# World's Largest Data Warehouse

## Sun™ Data Warehouse Reference Architecture — Scalability over 1 PB

Many enterprises use operational and transactional systems to find valuable business intelligence. As data warehouses grow to terabytes and beyond, companies are looking for scalable solutions that provide more features and use less datacenter floor space to reduce operational costs. Using the Sun™ Data Warehouse Reference Architecture for Structured and Unstructured Data, Sun built the world's large data warehouse. Able to support over 1 PB of structured and unstructured source data and scale to meet demand, the eco-friendly solution uses less storage, consumes less energy, and generates less heat than conventional solutions.

### Highlights

- Take advantage of a Sun data warehouse solution that uses 90 percent less storage, consumes 91 percent less energy, generates less heat and carbon dioxide, and costs less than conventional solutions
- Scale to petabytes of input data with a massively scalable data warehouse built with Sun technologies, the Solaris™ Operating System, Sybase IQ, and BMMsoft DataFusion
- Load raw transactional data faster with a solution shown to handle three million rows per second
- Store and search any type of data faster, including structured and unstructured data
- Ensure reliable operation with systems and software that are designed to run even during data loading and maintenance efforts
- Enable regulatory compliance for records retention and retrieval

### The need for smarter, bigger, and more cost-effective data warehouses

Companies the world over rely on operational systems, such as enterprise resource planning (ERP), supply chain management (SCM), and customer relationship management (CRM), as well as online transaction processing (OLTP) systems to manage the business. As a result, enterprises are turning to data warehousing solutions to collect, integrate, manage, analyze, and secure data so that it is available to users requiring ready access to information in order to make timely business decisions.

Unfortunately, many data warehousing solutions are unable to supply the features that companies need today. Traditional data sources, as well as electronic mail (email), electronic documents, articles, blogs and more, provide valuable information to an enterprise — yet this information often is not available for analysis through a data warehouse. With advances in Web and portal access technology, corporations are looking for ways to make both structured and unstructured information available to customers, suppliers, partners, and employees in order to foster greater collaboration, improve analysis, and find hidden data that can aid the decision making process.

In addition, databases are growing at a phenomenal pace, and many conventional approaches to data warehousing fail to provide the scalability and flexibility needed to grow and adapt to change. Indeed, the widespread use of the Internet and electronic commerce is causing systems to generate enormous amounts of data, with hundred of terabytes expected to give way to petabytes of information in the near future. As existing systems expand to accommodate growing data volumes, enterprises are feeling the strain of purchasing and managing large-scale systems that consume valuable datacenter floor space and are costly to maintain and power.

### A new approach to data warehousing

The Sun Data Warehouse Reference Architecture for Structured and Unstructured Data brings together industry-leading server and storage solutions with sophisticated database technology. High-performance and energy efficient Sun SPARC® Enterprise servers, underlying compression techniques from Sybase IQ analytics server and BMMsoft DataFusion work together to enable more data to be stored in less space and searched and analyzed in less time. Building on these components, the reference architecture can help take the cost, complexity, and risk out of deploying large-scale warehousing solutions.

## World's largest data warehouse solution

Sun created a data warehouse solution that pushes the limits of implementations. Based on the proven Sun Data Warehouse Reference Architecture for Structured and Unstructured Data, the solution supports over 1 PB of source data — over 34 times larger than the largest industry standard benchmark[1] and twice the size of the largest commercial data warehouse known to date[2]. Providing a massive solution in a small footprint, the data warehouse enables real problems to be solved in real time while reducing complexity and risk.

## Architected for massive data

As configured and tested, the data warehouse solution consists of one Sun SPARC Enterprise M9000 server and three Sun StorageTek™ 6540 arrays with a total of 250 TB of disk storage capacity that can handle over 1 PB of source data containing approximately six trillion SQL records and one billion e-mail messages, documents, and images.

To maximize data volume, all input data is compressed, enabling the solution to store more data in less space than traditional approaches. Indeed, the unique architecture of the system provides up to 90 percent storage savings versus other solutions processing the same amount of input data.

## A green solution for business intelligence

Using the latest eco-friendly Sun servers and storage arrays, the data warehouse solution delivers a high volume, high-performance, scalable architecture that uses 90 percent less storage, consumes 91 percent less energy, and generates less heat and carbon dioxide than conventional solutions.

Consider a comparable solution consisting of three servers and 29 storage systems that provide 1,500 TB of storage to house 1 PB of source data. Assuming maximally configured storage arrays each containing 176 disk drives that use 20 KWh of electricity to run and consume another 10 KWh for cooling, each cabinet requires 30 KWh of electricity to operate. Using 8.4 million KWh per year, these systems cost approximately $922K to run with electricity rates of $0.11 per KWh. Given that each KWh emits 1.34 pounds of carbon dioxide, such a solution could release over 5,000 tons of carbon dioxide into the air[3].

In contrast, the Sun data warehouse solution employs only a single, energy efficient Sun SPARC Enterprise M9000 server and three Sun StorageTek 6540 arrays. Together, these four systems require only 90 KWh of electricity to run. Using just over 788,000 KWh per year, the entire solution costs approximately $87,000 per year to operate with electricity rates of $0.11 per KWh. Consuming far less energy, the Sun data warehouse solution reduces carbon dioxide emissions to less than 480 tons — less than 10 percent of the pollution caused by traditional large-scale data warehousing solutions.

## Work with structured and unstructured data

The ability to quickly search and analyze a wide variety of information enables enterprises to mine more data and glean valuable information that can help provide a competitive edge. By taking advantage of innovative Sybase database technology, the Sun data warehouse solution provides the ability to seamlessly store, query, and quickly correlate structured and unstructured data.

## Store and search any kind of data

The data warehouse solution combines Sun, Sybase, and BMMsoft DataFusion products to provide text and media analysis for structured and unstructured data by consolidating transactions, email, documents, call center, voice, video, and multimedia data in a single database. The solution stores business email messages and attachments, and automatically extracts and stores text as searchable information. Integrated searching and cross-analysis across structured and unstructured data is possible on information stored in a single repository on Sun StorageTek systems.

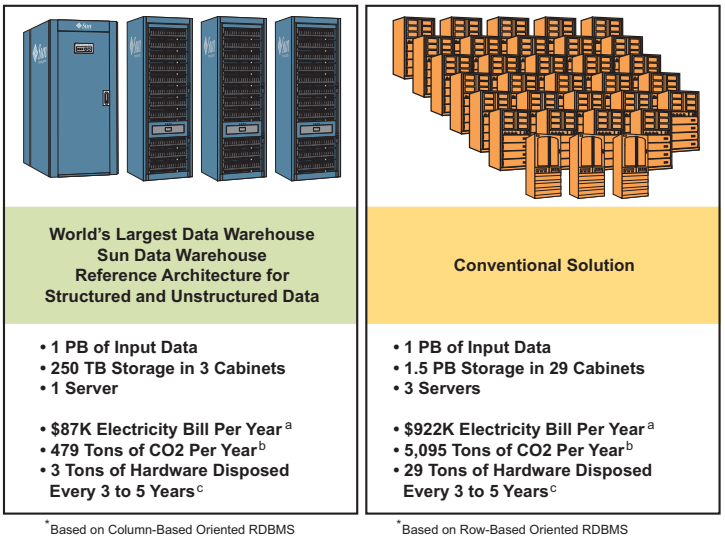


*Figure 1. The Sun data warehouse solution uses 90 percent less storage, consumes 91 percent less energy, generates less heat and carbon dioxide, and costs less than conventional solutions.*

## Analyze data faster with Sybase IQ

Unlike typical online transaction processing systems that often require optimization and tuning in order to run effectively, Sybase IQ is designed for high-performance analytics. Combining a column-based data structure with patented indexing and a scalable grid, Sybase IQ eases the process of adding and loading data, and makes it possible to perform analysis and reporting that were previously impossible, impractical, or cost-prohibitive. By integrating Sybase IQ, the data warehouse solution can deliver superior query performance with fewer hardware resources, enable more efficient data compression, and simplify maintenance and tuning efforts.

## Take advantage of massive scalability

With growing volumes and different types of information now able to be stored, searched and analyzed, data warehouse solutions must be able to process more workloads faster. The current data warehouse implementation is based on a previous version of the reference architecture that was independently verified to be able to simultaneously maintain query response time performance and data loading speed while increasing the query submission rate by nearly 500 percent[4].

Key elements in the reference architecture provide the scalability needed to process growing volumes of data faster:

- Sun SPARC Enterprise servers are reliable, vertically scalable systems that provide the benefits of mainframes without the cost and complexity. With symmetric multiprocessing scaling from one to 64 processors, memory subsystems as large as 2 TB, and fast I/O architectures, Sun SPARC Enterprise servers can help speed data searches and analyses by adding processing power as needed. Workloads can be spread across multiple servers to increase processing capacity, and support for dynamic domains enables massive consolidation and virtualization.
- The scalability inherent in Sun's server design, as well as Sybase IQ multiplex technology, offers near linear user and data scalability to support thousands of users and petabytes of data as additional storage and computing resources are added to the system.
- The ability to run multiple Sybase IQ readers across several servers and system domains using Dynamic Reconfiguration technology enables compute resources to be made available as needed, increasing asset utilization and processing power. A multinode, shared storage, parallel database system, Sybase IQ represents multiple instances of Sybase IQ engines running on a number of servers connected to a shared data repository. Each instance can access the entire database. As a result, organizations can put several Sun servers to work searching, processing, and analyzing data in parallel to perform more queries in less less time and increase query performance, without compromising data integrity.

## Ensure reliable operation

Companies depend on being able to access and analyze information stored in a data warehouse at any time in order to make faster and better business decisions. By integrating sophisticated technologies from Sun and Sybase, the data warehouse solution can help companies create an environment that runs 24x7. Indeed, Sun's large-scale data warehouse solution can be loaded in real time, enabling users to continue using the system even while data is loaded.

Deployed on a Sun SPARC Enterprise M9000 server running the Solaris™ Operating System, the data warehouse solution can take advantage of thousands of commercial off-the-shelf applications, Dynamic Reconfiguration technology, live upgrades, highly available patching, a hardened kernel and device drivers, and hot-plug and remote service capabilities to help ensure the system stays running even in the event of a failure.

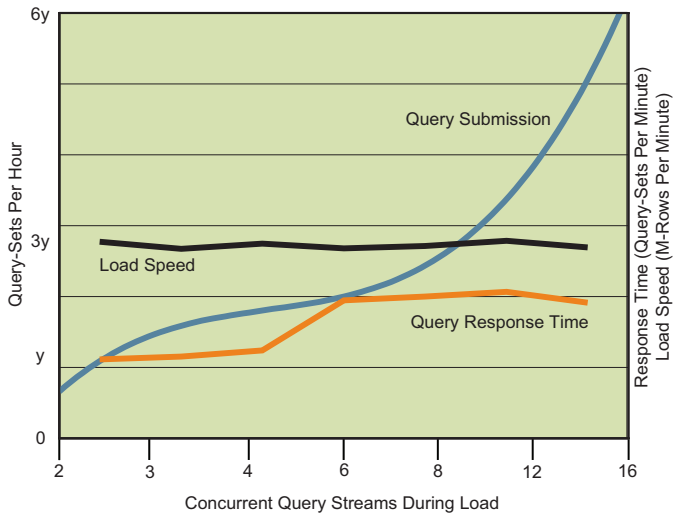


*Figure 2. The Sun data warehouse solution can scale to meet demand and perform more queries in less time.*

In addition, previous versions of the reference architecture demonstrated high availability by using the NonStopIQ method of Sybase IQ and a variety of storage services, including shadow and mirrored images, to foster data integrity and availability. Indeed, the NonStopIQ method of Sybase IQ offers full database backups that complete in seconds or minutes, near instantaneous application failover from primary to shadow databases with minimal disruptions and downtime, and the ability to off-load maintenance and backup tasks to shadow devices to keep production performance at peak levels.

## Satisfy regulatory requirements for record retention and retrieval

Many regulatory directives require data to be retained for years. The data warehouse solution can help enterprises store and retrieve growing amounts of data and comply with regulatory directives. With the ability to scale to 168 TB in a small footprint and stream data at 4 GB/sec, Sun StorageTek 6540 arrays excel at handling large data sets when fast access is key. Write Once, Read Many file system capabilities in the Sun StorageTek QFS software can consolidate, share, and store business data on unalterable media to ensure information is not tampered with or deleted.

In addition, Sybase IQ lets databases and devices be configured as read-only, and provides database and column-level encryption for added security. Enterprises can also deploy the Sun StorageTek 5320 NAS Appliance and Sun StorageTek Compliance Archiving Software which provide per-file retention periods and WORM protection, data mirroring, auditing, integrity checks, and more.

## Sun Reference Architectures

The Sun Data Warehouse Reference Architecture for Structured and Unstructured data is designed, documented, tested, and tuned so that customers can accelerate time-to-revenue as well as reduce the complexity, costs, and risks of deploying new technology in the enterprise. Sun Reference Architectures include:

- Recommended products from Sun and its partners
- Technical guides that provide architecture, sizing, and implementation information

Learn More

To learn more about the Sun Data Warehouse Reference Architecture for Structured and Unstructured Data and its components, visit sun.com/service/refarch, sybase.com, bmmsoft.com, or contact your Sun representative. For a prototype of this Reference Architecture, visit the Sun Solution Centers at sun.com/solutioncenters.

## Putting the reference architecture to work

Easily tailored to suit a wide range of needs, Sun's carefully designed reference architecture approach helps companies to implement large-scale data warehousing solutions quickly and gain access to valuable information. With a massively scalable infrastructure, enterprises can store more data, analyze information faster, and ensure regulatory compliance requirements are met while reducing risk.

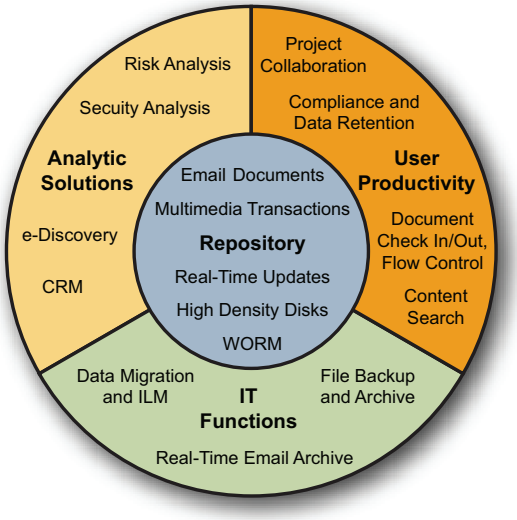


*Figure 3. The Sun, Sybase, and BMMsoft solutions enables a host of data warehousing solutions.*

1. See http://tpc.org/tpch/results/tpch_last_ten_results.asp as of July 9, 2007.
2. See "At Wal-Mart, World's Largest Retail Data Warehouse Gets Even Larger" at http://eweek.com/article2/0,1895,1675960,00.asp
3. See http://www.eia.doe.gov/cneaf/electricity/page/co2_report/co2report.html
4. *Performance Benchmark Report, One Trillion Rows, Sun-Sybase DW Reference Architecture*, InfoSizing, June, 2004.

a. As of August, 2007, the electric rate of $0.11 per KWh is used based on the Energy Information Administration at http://eia.doe.gov.
b. Typical pollutant factor of 1.34 lb. of CO2 per KWh (http://www.eia.doe.gov/cneaf/electricity/page/co2_report/co2report.html)
c. Based on the average weight of the storage and server equipment that would be needed.

**Sun Microsystems, Inc.** 4150 Network Circle, Santa Clara, CA 95054 USA **Phone** 1-650-960-1300 or 1-800-555-9SUN **Web** sun.com


SunWIN #: 509993 Lit #: SYSB13178-0 08/07