

The Five Pitfalls of Data Migration – and How to Avoid Them

WHITE PAPER



This document contains Confidential, Proprietary and Trade Secret Information (“Confidential Information”) of Informatica Corporation and may not be copied, distributed, duplicated, or otherwise reproduced in any manner without the prior written consent of Informatica.

While every attempt has been made to ensure that the information in this document is accurate and complete, some typographical errors or technical inaccuracies may exist. Informatica does not accept responsibility for any kind of loss resulting from the use of information contained in this document. The information contained in this document is subject to change without notice.

The incorporation of the product attributes discussed in these materials into any release or upgrade of any Informatica software product—as well as the timing of any such release or upgrade—is at the sole discretion of Informatica.

Protected by one or more of the following U.S. Patents: 6,032,158; 5,794,246; 6,014,670; 6,339,775; 6,044,374; 6,208,990; 6,208,990; 6,850,947; 6,895,471; or by the following pending U.S. Patents: 09/644,280; 10/966,046; 10/727,700.

This edition published June 2010

Table of Contents

Executive Summary	2
Failure to Follow Best Practices	3
Team Structure	3
Risk Mitigation	3
Data Discovery	3
Legacy Retirement	3
Agility	4
Audit and Validation	4
Mock Loads	4
Production Data	4
Data Quality	4
Cutover	4
Skipping Data Discovery	5
Master Data Discovery	6
Data Profiling	6
Legacy Retirement	6
Target Impact Analysis	6
Incomplete Data Movement Strategy	7
Reusability	7
Data Quality	7
Audit and Validation	8
Connectivity	8
Lack of Collaboration	9
Inadequate Tools	9
Conclusion	12

Executive Summary

Your company is planning an application data migration project. Perhaps you are modernizing your business processes, to save costs or become more competitive. This plan requires retiring legacy applications and implementing new applications to support your new approach to business. Maybe you're integrating data from a merger or acquisition. Whatever the reason, data migration is the lynchpin of the larger initiative in which your company is investing strategy, budget, and precious time. Your program will not succeed if delivery of the data is delayed or if the data does not correctly support your business. You can't afford to fail.

Five common errors can delay or even doom the outcome of a data migration project:

1. **Failure to follow best practices.** Data migration requires specific skills, tools, and plans unlike those necessary for other IT projects.
2. **Skipping data discovery or not understanding your data.** You must understand the source or target application data and address how to access it over time.
3. **Incomplete data movement strategy.** Moving the data must include strategies for access, validation, and audits.
4. **Lack of collaboration.** Business users and data stewards must be involved in verifying data and ensuring that it's fit for use.
5. **Inadequate tools.** The appropriate tools to support data migration include all needed processes and don't force IT to reinvent the wheel repeatedly.

This white paper outlines each of these pitfalls in more detail and provides advice on ways to avoid them. This is not a guide for managing an application data migration overall, but rather a set of advice on how to steer clear of the “gotchas” that plague many project teams.

Failure to Follow Best Practices

Many IT teams assume that they can apply the same best practices to application data migrations as they do to application code development. Although this approach seems to be logical, it actually leads to delays and failure. Why? Because the deliverable for data migration is unlike that of any other IT project. For application code development, the goal is to deliver code that adds new functionality or improves processes. For application data migration, the goal is to deliver a production-ready set of data for a new application. This is a key difference and one that you must take into account as you structure the project. The application data migration team should understand which practices are distinct to your project, including the following:

Team Structure

As you draw up a plan outlining the players and their roles, include everyone with an interest in the outcome: business stakeholders, business data experts or data stewards, and data architects, plus the team leader and developers. Large migrations should also include a project manager and a team dedicated to testing data quality and ensuring it works in the new application.

Risk Mitigation

The outcome of a data migration has enormous ramifications for the larger project it supports. As a result, it is critical for the data migration team and the larger project team to work together to align expectations, objectives, and timelines. For example, a project team implementing a new SAP ECC6 instance must include the data migration team in planning configuration changes as well as the approach to phasing in the live application. If communication breaks down, so will the project.

Data Discovery

Application data migration focuses much more closely than other projects do on data content in the initial analysis phase. It is necessary to determine what data to pull forward, what to do with data not being pulled forward, and how to map and enhance data for the target application needs. Mapping based on field names listed in DDL or copybooks is not sufficient, rushing through the data analysis may lead you to misidentify sources, and skipping basic profiling may conceal missing content. Instead, the data migration team must examine the data on both the source and target applications—accessing the data on all applications and across systems—to understand the existing content, rules, and relationships.

Legacy Retirement

Retiring source applications, servers, or databases delivers significant gains, so planning for retirement must be an integral part of the data migration. In addition, the team must examine any data not needed by the new application to determine how to archive it, for how long, and how to provide access to it if necessary.

Agility

During an application data migration, timelines can be aggressive, and discoveries about the state of the data can force significant changes in requirements with little notice. As a result, the data migration team needs an infrastructure that allows them to change and update access components and mappings quickly. They must be able to understand both the data and the requirements, map and test the data in multiple mock loads, and cut over the data to production with minimal impact on current business processes.

Audit and Validation

The focus on compliance and data quality once limited to financial services is now common in retail, health care, the public sector, and other areas. As a result, the data migration team needs an infrastructure and a solution that easily address audits of record counts, content, and mappings. They allow the team to show what was extracted and loaded, what was not, and the reasons for those choices.

Mock Loads

Data migration testing involves running multiple data loads into a test version of the target application, evaluating performance, and identifying the reasons any data failed to load before testing whether the data actually works in the target application. Testing the load process, and the target functionality, should begin early in the project.

Production Data

Given that the final deliverable of the data migration project is a production-ready set of data, the only way to accomplish the data migration is to extract and map actual production data. On most types of IT projects, developers test a new process or interface by using a specifically created test data set or a small subset of production data. For a data migration, this practice is a mistake. The data migration team is testing, cleansing, and enhancing production data. The team cannot meet its objectives if it is not using the real data.

Data Quality

You always need to ensure that the data will work in your target application. This verification often requires data cleansing, enhancement, and consolidation. If your target application implements new business functionality or consolidates business functions, plan to include data quality work in your migration project.

Cutover

The cutover process of loading data to production must be driven by business needs, risk tolerance, and customer experience. The cutover plan is one of the first things that the team needs to accomplish. It drives the entire approach. You may or may not need to minimize downtime. You may be able to move the majority of your transactional data weeks before your cutover date. You may be risk averse enough to want parallel processing. For example, a property and casualty insurance company developed a cutover process that specifically met its business needs. It migrated the data associated with each account to the new application on each customer's annual renewal date. The process required a full year, but actually had zero downtime and no negative customer impact.

Skipping Data Discovery

If the team that understood your old mainframe data has long since retired, your systems have grown organically over many years, or your company is trying to merge data from an acquired competitor, then you may have no repository of information about the legacy applications and their data. You don't know what you don't know, as the saying goes—and that can cause real problems in a data migration project.

One public sector organization, for example, needed to migrate data from a legal case management system. The data stewards identified the central case table as the source for the case data. However, a data profiling exercise revealed that roughly 20 percent of initial case records were stored in a secondary table and only entered into the central table later when their status changed.

A data steward misidentifying the actual source of data is not uncommon. Rarely does the actual data content look anything like the team is expecting. Data stewards and other users tend to see data only in an extracted, aggregated, or presentational form. Data migration development teams, who work with granular, table-level data, can also be challenged in knowing what data meets business needs. Although the data model, relationships, column names, and structures help provide the metadata necessary for a good understanding of data, they rarely offer the full picture. For example, in the COBOL world, data objects are “redefined,” while in Java, values are “overloaded.” For the past 30 years or more, developers have been naming a data object one thing when the actual data stored in the object is something entirely different.

Traditionally, the technical team wants to get the requirements, start coding, and iterate through refinements. This well-intentioned approach is how teams fall into the “lack of data knowledge” pitfall. For an application data migration, it's better for data discoveries (and their surprises) to occur primarily at the beginning of the project. If you do the data analysis up front, you minimize the need for costly extract and mapping rework. By avoiding the need to completely remap data or redesign the extract strategy, you also limit the risk of extended delays. This is one time when the IT team cannot skip data analysis in favor of moving on to the “real work.” Data analysis is the real work.

To avoid the impacts of not understanding your data that are caused by skipping or minimizing data discovery, the team should take the following steps:

Master Data Discovery

The first step in a data migration is not field-level mapping, but entity-level analysis to determine the master data entities needed for the target application. Identify the source of product, customer, vendor, or case data and validate it against other sources (which may include the target, in consolidations) of the same data. Evaluate the primary key conventions across databases, determine if matching or consolidation is needed, and analyze gaps across systems. For example, a team may need to understand why the billing system shows 20,000 customers but the provisioning system shows 20,500, and what superset or subset of data needs to be pulled forward.

Data Profiling

The next step is table- and column-level profiling. The team needs to evaluate inconsistencies, redundancies, inaccuracies, and referential integrity across tables and data sources. The resulting reports and metrics will provide the data points needed to further understand and analyze the content.

Legacy Retirement

As part of the data migration, the team must identify a minimal subset of data to map to the new target application. Older data sets, with stale and poorly understood data, result in exponentially more work. The team should also address long-term data access issues for data that is not being migrated, including compliance requirements, legal requirements, or historical trend reporting needs.

Target Impact Analysis

The team must consider how three issues of managing data impact the target application:

- Data governance after migration, ideally by reusing the knowledge and business rules developed in preparing the data for migration
- Aligning target reference and master data with source data, possibly through configuration changes on the target
- Ensuring performance of the target application, particularly in consolidations involving moving a large amount of master or transactional data to an existing application

Incomplete Data Movement Strategy

Moving business-critical data needs to be done purposefully and carefully. After all, this data runs your business. Many teams assume that a data migration is a simple mapping of data from one table to another. But it is so much more. Just to name a few examples: the business user will likely demand audits of the move, there will almost always be challenges about the quality of the data, and there will be frequent project changes and data discoveries that can create substantial delays. Also, if data isn't moved in a way that ensures that it supports the new application, your company's operations can be adversely affected. A data migration is risky, and the pitfall is not realizing how to mitigate that risk during the move. Below are four areas to focus on that will allow the application data migration team to move data that works on time.

Reusability

The risk of hand-coding data migration components, or of not creating a distinct set of business rules associated with the target application, is that the team will need to recode, rework, and recreate many of these components for subsequent migrations, data governance needs, or data quality projects. You may be able to move the data once, and confirm that it is correct, but when the next requirement arises to migrate to the same source, to cleanse, or monitor the same data, the effort will mostly likely be When you consider the amount of data discovery required to move the data correctly, this lack of reusability is a considerable waste of time and resources going forward.

One organization in the public sector consolidated data from 70 diverse source applications to a single SAP instance for financial reporting. By ensuring that the business rules for the target application were reusable, the organization was able to complete multiple distinct migrations with full confidence in the data quality and save time doing so.

Data Quality

The definition of data quality can differ based on context. For example, if ZIP code data for regional churn reporting is identified as 98 percent correct, that may be sufficient. For a billing system, on the other hand, data that's only 98 percent correct is completely unacceptable. In a data migration, the data must support the business processes in the target application. Data migration teams often make the mistake of assuming that the data is fine or good enough.

Data quality processes are iterative, requiring hands-on involvement from data content experts, most often business users. One common example involves matching and removing duplicates. Without involving content experts, a company may end up in the position of a U.S. health insurance company, which hired a team of hand coders to help migrate provider information to a new claims processing system. Because the migration team had little understanding of the data or the business processes, they missed many duplicate records and introduced errors into the target database. The erroneous data set moved into production, causing unprecedented delays and errors in claims payment processing, which in turn resulted in the loss of participating providers and subscriber groups. This mistake had a huge negative impact to this business.

By contrast, when a European retailer migrated off the antiquated mainframe application managing its global supply chain, it assigned a team of business data experts to help the data migration specialists develop matching algorithms for consolidating product data. Instead of having one product record for a specific brand and model of sneakers in red and another product record for the same brand and model in blue, the company was able to move data that had a single sneaker record, with optional attribute colors. It could now uniquely identify products. The result had a hugely positive impact on supply chain processes.

Data quality matters to your business processes. A data migration provides both an opportunity and an imperative for the team to evaluate the gaps, and the needs, for cleansing and deduplicating the data moving forward. The application data migration team must be able to address data quality with common best practices and tools. And the business team must be involved in each step of the data migration process.

Audit and Validation

Increasing governmental and industry regulations make it critical to build auditing and compliance procedures into data migration projects. The data migration team must be able to prove that the data from the source application is also the data that is being loaded to the target application—not only for regulatory purposes, but also to ensure business confidence in the new application. Several approaches can facilitate this validation:

- **Data quality scorecards.** As part of ensuring data quality, business users define the data quality metrics necessary for them to accept the data as production ready.
- **Source-to-target record validation.** The data migration includes a distinct process to validate that the records from the source system have moved to the target, and if not, why. This process may be simple record counts, but it may be more complex when data is transformed to meet the needs of the target application.
- **Validation to a secondary source.** The data migration team finds a secondary source of data trusted by the business users, then validates both to the original source and the secondary source. Note that this approach creates more work for the team and may uncover additional unexpected data issues.
- **Metadata management.** Being able to trace the lineage of the data as it moves through the data migration helps the data migration team resolve data issues as they arise and measure the impact of changes made to any mapping or process.
- **Documentation.** Well-documented mappings in the data migration enable auditors and business users to track changes.

Connectivity

If agility is key, data quality is iterative, and data discovery leads you to new paths and realizations, connectivity supports them all. Therefore, it must be agile, allowing the team to run extract jobs on short notice, change mappings as needed, and address data quality along the way. If possible, the team should rely on jump-start extracts to their source applications that are under their control and easy to modify.

A good example of the impact of not having an agile connectivity approach is that of a U.S. manufacturing company, which is consolidating dozens of applications into a single SAP instance as part of a global modernization program. For one source system, the team used existing monthly extracts from a legacy mainframe application. Although these extracts included all the data the team expected to use, they were unable to test changes that they were making to the source data. They were also unable to do a mock full-target load more than once a month. This restriction prevented the business data stewards from confirming the validity of their data quality updates on the source system and the business rules captured by the development team. They could only do a test run once each month. With no alternative to waiting a month for each test run and validation cycle, the project was significantly delayed.

Lack of Collaboration

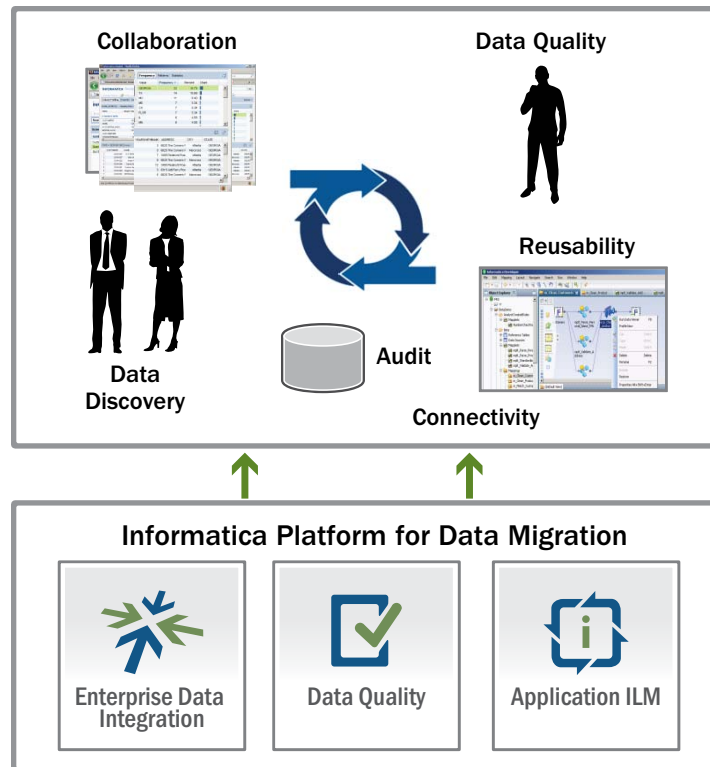
Consider this scenario: A data migration team discovers that the client data does not have a clear indicator or attribute to identify whether the client is a person or a company. The source system did not require the distinction, but the target system does. The developers spend weeks testing various string searches and algorithms and making choices based on content, manage to sort out 95 percent of the data, but struggle with the remaining 5 percent. Instead of trying to solve the problem alone, the developers could move ahead by asking the data stewards or the implementation team about the confusing data. They could ask whether a default value might work or what the impact of the current data quality might be—but they don't. Instead, they spend weeks working on this single issue. This story is not uncommon. Most data migration teams do not recognize the value of collaborating with the business users and the data stewards.

Avoiding this pitfall is simple: Involve the data experts, generally the business users, who may even be formally identified as data stewards on large projects. These are the people able to make necessary decisions about when data is “good enough,” what needs to be fixed at the source, what needs to be enhanced and how, and which transformations are working properly. They determine what data to pull forward and what to archive. They respond to data quality scorecards and approve data quality fixes. It's not enough to include business users only in the initial source to target the mapping process and acceptance testing. Data migration projects that succeed—that come in on time and on budget with the full confidence of end users—involve business users from start to finish.

Inadequate Tools

Too many teams either attempt to hand code a data migration or do not use the tools that they have effectively. Each data migration is unique and every business imperative is different, but certain common needs cut across all industries and all types of applications. Choosing the right tools and knowing how to use them dramatically improve the odds of success for every application data migration. The following table lists the known success factors for data migrations and the associated tools. It also maps them to the capabilities of the Informatica® Platform, which supports enterprise data integration needs from data warehousing and master data management to data synchronization and data quality while enabling strong collaboration between IT and business.

For data migration, the Informatica Platform includes data integration (DI) tools and capabilities supporting data movement, reusability, connectivity, and auditing. It also includes data quality (DQ) tools and capabilities for data discovery and cleansing. In addition, the Informatica information lifecycle management (ILM) family enables necessary archiving for legacy application retirement and target application performance, as well as the data subset functionality needed for an agile extract process and data masking to ensure security.



Informatica provides a comprehensive, unified, open, economical data integration platform to support data migration best practices.

The same platform can support your ongoing data integration needs, including data synchronization, data governance, and master data management. Many of our customers cite increased practitioner productivity as one of the largest benefits of the platform because their teams can quickly create data integration processes that provide high-quality data.

Success Factor	Product Family	Informatica Platform Capabilities
Data Discovery	DQ, ILM	<ul style="list-style-type: none"> Identify data quality problems at the start of the data migration to avoid delays Manage data growth Support regulatory compliance Protect sensitive data Safely retire legacy systems and applications
Data Movement: Reusability	DI, DQ, ILM	<ul style="list-style-type: none"> Access virtually any and all enterprise datatypes, including: <ul style="list-style-type: none"> Structured, unstructured, and semistructured data Relational, mainframe, file, and standards-based data Message queue data Eliminate the need for recoding by using a flexible, metadata-driven architecture that standardizes and reuses definitions across platforms and projects Reuse all profiling and rule specifications from business analysts and data stewards across all applications and projects
Data Movement: Data Quality	DQ	<ul style="list-style-type: none"> Analyze, profile, and cleanse data Define and model logical data objects Combine data quality rules with sophisticated data transformation logic Conduct midstream profiling to validate and debug logic as it's developed
Data Movement: Audit and Validation	DI, DQ	<ul style="list-style-type: none"> Consolidate metadata into a single integration catalog to increase insight into complex data relationships and engender trust in the data that drives business decisions Validate data transformation and data quality mappings using a distinct reconciliation process Profile, analyze, and create data quality scorecards Drill down to specific records with poor data quality to determine their impact on the data migration and how to fix them
Data Movement: Connectivity	DI, ILM	<ul style="list-style-type: none"> Discover and access all data sources—whether they're on premise, with partners, or in the cloud Access and update enterprise data without needing specialized programming skills in: <ul style="list-style-type: none"> Major enterprise and packaged applications, whether on-premise, outsourced, or hosted software as a service All major enterprise database systems and data warehousing environments Mainframe systems Midrange systems Message-oriented middleware (MOM) Industry-wide technology standards such as email, JMS, LDAP and Web services Access database changes as they occur with change data capture capabilities Locate data rapidly with flexible object filtering techniques to reduce errors and speed development with a point-and-click interface Create subsets of data as it moves into the data migration staging (or development) environment, effectively providing the team with a jump start to the project Mask key data for privacy and compliance as it moves into the data migration development and test environments
Collaboration	DI, DQ, ILM	<ul style="list-style-type: none"> Provide team with robust visual tools and powerful productivity tools to facilitate collaboration among architects, analysts, and developers Supply business analysts and data stewards with tools they can use to profile data themselves Profile, analyze, and create data quality scorecards Drill down to specific records with poor data quality to determine their impact on the data migration and how to fix them Monitor and share data quality metrics and reports by emailing a URL to colleagues Define data quality targets and valid reference data sets Specify, validate, configure, and test data quality rules Collaborate efficiently with IT developers to share profiles and implement data quality rules Identify anomalies and manage data quality exception records Track data quality targets on an ongoing basis Finally engage all the right people in improving data

Conclusion

Implementing the appropriate tools and processes allows the application data migration team to avoid common pitfalls that lead to cost overruns and delays. The Informatica Platform is a comprehensive, unified, open, economical data integration platform that includes all the necessary tools to follow best practices for data migration. Its many capabilities help your IT team manage and understand source data, both for discovery and for long-term access needs. It enables them to include robust and agile access, validation, and audit strategies in your data migration. And finally, the Informatica Platform lets your business users and data stewards verify and ensure that data is fit for the collaborations and tools that drive your day-to-day operations.

Learn More

Learn more about the Informatica Platform. Visit us at www.informatica.com or call +1 650-385-5000 (1-800-653-3871 in the U.S.).

About Informatica

Informatica Corporation (NASDAQ: INFA) is the world's number one independent provider of data integration software. Organizations around the world gain a competitive advantage in today's global information economy with timely, relevant and trustworthy data for their top business imperatives. More than 4,000 enterprises worldwide rely on Informatica to access, integrate and trust their information assets held in the traditional enterprise, off premise and in the Cloud.



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