informatica

White Paper

# Data Profiling

## Calculating Return on Investment for Data Migration and Data Integration Projects

This edition published September 2013

# Table of Contents

# Introduction

Industry experience has shown that data migration and data integration projects are prone to the same challenges and problems that are common to all IT projects. They suffer from time and budget overruns, tradeoffs between quality and deadlines, and outright project failures. Project managers face these issues on a daily basis and are often left with few options and no good solutions.

Viewed through a slightly different lens, these project challenges are quite similar to those faced by manufacturing organizations. Your IT organization is essentially an "information factory" that transforms your raw material (the data contained in applications, legacy systems, operational systems, mobile devices, and sensors, etc.) into a finished product—information— which is packaged in consolidated applications, new business intelligence deployments, Cloud-based solutions, an enterprise data warehouse, etc. Just as lean manufacturing methods have turned inefficient and problem-plagued manufacturing facilities around, the same methods should produce positive results in the "information factory."

# Lean Integration With Informatica

In all manufacturing processes, lean methods synchronize people, processes, and materials to eliminate wasted effort or resources. You can take the same approach to data quality and integration by applying the Informatica® Platform to transform organizational processes through automation and reuse and optimize performance in terms of cost, speed, and quality. By applying these lean integration principles early and often in information integration or data migration projects, your business can:

- Eliminate waste and reduce project lead time
- Increase the value of completed projects
- Drive significant, continuous improvement in project quality
- Reduce the cost of managing the information integration lifecycle

In manufacturing, understanding the quality of raw materials improves the efficiency of creating the finished product. In information integration and data migration projects, data profiling serves the same purpose. Data profiling is an essential initial step that can dramatically reduce the time it takes to plan and execute data integration projects. Before you can integrate data or use it in a data warehouse, CRM, ERP, or business analytics applications, you need a full understanding of its content, quality, and structure—not only as it relates to its original source but also in the context of what your integration or migration effort is hoping to achieve. Informatica Data Explorer helps you achieve that full understanding by quickly analyzing 100 percent of the data across multiple source systems, allowing you to accurately scope the size and complexity of your project.

A comprehensive data investigation and discovery product used by data analysts for planning complex data migration and data integration projects, Informatica Data Explorer provides a complete and accurate picture of all enterprise data through automated and process-driven data profiling. It enables data analysts to quickly discover hidden data quality issues, gaps, inconsistencies, and incompatibilities within and across data sources. The result is an accurate, baseline profile of data quality across your IT landscape. This profile is stored in a central open repository and can be used to accelerate the design and implementation of new applications, databases, and data quality programs.

## Key Benefits of Informatica Data Explorer

- Delivers accurate source system knowledge
- Improves corporate data quality and accuracy
- Enables efficient data migration, integration, and consolidation
- Helps expedite integration of multiple, disparate data sources
- Mitigates risk and reduces costly rework in data management projects
- Minimizes overruns in enterprise applications projects
- Improves productivity of data management projects
- Improves the success of data integration projects related to merger & acquisition activities
- Reduces development costs by fully understanding data content, quality, and structure

# The Critical Importance of Data Profiling

Despite careful initial planning, more than 38 percent of data migration projects are at risk of overruns in time and cost or of outright failure, according to Bloor Research. Although the increased cost or waste involved in these projects is significant—cost overruns averaged more than 30 percent of the migration budget in Bloor Research's 2011 work[1] —it is often dwarfed by the loss in business value caused by delaying or cancelling implementation of a new business system or data warehouse. As the *Harvard Business Review* explained in September 2011, the average cost overrun for IT projects was 27 percent, but that figure concealed an even more surprising one: "Graphing the projects' budget overruns reveals a "fat tail" – a large number of gigantic overages. Fully one in six of the projects we studied had a cost overrun of 200%, on average, and a schedule overrun of almost 70%."[2]

To maximize the business value of data migration projects, managers and planners must directly address the underlying causes of these overruns and failures to ensure that they complete projects not just on time but ahead of schedule.

One of the primary causes of data migration project overruns and failures is a lack of understanding of the source data prior to data movement. Bloor Research[3] suggests that data profiling software can dramatically increase the chances of project success. When Bloor Research surveyed data migration projects in 2007 and again in 2011, the differences were telling. In 2007, only 16 percent of the projects surveyed were completed on time and within budget. In 2011, that success rate had risen to 62 percent. Bloor Research speculates that the difference is the increased use of data profiling and data cleansing tools rather than hand-coded and manual efforts: "In 2007 only 10% of respondents' projects involved the use of data profiling tools. In 2011 that figure is 77%. While we cannot prove a causal link between this increase (and others) and the dramatically increased success rate of the data migration projects, these figures are highly suggestive."[4] The 2011 survey also revealed an increased use of data cleansing tools, from 11 percent in 2007 to 78 percent in 2011. This strongly suggests that the ability to better understand source data accelerates a project's time to value.

However, given the gains achieved by participants in the Bloor Research survey, a deeper look into the reasons why projects fail shows that most migration projects and data integration initiatives rely on external information to provide an understanding of the data. Much of this information—documentation, source programs, existing data models, and staff experience—is often outdated, incorrect, or missing. If the information is invalid, then it may take many iterations to develop new information and validate that it is indeed correct (i.e., that it actually represents the source data). In this scenario, as much as 50 percent of a total project's labor budget may be wasted on manual, outdated data analysis and diagnosis techniques, while poor understanding of the source data jeopardizes the project's overall success.

[1] "Data Migration, 2011", Philip Howard, Bloor Research, August 2011

[2] The Harvard Business Review Magazine, "why Your IT Project May Be Riskier Than you Think," Bent Flyyvbjerg and Alexander Budzier, September 2011

[3] "Data Migration, 2011," Philip Howard, Bloor Research, August 2011

[4] Ibid.

## Why Is Understanding Data Difficult?

- Manual data profiling is tedious, slow, labor-intensive, and error-prone
- Metadata documentation may be missing, incomplete, or badly out of date
- Unstructured or semistructured data in free-form text fields cannot be categorized accurately
- Source code for legacy systems may no longer be found
- Relationships between data elements are not always obvious
- Assumed relationships and dependencies may be wrong
- Databases are not static; over time they may be corrupted

Tackling the analysis of multiple data sources takes time—typically three to five hours per attribute, with more challenging cases taking up to 10 hours or more per attribute. An extreme example is provided by a major insurance provider where two months of effort resulted in the completed analysis of six attributes. An average of three to five hours per attribute is an accepted industry standard and consistent with a similar estimate on business rule extraction.[5]

There is, however, a more direct and efficient way to understand source data: by using Informatica Data Explorer to analyze it rather than relying on inaccurate or out-of-date metadata and documentation. By automating the discovery and identification of data and metadata, companies can readily highlight inconsistencies, redundancies, and inaccuracies across all their corporate databases. This deeper, more rapid understanding of the data helps speed the data migration project or data integration initiative to completion.

Informatica Data Explorer supports a six-stage process for examining source data in detail to provide a thorough understanding of its content, structure, quality, and integrity and then generates source-to-target attribute maps based on this understanding. The data profiling and mapping process rapidly produces an inferred data model and a set of source-to-target transformation maps and eliminates the unpredictable and time-consuming cycles of a manual approach.

With a clear understanding of the content, quality, and structure of the data, IT can clearly scope development efforts, set appropriate budgets, and mitigate the risks of downstream delays. Ultimately, the solution can help lower project cost, reduce project risk, accelerate business value, and deliver better results by:

- Reducing direct project costs such as labor and resources
- Increasing project value through early project completion
- Reducing unexpected costs and delayed benefits associated with overruns or cancelations
- Increasing data quality of the migrated or converted data without extending deadlines

[5] Terry Moriarty, "Getting Your Business Rules Automatically," Database Programming and Design, October 1997:74

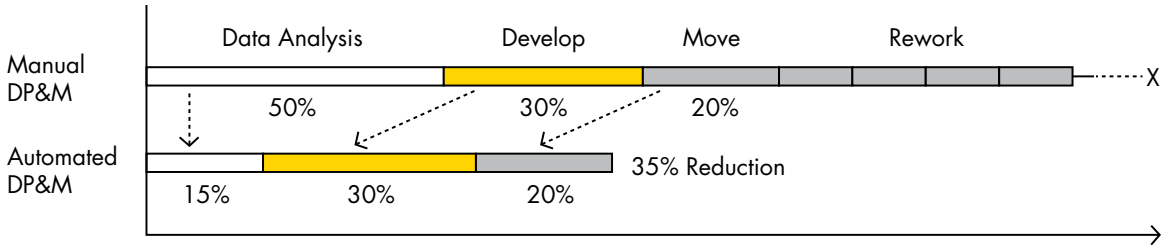


Figure 1: The graph above shows how using Informatica's Data Explorer immediately reduces timelines by cutting the time needed for data analysis. Even more time savings result from less rework and delays in later stages.

The following examples show how each of these four benefits affect total return on investment (ROI) for three projects of different sizes. While the examples are drawn from real-world Informatica deployments, they are composites rather than the experiences of a single customer.

## Example 1: Data Warehousing

An insurance provider moved data from a legacy application to a data warehouse and several data marts. The initial evaluation established that a full analysis of the legacy data could take a year or more to complete and might not produce useful results. The project was put on hold. Subsequently, the company deployed Informatica Data Explorer, enabling it to complete the data profiling and mapping process in six weeks and the entire project in ten weeks. An in-house project team did all the work; the fully loaded cost of $95 per hour is typical for the area.

## Example 2: ERP Implementation

In order to consolidate multiple legacy VSAM applications on a single ERP system, a manufacturing company needed to accurately and completely migrate human resource data, including benefits history, plus an large amount of manufacturing history data. The project was completed by a combination of in-house staff and external consultants. Most of the work was done by in-house staff, which is reflected in the average cost per hour of $75.

## Example 3: Data Consolidation

A major international transportation company was consolidating data from a recent acquisition. The estimate of four hours of analysis per attribute was believed to be low but because of the size of the project it could not be fully tested. Informatica Data Explorer was used to verify the accuracy of the initial data analysis and then to develop accurate data profiles and data maps. The project team included in-house staff plus consultants from a major systems integrator; the fully loaded cost of $100 per hour takes this staffing mix into account.

# ROI Scenarios

For these three example projects, ROI is demonstrated in four separate scenarios:

- Reducing direct project costs such as labor and resources
- Increasing project value through early project completion
- Reducing unexpected costs and delayed benefits associated with overruns or cancelations
- Increasing data quality of the migrated or converted data without extending deadlines

Although each scenario is illustrated for each example project, most projects will focus on a subset of these scenarios when establishing the expected ROI associated with the use of a data profiling solution.

These ROI scenarios use the term "full-time equivalent" or "FTE" in all staffing calculations. An FTE may be fractional, representing a person working part time on a project or may consist of various staff members contributing different skill sets as needed by the project.

## Reduction in Direct Project Costs

A properly planned project includes an allotment of resources for profiling data sources; designing the target systems and mapping specifications; and executing extraction, cleansing, and transformation processes. Failure to be thorough in the profiling and mapping phases will increase costs due to undiscovered problems that will eventually surface and need to be resolved— typically during the testing phase of the project when development resources may have already been redeployed. This causes a ripple effect in terms of disruption to the business; not only is the current project delayed for rework, but other projects may be impacted because staff must be reallocated. While manual data profiling and mapping takes an estimated three to five hours per attribute, Informatica customers report that they require at most 15 minutes per attribute—a significant savings.

Table 1 shows the difference in direct project costs attributable to the data profiling and mapping (DP&M) phase for the three example projects, comparing a manual approach versus an approach that utilizes Informatica Data Explorer. Historical project data shows that for the average $1 million data migration project, manual data analysis consumes 50 percent of the total project budget. Time, cost, and resource constraints in larger projects prevent many of them from undertaking a 100 percent analysis of the source data. In the scenarios in Table 1, each of the projects attempts only a 75 percent manual analysis of the source data, even though analyzing less than 100 percent of the data will result in an increase in project risk and a decrease in data quality. The scenario also doubles the total effort required for data profiling and mapping with Informatica Data Explorer as a contingency to allow for implementation, training, and startup time. Subsequent projects with an experienced team would not require this increase in the total effort.

Table 1: Reduction in Direct Project Costs

| | Formula | Data Warehouse | ERP Implementation | Data Consolidation |
|---|---|---|---|---|
| A. Initial project budget | Inputvalue | 1200000 | 3000000 | 12000000 |
| B. Fully loaded cost per FTE per hour | Input value | $95 | $75 | $100 |
| C. Number of attributes | Input value | 1750 | 4675 | 16200 |
| Manual analysis | | | | |
| D. Hours per attribute | Input value | 5.0 | 4.0 | 3.0 |
| E. Total house (75% analysis)* | C x D x .75 | 6,563 | 14,025 | 36,450 |
| F. Cost of manual DP&M ** | B x E | $623,438 | $1,051,875 | $3,645,000 |
| G. DP&M** as a percent of total budget | F / A | 52% | 35% | 30% |
| With Informatica Data Explorer | | | | |
| H. Hours per attribute | Input Value | 0.25 | 0.25 | 0.25 |
| I. Total house (100% analysis)*** | C x H x 2 | 875 | 2337.5 | 8100 |
| J. Cost of DP&M** using Informatica Data Explorer | B x I | $83,125 | $175,313 | $810,000 |
| K. Project budget using Informatica Data Explorer | A - F + J | $659,688 | $2,123,438 | $9,165,000 |
| L. DP&M** as a percent of revised budget | J / K | 13% | 8% | 9% |
| Redeuction in direct project cost | | | | |
| M. Net reduction | F - J | $540,313 | $876,563 | $2,835,000 |

* With manual approach only 75% analysis possible

** Data Profling & Mapping

*** Adjusted by factor of 2 for training and implmentation for the first project

## Business Value of Early Project Completion

Informatica Data Explorer also enables significant ROI by realizing business value through early project completion. This value is easy to compute if the business value of the project has been estimated. In Table 2, the business value per month of each of the three example projects has been multiplied by the number of months saved to show the total business value of using Informatica Data Explorer to accelerate project completion. Each month is assumed to have 175 working hours.

### Table 2: Business Value of Early Project Completion

| | Formula | Data Warehouse | ERP Implementation | Data Consolidation |
|---|---|---|---|---|
| A. Number of FTEs | Input value | 3 | 5 | 10 |
| B. Project business value per month | Input value | $60,000 | $90,000 | $200,000 |
| Manual Analysis | | | | |
| C. Total hours (months) | Table 1. E | 6,563 | 14,025 | 36,450 |
| D. Elapsed time (months)* | C / A / 175 | 12.5 | 16 | 21 |
| With Informatica Data Explorer | | | | |
| E. Total hours (100% analysis) | Table 1. I | 875 | 2338 | 8100 |
| F. Elapsed time (months) | E / A / 175 | 1.7 | 2.7 | 4.6 |
| Value of early project completion | | | | |
| G. Elapsed time savings (months) | D - F | 10.8 | 13.4 | 16.2 |
| H. Net value | B x G | $650,000 | $1,202,143 | $3,240,000 |

* 175 working hours per month

## Reduction in Project Overrun Costs

The traditional approach to data profiling and mapping is cumbersome and repetitive, requiring manual analysis of metadata, copybooks, documentation, and, in some cases, physical data. Based on this analysis and many assumptions, transformation specifications are produced and coded, and the data is extracted, transformed, and loaded into the new system. In more than 80 percent of data migration projects, this process doesn't work correctly the first time—a result often referred to as "code, load, and explode." The whole process is repeated over and over until the project either succeeds or is canceled.

Unfortunately, there is no way to predict how many iterations will be required and when the process will be completed. One of the key benefits of using Informatica Data Explorer is that the data profiling and mapping process is predictable. Furthermore, since the transformation specifications generated are 100 percent supported by the data, the project has a much greater probability of completing successfully the first time, without endless rework.

According to the Bloor Research study, in the projects that overran or were aborted, "the overrun nearly always involved both time and costs...In both cases the average overrun is approximately 30%, slightly more in the case of costs, and slightly less with respect to time." Therefore, Table 3 uses 30 percent to calculate the anticipated overrun cost for the three example projects in the absence of an automated data profiling and mapping product such as Informatica Data Explorer. This scenario also assumes that these overrun costs will be reduced by 85 percent when Informatica Data Explorer is used, a factor that has been validated in many data profiling and mapping projects. Although some projects might not overrun at all, good business planning requires taking this into consideration in the overall plan.

### Table 3: Reduction in Project Overrun Costs

| | Formula | Data Warehouse | ERP Implementation | Data Consolidation |
|---|---|---|---|---|
| A. Total planned cost for manual DP&M* | Table 1. F | $623,485 | $1,051,875 | $3,645,000 |
| B. Allowance for overrun | A x .30 | $187,046 | $315,563 | $1,093,500 |
| Reduction in project overrun cost | | | | |
| C. Net reduction | B x .85 | $158,989 | $268,228 | $929,475 |

* - Data Profiling and Mapping

For simplicity, this calculation does not consider additional project-specific costs such as delays in data analysis that can postpone a larger project or the need to maintain legacy systems longer than planned.

## Value of Improved Data Quality

IT projects often compromise quality for the sake of meeting a hard deadline. The wealth of information Informatica Data Explorer provides during the migration process can streamline subsequent data integration development efforts and ensure quality even in the face of a deadline. For example, the data profiles built in the initial stage of migration and integration efforts can be run from within Informatica PowerCenter when integration mappings are being produced. Additionally, data profiling can be run midstream during mapping creation to allow developers to double-check the accuracy of the integration routines they are developing, ensure accurate mappings for data integration tasks, and avoid costly rework efforts during the testing phase of deployment. While it is easy to show how Informatica Data Explorer allows project managers to meet or beat their project deadlines while ensuring high quality, accurate data in the target database, it is much more difficult to quantify this benefit. It is the business user who must put a value on accurate data. The conservative approach used in Table 4 assumes that improved data quality is worth at least 20 percent of the project business value per month for 12 months. Bear in mind that recovering from and correcting poor data quality would cost at least this much over the same timeframe.

Table 4: Value of Improved Data Quality

| | Formula | Data Warehouse | ERP Implementation | Data Consolidation |
|---|---|---|---|---|
| A. Project business value per month | Table 2. B | $60,000 | $90,000 | $200,000 |
| Value of improved data quality | | | | |
| B. Net value | A x 0.2 x 12 | $144,000 | $216,000 | $480,000 |

# Cost of Informatica Data Explorer

Table 5 shows the total cost of acquiring and implementing Informatica Data Explorer for the example projects. This cost is factored into the ROI calculations. These figures are based on Informatica list prices as of August 2013, are subject to change without notice, and do not represent a price quotation or an offer to sell.

Table 5: Cost of Informatica Data Explorer

| | Formula | Data Warehouse | ERP Implementation | Data Consolidation |
|---|---|---|---|---|
| A. Informatica Data Explorer | Input value | $160,000 | $240,000 | $260,000 |
| B. Source importers (VSAM or Relational) | Input value | $ - | $20,000 | $20,000 |
| C. Maintenance for one year | Input value | $32,000 | $48,000 | $52,000 |
| D. Training and consulting | Input value | $25,000 | $60,000 | $100,000 |
| Cost of Informatica Data Explorer | | | | |
| E. Total cost | Sum A to D | $217,000 | $368,000 | $432,000 |

## Total Return on Investment

One way of looking at the total ROI for each of the three example projects is shown in Table 6. The total ROI is the sum of the reduction in direct project costs, the reduction in project overrun costs, the business value of early completion, and the value of improved data quality. In all cases, the cost of Informatica Data Explorer is substantially less than the benefits realized.

Table 6: Total ROI

| | Formula | Data Warehouse | ERP Implementation | Data Consolidation |
|---|---|---|---|---|
| A. Reduction in direct project cost | Table 1, M | $540,360 | $876,525 | $2,835,000 |
| B. Value of early project completion | Table 2, H | $648,000 | $1,197,000 | $3,280,000 |
| C. Reduction in project overrun cost | Table 3, C | $158,989 | $268,228 | $929,475 |
| D. Value of improved data quality | Table 4, B | $25,000 | $60,000 | $100,000 |
| Total value | | | | |
| E. Total value | Sum A to D | $1,372,349 | $2,401,753 | $7,144,475 |
| F. Cost of Informatica Data Explorer | Table 5, F | $217,000 | $368,000 | $432,000 |
| ROI as a percent of investment in IDE | 100 (E-F) / F | 532% | 553% | 1554% |

# Conclusion

Businesses are becoming more data-driven in their decision making. In response, a comprehensive data integration platform enabling a more streamlined, automated approach to data integration and data migration projects is becoming a critical business requirement. Organizations need to look holistically at their data management practices and see them not as a set of discrete projects and teams but as an integrated, lean integration factory capable of meeting the demands of today and tomorrow efficiently, productively, and accurately.

Companies using Informatica Data Explorer in the early phases of their data migration and data integration projects typically achieve significant ROI by reducing the amount of effort to complete the project and, more importantly, delivering the project sooner than would otherwise have been possible.

- Managers can scope development efforts and budgets more accurately by achieving greater visibility across all data sources.
- Project teams reduce the risks of project overruns by understanding data content, quality, and structure before they begin to move and integrate data.
- Developers achieve greater productivity and greater understanding of business needs through a shared metadata repository that helps them collaborate more effectively with the business.

Informatica Data Explorer also delivers incremental benefits including lowering project costs, accelerating business value, reducing the risk of project overruns and failures, and increasing the quality of project results—all of which combine to back up research proving the value of data profiling tools.

## About Informatica

Informatica Corporation (Nasdaq:INFA) is the world's number one independent provider of data integration software. Organizations around the world rely on Informatica to realize their information potential and drive top business imperatives. Informatica Vibe, the industry's first and only embeddable virtual data machine (VDM), powers the unique "Map Once. Deploy Anywhere." capabilities of the Informatica Platform. Worldwide, over 5,000 enterprises depend on Informatica to fully leverage their information assets from devices to mobile to social to big data residing on-premise, in the Cloud and across social networks.


Put potential to work.™

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 informatica.com linkedin.com/company/informatica twitter.com/InformaticaCorp



IN09_0913_06744