, *mobile*, *cloud* and *big data*.

Read more from Rick van der Lans and Data Virtualization for Business Intelligence at Amazon.com.