



## **Increasing Enterprise Data Warehouse Performance, Longevity and ROI with the Vertica<sup>®</sup> Analytic Database**

*How Vertica complements and enhances your existing  
Enterprise Data Warehouse*

## Executive Summary

Considering the strategic importance of Enterprise Data Warehouses (EDWs) and the millions of dollars spent creating and maintaining them annually, it is imperative for organizations to easily gain access to the wealth of information retained within them. However, many EDWs become victims of their own success. Over time, satisfied users demand answers to new questions resulting in an increase in the volume and variety of data EDWs contain, and the number and complexity of queries executed against them. If the EDW becomes unable to cope with the requests placed upon it, and service level agreements cannot be met, companies often incur expensive re-design or DBMS and hardware migration projects, which can take many months.

*“Never make the assumption that your current EDW is not well designed enough to handle the workload. It may be that moving part of the workload somewhere else is going to save you from all of the expense of upgrading it.”*

Don Feinberg  
VP Research  
Gartner

This paper describes how by offloading certain analytic workloads from an existing EDW, based on popular solutions, like Oracle, DB2 and Teradata to a Vertica<sup>®</sup> Analytic Database, it is possible to increase the performance and longevity of the EDW and satisfy the ever-growing requests for information from the business.

## Evolution of the Enterprise Data Warehouse

Today EDWs are commonplace and a necessity if an organization is to remain competitive and successful. Originally they grew from the relational databases like Oracle and DB2, which at the time were designed more for OLTP systems, than for analyzing data. Without the data warehousing features that we rely upon today, such as partitioning, materialized views, bit-mapped indexes etc, the original data warehouses soon couldn't cope with the queries they had to answer. Consequently the users, unwilling to accept this level of performance decided to build their own data warehouses comprising only the data they were interested in, and the *data mart* was born.

Using data from existing data warehouses, and adding their own data to the data mart, users were delighted at being able to gain access to their information quickly and adjust it to their specific requirements.

Figure 1 – Data Warehouse and Data Marts



However, for the DBAs managing the data warehouse, it became a nightmare, as illustrated in Figure 1. Suddenly there were tens or even hundreds of copies of the data everywhere, with no control or checks being made on the quality of the data residing in these data marts.

With users adding their own data to the data marts, there was no longer a central definitive source of information within the enterprise and this had to stop. It did, and the Enterprise Data Warehouse was born. Slowly, organizations returned to centralizing their data within the EDW.

## The Pressures Today

Initially, centralizing all data into the Enterprise Data Warehouses worked extremely well. Information was now controlled in one central location; it was clean, accurate and available for queries. Users, realizing the wealth of information now available to them, started to exploit it for business opportunity. Which are the best performing retail stores? Who are the clients we should focus marketing budget on? Which areas have the highest costs? These questions and many more could be answered by the EDW.

The questions kept coming and the demands on the EDW increased and the volume of data contained within EDWs became massive, reaching very many terabytes. Organizations would purchase additional server hardware to cope with the demand. They would also employ as many data warehouse features as possible within the database to improve efficiency and performance, such as partitioning and aggregated views. Although disks have large storage capacity, they still were not that fast and many EDW's became very disk I/O bound. It wasn't that a query couldn't be answered, it was, but users were expecting a response in seconds, not 30 minutes or hours. Therefore we need to look at new optimization techniques.

## Optimizing the Enterprise Data Warehouse

A considerable investment has been made in your EDW and it would be wrong to think that the only solution to your current inability to meet your SLA's is to throw away the EDW and start again. Instead, what is required is to review your EDW workload to determine where the bottlenecks are, and typical ones are:

- Caused by specific applications running
- One of more users generating a very high load
- Excessive I/O demands caused by queries

Once you understand what is degrading performance in the EDW, it's time to look at what would be the most appropriate optimization technique to overcome this problem. Instead of the traditional approach of throwing hardware or applying software features at the problem, it is time to consider a new optimization technique of incorporating the Vertica Analytic Database into the EDW.

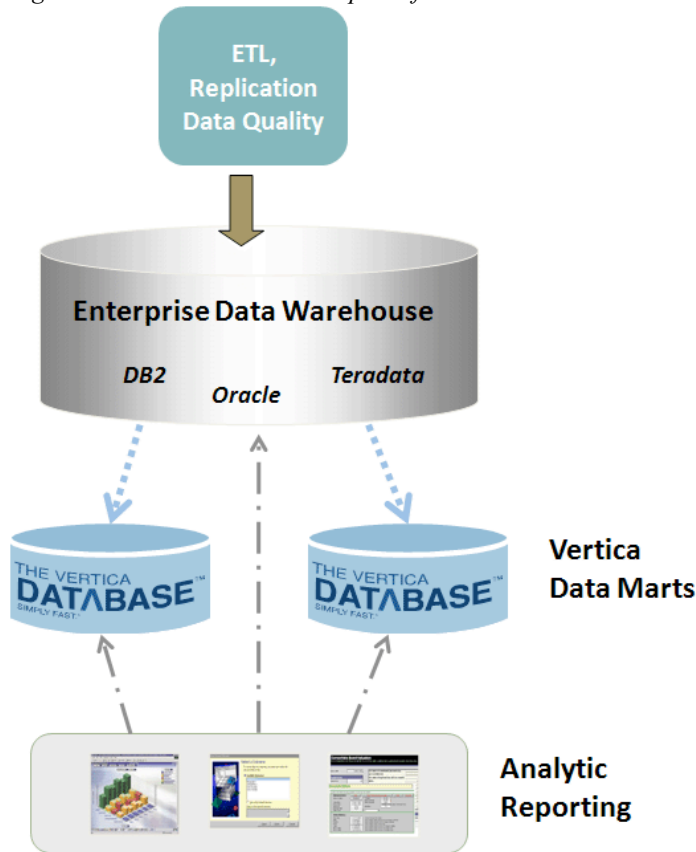
Now it's understandable that given the history of data warehousing, your first thought is that using Vertica within the EDW, is simply creating a data mart, and we all know the problems that has brought in the past.

Today, there is a big difference between the data marts described earlier in Figure 1, and the ones that you will be creating now. The Vertica data marts should be viewed as an optimization technique for the EDW, where it is sourced entirely from the EDW, as illustrated in figure 2.

*“Put data marts behind the EDW as an optimization technique. If I can move heavy analytic workloads off of the EDW on to a data mart, I can release the EDW to retain its performance on the everyday work it does and keep all the users happy.”*

Don Feinberg  
VP Research  
Gartner

Figure 2 - Vertica data mart as part of the EDW



Since these data marts can only be sourced from the EDW, they now sit *behind* the EDW.

Using this approach, those users or applications which make unacceptable demands on the EDW are asked to query the Vertica data mart, while existing users and applications are unaffected.

The investment in your existing EDW is retained, and those users or applications experiencing performance problems will now query the Vertica database, which satisfies their fast analytic processing requirements. By removing these burdensome workloads from the EDW, its overall performance increases, and it becomes better able to support new demands from the business

without costly re-designs or hardware or DBMS upgrades.

### Why Column Based Data Marts Make Economic Sense

Looking at this optimized solution raises the question, why is it necessary to use the Vertica Analytic Database, won't my existing DBMS suffice? It may, but the purpose of the data mart is to remove the very demanding workloads off the EDW onto a database system which is highly efficient. Vertica is optimized for an analytic environment, which is what is required for these query-intensive workloads. As a result, it is very easy and inexpensive to deploy (on off-the-shelf servers), requires very little storage (due to aggressive use of data compression) and answers queries 50x-200x faster than other databases with little-to-no tuning. The architectural features that enable this in Vertica include:

- Continuous Performance
- Column Storage
- Compression
- Clustering

## Continuous Performance with Column Storage

Vertica reduces DBMS disk I/O by only reading the columns referenced by the query, therefore it is not uncommon to see query response times 50 to over 200 times faster. This is achieved in Vertica because in a row-oriented DBMS, row values are stored contiguously, and to process a query, the DBMS must read data from every row and column (even columns not specified in the query). Bitmap indices, data cubes, materialized views, and other database exotica help, but for many databases, managing data structures to optimize performance for every query and use is prohibitively complex, and these data structures often impose dramatic storage space and update performance overhead. Vertica stores the values for each column contiguously meaning that it only needs to read the columns being queried. This approach dramatically improves query performance by eliminating unnecessary disk and memory I/O.

## Compression

CPUs are getting faster at a much greater rate than disk bandwidth is increasing, so Vertica replaces slower disk I/O with faster CPU cycles to encode data elements into a more compact form and query them. Vertica's innovative query engine operates directly on compressed data, meaning that it can actually require fewer CPU operations to process the compressed version of a table.

It is not uncommon to achieve compression rates of 50-90%. With the restrictions on the data center of power and floor constraints, using Vertica, you can store the same amount of data using considerably fewer resources. Therefore you don't have to worry about creating a data mart that requires vast amounts of physical storage and places additional demands on the data center.

## Clustering

Vertica can be deployed on inexpensive grid hardware, therefore there is only a small additional cost in using this optimization technique, despite queries being answered considerable faster than your existing EDW. If at any time, additional power is required, new servers can be easily added and Vertica automatically takes advantage of them without the need for expensive and time-consuming upgrades.

Vertica performance scales linearly to a large number of CPUs. Because Vertica does not depend on custom hardware, it will benefit over time from the dramatic performance improvements that commodity hardware has traditionally enjoyed. It is much harder for custom hardware solutions to reap these same benefits. Furthermore, because it is designed from the ground up for modern commodity hardware and runs on standard Linux (including kernel-based I/O and file systems), it will maintain compatibility and scalability over time.

## Loading the Data Mart

To avoid the issues seen previously with data marts, Vertica must be loaded using data from the EDW. This ensures that a central definitive source for data remains and that all copies of the data are identical.

Transferring information into the Vertica data mart can be achieved using many of the popular ETL tools. Vertica is capable of loading gigabytes of data per minute—in bulk, or on a 24x7 basis in “trickle” mode for real-time analysis, therefore, data can be placed in the data mart very quickly.

### Vertica Advantage

Although time is required to build and maintain the data mart, the customer example below shows the significant benefits that can be achieved from this optimization approach of using Vertica, rather than row-oriented OLTP databases for these analytic workloads. Here, queries against 1.5 terabytes of stock market trade and quote history are 270X faster when Vertica queries the data, and it is now possible to produce 33 times more reports, and data is available in real-time rather than being a day late. All of this can be achieved at a cost of 1/28<sup>th</sup> of the hardware being used by the EDW. Similar benefits can be reviewed for other industries and applications by visiting [www.vertica.com/benchmarks](http://www.vertica.com/benchmarks).

	Row-Oriented Data Warehouse DBMS	Vertica <sup>®</sup> Analytic Database	Vertica Advantage
Avg Query Response Time	37 minutes	9 seconds	270x faster answers
Reports per Day	30	1000	33x more reports
New Market Data Availability	Next day	1 minute	Real-time views
Hardware Cost	\$1.4M (2x 6 servers + SAN)	\$50,000 (6 HP ProLiant servers)	1/28 <sup>th</sup> of the hardware, Built-in Redundancy

With all data still residing in the EDW, it remains the definitive place where all changes to data are made. However, copying this data into a Vertica data mart, unleashes its full potential to the business.

### See for yourself...

The Vertica Analytic Database supports SQL and integrates with 3<sup>rd</sup>-party ETL, analytic and BI reporting tools and applications via JDBC, ODBC and specific language bindings. Therefore, using all your existing SQL knowledge and technology, a Vertica database can be very quickly created and loaded with data.

If you would like to learn more about the Vertica Database or if you would like to evaluate it yourself, then visit the following links:

<b>Gartner on Vertica and EDWs</b>	<a href="http://www.vertica.com/gartner">www.vertica.com/gartner</a>	Watch a recording of Don Feinberg of Gartner explain why supplementing an EDW with column-DB based data marts improves the ROI on EDWs
<b>Vertica Resource Library</b>	<a href="http://www.vertica.com/resourcelibrary">www.vertica.com/resourcelibrary</a>	White papers, demos, webcasts, system requirements
<b>Vertica Benchmarks</b>	<a href="http://www.vertica.com/benchmarks">www.vertica.com/benchmarks</a>	See customer-submitted cost and performance comparisons between Vertica and other databases
<b>Vertica for the Cloud</b>	<a href="http://www.vertica.com/cloud">www.vertica.com/cloud</a>	Get a Vertica database instance provisioned instantly on the Amazon Cloud and use it on a month-to-month basis
<b>Vertica Customers</b>	<a href="http://www.vertica.com/customers">www.vertica.com/customers</a>	See who's using Vertica
<b>Request a Vertica Evaluation</b>	<a href="http://www.vertica.com/download">www.vertica.com/download</a>	Request a free evaluation copy of the Vertica Analytic Database to download and install

## About Vertica Systems

Vertica Systems is the market innovator for high-performance analytic database management systems that run on industry-standard hardware. Co-founded by database pioneer Dr. Michael Stonebraker, Vertica has developed grid-based, column-oriented analytic database technology that lets companies of any size store and query very large databases orders of magnitude faster and more affordably than other solutions. The Vertica Analytic Database's unmatched speed, scalability, flexibility and ease of use helps customers like JP Morgan Chase, Verizon, Mozilla, Comcast, Level 3 Communications and Vonage capitalize on business opportunities in real time. For more information, visit the company's Web site at <http://www.vertica.com>.