INFORMATICA®
The Data Integration Company™

The Business Case for Superior Data Integration Performance and Reliability

How Your IT Organization Can Handle Growing Data Volumes and Business Demands by Extending Your Informatica PowerCenter Environments

WHITE PAPER

This edition published September 2009.

Table of Contents

Executive Summary

For more than a decade, IT organizations have strived to improve the performance and reliability of their enterprise data integration infrastructure in the face of growing data volumes and business demands. They have used a variety of approaches that have proved time-consuming and costly. For instance, third-party data accelerators may speed performance, but they also introduce a new layer of complexity and administrative overhead. Manual development and tuning of load-balancing algorithms in a multi-node grid computing environment may improve reliability, but they consume scarce IT staff resources and can be prone to errors.

Improving data integration performance and reliability doesn't need to be a large-scale and painful IT initiative. Users of Informatica® PowerCenter® enterprise data integration software have at their disposal an array of innovative features and options engineered to enhance performance and reliability.

This white paper examines the importance of data integration performance and reliability and explains how features and options of PowerCenter can help your IT organization handle increasing data volumes and meet business demands for timely, trusted information. These features and options include:

- **64-bit Architecture:** Harness the power of 64-bit server technology with an upgrade from 32-bit versions of PowerCenter.
- **Grid Computing:** Distribute data integration workloads across multiple servers in a grid environment with the PowerCenter Enterprise Grid Option™.
- **High Availability:** Ensure continuity of data integration processes in the event of a hardware, software, or network outage with the PowerCenter High Availability Option™.
- **Partitioning:** Speed execution of data integration workloads with dynamic realignment in parallel environments with the PowerCenter Partitioning Option™.
- **Pushdown Optimization:** Exploit database processing power for large data volumes and complex transformations with the PowerCenter Pushdown Optimization Option™.

These features and options enable your IT organization to build on PowerCenter's industry-leading performance to gain new flexibility in building data integration infrastructures that can reliably scale in the face of growing data volumes and real-time business needs.

This white paper also illustrates how real companies are maximizing their PowerCenter investments. For example:

- More than 88 percent of companies that tuned their 64-bit versions of PowerCenter doubled its performance in terms of throughput and scalability
- Companies upgrading to 64-bit versions of PowerCenter cite two key drivers: very large data volumes and service level agreement (SLA) compliance
- On-line marketer LinkShare realized massive data throughput for near-real-time operations using its 64-bit version of PowerCenter in a highly available grid
- Brazilian telecommunications provider Oi slashed in half the time needed to refresh its 20-terabyte data warehouse with the PowerCenter Pushdown Optimization Option

The Need for Greater Data Integration Performance and Reliability

Your enterprise data integration infrastructure needs to provide the performance and reliability your IT organization needs to meet the demands of the business.

Data integration technology is no longer merely a tool to update a data warehouse with a weekly batch load. Today, data integration technology lies at the center of mission-critical processes in every industry. Companies depend on data integration technology to power data transformation and exchange among core financial, customer, and supply chain operational systems.

Other companies operate near real-time enterprise data warehouses (EDWs) and operational data stores designed to deliver timely and consistent information to decision makers, support staff, and external customers and partners. Data synchronization, consolidation, and migration initiatives demand high performance and reliability to support 24x7 global operations.

As IT organizations strive to keep pace with growing business requirements, they face another challenge—growing data volumes. A June 2009 Forrester Research report says that in 2009, Forrester sees anecdotal evidence that approximately more than two-thirds of today's EDWs contain between 1 TB and 10 TB of data. "Forrester estimates that by 2015, a majority of EDWs in large enterprises will be 100 TB or larger, with petabyte-scale EDWs becoming well-entrenched in such sectors as telecommunications, finance, and Web commerce,"[1] Forrester's report says.

The economic downturn has raised the stakes for data integration performance and reliability. In an uncertain climate, speed in reacting to changing market conditions is critical. Agility in meeting customer demands for value and service is paramount as consumers curtail spending.

> "Forrester estimates that by 2015, a majority of EDWs in large enterprises will be 100 TB or larger, with petabyte-scale EDWs becoming well-entrenched in such sectors as telecommunications, finance, and Web commerce."
>
> — "Massive But Agile: Best Practices For Scaling The Next-Generation Enterprise Data Warehouse," Forrester Research, Inc., June 2009

[1] Forrester Research, "Massive But Agile: Best Practices for Scaling the Next-Generation Enterprise Data Warehouse," June 2009.

Performance and reliability are vital to meeting these objectives—and to cope with surging growth in data volumes and complexity. A high-performing and reliable data integration technology is one that handles:

- **Any data volume.** he explosion in data volumes is forcing enterprises to address the performance and scalability of their data integration technologies. This issue will remain at the forefront as organizations accumulate petabytes of data generated from their operations worldwide and struggle to leverage that data for insights and decision making.
- **Any data type.** As volumes grow, so does data complexity. The typical enterprise manages dozens or hundreds of mission-critical source systems in a heterogeneous environment. This means an overload of datatypes, structures, and formats, which can compromise performance and overwhelm ad hoc data integration solutions.
- **Any data latency.** IT organizations must manage a wide spectrum of time frames and latencies for data integration, depending on the application and use. The timing can range from weeks and days to seconds. IT organizations need the flexibility to deliver trusted, high-quality data whenever applications or users need it—whether in real time, batch, or changed data capture (CDC).
- **Any role.** Many different people are involved in data integration projects: data stewards, data analysts, architects, administrators, and developers. They have different skill sets and different tasks to accomplish. At the same time, they all need to work together to share artifacts and tasks, increase cross-team productivity, and ensure that IT results align with business needs.
- **Any time.** Data must be made highly available. Downtime in data integration processes can mean costly interruptions, customer dissatisfaction, and violation of service-level agreements.

The recession has forced business and IT alike to scrutinize operations for inefficiencies and target affordable, low-risk investments with the potential for high payback and a tangible impact on business fortunes. Amid the downturn, business leaders are looking to IT to help manage initiatives that enable them to weather the challenging times and to lay a foundation for long-term prosperity.

Data is at the core of many of these business initiatives. A high degree of data integration performance and reliability can help IT organizations enable their companies to:

- Manage risk and compliance
- Improve operational efficiencies
- Streamline and quickly realize value from mergers and acquisitions
- Modernize the business
- Strengthen customer relationships

Managing Risk and Compliance

With risk management on the radar, companies have to quickly zero in on market and business dynamics that could jeopardize their positions. Accessing and monitoring activity in near real time from vast and fast-moving data flows is increasingly important to mitigating risk and ensuring compliance with SLAs and regulatory requirements.

Improving Operational Efficiencies

In an economic downturn, companies have to do more with less. They're looking for ways to coax greater efficiencies from business processes and IT systems, even as spending is curtailed. Greater data integration performance can compress business cycles by making more data available, faster, across more channels. Improved reliability translates into higher productivity and minimizes the risk of costly downtime.

Streamlining Mergers and Acquisitions

In a merger or acquisition scenario, companies need to quickly consolidate data from the new entity. Speed is critical. Performance comes into play in such areas as real-time synchronization between disparate systems and processing data volumes that can double in a merged entity. Scalability and reliability are essential to fully and swiftly leverage the combined data assets of two entities.

Modernizing the Business

Many IT organizations are transitioning from legacy environments that are increasingly expensive to maintain, complex to manage, and not flexible enough to accommodate today's business demands. A transition to highly reliable grid computing environment offers an opportunity to increase data integration performance by distributing transformation processing across multiple nodes of low-cost commodity hardware.

Strengthening Customer Relationships

Every customer counts—especially in a tough economy. To attract and retain customers, companies must understand their customers better. They need more speed and precision in analyzing customer needs and meeting their expectations. Improving customer service, satisfaction, and cross- and up-sell rates requires a single view of the customer. To deliver this single view, IT organizations need to create a customer data hub, consolidating, synchronizing, and managing customer data from different applications and systems. And to meet the requirements of today's ultra-wired businesses and consumers, IT organizations must make customer data available on demand across diverse channels.

> "In these times of economic uncertainty and intense competition, the cost of each misstep is magnified, and each missed opportunity could bring the enterprise a step closer to disaster."[2]
>
> — Carl Olofson
> Research Vice President,
> Information Management and
> Data Integration Software, IDC

[2] IDC White Paper sponsored by Informatica, "Maximizing Opportunity and Minimizing Risk Through Enterprise Data Integration: Strategies for Success in Uncertain Times," Doc no. 217393, May 2009.

Optimizing Data integration Performance and Reliability

Data integration performance is usually defined in terms of throughput and scalability:

- **Throughput** measures how rapidly a given quantity of data can be processed—for instance, reading, transforming, and loading 10 GB of data in x number of minutes.
- **Scalability** refers to the capacity to accelerate throughput, and/or handle greater data volumes, by modifying the software and hardware environment in which data integration processes are executed. For example, in the ideal of near-linear scalability, doubling the number of server processors will in theory double throughput.

Exponential data volume growth has made performance harder to measure. For instance, an SLA may require that IT process 10 GB of data in x minutes. When your data volumes are growing from 10 GB to 100 GB to 500 GB and into the terabytes range, it's harder to process more and more data at a consistent pace.

You don't need to rip and replace your existing systems. All PowerCenter features and options are designed to work with what you own now, so you can make the most of your existing IT infrastructure investments.

Improving data integration performance is not always straightforward. Your IT organization needs to consider use, datatype and complexity, partitioning options, sources and targets, hardware and network infrastructure, and other factors. Rarely do you find a one-size-fits-all solution to accelerating throughput and building scalability.

To a lesser extent, improving the reliability of a data integration infrastructure also depends on variables. In some cases, simple failover of data integration processes to a secondary compute node may be sufficient. In certain mission-critical uses, IT organizations may be inclined to build in more advanced capabilities for resilience and recovery.

An array of features and options available with Informatica PowerCenter gives your IT organization flexibility in strategically implementing performance and reliability enhancements that zero in on your most pressing business and IT needs. Informatica Professional ServicesSM personnel and sales engineers are available to assist with strategic assessments, technology selection, and proofs of concept.

Whatever course you chose, you don't need to rip and replace your existing systems. PowerCenter's features and options are engineered to complement your existing environment and help your organization improve the use of all its IT resources—servers, data integration and other enterprise software, and the IT personnel who run your data infrastructure. These technologies include:

- 64-bit architecture
- High availability
- Grid computing
- Pushdown optimization
- Partitioning

Let's take a look at how these technologies may be implemented to help your IT organization achieve its objectives and capture the payback from greater data integration performance and reliability.

64-Bit Architecture

If your IT organization upgraded its servers and operating systems in the past few years, chances are you're on a 64-bit architecture. Now standard in low-cost x64 servers, 64-bit technology expands beyond the 4-GB addressable memory limit of 32-bit systems and enables more compute processes to execute in memory, versus the slower disk.

But 64-bit hardware is only half the equation. To harness the power of 64-bit technology for data integration, it's necessary to upgrade to a 64-bit version of Informatica PowerCenter. The latest version of PowerCenter ships with built-in support for 64-bit architecture.

It's important to recognize that performance gains realized from upgrading from 32- to 64-bit technology will depend on the type of data integration process involved. The acceleration of simple, routine mappings that do not depend heavily on memory may be nominal.

But performance gains of two and three times faster are common for larger, memory-intensive transformations. From a business perspective, these can involve real-time data sourcing and integration to support mission-critical customer-facing and financial systems and enterprise data warehouses.

Tuning a 64-bit PowerCenter deployment can deliver even greater performance. More than 88 percent of Informatica customers who tuned 64-bit PowerCenter deployments increased performance by two times or more, according to a customer survey conducted in May 2009 (see **Figure 1**).

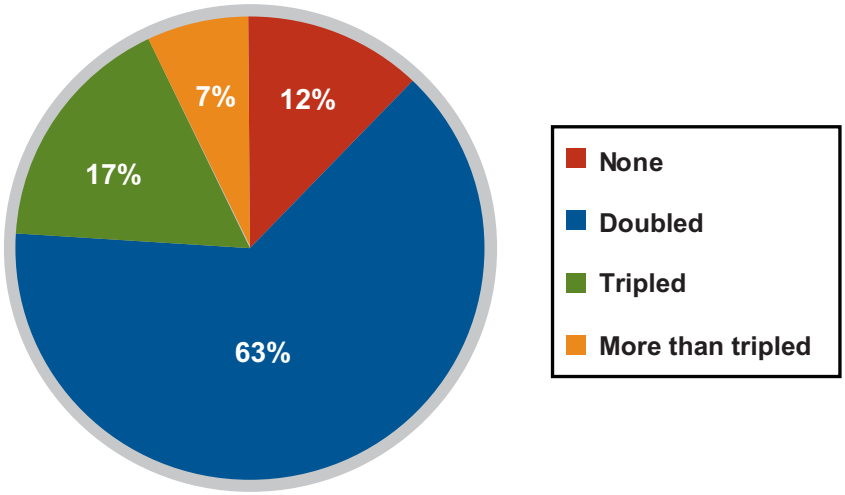

Figure 1. Degree of Performance Improvements that PowerCenter Customers Observed from 64-bit Performance Tuning

"PowerCenter helps produce reliable data for L'Oréal's HR function, whether for the social balance sheet in France or the multiple indicators that help the HR teams run their daily activities. The transition to the 64-bit version will result in reduced processing times, which are essential to deliver data on time within a group that operates on five continents."

— Philippe Bot
Director of HR Corporate Information Systems, Office of Human Relations, L'Oréal

The survey also found that processing very large data sets and meeting SLAs for performance and scalability were key reasons why customers upgraded to a 64-bit version of PowerCenter. (See **Figure 2**. Note: the numbers in the chart do not total 100% because multiple answers were allowed).

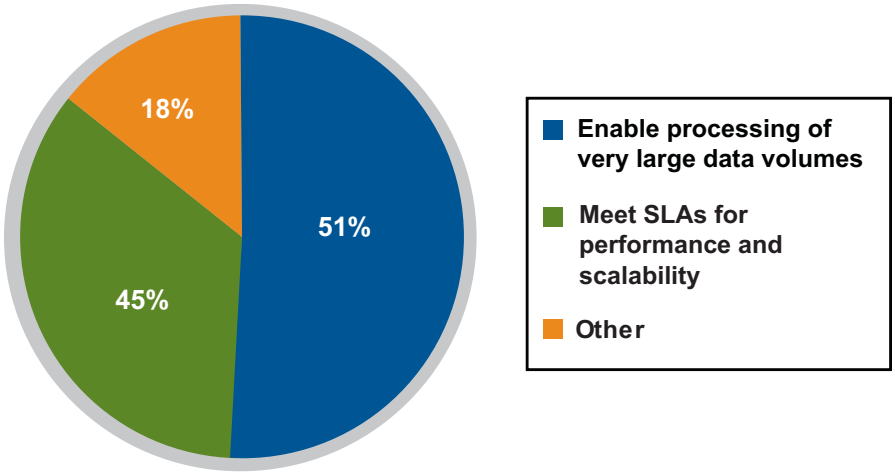

Figure 2. PowerCenter Customers' Business Drivers for Upgrading from 32- to 64-bit Processing

High Availability

In concert with high performance, high reliability is increasingly important in a volatile economic climate. Your company simply cannot afford the cost and risk of interruption to data integration processes. In addition, many organizations now run data integration in mission-critical roles, powering data exchange across operational systems, business units, and geographic locations.

The PowerCenter High Availability Option can give your IT organization peace of mind. This option helps ensure reliability by enabling data integration processes to seamlessly fail over to available servers in a multinode environment in the event of a hardware, software, or network failure. The High Availability Option:

- Minimizes risk of costly downtime and SLA violations
- Supports mission-critical requirements of operational data integration
- Automates availability, resilience, recovery, and restart

Grid Computing

Grid-based infrastructures, consisting of low-cost commodity servers, have emerged as an economical alternative to traditional symmetric multiprocessing (SMP) servers. Many IT organizations have transitioned to grid-based data centers to support operational applications and databases.

With load balancing technology, these distributed architectures typically improve performance by distributing compute processes across multiple servers according to available processor, memory, and disk resources.

With the PowerCenter Enterprise Grid Option, your IT organization can take advantage of the greater performance and scalability of a grid environment for data integration. This technology automatically optimizes data integration across multiple servers, sparing developers from manually configuring PowerCenter to function in a grid environment. The Enterprise Grid Option:

- Improves performance by automatically balancing data integration loads across available grid resources
- Allows unlimited PowerCenter scalability with additional grid servers
- Increases capacity to handle peak processing spikes and large data volumes

CASE STUDY: MASSIVE DATA THROUGHPUT

Founded in 1996, LinkShare had over a decade built a global, 24x7 data infrastructure with a near-real-time operational data store and enterprise data warehouse. The massive system involved more than 700 data integration sessions, 500 mappings, and 100 workflows.

As LinkShare grew its pay-per-action Web marketing network to millions of partnerships among Fortune 500 and other companies, its data volumes soared—and so did its need for greater data integration performance. The company deployed the 64-bit version of PowerCenter in a highly available grid.

With a PowerCenter upgrade and a revamped infrastructure of Linux-based IBM servers, LinkShare achieved breakthrough data integration performance—and capacity to grow into the future.

"We were able to really, really increase our throughput substantially through the combination of load balancing, source partitioning, dynamic partitioning, session on grid—all in a 64-bit architecture," says Dave Ramos, LinkShare director of Business Intelligence and Analytics. "You always want bigger, better, faster, and Informatica's definitely been able to provide that for us."

Pushdown Optimization

Increasing the performance of data integration processes often means leveraging all available Increasing the performance of data integration processes often means leveraging all available resources. With the PowerCenter Pushdown Optimization Option, your IT organization can tap into the power of your relational databases to increase performance.

This innovative technology can dramatically speed data integration by splitting transformation processing between a PowerCenter data server and a relational database in which source and target data resides. By "pushing down" processing to the database level, this option gives you new flexibility to push the performance frontier and scale to peak demands. The Pushdown Optimization Option:

- Exploits database processing to improve performance and scalability
- Maximizes utilization of IT investments
- Dynamically creates and executes database-specific transformations

CASE STUDY: TELCO DOUBLES PERFORMANCE

Oi, the leading Brazilian telecommunications provider, has doubled its data integration performance with the PowerCenter Pushdown Optimization Option.

Oi relies on the Informatica data integration platform to support a Teradata-based enterprise data warehouse geared to drive sales and marketing with a single customer view. With greater performance, Oi sales and service professionals have faster, accurate data on the company's 31 million customers.

By adding the PowerCenter Pushdown Optimization Option to its environment, Oi slashed the hours required for daily loads of more than 1 TB of data into the 20-TB warehouse in half. "This performance improvement allows us to provide quicker answers to marketing and sales teams," says Vera Duarte, Oi enterprise IT manager.

The option enabled Oi to achieve immediate performance gains and scalability to handle increasing data volumes. It builds on Oi's success with the Informatica Platform, instrumental in integrating heterogeneous data from 16 separate companies when the company was formed.

Oi also uses the PowerCenter Enterprise Grid Option to process 150 million call detail records daily in a 64-bit architecture based on two 16-core Linux servers, well within SLA requirements. And the Informatica Platform helps power a customer-facing billing system that has reduced credit recharge time by 60 percent and improved response capacity by an incredible 400 percent.

Partitioning

Partitioning is another weapon in the performance battles against mounting data volumes and to accelerate data-driven business processes. The PowerCenter Partitioning Option divides data processing into subsets that run in parallel across available CPUs in multiprocessor and grid-based systems.

By spreading the computational load across available IT resources, large data volumes can be processed faster. Data integration processes are executed in parallel, rather than sequentially. And unlike manual data partitioning, the PowerCenter Partitioning Option ensures data integrity because its parallel engine dynamically realigns data partitions for set-oriented transformations. The Partitioning Option:

- Enables hardware and software to jointly scale to handle large data volumes
- Supplies parallelization flexibility with multiple partitioning mechanisms
- Equips developers with design and tracking tools to optimize performance

Conclusion

In time, the recession will end. Economic growth will resume. But there's no end in sight to the growth of your data volumes. To help your IT organization continue to meet business demands and SLAs, it needs superior performance and reliability of data integration processes.

Think a few years down the line. Or 5 or 10 years—how much data will your organization have? How important will leveraging that data rapidly and reliably be to achieving your business objectives? What new business requirements will your organization face that demand timely, trusted, and consistent data?

Superior data integration performance and reliability are distinct strategic advantages that bolster chances for both short-term survival in uncertain economic times and long-term prosperity in the years to come.

Informatica makes it possible for your organization to readily and economically take advantage of PowerCenter features and options for high performance and reliability. With them, you gain a powerful means of harnessing the power of data for greater agility, insight, and competitive advantage, in good times and bad.

Learn More

Learn more about Informatica PowerCenter features and options for superior data integration performance and reliability. Visit us at www.informatica.com/powercenter or call 1.800.653.3871.

About Informatica

Informatica Corporation (NASDAQ: INFA) is the world's number one independent leader in data integration software. The Informatica Platform provides corporations with a comprehensive, unified, open, and economical approach to lower IT costs and gain competitive advantage from their information assets. More than 3,600 companies worldwide rely on Informatica to access, integrate, and trust their information assets held in the traditional enterprise and in the Internet cloud.

INFORMATICA®

Worldwide Headquarters, 100 Cardinal Way, Redwood City, CA 94063, USA
phone: 650.385.5000 fax: 650.385.5500 toll-free in the US: 1.800.653.3871 www.informatica.com

Informatica Offices Around The Globe: Australia · Belgium · Canada · China · France · Germany · Japan · Korea · the Netherlands · Singapore · Switzerland · United Kingdom · USA

7037 (09/01/2009)