

TECHNOLOGY AUDIT

Vertica Analytic Database









Vertica Systems


OVUM BUTLER GROUP VIEW

ABSTRACT

Vertica is one of a growing band of vendors offering column-based analytic databases for data warehousing. Its Analytic Database is designed specifically for storing and querying large datasets. Vertica's differentiator is that it combines a columnar database engine with massively parallel processing (MPP) and shared-nothing architecture, aggressive data-compression rates, and high availability. Column-based databases can be slow when it comes to deleting and updating data, and Vertica addresses this by taking advantage of a hybrid store that handles write, update, and insert operations. The store also makes the data available for queries in-memory. In addition, the product benefits from smart capabilities for data access and disk I/O. Ovum believes that Vertica's technology could be applicable across a range of markets among companies that have a mix of analytical requirements.

KEY FINDINGS

- | | |
|--|---|
|  Column-oriented database engine augmented by row-based in-memory capabilities. |  Distributes data across a grid in specialised data structures used in place of tables called "Projections". |
|  Massively parallel-processing capabilities with built-in failover and high availability. |  Offers high levels of compression and encoding and can execute on encoded data. |
|  The appliance option has no specific performance advantages over other deployment options. |  Does not fully support SQL extensions from leading vendors such as Oracle and Microsoft. |
|  Load-balancing is achieved through Linux's IPVS capability. |  As a young company Vertica does not have a big presence outside the US. |

Key:  Product Strength  Product Weakness  Point of Information

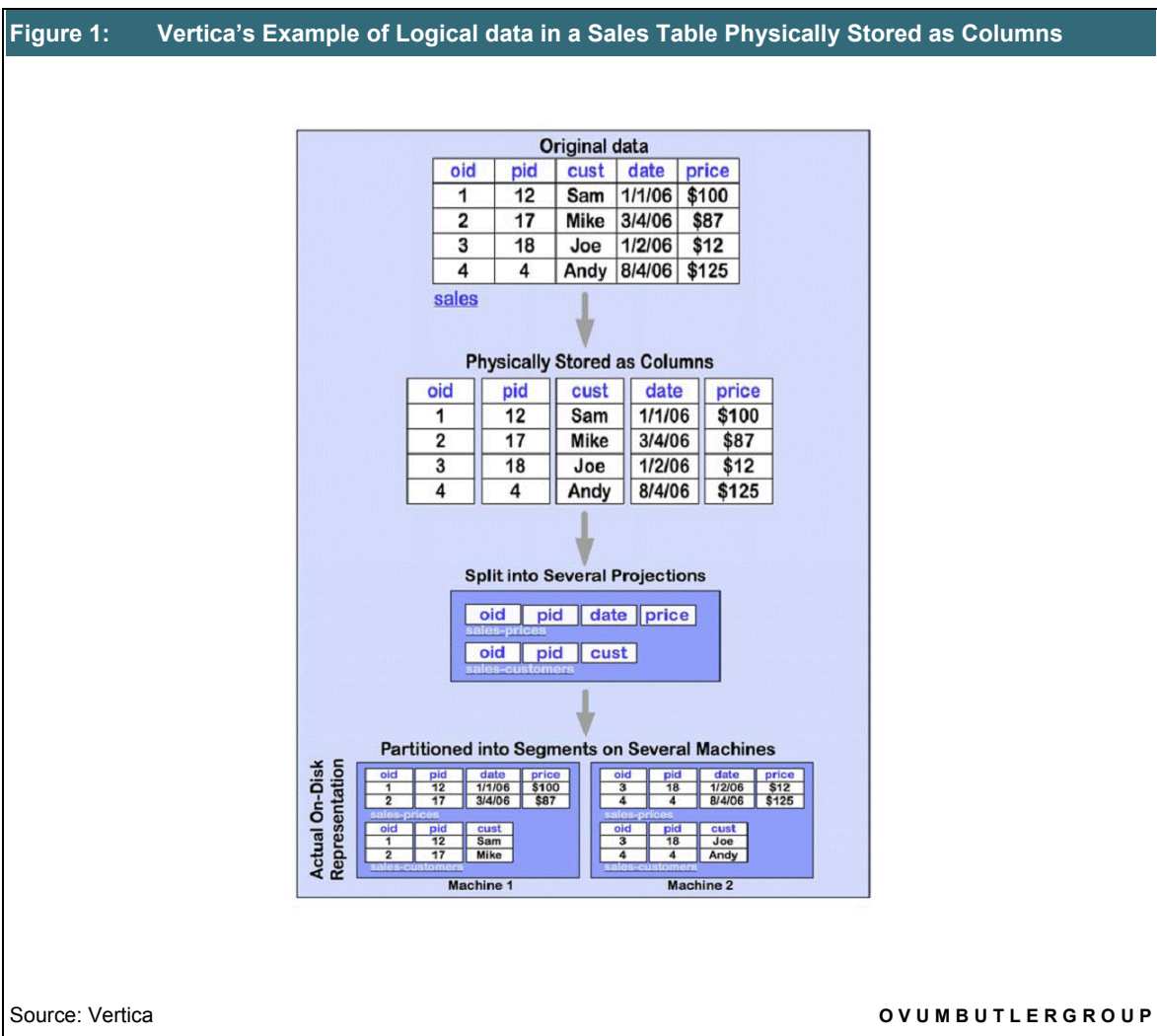
LOOK AHEAD

Planned enhancements this year will focus on enhancing the analytics capabilities of the product, its performance and scalability, and its internationalisation features.

FUNCTIONALITY

Product Analysis

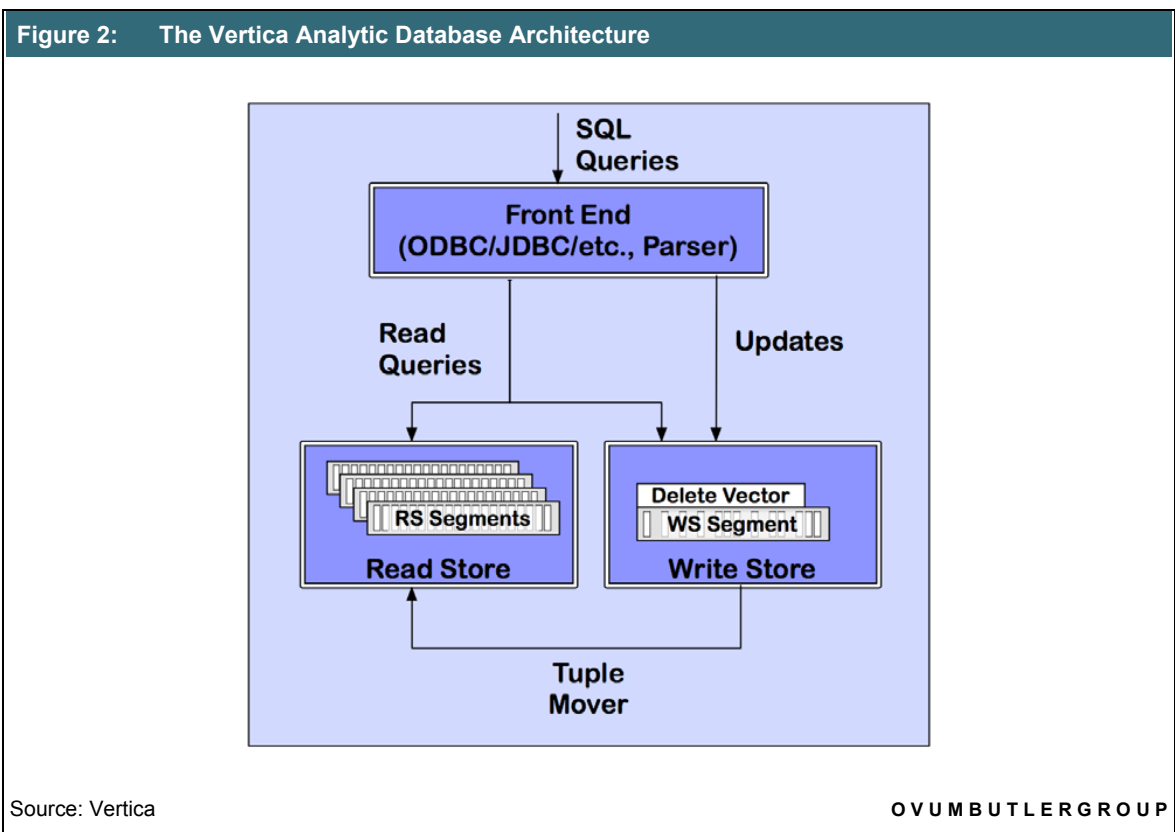
The Vertica Analytical Database Management System is optimised for storing and querying large datasets for data warehousing. Its database engine organises data into columns rather than rows. This differs from traditional row-based relational database management systems (RDBMS) that were originally designed for transaction processing. In these, a read operation involves reading all of the fields in a row even when only a fraction of the record is required. While this is not of any consequence in transaction processing, it impacts query performance in data warehousing. Analytical applications usually require a small fraction of data to be read from millions of records/rows and the efficiency of the read operation can make a big difference to performance. Column-based databases hold values for the same record attribute in one column that can be read without having to read the full record. This can improve the performance of analytical applications, hence the burgeoning market for analytical databases.



The internal aspects of the column-based database are hidden from the user through standard support for SQL, ETL tools, and business-intelligence (BI) applications.

Vertica combines the benefits of a columnar database engine with massively parallel-processing (MPP) and shared-nothing architecture on clusters of commodity hardware to provide relatively low-cost performance on multiple-terabyte data sets.

Since version 2.5 Vertica has had the ability to support any type of database schema design.



The built-in parallelism is achieved in part by segmenting large tables across nodes of a cluster and replicating smaller tables such as dimension tables on each site of a cluster to reduce network traffic.

In the database world, for every performance gain there is usually a trade-off. In the case of column-based databases, it is delete and update operations that can be slow and complex. However, this is not the case with Vertica and its hybrid storage architecture that efficiently trickle-loads and updates data to and between an uncompressed write-optimised column store (WOS) and a compressed read-optimised column store (ROS) (see Figure 2).

For bulk-loading, data is pulled into the WOS in-memory (RAM) in an uncompressed format and then written out in bulk, based on a configuration parameter specifying how frequently to move data (every minute, for example) or how much data to move (1 gigabyte at a time, for example) to a compressed ROS.

Each compressed volume of data is then merged into a file. Access to these files is fast because it doesn't involve any sorting and does not compromise the compression scheme. It is also possible to bulk-load directly to the ROS. Queries are executed by reading data from both the WOS and ROS stores. Because data snapshots are transferred and loaded between the two stores, in-memory conflicts between read and write are minimised. Interestingly, the data in memory is often stored in a row-oriented structure, and is not in the columnar style of Vertica's core database engine. This unique hybrid approach by Vertica is designed to get around the slower speeds of loading and updating columnar databases. Vertica refers to this hybrid architecture as "continuous load and query".

Vertica's Analytic Database has been architected from the ground up to handle large query-intensive workloads, with good performance achieved without the use of proprietary hardware. The product also provides intelligent capabilities for improving disk I/O efficiency. Ovum believes that Vertica's product will be attractive to companies looking to stretch performance limitations and budgets associated with heavy-duty analytics.

Product Operation

Vertica has good compression capabilities in common with other column-oriented databases. Because the columns hold data of the same type, more efficient compression algorithms can be applied to the columns that can reduce the size of the data by as much as 90%. By contrast, in row-oriented databases each row contains fields of data of distinct types, so compression cannot be optimised to the same degree. Vertica achieves further reductions in the size of data through encoding techniques where codes are used to denote unique data values, and are then mapped to the real data values.

Vertica can perform queries on encoded data and so reduce the need for data storage and processing requirements and help lower the space required for data movements within the data warehouse. There are other advantages, for instance simplification of schema design, and lower overhead requirements, such as for temporary storage of data, for indexes and other database-tuning parameters. These help reduce the overall database size requirements.

Vertica's database radically reduces disk I/O by only having to read the encoded data in columns referenced by the query. So rather than having to package/unpackage all the data from disk it makes the execution run much faster as a result of gains in in-memory speed and movement. It therefore reduces the number of actual disk reads and time required to satisfy a query.

Performance can also be bumped up through the use of parallelism across clusters of shared-nothing computers where data is typically partitioned horizontally (different data rows are shipped to different server nodes) with queries resolved independently on each server and the results combined centrally. This means that Vertica databases can be scaled up simply by adding additional servers to the cluster.

Another key design feature of Vertica is the way it distributes data across a grid in specialised data structures called "Projections" used in place of tables. The Projections are selected based on an analysis of the data and queries executed against it. The data in each Projection can optionally be stored in "overlapping" views, each of which is sorted by at least one of the columns in the view. Grouping data separately, in terms of sort orders on different nodes (Projections) of the grid, enables companies to be flexible in their storage strategies with columns sorted differently in each Projection. Like conventional relational DBMS indexes and materialised views the pre-sorting of columns helps bump up query performance, but it does run the risk of redundancy, which can impact data load and update times. Vertica tries to overcome this by building "active redundancy" into Projections and using of parallelisation to afford better concurrency and support for failover and recovery. For example, a query that has a join across several columns is directed to the Projection that has a particular column pre-sort order needed to most efficiently process the data.

Vertica has no specialised nodes, all nodes are peers and queries or load can be sent to any node.

Vertica's FlexStore broadens the applicability of column stores by allowing columns to be grouped together based on size and usage to be exploited for fast loading. Another way of looking at FlexStore is that it provides multiple ways of organising data for different queries. It also analyses storage media to optimise storage based on frequency of accessing columns of data. Columns are then stored on the fastest or slowest available storage media based on usage. Vertica refers to this as Intelligent Disk Utilization, which constitutes the first phase of its plans for information-lifecycle management (ILM) that will eventually lead to Vertica offering solutions for archiving.

Vertica includes capabilities for automatic design and administration. It analyses the types of data and queries to recommend a physical database design to provide the best performance for the user's needs. This can then be finely tuned manually by the database administrator but minimises the time that the DBA spends on physical database tuning.

The Vertica database supports SQL, ODBC, JDBC, and the majority of ETL and BI reporting products. It also supports virtualisation. It can handle common SQL Server and Oracle SQL extensions but it does not claim to be 100% compliant with them.

Vertica 3.5 has built-in support for Hadoop Map Reduce. It has a connector that allows access to the Vertica analytical database via queries in Hadoop. This capability can be used for data manipulation or to perform time-series analysis on large volumes of data using languages other than SQL.

Other features include support for Perl and Python (over ODBC) and load-balancing via Linux's IPVS capability. In addition, Vertica offers a console for administration purposes that is based on a version of the open-source Webmin tool adapted for Vertica. It offers functionality such as database creation, and stopping and starting. It also includes an open-source monitoring tool called Ganglia that provides performance-monitoring capabilities such as memory and disk usage and alerts if any preset thresholds are reached. It can follow performance for a specific query or can provide collective performance statistics.

Product Emphasis

Vertica's strategy is to focus on the market for high-end analytics performed on relatively large volumes of data. Vertica's offering has been designed from the ground up for performance. To that end the company has developed built-in intelligent capabilities for duplication, grouping, and storage of data. This optimising software approach minimises disk I/O and has been very successful for Vertica as evidenced by the company's growing list of customers.

Vertica operates in a competitive market with other column-based DBMS vendors including Sybase, Infobright, and ParAccel. Vertica and some of the other newer players, notably Infobright and ParAccel, claim to have more flexible modern architectures. However, Sybase has invested heavily in its IQ Analytics Server in recent years and has achieved stellar growth in revenue as a result. Vertica is still building its presence outside its core US market, and in Ovum's view needs to increase its profile in Europe and other geographies to achieve a bigger global market share.

Ovum believes that high-performance column-based databases are now established as a separate market in analytic data warehousing, driven by companies looking to replace traditional database platforms in favour of compression-based column-oriented architectures for better price/performance. Vertica is well positioned to take advantage of this demand.

DEPLOYMENT

Vertica's Analytic Database runs on an inexpensive grid of servers from HP, Dell, IBM, and others, and on Linux (Red Hat Enterprise, Fedora Core, and SuSE) with standard configurations. Grid connection is provided via Ethernet interconnects that connect nodes across a cluster, with each node having its own local disk-storage facility. While network storage is supported, there is no explicit need to implement a storage-area network (SAN) because of the enormous compression gains that Vertica delivers. The net result is that Vertica's product tends to have a much smaller footprint on the network than some other databases.

Purchasing options include a software-only offering, a pre-packaged commodity hardware appliance option, a cloud offering, or a virtualised solution. The appliance option comes with pre-configured platform-optimisation but does not offer any other performance-related features such as superior network connections.

The cloud deployment option is designed to be hosted on Amazon's Elastic Compute Cloud (Amazon EC2) cloud-computing environment. It is based on a pay-as-you-go model, similar to a software-as-a-service (SaaS) subscription, with optional infrastructure management provided by RightScale. The hosting model is intended to speed up and lower the cost of deploying Vertica's analytic database.

According to Vertica, the average deployment time from sale licensing to production rollout is about three months, faster for the Amazon EC2 implementation. The sale doesn't usually come with a heavy services element. Instead Vertica has concentrated on making installation and migration quick and easy, claiming it can pre-load a server cluster "in minutes" and that technically the only need to employ services help would be the schema design and optimisation associated with Projections.

The software comes with documentation and there are online training resources such as walk-throughs with real-life examples developed using best-practice guidelines. Vertica also provides on-site training for customers.

PRODUCT STRATEGY

Vertica targets all markets, vertical and horizontal. The solution is particularly suited to organisations that have a lot of data-analysis requirements, irrespective of the size of the organisation. For those companies, Vertica aims to deliver a rapid return on investment mainly by lowering the cost of ownership. The company has customers in the telecom, financial services, Internet, and healthcare sectors, as well as marketing analytics and embedded SaaS.

Vertica sells its product directly and through OEM relationships. Key business partnerships and alliances include MicroStrategy, IBM Cognos, SAP BusinessObjects, Talend, Tableau, HP, Sun, RedHat, and Novell. Vertica has technology partnerships with a number of companies including IBM, HP, RedHat, MicroStrategy, Tableau, and Talend.

Pricing for Vertica's Analytic Database is based on the amount of data (per terabyte) that the system stores. This is intended to give customers freedom to custom-configure their hardware set-up because there are no per-CPU/core charges. The project value of Vertica installations can be anything from US\$50,000 to US\$5,000,000. According to Vertica, the cost of licensing typically accounts for 80% of the cost of deployment. Pricing for the Vertica Analytic Database for the Cloud starts at \$2,000 per month to store and query 500 gigabytes of data. This makes the offering suited to small-to-mid-sized datamart implementations, and will also make it attractive to providers of outsourced analytic services.

Technical support is provided on a 24x7 basis via telephone and email.

Vertica Systems releases two major versions of the product each year with service packs provided throughout the year as and when needed. Upcoming releases planned for Q1 and Q4 2010 will be focused on enhancing the analytics capabilities of the product, its performance and scalability, and internationalisation features.

COMPANY PROFILE

Vertica Systems is a supplier of high-performance analytical database-management systems. It was founded in 2005 and is headquartered in Billerica, Massachusetts. It has about 80 employees, most of whom are based in the US, with some representation in the UK and India. Vertica has had financial backing from venture-capital firms such as Bessemer Venture Partners, Highland Capital Partners, Kleiner Perkins Caufield & Byers, and New Enterprise Associates. Even though Vertica was founded in 2005, the first version of its analytic database was stealthily released in December 2006 to the “early-adopter” market. It was only with the launch of version 2.0 in February 2008 that it started to do some real marketing and publicity.

One of the company’s co-founders is Dr Michael Stonebreaker, a “Codd-like” celebrity in the database world who invented Ingres and was the CTO (a position he holds at Vertica now) of former database pioneer Informix. Vertica’s other co-founder is Andrew Palmer, a business executive with experience in both software and biotechnology.

The US is the primary market for Vertica with only 20% of its revenues coming from the rest of the world. This is likely to change as the company expands its overseas operations. Vertica has gained over 100 customers since it first started. The customers are mostly departments in organisations that either deal with relatively large volumes of data or have many analytical requirements. One example is Comcast, a US cable company that runs a number of very large network analytics data warehouses on Vertica. Another customer is Verizon, a US telecoms giant that has built several large call-data record (CDR) marts to support billing assurance, compliance, and other initiatives. Other customers include BlueCrest, HMetrix, JP Morgan, Mozilla, Sonian, and Sprint.

As a private company Vertica does not publish its financial results.

SUMMARY

Vertica’s Analytic Database has been architected from the ground up to handle query-intensive workloads. Vertica delivers good performance using a columnar database engine combined with an innovative hybrid data store for improved data-loading, and smart capabilities for optimising data access and storage. It does so without the need for proprietary hardware, and the database runs on clusters of industry-standard servers that are relatively inexpensive. Ovum believes that organisations can derive value from Vertica’s Analytic Database by taking advantage of the solution’s smart capabilities and its agility to rapidly analyse large data volumes.

Table 1: Contact Details	
<p>Vertica Systems 8 Federal Street Billerica MA 01821 USA Tel: +1 978 600-1000 Fax: +1 978 600-1001 Email: info@vertica.com www.vertica.com</p>	<p>Source: Vertica</p> <p style="text-align: right;">OVUMBUTLERGROUP</p>

Headquarters

Shirethorn House,
 37/43 Prospect Street,
 Kingston upon Hull,
 HU2 8PX, UK
 Tel: +44 (0)1482 586149
 Fax: +44 (0)1482 323577

Australian Sales Office

Level 46, Citigroup Building,
 2 Park Street, Sydney,
 NSW, 2000,
 Australia
 Tel: + 61 (02) 8705 6960
 Fax: + 61 (02) 8705 6961

End-user Sales Office (USA)

245 Fifth Avenue,
 4th Floor, New York,
 NY 10016,
 USA
 Tel: +1 212 652 5302
 Fax: +1 212 202 4684

Important Notice

This report contains data and information up-to-date and correct to the best of our knowledge at the time of preparation. The data and information comes from a variety of sources outside our direct control, therefore Butler Direct Limited cannot give any guarantees relating to the content of this report. Ultimate responsibility for all interpretations of, and use of, data, information and commentary in this report remains with you. Butler Direct Limited will not be liable for any interpretations or decisions made by you.

For more information on OVUM Butler Group's Subscription Services please contact one of the local offices above.

