Enabling Self-Service BI with a Logical Data Warehouse

Ravi Shankar

Ravi Shankar is the chief marketing officer at Denodo and is responsible for their global marketing efforts including product marketing, demand generation, communications, and partner marketing.
rshankar@denodo.com

ABSTRACT

Logical data warehousing enabled by data virtualization technology acts as a single holistic virtual repository for all enterprise data, providing a layer of abstraction between data sources and BI tools. This allows business users to access all-inclusive and up-to-the-minute data from source applications so they no longer have to depend on IT to build custom reports that span multiple systems. By leveraging a single source for data, business users can use any tool, or any number of tools, without creating data silos that impede enterprisewide insight.

This article introduces the logical data warehouse and covers its performance, security, governance, and potential impact on source systems.

INTRODUCTION

Today's enterprises want business users to be able to dive into the data and run analytics without having to submit a formal request to IT. One challenge facing the IT teams at these enterprises is that different departments within a company need different BI tools, and they need the tools to connect to different data sources.

For example, the marketing team might use Tableau to analyze the performance of marketing campaigns based on data from the CRM and marketing automation systems while the sales team might use QlikView to analyze sales performance using data drawn from POS and

demand-forecasting systems. Although many of these sources will be replicated to the data warehouse, some may not be. Siloed data sources prevent analysts from having an enterprisewide view of all available data.

Many companies are trying to ensure that all data is replicated to the data warehouse, but some data, such as unstructured social media data, is just not made for the traditional data warehouse. Some data stores are too large or too expensive to relocate into a data warehouse and so many companies are investing in cheaper storage systems such as Hadoop. Companies can set up ad hoc, point-to-point connections to multiple data sources as needed, but this method is not sustainable.

One solution to the proliferation of data sources is logical data warehousing. Built on the foundation of the traditional, physical data warehouse, logical data warehouses can provide access to all the data sources in an organization while leaving the source data exactly where it is.

DEFINING THE LOGICAL DATA WAREHOUSE

Gartner analyst Mark Beyer coined the phrase "logical data warehouse" in 2008 while discussing technologies of the future with a client. He shared the story on the Gartner blog (Adrian, 2011). Beyer describes how he suggested the phrase as a way to describe the future data warehouse. He held onto the idea and vetted it with clients and vendors before releasing research on its viability in 2011. His basic premise was that the logical data warehouse would be an extended version of the enterprise data warehouse, adding semantic data abstraction and distributed processing. As he explains in his post, the phrase "logical data warehouse" seemed accurate because "it focuses on the logic of information and not the mechanics that are used."

How do we decouple logic from mechanics? By virtualizing the data sources. For the last five years, vendors have been using data virtualization to establish logical data warehouses and today there are quite a few working examples in the field.

Unlike their traditional, physical counterparts, logical data warehouses do not contain any data.

Unlike their traditional, physical counterparts, logical data warehouses do not contain any data. Instead, they contain the intelligence and the logic for accessing the various sources, including the necessary security credentials. This access logic abstracts data consumers from the mechanical realities of where the data is stored. In a logical data warehouse, the data sources can be many and varied. They may include structured, semistructured, or unstructured databases; flat files; cloud-based storage repositories; and, of course, one or more physical data warehouses.

Rather than replicating data, the logical data warehouse creates views drawn from across the applicable sources, and most important, it does so in real time.

The logical data warehouse sits, therefore, in a layer between data sources and data consumers. In this architecture, as long as the data consumers operate through the logical data warehouse,

they will not create data silos and they are free to use whichever tool they wish. Each application accesses the same data, so if an analyst desires a view across the enterprise, the logical data warehouse presents no technical impediment.

KEY CAPABILITIES OF LOGICAL DATA WAREHOUSES

Capabilities including query optimization, caching, and resource throttling allow logical data warehouses to process very large volumes of data with subsecond response times. When these capabilities are implemented as described here, they will not adversely affect the source systems.

Query-optimization features based on data-source particularities. Some logical data warehouses can accommodate the query capabilities of each individual source. For example, certain source platforms (such as Oracle Exadata, SAP Hana) or data warehouse appliances can handle operational and analytical queries simultaneously with acceptable performance, whereas others would be burdened. Logical data warehouses should know the difference and only push down operations to systems that can support them without affecting performance.

Extensive caching capabilities. Logical data warehouses enable the cache to be on disk, in memory, or in elastic storage networks. The cache should be batch-loaded during off-peak hours and incrementally cached during the day to minimize impact. Such solutions also enable partial caching to reuse the results of recent queries or to cache frequently used data. They also provide the ability to cache the results of individual, costly queries as well as full caching of certain views; this prevents multiple queries on those views from hitting the data source at once.

Advanced scheduling capabilities. Logical data warehouses support hybrid scheduling that embraces real-time, cache, and batch execution strategies, adjusting the system's use of each mode on a second-by-second, case-by-case basis. Logical data warehouses also provide users with the option to schedule any view in batch mode at any time.

Resource throttling. Logical data warehouses perform data source throttling, which limits the number of concurrent requests sent from the data warehouse to individual data sources within specified time frames and thresholds. This is a critical capability because it enables system engineers to avoid performance problems in source systems during known periods of peak activity, such as new product releases or end-of-quarter expense reporting.

Dynamic policy configuration. To enable stakeholders to manage the impact on source systems, logical data warehouses allow users to set policies based on diverse factors including the particular view, query type, the role of the user executing the query, the time of day, and monitoring status. Through such policies, users can set a maximum number of concurrent queries for a particular application, prioritize certain applications, or specify that executives' queries should have a higher priority than those of administrative users.

Monitoring capability. This is critical for the logical data warehouse, as it enables practitioners to flexibly apply each of the above strategies, adjusting based on a real-time analysis of source-system impact. Logical data warehouses permit stakeholders to measure the response time for any query and assess

the effect of any applied resource throttling mechanism or caching strategy. The logical data warehouse can also maintain a historical view of usage, performance, and service levels to support ongoing optimization.

Logical data warehouses can be particularly effective in maintaining security across diverse systems because they sit between data sources and data consumers.

SECURITY AND THE LOGICAL DATA WAREHOUSE

Security and privacy are critical within a logical data warehouse environment because a large number of users across multiple departments with different access privileges will all have access to the same data sources. This might seem like a daunting problem, especially because each of the underlying data sources will have its own security model, which will vary in sophistication. Physical data warehouses tend to have mature security and privacy capabilities—especially when they are provided by established vendors such as Teradata, Netezza, and Microsoft—whereas newer data technologies, such as Hadoop and Spark, tend to have relatively immature security models. HDFS, for example, has an "all or nothing" policy; if you can access some of the data, you can read all of it.

However, this problem does not hamper logical data warehouses. In fact, logical data warehouses can be particularly effective in maintaining security across diverse systems because they sit in a layer between data sources and data consumers. Within this layer, logical data warehouses provide security abstraction and normalization capabilities, which impose a strong, unified security model on all the underlying data sources. Logical data warehouses can provide row and column security as well as masking at the value level—regardless of the data source. In a logical data warehouse, data stored in HDFS can be protected with the same level of security as data stored in a best-of-breed physical data warehouse.

Logical data warehouses can perform role-based authentication at the guest, employee, and corporate level with data-specific permissions that include row- and column-level masking, schema-wide permissions, and policy-based security. They also match user identities via an LDAP Active Directory and encrypt cache or swap data at rest. The logical data warehouse encrypts incoming source data in motion using SSL/TLS protocols and authenticates users through pass-through Kerberos, Windows SSO, OAuth, or SPNEGO authentication. When sending that data on to consumers, the logical data warehouse secures the data in motion using the same SSL/TLS protocols and manages authentication via standard JDBC/ODBC security, Kerberos, Windows SSO, or Web Service Security.

DATA GOVERNANCE AND THE LOGICAL DATA WAREHOUSE

Just as logical data warehouses are well suited to securing diverse sources because they provide a unified framework for security management, they are similarly well equipped to provide a framework for establishing data governance.

Although some security features also have a direct bearing on data governance, such as masking data from those without sufficient privileges, data governance requires a few additional considerations.

Because they centrally manage access to the various sources, logical data warehouses are able to record the lineage of the data from source to target. This helps companies know where their data comes from, which is a key part of establishing its veracity for data stewards, consumers, and executive sponsors. With a detailed data lineage, stakeholders can trace the origin of every transformation for every piece of data in a model.

PERFORMANCE OF LOGICAL DATA WAREHOUSES

In terms of performance, how do logical data warehouses compare to traditional physical data warehouses? Of course, the answer will vary from one vendor's solution to another, but in at least one case, performance was found to be comparable (Denodo, 2017). Using queries drawn from the standard TPC-DS benchmarking test for decision support systems, the performance of a Netezza data warehouse was compared against the performance of a proprietary logical data warehouse architecture that federated an Oracle database and a SQL Server database along with the Netezza data warehouse.

It was found that for typical reporting and analytical queries involving aggregation operations over very large data sets, the logical data warehouse with virtualized access to data across different sources came in almost at the same speed as the physical data warehouse that housed all of the data within it. For a query of "total sales by customer," returning 1.99 million rows, the Netezza data warehouse took 20.9 seconds, and the logical data warehouse solution took a half-second longer at 21.4 seconds.

THE LOGICAL DATA WAREHOUSE AND BI SELF-SERVICE

By establishing a layer that enables real-time, secure access to trusted data across a myriad of heterogeneous sources, the logical data warehouse provides a foundation for self-service BI by offering both basic self-service functionality and a tool-agnostic platform.

By federating the data into a single, virtual repository, logical data warehouses make it possible for BI consumers to perform simple "Google-like" searches across all the data in the enterprise. Whether data is stored on a server in a physical data warehouse or in the cloud as part of an HDFS system, the logical data warehouse enables users to browse the relationships between data entities, check the data lineage, and build queries.

On its own, a logical data warehouse makes it possible for stakeholders to use whichever BI or visualization tools they wish without creating data silos or needing constant IT support. As long as each tool points to the logical data warehouse, each tool will have access to the same data (depending on access privileges), whether it is a narrow data set intended for a specific department or a view across the entire enterprise.

With a logical data warehouse, companies can also create "logical data marts," which are views on top of the logical data warehouse. Logical data marts, just like the logical data warehouse, are very easy and fast to create. The beauty of logical data marts is that business users can create them without depending on IT, unlike

a physical data mart, which requires technical help to set up. The logical data marts are easy to maintain and update, and they provide up-to-the-minute views of the data without any lag time, unlike the physical data marts, which are refreshed only periodically.

The beauty of logical data marts is that business users can create them without depending on IT.

Vendors are continually evolving their support for logical data warehouses. Cisco Information Server and the Denodo Platform both push queries down to individual data sources to minimize the number of results returned; the Denodo Platform even does this dynamically. This includes variations such as full aggregation pushdown, where the complete query is pushed down to a source; and partial aggregation push-down, where a subset of the query is pushed down to the source and the rest is executed within the logical data warehouse.

Support is also being added for applications of the logical data warehouse architecture to different use cases. Such use cases include bridging a traditional data warehouse with a Hadoop system that contains historical data offloaded from the data warehouse to leverage Hadoop's lower cost or integrating master data from master data management systems with related transactions in the data warehouse.

CASE STUDY: A LOGICAL DATA WAREHOUSE FOR SELF-SERVICE BI IN ACTION

Seacoast Banking Corporation of Florida is one of the largest community banks headquartered in Florida; it provides commercial and retail banking, wealth management, and mortgage services.

Seacoast's internal users from core banking groups such as loans, deposits, and business internet banking had to request custom static reports from the IT team for operational purposes. However, a large amount of the operational data resided in a hosted data warehousing platform that made generating these reports extremely inefficient. For example, creating analytical reports such as trend analyses used to take two-to-three days.

Seacoast wanted to provide its business users self-service capabilities to interact directly with the data so they could create their own custom reports, based on the company's changing needs. The bank decided to unify its data assets by moving away from its hosted physical data warehouse to a logical data warehouse. Such an architecture would also provide them with the agility to make changes to the data model as newly acquired banks and systems were integrated with those of Seacoast.

Using the Denodo Platform, Seacoast now integrates its six separate data sources from a mix of cloud and on-premises sources. The aggregated data is then exposed as views or virtual data marts to BI systems for dashboarding, reporting, and analysis for loans, online business and consumer banking, and other purposes.

Running analytics on a single logical layer provided more flexibility to the firm than it had

thought possible. Now, business users have the flexibility to use analytics and reporting tools to administer credit, mitigate risk, and manage internal operations, all in almost real time. The interactive self-service reports now take less than two hours to create instead of the two days it took to build the previous static reports (Vizard, 2017).

CONCLUSION

Companies looking to implement self-service BI would do well to consider a logical data warehouse. It enhances the capabilities of traditional data warehouses, federates data from myriad heterogeneous sources, provides this data to BI tools in real time, and allows BI stakeholders to use any BI tool they wish without creating data silos.

Logical data warehousing also has the potential to save costs. According to Gartner, organizations that use data virtualization will spend 40 percent less on building and managing data integration processes for connecting distributed data assets. Logical data warehousing has enormous potential for improving the efficiency of data access—the necessary first step in improving the efficiency of business. ●

REFERENCES

Adrian, Merv [2011]. "Mark Beyer, Father of the Logical Data Warehouse, Guest Post," http://blogs.gartner.com/merv-adrian/2011/11/03/mark-beyer-father-of-the-logical-data-warehouse-guest-post/

Denodo Technologies [2017]. "Data Virtualization for Logical Data Warehouse," Architect to Architect Ebook Series. http://www.denodo.com/en/document/e-book/architect-architect-ebook-series-data-virtualization-logical-data-warehouse

Vizard, Mike [2017]. "Seacoast Bank Cashes in on Data Virtualization," January, http://www.baselinemag.com/virtualization/seacoast-bank-cashes-in-on-data-virtualization.html