INFO SIZING

Technology Evaluation Report

BMMsoft EDMT® Server

Building a

Petabyte Data Warehouse

Using the

Sun Data Warehouse Reference Architecture

with

Solaris™ 10 OS, Sybase® IQ

Technology Evaluation
BMMsoft EDMT® Server

Executive Summary

Businesses are constantly looking for ways to analyze information and make decisions in a timely manner. Local governments and regulatory agencies mandate upon businesses that they maintain vast amounts of historical data about their activities and their communications, both internal and external. Data warehouses are rapidly becoming mainstream and are providing a vital tool upon which enterprises rely in order to comply with mandates and make timely business decisions. Ideas and techniques for handling and using data warehouses have evolved since the concept emerged, just over a decade ago. To survive and prosper in today's hyper-competitive global market, businesses must adopt and rapidly integrate advanced data warehouse technologies.

While data warehouses are used differently across businesses and industry sectors, the need to consolidate and correlate structured and unstructured data is becoming commonplace. Many if not all the data sources within an organization are recorded in the data warehouse and, as a result, data repositories are growing at a phenomenal pace. As data volumes grow, enterprises are looking for ways to store more data in less space and to analyze more data in less time, while meeting the need for optimal datacenter floor space and reduced energy consumption.

In this environment where multiple, apparently conflicting challenges must be solved, **BMMsoft EDMT® Server** is providing a combination of state-of-the-art technologies that address these challenges. These technologies include the ability to create a unified data warehouse holding structured and unstructured data, to store this universe of information in a highly compressed form, to correlate and analyze this data without structural boundaries, and to achieve this with less computing resources and in less time than conventional technologies.

The capabilities of the real-world solutions provided by **BMMsoft EDMT ® Server** were formally tested during a **Guiness World Record**[1] demonstration involving over a petabyte (1,035 TB) of actual business data.

[1] Guinness World Record, World's Largest Data Warehouse, Sybase IQ
http://www.sybase.com/guinness

125 West Monroe Street • Colorado Springs, CO 80907 • 719-473-7555 • www.sizing.com

Real-World Capabilities

The real-world capabilities of the **BMMsoft EDMT ® Server** were showcased during the independently verified[(2)] creation and querying of a **Petabyte Data Warehouse**.

This demonstration was based on the **Sun Data Warehouse Reference Architecture,** powered by a Sun Enterprise M9000 server running Solaris™ 10 OS and connected to three Sun StorageTek™ 6540 arrays. The **BMMsoft EDMT ® Server** was driving a Sybase® IQ analytic server.

The following significant achievements were demonstrated:

- It demonstrated its capabilities for **quick deployment** based on the simplicity and self tuning ability (few "knobs") of the system. Once the hardware was operational, it took all of 3 days to build an environment ready to accept a petabyte of data, most of which was spent initializing storage partitions.
- It loaded **1 Petabyte of raw transactional data** (6 Trillion stock quote records) in a fully indexed star schema; creating a new, independently verified record for the **world's largest data warehouse**.
- It reached a **load speed of 285 billion rows per day** (3 Million rows per second) for "***T***" (Transactional) data and sustained that database population pace for a period of over 3 weeks.
- It reached a **load speed of 67 million documents per day** for "***EDM***" (Emails, Documents and Multimedia) data. In total, 185 million mail messages, attached documents and multimedia files (over 72 Terabytes of data) were loaded in less than 3 days while consuming a fraction of the available CPU power, leaving over 93% of the M9000's processing power for other activities.
- It achieved an **average Ready-Time of less than 2 seconds** for data freshly added to the data warehouse.
- It showed an **85% data compression ratio** by storing a Petabyte of raw transactional data in less than 260 Terabytes of actual disk space. The substantial reduction in the number of disk drives needed for storage translates directly into at least **90% reduction in CO_2 emission** over the lifetime of the **BMMsoft EDMT ® Server** platform.

[(2)] Performance Sizing Report, Petabyte Data Warehouse, February 29, 2008, InfoSizing http://www.sun.com/service/refarch/datawarehouse/WLDWAuditReport.pdf

125 West Monroe Street • Colorado Springs, CO 80907 • 719-473-7555 • www.sizing.com

Real-World Solutions

BMMsoft EDMT ® Server physically collects a company's unstructured information (documents, emails, multi-media attachments, blogs, instant messages, etc.) into a single data store. The resulting Unified Data Warehouse (**UDW**) can be accessed through a real-time query view of the entire information space.

During the querying of the independently verified **Petabyte Data Warehouse**, the **BMMsoft EDMT ® Server** was used to produce answers to real-life business questions, most of which could not be answered by more conventional enterprise data warehouses.

Following are some of the record breaking queries that were executed during the **Petabyte Data Warehouse** demonstration, and how they apply to other real-world scenarios.

- **Mining Unstructured Data**

  The "Popular Stocks" Query: Mining the unstructured content of the Unified Data Warehouse, this query identifies which securities are most discussed among traders. All correspondences, such as email, blog entries or instant messages are searched to determine which securities were the most frequently mentioned on a particular day.

  While the context of the test was borrowed from the financial sector, this query demonstrates that such a data mining tool can provide a vast array of new opportunities. During this test, new emails were loaded in the UDW while the queries were executing. The delay between the insertion of a new email and its discovery by a query, also called the "ready time", was found to be less than 2 seconds. This ability to simultaneously load large amounts of new unstructured data while querying in real-time the growing UDW is one of **BMMsoft EDMT ® Server**'s strengths.

  Real-World Solutions: A similar query could identify "insider trading" by finding securities for which a prevailing recommendation to "buy", "hold" or "sell" was discussed prior to the company's financial report. In another business environment, this query could be mining millions of emails and blog entries looking for the source of a leak or for the origin of a rumor affecting the company's reputation in the marketplace. In the context of litigation, the complete record of all electronic communications from the parties involved could be searched to help bolster the cause of action or uncover the "smoking gun". And in a homeland security context, on-line postings world-wide could be loaded in real-time into the UDW and similar queries could help identify terrorist plans or money-laundering activities.

125 West Monroe Street • Colorado Springs, CO 80907 • 719-473-7555 • www.sizing.com

- **Mining Transactional Data**

  The "Portfolio Growth" Query: Mining the transactional content of the Unified Data Warehouse, this query determines the hypothetical growth of a portfolio over the past year. The query targets ten securities and assumes the following trading strategy: when the 20-day moving average of a security crosses over the 5-month moving average, a tenth of the portfolio is invested; and when the 20-day moving average crosses below the 5-month moving average, the position is sold.

  Beyond its real-world applications, this query demonstrates two of **BMMsoft EDMT ® Server**'s strengths. It shows the ability to concurrently process and answer multiple streams of queries (up to 50 streams in the test). And, more importantly, it keeps responses times at just a few seconds, providing answers in real-time. This was all achieved despite the query's very high degree of complexity (130 lines of standard SQL) and the 1,000 terabytes (one petabyte) of real data being searched. While part of the credit for this achievement belongs to the underlying Sybase® IQ engine, it is **EDMT ® Server**'s unique ability to exploit IQ's querying capabilities that made it possible.

  Real-World Solutions: A similar query could be most useful in times when billions of dollars are being spent on complex government programs with inadequate tracking and auditing. The query could be applied to the detection of fraudulent investment schemes (such as Madoff's) by verifying that paid returns are the result of gains derived from actual purchases and sales within existing portfolios. In another example, the transactional data for orders between car manufacturers and their part suppliers could be loaded in the server. The capabilities demonstrated by this query could then be used to identify healthy areas of the industry that can be best rescued by an influx of capital, or areas where failures are chronic, decay is unavoidable and new funding would be wasted.

- **Mining EDMT Data**

  The "Secret" Query: Mining the complete set of transactional (T) and unstructured (EDM) data from the Unified Data Warehouse, the query determines if there is any correlation between quotes (BIDS) tendered and emails urging secrecy about the same stock.

  Aside from its practical applications in the financial world, this query exemplifies the search for the proverbial "needle in the haystack". It shows **BMMsoft EDMT ® Server**'s ability to correlate billions of emails, attachments, blog entries and instant messages with trillions of conventional business records. This combination of two structurally different data universes to produce real-time answers to highly complex queries is the key to creating a state-of-the-art Unified Data Warehouse.

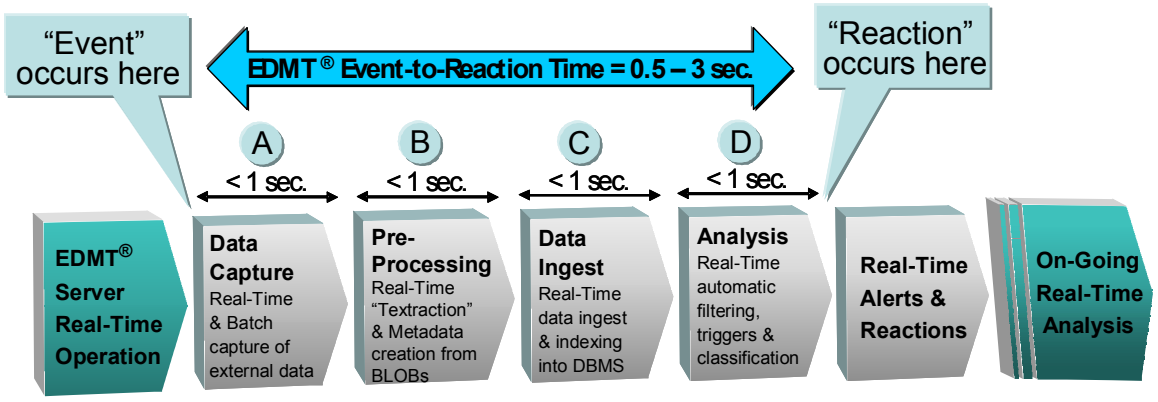
125 West Monroe Street • Colorado Springs, CO 80907 • 719-473-7555 • www.sizing.com

Real-World Solutions: A similar query could draw a correlation between the level of increase in the sale of a new product and the number of on-line postings containing a positive mention of that product. Conversely, the ability of the server to include new on-line postings and new sales data in the query in real-time could give a consumer product vendor an early warning about a manufacturing defect with a potential for a product recall. In another example, EDMT cross-analysis could make real-time use of blogs, emails and other on-line "chatter" to detect early "rumors" and "threats" and correlate them with transactional data (i.e., stock trades, purchases of chemicals or shipments of weapons and ammunitions). By effectively integrating all modern communications and social networking into the Business Intelligence (BI) framework, EDMT cross-correlation could provide fact-based early warning of "hard-to-catch" events representing potential threats of terrorist attacks, criminals operations or financial meltdowns.

Real-Time Analysis – Critical in Today's World

Favoring the prevention of problems over repairing the resulting damages is common sense. Applying common sense to real-time processes and events requires a new breed of tools, a new class of real-time applications that can detect events, problems or opportunities and react, prevent or exploit them in real-time.

EDMT ® Server leads this new breed of tools. Its real-time claims were demonstrated by measuring between 0.5 and 3 seconds of latency between "Event" and "Reaction".



EDMT ® Server's real-time latency between "Event" and "Reaction"

The diagram above shows that during the short, real-time latency, data is (A) captured (B) pre-processed, (C) ingested, (D) analyzed and, if warranted, data is acted upon.

125 West Monroe Street • Colorado Springs, CO 80907 • 719-473-7555 • www.sizing.com

A New data is accessed once by **EDMT ® Server**. The data is then captured in real-time, and no additional access to the original data will be needed. Unlike other federated data solutions that have to keep accessing the "source" every time the data is needed, **EDMT ® Server** maintains and relies on its own record of the data.

B For many data types, the pre-processing done by **EDMT ® Server** prior to ingesting the data into its own data store consists in generating meta-data about the data source. This process of "textraction" is extremely efficient and designed to accommodate new data types easily without making changes to the structure of the underlying data store. A new "textractor" can be added at any time to respond to the introduction of a new data type, thus making the process "future-proof".

C Once done with pre-processing, **EDMT ® Server** inserts the new data into its own data store, alongside the hundreds of terabytes of already captured data. This ingestion process includes updating all tables and indices affected by the new information, thus optimizing the access paths for future use of the data.

D On a on-going basis, **EDMT ® Server** performs complex, real-time analysis of the vast amount of data in its data store. This real-time analysis of information of diverse data types, coming from multiple sources and correlated with hundreds of terabytes of existing data is particularly critical when dealing with applications that cannot afford less than real-time, such as fraud detection, security threat (e.g., military, climatic or terrorist), medical emergency, product recall, etc.

Full Disclose: Francois Raab is an independent industry consultant and certified benchmark auditor who also serves as a member of the BMMsoft Advisory Board.

125 West Monroe Street • Colorado Springs, CO 80907 • 719-473-7555 • www.sizing.com