

# Benefits of Transformational Data Integration

**DataMirror**<sup>®</sup>  
The experience of now.<sup>™</sup>

# Benefits of Transformational Data Integration

## Executive Summary

In the new economy, data drives all business. Dynamic databases for real-time pricing and inventory make today's e-Business experience. Data is precious and making sure it's available to your customers 7 days a week, 24 hours per day is critical. Companies across all industries look to manage and analyze their data to help them achieve competitive and customer insight.

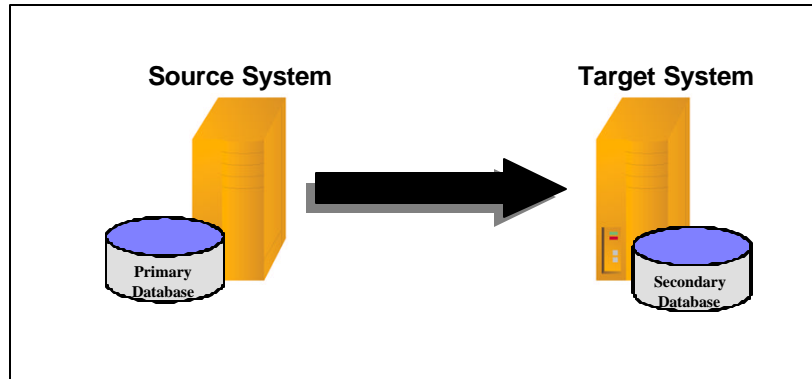
Only companies that can effectively manage, integrate, distribute and utilize their data assets will survive and prosper. Businesses require the ability to manage all the data in the organization no matter where it resides—whether on web servers, legacy and operational systems, diverse data stores or applications. Companies must build an effective data integration infrastructure that ensures data is in the right place at the right time, and in the right format for business intelligence, e-Business and other business applications.

The challenge is how to access, manage and manipulate the mountains of business data collected by computers every day—customer data, financial data and Internet click-stream data. According to some sources, the world has generated more data in the last 30 years than it has in all the preceding 5,000 years. This data often resides in different databases on different hardware platforms and operating systems. The proliferation of mixed system environments and incompatible data stores have made the goal of ubiquitous data access seem difficult or even impossible to attain.

The need for capture, transform, and flow (CTF) data integration software that can integrate data across a range of databases and computing platforms is critical. Organizations can choose from a number of technologies that provide data movement on an enterprise-wide basis. This white paper outlines the terms and concepts associated with moving, sharing, enhancing and integrating enterprise data. Additionally, this paper discusses capture, transform and flow (CTF) technology available today that enables bi-directional cross-platform, cross-database integration between diverse computing platforms and database technologies. It offers an informative approach to evaluating real-time data capture, transform, and flow data integration solutions.

## Data Integration Defined

Data integration, in its simplest form, is the capture and movement of data from one database or flat-file on a source system to another database or flat-file on a target system. Most companies are already performing some type of data integration, whether it is backing up network servers, performing mirroring for high availability, or replicating key databases to business intelligence applications. But with the sheer number of integration tools on the market, along with the different approaches each employs, choosing an effective solution can be both difficult and time consuming.



**Figure 1:** Simple integration from source system to target.

For example, basic database copy tools are available that take entire databases and move them on a full-refresh basis. As databases grow, however, the time to refresh the target machine increases. Even if the database has relatively few changes since the last update, the tool sends the entire database. There are no selection or filtering capabilities and the tools do not integrate on a net-change or change-only basis. Database snapshots are another method of transferring data between systems. Basically, a "picture" of a source database is sent at a given point in time. Snapshots do not move the entire database but rather simply capture portions of database files (i.e. specified columns). The growing number of users requiring access to data and the increasingly diverse ways in which companies use data for competitive advantage has created a need for more sophisticated real-time capture, transform and flow integration tools.

## Capture, Transform and Flow (CTF)

### Change Data Capture

Today, more and more businesses realize that they cannot achieve point-in-time consistency without continuous, real-time change data capture. There are several techniques used by data integration software to move data. Essentially, integration tools either push or pull data on an event driven or polling basis.

Push integration is initiated at the publisher (source) system for each subscribed target. This means that as changes occur, they are captured and sent, or "pushed" across to each target. Pull integration is initiated at the target by each subscribed target. In other words, the target system extracts the captured changes and "pulls" them down to the local database. Push integration is more efficient as it can better manage system resources. As the number of targets increases, pull integration becomes resource draining on the publisher system, especially if that machine is a production machine that may already be overworked.

Event driven integration is a technique that involves events at the source initiating capture and transmission of changes. Polling involves a monitoring process that polls the status to initiate capture and application of database changes. Event driven integration conserves system resources as integration only occurs after preset events whereas polling requires continuous resource utilization by a monitoring utility.

But in order to compete with an information-driven Internet era, organizations must employ solutions that offer the option of updating databases as incremental changes occur, reflecting those changes to subscribed systems. With advanced CTF solutions, every time an add, change or delete occurs in the production environment, it is automatically integrated or “pushed” in real-time to the subscriber system. By significantly reducing batch window requirements and instead making incremental updates, users regain computing time once lost.

Beyond real-time integration, change data capture can also be done periodically. Data can be captured and then stored until a predetermined integration time. For example, an organization may schedule its refreshes of full tables or changes to tables to be integrated hourly or nightly. Only data that has changed since the previous integration needs to be transformed and transported to the subscriber. Subscribers can therefore be kept current and consistent with the source databases.

**Transformation**

The way companies think about data, and the way it is represented in databases, has changed significantly over time. Obscure naming conventions, dissimilar coding for the same item (e.g. number representation as well as character based codes), and separate architectures are all commonplace. Software that can transform data across multiple computing environments and databases can remedy these problems while distributing or consolidating the organization’s information resources.

Companies are beginning to realize the benefits of sharing data between enterprise resource planning (ERP) systems and relational data stores housed in databases including Oracle, Sybase, DB2, and Microsoft SQL Server for e-Business, business intelligence or other distributed data applications. The problem is that ERP systems use proprietary data structures that need to be cleansed and reformatted to fit conventional database architectures. Rows and columns may have to be split or merged depending on the database format. For example, an ERP system may require that “Zip Code” and “State” be part of the same column while your company’s data structure in Oracle may have the two columns separated. Similarly, your company may have “Product Type” and “Model Number” in an inventory database as one column, whereas the ERP system requires them to be split. Data transformation and integration software can accommodate these requirements in order to make your data more useful and meaningful to users.

<b>Publisher</b>	<b>Transformation process</b>	<b>Subscriber</b>
Smith, M. \$10.00 U.S. 20” 1B	Two-field Consolidation and Rearrangement Euro Conversion Unit Conversion Value Substitution	Mary Smith 9.333 Euros 50.8 cm In Stock

**Figure 2:** Sample data transformations. e-Business and data warehousing applications often require operational data to be reformatted, enhanced and standardized in order to optimize performance and make content more meaningful to end users or e-Business customers.

Other applications of data transformation software include changing data representation (US dollars converted to British Sterling, metric to standard, character columns to numeric, abbreviations to full text, number codes to text), visualization (aggregate, consolidate, summarize data values) and preparation

for loading multidimensional databases. Transformational data integration software can conduct individual tasks such as translating values, deriving new calculated fields, joining tables at source, converting date fields, and reformatting field sizes, table names and data types. All of these functions allow for code conversion, removal of ambiguity and confusion associated with data, standardization, measurement conversions, and consolidating dissimilar data structures for data consistency.

## Flow

This refers to replenishing the feed of transformed data bi-directionally and in real-time between multiple operational systems and one or more subscriber systems. Whether the data is being moved onto a web server, data warehouse, or several data marts, the flow process is a smooth, continuous stream of bits of information as opposed to the batch loading of data conducted by ETL tools.

## Bi-directional Integration

Integration back and forth between two or more systems is referred to as bi-directional integration. The most common use for bi-directional integration is in data synchronization applications. Centrally administered inventory systems, for example, can have changes in product levels reflected in real-time from all branch sites. Bi-directional data movement facilitates an "update anywhere" architecture for fully distributed business solutions.

## Synchronous vs. Asynchronous Integration

Mature integration tools are based on either synchronous or asynchronous architectures. The most advanced capture, transform and flow (CTF) solutions use an asynchronous architecture; meaning that synchronization can occur while other processes continue on the publisher (source) system. This is in contrast with the two-phase commit logic inherent in synchronous distributed database management systems (DBMS).

Two-phase commit architecture guarantees that all database copies are synchronized, regardless of location, but any update failure can cause a transaction to be rolled back, necessitating a complete database refresh. This approach can be extremely time-consuming, especially with large databases. Moreover, as the number of nodes within a distributed DBMS increases, two-phase commit logic becomes unworkable as all subsequent updates are frozen until the commit process is completed for the current transaction. Essentially, system usage comes to a halt each time an update occurs.

Asynchronous integration provides reliable delivery of data while preventing any possible transaction deadlock between multiple database engines. Asynchronous architecture also allows data recovery in case of communication failure and offers integration on a predetermined schedule to avoid network or resource interruption.

## Data Staging vs. Peer-to-Peer Integration

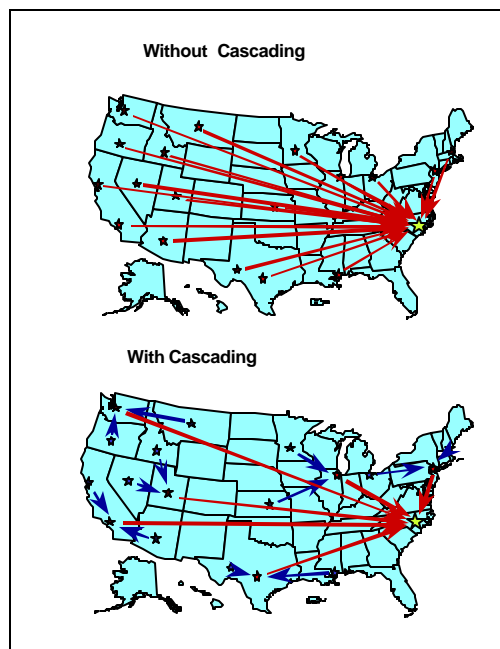
Many integration tools require intermediate data staging. This involves replicating data from an operational system to a temporary holding area on either the same system or, more commonly, on a separate system. Usually implemented in data warehousing applications, data stages are where transformation and enhancement routines are performed on the data. While there are good reasons for implementing a data stage, including a single point of control, systems administrators should keep in mind that staging adds another layer of complexity to the integration infrastructure and may require additional hardware expenditures for another server and gateway technologies.

Even if gateway tools do not stage the data, these solutions may be less efficient at moving data between publisher and subscriber systems. Gateway solutions require greater system resources and

network bandwidth due to the additional network hop required (one hop to “pull” the data from the publisher and a second hop to “push” the data to the subscriber). Additionally, the gateway itself represents another point failure in the integration network. In many cases, direct peer-to-peer integration provides a high performance alternative that does not require data staging or the acquisition of additional hardware or gateway technologies.

### Cascading Integration

Tools that support cascading integration enable the most efficient movement of data throughout a large organization. Cascading integration involves a source system transmitting data to a target system which, in turn, serves as a source for the next system in the integration chain. Let's say a company has 12 branch offices and a head office. If it takes an average of 15 minutes for each site to send its nightly data update to head office, the total integration time will be three hours. In a cascading integration environment, this time can be cut significantly. If three of the remote sites served as cascade points for the three offices closest to them, the time required to complete the integration process—and the accompanying communication costs incurred—could be reduced dramatically.



**Figure 3:** Cascading integration streamlines the integration process by enabling organizations to select regional cascade points.

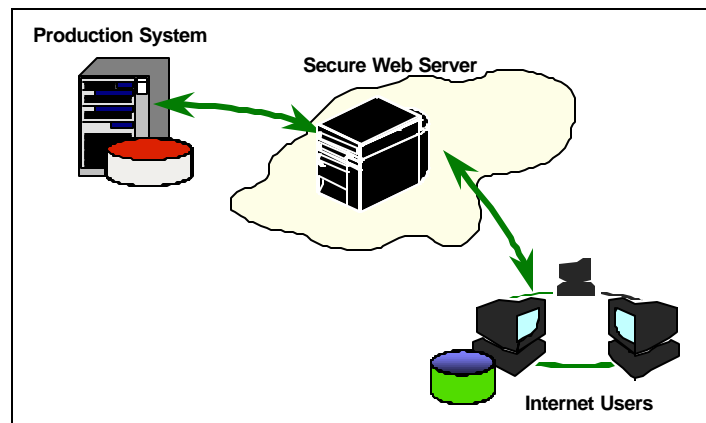
### Business Applications Of Data integration

There are many applications that capture, transform and flow data integration software enables. Real-time business intelligence/data warehousing, e-Business, data distribution, and workload balancing are all solutions that may be enabled by data integration technologies. This section provides a brief overview of how data integration and CTF technology can help companies implement a host of business applications that can help contribute to competitive insight and advantage in the new economy.

## e-Business

Today's data-rich e-Business environments require a real-time, bi-directional data infrastructure that seamlessly integrates dynamic data between operational systems and web servers—even across diverse hardware and software. Any successful e-Business venture will depend on reliable, secure, and scalable server technology coupled with data integration software that enables e-Business applications to integrate with operational systems at the data level.

e-Business involves both front-end or back-end processes. Every business-to-consumer e-Business transaction generates multiple business-to-business transactions on the back-end—credit checks, automated billing, purchase orders, stock updates and shipping. The challenge is how to integrate operational systems and enterprise data with web applications, and enable customers, partners and suppliers to transact directly with your corporate systems—inventory, accounting and purchasing. Many companies are discovering that business data integration and e-Business go hand in hand. Business data integration initiatives are being driven by e-Business needs.



**Figure 4:** Illustration depicting data moving from operational data stores to web server where e-Business applications would process information and orders.

Integration software is capable of feeding secure web servers with operational data for electronic business applications. For example, a company selling computers on the Internet will require a good deal of data movement to have the process run as efficiently as possible. For each transaction, a credit card database must be queried, inventory levels reflected, shipping and receiving database updated, and perhaps a customer database loaded. Integration software is capable of meeting all the data movement requirements of these types of electronic business applications, even across dissimilar hardware and software. Transformational data integration software is capable of reformatting database schema and filtering, cleansing and enhancing the data while moving it to and from e-Business applications on other platforms.

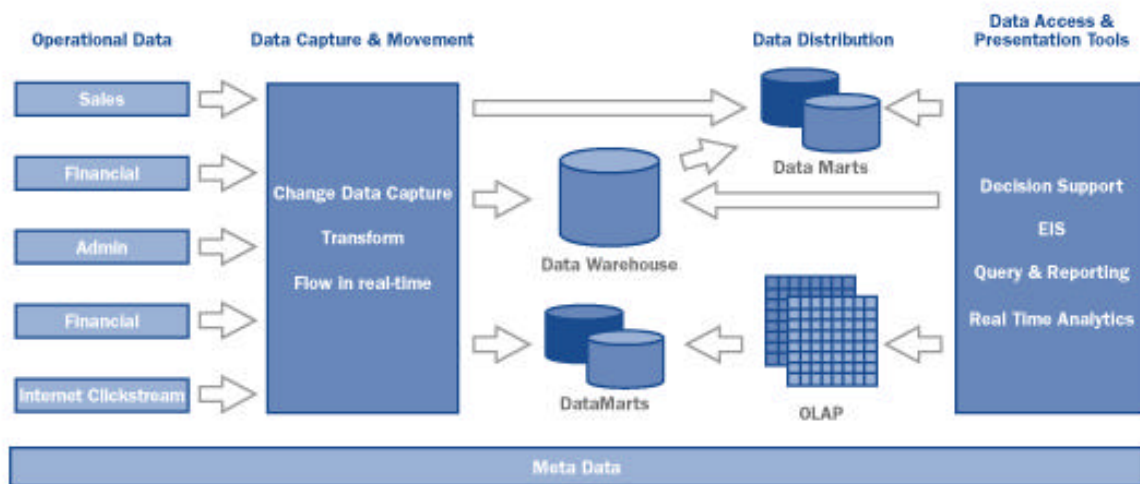
Web site activity from e-Business generates a lot of extremely valuable customer data. There is a clear need to transport e-Business and Internet clickstream data into data warehouses and data marts for business intelligence. A flexible, scalable data infrastructure based on open standards should be able to simultaneously handle the heterogeneous real-time integration needs of e-Business, business intelligence and other distributed data applications. Data integration for e-Business is about linking the elements of the business together into a cohesive whole—different computing platforms, operating systems, database technologies and applications—and giving businesses the ability to manage all the data in the organization no matter where it resides. With a scalable, flexible and resilient data integration

solution in place, companies can begin to realize the benefits of electronic business sooner, while gaining the long-term assurance that they can quickly adapt as e-Business initiatives and demands change and grow.

### Real-time Business Intelligence

Currently, the dominant method of replenishing data warehouses and data marts is to use extraction, transformation and load (ETL) tools that “pull” data from source systems periodically—at the end of a day, week, or month—and provide a “snapshot” of your business data at a given moment in time. That batch data is then loaded into a data warehouse table. During each cycle, the warehouse table is completely refreshed and the process is repeated no matter whether the data has changed or not.

In the Internet era, more people are beginning to realize the limitations that snapshot copy replenishment presents and demand better alternatives. Snapshots do not involve entire database movement but simple captures of parts of database tables; for example, specified columns. As well, not each individual change is made to a record between copy processes.



**Figure 3:** Typical data warehouse implementation utilizing capture, transform and flow (CTF) technology for real-time replenishment.

The Internet era is about having absolutely current and up-to-date business intelligence information. Data is a perishable commodity: the older it is, the less relevant. Businesses need tools that can provide real-time business intelligence and an absolutely current and comprehensive picture of their organization and their customers—not last week or last month, but right now.

Transformational data integration software enables users to capture, transform and flow data in real-time between many different databases and computing platforms for loading and replenishing business intelligence applications. The data can be cleansed, summarized and enhanced while consolidating it into an operational data store or directly into data marts or data warehouse tables.

## Enterprise Application Integration (EAI)

Over time, most organizations have invested in a combination of best-of-breed application packages and in-house development. Adding or subtracting software from this solution set can often require considerable effort to integrate changes with the existing infrastructure. Integration of operational systems and enterprise data with new business applications like e-Business and business intelligence is a problem that costs companies over \$100 billion per year. Often, application interfaces and collection routines can grow more complex than the applications themselves. Companies are realizing the value of re-engineering their knowledge management systems through enterprise application integration initiatives.

Enterprise application integration requires linking applications at two levels: the data level and application interface or business model level. Although message brokering solutions provide integration at the business model level, a data brokering solution that provides real-time data integration, transformation and transport is often required for a complete solution. Data integration software is complementary to message brokering EAI solutions.

Data integration tools can provide data level integration for both API and non-API applications. Many custom applications were not designed to facilitate EAI, but rather to function independently as stovepipe solutions that do not allow open access to the various levels and services of the application. Since most message brokering software use APIs to facilitate application integration and interfacing, they cannot easily support integration of many in-house applications. Data integration software can often provide the flexibility to enable enterprise application integration for many in-house applications that are not supported by message brokering software.

Companies need to ensure that data is continuously available to all business applications in real-time—regardless of platform, databases, application oriented data structures or geography. Companies should select an open solution that enables applications on a wide range of computing platforms to share information at the data level. Full-featured and high performance integration software provides the real-time, bi-directional integration and transformation capabilities required to move data between different schemas on many different databases and hardware platforms.

Transformational data integration tools usually require no programming changes to existing applications or databases. This ensures a risk-free implementation. Moreover, there is no need for costly gateway technologies or data stages that may require additional hardware. Choosing a direct peer-to-peer integration infrastructure with built-in transformational capabilities can help companies achieve rapid implementation and ROI, making Enterprise Application Integration a cost-effective reality.

## Customer Relationship Management (CRM)

According to META Group, results from our recent industry studies indicate that customer data leads integration challenges by nearly a three-to-one margin over other types of data, and that over 75% of businesses are attempting to integrate their CRM and e-business solutions. Enabling technologies will include a new breed of point-to-point information routers that securely synchronize data among applications.

Transformational data integration solutions with real-time capabilities can enable this kind of seamless point-to-point integration of customer data stored in BackOffice or legacy systems. Companies are just beginning to realize that real-time data integration is the cornerstone of all CRM initiatives and a key to

21st century competitiveness. Real-time integration can deliver enormous value to CRM vendor solutions and customers by ensuring that all computing systems and data stores across the business are fully integrated into the CRM backbone. End-to-end integration with out-of-the-box flexibility is absolutely imperative in order for companies to gain a 360-degree view of customer relationships. This comprehensive view of all customer interactions with the organization is necessary for companies to become truly customer-centric.

Leading CRM suites are starting to utilize data integration technology as part of an end-to-end package. This empowers users with real-time information such as inventory levels, delivery schedules and customer status that would otherwise be difficult to access. CRM solutions that include cross-platform data integration components can be used to ensure that data from BackOffice and legacy systems is delivered directly to users within the CRM application—adding considerable value to the CRM solution and to its end users.

### Data Distribution

Data distribution allows companies to move data from one publisher (source) system to multiple subscriber (target) systems—regardless of computer hardware or database. Organizations can integrate entire databases from the publisher machine to subscribers or integrate location specific data, such as daily price updates, to branch offices or retail sites. For example, parts of centrally administered database files can be integrated to remote locations for local query and analysis purposes. Transformational data integration software satisfies the data selection, filtering, enhancement and movement requirements of data distribution projects.

### Workload Distribution

Companies often suffer from strained system resources as transaction levels rise and query and analysis activity increases. Integration software is well suited for improving system performance and reliability by distributing batch jobs and query and analysis activity to a replica server. This allows business users to manipulate data on a secondary machine without affecting performance on operational systems.

## Other Considerations

### Ease of Implementation & Flexibility

A serious consideration when evaluating integration tools is the work required to set up and initiate the solution. Organizations should ensure that no programming changes to existing applications are required and that no coding must be done to initiate the process. Integration software is meant to avoid resource intensive and costly custom programming. Effective data integration tools must also provide the flexibility to employ multiple integration modes including:

- **CONTINUOUS MIRRORING:** Also known as real-time integration, continuous mirroring enables organizations to update databases as changes (adds, updates, deletes) occur and reflect those changes to target systems.

- **CHANGE DATA CAPTURE:** Captures database changes as they occur and stores them until a predetermined integration time such as nightly or hourly. Only records that have changed since the last update are sent to target databases.
- **FULL COPY REFRESH:** Integrates an entire database copy to target systems. Full copy refreshes are done to resynchronize databases after an outage or upon initial synchronization.

It is paramount to consider the administration of your integration network. The solution must be capable of administering current and future integration needs. A platform independent, web- or wireless-enabled graphical front-end to the entire integration network is crucial. Platform independence is so important because it is difficult to know what your company's platform of choice might be even one year down the road. A graphical front end makes it easier to see and perform actions on the entire integration network from a holistic perspective. Web-enabled administration ensures that corporations can access their integration network from any workstation anywhere in the world. Considering these factors, an efficient administrator component that fits your company's current and future IT strategy will substantially reduce the total cost of ownership while increasing the productivity of the integration network.

Companies should choose a tool that provides maximum flexibility and ease-of-use in order to ensure rapid ROI and a long solution life cycle. Why select a niche tool to enable one specific application, or custom-code a solution in-house, when there are flexible, open integration solutions that will enable a range of powerful business applications for years to come?

### Metadata Management Capabilities

Metadata is information about data. It allows business users as well as technical administrators to track the lineage of the data they are using. Metadata provides information about where the data came from, when it was delivered, what happened to it during transport, and other descriptions can all be tracked. Essentially, examining metadata enhances the end user's understanding of the data they are using. It can also facilitate valuable "what if" analysis on the impact of changing data schemas and other elements. For administrators, metadata helps them ensure data accuracy, integrity and consistency.

During data transport, advanced capture, transform and flow (CTF) replenishment solutions store metadata in tables located in the publisher and/or subscriber database. This is an attractive feature to companies wanting to share metadata among heterogeneous applications and databases. Most tools and databases manage metadata differently. That is, they store the metadata in distinct formats and use proprietary meta-data to perform certain tasks. An open CTF solution allows organizations to distribute metadata in different formats using published industry standards. Using CTF technology based on open standards such as the Metadata Coalition Standard for Metadata Integration, Open Information Model (OIM), or XML Interchange Format (XIF), metadata can be easily integrated to another repository in the required format. This addresses the challenge of standardizing metadata. Without this functionality, companies often must resort to custom development of a tool capable of entering metadata in a variety of formats—a development task that often proves time-consuming, difficult to accomplish, and may negatively impact time to market.

### Buy versus Build

Despite the buy-versus-make recommendations of most analysts, some companies still choose custom coding to handle the integration and transformation of data. In this situation, businesses can write customized programs to integrate data. However, given the productivity gains of using a mapping-based integration tool over scripting, it is difficult to justify allocating the time and resources required for

developing custom applications. Mapping-based tools involve built-in administration utilities that provide quick and easy definition of the integration process. Changes in the integration schema are simply re-mapped using the same front-end that defined the original process.

Some companies who choose to bring the task in-house and custom code extraction routines soon discover that it is a difficult problem with its own challenges. Why allocate a programmer to do punishing extract programming when your organization can use an easy-to-implement mapping-based tool? IT and programming resources are at a premium and most organizations would prefer to have their programmers working on building or integrating new e-commerce applications and web-based services than writing extract routines. Packaged solutions can provide businesses with many advantages, including flexibility, scalability, and upgrade support for the latest database versions upon their release.

## Summary

Today's computing environments are both dynamic and volatile. e-Business and the proliferation of mixed system computing environments can place enormous strain on IT resources. Many applications require the movement of data between heterogeneous systems and databases. Transformational data integration software with advanced real-time capture, transform and flow capabilities can simplify development of business solutions that require data sharing, distribution and consolidation—like e-Business applications, data warehouses and data distribution projects. Deciding which integration architecture is right for your company should be a three-step process. First, organizations need to clearly understand the type and volume of data that they need to support. Secondly, they should explore all the alternatives, and finally, test the technology thoroughly with the vendor. This due diligence should ensure that the integration tool you acquire today will continue to deliver value tomorrow and well into the future.

Copyright © 2000 DataMirror Corporation. All rights reserved. DataMirror is a registered trademark of DataMirror Corporation. All other brand or product names are trademarks or registered trademarks of their respective companies.

## About DataMirror Corporation

DataMirror (Nasdaq: DMCX; TSE: DMC) delivers solutions that let customers integrate their data across their enterprises. DataMirror's comprehensive family of products unlocks *The experience of now™* by providing advanced real-time capture, transform and flow (CTF) technology that gives customers the instant access, integration and availability they demand today across all computers in their business.

Over 1,400 companies use DataMirror to integrate their data. Real-time data drives all business. DataMirror is headquartered in Toronto, Canada, and has offices worldwide. DataMirror has been ranked in the Deloitte and Touche Fast 500 as one of the fastest growing technology companies in North America.



DATAMIRROR RESOURCE CENTER  
SOFTWARE | SOLUTIONS | BEST PRACTICES  
[www.datamirror.com/resourcecenter](http://www.datamirror.com/resourcecenter)

FOR MORE INFORMATION CALL 1 800 362-5955