HRG Assessment: Mastering Change
The Big Data Solution (EDMT, Sybase IQ, and DL980)

Change is a constant and the rate at which change occurs is increasing. The business environment is changing - helping some - hurting others. The challenge is to find opportunity in change, take advantage of change to prosper, and learn to deal with broader based competition, global demographics, and continual change in regulations.

The current business environment is dynamic, highly competitive, and increasingly fast paced. Decision makers have little time for reflection and even less time for making effective decisions on best available information.

Big Data integration and analytics is a required part of any viable business strategy looking to overcome the challenges of managing an environment where increasing rates of change and business model innovation are the new normal. In the current business environment for a strategy to be effective it must recognize the importance of Big Data and include an investigation of the requirements to ingest, index, and integrate structured and unstructured, streaming and static data from a variety of sources.

A few things to consider when developing a Big Data Strategy:

➢ Data Variety - structured, unstructured, and real time.

➢ Data Volume - terabytes and petabytes of data and billions of transactions or files.

➢ Data Velocity – can you ingest and store the volumes of data at the rate required.

➢ Choice of database –business critical Big Data should be stored in an ACID-compliant repository.

➢ Running analytics and searches on (and across) structured and unstructured data from multiple sources.

Good information based on best available data is central to analytics, planning, and decision-making. The key to developing and implementing a Big Data strategy is understanding the business, knowing what data is available, defining what information is needed, and what results are desired. The ability to quickly and effectively respond to competitive and market place change is only one of the benefits of implementing an effective Big Data strategy.

With change being the only thing you can count on managers and executives need real time interactive decision support to model "What if" scenarios that leverage Big Data. Recognizing the importance of Big Data is significant for ensuring management and executive effectiveness at a time when change is the new normal. Predictive analytics run on static and streaming data provides real competitive advantage. The combination of BMMsoft (EDMT), Sybase IQ, and HP's DL980 represents a high value proven solution for the real time use of Big Data to gain maximum impact through actionable results.

Big Data Sources

Executives understand the benefit of having access to the complete range of corporate data without having to be exposed to its underlying complexity. Today, no data, regardless of type, location, or format is beyond integration. It is the nature of businesses to grow and to accumulate data, and those businesses that interpret historical business and market data accurately and apply that knowledge to the analysis of current business and market data in the shortest elapsed time will tend to be market leaders. As companies collect increasing volumes of data from their business operations, websites, marketing initiatives, sales campaigns, and other sources (Big Data) keeping track of that information and making it useful to the business through data analytics can provide significant distinctive competitive advantage. Some examples of the sources of Big Data follow:

- ➢ The continually increasing volumes of data and transactions originating from personal mobile computing devices including smart phones, tablets, and other untethered mobile computing devices.

- ➢ Social media sites used for sharing ideas, opinions, images, documents, and promoting products and services.

- ➢ Graphics applications, imaging applications, as well as, continually improving presentation and visualization capabilities are moving into the main stream. The use of virtual reality and remote visualization applications will become more wide spread. These types of applications will increasingly be used for the visualization of complex data and data interrelationship modeling for interactive real time analysis and "what if" scenario testing in certain sectors of industry and government.

- ➢ Imaging in healthcare, life sciences, petrochemical, energy, insurance, social networking, on-line advertising, media and on-line publishing, and new product design and development is driving significant data growth. The higher the image resolution the larger the image file size. Complex documents and e-mails with imbedded images and attached images contribute to the volumes of data being created.

- ➢ Emerging technology developments facilitate organizational and business change, drive productivity, increase creativity, speed new product development, and create data.

- ➢ Person-to-person Web, e-mail, and SMS based interactions, and device-to-device transactions create data.

- ➢ Astronomical observations and imaging, NOAA weather data and imaging, CERN experiments, SCADA infrastructure sensors, ATM cameras, TSA cameras, traffic monitoring cameras, and other surveillance devices create data.

Big Data Benefits

The effective use of Business Intelligence (BI) and business analytics is a key factor in driving business success. Each business must identify its own unique BI requirements and data sources as part of their business planning, strategic review, and self-assessment. Information management, BI, and business analytics software gather, store, access and analyze corporate data. Bringing Big Data into the mix dramatically impacts the use of business analytics as a strategic and tactical tool as well as the results produced. When run against the best available data these software applications enhance customer profiling, custom support, market research, market segmentation,

product profitability, statistical analysis, inventory control, order fulfillment, and overall customer satisfaction and retention.

Once the required data sources have been identified the data needs to be ingested, integrated, indexed, and archived. This data in its original form is likely to be a mixture of structured, unstructured, real time, static, e-mails, documents, multimedia (video, voice, and images), database transactions, SMS, and sensor data. After this has been accomplished you can search, run analytics, and produce results based on this body of heterogeneous data and begin to realize the true benefits of Big Data analytics such as:

➢ Identification of new business opportunities

➢ Better product innovation, shorter development cycles, and reduced time to market

➢ Better focused / more effective sales and marketing initiatives

➢ Better targeting of prospective customers

➢ Improved customer satisfaction

➢ Improved customer retention

➢ More accurate and timely risk assessment and risk avoidance

➢ New market / new opportunity identification

➢ Reduced risk from spurious litigation

➢ Better regulatory compliance

Risk Management

In the financial services and insurance industry sectors risk management is about consciously taking the risks we want to take, for a fair price, without taking on too much risk or being blindsided by the unforeseen. Risk management helps a firm maximize its return and general prosperity. Poor risk management practices can quickly lead to ill-conceived financial commitments, undetected fraud, insolvency and collapse. Vital to business success and mandated by regulatory law, risk management is a firm wide, board-level concern.

Enterprise risk management or corporate risk management is focused on investigation, fraud examination, and identification of security threats, intellectual property loss, environmental risks, product deficiencies and many more potential areas of enterprise or corporate liability exposure.

Counter Terrorism

The National Counter Terrorism Center (NCTC) has at its core the requirement for the integration and safe sharing of data regardless of hardware platform, operating system, or data format. The NCTC drives the integration of federal IT resources and data across agencies and departments of the federal government. The result is the delivery of actionable information and alerts when and where needed to produce desired outcomes.

Genome Sequencing, Electronic Medical Records, and Personalized Medicine

In genome sequencing the raw "DNA base" data produced by the sequencing appliance results in the generation of extreme volumes of data (multiple terabytes), that must then be stored for further analysis. After the assembled and aligned genomic image data has been stored, secondary analysis and research on the data can be performed. This area of secondary analysis is a key focus for researchers and technology providers.

The field of personalized medicine relies on the integration and analysis of genomic sequencing and clinical history data. This data comprises medical imaging data, text, e-mails, documents, database transactions and genome sequence data. All of these data types in combination comprise what is known as Electronic Medical Records (EMR) and contribute significantly to the effectiveness of personalized medicine. Security, privacy, disaster recovery, availability, and other emerging concerns need to be addressed to pave the way for fulfilling the promise that personalized medicine holds.

Personalized medicine is based on the combination of personal, historic, clinical, and genomic data (when available) in a single source. This single source could take the form of a machine-readable card, subcutaneous RFID chip, or some other device that an individual could carry with them. This device will contain the individuals' unique genome encoded along with other relevant clinical information. When this individual goes to a pharmacy, for example, a pharmacist will be able to cross check the encoded genome, clinical historical data, and existing prescriptions in order to ensure that any new prescription does not conflict with existing conditions or treatments. This type of information will be a game changer in terms of enhancing the physician's effectiveness. This in turn will facilitate proactive prevention and treatment through life style counseling, personalized medicine, and even unique custom personalized drug therapies.

eDiscovery

Electronic Discovery refers to preservation, collection, review and production of electronic records in connection with litigation or regulatory requirements. Specific rules regarding the production of Electronically Stored Information ("ESI") have been established. The federal rules (and their state counterparts) require potential litigants to have a clear and comprehensive understanding of their ESI so that they can search and produce such ESI if necessary as part of the litigation process. ESI can include image data, audio files, video, e-mails, documents, database transactions, SMS, and text. If a litigant does not know where its ESI is stored, and has no consistent document retention policies this adds unnecessary expense, lengthens the time necessary to comply with discovery requirements that typically demand fast production times. The lack of ESI and consistent document retention policies has in some cases resulted in sanctions and the award of substantial punitive damages.

Audit and Compliance

An audit is the review and examination of a company's financial statements, assets and liabilities and may include a physical audit. The auditor must report his findings based upon the evidence examined expressing an opinion on and providing credibility to the financial statements. In order to facilitate the audit process thereby limiting related costs one approach is to establish a document retention policy and establish an electronic stored information archive. By anticipating an auditor's information requirements and then archiving that data the time and labor required to perform a financial audit may be reduced and controlled. This same line of reasoning applies with regard to industry and government regulations and providing requisite proofs of compliance. This level of preparedness will result in the reduction of time and costs required to support either a financial audit or a regulatory compliance audit. The type of data captured and archived will be similar to that required for eDiscovery.

Big Data Solution

The combination of BMMsoft EDMT® Solution, Sybase IQ, and HP DL980 represents a significant partner based solution in meeting the Big Data challenge. The Big Data challenge is ingesting, indexing, archiving, managing, and analyzing high volume, high velocity, structured, unstructured, static, and streaming data from an increasing variety of sources and locations. The level of compatibility and fit between the three products is impressive. From

everything we have read and based on a thorough review of all publically available materials augmented with a series of open ended discussions with each partner and selected customers this team play gets very high marks.

BMMsoft's use of Multiplexing fits hand in glove with Sybase IQ PlexQ™ technology and both parties benefit from the full mesh connectivity between EDMT and Sybase IQ. Both BMMsoft and Sybase provide products with high data compression quotients where substantive data compression of 80% is the norm.

BMMsoft EDMT delivered record setting mixed data ingest speeds of more than 14 TB per hour running on an HP DL 980 8-socket, 8-core system for a total of 64 cores. During recent discussions it was disclosed that this particular DL 980 was straight out of the box with no special tuning and no virtualization. The only changes made that are worth noting were the setup of LUNs for the SAN and the addition of multipath software, aside from that this was pretty much a pure vanilla DL980. BMMsoft told us that the "'standardness' of RHEL6 on the HP DL 980 made the installation of EDMT/Sybase IQ a breeze - it worked out-of-the-box and we used the standard configuration file for 'big box'."

When EDMT is coupled with Sybase IQ ingesting real time streaming market tick data, archiving, and then running in database analytics the performance and throughput delivered is a testament to this close partnership. In addition, normal business level queries typically return results in the sub second range while highly complex queries can return a result in from 1 to 300 seconds.

EDMT is able to deliver when it comes to enterprise wide searches across disparate mixed data types due to its high speed Extract, Transform, and Load (ETL) and indexing capability. Using this solution virtually any type of electronic data regardless of format or location can be ingested, indexed, and queried in close to real time. When ingesting streaming data, such as financial market data, EDMT does not perform the extract and transform operations which means that even higher rates of data ingest can be realized making ingested data available for search, query, and analysis in close to real-time.

Because both BMMsoft and Sybase IQ use column data stores there are significant benefits to be had in terms of data compression, query, analytics speed, and data security. It is inherent in the design of most column oriented data bases that even if a hacker gained access to the raw data store they would be unable to make sense of the data without having access to the index. Without the index the data would appear as a series of vertically shredded bits in much the same way that a paper shredder makes printed documents indecipherable.

The BMMsoft EDMT® Solution

In 2007 BMMsoft EDMT stored 1 petabyte of data setting a Guinness record, and in November 2011 EDMT ingested 14 Terabytes per hour on an HP DL980.

The name EDMT stands for Emails, Documents, Multimedia, and database Transactions. The BMMsoft EDMT® Solution in combination with Sybase IQ 15.4 running on the latest generation HP ProLiant DL980 delivers the performance, reliability and scalability required for real-time ingesting, indexing, accessing, and cross analysis of extreme volumes of heterogeneous Big Data.

Both EDMT and Sybase IQ are enterprise-ready Big Data solutions that comply with the ACID rules for database architectures providing high-quality data that is critical for enterprise applications and reliable verifiable analytic results based on clean and complete data. Furthermore, EDMT uses unadulterated or "accent-free" SQL that ensures precise data definition, ACID compliant data relationships, and precise queries results.

The EDMT® Solution ingests, indexes, and transforms structured and unstructured data using up to 3 million independent data processing channels that can then be stored and managed by a single analytic data repository such as Sybase IQ. The EDMT® Solution automatically creates and stores metadata for the ingested content. The BMMsoft EDMT Universe connector for BusinessObjects, lets BusinessObjects users access, view, and analyze all of the structured and unstructured data in the EDMT archive.

The BMMsoft EDMT® Solution is a standalone application specifically designed for ingesting, indexing, archiving, and searching of large volumes of structured and unstructured data. Using the EDMT® Solution businesses can run real-time text analysis, SQL analysis, and cross-analysis of mixed data types. With the EDMT Solution near real-time monitoring, instant cross-analysis of new and historical data and real-time response to market changes, product problems, customer dissatisfaction, litigation, audit, fraud threats, and competitive threats can all be handled from a single system with no impact on production systems. Using EDMT customers can capture, index and store data, and enforce data retention policies to meet EIS and regulatory compliance requirements while leaving the source data in its original state.

EDMT's high availability features support the implementation of active-active disaster recovery sites. Remote replication to multiple sites ensures that problems with one or more replication targets or channels will not impact replication. Replication to sites that have "fallen behind" will automatically resume from the failure point and eventually catch up with the working sites. According the BMMsoft the latency between the primary site and the secondary site(s) can be configured to be as low as 2 seconds.

EDMT captures e-mails (with attachments), files and database transactions in real time. According to BMMsoft most basic searches return results in under one second, and complex cross analyses are typically completed in 1-300 seconds.

EDMT delivers up to 90% data compression rate according to BMMsoft thereby reducing storage requirements, and energy use in contrast to traditional RDBMS solutions. Furthermore BMMsoft states that for $1 EDMT can store, search and cross-analyze 10,000 e-mails or 1 million SMS (short message service) text messages or 5 million (150-byte-wide) database records and that the cost of archived data is approximately $500 per terabyte.

Sybase IQ 15.4

With Sybase IQ 15.4, Sybase introduced a native MapReduce API, Hadoop integration, Predictive Model Markup Language (PMML) support, and an expanded library of statistical and data mining algorithms that leverage the power of distributed query processing across a Massively Parallel Processing (MPP) grid based on Sybase IQ's PlexQ™ technology.

New Sybase IQ APIs enable the implementation of proprietary algorithms that run in-database. Sybase claims that running proprietary algorithms in database delivers greater than 10 times the performance acceleration as compared to existing approaches. Additional improvements have been made for text data compression and bulk data loading interfaces.

Column vs. Row oriented

Highly complex query environments that support strategic and operational decisions can be used to gain a competitive edge by better understanding customers, competition, risk positions, revenue leaks, and fraud. Sybase IQ is a column oriented database that provides significant storage compression, query speed and performance advantages when compared to traditional row oriented databases. Column oriented DBMS outperform traditional row based database management systems on average by a factor of 100 times. Real-time data analysis can be used

to grow the business and investigate additional sources of revenue. What if'' scenarios and ad hoc queries can be run on the data to quickly examine trends, provide analysis, and gain significant competitive advantage.

Sybase PlexQ™

The Sybase PlexQ™ and BMMsoft EDMT Multi-node designs are very tightly coupled. The Sybase PlexQ™ shared disk / shared database is made possible because Sybase IQ is not a partitioned database. This provides exceptional scalability and flexibility. With a Sybase PlexQ™ you can add storage without being required to add additional servers just as you can add servers without having to add storage. This architecture allows the addition of servers of any size to the PlexQ™ grid. In contrast shared nothing MPP does not offer this level of configuration flexibility.

Diagram of Sybase IQ PlexQ™ technology
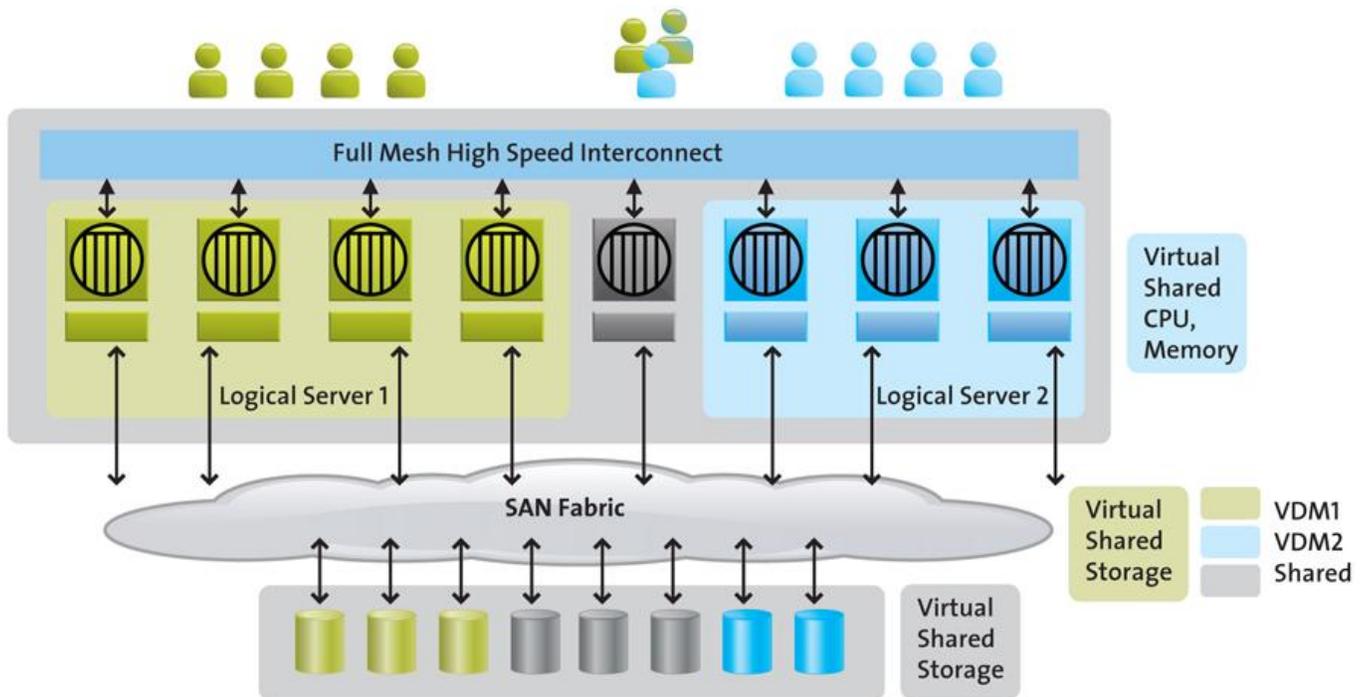
Scalability

An additional benefit of this PlexQ™ grid architecture is that you can scale the system as required (scale on demand). The fact that you can add servers, mixing servers of different sizes, in PlexQ™ grid provides a linear increase in performance for data loading (ingesting), running searches, queries, and analytics. With Sybase IQ a single query can be spread across any number of PlexQ™ grid nodes. Sybase PlexQ™ delivers practically unlimited horizontal scaling without any restrictions related to data partitioning.

Flexibility

EDMT Solution and Sybase PlexQ™ both use MPP "shared disk" architecture. This means that each server can see all of the shared storage in Sybase IQ because data is not partitioned or assigned to individual servers. Adding or removing servers ("nodes") with Sybase PlexQ™ is straight forward not requiring data reorganization or re-partitioning. Servers can typically be added or removed from the Multiplex in one minute or less. Servers are

connected to the Fibre Channel Switch and thereby gain access to all LUNs (storage units) used in the IQ Multiplex. A single query in either EDMT Solution or Sybase IQ PlexQ™ grid can be spread across all nodes, or a predefined subset of all nodes, or use just one server. This functionality provides substantial flexibility in terms of workload management and user isolation. All available nodes can be regrouped within 1 second (according to BMMsoft) to provide support for a big query.

Availability and Disaster Recovery

The nonstop functionality of both Sybase PlexQ™ and EDMT Solution is closely tied to their multiplex capability. With regard to availability and fault recovery, if any node in a Sybase PlexQ™ fails or is taken off-line for any reason the shared data store is not affected. All of the "shared data" is accessible as long as there is at least one Sybase PlexQ™ node functioning. In the event of a node failure or replacement there is no requirement for the admin to take any action to reestablish access to "lost data" because all of the nodes in the Sybase PlexQ™ MPP environment "see" and access all of the shared data directly.

For disaster recovery the disaster recovery site would have 100% of the data that the primary site has. EDMT is priced per-core and not per-TB therefore the cost of a single server disaster recovery site should be minimal. The EDMT Multi-site Replication feature is used to replicate all shared EDMT data to the disaster recovery site.

HP DL980

When you factor in the latest HP DL980 servers this technology partnership gets even better. HP's new DL980 features either 4 or 8 Intel 10-core CPU's. To put this in perspective consider the fact that the record setting ingest rate of 14 TB per hour was set running EDMT on 8-core Intel chipsets and not the newer 10-core engines. Based on recent discussions with the vendors and on our own analysis we expect that if the same test were run on the current generation DL980 that we should see a 20% improvement. When this test is run again it will be interesting to discover how conservative our current assessment actually is.

The HP DL980 with ten-core Intel® Xeon® Processor 7500/6500 Series gains substantially from the new PREMA architecture which incorporates design elements of the HP Superdome architecture with Smart CPU Caching, dynamic QPI routing, and a redundant system fabric. The new DL980 can support up to eight processors (80 processor cores) and 2 terabytes of memory (128 DDR3 DIMM slots).

The HP DL980 is highly scalable and appears to be well suited for most enterprise class and HPC type workloads. With HP Integrated Lights-Out 3 (iLO 3) server management software server a level of manageability, scaling, resiliency is added that is a good fit for most enterprise class compute environments. On the high end of the EDMT spectrum, with 4K and 16K systems, the HP DL980 is the recommended "building block" when 10s and 100s of enterprise-grade servers are needed.

With the addition of Red Hat Enterprise Linux 6 (RHEL6) the HP DL980 benefits from RHEL6's enhanced performance, reliability, scalability and security not to mention the Red Hat worldwide, mission-critical 24x7x365 support.

The storage used for this BMMsoft, Sybase, and HP Big Data team play is the HP P2000 set up to run in a SAN environment. For this solution offering each HP P2000 is set up as a RAID-6 device with 12-terabyte capacity.

The maximum number of nodes that can use a single Sybase IQ Multiplex MPP is 12,000. When new storage is added all nodes will immediately see the entire shared database. Storage devices can be of different type, disk size, and disk speed. With the BMMsoft EDMT® Solution customers can connect as many as 192 HP ProLiant DL980

servers and 40 PB (40,000 TB) of storage in a single 15,360 core system making this a truly capable Big Data solution. The HP P2000 is an excellent storage platform choice for this solution offering.

Conclusion

For Big Data problems that require ingesting extreme volumes of heterogeneous data the BMMsoft EDMT Solution gets very high marks from HRG. In addition the partner solution offered by the combination of BMMsoft, Sybase, and HP brings together an excellent solution set that is fully capable of effectively addressing most of today's Big Data Challenges. Each of the products that comprise this solution is in its own rights a best in class solution. However, in combination BMMsoft EDMT Solution, Sybase IQ 15.4, and the HP DL980 provide extreme levels of scalability, flexibility, and availability not to mention record setting data ingest speeds that you would be hard pressed to find elsewhere.

Harvard Research Group

Harvard, MA 01451 USA

Tel. (978) 456-3939
Tel. (978) 925-5187

e-mail: hrg@hrgresearch.com

http://www.hrgresearch.com