Protect your data with real-time data replication
April 2008

IBM Information Management software

Selecting a data distribution solution for IBM System i environments

Contents

Why distribute data?

Today, more than 200,000 businesses in more than 100 countries have come to depend on the stability and security of IBM System i™ servers. Renowned for its reliability and scalability, System i servers are used to support a business' most demanding operations. Not only do they handle large workloads and data volumes but they are also commonly used to run the world's most critical business applications including MAPICS, J.D. Edwards, BPCS, ADP, Geac, SilverLake, and many others.

With the widespread deployment and broad use of System i applications, as well as their reputation for performance and reliability, it's no surprise why DB2 applications are used to power the largest banks in England, hospitals in Germany, manufacturing sites in China, and retail stores in America.

Given the importance of these applications and the critical nature of the business operations they run, it is imperative to protect the security and scalability of System i applications while at the same time recognize that these systems house critical operational data that must be distributed throughout the organization.

System i organizations are anxious to find effective ways of leveraging their systems and applications and, at the same time, make efficient use of their hardware resources and increase their network performance. One of the most successful ways to achieve this is through data distribution. With data distribution, transactional data from enterprise resource planning (ERP) systems, customer relationship management (CRM) systems, point-of-sale (POS) systems and other mission-critical applications that run a business can be mirrored in a distributed environment. This allows users to access a single up-to-date view of critical business information that is spread throughout applications that are physically housed in different departments, regions or business units. Users can access applications based on function, business unit or geographic region to improve performance and maximize IT resources and bandwidth.

There are several ways organizations are using data distribution to make the most of their System i investments.

Workload Distribution

According to Gartner, the rising energy demands of technology have raised concerns over electrical power consumption. Organizations need to better utilize their existing equipment, and workload balancing is a viable solution. Through bi-directional mirroring, workload distribution allows data to be on more than one machine with users divided between them. This can help reduce the cost of maintaining a fragmented IT environment by allowing incompatible applications to co-exist. Data and System i objects remain current and synchronized by capturing changes as they occur in the source application and distributing them to one or more secondary applications. Users accessing each dispersed application have access to the same information for consistency across the organization. Organizations can then maximize the productivity of dispersed applications and minimize network bandwidth usage by allowing users to access local applications with accurate corporate data.

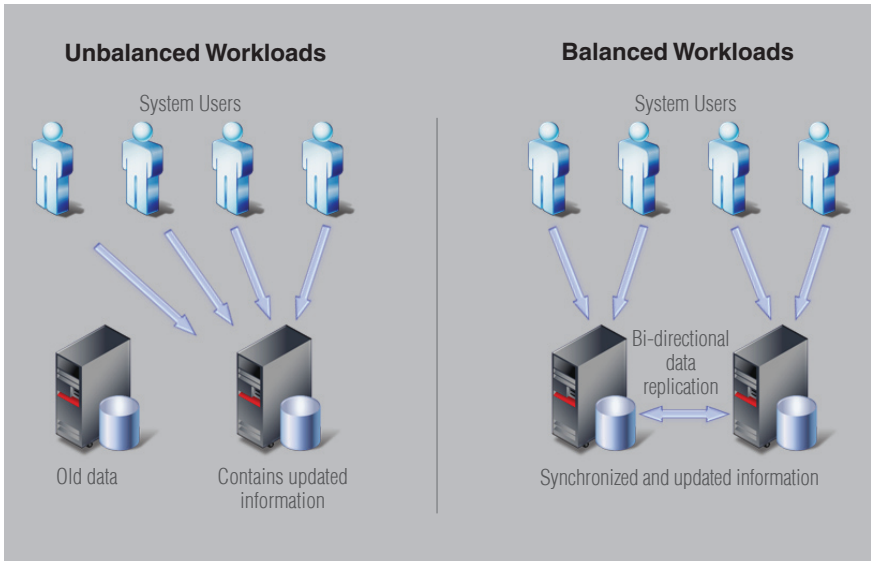

Figure 1: Architectural overview of workload distribution

Workload balancing of read-only processes

As data is accessed from production applications for reporting, querying, and analysis, users often experience slower response times due to high overhead, delayed application processing, and reduced system performance.

When workload balancing is put in place, data is mirrored within enterprise applications to a secondary system. Users can run reports and resource-intensive queries on the secondary system quickly and efficiently without impacting the performance of the source enterprise application, yet still performing reports on the most recent data.

They can also offload resource-intensive processes to a secondary system, allowing the primary system to perform its core functions. For instance, sales staff can still access the organization's primary customer resource management system while IT is offloading data to a secondary box for reporting or backup processes. This can possibly help reduce the cost for companies by balancing workloads across existing resources instead of purchasing new systems.

Uninterrupted business operations during rolling upgrades

In an environment with both a production and test environment, realistic and live test data can be refreshed and mirrored to the test system through data distribution. Changes can be made on the test system without affecting the production data. This also allows users to continue using legacy applications during migrations to the new system. The result is easier administration and maintenance since business operations will not be interrupted as systems are taken offline for maintenance activities. This also reduces the risk to applications as rolling upgrades allow users to continue to use the primary server as users are slowly migrated to the new environment. Full testing can be completed with the old and new applications working from the same data.

Challenges to data distribution

There are several challenges that IT departments face while trying to implement data distribution.

Batch process limitations

Integration tools move batches of data between production systems. For many organizations; running a nightly batch process when systems are not running at full capacity, works well. However, as businesses are operating worldwide through all time zones and as more business is conducted 24/7 over the Internet, finding a time when production systems can be taken offline to run a batch extraction of data becomes increasingly difficult. In addition, conducting batch integration means that data used by applications or for reporting between batch loads is outdated and stale.

Risk to production systems

Integration tools often require changes to production systems – for example, adding a date/time stamp to allow incremental updates. For many corporations, the additional risk this would pose to their mission-critical systems is intolerable. In addition, this makes integration processes more difficult to maintain since database and application changes must be maintained as applications get upgraded and integration requirements become more complex and require additional data sources.

Keeping systems in sync

Data stored throughout a distributed environment must be kept up-to-date and consistent. It is critical to synchronize data between various applications such as customer relationship management (CRM) tools, financial systems, and enterprise resource planning (ERP) applications to ensure all critical data is consistent throughout the organization.

Maximizing the performance of it resources

As distributed applications are used to make the most out of limited IT resources, application and network performance as well as response times become concerns in implementing applications in a distributed environment. Data must be shared to minimize impact on systems and must appear seamless to both internal and external users.

Data integrity throughout distributed systems

The integration process must be able to distribute data changes and guarantee data consistency to ensure all users, regardless of their region, business unit or department. If data is entered incorrectly or not in the order it occurred in the original source application, users risk acting on incomplete or inaccurate information. Secondary applications can even reject data if it is not in the proper format or processed in the correct order.

Conflict detection and resolution in a distributed application

With changes to data being done on each system throughout the distributed application, there are times when multiple users may be simultaneously accessing the same information. Applications must be able to integrate data as it changes and be able to detect and resolve conflicts as they occur to ensure data integrity.

Awareness of data distribution status

System i is used to run mission-critical applications. Visibility into the data distribution process is imperative not only for peace of mind, but also for problem detection and diagnosis.

IBM Information Server

A revolutionary data integration platform

IBM Information Server is a revolutionary data integration software platform that profiles, cleanses and integrates data from heterogeneous sources. It provides a metadata-driven approach to data integration by acting as a buffer between operational systems and enterprise data warehouse. It is the industry's first comprehensive data integration platform that includes services for data discovery, transformation and cleansing. This ensures that the right data is identified, aligned and delivered to the data warehouse model, business intelligence applications or business processes.

IBM Information Server enables customers to understand and integrate their data assets in order to deliver consistent, complete and trustworthy information for key business and IT projects. It helps customers expedite their data governance and master data management projects with trusted data.

IBM Information Server capabilities:

The IBM Information Server platform consists of many technologies including data profiling and cleansing, transformation, federation, replication and event publishing; all built on a common platform. The unified platform delivers four integration functions:

Understand

This function enables users to gain a deep understanding of their data by discovering the content, quality and structure of their source systems. Truly understanding the meaning, relationships and lineage of data upfront, helps speed downstream development tasks, reduces the risk of proliferating bad data and eliminates the cost of scrap and rework due to lack of understanding of data sources.

Cleanse

This function involves cleaning problems that users find in information by standardizing and matching records across different systems to get a clean, accurate and consistent view of data.

Transform

This function takes information out of its original context within individual source systems and allows information to be used in a new context to solve new business problems. This often involves combining information from many systems to get a more aggregated, enriched and complete view.

Deliver

This function enables the delivery of information across the enterprise based on the requirements of the business. Delivery mechanisms include federation, replication/synchronization and changed data capture.

Underlying these functions is a common metadata and parallel processing infrastructure that provides leverage and automation across the platform. Each product in the portfolio also provides connections to many data and content sources, and the ability to deliver information through a variety of mechanisms. Additionally, these functions can be leveraged in a service-oriented architecture through easily published shared services.

Change Data Capture to enhance IBM Information Server

There are a number of approaches to data distribution dilemmas, but they all involve capturing change data in a source system for distribution to one or more target systems.

Many organizations use data integration tools that query the systems for changed data. To keep the systems in sync, these queries must be run frequently throughout the work day. These potentially large queries may slow application response times and network performance in large systems.

Some organizations take a trigger-based approach to capture the changed data: a change to a specific database table will invoke the trigger, which usually writes the changed data to another database table known as a change table. However, this is not a feasible approach for customers with high volume transactional environments as the overhead required to write and scan the change table for new entries is often unacceptable.
Some organizations have very strict database policies, according to which no changes are allowed to the database. This means they cannot add triggers to the database tables.

Other organizations adopt a log-based Change Data Capture (CDC) approach. Rather than using triggers or performing queries against the database, real-time, log-based CDC technology addresses the primary challenges with any data integration initiative: real-time performance and low impact on source systems.

The best log-based CDC solutions take advantage of the System i DB2® journal to track changes in the database. And since operational systems already have massive data volumes to process, CDC solutions do not add extra load or require any changes to existing source systems as its reads the logs directly.

Further, data distribution solutions replicate both data and object-level changes which come from both the database journal and the audit log. The best CDC solutions ensure that changes are replicated to the secondary system in the order that they were applied on the primary, thereby ensuring data consistency.

Unlike batch processes, data distribution solutions replicate business transactions in real-time through remote journaling as well as a proprietary scrape and send process. With scrape-and-send, two independent processes must take place. All database changes must first be captured into a log – or must be "journaled." The appropriate journaling configuration for a customer's business environment generates minimal CPU overhead on the source node and minimal performance impact to the customer's business applications. The second process reads the transaction log sequentially and pushes transactions to the backup node. As each business transaction is received on the backup node, it is applied to the backup volume and a confirmation is sent back to the production node. The strengths of real-time asynchronous mirroring include low communications bandwidth, low impact on business applications and high throughput.

Real-time, log-based CDC and data distribution combined

To maximize the benefits of distributing data in System i environments, IBM solutions provide high performance, real-time, log-based CDC to distributed DB2 environments.

IBM InfoSphere™ Change Data Capture, a real-time, log-based CDC solution, meets the challenges raised in data distribution implementations. The powerful real-time data integration solution mirrors critical DB2 application data and objects in real time from a primary system to one or more secondary systems so organizations can distribute processing load and network use across multiple systems.

IBM InfoSphere Change Data Capture business benefits are many:

Fast performance & scalability

For customers with large volumes of data, performance is a critical differentiating factor in selecting a data integration product. DataMirror solutions achieved more than two billion transactions per hour on IBM's eServer i5 systems at a benchmarking test at the IBM Benchmarking Center in Rochester, Minnesota. The benchmark results are a testament to the delivery of high throughput and impressive scalability for large data volumes.

Data integrity

Data distribution is useless if replicated data is out of sequence. As InfoSphere Change Data Capture technology scrapes transactions from the data journals, they are sent across a TCP/IP communications link using a send-and-receive process. If errors occur during replication, the solution can recover at the point of failure – ensuring the most efficient recovery process and it restarts at the same point to guarantee every transaction on the source gets replicated in the same order on the target.

Bi-directional conflict resolution

The solution provides workload balancing and data distribution that allows both the source and target files to be updated by customer applications and this raises the need for conflict detection and resolution. A conflict is defined as any operation attempted on the target where the target system is not in the same state as the source system. If a conflict is detected, the add-on module provides four rules for its resolution: source wins, target wins, none (file is suspended), or custom-defined through a user exit. Thus, users have the ability to easily manage and resolve conflicts within their business environment.

For more information

For more information about IBM Information Server, contact your IBM marketing representative or visit **ibm.com**/software/data/integration

IBM Software Group
Route 100
Somers, NY 10589

Printed in the United States
April 2008

TAKE BACK CONTROL WITH **Information Management**

IMW14007-USEN-01