



---

## **THE NEW DATA INTEGRATION LANDSCAPE**

Moving beyond ad-hoc ETL to an enterprise data integration strategy

---





---

## Contents

---

<b>Introduction: The expanding scope of ETL .....</b>	<b>1</b>
<b>Moving from ad-hoc ETL to enterprise data integration.....</b>	<b>2</b>
<b>Data integration defined.....</b>	<b>2</b>
<b>How best to move forward? .....</b>	<b>2</b>
<b>High-level guide to data integration programs .....</b>	<b>3</b>
Data cleansing and enrichment .....	4
Data warehousing/marts (ETL).....	4
Cross system data consistency (data synchronization) .....	5
Data migration/consolidation: legacy system, ERP and RDBMS .....	5
Master data management (MDM) .....	6
<b>High-level guide to data integration capabilities .....</b>	<b>7</b>
<b>High-level guide to data integration services .....</b>	<b>8</b>
<b>Enterprise connectivity.....</b>	<b>9</b>
<b>Bringing it all together .....</b>	<b>10</b>
<b>Your data integration strategy .....</b>	<b>11</b>

---

Content for *The New Data Integration Landscape: Moving beyond ad-hoc ETL to an enterprise data integration strategy* was provided by Mark Torr, Director of the Global Technology Practice at SAS.

---

## Introduction: The expanding scope of ETL

For years, the key to success for any business intelligence solution has been the process known as extract, transform and load (ETL). Selecting the right tool to bring data from disparate sources and transform it before loading into a target destination was the critical factor in building a data warehouse or data mart to support an organization's business intelligence projects. In fact, it was so important that the process became synonymous with the tool, and the technology became known as ETL technology, which spawned many ETL tools.

Numerous organizations have struggled through the process of selecting tool after tool to gain access to new data sources as limitations in the previously chosen tools became apparent. This organic growth of tools as departments operated unchecked with their tool of choice, coupled with mergers and acquisitions, caused many organizations to end up with several non-integrated ETL tools.

Likewise, some organizations have failed to see the benefits of tools over custom coding, which has resulted in small armies of programmers building and maintaining code. The problem with using several tools or custom code is that it significantly increases the total cost of ownership in terms of maintenance, training and time lost in regaining familiarity with a rarely used tool. Using several tools also can lead to very fragmented metadata, which turns compliance and other issues into chores rather than automatic processes delivered through self-documenting metadata.

In addition to the proliferation of ETL tools, building and maintaining a data warehouse or a data mart is no longer the only activity taking place in organizations when it comes to data. And business intelligence, while still a powerful driver, no longer stands alone. Organizations are finding it increasingly necessary to take on additional non-warehouse projects, such as system migration, consolidation and synchronization as a result of mergers, acquisitions, corporate break-ups and an overall need to modernize older systems. While ETL processes support some of these projects, many others demand new technologies. For example, master data management and real-time synchronization/data quality, which are needed to maintain integrity of operational systems, are fast emerging as critical themes in most organizations and they require new technologies.

This new, expanded scope has led to the emergence of data integration, which should be a strategic topic in all organizations because it affects everything the business does.

Numerous organizations have struggled through the process of selecting tool after tool to gain access to new data sources as limitations in the previously chosen tools became apparent.

The problem with using several tools or custom code is that it significantly increases the total cost of ownership in terms of maintenance, training and time lost in regaining familiarity with a rarely used tool.

---

## Moving from ad-hoc ETL to enterprise data integration

---

It is time to move forward from an ad-hoc approach and look at data integration as something that can contribute significantly to competitive advantage. It is time to perhaps think about standardizing as much of your data integration, including ETL, with one “system-neutral” vendor in order to leverage synergies such as shared business rules and metadata across the spectrum of data integration. You will see reduced costs for training and maintenance as well as many other benefits on the operational and BI fronts from having one, consistent, integrated set of technologies. It is time to ensure that the experiences of the ETL era are understood and establish a new way to success.

---

## Data integration defined

---

Data integration can be seen as the convergence of multiple technologies and the emergence of some new ones. Broadly speaking, data integration brings together technologies that typically are needed for the operational side of the business with technologies that are needed for the BI/decision support side of the business. Data integration deals with incorporating all types of organizational data into a unified whole.

Data integration cannot be seen as just “a means to an end” because in many cases data integration is not directly driving things such as BI and analytics; it is supporting operational processes or keeping operational systems in sync. It is this shift in focus that perhaps best characterizes data integration and it is also the reason that data integration technologies from RDBMS vendors are somewhat limited; they are still too focused on the BI world.

A cohesive, data integration strategy requires a major focus on the non-BI aspects affecting organizations today, at a time when many organizations have not yet resolved the issue of ETL/data integration for the purpose of supporting data warehouses and data marts.

---

## How best to move forward?

---

As with all things, a data integration strategy brings organizations some “buy” vs. “build” choices. Because most vendors in the market have evolved their portfolios through mergers and acquisitions, there is a third choice: buy and integrate the tools even if they are from a single vendor. This is the same tool integration that would be required with a “piecemeal” approach if you bought from several vendors to meet all your needs.

Organizations looking to establish a data integration strategy should learn how all of the capabilities were added to a portfolio (integrated through in-house development or purchased through acquisition), and if things such as metadata, business rules, etc., can be shared.

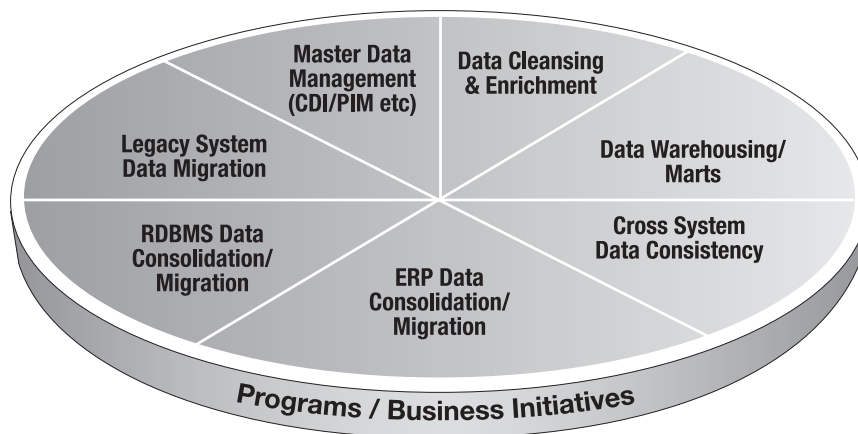
If they cannot be shared, when and what will the migration steps be? Organizations should be careful not to be “taken in” by descriptions of manual steps needed to get the bigger picture. Manual steps introduce overhead and risk, and hidden costs and risks suddenly can become very apparent.

## High-level guide to data integration programs

If you are new to data integration, you may be wondering just what you should expect from a data integration solution. Perhaps we should start by establishing what data integration is not.

Data integration is not about enterprise application integration middleware, although it does make use of that in some cases. It is not about message queues and application servers, even though these are important parts of the infrastructure that will support certain aspects of data integration. There seems to be a desire to force these topics and what is commonly seen as “middleware” into the broader data integration domain because it suits certain vendors. Do not be confused — tying your infrastructure to your data integration vendor creates a lock-in that could be difficult to get out of.

A comprehensive universal data integration solution should enable the successful completion of many different programs or business initiatives. It also should enable reuse of common services across each of these programs and be able to operate in various modes of latency. In this paper we will begin by providing an overview of the data integration programs and business initiatives, including why they are important.



*Figure 1: A comprehensive universal data integration solution should provide the capability to execute a variety of data integration programs and/or business initiatives.*

## **Data cleansing and enrichment**

Any comprehensive universal data integration solution should provide the capability to cleanse and enrich (augment data with third-party external data to improve its completeness) data. The importance of data cleansing and enrichment is on the rise and organizations need to be thinking about it. One of the main drivers (though not the only one) is the increasing utilization of data in automated processes where bad data quality can lead to immediate and very high unwarranted costs. An example of this is sending customer mailings to the wrong address or multiple mailings to the same customer who is in the system several times with different spellings. In the past this was not as much a problem because automation was limited. Now, however, automation is on the rise and human intervention on the decline. Therefore, data needs to be as accurate as possible.

In addition to this automated use of data, there is an increasing desire to ensure that information entering operational systems is more accurate. This reduces the downstream costs (monetary and time) of having to clean the data and ensures a better customer/supplier experience in call centers, which can be the difference between a happy customer and one who is frustrated. Data cleansing does not end there though because compliance means you have to get your data into order to ensure the maximum accuracy of reporting. All of this means that you should be looking for a data integration solution that includes integrated data cleansing and enrichment to support data quality processes such as profiling, householding, deduplication, data quality, business rule creation, and cleansing of data (where required).

These rules should be callable through custom exits, messages placed onto message queues, or Web services to trigger the process and deliver what can be referred to as real-time data quality integration. A classic example is the checking of names and addresses at the point of entry into an ERP system through the use of a custom exit, in order to build in data quality from the start.

## **Data warehousing/marts (ETL)**

Any data integration solution should provide the capability to build and maintain data warehouses/data marts via the ETL process. The solution needs to leverage the data connectivity capabilities that were previously mentioned and have fully integrated metadata. Such a solution also should include support. By support we mean technical support and help from professional services as a part of the solution, as well as extensions through custom coding so that organizations have the flexibility to do more than the tool delivers but will not lose the support of the vendor when they use custom code, thus reducing risk. In addition, the solution must allow for the reuse of data quality business rules that are provided by the data quality component of the data integration offering. Data quality must take “center stage” in any integration strategy.

Data quality must take “center stage”  
in any integration strategy.

## **Cross system data consistency (data synchronization)**

Any data integration solution needs the capability to reflect changes made between systems across the enterprise. There are two types of cross system data consistencies. The first type is the movement of “changes” made in one or more systems to other systems in batch/near real-time. The second type is the movement of “changes” made in one or more systems to other systems in real-time.

The first type is just another application of the ETL process using change-data capture and a scheduled process to move data around. This process can be scheduled nightly, every 30 minutes, every five minutes or even more frequently depending on the needs of the organization and the amount of data to be moved. However, it typically involves the movement of multiple “transactions” or “records” concurrently.

The second type involves the movement of “individual transactions” or “records” to synchronize status across multiple systems as the transactions occur and in real-time. Technologies such as message queues and brokers are often used in such circumstances. Here, a real-time server needs to be invoked using custom exits, messages placed in message queues, change brokers or Web services to trigger the process.

Again, we should underscore the importance of data cleansing and enrichment in cross system data consistency. Although bad data in one system is not good, the proliferation of bad data through automatically synchronizing systems can have a devastating effect. Organizations should ensure that any cross system data consistency efforts also include the application of data quality business rules to maintain the quality of data throughout all systems.

## **Data migration/consolidation: legacy system, ERP and RDBMS**

Any data integration solution should provide the capability to migrate data from multiple existing systems to one or more new or existing systems. One could argue that in its most primitive form this is just the application of the ETL process, along with data quality, to a target other than a data warehouse or a data mart. But why migrate bad data forward? Why not clean and enrich it, and deliver business value from what is normally perceived as strictly an IT project?

Organizations should be looking to build up data quality business rules over time (and from data migration project to project) that can be applied whenever a migration takes place in order to get reusable, immediate and low-cost business benefits. These same rules should be usable when supporting data warehouse and data mart creation/maintenance.

While a one-off migration often might take place, it will be very hard to achieve on the operational side where the source system might live. This is because organizations often have operational business applications that need to be migrated. That movement forward will first involve migrating the data to a new system and verifying its correctness (again, this is where metadata becomes vitally important) before establishing an ongoing data synchronization process between the old and the new, and placing the business application on top of the new system for acceptance testing. Once you are satisfied that the data in the new system is up-to-date and that the business application is operating as expected on the new system, the old system can be turned off and data synchronization ended.

### **Master data management (MDM)**

Any data integration solution should provide the capability to handle the new and emerging topic of master data management. Master data management is the practice of creating a single “perceived” truth by mapping multiple disparate definitions of items, such as names of customers and products, that are held in various systems.

Thus, when a user asks for “customers,” all customer names are returned to any application in a common format using a standard, company-accepted definition without the users having to understand the underlying structure in the various silos throughout the organization.

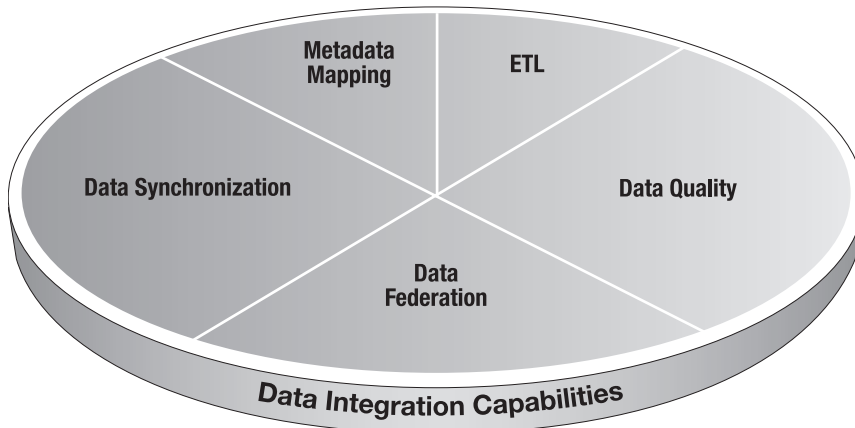
Tied closely to master data management are emerging topics such as Customer Data Integration (CDI) and Product Data Integration (PDI) that build on the basic technology and deliver a number of common mappings and definitions to get organizations up-and-running quickly. Where MDM is a topic of concern, organizations should look for the development of these more advanced solution areas that incorporate a true metadata management framework within traditional reference data management and that speed up the time to deployment, thereby reducing overall costs.

Ultimately, CDI and PDI are examples of real implementations that solve specific problems in the broader MDM space. Many organizations will need to solve one of these specific sets of problems first. However, more forward-thinking organizations will have a broad MDM strategy that leverages many of the same technologies and capabilities to achieve common results within the enterprise.

If a vendor says its software can do MDM but they do not deal with topics such as CDI or PDI, then you could be buying into a very limited solution that will require a lot of ongoing manual and expensive custom development.

## High-level guide to data integration capabilities

In order to provide the ability to handle all of the aforementioned data integration programs, a universal data integration solution needs to deliver a certain set of capabilities that are combined to meet the requirements of a program. A basic set of capabilities as shown below are required to fulfill all the needs:

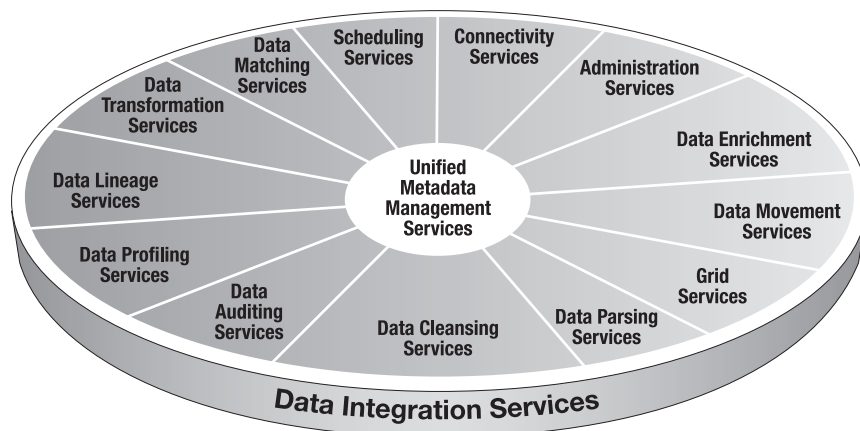


- ETL: Extract, transform and load.** A comprehensive data integration offering needs ETL capabilities. Many people automatically associate ETL with building or maintaining data warehouses or marts but that is not the only use for ETL technology. When seen as a process, ETL is used in almost all of the data integration programs to access data, transform it and load it. The ETL technology brings together underlying services, normally through a design interface, to build re-usable services and support various data integration programs. Note that by pushing the processing to the database, ETL technologies can be used just as easily for Extract, load and transform (ELT).
- Data quality.** A comprehensive data integration offering needs data quality capabilities that can be deployed from the operational world to the data warehousing world. A good offering would provide a graphical environment for data stewards that allows them to bring together the underlying services to profile, parse, enrich, cleanse and match data to create the business rules to be applied either in real-time/batch as a part of an integrated warehouse process, a data migration/synchronization, a master data management solution, or just to cleanse RDBMS/ERP data in place or at the operational edge as data is entered.
- Data federation.** A comprehensive data integration offering needs to provide a data federation capability that allows data to remain in place and be integrated and surfaced as needed. Due to its dynamic nature, this technique can lend itself to solving potential problems where there is no need to access large amounts of data or data from many underlying systems. Data federation and Enterprise Information Integration (EII), along with data synchronization, are often the underlying technologies employed with certain styles of master data management, and are also often used with BI solutions where more operational or real-time views of data are required.

- **Data synchronization.** A comprehensive data integration offering needs to provide data synchronization capabilities to enable data to be moved from one place to another by enabling the use of message queues, triggers, Change Data Capture, and more.
- **Metadata mapping.** A comprehensive data integration offering needs to provide capabilities that allow for impact analysis and change analysis. It also needs to provide capabilities to the import metadata from various systems and exchange metadata with other systems.

## High-level guide to data integration services

Besides the ability to create your own data integration services that can be called in real-time, near real-time or in batch, a good data integration solution needs to provide a rich set of services out-of-the-box that come together to deliver the capabilities previously covered in this paper. Which services are used depends on what capability of the data integration platform you require. A set of them is shown below.

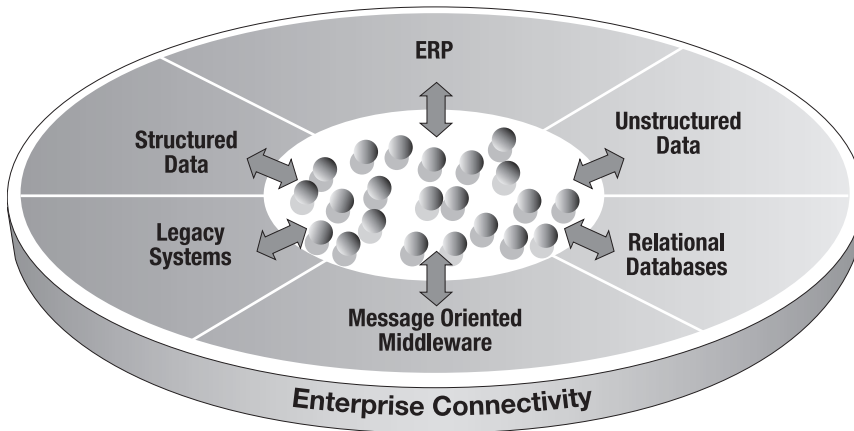


At this level, we can clearly see metadata services are very important in any data integration strategy and that is because it should be pervasive as a key enabling and documentation mechanism. Metadata should be pervasive through all data integration programs being undertaken to promote governance, reuse and productivity.

Data integration, at its core, is about relating multiple data sources and bringing them together to make your data more valuable. Metadata provides the definition across data sources that make this possible. In addition, metadata allows you to trace what moved when, how it was changed, what business rules have been applied, and what impacts those changes might have. These are critical issues facing all organizations. Failure to place enough emphasis on metadata will result in problems later on, often at a great cost to the organization.

Failure to place enough emphasis on metadata will result in problems later on, often at a great cost to the organization.

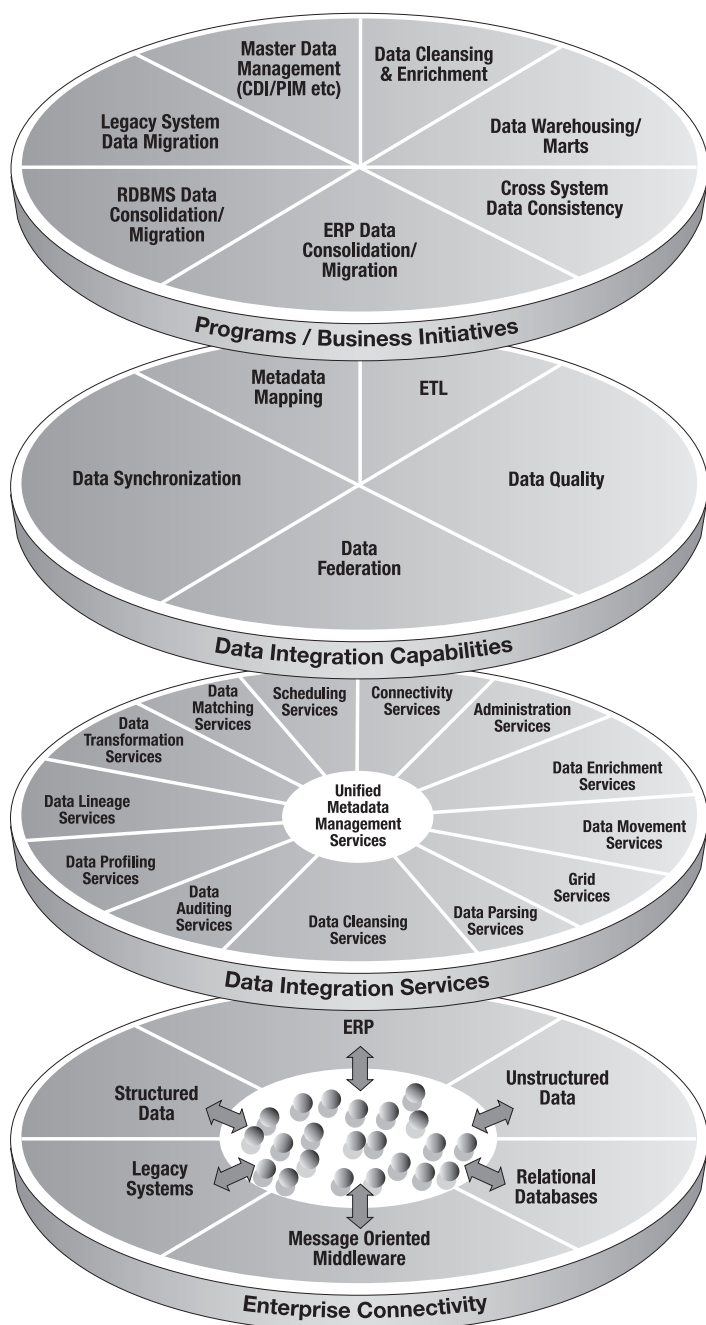
## Enterprise connectivity



Any data integration solution should provide connectivity through both native access using standard utilities and open standard access (such as ODBC) to all major structured data sources, including relational databases, flat files, ERP systems, and mark-up languages such as XML for reading and writing. The data connectivity capabilities should facilitate the access of information on many different systems such as z/OS, UNIX and Windows, preferably without having to make use of intermediate files and extracts. Support for connecting to and reading and writing data from message queues and the ability to receive and send data to and from Web services should also be provided by the solution to provide complete connectivity. In the longer term, the solution needs the ability to evolve to support unstructured data sources.

## Bringing it all together

When you combine all of these layers you see that a universal data integration solution is both wide and deep, providing the flexibility to meet your requirements now and in the future. It delivers a comprehensive set of reusable services and promotes the reuse of services through integrated metadata and governance. In selecting a data integration vendor, organizations need to look beyond the GUI and see what lies under the covers for today and for their needs tomorrow.



## Your data integration strategy

There is no doubt that organizations will set priorities on which data integration tasks take precedence. But how many will avoid past mistakes made with ETL and ensure that a short-term strategy is backed by longer-term integrated possibilities?

SAS—in conjunction with DataFlux, a wholly owned subsidiary of SAS that focuses on the data quality aspects of data integration and real-time data integration—delivers integrated solutions that can be brought together, incrementally and in a variety of ways, to suit the needs of any organization. You can begin with a solution to address master data management, or with technologies to build data warehouses and data marts, or to perform rudimentary data profiling.

The important thing is that, whichever direction you subsequently take, SAS and DataFlux can deliver all the technologies you need to establish a data integration strategy while realizing the benefits of shared business rules, shared metadata and integrated technologies. These savings combine with the associated cost reductions when employees need less training and business rules can be reused. In addition, you'll have less inherent tool and metadata integration and fewer maintenance and management problems, which are alleviated by an integrated, comprehensive approach.

If you are not doing so already, today might be a good time to start deciding where the future of your data integration strategy lies. All the topics in the preceding landscape should live and work together to give you maximum benefit and value. Previously established piecemeal standards need to be challenged – and time is not on your side.

The most successful organizations will have a clear and precise strategy in place that recognizes data integration as a fundamental cornerstone of their competitive differentiation. Those who succeed will be the leaders who can address all their needs by using one integrated offering, thereby having the flexibility to react to new challenges quickly. Those who hesitate will be left behind in a sea of complexity and cost.

Data integration should be complete, flexible, integrated and proven. SAS and DataFlux provide all these strengths and stand ready to help you address your challenges today.

The most successful organizations will have a clear and precise strategy in place that recognizes data integration as a fundamental cornerstone of their competitive differentiation.



THE  
POWER  
TO KNOW.

SAS Institute Inc. World Headquarters +1 919 677 8000

To contact your local SAS office, please visit: [www.sas.com/offices](http://www.sas.com/offices)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2007, SAS Institute Inc. All rights reserved. 102446\_403251 .0407