Five Ways Data Virtualization Can Enhance Your Data Warehouse Investment

Prepared for Rocket Software by:

David Loshin
Knowledge Integrity, Inc.
February, 2015

www.knowledge-integrity.com
(301) 754-6350

Understanding the Challenges of the Analytics Architecture

The architecture of the venerable enterprise data warehouse, while deeply-rooted in the need for performance, reflects the design decisions made at the dawn of the age of reporting and analytics. In the mid-1990s, ensuring the performance of production transaction processing systems and maintaining sub-second response time remained the highest priority for the analytics architecture. And while the desire for reporting and analysis led to the creation of alternate data organizations for the data warehouse, the potential drain on computing resources motivated early designers to segregate the data on a separate platform, with its own specialized data models and applications.

This decision, wise at the time, has created an entire ecosystem of 'applicationware,' hardware dependencies, and skills requirements to support the objectives for reporting and analytics. As the speed and efficiency of computing resources has improved over time, though, the performance drivers have changed as well, exposing a different set of challenges that need to be considered and addressed. These challenges can be divided into three key areas:

- **Platform Challenges**, which aside from the physical system segregation includes physical limitations in data warehouse storage capacity, horizontal and extra-enterprise data dependencies, the existence of alternative architectures for reporting and analysis, and the need for data synchronization within narrowing time windows, all within the constraints of a decades-old design paradigm.
- **New/Emerging Opportunities**, associated with evolution of technology, data awareness, and the thirst for more powerful predictive and prescriptive analytics, such as discovery analytics (including interactive visualizations, event stream analytics, or collaborative interactions), the growing data distribution and diffusion as dependence on cloud computing grows, the role of the ubiquitous mobile devices and their rampant creation and injection of data, as well as the desire to capture and analyze Big Data.
- **Environmental Challenges** comprised of exploding data volumes, diversity of forms in which data is generated, the various speeds at which information is streamed, and a more mature demand that organizations provide a real-time comprehensive view of actionable information.

Enterprise Data Warehouse vs. Hadoop

These challenges lead many organizational architects to consider abandoning their enterprise data warehouse while seeking greener pastures (with correspondingly green technology). And in some camps, there is a perception that the emergence of Big Data (in general) and Hadoop (in particular) is sounding the death knell for the enterprise data warehouse as we know it. With organizations aching to adopt Hadoop, it may seem that these enterprises are prepared to abandon their decades-long investment in infrastructure, software, staffing, and development.

As a replacement platform, Hadoop (as well as other high performance NoSQL tools) can be used to simplify the acquisition and storage of diverse data sources, whether structured, semi-structured (web logs, sensor feeds), or unstructured (social media, image, video, audio). In addition, data

distribution and parallel processing can speed execution of algorithmic applications and analyses, and provide elastic augmentation to existing storage resources.

However, at the current level of system maturity Hadoop does not necessarily address our aforementioned challenges. While there is a promise of linear scalability, migrating reporting and analytics to a big data platform does not address data dependencies and synchronization requirements. Data sets will still need to be moved from their origination points to a separate analytics system. Re-platforming from an existing EDW to Hadoop may incur significant costs, especially in terms of reprogramming vast quantities of production-class SQL queries, end-user reporting tool configurations, and coded solutions for analytics.

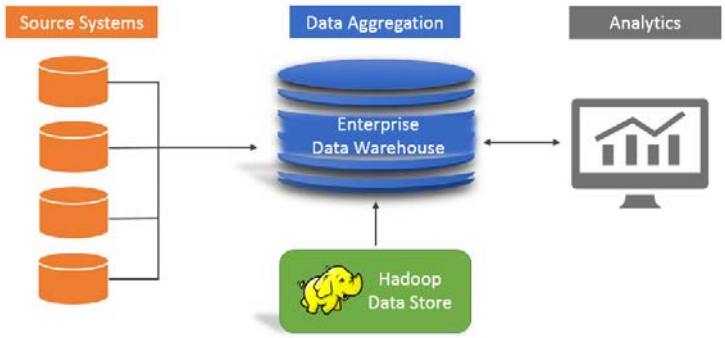

Figure 1: Hadoop and the EDW.

So despite the apparent (and justified) benefits of the growing capabilities of the different big data platforms, a more reasoned and responsible approach would blend consideration of new technologies like Hadoop with new options for extending the value of the existing information architecture investment. Consider that:

1) The production-hardened enterprise data warehouse in its various configuration still presents opportunities for significant value, especially in the context of the existence of tested queries and applications for accessing, organizing, and analyzing data.
2) The emergence of production-class data federation and data virtualization tools extends data accessibility across the enterprise without sacrificing the effort in development of existing reports and analyses. At the same time, optimizations, in-memory computing, and caching reduce the data latency that originally motivated system segregation. Not only does this reduce the need for additional staging areas and costly ETL, it also enables reporting and analysis to be more tightly-coupled to data sitting in its original source location diminishing the synchronization challenge.
3) Increasing mainframe utilization through data virtualization can amortize the per-user costs and prolong the lifetime of the EDW, as well as enhance the continued advantage of existing investments.

This raises the question: do you want to continue moving data from original sources and staging platforms to a segregated system, or do you want to examine ways of keeping the data sets where they are and redevelop around new services interfaces layered using data virtualization?

What is Data Virtualization?

The key to balancing the existing EDW's value while incrementally new analytics componentry is data virtualization. Data virtualization tools enable independently designed and deployed data structures to be leveraged together as a single source, in real time, and with limited (or often no) data movement. According to noted data virtualization expert Rick van der Lans, "*data virtualization is the technology that offers data consumers a unified, abstracted, and encapsulated view for querying and manipulating data stored in a heterogeneous set of data stores*."[1]

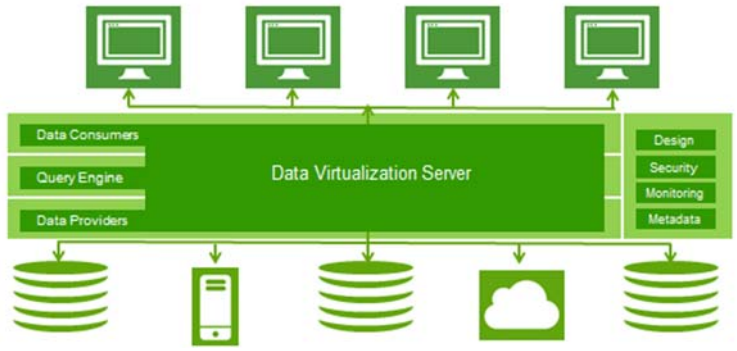

Figure 2: Data virtualization server

Data virtualization tools specifically adapted to mainframe environments (such as the z class IBM mainframes) use a special mainframe processing engine (one example being the IBM System z Integrated Information Processor, or zIIP) to handle data transformation and facilitating access to the data store on the mainframe.

Not only does this eliminate a significant amount of mainframe processing, but it also provides a low latency method to satisfy the data requests for downstream business intelligence and visualization tools. At the same time, the data virtualization methodology uses federation techniques to access data on external platforms (in internal relational database management systems, web/mobile data, data in the cloud, and with varying degrees of imposed structure) to create composite views of the information that is not in the data warehouse.

Data virtualization provides an abstracted view of organized data potentially drawn from heterogeneous sources, and using the right tools, can be deployed on mainframe's integrated processors as long as it:

- Provides support for SQL queries

[1] "van der Lans, Rick F., "Data Virtualization for Business Intelligence Systems," 2012 Morgan Kaufmann

- Does not impact OLTP response time
- Does not incur additional costs for storage, processing

Virtualizing a data warehouse deployed on a mainframe using a specialty processing engine, allows you to leave the mainframe data in place, avoiding the cost and complexity of data movement. The integrated processor uses the existing storage capacity of the mainframe, which reduces network bandwidth demand while providing real-time integration with transaction data. When the data virtualization tool can federate to big data storage environments like Hadoop/HDFS or NoSQL platforms, it enables programmers to use modern APIs such as MongoDB without demanding that the data be offloaded from the mainframe.

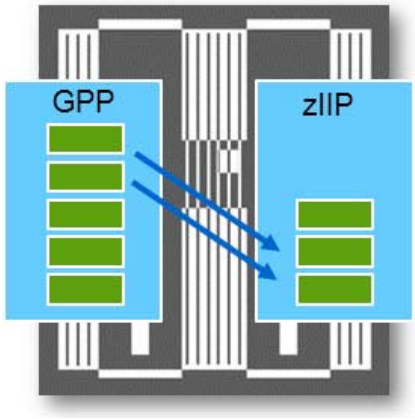

Figure 3: Mainframe using zIIP specialty engine.

Five Data Virtualization Use Cases for the Enterprise Data Warehouse

In this section we will discuss data virtualization use cases for enhancing the existing data warehouse environment, including (but not necessarily limited to):

1. **Storage augmentation and system federation** – By enabling a uniform method of accessing logical views of data sourced from different platforms in place (including Hadoop), data virtualization can help create composite views of data that are not persisted within the confines of the data warehouse.
2. **Streamlining extraction, transformation, and loading** – Data virtualization provides two complimentary benefits for loading data into the data warehouse. First, the way that federation enables access to other data sources reduces the need for bringing the data into the data warehouse before using the data to satisfy new business requests. Second, data virtualization reduces the hardware, software, and programming costs of data integration and loading by limiting network bandwidth contention, shrinking the costs of duplicated storage, and speeding execution time through the use of caches that use in-memory capabilities to eliminate data latency.

3. **Rapid prototyping for new development** – Enabling access to heterogeneous data sources using data virtualization accelerates assessment of data warehouse requirements yet streamlines integration without having to load data first. This facilitates rapid prototyping of reports and analyses and assess their respective business suitability prior to doing the work needed to extract, transform, and load the data first
4. **Increase breadth of data accessibility** – Under the right circumstances, data virtualization tools can enable access to both structured and unstructured data sources, as well as data in non-relational formats such as the various NoSQL data management schemas. This allows one to create composite representations of information that are not typically available in a relational data warehouse, as well as query a broad set of data sources in real-time.
5. **Substitution using virtual data marts** – When the enterprise data warehouse is unavailable (either for routine maintenance or because of unscheduled down time), accessing composite sources using data virtualization can function as a substitute for reporting and analytics until the EDW is back up and running.

There is a common theme that flows across all these use cases – leveraging data virtualization as a strategic tool for enabling, extending, or continuing accessibility to an enterprise data warehouse. Over time, these use cases demonstrate how data virtualization enables the eventual incorporation of a wide variety of data assets for analytics. In some cases, data virtualization can help make the case for streaming data into the mainframe-based EDW when it is more efficient than migrating to a new platform.

Summary: Augment Enterprise Data Warehouse with Data Virtualization

There is no doubt that the attraction of emerging data management paradigms such as NoSQL and Hadoop will prove to be a strong motivating factor in corporate reengineering and re-platforming data warehouses from their heritage environments. However, at the same time, it would be irresponsible to abandon the resources and time invested in developing production systems that are more than adequate to address a healthy proportion of today's business reporting and analytics needs. And as the vision for the future analytics environment takes shape, you will see that there is enough room for both emerging technologies and trusted heritage environments. The trick will be to balance the continued expanded use of the traditional systems with the design, development, and deployment of newer systems.

As we have seen, data virtualization provides a way to bridge these technologies. Data virtualization can help revitalize the data warehouse (particularly those involved with mainframe data) through a variety of hybrid approaches for data accessibility, thereby extending the useful life of existing platform investments. Data virtualization can be a key component of the strategy for continuing to extract value out of the years of underwritten costs. Adopting a strategy that retains the use of existing mainframe capabilities will preserve the investment in the development of SQL and code for reporting and analysis.

When it comes to organizations dependent on mainframe data, before disavowing the trusted data warehouse, consider some of these questions:

- What is your current volume of persisted data? Does that overburden the existing environment?
- What are your expectations for data volume growth?
- Has there been a significant investment in developing SQL queries for reports, ad hoc analyses, and other types of analytical applications?
- How much specialized code would have to be developed to replicate that functionality on a new environment such as Hadoop?
- Do you have measurable statistics for comparing the total costs of operation for both the existing and any proposed replacement environments?
- Are there ways of incrementally introducing new technologies like Hadoop for storage augmentation and pilot algorithmic analytics that dovetail with current mainframe reporting?

Each of these questions refers to dependences on existing production systems some of which deployed on a mainframe, with an expectation for incorporation of emerging tools for enhancement and growth. That suggests that an effective strategy for saving your data warehouse investments in the near-term and medium-term will incorporate a hybrid architecture combining the mainframe and data virtualization to provide the transition environment for the future of reporting and analytics.

About the Author

David Loshin, president of Knowledge Integrity, Inc, (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of data quality, master data management, and business intelligence. David is a prolific author regarding best practices for data management, business intelligence, and analytics, and has written numerous books and papers on these topics. Most recently, he is the author of "Big Data Analytics" (Morgan Kaufmann 2013). His books have been hailed as resources allowing readers to "gain an understanding of business intelligence, business management disciplines, data warehousing, and how all of the pieces work together." Visit http://dataqualitybook.com for more insights on data management .

David can be reached at loshin@knowledge-integrity.com.

About Rocket Software

Rocket Software is a leading global developer of software products that help corporations, government agencies and other organizations reach their technology and business goals. 1,100 Rocketeers on five continents are focused on building and delivering solutions for more than 10,000 customers and partners – and five million end users.

Rocket Data Virtualization enables mainframe relational and non-relational data to seamlessly integrate with Big Data, Analytics, and Web/Mobile initiatives; eliminating the need to move or replicate data, and with significantly reduced costs, complexity and risk.

- The industry's only mainframe-resident data virtualization solution for real-time, universal access to data, regardless of location or format.
- Support for Data Providers - IBM Big Insights, Hadoop, MongoDB, DB2, Oracle, SQL SQL Server, VSAM, IMS, Adabas, and others
- Support for Data Consumers – Cloud, Mobile, Analytics, Search, ETL, as well as ODBC, JDBC, REST, SOAP, JSON, HTTP, HTML, XML
- Reduced mainframe TCO -engineered to divert up to 99% of its integration related processing to the System z Integrated Information Processor (zIIP).
- Universal DB2 Support - applications using DB2 can now seamlessly integrate with any non-DB2 data source with the same ease of functionality
- Asymmetrical Request/Reply - any mainframe application (Batch, Started Task, IMS, IDMS, Natural) can interface with DV to request data for itself or another applications

Our customers tell us that IBM System z—the mainframe—is still the best platform in the world for running their critical business applications. And those applications generate and access large data volumes—big data. Increasingly, those applications and data must connect with other applications within the enterprise and even outside the enterprise. Rocket has deep domain expertise and world-class technology to keep the data where it belongs and move the analytics closer to the data.