January 2006

Why Data Serving on a Mainframe

Contents

Introduction

One of the few constants in the world of Information Technology is change, and that the rate and pace of change increases over time. As clients evolve to become on demand businesses, they need to closely integrate their people, processes and information across the enterprise and across its ecosystem of partners, to speed up core business processes, reduce costs and enable more rapid "sense and response" reflexes to changing market and business conditions, opportunities and threats.

A key element of this approach is IBM's vision of Information on Demand. This focuses on enhancing business value and lowering risk through the capturing, creating, integrating, analyzing and optimizing of all types and sources of information throughout their lifecycles. This approach is designed to help businesses become more reactive based on the information they obtain. The benefits of this approach are numerous and include the ability to make decisions based upon accurate information and real time monitoring and control, the ability to better manage risk and compliance, collaborative benefits across the ecosystem through sharing of information with partners, and reducing overall IT costs through the simplification of the infrastructure.

One key factor that underpins information on demand is that businesses may have to re-evaluate their approaches to data serving. The purpose of this paper is to review the impact that these changes have on data serving and the requirements that this generates as you consider the way you deploy data serving technologies. It then reviews how the IBM mainframe can help you address these requirements and can play a significant role as the data serving hub of the enterprise.

Ever since computing started to be used commercially, data serving has been a key element in business use of information technology. Ensuring that applications and their users have access to the data they need is critical to supporting business processes. Historically, the mainframe has performed the role of a central data server for many large enterprises, taking advantage of its typically high data throughput, scalability and strong security and resiliency capabilities. This has made it an excellent choice as a data server to help support mission critical core business applications. However over time, as businesses and their infrastructures have evolved with multi-tier solutions and technology spread across the company, so corporate data has become more fragmented across the infrastructure.

Today, as businesses continue to evolve, and the reach of their technology expands within the business and out to customers, partners and suppliers, there is an increasing need to re-assess their approach to corporate data serving, being driven by two of the key requirements. Firstly, there is the need to present a common view of data to many users, using many different applications both within and outside of the business to help control costs and improve responsiveness to customers and suppliers. Secondly, and somewhat conversely, there is the need to protect the data from unauthorized access, driven by regulations, customer service and to maintain competitive advantage.

A common view of data

Probably the best way to demonstrate the driving forces and the complexity behind this issue is to look at a specific example. Let's take something as simple as an individual's bank account. It is not that many years ago that all transactions related to the account were processed within the bank. If you withdrew cash from the account you had to go to a branch to do so. If you bought something from a retailer, you paid by check, and those checks were sent to the bank to be cleared. Access to the data was only required by a small number of applications and a few users.

Today's picture is totally different. It is much more complex and involves many more parties and places new strains on your approach to data serving. If you need cash, you can still go to your branch, but typically you would go to an ATM. You are no longer limited to normal working hours, or your bank's location; you can access funds 24/7 around the world. The ATM you are using need not be part of your bank's network; it may be part of a competitors network. If you want to work with your account, you have access to the information by phone, or increasingly over the Internet, again at any time, from anywhere. When you make purchases, you use debit or credit cards, with retailers gaining authorization for payment electronically day or night. Finally, of course, the bank may use the data to customize offerings to meet your needs.

This is, of course, just one of many examples that span almost all industries, and as businesses become more 'on demand' we expect that this issue will become even more prevalent. We see many other examples today. Consider the impact that the Internet has had for data access for online retailers, or the requirement for access to patient information, whether you are the local doctor or in a Trauma Center across the country, or the impact on access to reservation systems for online bookings. The list is large and keeps on growing!

So what does this dramatic change in data accessibility require? The data needs to be accessed in many ways, by multiple applications and available to a large number of authorized users and applications around the clock and around the globe. It must be updated in real time, constantly reflecting the most recent transactions. Managing this across multiple fragmented databases, or keeping multiple copies of the data aligned is extremely complicated and can be very costly to manage. Having fewer copies of the data that all the applications and users access could be a logical way to help simplify the problem and increase efficiency. However, to achieve that, it would likely require a server that is highly scalable, very reliable and security-rich, and designed to support very high I/O rates with very high bandwidth for connectivity and for tens of thousands of users.

Making the most of your data and protecting it.

There is no doubt that the vast amount of corporate data that many companies store can be a significant asset to the business when fully exploited. It may be used to help drive down costs through exploiting new channels such as the Internet, allowing customers direct access to the data they need so as to help reduce the staffing costs needed to provide the service to your customers. It can be used to develop better information, to understand more about individual customers so that you can customize services to their needs, or to identify and understand trends in the market providing the potential for competitive advantage.

It is clear that making better use of your data can help provide significant benefits. However, increased dependency on data access for business processes along with opening up data access to a much larger number of users brings with it a critical need for data availability and protection. In today's world, the Internet has meant that users expect to be able to access information anytime of the day, from anywhere in the world. This means that your systems need to be available 24/7, 365 days of the year. If a customer cannot access the information they need, not only may you have a customer service issue, but there is a real concern that a competitor's Web site may be only a couple of 'clicks' away.

It is of course impossible to generalize about the cost of 'downtime' of applications as there are so many factors involved. Surveys have shown how cost of downtime varies by industry, from tens of thousands of dollars to millions of dollars per hour, and it would probably vary from business to business depending on size, business process design and many other variables. There are many other factors to take into account. How critical is the application to the business? Is the application internal to the business or open to customers and partners? How many users are affected? Are the transactions lost completely or deferred to a later time? If the transaction is lost, is it a customer transaction that could be gained by a competitor? What impact does that have on future potential transactions from that user?

Probably the only generalized comment that can be made regarding the cost of downtime is that there are costs to the business, sometimes substantial costs. What is also probably the case is that as businesses become ever more dependent on IT within their business and in interfacing with customers, partners and suppliers, the cost of downtime is likely to rise.

Along with the need for high availability, there is a very critical requirement for security designed to protect the data and maintain privacy. The requirements span many aspects of IT security. These aspects include helping to secure data through encryption, whether that data is being transmitted across a network or 'at rest' being stored on media such as tape; providing user identification and checking that users only have access to data that you want them to be able to access; and helping to secure a key corporate asset, the data, and the applications and the infrastructure that support the data.

There is no doubt that the public at large has a growing concern regarding protection of data. This is driven by an understanding of the amount of data that is stored by businesses about themselves and by the increasing need to provide sensitive information, such as credit card details, electronically across networks, where the ramifications for any mis-use of that data is considerable. This issue has been well understood by the legislature and there have been a number of regulatory requirements put in place designed to address data protection with varying degrees of severity by industries and countries. In some cases, the costs for non-compliance can be significant. However, nothing illustrates the importance of data protection and security features more than the highly public, headline news that has resulted from security breaches in recent years. These occurrences can have an impact on business reputation and customer trust and may have the potential to affect the both the bottom line and shareholder value.

Addressing the objectives for corporate data serving

We have seen so far that there are three critical objectives that have to be addressed for today's corporate data serving environment. Firstly, the objective to support fewer copies of data that is accessible by multiple applications across the enterprise can require a highly scalable processing environment. Secondly, the objective for access at anytime, from anywhere, can require a system designed for extremely high availability. Thirdly, the objective for data protection and privacy features can require strong security capabilities.

The latest mainframe system from IBM, the IBM System z9™, continues to evolve the technology capabilities for which the mainframe is renowned. Its strengths in areas such as scalability, availability and security are just some of the reasons why the IBM mainframe has for many years been a leader in data and transaction processing, supporting many mission critical applications. Indeed, its presence in the finance industry is a testament to these capabilities. Let's now look at each of these areas and see how the IBM System z9 performs.

Let's look first at scalability. Mainframes have for decades been viewed as 'large systems' and the latest System z9 is no exception. The combination of increased processing power as well as the ability to have 54 processors in a single server produces a top of the range model, the S54, which is almost double the capacity of the top of the previous IBM eServer™ zSeries® 990 (z990) range. Then, if you require even more capacity, in theory, up to 32 System z9 servers could be 'coupled' together utilizing Parallel Sysplex® technology, with a single application being run across multiple servers.

There is of course much more to scalability than just raw processing capacity; you need to be able to keep the processors busy. IBM Mainframes have for many years been focused on what we refer to as a balanced system design. As the name implies, this design approach is about balancing components within the server, so that in addition to scalability in processing power, there is also scalability in memory and in I/O capabilities so that the large quantities of data can be made available to support a high volume of transactions. To enable a high volume of transactions, IBM mainframes utilize a combination of virtualization and workload management capabilities designed to enable resources to be shared across applications and reallocated dynamically between them based on transaction volume and on preset prioritization according to your business needs. This combination of a balanced system design, virtualization and workload management capabilities help IBM mainframes to be capable of running at near 100% for long periods of time.

One other key element of scalability is the ability to apply additional capacity dynamically. IBM mainframes provide a broad set of functions that we refer to as Capacity on Demand. One of the capabilities that this provides, with the right pre-planning and configurations, is additional capacity that can be added either permanently or temporarily without the need to stop the system, so that you may be able to change the capacity of the system over time to match your business requirements without having to interrupt the service you provide to users.

Let's now turn to availability. In today's IT market, almost every platform from every vendor has very strong availability claims and it is easy to think that all platforms are pretty much the same. However, it is important to realize that it is possible to measure availability in many different ways. For example you could measure at a highly granular level such as processor availability, or you could measure at the server level, or you could measure the entire system. Each of these approaches, and many more, are totally reasonable, but the ramifications of the different approaches can vary substantially.

If we take the personal computer that almost all of us have on our desks, I expect the majority of us have experienced some kind of failure, the system gets 'locked', an un-recoverable error, the 'blue screen', and so on. But what caused the failure? Was it a hardware problem, or the operating system? Was it a problem in the middleware or in the tools and utilities that help support the system? From your perspective as a user, the origin of the failure is not the issue and wherever the failure takes place, the fact that the rest of the components are 100% available does not help you. Whatever the cause, the effect is the same: you cannot access your applications and proceed with the activities you planned. The ramifications are probably at least frustration, and may lead to lost productivity and missed deadlines, which may also have financial implications.

Of course in this environment, there is normally only a single user affected, but with a large corporate data server there may be tens of thousands of users affected, both inside and outside of the enterprise and the ramifications are probably much more severe. IBM mainframes' design point for availability is application availability, with availability capabilities and considerations taken across all the system elements. There is, of course, a strong focus on hardware availability, with redundancy built into a number of components from processors to power supply. But there are also availability characteristics built into the operating systems, the middleware, the tools and utilities, and the storage and networking technologies designed to enable you to access to your applications and data.

In the scalability discussion, we looked at the mainframe's Parallel Sysplex capabilities that enable a single application to span multiple servers. This capability is primarily about availability, designed so that in the very rare event of an individual server failure, the application can continue to run across the other servers within the sysplex. Not only is this approach designed to provide continued application availability in the event of such a failure, the workload management capabilities built into this environment are designed so that the reduced resource can be prioritized to your most critical workloads. IBM System z™ technology is designed for up to 99.999% availability with Parallel Sysplex clustering.

Finally, let's look at security. IBM mainframes have for many years provided a security-rich environment and have been used by enterprises around the world to support their mission critical applications. This includes the financial industry, where security requirements and applicable regulations can present special challenges. Indeed, IBM mainframes have received some of the highest levels of industry security certification. The mainframe's strengths in security stem in part from its history of supporting sensitive data for large enterprises, resulting in security features being built into its design for many decades, as well as a systems wide approach with security capabilities built into hardware, operating systems, key middleware and so on.

The mainframe's security capabilities fall largely into three categories. First, there is user identification and authorization, designed so that only eligible users can access their applications and data. As the mainframe has for many years been designed to run multiple applications simultaneously on the same server, these capabilities are well evolved. When deployed, Multilevel security (MLS) is designed to be an efficient way to help reduce the number of copies of data required while increasing the effectiveness of access management. Second, there are intrusion detection services, designed to help detect and prevent unauthorized intrusion to applications or data. Third, there is encryption, so that data that is being transmitted over the network or stored on tape for archival or distribution purposes can be protected, should that data get into the wrong hands.

IBM System z and IBM DB2® for z/OS® – synergy and strength

The foundation for these traditional strengths for data serving on the mainframe lies with the tight integration that System z and z/OS share with the database – DB2 for z/OS. DB2 for z/OS is written to exploit the System z platform and as a result can offer advanced features and function. IBM DB2 for z/OS delivers rich function for highly scalable, industry-leading high availability IT infrastructure for your enterprise data and on demand business applications. The combined power and capacity of IBM System z with the high performance and availability of the z/OS operating system, and the strength of the DB2 for z/OS data server can expand and extend your IT infrastructure and the business value of your data. The combination of DB2 and System z provides a unique competitive advantage for on demand environments by providing a flexible, cost effective and optimized foundation for Information on Demand. This foundation allows you to better manage risk, supports your efforts to demonstrate compliance with policies and standards, and helps to simplify management of your information infrastructure. These capabilities are important to enable customers to use their core business data to drive insight and gain competitive advantage.

DB2 for z/OS – foundation for enterprise data serving

DB2 for z/OS Version 8 is the most robust release to date with strong customer acceptance. Key improvements include enhanced scalability, easier application development and porting, even stronger security, and significant enhancements to help move closer to the goal of continuous availability. Management of very large databases and 64-bit virtual storage support make systems management simpler, while continuing to lead the industry in scalability and availability. This new version broke through many old limitations in the definition of DB2 objects, including SQL improvements, on-line schema evolution, longer names for tables and columns, longer SQL statements, enhanced Java and Unicode support, enhanced utilities, support for more log data sets, support for more partitions, and more. Extensive SQL enhancements provide for even tighter DB2 family compatibility, which in turn helps ISV and application developers deliver and port applications quickly and efficiently. DB2 continues to provide industry-leading availability by minimizing planned downtime. Database changes, such as adding a partition and upgrading to a new version can now be done without an outage – a breakthrough for availability. Multilevel security at a row level, improved flexibility for on demand business and new encryption options provide a breakthrough for security. Version 8 and its enhanced suite of tools is a significant milestone, addressing significant customer requirements and allowing for ongoing incremental database and application growth.

DB2 has used the function of the System z and z/OS platform extensively for many years and plans to continue to do so with future releases. The tight integration between hardware, software, and z/OS drives unmatched capabilities that are core to delivering a flexible, cost effective IT infrastructure that can deliver capabilities needed to help you to better manage risk, to support efforts to demonstrate compliance with policies and standards, and to achieve lower cost thru an optimized IT infrastructure. These capabilities are important to support key information-intensive applications that provide business advantage. Our plans are for future releases of DB2 for z/OS to build on the foundation of DB2 for z/OS Version 8, providing increased integration with the DB2 family and with the System z9 platform itself. New trusted security context and database features are expected to expand on the concept of granular security first offered in DB2 for z/OS V8. Fast table replacement (a.k.a. cloning) is targeted to help improve the availability of many Web-based applications. Improvements are planned to SQL, optimizers and algorithms to help improve performance.

We have seen how DB2 and the mainframe have strong capabilities to help address today's data serving requirements. However, the mainframe's evbolution in recent years has involved much more than just these traditional capabilities. The introduction of a broad set of open and industry standards such as Linux® and Java™ along with the introduction of new middleware such as WebSphere®, has provided a much more flexible environment that can allow greater interoperability with other platforms and may help improve access to the wealth of data on DB2 and to the mission critical applications that are key elements of the mainframe environment.

Summary

In today's on demand world, providing the support for internal and external clients requires an increasing dependence on IT, with more users and more applications requiring access to a common set of data in real time that is accessible anytime, from anywhere across the globe. So how do you meet these needs? The answer will, of course, be dependent on your individual situation and requirements; there is no 'one size fits all' answer for data serving. However, if you feel that meeting your individual situation requires consolidating to fewer copies of data, running on highly scalable servers, while also providing a high level of security and availability to provide the service your users demand, then the new IBM System z9, which builds upon the inherent strengths of the IBM mainframe to deliver industry-leading data and transaction serving capabilities, may well be the right choice for your data serving needs.

IBM Corporation
New Orchard Rd.
Armonk, NY 10504
U.S.A.

Produced in the United States of America
01-06

ZSW01811-USEN-00