Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

WHITE PAPER

Snap-in, Grid-Based Data Integration: IBM Information Server Blade

Sponsored by: IBM

Carl W. Olofson Jean S. Bozman

February 2008

IDC OPINION

The many databases in any enterprise IT environment have added a level of complexity not only to their management but also to finding, collecting, and distributing key information from them. The problem is made more difficult in that most enterprise data integration solutions currently available require careful choices in terms of the servers that will best accommodate the enterprise data integration software and meet the required service levels, and also require knowledgeable installation and configuration of the servers, the software, and the network connections. Challenges include the following:

- Choosing and configuring the right hardware environment for data integration
- Assembling and setting up data integration software, which is often a complex, error-prone process
- Tuning and managing the data integration server for performance

IBM Information Server Blade in a grid computing environment represents an important approach to solving all of these problems and to providing a unified view to data that is located within many business units across the enterprise. IBM Information Server Blade is a preconfigured "appliance"-style offering with components that are configured based on customer best practices and deployed in a grid computing environment to deliver both ease of setup and highly scalable performance. Importantly, it is also extensible, allowing customers to add server nodes, thereby providing more computing resources, as needed.

METHODOLOGY

This white paper was developed using information from three discrete sources: background knowledge from years of IDC research in the area of information management technology, detailed briefings and materials provided by IBM regarding Information Server and the Information Server Blade, and interviews with IBM customers who use the Information Server technology.

Specifically, IDC has relied on information gathered over the years from market analysis and technical evaluations of information management software that is used for data integration to provide overall commentary and background information regarding the problem space that Information Server Blade addresses. This is important, as the IBM offering is new and still in the early adoption stage. Detailed briefings and supporting materials from IBM have provided the information necessary to describe the Information Server Blade. The analysts also conducted a series of interviews with key users of this technology in order to understand how it is used in actual practice.

IN THIS WHITE PAPER

This white paper explores the problems confronting IT organizations today as they seek to bring data together from disparate sources in order to achieve better information coherence, operational efficiency, and more effective overall governance. Specifically, this paper considers the problems posed by the complexity of enterprise data integration technology in terms of its configuration and maintenance. It looks at the utility of a grid approach in providing scalability and manageability for enterprise data integration and considers how the appliance approach, with its preconfigured hardware and software, can further reduce cost and risk for customers by taking the effort and guesswork out of setting up an enterprise data integration server.

This paper considers the solution offered by IBM, first with respect to the Information Server technology and then with its appliance offering, the Information Server Blade. It examines IBM's renewed commitment to the IBM Information Server and how the blade offering, inspired by customer examples, has reduced the risk and complexity of deployment. This paper also offers concrete examples of how the IBM Information Server is delivering real value to customers.

SITUATION OVERVIEW

Key Trends in Support of Coherent Information Management

Enterprises are increasingly challenged by business requirements that demand timely access to key information. These requirements include both decision support and better, more nimble automated processes. Yet at the same time, most large enterprises have found that as their application portfolios grow, adding to their inventory of potentially valuable business information, value-rich information is locked up with the data that is managed by individual application databases. This means that data that could potentially enable applications to run more efficiently, avoid errors, and provide business users with timely information for critical business decisions is scattered about a broken maze of unconnected application databases.

The result is that applications lacking a common view of critical elements (e.g., customers, inventory, sales) may act inconsistently, and even detrimentally, from a business perspective. It means that businesses cannot provide accurate accounting and may not only make poor management decisions but also be subject to penalties for violations of various reporting regulations, such as Sarbanes-Oxley. Lack of coherence in the information provided by IT can cause employees to fail to act as a team and can cause inconsistent, unreliable, and incoherent behavior on the part of the company in the eyes of customers, partners, and investors.

IT managers and executives agree that the answer to these problems is to build systems that can ensure that information provided by IT is consistent, complete, and timely. In effect, they are building a way to access data that shows a consistent view of the state of their business, over time, gaining deep insights into the way the company operates, and providing information as to how it can be continually improved. The challenge is to accomplish this while dealing with the existing, crazy quilt of disparate application databases. At the same time, data governance requirements demand that organizations guarantee certainty that only those who are authorized to read or update information have in fact done so. To meet this challenge, IT organizations increasingly are turning to data integration technologies that provide dynamic data movement and transformation, delivering the right data at the right time into blended data environments for processing and reconciliation, with appropriate levels of auditing and security. In short, the requirement is for:

- Information that is always in agreement, especially the key master data
- Information that is accessible to whoever is eligible to use it and unavailable to those who are not permitted to see it
- Information that is available to whoever needs it when it's needed

These are the essential elements of a coherent information environment.

Enterprise Information Integration

To deliver such an environment, IT organizations turn to enterprise data integration. This is a collection of coordinated technologies that take data from disparate sources, transform it, reconcile it, cleanse it where necessary, and coalesce it in structures such as databases or virtualized data delivery systems where it can be used to provide business intelligence, condition or enhance automated business processes, synchronize application data, or deliver integrated data in a timely manner to users and applications. In effect, enterprise data integration technology acts as a "lens" that concentrates important data, gathering it from data sources across the enterprise and making it easier to access a common view of this data — and then importing and encapsulating that data into an appliance that is easy to deploy and to maintain.

Enterprise data integration normally consists of an infrastructure of data integration technologies and a means of dynamic delivery of the integrated data. The elements of an integrated, strategic data integration infrastructure include:

- Bulk data movement technologies that extract data from the sources, transform it into the target format, reconcile conflicts both with various data streams and with target data, and then load it into the target database. This technology is generally called extract, transform, and load (ETL), though it is sometimes called ELT, depending on where the transformation takes place.
- Data synchronization or "real-time" data movement technologies that move time-sensitive data, whenever it changes, from sources to targets, again providing transformation and conflict resolution, and usually driven by a "change data capture" (CDC) capability.

- Data quality technology that detects data inconsistencies, including logical inconsistencies in units of data, as well as format errors and other application-specific problems with the data, and corrects or rejects the errant data so that errors and inconsistencies are not propagated to the target database.

- Security safeguards that ensure that as data moves through the system, common access and update rules are enforced to ensure that data is visible to and changeable by only those authorized to see or change it. Auditing is also involved to certify that those security rules have not been violated.

- Development and management elements, including tools for defining the data, its business meaning, the data sources and targets, the transformations, integrity and validity checks, and security parameters. These tools populate common metadata that drives all the operational elements of the enterprise data integration system.

When the picture is completed with technology that can offer the integrated data either directly from the data sources (commonly called data federation) or from some integration point, such as an operational data store (ODS) or an integration cache to the requesting user (through a query or reporting tool) or application, the result is enterprise information integration (EII). More complete solutions also blend content search and retrieval as part of their EII offerings. Some vendors only provide data delivery from sources without any attempt at synchronization or reconciliation. This is commonly called "integration on the glass." Such vendors call it EII, but it should be understood that robust EII requires full data synchronization and reconciliation of the various sources. Otherwise, the delivered data, blended from differently managed sources, is suspect in terms of its integrity. A complete enterprise data integration capability is a necessary prerequisite of a full EII solution.

Virtualization, Grid Computing, and Appliances

As may be easily imagined, enterprise data integration represents a daunting configuration challenge for larger IT systems due to its demanding yet highly variable workload and the premium placed on performance and scalability. In general, demanding and highly variable workloads have been difficult to manage and provision in the past, usually leading to massive overprovisioning in order to ensure that such workloads have the resources that they will need and that those resources will be available when they need it. This is an unfortunate effect of any environment in which fixed resources are statically assigned to individual workloads, such as applications, databases, or data-intensive activities such as enterprise data integration. Today's organization needs more flexibility in infrastructure, just as it needs to be sure that all data sources have high levels of data integrity, accuracy, availability, and security.

In recent years, however, approaches to systems configuration and provisioning have emerged that address this problem. These approaches are known as virtualization, grid computing, and appliance delivery — and all three are being widely discussed within organizations today, even though some are only in the early stages of adopting these technologies and harnessing the benefits they bring in cost reductions, energy efficiency, and improvement in data delivery to end users.

Virtualization, in general, is the rendering of an array of system resources in abstract form so that applications are deployed upon, and bound to, the abstract entity rather than an actual systems or storage resource. This allows systems and storage resources to be altered, moved, expanded, added, and removed with minimal impact to the application. As implemented in the IT world, mainframe and scalable RISC-based and Itanium-based servers are already highly virtualized systems, on which multiple workloads can be reprovisioned, as needed, within multiple logical (software-defined) partitions — abstracted above the underlying hardware layer. But virtualization in the x86 world has been rapidly accelerating in recent years, as customers seek to consolidate workloads that have been scattered across many volume servers, with the goal of reducing operational costs, reducing management complexity, and improving energy efficiency.

Virtualization is typically employed for server consolidation because it permits a single server to seem, from the point of view of applications, to be many servers, by means of hosting multiple operating system images on a single hardware system. Thus, applications that had been deployed on individual servers may be deployed on a single server, even though the applications operate as if each were on its own system. But if more capacity is required to service these applications, it may be necessary to reprovision the workload onto more scalable servers and to restart the application on new hardware — whether it be rack-optimized servers or blade servers. Similarly, storage can be provisioned in such a way that applications that target individual files on individual drives may "see" the files in a logical view of storage resources, although the files themselves may span multiple volumes and often may be "striped" (broken up and distributed across spindles to reduce head contention) and "mirrored" (copied onto other drives to ensure constant availability in case of a drive failure).

In short, virtualization insulates applications from the physical details of their deployment so that those physical elements can be changed without affecting the applications. It is often thought of as rendering one physical resource from many, separate physical resources (such as a server or storage array) as multiple logical resources (such as logical servers or storage volumes). However, it is also possible to use virtualization to see one resource when the application is accessing a partition, or "slice" of a server's resources.

Grid computing enables multiple system and storage resources so that they seem to be a single continuum of system capacity or storage capacity — or both. For this reason, applications deployed on the grid can have their elements opportunistically executed on whatever server happens to be the best one, at any given time, to host the workload. An application grid is normally set up by redundantly loading application components on multiple servers, usually in a cluster, and then executing those components through a governor that uses load balancing to assign servers dynamically, launch the component instances on those servers, and orchestrate tasks that involve combinations of component services. This approach favors complex tasks with highly variable work volumes, especially where parallel processing can be employed to optimize execution.

Enterprise data integration is a good candidate for deployment as an application grid. It is important to note that while grid computing and virtualization are related concepts, they are very different in practice and do not solve the same problems. Virtualization aids in software deployment flexibility, but it does not address high availability or scalability. Grid computing addresses hardware deployment flexibility, as well as high availability and scalability. Deploying a grid can be very complicated business, even for IT organizations with advanced IT skill sets. Even for a well-bounded problem domain such as enterprise data integration (the data is not bounded, but the tasks are), deploying the execution elements as an application grid requires a good deal of time and expertise. Increasingly, complex software packages are being preconfigured and delivered as appliances. This approach overcomes the problem of installing and configuring the software, over and over again, on each hardware resource. It also allows the deployed software to work in a grid configuration by setting up the software on certified hardware in advance and by enabling it to connect together to provide a grid-based capability immediately upon deployment.

Key Challenges: What Users Are Saying

A succession of surveys and customer conversations has revealed steady evolution in the thinking of IT managers and CIOs toward more strategic data integration. In 2000, most companies reported that their top 2 uses of data integration technology were to populate data warehouses and to perform data migration projects. Data synchronization was also done, but typically on a very limited basis (one-to-one) with the number 1 technology chosen being in-house-developed programs and scripts.

These approaches were not scalable and did not address the problem of increasing data complexity in the IT environment. In fact, they often contributed to such complexity because application changes or version upgrades often required changes to the data integration software that moved its data. This additional burden meant that version upgrades were increasingly postponed, and IT managers were increasingly reluctant to take on new database projects, thereby creating a data management environment that tended to inhibit business agility.

The additional challenge of compliance requirements such as Sarbanes-Oxley has forced many enterprises to come to terms with the confused, unmanageable "rat's nest" of data in their midst. To deal with this problem, and to try to use data integration to reduce rather than increase operational complexity, IT managers in the past five years have shown a distinct preference for strategic rather than tactical approaches to data integration and for packaged software over in-house development.

Recent surveys have shown that a majority of IT managers today regard data integration as a key component of their overall information strategy; the topic has gained top management visibility and requires such important elements as multiple database support, "real-time" integration, and data quality features. Data quality, which had been a small, slow-growing niche a few years ago, is now regarded as a critical part of strategic data integration, and that market segment has accelerated tremendously in terms of revenue growth.

The most successful strategic data integration approaches are those that are implemented incrementally, in which each project builds upon the one before it. This involves the use of a single vendor's data integration technology and robust use of metadata to capture data definition information at each stage. Where some enterprises have run into significant challenges in this regard, however, is with the problem of scalability for large-scale data movement. In order to achieve such scalability, enterprises have put a great deal of effort into configuration and deployment of large-scale integration servers, and the manual effort in maintaining such systems has been an inhibiting factor in their adoption.

Critical Success Factors for Coherent Information Management

Since the presence of multiple application databases is likely to be a permanent fact of life for most large enterprises, and since the volume of data they manage is constantly and rapidly growing, a strategic approach to data integration is the only path to coherent information management, which delivers manageability, good governance, and business agility.

Virtualization and grid computing are deployment options that can enhance the scalability and robustness of a strategic data integration platform, but they come at the cost of a good deal of IT staff effort devoted to their configuration, deployment, and ongoing management. Three elements can be employed to address this problem:

- Deployment of components that work well together, including a simple, elegant design that eliminates the need for the kinds of software and hardware setup and configuration tasks that would introduce additional cost and risk into the picture.

- Grid-based hardware deployment to ensure system scalability and availability. The ability to leverage best practices learned over years of customer projects regarding scale-out grid deployments is one important factor leading to successful deployments of enterprise grid technology to support resource pooling for IT infrastructure.

- Hardware that employs the latest advances in technology around compute capacity versus power consumption and cooling requirements. These include support for multicore and virtualization technologies that support more efficient deployments of server hardware into the networked computing environment. The trend toward denser server "nodes" within a grid or cluster, supporting multicore processors and increased memory within each node, allows each server to take on more workload, thus improving server resource utilization.

Moore's law, combined with intense competition within the worldwide server marketplace, has resulted in highly advanced platforms at commodity prices that can now support scalable enterprise workloads.

But the hardware elements must be put together, the software must be installed and configured on them, and the whole thing requires a good deal of staff time to manage. How can organizations take this approach and still manage ongoing costs? Because IT skill sets vary greatly from site to site, and from organization to organization, the "best practices" gleaned from previous deployments must be leveraged — and preserved — so that new customers can gain similar benefits without climbing a steep learning curve.

By taking advantage of customer experience and employing best practices in such deployments, organizations can build and deliver appliances that provide preconfigured hardware and software with utilities that make them easy to manage.

IBM Information Server Blade

Like most vendors, IBM has offered a wide range of data integration technologies, including ETL, data quality, "real-time" data movement, and data federation. Also like most vendors, IBM offered these products as separate modules that had to be integrated separately. With IBM Information Server, IBM has integrated them into a platform for enterprise data integration.

Some customers have found that the best way to get the performance and scalability they need from IBM Information Server is to deploy it in a grid configuration. Such a deployment is complex and takes some time to set up. Nonetheless, several have done so, and with impressive results. By taking the approach of integrating all the individual components of this grid-based solution, IBM is reducing operational costs associated with asking customers to take on the task of integration, which can be complicated by the variation in IT skill sets from customer site to customer site.

IBM has a well-established history of learning from customer experience and applying best practices to enrich and improve its product offerings. By studying the successful grid implementations of IBM Information Server by these leading customers, IBM was able to come up with a reference configuration of hardware and software, as well as standard software installation and deployment parameters.

But IBM went beyond that — implementing the Information Server solution on a bladed server hardware platform and bringing these elements together to produce an appliance product: a preinstalled and preconfigured instance of IBM Information Server on an HS21 blade with two dual-core x86 processors, between 4GB and 8GB of RAM, and up to 12 76GB drives of integrated storage, running Red Hat Enterprise Linux 4. The blade runs on IBM's BladeCenter blade server chassis and is governed by the IBM Tivoli Workload Scheduler Load Leveler.

This appliance-based solution represents the sum of lessons learned from the best practices of leading users of IBM Information Server who have deployed the product in a grid configuration. It can be installed by customers with a bare minimum of setup, since installation and configuration of the product, the operating system, and the grid software have already been done. Its grid computing approach is intended to ensure system deployment flexibility (the blades, based on BladeCenter technology, are deployed by sliding them into the chassis, and since they are preconfigured, they are ready to deploy), high availability, and scalability.

Of course, it is still up to the user site to define its data and the rules that are to govern the integration of the data. Still, the effort, guesswork, and risk involved in the initial installation and setup are reduced by taking a holistic, system-level approach to delivering a data analysis solution. IDC also notes that the IBM BladeCenter product is capable of running Information Server blades alongside other blades (x86 and POWER blades) that use the same chassis and blade management software. As a result, organizations can leverage their overall IT investments in BladeCenter systems to host multiple workloads.

CHALLENGES/OPPORTUNITIES

IBM enjoys a key advantage in supplying a hardware/software combination consisting of products that derive from the company's extensive portfolio of servers, storage, software, and services. As with any appliance product, a key dimension of the offering will involve software and hardware upgrades that are as easy to administer as the original product installation.

However, the marketplace around business intelligence and data analysis is a highly competitive one. While IBM is early to deliver a grid-based solution in this space, other competitors may appear in the future. For this reason, IBM must continue to update its Information Server solution, ensuring that it delivers the most accurate, unified "view" of data gathered from all corners of the extended enterprise. IBM has the opportunity to build on its knowledge base, gained from multiple customer deployments of Information Server technology, to evolve the offering, and to continue to differentiate Information Server in terms of its "reach" into the enterprise, as well as its ease of use and support for access by business users within the organization.

CONCLUSION

Enterprises are challenged to make sense of the data they have under management and to bring it together to form a coherent base of information upon which to make decisions and coordinate operations. With the growing demand for timeliness intensified by globalization, competition, and governmental oversight and regulation, these pressures are becoming more acute, over time. To deal with them, IT organizations are turning to strategic data integration technology.

But if integrated data is to be made an element of day-to-day operations and decision support, then the technology that delivers it must be robust, scalable, and always available. For these reasons, a grid-based technology provides an innovative approach to achieve these operational goals.

As described in this paper, IBM has used the best practices of its leading users of IBM Information Server in a grid configuration to develop and deliver an appliance, now called the Information Server Blade, which seeks to deliver just such capabilities in a bladed server form factor for rapid reconfiguration and IT flexibility. IBM has analyzed customer experiences in deploying this technology and has used that feedback to bring the business benefits of the Information Server grid solution into a blade server form factor to broaden access to a wider group of customers and to reduce the cost of entry to this enterprise data analysis resource.

CASE STUDY

MGM MIRAGE is a company that owns and manages hotel and casino properties in a number of states, including Nevada, Mississippi, Michigan, and Illinois. Some of its best-known properties include the Las Vegas casino hotels MGM Grand Las Vegas, Luxor, Bellagio, The Mirage, New York-New York, and Mandalay Bay. In 2006, the company's use of IBM Information Server technologies earned it the DM Review World Class Solution Award for Data Integration.

For the past two years, since the acquisition of the Mandalay Bay property in Las Vegas, MGM MIRAGE has been working to find a more scalable way to integrate data from its rapidly growing line of properties. Having started with technology licensed from the former Ascential Software (now part of IBM, and representing the base of IBM's Information Server offering), the company needed to build a permanently scalable data integration environment.

MGM MIRAGE turned to the grid-based approach to channel the rapidly growing customer data from its increasingly large properties portfolio into a data warehouse. As a result of the company's efforts, processes that formerly ran for 33 hours before completion were run in just three hours, with the confidence to just let them continue to grow, since the system's capacity scales linearly. Currently, MGM MIRAGE is running 10 million customer records a day in its 2.5TB warehouse, which had held 300 million customer accounts as of October 2005. All customer records are run, through matching and survivorship processes, twice each day to ensure that all records belong in the warehouse and that there are no duplicates.

If any kind of failure of a node in the grid is encountered during these checks, the failed node can be replaced by another. The new hardware is then fully integrated into the system and available for use within seven minutes. This is done with virtually no effort on the part of administrators because the new server blade can be easily activated.

Previously, before adopting the grid approach, the company had to dedicate about 10 hours a week to administer the data integration servers. Now, it spends just minutes per week on routine administration and spends the rest of its time executing and supporting development projects that add functionality to its environment.

Its current grid environment, which the company developed itself, involves somewhat smaller blades than the IBM Information Server Blade. It has eight nodes with two CPUs per node and 4GB of RAM. Senior Database Developer Charles Gladu says the company is "thrilled" that IBM is now offering the Information Server Blade because it will make this kind of grid deployment accessible to many more customers who might not have considered it in the past because "grid deployment can be intimidating."

Not everyone can manage the kind of creative solution on their own that has yielded MGM MIRAGE such stunning success in building scalable data integration. If IBM Information Server Blade works as advertised, all a user will need to do is slide it in and turn it on.

Copyright Notice