An Oracle White Paper
October 2011

Successful Data Migration

ORACLE

Part 1: The Importance of Data Quality

Plotting a Smooth Path to Data Migration

Businesses spend billions of dollars migrating data between information-intensive applications. Yet up to 75 percent of new systems fail to meet expectations, often because flaws in the migration process result in data that is not adequately validated for the intended task.

Because the system itself is seen as the investment, any data migration effort is often viewed as a necessary but unfortunate cost, leading to an oversimplified, underfunded approach. With an understanding of the hidden challenges, managing the migration as part of the investment is much more likely to deliver accurate data that supports the needs of the business and mitigates the risk of delays, budget overruns, and scope reductions that can arise.

Why Migrate Your Data?

Data migrations generally result from the introduction of a new system. This may involve an application migration or consolidation in which one or more legacy systems are replaced or the deployment of an additional system that will sit alongside the existing applications. Whatever the specific nature of any data migration, the ultimate aim is to improve corporate performance and deliver competitive advantage.

Accurate data is the raw material that maximizes the value of enterprise applications. However, when existing data is migrated to a new target application, it can become apparent that it contains inaccuracies, unknowns, and redundant and duplicate material. And although the data in the source system may be perfectly adequate for its current use, it may be wholly inadequate, in terms of content and structure, for the objectives of the target system.

Without a sufficient understanding of both source and target, transferring data into a more sophisticated application will amplify the negative impact of any incorrect or irrelevant data, perpetuate any hidden legacy problems, and increase exposure to risk.

Does Data Get the Attention It Deserves?

Data migration is usually part of a larger project deliverable, and typically the majority of business attention is focused on the package selection and configuration rather than on ensuring that the data that populates the new system is fit for the purpose. There are some clear reasons why data migration subprojects tend to be "planned" so cursorily. Choosing the new system is an exciting, strategic business activity that usually entails working with new technologies, suppliers, and opportunities. In short, it is the sexy part of the project. In contrast, data migration planning is seen as a simple matter of shifting data from one bucket to another via a process that is a necessary administrative burden and an extra cost. Thus, planning is often left until too late and the required resources and the difficulty of the migration are frequently underestimated. Migration is regarded as a mundane and thankless task, and in some instances, people know they are migrating themselves out of a job.

Key Drivers of Data Complexity

A combination of trends is accelerating the need to manage data migration activity more effectively as part of a corporate data quality strategy:

- **Corporate growth**. Mergers, acquisitions, and restructuring of disparate systems create new data and new datasources.
- **Compliance.** Data must be validated against regulations and standards such as Basel II and Sarbanes-Oxley (SOX).
- **Data volume.** Escalating amounts of data are increasing the burden of data management.
- **Data diversity.** Introduction of data in new formats—such as RFID, SMS, and e-mail—increases complexity.
- **Data decay.** Data is volatile; customer data typically deteriorates at a rate of 10 percent to 25 percent per year.
- **Data denial.** Organizations are often unaware of their data quality issues and lack the expertise or a senior sponsor to champion decisive action.
- **Technical advances.** Proliferation of new data devices, platforms, and operating systems also contributes to complexity.
- **Economic factors.** With pressure on margins, all corporate data must help organizations. compete more effectively

Migration Strategies

Organizations planning a data migration should consider which style of migration is most suitable for their needs. They can choose from several strategies, depending on the project requirements and available processing windows, but there are two principal types of migration: *big bang migrations* and *trickle migrations*.

Big bang migrations involve completing the entire migration in a small, defined processing window. In the case of a systems migration, this involves system downtime while the data is extracted from the source system(s), processed, and loaded to the target, followed by the switching of processing over to the new environment.

This approach can seem attractive, in that it completes the migration in the shortest-possible time, but it carries several risks. Few organizations can live with a core system's being unavailable for long, so there is intense pressure on the migration and the data verification and sign-off are on the critical path. Businesses adopting this approach should plan at least one dry run of the migration before the live event and also plan a contingency date for the migration in case the first attempt has to be aborted. The reality is that few organizations ever do this. Big bang migrations are most often planned as a one-off requiring a shutdown over a weekend or a public holiday, meaning that the quality of the migrated data is often compromised.

Life Insurance Case Study

A well-known life insurance company had allocated an entire weekend to perform a big bang data migration of its back-office application to a new system. Analysis and benchmarking by the IT services company commissioned to carry out the work revealed that the whole process would take more than two weeks. The migration solution architect commented, "A major rethink of the migration strategy was required!"

Trickle migrations take an incremental approach to migrating data. Rather than aim to complete the whole event in a short time window, a trickle migration involves running the old and new systems in parallel and migrating the data in phases. This method inherently provides the zero downtime that mission-critical applications requiring 24/7 operation need. A trickle migration can be implemented with real-time processes to move data, and these processes can also be used to maintain the data by passing future changes to the target system.

Adopting the trickle approach does add some complexity to the design, because it must be possible to track which data has been migrated. If this is part of a systems migration, it may also mean that source and target systems are operating in parallel, with users having to switch between them, depending on where the information they need is currently situated. Alternatively, the old system(s) can continue to be operational until the entire migration is completed, before users are switched to the new system. In such a case, any changes to data in the source system(s) must trigger remigration of the appropriate records so the target is updated correctly.

Often a data migration results from the introduction of a new additional system: an enterprise application or perhaps a data warehouse to support business intelligence. In such a case, the initial data migration is not the end of the matter, because the source and target systems will coexist and the congruence of their data needs to be maintained. If the target system is purely analytical, it may be possible to satisfy this requirement by rerunning the entire data migration periodically, providing either a full or a partial refresh of the data, in which case either approach to data migration can work. If changes may be made to the data in both the source and target systems, synchronizing them becomes far more complex. In such a case, organizations should consider whether the best route to achieving the migration is actually to deploy real-time data integration processes from the outset and trickle-load the new system.

Part 2: Formulating a Strategy

Challenges and Pitfalls of Classic Data Migration

A data migration project typically starts with a broad brief from the business to the IT team that leads to a technically focused migration in which more data is moved than necessary, at a greater cost over a longer period of time than was forecast, resulting in multiple revisions at numerous stages.

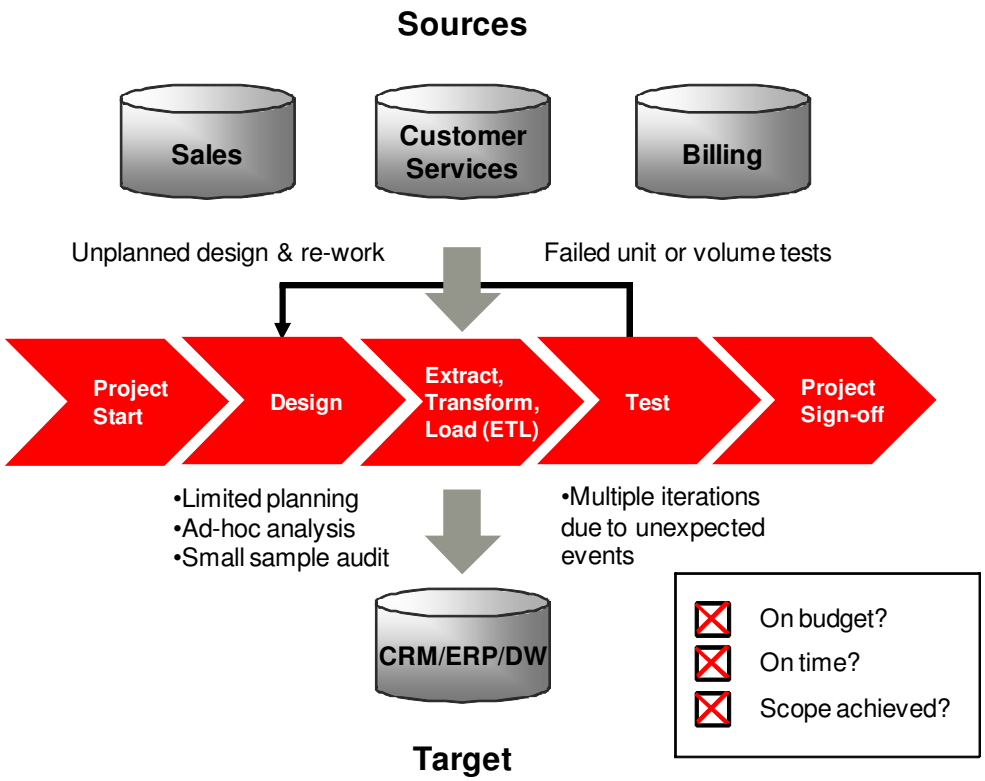

Figure 1. A typical data migration project

It is usually assumed that the migration team knows the existing systems and has schemas for the target application and that programming work can utilize spare staff capacity so that upload files can be delivered to the target system on the appropriate days.

The next section covers some pitfalls that can be avoided.

Mapping a Faster Route to the Unknown

In classic situations, the main priority is the safe physical transfer of data from the source(s) to the target without disruption of the business. The focus is on preparing detailed mapping specifications or rules for moving source data to the target, based on

- Ad hoc queries of the source system(s)
- Analysis of a small sample set of source data
- Little knowledge or documentation of how the source system(s) work
- Little knowledge of how the target system works—these parameters may move as modifications are made during implementation
- Insufficient access to the target system

- Much IT input and little business input

Consequently, mapping specifications and code are often metadata-driven, not content-driven. Inadequately checking the actual content of the data means that many assumptions are made, resulting in significant errors and a high rework rate.

Testing

Testing of a data migration suite is itself often a moving target. The deployment of the new system often gets squeezed by other business priorities, leaving little migration testing time. Unit testing should identify holes in what has been built, but because the unit testing is conducted on a small sample of data, the results are unlikely to be representative of the whole data set.

Load and Explode

The first true test of the migration suite's readiness for production delivery is usually the first full-volume migration test. Suddenly, the full volumes expose unexpected scenarios in the data, sparking numerous unexplainable problems. Live testing then has to be delayed until all the data can be processed through the technical data migration suite.

This *load and explode* syndrome can take the project over budget and beyond deadline, provoking an urgent search for answers: How did this happen? How can the problems be remedied? How can delays be handled? How will the additional costs be met? How can the situation be explained to stakeholders?

Metadata and Content Data

For the purposes of data migration, metadata can be defined as the information that describes the location of the source (database name, filename/table name, field/column name) and the characteristics of each column, such as its length and type (character, numeric, date). Content is what is contained within each actual occurrence of each field.

A metadata-driven migration assumes that the content reflects its description, which may not actually be the case. For example, a source described as telephone numbers may contain only a few telephone numbers or something else. The real telephone numbers may be stored elsewhere. Only content analysis, profiling, and auditing can confirm the actual content. These processes must ascertain the migration rules that should, in fact, be applied.

Unfortunately, because the source data was never thoroughly validated from the outset, the quest for answers usually means a meticulous review of the source data, giving rise to a significant number of iterations, code amendments, and then retesting—all costing extra time and money.

The Risks of Overlooking Data Content

Basing rules on examination of small source data samples and relying on metadata descriptions is a major risk that is likely to mean that

- Time and budget estimates will fall short of actual needs

- The target system will not perform effectively
- Workarounds will need to be implemented and resourced
- Remedial data cleansing work will need to be devised and resourced
- The costs of missing the deadline will include maintaining the team and incurring continued running costs of legacy systems and downtime on the target application
- The new system will be blamed, making it harder to gain user acceptance
- Management confidence will be questioned

The People Perspective

Aside from the technical factors, involving the right people is vital, for several reasons:

- Access to experts who really understand the history, structure, and meaning of the source data is likely to be limited.
- In-house expertise on how the new target application will work will be minimal.
- The target system vendor is unlikely to be closely involved with the data migration.
- Outsourcing any part of the migration will bring extra communication challenges.
- Getting the right mix of business and IT at the right levels across the project is important.
- Appropriate sponsorship from a senior business executive is essential.
- The business users who will work with the new system are unlikely to have an in-depth appreciation of the IT or data quality issues.
- Excellent communication across the team is a prerequisite for success.

Part 3: Discovering Your Data

Discovering Missing Links with Profiling and Auditing

The most effective way of delivering a data migration program is to fully understand the datasources before starting to specify migration code. This is best achieved with a complete profile and audit of all source data within the scope at an early stage, and it can deliver tangible benefits:

- With complete visibility of all source data, the team can identify and address potential problems that might have remained hidden until a later stage.
- The rules for planning, mapping, building, and testing migration code can be based on a thorough analysis of all source data rather than a small sample set.
- Decisions can be based on proven facts rather than assumptions.
- Early data validation can assist with the choice of the migration method.

- Establishing an in-depth repository of knowledge about the sources to be migrated enables organizations to deliver more-accurate specifications for transferring data faster.
- Full data auditing can reduce the cost of code amendments at the test stage by up to 80 percent.

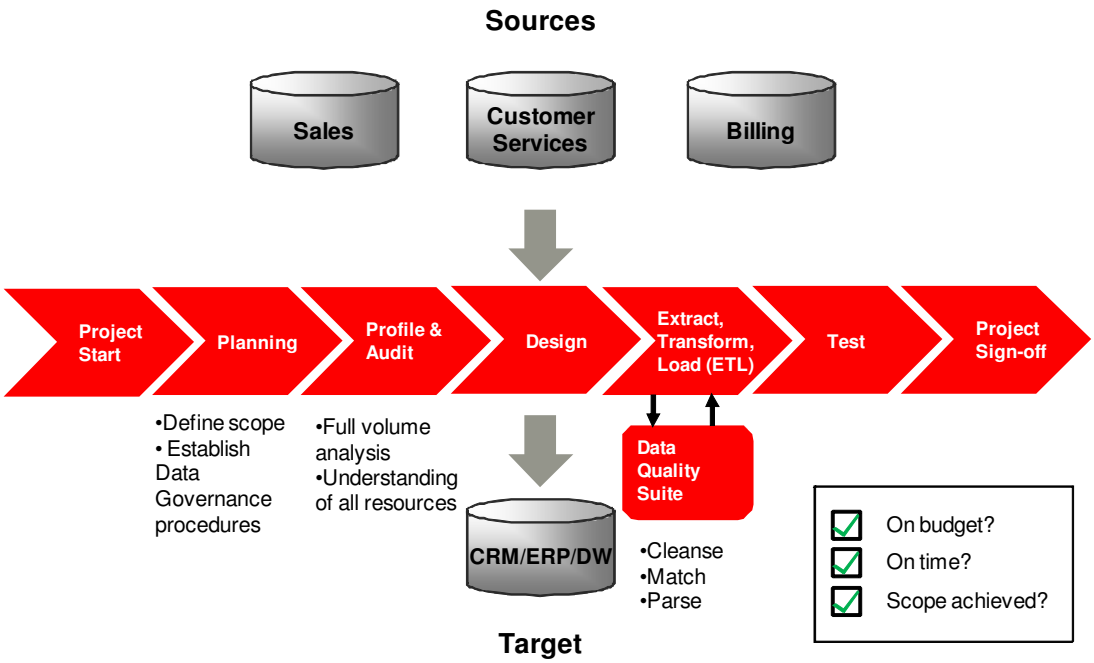

Figure 2. An effective data migration project

Redefining Data for Migration

Regardless of structure, type, or format, source data intended for migration should be validated in terms of the following key attributes:

- **Relevance.** Is it relevant to its intended purpose?
- **Accuracy.** Is it correct and objective, and can it be validated?
- **Integrity.** Does it have a coherent, logical structure?
- **Consistency.** Is it consistent and easily to understand?
- **Completeness.** Does it provide all the information required?
- **Validity.** Is it within acceptable parameters for the business?
- **Timeliness.** Is it up to date and available whenever required?
- **Accessibility.** Can it be easily accessed and exported to the target application?
- **Compliance.** Does it comply with regulatory standards?

Benefits of Data Validation

The case for data validation as part of the migration process is unequivocal. Benefits include

- Improved corporate performance
- Increased competitive advantage
- Maximized ROI from enterprise applications
- Efficient and effective business processes
- Reduced uncertainty and risk
- Lower costs of iterations and rewriting code
- Reduced delays and wasted time
- Reduced unexpected costs
- A measurable and accurate view of data
- Improved customer service
- Better accountability and ability to meet compliance targets
- Additional value for shareholders and stakeholders

Part 4: Essential Steps to Success

Phase 1: Planning

If the business has given the data migration project a broad migration scope, the first stage is to define what is achievable in terms of what the datasources will support and what is reasonable. With the scope refined, a timeline, resource plan, and budget can be put in place. Data migration projects are complex, and a key aim is to migrate the smallest amount of data required to run the target system effectively. Seldom will all source data be required, so scoping needs to be approached firmly to filter out any surplus data.

Due Diligence

The scoping refinement is integral to the success of the migration. High-level analysis of the source and target systems should be conducted in close consultation with the business users who will be directly affected by the data migration process and who have in-depth knowledge of the business issues—past, present, and future.

Budget and Timeline

Refining the scope makes it easier to plan how the project will be resourced and to secure agreement from the business. Estimates should include all time and material costs of auditing and profiling data,

developing mapping specifications, writing migration code, building data transformation and cleansing rules, and loading and testing data. A typical project is managed over six months to two years. Realistic deadlines must be set in line with external dependencies and include appropriate levels of contingency.

Data Governance

All corporate data ultimately belongs to business users, with technical support from the IT department. Therefore, migrations require a data governance board led by business users committed to the success of the new system and empowered to make decisions and drive actions.

Phase 2: Understanding the Data

Top-Down, Target-Driven Priorities

By applying a top-down, target-driven rationale, you can prioritize scoping decisions in line with the value to the organization, using criteria such as region, line of business, and product type. Refining data outward from the core saves time and effort.

The core source tables/files having been identified, profiling and auditing tools are now used to look at the data content of all potential sources to understand the data and identify what needs to be migrated. This stage is for detecting possible conflicts and drilling down to a detailed level to resolve any issues and inconsistencies.

Key benefits of profiling and auditing are that they enable you to

- Create a single repository for all analysis, regardless of source system
- Gain clear visibility into and access to all data problems, with the ability to investigate anomalies to any required depth
- Identify unknown data issues by making it easy for nontechnical staff to find the answers to questions they didn't know they needed to ask
- View any inconsistency across the full scope of the data to be migrated and assess its impact on the whole migration project
- Establish a single way of conducting analysis across the project
- Remove dependence on technical source system owners and their time
- Use a simple, business-friendly interface to review issues
- Ask questions about technical and business inconsistencies through the same user interface

Can SQL Deliver the Answer?

The primary limitation of using SQL as a profiling tool is that it answers only the exact questions asked of it and does not look beyond them. This places the delivery of any analysis request in the hands of a skilled SQL technician, creating further delays and resource issues.

A robust profiling and auditing tool gives non-SQL specialists the ability to discover what they need to know, including answers to questions they didn't even know they needed to ask.

Cost/Benefit Analysis

Every single datatype marked for migration will have a cost and a benefit. The more data to be migrated, the greater the project cost, so stringent decisions will need to be made about the level of source information to be included in or excluded from the migration suite. Historical data, for example, can greatly inflate project costs.

The key benefits of full-volume analysis include the following:

- Enables potential issues to be identified before any mapping specifications have been defined or code written
- Gives the best chance of identifying anomalies that wouldn't otherwise be revealed until full-volume testing
- Ensures that mapping specifications are based on full rather than sampled knowledge, resulting in far fewer iterations of the build and test phase
- Enables business rules defined in the target to be tested for effectiveness against all sources

Phase 3: Designing and Building

Developing the Mapping Specifications

The results of the data audit are applied to the agreed scope to develop a series of rules for transferring all designated source data and ensuring that it is accurately manipulated to fit the target. Before any actual code is written, the mapping specifications and supporting documentation must be clearly understood and signed off by the business.

Small Is Beautiful

Data migration projects run more efficiently when segmented into increments. Few projects require a big bang approach, and source data can be audited, mapped, tested, and transferred in staggered phases. This makes it easier to stick to budgets and deadlines, conduct regular progress reviews, adjust the rules, and deliver better results.

Once the mapping specifications have been converted into migration code, they should be independently verified against the rules. This will rapidly identify errors in the test environments and simplify the process of making key decisions for going live with migrated data. Traditional extract, transform, and load (ETL) tools are powerful, but these applications are usually limited when dealing

with free text fields and more-complex requirements such as fuzzy matching. In such cases, a data quality tool with parsing and matching capabilities is required to separate and restructure the detailed content for delivery to the target. The ideal solution would be an integrated tool that encompasses transformation, cleansing, and matching functions.

Phase 4: Executing

Data is extracted from the source system, transformed, cleansed, and loaded into the target system, using the migration rules.

Phase 5: Testing

Unit, system, volume, online application, and batch application tests need to be carried out before the conversion can be signed off by the business. A key objective should be to get to the full-volume upload and online application test stage as early as possible for each unit of work—in many cases, several units of work may need to be completed to achieve this before an online test can be done. This strategy helps avoid storing up issues until too late in the development cycle, when they are more expensive to rectify. Another major risk is that data is changing in the source systems while the migration is in development. Having created a profile and audit of the sources, it is possible to rerun the audit at any point to assess any changes and take appropriate action. This should be done before all major project milestones, to facilitate continue/stop and revise decisions.

Golden Rules for Successful Data Migration

- Clearly define the scope of the project.
- Actively refine the scope of the project through targeted profiling and auditing.
- Minimize the amount of data to be migrated.
- Profile and audit all source data in the scope before writing mapping specifications.
- Define a realistic project budget and timeline, based on knowledge of data issues.
- Secure sign-off on each stage from a senior business representative.
- Prioritize with a top-down, target-driven approach.
- Aim to volume-test all data in the scope as early as possible at the unit level.
- Allow time for volume testing and issue resolution.
- Segment the project into manageable, incremental chunks.
- Keep a total focus on the business objectives and cost/benefits throughout.

Phase 6: Follow-Up and Maintenance

Once established within the IT infrastructure, data audit can be implemented at any time, on any set of data, and at any point in the data migration cycle to assess whether the project is on track and still

within its scope. The ongoing use of data quality tools can then maintain data in a consistently high-quality state of readiness for future data migrations and other requirements.

Conclusion

Whatever the reason for the data migration, its ultimate aim should be to improve corporate performance and deliver competitive advantage. To succeed, data migrations must be given the attention they deserve, rather than simply being considered part of a larger underlying project. Lacking this and without proper planning, there is a high risk that the project will go over budget, exceed the allotted time, or even fail completely. A fully integrated migration environment should address the following four key areas:

- **Understand.** Comprehensive profiling and auditing of all datasources from full-volume samples can eliminate unexpected scenarios during the migration.
- **Improve.** Poor data quality in source systems should be addressed before or during the migration process. Modern data quality software can be used to restructure, standardize, cleanse, enrich, deduplicate, and reconcile the data. Rather than involve custom code or scripts that can be understood and maintained only by the IT department, the technology should be easy to use and aimed at the analyst or business user who understands the problem to be solved.
- **Protect.** Migrated data will naturally degrade over time, until it becomes a problem again. Maintaining and improving the quality of this data is vital to increasing the value that can be derived from the information. Enterprise data needs to be protected from degradation due to errors, incompleteness, or duplication. Implementing a data quality firewall to police data feeds—in both batch and real time—is critical to maintaining the integrity and therefore the value of the application.
- **Govern.** Regularly tracking and publishing data quality metrics to a dashboard enables senior executives and business users to monitor the progress of data migration projects or data quality initiatives.

Following a structured methodology will reduce the pain of managing a complex data migration, but the correct choice of technologies will go a long way to promote a successful outcome. A range of software from different suppliers, plus a lot of technical know-how, has been necessary to successfully accomplish a data migration, but the architecture becomes difficult to manage and performance deteriorates with numerous different interfaces between applications. The ideal solution is a software tool that supports the whole data migration lifecycle, from profiling and auditing the source(s) through transformation, cleansing, and matching with the population of the target. It needs to be flexible and highly scalable, require minimal technical expertise, and be intuitive so that business and technical staffs can work collaboratively. Users should be able to implement complex business rules for data migration or data quality assurance without requiring coding skills or specialist training. If you follow the steps outlined in this white paper and adopt a single technology to manage the data migration, your project will have the best chance of success.

Successful Data Migration
November 2011

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200

oracle.com

Oracle is committed to developing practices and products that help protect the environment

Hardware and Software, Engineered to Work Together