Delivering information you can trust

June 2007

Best practices for high data quality in an SAP environment

Contents

Executive summary

Enterprise-wide application initiatives involve major investments in software, services, updated business processes, staff training and more. To accomplish their objectives, enterprise resource planning (ERP), customer relationship management (CRM), supplier relationship management (SRM), supply chain management (SCM) and business intelligence applications must operate from consistent information, which is often integrated from across the enterprise.

SAP® applications also need high-quality information from across the enterprise. Consistent and accurate data is essential to the success of an SAP application. Yet, at the beginning of an SAP deployment, most data to be used by the SAP environment is derived from existing data within the enterprise—and likely maintained in multiple independent systems. To integrate the data—in a consistent manner—often poses challenges that include aligning each source of master and transaction data to SAP requirements, standardizing data into a consistent format, harmonizing it to identify and remove duplicated data, filtering out non-relevant master data and more.

Fortunately, solving the data quality problem is a relatively small investment compared to the total cost of the SAP implementation project. IBM estimates that to manually migrate data would require 15–30 percent of an SAP implementation budget in order to provide the clean, integrated enterprise data necessary to support the SAP application. Using IBM technology and methodologies, however, can help to dramatically reduce this number in the first phase and potentially provide ever greater savings in subsequent phases. Yet data quality has a significant impact on the success of an SAP deployment.

The data integration and quality improvement process can be manual or automated. The manual approach may require large numbers of people, and results can vary based on the skills of the people doing the work. In addition, this approach makes it nearly impossible to maintain an audit trail and provide the ability to reuse data integration objects for subsequent phases of the implementation.

IBM provides a data integration methodology and software for automating the data integration process in an SAP environment. Based on business rules for transformations and algorithms for eliminating duplicate data, automated data integration systems provide automated documentation (metadata) of how transformations were conducted. If problems are encountered when loading data into SAP applications, metadata can direct efforts to change the transformation rules so the process can be repeated with better results. Defining these rules sets the stage for applying the same business rules in real time to new data as it enters the SAP software.

This white paper presents an overview of best practices—gleaned from many years of IBM teams working closely with customers deploying SAP—to overcome data integration challenges and boost data quality. By producing high-quality data for new SAP deployments, instance consolidations or major upgrades, IBM® Information Server software can not only help dramatically improve the effectiveness of SAP applications, but also help boost user adoption of those systems.

Through the automated information integration tools described in this white paper, IBM consultants have seen savings ranging from 47–81 percent for initial data integration activities. IBM estimates that savings during subsequent integration phases can reach up to 90 percent because of the large amount of integration infrastructure—business rules, processes and so on—that can be reused.

SAP environments depend on high-quality data

No one would argue that the success of SAP installations depends on the quality, consistency and accuracy of the underlying data, because data is always on the critical path of a successful implementation. After all, SAP applications are used to inform key business decisions, so the underlying data must be reliable.

Yet, those who have deployed an SAP application suite know that establishing integrated, consistent, accurate data can be challenging. In fact, one major risk factor for SAP deployments meeting their objectives is a lack of user acceptance because of data quality issues.

The difficulty arises because SAP applications touch so many parts of an enterprise. Most enterprise IT environments are composed of a collection of heterogeneous systems, each having very different data definitions and degrees of accuracy. If the data integration process does not fully accommodate the data complexities, users will not have a clear, consistent view of operations across their enterprise.

Assembling clean, consistent and accurate information from across the enterprise involves a wide range of activities:

- *Understand the legacy data—its meaning, its sources and the practices used to maintain it*
- *Standardize data that is entered in many different ways into a consistent format*
- *Cleanse the data by removing duplicates and creating a single instance of each master data object, such as customer, vendor, material and so on*
- *Augment the data to add required data that is not present or used in legacy systems, but required to run SAP business processes*
- *Resolve inconsistencies among legacy data sources*
- *Help ensure that data in various repositories—data warehouses, datamarts, master data, operational data stores and transactional applications—is accurate*

While SAP provides application programming interfaces (APIs) for loading data from across the enterprise, IBM provides the products and services to help SAP project teams implement data quality improvement initiatives. Data quality is a core discipline of the IBM Information Server platform, which is designed to integrate information from across an enterprise and deliver consistent, reusable and trusted information. For more information about the specific products, visit **ibm.com**/software/data/integration/info_server/

Incorrect or inconsistent data in SAP applications can have adverse effects across multiple parts of the enterprise. It may force workers to spend time trying to understand the meaning of inconsistent data. It can have a negative impact on customer care and customer satisfaction, and even carry legal ramifications in the form of restated financials or law suits.

Data quality considerations when changing or updating an SAP environment

For enterprises using SAP applications, the focus on data quality usually sharpens when major system changes occur—generally when new information sources are brought into the SAP environment. Triggers for enterprise data integration activities include mergers and acquisitions, streamlining or consolidating business processes, new SAP deployments, consolidation of multiple systems and major SAP upgrades:

- ***New SAP application deployments:*** *SAP is well known for the consistent evolution of its application suite. But any time new SAP modules are added, it is likely that new information must be obtained from other enterprise systems.*
- ***SAP instance consolidations:*** *Increasingly, enterprises will consolidate SAP applications to gain economies of scale, enhance inventory reporting, obtain a single view of the customer and improve the consistency of information across business units and geographies. Consolidations can also be driven by the desire to migrate from non-SAP systems or siloed applications that serve limited purposes.*

- ***Major SAP version upgrades:*** *As support ends for older versions of SAP R/3®, version upgrades often lead to data integration activities.*
- ***New deployments or expansions of SAP NetWeaver® Business Intelligence systems:*** *Assembling relevant data is essential to making sound business decisions from analytical software such as SAP NetWeaver Business Intelligence (SAP NetWeaver BI). This data might reside in both SAP applications and non-SAP applications from vendors like Oracle, PeopleSoft, Siebel and other legacy sources.*

Each of these scenarios brings data quality challenges. When data from multiple systems is combined, issues arise about consistency of definitions and practices for creating the data. This is also true when data from a legacy SAP application is combined with data from other enterprise systems.

Understanding the source systems and practices associated with data creation is a first step in creating a data migration plan. Vast amounts of information can be involved, making it important to profile and analyze source data, cleanse and correct information as it enters the SAP application and eliminate duplicate records before the combined data is used for decision making.

Data integration alternatives—manual and automated

Two common approaches to the data integration process are manual and automated. Some companies outsource data integration to low-cost regions of the world where transformations are done by hand. Others attempt to analyze and transform data manually with their own staff using spreadsheets and programming languages. Either way, manual integration can require large numbers of people, and results often vary based on the skill set of the people doing the work. Maintaining an audit trail for how the work was done is nearly impossible. Because information is dynamic, data migration is not a one-time activity, but an ongoing process to maintain data quality. The manual process becomes inefficient and cannot work under high data volume, so automated processes become attractive.

Automated data integration is based on rules for transformations and algorithms for aligning data and for creating high-quality data for SAP. Automated systems enable an iterative approach to migration—if the transformed data is not quite correct, it is easy to change the transformation rules and reload the data. These systems also provide a clear audit trail that shows how transformations were conducted, so if problems are encountered when loading the data, the rules can be changed and the process repeated. This iterative methodology is a very effective way to fine-tune the data cleansing process. Once those rules are finalized, they can be applied in real time to new data as it enters the SAP system, thus leveraging the initial integration investment over years of use.

Speed is another advantage of the automated data integration approach. Because source data continues to change during the weeks required for manual data integration, the manual processes always operate on outdated information. On the other hand, the speed of an automated process enables all the source data to be integrated and uploaded at once, which leads to a best practice for SAP: Load SAP systems early and often.

IBM approach to enterprise data integration

Initial conversion of data into SAP

- ***Discover****—Analyze source systems to identify the best sources, determine how each type of information will be prepared to form a consistent whole, and scope the size of the project*
- ***Prepare****—Apply that knowledge to execute data quality processes*
- ***Deliver****—Apply additional rules to refine the data, then load the data into the SAP environment*

Ongoing data quality processes

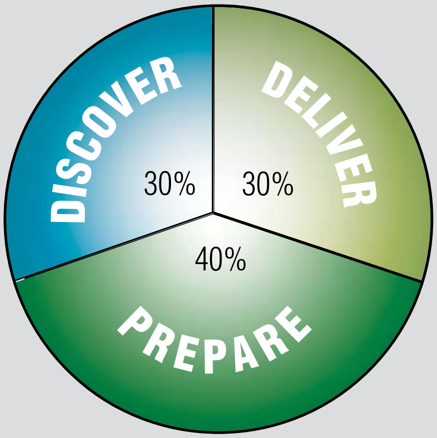
- ***Control****—Maintain high data quality as data is entered by users and provided by other applications*
- ***Monitor****—Proactively spot new data quality issues being introduced into the SAP environment*

The IBM automated enterprise data integration process

IBM has established an enterprise data integration process based on an extensive amount of experience helping enterprises solve data quality issues in hundreds of SAP deployments. Within the IBM methodology, three phases comprise the initial conversion process to help ensure clean data gets into an SAP environment. Following a logical step-by-step methodology, the conversion begins with a review of all legacy systems to understand the information that is currently available. The integration team must identify which systems have the best information, decide how to extract that information from each system, and then clean up and transform the data into a consistent set for loading into the SAP environment.

With an automated process, the work involved in the initial conversion is typically evenly distributed across these three phases, as shown in Figure 1. People accustomed to performing data integration by hand may find this surprising because most of their work in done in the Deliver phase. But with automation, the data transformation and loading work is on par with the other phases.

Figure 1: With automation, initial conversion is typically divided equally across three phases



Source: IBM estimates

Using traditional manual data integration methodologies, IBM estimates that as much as 15–30 percent of total an SAP project budget could be spent on data integration activities. By adopting the automated information integration tools described in this white paper, IBM consultants have seen savings ranging from 47–81 percent for the initial integration activities. Estimated savings during subsequent integration phases (such as additional SAP deployments in other parts of the enterprise) may be even higher—73–90 percent—because of the large amount of integration infrastructure such as business rules and processes that can be reused.

After legacy data is loaded into the SAP environment, the IBM approach helps to enable two ongoing (post go-live) activities that help maintain clean data in the SAP application as new data is added and changes to existing data are made.

The following sections review each phase in more detail.

Discover phase

The first step in the IBM automated enterprise data integration process is to identify all the legacy sources of data that will be used in the SAP environment. In a large enterprise, information about customers, for example, can be contained in hundreds of systems. The goal of the Discover phase is to identify the best sources of data and create specifications to migrate that data into a consistent whole within the SAP environment—to match similar information, eliminate duplicate records and clean up the legacy data. The outcome is a collection of best-of-breed records from all legacy systems.

The most popular discovery tool in use today is Microsoft® Excel. Although its advantage is minimal setup time, the manual discovery process itself is time consuming and error prone. The risk of the manual approach is that the conclusion from the Discover phase may be an inaccurate picture of legacy systems, which can lead to delays, cost overruns and a low level of confidence in the quality of the data uploaded to the SAP environment.

IBM automated enterprise-class discovery tools, on the other hand, help eliminate surprises and provide a solid foundation to scope the initial effort and define roles and responsibilities for the work that follows. IBM Information Server software helps the data migration team understand the content, quality and structure of source systems. It also identifies high-risk areas, which can help the team take action to improve the likelihood of the project achieving its goals (Goals may include a level of data quality as well as project costs and timeframes). Automated IBM tools also provide ways for business users to understand how the data will be transformed prior to loading into the SAP environment and to participate in creating the transformation rules.

The Discover phase has two steps. First, a preliminary assessment of legacy data sources helps select the data sources that will supply the *core attributes* of the SAP master data object. These attributes are required by the SAP applications regardless of industry—attributes such as customer name and address. The second step is a more detailed assessment to identify how all the SAP attributes will be created—SAP has over 1,000 attributes, some used only in specialized industries and others used across all industries.

Discovery step 1: Quick, high-level assessment for planning purposes. Used to create project plans and budgets, a preliminary assessment provides a general understanding of data transformation requirements and their anticipated complexity. The project team scopes the amount of hardware and software required to build the business rules and migrate the data into SAP, defines roles and responsibilities, identifies risks and so on.

When early assessments of project size are neglected, teams can encounter roadblocks during the migration phase that may be caused by insufficient capacity or the lack of software to perform needed functions. Therefore, best practices suggest making sizing estimates available as early as possible in the project—before plans are defined, budgets set and schedules for rollout determined.

Also, if assumptions are wrong about the amount of work needed to prepare the data for inclusion into the SAP environment, all aspects of the project could be affected. Surprises at the Deliver phase of the project—where the team is actually building the data transformation and user requirements—can cause project goals to be missed.

Discovery step 2: Full baseline assessment of source data. Once the overall plan is established, the next step is a full assessment of how data in the source systems will be combined to supply all the attributes required in the SAP environment. This process involves detailed analysis, so the timing depends on the scope of the assessment.

The output of the discovery phase is a complete specification of how the data will be migrated from the legacy systems—which systems will provide the data, what processes will be followed to transform the data and how much effort will be required. This step also identifies all the challenges anticipated during the actual migration.

The following activities are typically performed on legacy data prior to uploading into an SAP environment:

- ***Master data relevancy.*** *Obviously, not all legacy data will be used in the SAP environment. In this activity, the project team determines specific source data that will be migrated to SAP—what data is relevant and what is not. For example, objects that have become obsolete are probably not desired in the new SAP implementation. But the selection process can be complex. For example, because it is often hard to remove data from legacy systems, obsolete records may simply contain a notation such as "do not use." Therefore, the team must define techniques and implement processes to filter out those records during the Deliver phase.*

- ***Data augmentation.*** *SAP may have business processes that were not a part of the legacy systems, so the proper data may need to be created during the transformation process. This may require that some information, such as a Dun & Bradstreet number, be obtained from outside sources.*

- ***Data transformation*** **(standardization).** *Because source data likely resides in multiple legacy systems, this step determines how to transform the data into the master data object within the SAP application. For example, source systems may not contain all the fields required by the SAP environment—structures are likely to be different. Even if the field exists in the source systems, conventions used to represent data will generally differ from those used in SAP, requiring source data to be standardized into a common form. Transformation operations performed on the data help to accomplish the standardization. The IBM process includes an iterative gap analysis, which compares the fields required by SAP with the fields in the source systems. Typically the gap narrows as transformations are refined.*

  IBM Information Server performs this standardization and matching analysis to help ensure that information behind business decisions reflects the facts in the real world and provides an accurate view across the enterprise.

- ***Data matching and de-duplication*** **(harmonization).** *Duplicate records are common in enterprise systems. Once data is standardized, multiple techniques are used to identify duplicate records across source systems. IBM Information Server can match objects accurately—names against names, addresses against addresses and so on. For example, IBM software can recognize "IBM Corporation," "IBM Corp." and "International Business Machines" as the same customer. But there may be times when a user must interpret the data: Is this record really a duplicate of the other record? Two customer records might be very close, yet one is a bill-to customer record and the other is a ship-to record. Resolving that might require identification by a knowledgeable user.*

- ***Cross-functional completeness (full assessment only).*** *This check helps ensure that each SAP process contains source data. Project teams create rules to measure data quality and audit the data as it is being migrated into the SAP environment. These rules can also be continuously applied to new data as it is added to SAP .*

Prepare phase

IBM Information Server is the unified data integration platform used in the Prepare and Deliver phases. Operating much like an assembly line, it contains integrated data quality functions (preparation, cleansing and delivery activities). In the Prepare phase, IBM Information Server processes source data according to the rules defined in the Discover phase. These rules transform the unstructured, free-form source data and standardize it into a consistent format. The data is harmonized and augmented, identifying any required missing data and filtering out master data that is not required by the SAP environment.

A *metadata repository*, shared across all IBM data quality processes, contains the rules to direct this phase. Thus, if a business user creates a rule to transform data from a particular legacy system, that rule is saved in the metadata repository. IBM tools, such as IBM Information Server, can then use those rules to direct each step of the automated data integration process.

The metadata repository is the one central place where the data integration process is defined. It provides an audit trail that can be followed even when all records are not migrated during the transformation process. And if a user asks how a data field was derived, the project team member can simply go to the repository and find the business rule. For example, assume the legacy systems had three representations of customer number: CUST_NO, AP_NO and ACCOUNT_NUMBER. The metadata repository would show that those three fields were all linked together to produce a single customer number in the SAP environment.

The rules created in the Prepare phase will be used to evaluate the data prior to loading to make sure it is correct. These rules will also be used in the Monitor phase after the SAP environment goes live.

Next, all the best-of-breed records from the source data are handed off to the Deliver phase within IBM Information Server.

Deliver phase

IBM Information Server applies the remaining business rules to the data and loads the resulting data into the SAP environment.

Business rules applied in this phase require that transformations in the Prepare phase are completed. Examples include:

- *Code and date conversions*
- *Data type conversions*
- *Data mapped to values used in the SAP environment*
- *Faulty logic identified*
- *Additional data augmentation*

After applying the additional Deliver phase rules, the data is stored in a *master data integration (MDI) hub*—a staging/provisioning area for data that has been processed through all the data quality activities. The MDI hub can drive multiple targets, including the SAP environment, a master data management (MDM) application, a data warehouse and departmental datamarts.

Loading data into the SAP environment is an iterative process because SAP software may reject early attempts, providing numerous error messages. The team fixes the problems and makes another attempt to load the data, repeating the process until an acceptable percentage of the data is loaded.

In automated integrations, spot assessments of the data using the rules created during the Prepare phase can help identify and resolve issues with the data. In manual data integration activities, post go-live data maintenance activities are common but can lead to expensive delays in the completion of the project. Adoption of the IBM automated information integration process all but eliminates post go-live corrections.

Control phase

Without ongoing efforts to maintain data quality, the quality of SAP data degrades with use (typically 2–5 percent per month). For example, as new data is added, it inevitably includes duplicate records, which can have a serious impact on the business. For example, a duplicate customer record could cause replacement parts to be sent to the wrong address. A customer service rep looking at a duplicate record could miss recent activity or the special handling needs of a customer. Even incorrect inventory levels might be reported.

Controlling data quality can help the enterprise continue to reap the optimal return on its investment in the SAP environment. Maintaining effective operations requires ongoing data quality processing at the point of data entry, optimizing new incoming data in real time.

Using IBM data quality software, all the rules and processes defined for the initial data conversion project can be leveraged to help maintain high-quality data even as new data is being entered into SAP. By eliminating much of the work to set up real-time data quality processes, the overall cost of data quality initiatives can be reduced.

IBM Information Server can check the quality of new data in real time as it enters the SAP environment—whether it originates from a legacy system or an agent enters it using the new SAP application. IBM Information Server connects to the SAP environment through SAP-certified interfaces, so integrity is maintained while the IBM tools appear as if they are integrated into the SAP workflow.

Through a certified SAP interface, IBM Information Server Data Quality Module for SAP provides two continuous data quality capabilities: duplicate prevention and error-tolerant searching. Using powerful matching techniques from IBM Information Server, the Data Quality Module identifies potential duplicate records (based on probabilistic matching criteria) when the user enters a new record into the SAP application. The module gives the user the opportunity to continue to enter the information as a new record or modify an existing record. Thus, users can easily do their part to maintain high-quality data in the SAP environment.

To complement the exact search capabilities of SAP applications, IBM also offers an error-tolerant search engine. Using the same technology that is used to detect duplicates, IBM OmniFind™ Enterprise Edition software enables users to find an object in the database even if a data-entry mistake prevents an exact match. This helps improve user productivity when searching for items such as customers and business partners.

Monitor phase

The IBM automated information integration process includes ongoing SAP data quality monitoring. IBM Information Server can monitor service-level agreements, auditing and data governance requirements such as those required for compliance with the Sarbanes-Oxley Act. The same rules and processes defined for the initial data conversion project can be leveraged here. Not only is data quality assessed periodically, but data quality trend analysis over time can highlight additional problem areas.

Repurposing SAP data in external data warehouses

Frequently the need arises to analyze core SAP data together with data from other systems throughout the enterprise. While SAP environments have extensive reporting capabilities, such as those in SAP NetWeaver BI, this analysis and reporting need may be filled by an enterprise data warehouse or a departmental datamart such as IBM DB2® software.

When an external application needs data from other systems, the MDI hub created in the Deliver phase can provide the data. Thus, the external application can leverage the same IBM data quality processes used to build and maintain SAP-resident data. Leveraging the IBM data quality process provides a cost-effective approach to enabling high-quality data to external analysis applications.

For more information

To learn more about the technologies and products behind IBM information integration solutions, contact your IBM marketing representative or IBM Business Partner, or visit **ibm.com**/software/data/ips/solutions/sap.html

IBM Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
June 2007

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. Offerings are subject to change, extension or withdrawal without notice.

All statements regarding IBM future direction or intent are subject to change or withdrawal without notice and represent goals and objectives only.

The information contained in this document is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this document, it is provided "as is" without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this document or any other documents. Nothing contained in this document is intended to, nor shall have the effect of, creating any warranties or representations from IBM Software.