

HRG Assessment: Cross-Enterprise Data Integration

Executive Summary

Executives understand the benefit of having access to the complete range of corporate data without having to be exposed to its underlying complexity. In the future, users will no longer have to be limited by or concerned with either the physical or virtual location of data, which application it comes from, or even if the data is structured or unstructured. As Data Integration matures and evolves, it will accommodate an increasingly wider range of data types. Ultimately no data, regardless of type or location, will be beyond integration. Initially, operational data coming from transactional systems as well as historical data coming from data warehouses and data marts will be integrated into a single, actionable view. In the future, unstructured data will increasingly be brought into that view such that it will play a larger and larger part in contributing to the benefits of cross-enterprise data integration.

Data Integration that leverages the full capabilities of metadata (data about or describing data) allows business managers and executives to live with complexity without having to be overwhelmed by it. The business results that come from integrating cross-enterprise data into a single actionable view are substantial. Data integration drives real long term cost reductions, improved operational efficiency, enhanced competitive advantage, streamlined regulatory compliance, and increased productivity.

Ideally, the people in an organization who are thinking about what data is and how to leverage it should not have to worry about how to get it. People who build, deploy, and use applications know that the real objective of an application is converting data into actionable or useful information. There have been notable failures of large ERP and CRM applications wherein organizations missed out on the promised benefits of such solutions because of data-related problems. Data brings with it a unique set of requirements, needs, and expectations. The automation of business logic embedded in applications must undergo continual change if businesses are to evolve.

Data integration drives real long-term cost reductions, improved operational efficiency, enhanced competitive advantage, streamlined regulatory compliance, and increased productivity. But until now, enterprise-wide data integration has been virtually impossible. Data in corporations is in every conceivable format, structured and unstructured, and in diverse heterogeneous environments. And to add to the complexity it comes in all speeds: real-time, near time, historical, and archival.

Dealing with this diverse data on an enterprise basis is now possible with the Sybase Data Integration Suite. It provides the modeling, data transformation, data manipulation and data integration tools to enable IT to deliver to its users the competitive edge knowledge base applications they need.

Data Integration Drivers

The need to create, achieve, and maintain distinctive competitive advantage is driving the move to implement cross-enterprise data integration. Beating the competition is key; getting more out of your data in order to advance the business is a requirement. Additional data integration drivers include: globalization, government-focused compliance or risk mitigation, and reducing operational risks (lowering the risk of being in business). These are only few of the reasons why IT and business professionals should be concerned with the integration of complex, highly distributed, and heterogeneous cross-enterprise data flows into a single view.

Also, cross-enterprise data integration projects typically involve several different integration techniques used in conjunction with one another. This could include various combinations of the primary techniques of ETL (extract, transform, and load) for data consolidation, replication, federation, and real-time event messaging. Three examples of business scenarios that require such combinations are quarter-end financial reporting, front-office financial risk management, and operational business intelligence.

Delivering end of quarter reports without the bottlenecks

When implementing a data mart or data warehouse for financial reporting, the traditional technique to use is consolidation via an ETL tool. The classic scenario - the financial data mart needs to be updated every day so the CEO can track progress against goals. This works well until the end of each quarter when the CEO wants a report based on that data mart twice a day. Therefore the data mart needs to be updated twice a day. If an update is done using ETL twice a day during the busiest week of each quarter the impact on system performance and IT personnel will be significant if not outright painful. In this case using ETL by itself is not efficient rather it would be better to use consolidation and replication techniques in combination. Replication could be used to stage the data to a non-critical path server and then ETL could be run against that server.

The only way to figure this out is to understand the end user requirement. The end user requirement in this case in the CEO's words is "I need the data once a day but during the end of the quarter I need it twice a day." Increasingly, many end-user requirements are ad hoc. The root cause is a shift in many organizations to intra-day decision-making, which means data warehouses need to be refreshed more frequently. The focus is increasingly on operational decision-making. Operational decision-making requires data from multiple points in the operational process. This highlights the trend toward data integration projects involving multiple techniques.

Risk moves from middle to front office

Financial services companies are comprised of three parts: front office, middle office, and back office. The front office is where revenue is captured and includes sales and corporate finance. The middle office has traditionally been where risk and IT have been managed. The back office is where administrative and support staff resides, and is where settlement, clearance, record maintenance, regulatory compliance, and accounting functions occur. In many large financial institutions, risk is moving from middle office to front office.

Risk now has to be looked at more cohesively or more holistically across middle and front office. In the traditional middle office environment, risk assessments occur once each day. However, with front office risk, assessment needs to be made much more frequently because the frequency of transactions is rapidly increasing due in large part to automatic and algorithmic trading. Therefore, this increased risk exposure requires access to both real-time and historical information for successful management. Also, since risk now needs to be assessed closer to both the

front office and middle office, there is an emerging requirement to have a shared common view of the data (transaction data, market data, customer data, reference data, and corporate actions data). The front office may be more reliant on current transaction and market data, whereas the middle office may make more use of historical reference data.

Given these needs, there are several data integration techniques that can be used in tandem to move data into a high-speed consolidated warehouse. Using ETL, data is extracted from historical systems that don't change. Real-time event messaging can be used to push information out from systems that change occasionally or periodically. From this point risk assessment and analysis is done and reports are generated. Federation is a good way to distribute those reports and with grid capabilities the reports can be loaded into cache on the grid and then accessed / downloaded when and where needed without impacting the report server. The federation technique gives faster access to the data (reports) by providing locality of cache. This locality of access also enables the application using the data to run faster because it gets the data more quickly.

Business Intelligence and Operational Business Intelligence

Traditionally, business intelligence based decision-making has been done on the basis of static historical data where latency has not been an issue. For example, business analysts who are trying to come up with better products and services or to enhance the business' ability to compete leverage integrated historical data sources to learn in which region the business has been doing well. They then look for trends and correlations among products, sales, and promotions that provide insights into the reasons for success and can be replicated in other areas. In this way, businesses benefit through improved intelligence based on historical business performance data that has been appropriately pulled together for analysis.

The requirements for operational business intelligence are significantly different from those of this more traditional business intelligence process. Operational business intelligence is used for intra-day and intra-hour decision-making and as such needs to take into account current near-time/real-time data in analysis and problem resolution. For data integration solutions to be effective in this environment, they need to accommodate the integration and use of both historical and real-time data sources that come from all across the organization. One possible method to provide such a solution is to use ETL to consolidate the historical data combined with a federation layer that provides access to data needed on a real-time basis.

Given this increasing need for reduced data latency and more timely decision-making, the ability to support operational business intelligence and data analysis are key requirements for any data integration solution provider.

What's Next?

Most organizations can benefit from achieving a single, coherent view of their enterprise-wide data resources. This view in many cases will be based on a blend of heterogeneous data (relational, delimited text, etc.) from various sources and locations including real-time, historical, and event-driven. There is substantial untapped potential for innovation that can be realized by integrating cross- enterprise heterogeneous data to enable the creation of the next generation of focused, effective applications. Some of the first of these next-generation applications will be sales, marketing, and customer experience focused.

Next Generation Data Driven Applications

A generation of new applications are being developed that combine real-time and near real-time information from multiple sources. Some of these next generation applications are:

Operational Business Intelligence

Synthesizing analytical information with data from an organization's operational processes to provide "just-in-time" information at the point of action. Examples include:

- Process analytics and tracking
- Customer analytics
- Real-time business intelligence
- Business performance analysis
- Operational unified views

Risk Management

Combining data from multiple sources to reduce the uncertainty inherent in many key business decisions. Examples include:

- Risk profiling
- Insurance risk
- Financial risk
- Operational risk
- Scenario building

Development / Deployment of other Data-Rich Enterprise Applications

The data explosion offers opportunities for companies to innovate, develop and launch new data-driven products and services. Examples include:

- Personalization services
- Competitive analysis
- Single View of the Customer

The Data Integration Alternatives

Stand-Alone Products

Delivering the business value that cross-enterprise data integration promises is a task that is too complex and difficult for the majority of stand-alone products and vendor-centric data hubs. There are stand-alone point products that can provide some degree of control, but in the end these will result in a fragmented approach to building data flows. For an organization taking on a fairly simple project, such as pushing data from transactional systems into a data mart for reporting and analysis, today's stand-alone tools may prove adequate. However, as organizations begin to build more sophisticated data flows, making enterprise-level changes as the business grows, a stand-alone product may not present the options or flexibility required for growth. In many cases, the stand-alone

point product solution that works well on a departmental project basis is not likely to be the most practical and effective enterprise-wide data integration approach.

In the current market, many vendors offer what amounts to a single unique data integration technique rather than a total solution, while some of the larger players are investing in several data integration areas. The market is quite fragmented and for the next couple of years it will be going through a consolidation and shake out. Today data integration techniques are on a convergence course and many stand-alone point solution vendors will be acquired, expand, or go out of business.

Data Hubs

Aside from the stand-alone point products that were touched on previously, there are basically two approaches to enterprise data integration. One is the data hub approach (e.g., Oracle and Microsoft) and the other is the data integration suite (e.g., Sybase). Data Hubs on the whole tend to be more vendor-centric and less open while Data Integration Suites will typically be more open, flexible and less restrictive.

Data hub solutions, while nice in theory, are not a best fit for the clear majority of companies that have disjointed heterogeneous data environments made up of a diverse mix of hardware systems, operating systems, and data types. Data hubs, rather than integrating a data environment, actually force the consolidation of data on to the vendor's platform whether it is Oracle, Microsoft, or one of the other lesser-known data hub vendors. Therefore the Data Hub option is an excellent choice to consolidate your enterprise data onto a closed single vendor database centric platform. The two most prominent Data Hub vendors today are Oracle and Microsoft.

Oracle's hub solution drives customers to move all of their data onto the Oracle hub architecture, thereby locking them in to a single-vendor, and relatively expensive, solution. However, with over 10 million installations of open source databases worldwide, the Oracle-centric view of the world leaves a lot to be desired. Oracle would prefer that customers ignore the fact that most all enterprises today have heterogeneous data sources from multiple vendors and custom in-house solutions. Customers who follow Oracle's recommendation may realize when it is too late that either they have left a lot of data value untapped and isolated or that the expense of integrating that data is prohibitive.

Microsoft's hub strategy is very similar to Oracle's and in addition Microsoft is a late entry to the data integration arena. Their approach generally ties an enterprise into their platform. Some data can be integrated with Microsoft, but in the end the Microsoft approach tends towards a homogeneous, all-Microsoft, environment.

Technique, Technology, and Product Convergence

In this new world of data integration technique, technology, and product convergence, data is collected and stored in all different forms: delimited text, spreadsheets, word documents, SQL, VSAM, ISAM, DB2 Oracle, Sybase IQ, Sybase ASE, and more. All of this data must be extracted, transformed, and loaded into a data warehouse for analysis and reporting. Once data is integrated it can be made accessible either through a relational format, a file system format, or a web services format.

Organizations are now considering using multiple tools in conjunction with each other to solve their data integration problems. They may build a data warehouse using ETL tools and then use EII to federate the information between the warehouse and current data. Each of these tools is likely to be used in different projects, and so it is in each organization's best interest to become familiar with a single set of tools that addresses the

complete problem set. Within each of these organizations, there is a need to model, manage, and administer these components with a single tooling that allows users to learn a single set of tools and then leverage that knowledge across different projects.

The need for a set of data integration software tools that work well together is driving the ongoing convergence of data integration techniques, technologies, products, and the ongoing development of Data Integration Suites by IBM, Informatica, and Sybase. The Sybase Data Integration Suite will provide a single tooling that enables modeling and development of each of the different components or techniques into a single solution. Over time, this tooling will become more tightly integrated, sharing a common metadata layer or repository.

The Sybase Data Integration Suite

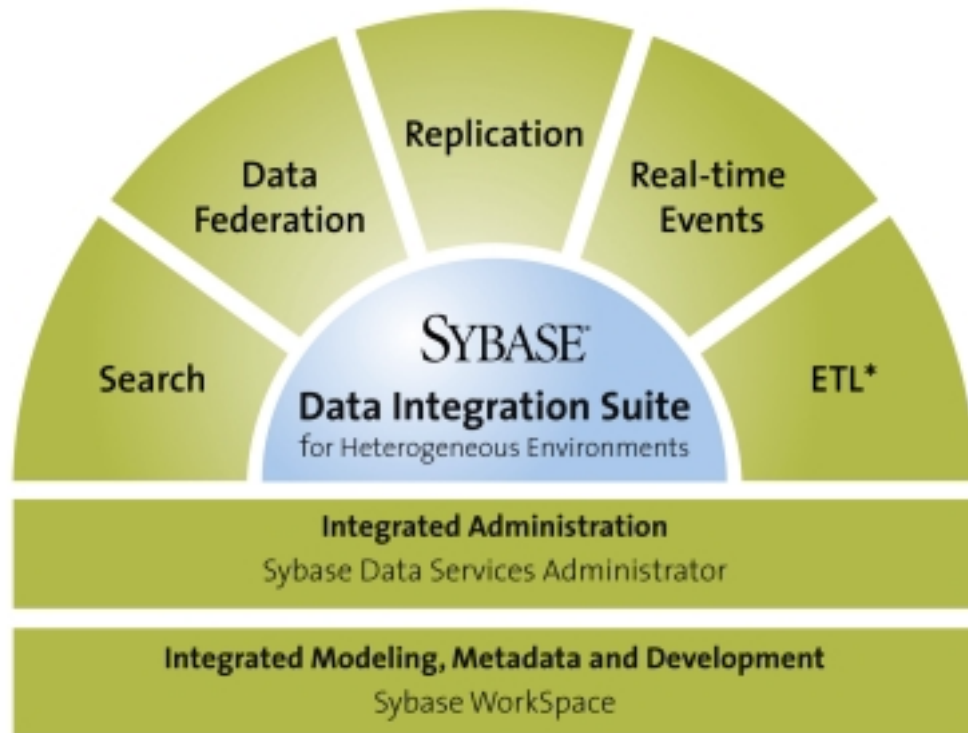
At their August Tech Wave 2006 event Sybase Data Integration Suite (now generally available) was introduced for the first time. This suite, when fully fleshed out some time in 2007, will combine Data Federation, Replication, Real-Time Events, Search, and ETL in a single scalable suite, using a common development environment, metadata management and administration. The Sybase Data Integration Suite will allow enterprises to create data flows from a range of heterogeneous data sources thereby unleashing the latent value of their data assets to improve the speed and quality of decisions across the organization.

Before Sybase Data Integration Suite, organizations were faced with having to buy - for example - ETL from one vendor and replication from another vendor. Sybase through its suite offers common modeling, common metadata, and common management across these different integration techniques. This is a real boost to IT productivity because it allows identification of reports that are impacted by changes at the database or data source level. This is very difficult without common metadata and common modeling.

Sybase's vision is to enable customers to look at end-to-end data flows in the organization regardless of the component processes, and then provide a single way to model and implement the data flows based on a common metadata model that will be shared among all the different techniques.

Sybase's plan is to enable organizations to run their own infrastructures. The trend of customers wanting to run their own infrastructures is driven by cost pressure and is expected over time to impact even the Fortune 100 that so far have been focused on professional services centric offerings.

In today's competitive environment, hardwiring a particular method of integration to a given application is typically not recommended. If only historical data is being dealt with then this may be acceptable but what happens in the future if there is a requirement to add real time data to the integration mix. For example; if a business has a data warehouse with extremely clean data and then acquires another company they will need to blend the loose data from the newly acquired company with existing clean data in order to generate a set of consolidated financials to meet government requirements. What is needed in this case is choice and flexibility regarding data type, data format, and data migration without sacrificing the ease, speed, low cost, and low complexity that come with a more simplified method. In this case an integrated tool driven approach is recommended. This will allow design of the lifecycle of the dataflow in a flexible way by modeling metadata and then reusing the models and metadata themselves as required. In this way better consistency can be achieved while removing variability from the process thereby resulting in a faster more controllable outcome.



**Sybase ETL is currently available as a standalone product, sold separately from the Data Integration Suite.*

The components of the Sybase Data Integration Suite (depicted above) are:

Data Federation

Sybase Data Federation (formerly Avaki EII) integrates data from multiple distributed sources in order to provide standardized access to integrated views of the data through a single data layer. Data Federation delivers data from original sources without the requirement to move or copy. Data Federation allows the user to look at information across the company / organization in order to obtain a single coherent view of the customer.

Replication

Sybase Replication (also sold as Sybase Replication Server) provides transactional near real time near simultaneous synchronization between various databases and locations to ensure reliable data availability across the enterprise. Organizations can reach simultaneous synchronization through two-phase commit but for most applications two phase commit is largely impractical

ETL

Sybase ETL (formerly Solonde) -*Extract, Transform and Load*- is a data consolidation technique that allows you to take information from multiple parts of an organization bringing together data or copies of data from multiple sources into a central location or repository for analysis and reporting. Available as a stand-alone product, Sybase ETL will be included in the next version of the Sybase Data Integration Suite)

Real-Time Events

Sybase Real-Time Events (RTE, formerly Real-Time Data Services) is a distribution technique and is actually a subset of the functionality of Sybase Replication Server that updates in micro batches rather than synchronizing by replicating an entire database. RTE and Replication Server keep timely data in support of Operational Analysis / Intelligence. RTE uses real-time event triggers in the data that, when a threshold or event is reached, cause the required data element to be pushed data from heterogeneous enterprise databases to messaging architectures (message bus), eliminating the “information lags” created by batch updates or intermittent polling processes. RTE can be used to connect any data source directly to any application or data warehouse requiring dynamic near real-time event triggered updates. Future versions will include BAM and a rules engine.

Search

Sybase Search (formerly OmniQ) uses a context sensitive search technique that can search for and find relevant information within the volumes of unstructured information stored in organizations centralized repositories, libraries, file systems and existing document management systems. Sybase Search along with the Sybase Avaki EII tool constitutes the data federation capability of the Sybase Data Integration Suite.

Metadata Benefits

Today many enterprises are using a variety of disconnected techniques for data integration such as data federation, replication, and ETL. These enterprises can now benefit by adopting a metadata approach to defining and modeling data transformation. Metadata has been used to model data transformation (for example using ETL) and now it is time to begin to focus on integrating these various data transformation models into data flows. One group who will benefit from this will be the business users who will be able to have a single view of the customer and products across the organization. They will be able to pull together data from multiple different sources in order to meet compliance requirements (e.g., SOX, HIPPA, and Basel II) without having to create more work / cost by pushing compliance down to individual divisions of the company.

Meta Data approach to data integration

With their metadata driven approach to cross-enterprise data integration, Sybase will be able to drive down TCO for most IT infrastructures. With this approach, there is very little if any recoding required because most of the metadata of the source and the target structures is already in the system and can be completely reused. Additionally Sybase delivers an extendable data modeling capability as part of PowerDesigner that includes business process modeling.

For example, in a data federation context, the Data Integration Suite will provide the tools needed to model the sources, the federated database itself, and all of the data mapping. From this metadata, a significant portion of the implementation details can be automatically generated via the modeling environment (PowerDesigner). Multiply that automation by all of the contexts that the various data integration techniques represent – replication, real-time

events, ETL, etc. – and the significance of the simplification and automation that the centralized Data Integration platform provides becomes quickly apparent. For managing complex data environments, this is a huge step forward.

Modeling , Metadata, and Sybase PowerDesigner

Metadata is data about data, more specifically the data that describes the structure and workings of an organization's use of information and the systems it uses to manage that information. The modeling techniques that produce and manage metadata and provide a complete description of the systems within an enterprise are data models, process models, information liquidity models, application models, and requirements.

Sybase is creating an integrated metadata layer or repository as a foundation for their Data Integration Suite via its PowerDesigner data modeling product. Sybase will drive development for the Data Integration Suite components from this metadata layer. This metadata layer will initially be populated with metadata models to create the analysis and design that will be used to drive data integration through the suite's toolset. The Information liquidity Model, developed for use with Replication Server, lets you use a source data model and a target data model to model the mappings and from the resulting metadata you are able to generate the implementation routines for replication server. From the metadata layer you will have access to the entire Data Integration Suite toolset and be able to generate the required implementation routines.

In the first version of the suite, different data integration components will have different sets of metadata. In the second version, the components in the suite will take advantage of an integrated, common metadata layer or repository. Version 2 of the Data Integration Suite will provide enhancements to metadata management that will include access rights, ownership, and permission definition. Additionally, the metadata management capabilities of the Data Integration Suite 2.0 will provide support for data lineage and impact analysis.

The industry-leading data modeling tool Sybase PowerDesigner, which provides key design functionality for Sybase Workspace, has been in use for more than a decade. PowerDesigner now provides metadata management capabilities and supports standard modeling techniques to produce metadata.

Sybase WorkSpace

Sybase WorkSpace is a key part of the Sybase Data Integration Suite, providing its capabilities for integrating the tools that comprise the Suite and its cross-enterprise dataflow creation capabilities. WorkSpace is the Sybase framework that includes the PowerDesigner modeling capability as well as a development environment where administrators can develop, create, monitor, and deploy services to target run-time environments. Sybase Workspace brings with it the flexibility and openness of Java/Eclipse and the ability to support a wide variety of platforms. WorkSpace includes and supports a common integrated administration layer that manages all of the Suite's components.

Sybase Data Integration Suite Road Map

Today the Sybase Data integration Suite includes Data Federation, Data Replication, Real-Time Events, and Search capability. In the next version ETL (Extract Transform Load) capability will be added along with enhanced integration capability via metadata and modeling. The third release will provide seamless integration with development and management.

Sybase Common Services and Data Services Administrator will be included in the suite at no charge and a limited copy of Workspace will be provided with each license of Replication, Real Time Events and Federation. Additional Workspace licenses will be sold separately.

Organizations who invest in the Sybase Data Integration Suite will be able to license only the components they need for their specific environments. Additionally, Sybase Replication Server will continue to be made available either as a part of the Data Integration Suite or as a stand-alone product.

Conclusion

Sybase Data Integration Suite earns high marks on its:

- ✓ Ability to deal with near real time high volume transaction work loads without missing a beat.
- ✓ Openness – organizations are not forced to buy Sybase ASE or IQ – they can leave their data where it is in its original format.
- ✓ Expected low total cost of ownership and extremely competitive pricing.
- ✓ Innovative use of metadata to create enterprise-wide data-flows that are required by many next generation applications.

In a world that is increasingly using open systems and open source software any closed vendor-centric single technology focused data integration offering for most mid to large size companies is a less than optimal choice.

Data integration drives real long term cost reductions, improved operational efficiency, enhanced competitive advantage, streamlined regulatory compliance, and increased productivity. But until now enterprise-wide data integration has been virtually impossible. Data in corporations is in every conceivable format, structured and unstructured, and in diverse heterogeneous environments. And to add to the complexity it comes in all speeds, real-time, near time, historical, archival.

Dealing with this diverse data on an enterprise basis is now possible with the Sybase Data Integration Suite. It provides the modeling, data transformation, data manipulation and data integration tools to enable IT to deliver to its users the competitive edge knowledgebase applications they need.

Harvard Research Group is an information technology market research and consulting company. The company provides highly focused market research and consulting services to vendors and users of computer hardware, software, and services. For more information please contact Harvard Research Group:

Harvard Research Group™

PO Box 297

Harvard, MA 01451 USA

Tel. (978) 456-3939

Tel. (978) 263-3399

e-mail: hrg@hrgresearch.com

<http://www.hrgresearch.com>