



## **Data Warehouse in the Enterprise**

A Competitive Review of Enterprise Data Warehouse Appliances and Technology Solutions

SQL Server Technical Article

**Published:** January 2009

**Applies to:** SQL Server 2008

**Summary:** In this white paper, we discuss the data warehouse products from three traditional vendors, as well as the newer appliance and column-based vendors. We compare the strengths and weaknesses of these products to Microsoft® SQL Server® 2008 and shows that Microsoft provides the best choice of data warehouse solution among all vendors in the marketplace because of its broadest offerings, openness, and cost-effectiveness.

# Copyright

The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

This white paper is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED, OR STATUTORY, AS TO THE INFORMATION IN THIS DOCUMENT.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in, or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

© 2009 Microsoft Corporation. All rights reserved.

Microsoft, SQL Server, Windows, and Windows Server are trademarks of the Microsoft group of companies.

All other trademarks are property of their respective owners.

# Contents

- Introduction..... 5
- Main Vendors..... 6
  - IBM ..... 6
    - Products ..... 6
    - Operation ..... 7
    - Market Position..... 7
  - Teradata ..... 7
    - Products ..... 8
    - Operation ..... 8
    - Market Position..... 8
  - Oracle..... 9
    - Products ..... 9
    - Operation ..... 9
    - Market Position..... 10
  - Sun ..... 10
    - Products ..... 10
    - Operation ..... 10
    - Market Position..... 10
- Data Warehouse Appliances..... 11
  - Row-Based Data Warehouse Appliance Vendors ..... 12
    - Netezza..... 12
    - Greenplum ..... 12
  - Column-Based Data Warehouse Appliance Vendors..... 12
    - Sybase IQ..... 12
    - EXASOL ..... 12
    - ParAccel..... 13
    - Vertica ..... 13
    - HP Neoview ..... 13
    - Kickfire ..... 13

Conclusion..... 13

## Introduction

The SQL Server 2008 database system provides a comprehensive and scalable data warehouse (DW) platform that enables organizations to integrate data into the data warehouse faster and scale and manage growing volumes of data and users, while delivering insights to all users. SQL Server 2008 enables customers to quickly build their data warehouse by providing development teams with Business Intelligence Development Studio, which provides a productive and collaborative environment for building solutions. Customers can easily manage large volumes of data because of the increased scalability, manageability, and performance provided by SQL Server 2008. Finally, SQL Server 2008 delivers better business insight by providing deep analytical capabilities, rich visualization, and enterprise reporting to all employees.

IBM has a data warehouse product stack that is repositioned as a data warehousing appliance. Balanced Configuration Unit (BCU) server hardware is sold with DB2 Data Warehouse Edition (DB2 DWE) software to create a data warehouse appliance-like system. The IBM enterprise data warehouse solution stack consists of four different, fragmented software products on two different hardware platforms. The result of the fragmented product offering is a licensing model that is both expensive and complicated. The products themselves are also complicated, and require extensive tuning to reach optimal performance. The complicated nature of the products and a lack of experienced professionals to work with these systems lead to increased training costs and increased ongoing staff costs.

Teradata could be considered the original data warehouse appliance vendor. Teradata sells data warehouse solution packages that include consulting, support, hardware, and software. Teradata is challenged in terms of performance and is expensive, both in terms of initial licenses and ongoing support costs. To achieve high performance and availability, Teradata requires expensive proprietary hardware, and to scale down to smaller applications is economically unfeasible. To provide a complete solution, Teradata relies on partnerships for tools such as extract, transform, and load (ETL), as well as reporting, analytics, backup, and advanced replication. This reliance complicates licensing and causes problems because competitors are increasingly acquiring these partners. Perhaps because of these problems, Teradata has experienced tremendous pressure from competitors in the data warehouse market and recently released Teradata Accelerate bundles offering lower-cost hardware and software packages to enable mid-size companies to get started in data warehousing.

Oracle is also developing more specialized data warehousing solutions. The latest HP Oracle Database Machine combines Oracle software and HP hardware for its data warehouse. However, Oracle data warehouses are complex to deploy if customers do not already have the necessary skills. Moreover, the licensing and support costs are expensive, the benchmark figures are difficult to replicate in the real world, and there is little support for smaller data warehouse implementations.

Data warehouse appliances (such as Vertica, Netezza, Greenplum, HP Neoview, ParAccel, EXASOL, and Sybase) are proprietary integrated storage, server, operating system, database, and software specifically designed for data warehousing performance. Data warehouse appliances are a new development in this

field. There are two types of storage used for data warehouse appliances: traditional row-based storage and column-based storage. As implied by the name, column-based storage stores records in columns rather than rows. Data warehouse appliances typically provide performance and scalability, and column-based appliances are optimized for aggregation. However, data warehouse appliances are not mature systems and the vendors are typically very small companies with no established track record, installed base, or research and development resources to evolve. Moreover, data warehouse appliances require third-party software to provide a complete solution. Finally, column-based appliance systems provide good performance only for aggregation. While aggregation queries are important, they are typically a small subset of total data warehouse queries.

Microsoft SQL Server 2008 is a leader in the data warehouse field. In fact, the Gartner DW Magic Quadrant '07 defined SQL Server as a leader in data warehousing. Microsoft has accelerated its development in SQL Server 2008 with improved scalability, security, productivity, and total cost of ownership. Furthermore, Microsoft has also shown its vision and commitment to the future of data warehousing with the acquisition of DATAAllegro. DATAAllegro's data warehouse appliance installations boast some of the largest data-volume capacities in the industry with hundreds of terabytes of data on a single system.

Microsoft provides the broadest offerings from the breadth and depth of data warehouse technology solutions, the wide-ranging analytical solutions provided by SQL Server 2008 Analysis Services, the highly scalable reporting delivery system to all users provided by SQL Server 2008 Reporting Services, and ubiquity of usage in the Microsoft Office system. Furthermore, SQL Server runs on standard nonproprietary operating systems, software, and hardware. This standardization and openness improves flexibility and ease of support, and ensures that customers do not need additional resources to support the underlying infrastructure of their data warehouse deployment.

## **Main Vendors**

This section includes the major data warehouse vendors that compete with SQL Server 2008.

### **IBM**

IBM is one of the mature vendors of software and hardware technology and is well-known in the data warehouse market. In addition to hardware and software, IBM offers infrastructure services, hosting services, and consulting services. IBM is repositioning itself as a data warehouse appliance vendor with its InfoSphere Balanced Warehouse.

### **Products**

IBM has grouped its data warehouse products under the InfoSphere brand, and they have two main products, InfoSphere Balanced Warehouse and InfoSphere Warehouse.

The InfoSphere Balanced Warehouse is an extension of the Balanced Configuration Unit (BCU) data warehouse appliance, which combines server hardware with DB2 Data Warehouse Edition (DB2 DWE) software to create a data warehouse appliance-like system. The InfoSphere Balanced Warehouse is not a true data warehouse appliance, but is a package of storage, hardware, operating system, and software that is designed to be complementary and provide an out-of-the-box solution.

IBM provides different products for different sizes of business. The InfoSphere Balanced Warehouse is available in C-Class, D-Class, and E-Class variants, and the InfoSphere Warehouse is provided in Starter, Intermediate, Base, and Enterprise versions. InfoSphere Warehouse is provided as a software-only solution to be installed onto existing hardware and operating systems, and is based on IBM DB2 9.5. InfoSphere Warehouse can be installed onto Linux, UNIX, or the Windows® operating system, and is designed for mixed workloads covering OLTP, data warehousing, and business intelligence (BI). It includes a database, OLAP cubes and analytics, ETL, and data mining.

## Operation

IBM has been careful to use the InfoSphere branding and avoid mentioning DB2. The size of the database can range from 1.3 to 5 times the size of the source data, based on the expertise of the administrator. Therefore, a team of skilled administrators is imperative, but there is a shortage of skilled IBM DB2 DWE experts and this can have a significant impact on deployment and maintenance costs. Without this skilled staff, the cost and complexity of deploying and IBM solution is daunting for organizations, and even when the staff is in place, there is a high initial licensing cost and high ongoing support costs.

## Market Position

IBM has created products for all sizes of business from small companies to enterprise class solutions; however, the products are not the same. For example, an InfoSphere Warehouse Starter edition does not have the functionality of the Enterprise edition, regardless of the improvements in hardware. The lower-priced offerings are also limited in functionality, and an expanding company cannot simply scale up by adding hardware. On the other hand, Microsoft offers one software product, SQL Server 2008, that is able to scale to meet the varying data warehouse demands from small to the largest enterprise customers.

## Teradata

Teradata is also one of the well-known vendors in the data warehouse market. Formerly part of NCR, they became an independent company in 2007. They provide a solution comprising storage, hardware, and software, which can be deployed on an operating system of your choice and can be considered the original data warehouse appliance.

## Products

Teradata is a massively parallel processing (MPP) system, using a shared nothing architecture that is scalable in different dimensions of DBMS workloads (data volume, breadth, number of users, and complexity of queries). Shared nothing architecture, as the name implies, shares nothing between nodes. The data is split between nodes and each node acts independently. This removes potential bottlenecks and, in theory, allows limitless scaling up through the addition of nodes. In addition, the nodes provide high availability because the Teradata Database can fail over workloads between nodes.

The Teradata solution combines hardware and software and is offered on Intel-based servers. The networking is provided by the BYNET messaging fabric, to provide scalability, fault tolerance, and speed. Teradata solutions can be deployed on UNIX SVR4.2 MP-RAS (Teradata's proprietary UNIX system, a variant of System V UNIX from AT&T), Windows Server® 2003, and SUSE Linux Enterprise Server.

Teradata 12.0 includes high-performance parallel architecture, a suite of data access and management tools, and data mining software. In addition, Teradata is including data warehouse training and consulting advice as an integral part of their product delivery. Furthermore, Teradata can provide consulting services and has an online support program, certified training courses, and a certified professional program.

## Operation

There is a partner network to extend the functionality of the data warehouse, including Microsoft for BI (Reporting and Analytics); GoldenGate for replication; Hyperion, Business Objects, and Microstrategy for reporting and analytics; and Ward Analytics for performance analysis. The risk of partnerships to complete the BI solution is that the partners themselves can fail or be acquired by competitors and, since Hyperion was acquired by Oracle and Business Objects was acquired by SAP, there is no guarantee that these partnerships will continue. Moreover, Teradata is a proprietary black-box solution requiring specialist hardware and skills, and these can be expensive to acquire.

## Market Position

Teradata is a functional system, but it is also very expensive. Microsoft currently offers comparable functionality at a fraction of the price in the sub-30TB range, and the emerging appliance vendors are competing on larger systems in the hundreds-of-terabytes range, again at a much lower cost than Teradata. Furthermore, with the acquisition of DATAlegro, Microsoft will move into the very large data warehouse market in the hundreds-of-terabytes range and provide the stability, support, and maturity of a large vendor, but at a considerably lower price. This competition in price and functionality at all levels is causing a shrinking market share for Teradata, which might force Teradata to lower prices. It is still to be seen whether Teradata can survive in a more competitive market, particularly without the backing of NCR, its previous parent company.

Teradata is an expensive system to purchase, with high developer costs, high annual support costs, and high upgrade costs. To scale a Teradata system down to smaller applications is economically unattractive.

## Oracle

Oracle is also one of the well-known data warehouse vendors with a wide-ranging support, education, and consultancy infrastructure.

### Products

Oracle's most recent announcement has been the HP Oracle Database Machine, which is being positioned as Oracle's comprehensive data warehouse product offering. A key component of the Database Machine is the Exadata Storage Server, which is based on HP hardware, Cisco Infiniband high-speed networking, and Automated Storage Management (ASM). The HP Oracle Database Machine includes software, servers, and storage and is claimed to be 10 times faster than conventional data warehouse systems. Aimed at large, multiterabyte data warehouse deployments, the Database Machine combines a range of technology solutions from Oracle and its partners, including the Oracle Exadata Storage Server, Oracle Database 11g, Real Application Clusters, Oracle Enterprise Linux, Infiniband networking, and related hardware. All these components result in very high acquisition cost compared to traditional data warehouse solution from Oracle.

Previously, Oracle's data warehouse appliance solution strategy has been focused around the Oracle Optimized Warehouse Initiative, which consists of prevalidated hardware configurations and preinstalled database software.

Oracle keeps changing its data warehouse strategy and sending contradictory messages to customers by promoting both its commodity hardware scale-out solution using RAC technology and its proprietary scale-out solution using Exadata.

### Operation

Oracle data warehouses require extensive tuning to optimize storage, achieving storage figures of 1.5 times the size of the source data. Without optimization, the storage can reach 5 times the size of the source data. Despite Oracle's performance claims, no formal benchmarks have actually been published for the Database Machine to validate such claims.

Oracle typically has many optional features that are not included with the base versions of its products; therefore, the final price is often far higher than published prices. Oracle maintenance costs are high, and the overall TCO is poor. Oracle typically charges an annual maintenance fee that is 22% of the price of the product, and, in subsequent years, this may not be based on the original purchase price, and it might rise further.

## Market Position

Oracle has a long-standing partnership with HP and relies on this partnership for products such as the HP Oracle Database Machine. This relationship has lasted because the two companies do not directly compete, but when HP acquired Compaq, they had access to the Tandem NonStop SQL database, and have since developed Neoview, a data warehouse appliance that is in direct competition with the HP Oracle Database Machine. It remains to be seen whether this relationship will survive. Oracle plans to sell, install, and eventually support the Database Machine Product, but Oracle has a lack of experience in hardware support and deployment.

## Sun

Sun is not a data warehouse provider as such, but it is selling the Greenplum data warehouse appliance packaged with its X4500 hardware as the Sun Data Warehouse Appliance.

## Products

Greenplum is a modified version of the open-source PostgreSQL database, with modifications to process parallel queries and manage the parallel workload. The Sun/Greenplum offering is aggressively priced and can produce high-performance results. Sun claims that Sun Data Warehouse Appliance is the most cost-effective high-performance data warehousing solution and that it is more energy-efficient and therefore generates less carbon dioxide and heat than other systems.

## Operation

Sun also partners as a hardware provider with other data warehouse providers such as Oracle, Sybase IQ, and Kickfire, and it is unclear how these partnerships will be affected by Sun becoming a direct competitor because of Sun's acquisition of MySQL. The Sun partnerships are somewhat haphazard. The Sun Data Warehouse Appliance is in partnership with Greenplum, but the Sun strategic data warehouse platform is Sybase IQ. Sun also partners with many other data warehouse vendors, making it unclear what their long-term goal is.

Because Greenplum is such an emerging product, there are very few skilled professionals that can implement and support it. The lack of supply artificially inflates their consulting rates, and it increases the risk of successfully completing implementation projects.

## Market Position

Although Sun is a large company, it has little experience in this area. Greenplum is a small, new company and has little support or consultancy operations. This new technology has yet to mature, and it has neither a broad customer base that would confirm the claimed performance, nor the ability to manage the mixed workload of a modern, enterprise-scale data warehouse. With these factors in mind, the Sun Data Warehouse Appliance should be viewed as a high-risk system that should only be used by

organizations that can fully self-support their systems. Sun in the past several years has experienced several downturns in their stock price, laying off people, and removing key initiatives such as cloud services. It remains to be seen whether, in the near term, Sun can build the data warehouse professional services organization capable of supporting the large potential customer base for this new offering.

## Data Warehouse Appliances

A data warehouse appliance is an integrated set of servers, storage, operating system, database, and software specifically optimized for data warehouse applications. Data warehouse appliances generally target the mid-to-large-volume data warehouse market, based on claims of low cost and high performance and scalability to data volumes in the terabyte (TB) to petabyte (1,000 TB) range.

Most data warehouse appliance vendors use massively parallel processing (MPP) architectures to provide fast query performance and data platform scalability. MPP architectures consist of independent processors or servers executing in parallel. Most MPP architectures implement a shared nothing architecture where each server is self-sufficient and controls its own memory and disk resources. Data warehouse appliance-based solutions claim lower total cost of ownership, reduced maintenance, and high performance as their key strengths.

In this section, we have included both row-based data warehouse appliances and column-based databases. Although these systems operate quite differently, column-based systems are typically marketed as data warehouse appliances, and both systems operate in the same market segment.

Column-based systems organize records into columns rather than rows. This means that when a row is retrieved, you have one attribute sliced across all records. This system provides very good performance in OLAP systems where aggregation performance is crucial. However, this performance comes at a cost. Column-based systems are only effective on read-only databases, because the workload to modify a row is increased significantly over a row-based system. In addition, to achieve the high performance figures it is necessary to avoid joins, limit the columns used in the query, and ensure that the data can fit into available memory.

Most data warehouse appliances share common weaknesses. They are typically produced by small, new companies with few real-world customers to offer true benchmarks and provide feedback. It is very difficult and expensive to recruit or train professionals on these niche systems and, furthermore, there is little or no support and consultancy provided by the companies themselves. Where it is offered, consultancy is typically from expensive, inexperienced third parties. By purchasing a fully integrated system, you are locked into that technology and cannot change the hardware or operating system in the future without replacing the whole system. Plus, the risk that these companies will not continue to exist in these tough economic times is also uncertain.

Column-based systems particularly are very specialized and are not appropriate for a real-world primary data store. It is not appropriate to compare the performance against a relational database when most real-world database applications require a balance of read and write access in querying, for example, when performing data loads or updates.

## Row-Based Data Warehouse Appliance Vendors

### Netezza

Netezza Performance Server is a data warehouse appliance combining database, server, and storage, based on Linux and PostgreSQL. It uses a patented massively parallel architecture. Netezza has posted impressive performance benchmarks and is competitively priced; however, it is based on proprietary technology, it does not work well in mixed workloads because it has no indexes, and there are some doubts in the marketplace about boardroom commitment, with executives selling large volumes of shares and the Chief Executive Officer resigning. It is also a proprietary stack offering with no hardware flexibility. Furthermore, Netezza offers no integrated BI tools.

### Greenplum

See the section on Sun, above.

## Column-Based Data Warehouse Appliance Vendors

### Sybase IQ

Sybase IQ Analytic Server is the oldest column-based data warehouse solution (about 15 years old) and it runs on several versions of UNIX, Linux, and Windows. The Sybase Analytic Appliance is a combination of Sybase IQ Data-Warehouse software, Sybase PowerDesigner, Sybase ETL, and Microstrategy8 BI software deployed on the IBM POWER Systems hardware platform using AIX.

The continuing support of Sun is no longer assured as it has its own product and works with other data warehouse vendors. As with all column-based systems, Sybase IQ performs poorly when data is modified.

### EXASOL

EXASOL supplies EXASolution, a shared nothing MPP running on EXACluster, its own Linux operating system. EXASolution is available as software only, or with appropriate hardware. EXASOL has some of the best results in TPC-H6 benchmarks, but it is important to note that TPC rules limit a direct comparison of actual production performance. EXASOL has high performance, but for a very narrow workload, with very high hardware costs, running a proprietary database on a proprietary operating system. For the 1TB TPC-H benchmark, they had 640 GB of memory, 32 TB of disk storage in 240 disks, and 40 data cluster servers.

## ParAccel

ParAccel Analytics Database is based on the PostgreSQL database with an MPP engine running on Linux (although a Windows version is in development), with storage provided by EMC. ParAccel is a very small organization with only about 50 employees, no extensive support system, and few customers. They have posted good TPC-H benchmarks, but it is very expensive to achieve this level of performance. EXASOL has since beaten these benchmarks, and it remains to be seen whether this performance can be translated to the real world.

## Vertica

Vertica was founded in 2005 and offers application-specific data marts based on the C-Store column-oriented database technology developed as an open-source project at MIT. Vertica runs on Intel-based Linux servers. Vertica claims it has very high performance, but this is yet to be verified and, as with all column-based systems, it is built and designed specifically for OLAP, making it inappropriate for a wider range of queries. The claimed performance benefits are only seen in cache-friendly operations, such as the sum or average of a column.

## HP Neoview

HP Neoview is a data warehouse appliance based on the Tandem NonStop SQL database, acquired as part of Compaq. Neoview is supplied as a complete solution of storage, server, operating system, and database running on HP hardware. HPNeoview is an emerging product in a mature market with only a handful of customers, and HP is unlikely to want to jeopardize its relationship with existing data warehouse partners such as IBM, Microsoft, and Oracle. Therefore, HP's strategy for Neoview is as yet unclear.

## Kickfire

Kickfire released its first product in 2008 using custom hardware and the MySQL database on Linux. Unlike most column-based systems, Kickfire does not support MPP or clustering, but instead relies on a specialist SQL-on-a-chip solution. These limitations restrict its scalability. Like most column-based solutions, the TPC-H results are good, but, as with all column-based systems, it is inappropriate for a wide range of queries.

## Conclusion

Microsoft SQL Server 2008 provides an enterprise-level scalable, mature data warehouse platform that can be deployed on systems running Windows Server and generic hardware. Using standard systems, you can use existing hardware, skills, and partnerships, and consolidate your systems, improving integration, development, security, and support, and reducing costs.

SQL Server's leadership has been recognized by many analysts. SQL Server is a leader in Gartner's Data Warehousing2 Magic Quadrant. They have highlighted that the use of SQL Server for data warehouse is accelerating. They have also praised Microsoft's worldwide support, and they have noted that SQL Server scales to large data volumes with little effort. All of this has led Gartner to conclude that SQL Server represents good value for money.

SQL Server has the most integrated BI platform with enterprise-class ETL, OLAP, reporting, and data mining all included at no additional cost. Furthermore, through integration with Microsoft Office, Microsoft provides users with straightforward, user-friendly access to business information to extend its vision of ubiquitous BI.

SQL Server already scales up to many terabytes, but, with the acquisition of DATAlegro, Microsoft has shown its commitment to scale up still further, with real customers running today with loads close to half a petabyte. DATAlegro is a data warehouse appliance vendor using a row-based data warehouse to provide high performance over the full range of queries. DATAlegro has a commitment to standards-based systems and has therefore avoided proprietary hardware and operating systems that can leave customers at a technological dead end.

Microsoft SQL Server 2008 avoids many of the problems of its competitors. It is a mature, stable system with a proven track record. Microsoft has an extensive support, consultancy, and education portfolio, and there are many well-trained and highly skilled SQL Server professionals. The hardware and operating systems are likely to already be in use and fully supported within your organization, and with the acquisition of DATAlegro, there is a clear ambition to scale up even further. This level of performance and functionality, combined with the lowest total cost of ownership, makes SQL Server the best choice for your data warehouse needs.

**For more information:**

<http://www.microsoft.com/sqlserver/>: SQL Server Web site

<http://www.microsoft.com/sqlserver/2008/en/us/data-warehousing.aspx>: Microsoft SQL Server 2008 data warehousing

<http://mediaproducts.gartner.com/reprints/microsoft/vol7/article3/article3.html> : Gartner DW Magic Quadrant'07

<http://www.microsoft.com/presspass/press/2008/sep08/09-16DAPR.msp>: Acquisition of DATAlegro

<http://www.datallegro.com>: DATAlegro

[http://www.gartner.com/DisplayDocument?ref=g\\_search&id=766714](http://www.gartner.com/DisplayDocument?ref=g_search&id=766714): Gartner report on Oracle's Exadata

<http://www.tpc.org>: Transaction Processing Council

Did this paper help you? Please give us your feedback. Tell us on a scale of 1 (poor) to 5 (excellent), how would you rate this paper and why have you given it this rating? For example:

- Are you rating it high due to having good examples, excellent screen shots, clear writing, or another reason?
- Are you rating it low due to poor examples, fuzzy screen shots, or unclear writing?

This feedback will help us improve the quality of white papers we release.

[Send feedback.](#)