

Data Warehouse Appliances: Lessons from the Trenches

By Jerry Locke & Steve Dine

June 4, 2009

Over the past few years data warehouse (DW) appliances have become a viable option for business intelligence (BI) programs looking to load and process large data volumes as well as reducing query execution times. However, like most architectural decisions, there are many factors to consider when choosing to implement a DW appliance that aren't usually highlighted during the sales presentation. This article will present some observations and lessons learned from a recent project in which we implemented a DW appliance for both the data warehouse and the data mart layers of the client architecture.

What is a DW Appliance?

First let's review some of the characteristics of a DW appliance. DW appliances often leverage integrated hardware with a high-speed back-plane, embedded operating system, scalable storage and analytic database technology to deliver a compelling price to performance solution. In some cases, vendors leverage commodity off-the-shelf while others design their hardware specifically for their requirements. Most utilize a massively parallel processing (MPP) database architecture to divide the work across computing nodes with their own processor, memory and disk. What many refer to as shared nothing architecture. Many also usually include a proprietary bulk loader and data compression algorithms. The one common characteristic across all appliances is that they are designed to handle large scale query processing.

Why a DW Appliance?

The impetus for considering a DW appliance for our recent project was that the client had a limited time window to load large quantities of data, were routinely querying large data volumes that often required full table scans and multi-pass queries, and had limited database administration resources. Prior to the decision to move to a new database platform, they worked with their existing vendor to reduce load times and tune existing queries. Unfortunately, after many attempts, the vendor was unable to meet their requirements. They also had budgetary and resource constraints that prevented them from turning to the more traditional MPP database vendors.

Lessons Learned

The decision to move to a new database platform was not taken lightly. DW appliances are still relatively new compared to traditional relational database management systems (RDBMS), such as Oracle and Teradata. Many appliances also rely on integrated hardware which adds another level of complexity to the solution. However, given these risks, and much research, our client felt that there were few alternatives available. They also believed that the potential reward outweighed the risk to the project timeline and scope of delivery. Ultimately, the project proved to be a success but there were many lessons learned along the way. When considering a DW appliance we recommend that you consider the following:

1. Mixed Query Loads

It is important to keep in mind that there are certain types of queries that each of the appliance vendors excel at processing, especially those that involve full table scans. These are usually the queries that they target during a proof-of-concept (POC) against a traditional

RDMBS. While DW appliance vendors are starting to add functionality to support varying types of queries, such as prioritization, scheduling, short query bias and materialized views, many are still limited in these capabilities. When evaluating appliance vendors make sure to include different types of queries, not simply full-table scan based sets.

2. *Availability of Native Database & ETL Connectors*

It takes time and expense to build native connectors for different ETL tools, applications and databases. In many cases, Open Database Connectivity (ODBC) drivers have comparable performance to native drivers, but it depends on the driver, as not all ODBC drivers are created equal, and the action being performed. In our case, the vendor did not have a native driver for the ETL tool and the load performance was dismal. During the POC, the appliance vendor used their bulk loader and the client's ETL tool was never used. On other projects, we have run into cases where the available ODBC driver did not support certain database functions. Make sure to ask your vendor about which native drivers are available and make sure that they are used during the POC.

3. *SQL (Structured Query Language) Differences:*

The structure query language (SQL) utilized by different database vendors are similar but not the same. While most database vendors follow American National Standards Institute (ANSI) for SQL, they are often on different versions of the standard. They also implement additional commands and functions that are unique to their specific software. This also applies to the databases that are embedded in the DW appliance. Be aware that there will likely be differences in SQL between the previous database, as well as the new database and project team resources will need time to adjust to the new commands.

4. *Existing tool set interoperability:*

Many organizations already have a significant investment in data modeling, metadata and reporting tools. However, these applications do not support an infinite variety of software. Many of the data warehouse applications use open source databases like MySQL and PostgreSQL. Make sure that your tools work with the software on the DW appliance. In our case, the data modeling tool did not create PostgreSQL DDL nor could it reverse engineer a schema from the database. Also, BI tools will often support extended SQL commands and analytic functions, but only for a subset of database vendors. Check to see that your BI vendor supports the new database.

5. *Stored Procedures:*

Many existing DW architectures leverage database stored procedures for ETL processing and/or BI reporting. The DW appliance database is unlikely to support the store procedure language from your existing vendor, and may not have one at all. This translates into a potentially more complex and longer implementation if you are migrating from existing data warehouse. It can also throw some roadblocks into an implementation if you are leveraging extraction, loading and transform (ELT) and reliant on the database for complex processing.

6. *Hardware versus Software Appliances:*

Some of the DW appliance vendors provide the software stack that can run on a variety of off-the-shelf hardware. While this can be beneficial from both initial investment and replacement cost standpoints, it does segregate the support responsibilities across multiple vendors. In our case, when we ran into performance issues and unexpected behavior, the vendors simply pointed fingers at each other. The software provider wanted the hardware vendor to perform diagnostics and vice-versa. Before venturing into this scenario, make sure a primary vendor is established and have that vendor be responsible for the service level agreements.

Conclusion

It's likely that DW appliances will continue to gain market share in the BI industry. They can solve large database challenges with a good price-to-performance ratio, reduced maintenance and high performance to address business analytics on growing data volumes. However, with all of these benefits, it is important to be mindful of the possible issues with bring in this technology. Any new application will inherently bring on risk. We hope our experiences mentioned above will mitigate the risks of implementing your DW appliance.

About the authors:

Jerry Locke is a principle consultant with Datasource Consulting, LLC. He is an experienced data warehouse architect, ETL developer and analyst with more than 11 years in designing, implementing and managing enterprise Business Intelligence Software. His ETL experience spans data modeling, process design and development, data analysis and data cleansing. He also has extensive experience integrating data to and from Oracle, DB2, SQL Server, OLAP and Informix databases. Jerry earned his bachelor's degree from Arizona State University.

Steve Dine is the President and founder of Datasource Consulting, LLC. He has more than 12 years of hands-on experience delivering and managing successful, highly scalable and maintainable data integration and business intelligence solutions. Steve is a faculty member at The Data Warehouse Institute and a judge for the Annual TDWI Best Practices Awards. He is the former director of global data warehousing for a major durable medical equipment manufacturer and former BI practice director for an established Denver based consulting company. Steve earned his bachelor's degree from the University of Vermont and a MBA from the University of Colorado at Boulder.