The New Data Integration Landscape
Moving Beyond Ad Hoc ETL to an Enterprise Data Integration Strategy
### Table of Contents

- **Introduction: The Expanding Scope of ETL** ........................................... 1
- Moving from Ad Hoc ETL to Enterprise Data Integration .......................... 1
- **Data Integration Defined** ................................................................. 2
- **How Best to Move Forward?** ......................................................... 2
- **A High-Level Guide to Data Integration Programs** ........................... 2
  - Data Cleansing and Enrichment ....................................................... 3
  - Data Warehousing and Data Marts (ETL) ........................................... 4
  - Cross-System Data Consistency (Data Synchronization) ...................... 4
  - Data Migration and Consolidation: Legacy Systems,
    - ERP and RDBMS ................................................................. 4
  - Master Data Management (MDM) ................................................... 5
- **High-Level Guide to Data Integration Capabilities** ........................... 5
- **High-Level Guide to Data Integration Services** ................................ 7
- **Enterprise Connectivity** ............................................................... 8
- **Bringing It All Together** .............................................................. 8
- **Your Data Integration Strategy** .................................................... 10
Introduction: The Expanding Scope of ETL

For years, the key to success for any business solution has been data. Extract, transform and load (ETL) has traditionally been an important part of this success. Selecting the right solution to bring data from disparate sources and then transforming it before loading it into a target destination were the critical factors in building a data warehouse or data mart to support an organization’s data project. Numerous organizations have struggled through the process, selecting one tool after another to access new data sources as limitations in the previously chosen tools became apparent. Departments often operated in isolation when choosing tools, which, coupled with mergers and acquisitions, resulted in many organizations using several ETL tools that were not integrated.

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 can also lead to 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 are no longer the only data activities taking place in organizations. And business intelligence, while still a powerful force, no longer stands alone. Organizations are finding it increasingly necessary to take on additional projects – such as system migration, consolidation and synchronization as a result of mergers, acquisitions and corporate break-ups – and an overall need to modernize older systems. While ETL processes support some of these projects, many of them demand new technologies. For example, master data management and real-time synchronization/data quality, which are needed to maintain the integrity of operational systems, are emerging as critical themes that 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 organization does.

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 think about standardizing as much of your data integration as possible, including ETL, with one “system-neutral” vendor so that shared business rules and metadata can be used 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 and integrated set of technologies.

Numerous organizations have struggled through the process, selecting one tool after another to access 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.
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 helps you incorporate 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. It is also the reason that data integration technologies from RDBMS vendors are somewhat limited; they are still too focused on the BI world.

How Best to Move Forward?

As with all things, a data integration strategy allows organizations to decide between buying and building solutions. Yet, since most vendors in the market have evolved their portfolios through mergers and acquisitions, there is a third choice: buy and integrate the tools from a single vendor. This same tool integration approach would be required if you bought from several vendors.

Organizations interested in establishing 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 such things as metadata, business rules, etc., can be shared. If they cannot be shared, when and what will the migration steps be? Organizations should weigh the number of manual steps needed to achieve data integration because manual steps introduce overhead and risk. The hidden costs and risks suddenly can become very apparent.

A 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. We will start by establishing what data integration is not.

Data integration is not about enterprise application integration middleware, although it does make use of it 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. It seems there is 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 leave.

A comprehensive universal data integration solution should enable the successful completion of many different programs or business initiatives. It should also 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 data integration programs and business initiatives, including why they are important.

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

Data Cleansing and Enrichment

Any comprehensive data integration solution should provide capabilities to cleanse and enrich data. It is an increasingly important step given the expanding use of data in automated processes and how bad data quality can lead to immediate and unwarranted costs. For example, consider the cost of sending customer mailings to the wrong address or multiple mailings to the same customer whose name is spelled several different ways. As automation increases and human intervention continues to decline, the consequences of bad data are greater and felt more acutely. 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 accurate. Accurate data reduces the downstream costs (both monetary and time) of having to clean the data; it also ensures a better customer/supplier experience in call centers, which are often the difference between a happy customer and a frustrated one. Regulatory compliance mandates have also increased the importance of high-quality data to ensure accurate reporting. Given these demands, you should look for a data integration solution that includes integrated capabilities to support data quality processes such as profiling, householding, deduplication, business rule creation and data cleansing.

These rules should be callable through custom exits, messages placed onto message queues or Web services to trigger the process and deliver real-time data quality integration. A classic example is checking names and addresses at the point of entry into an ERP system with a custom exit so that data quality is built in from the start.
Data Warehousing and Data Marts (ETL)

Any data integration solution should build and maintain data warehouses and data marts via the ETL process. The solution should use the data connectivity capabilities that were previously mentioned and have fully integrated metadata. Such a solution should include technical support and help from professional services. It must also have extensions through custom coding so that organizations have the flexibility to do more than the tool delivers but without losing the support of the vendor when custom code is used. In addition, the solution must allow the reuse of data quality business rules that are provided by the data quality component of the offering. Data quality must take center stage in any integration strategy.

Cross-System Data Consistency (Data Synchronization)

Any data integration solution must 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 or 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 depending on the needs of the organization and the amount of data to be moved: nightly, every 30 minutes, every five minutes or even more frequently. 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 is 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 and Consolidation: Legacy Systems, ERP and RDBMS

A data integration solution should 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 an ETL process along with data quality to a target other than a data warehouse or a data mart. But why migrate bad data? 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 increase the number of data quality business rules over time (and from one data migration project to another) that can be applied.
whenever a migration takes place for reusable, immediate and low-cost benefits. These same rules should be used when supporting the creation and maintenance of data warehouses and data marts.

While a one-off data migration might take place, it will be very hard to achieve on the operational side where the source system lives, since organizations often have operational business applications that need to be migrated. That movement will first involve migrating the data to a new system and verifying its accuracy (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)**

Data integration solutions should provide master data management capabilities. Master data management is the practice of creating a single 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 users having to understand the underlying structure in the various data silos throughout the organization.

Tied closely to master data management are emerging topics such as customer data integration (CDI) and product information management (PIM) that build on the basic technology and deliver a number of common mappings and definitions to help organizations get up and running quickly. Where MDM is a topic of concern, organizations should develop these more advanced solution areas that incorporate a true metadata management framework within traditional reference data management, which speed time to deployment and reduce overall costs.

Ultimately, CDI and PIM are examples of 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 uses many of the same technologies and capabilities to achieve common results within the enterprise.

**High-Level Guide to Data Integration Capabilities**

To handle all of the aforementioned data integration programs, a data integration solution should deliver a certain set of combined capabilities. A basic set of these capabilities (as shown below) are required to fulfill that need:
Figure 2: A basic set of integrated capabilities are required to implement a successful data integration initiative.

- **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 data marts, but that is not the only use for ETL technology. When seen as a process, ETL is used in almost all data integration programs to access data, transform it and load it to various targets. The ETL technology brings together underlying services, normally through a design interface, to build reusable services and to 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 that allows data stewards to bring together the underlying services to profile, parse, enrich, cleanse and match data to create the business rules. These rules can then be applied either in real time or batch as a part of an integrated warehouse process, a data migration/synchronization or a master data management solution. It should also be able to cleanse RDBMS/ERP data in place or at the operational edge as data is entered.

- **Data federation/data virtualization.** A comprehensive data integration offering needs a data federation capability that allows data to remain in place as it is integrated and surfaced. Due to its dynamic nature, this technique can solve 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 used with certain styles of master data management. They are also used with BI solutions where more operational or real-time views of data are required.

- **Data synchronization.** A comprehensive data integration offering must provide data synchronization capabilities to enable data to be moved from one place to another using message queues, triggers, change data capture (CDC) and more.
• **Metadata mapping.** A comprehensive data integration offering should allow impact analysis and change analysis. It also needs capabilities for importing metadata from various systems and exchanging metadata with other systems.

**High-Level Guide to Data Integration Services**

In addition to allowing you to create your own data integration services that can be called in real time, near-real time or in batch processing, a good data integration solution needs to provide a rich set of out-of-the-box services that deliver the capabilities previously covered in this paper. Which services are used depends on the capability of the data integration platform you require (a set of services is shown below).

![Diagram showing various data integration services](image)

*Figure 3: A wide variety of out-of-the-box services should be included with your data integration solution.*

At this level, we can clearly see metadata services are very important in any data integration strategy because they should be pervasive as a key enabling and documentation mechanism. Metadata should be pervasive through all data integration programs 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 which data moved when, how it was changed, the business rules that were applied and the effects those changes might have. These are critical issues facing all organizations. Failure to place enough emphasis on metadata will result in problems later, often at a great cost to the organization.
Enterprise Connectivity

A data integration solution should provide connectivity using both native access with 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. The solution should also 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. When selecting a data integration vendor, organizations should look beyond the GUI to see what lies beneath the covers for both current and future needs.
The New Data Integration Landscape

Figure 5: A comprehensive data integration strategy depends on the integration of business initiatives, data integration capabilities, services to provide those capabilities, and last but not least, enterprisewide connectivity.
Your Data Integration Strategy

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

SAS focuses on the data quality aspects of data integration and real-time data integration to deliver 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 sophisticated data profiling.

The important thing is that whichever direction you take, SAS enables you to establish a data integration strategy while realizing the benefits of shared business rules, shared metadata and integrated technologies. These capabilities are combined with the associated cost reductions when employees need less training and business rules can be reused. In addition, tool and metadata integration issues are alleviated by an integrated, comprehensive approach, which means fewer maintenance and management problems. If you are not doing so already, today might be a good time to start thinking about the future of your data integration strategy. All the topics in the preceding landscape should work together to provide 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 provides all these strengths and stands ready to help you address your challenges today.
About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 60,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®.