Data Migration Issues and Keys to Success

A.Mahendiran[1]

ABSTRACT

Data migration is the process of transferring data between storage types, formats, or computer systems. Data migration is actually the translation of data from one format to another format or from one storage device to another storage device. Usually it is performed programmatically to achieve an automated migration. Data migration is necessary when a company upgrades its database or system software, either from one version to another or from one program to an entirely different program. Unfortunately, most data migration projects don't go as smoothly as anticipated. More than 80 percent of all IT projects either overran or failed, resulting in unexpected costs. One of the primary reasons for this extraordinary failure rate is improper *Migration Analysis,* the lack of knowledge about the source data early on in these projects. Using conventional approaches to data profiling and migration can create as many problems as they resolve - data not loading properly, poor quality data and compounded inaccuracies, time and cost overruns and, in extreme cases, late- stage project cancellations. Comparing and synchronizing models of the source and target databases is one way to migrate changes between database environments. Another way to migrate changes is to apply the deployment script that was used to change the source database to a target model of the target database. Issues related to the transformation of data from the old databases to new applications and the keys to successful data migration are discussed

[1]Department of MCA , School of Computing , SASTRA University, Tanjore – 613 402. Email : mahi_thrc@ yahoo.com

Keywords: Migration Analysis, Data Profiling, Data Mapping, Normalization, Column Profiling, Dependency Profiling

1. INTRODUCTION

Introduction to Data migration

Data migration is a set of activities that moves data from one or more legacy systems to a new application [1] the purpose of data migration is to preserve core business knowledge and make it accessible from the new application.

Data migration typically involves planning and scoping the project, extracting data from the source application, cleansing to repair corrupt data or invalid records, removing duplicates, transforming the source data to new data [2].

The migration workflow has to meet the following demands

- Minimize risk.
- Stay on budget.
- Deliver in due time.
- Keep downtime to a minimum.
- In case of failure, be able to return to the source system.

2. PHASES OF DATA MIGRATION

Data migration typically has four phases:

1. analysis of source data
2. extraction and transformation of data
3. validation and repair of data, and
4. use of data in the new program.

During each phase, the data migration software works its electronic magic, performing the necessary machinations before moving the data on through the process. Perhaps the most sensitive phase is validation and repair. In this phase, data is evaluated for potential problems, which are flagged and identified to the user. Unresolvable problems can be identified at this stage, along with untranslatable data, so they can be set aside and not gum up the whole data migration works.

3. Data Migration Strategies

Data profiling and mapping consist of six sequential steps, three for data profiling and three for data mapping, with each step building on the information produced in the previous steps. The resulting transformation maps, in turn, can be used in conjunction with third-party data migration tools to extract, scrub, transform and load the data from the old system to the new system.

Data sources are profiled in three dimensions: down columns *(column profiling)* ; across rows *(dependency profiling)*; and across tables *(redundancy profiling)*[3].

Column Profiling

Column profiling analyzes the values in each column or field of source data, inferring detailed characteristics for each column, including data type and size, range of values, frequency and distribution of values, cardinality and null and uniqueness characteristics. This step allows analysts to detect and analyze data content quality problems and evaluate discrepancies between the inferred, true meta data and the documented meta data.[6] [7].

Dependency Profiling

Dependency profiling analyzes data across rows - comparing values in every column with values in every other column - and infers all dependency relationships that exist between attributes within each table. This process cannot be accomplished manually. Dependency profiling identifies primary keys and whether or not expected dependencies (e.g., those imposed by a new application) are supported by the data. It also identifies "gray-area dependencies" - those that are true most of the time, but not all of the time, and are usually an indication of a data quality problem. [6][7]

Redundancy Profiling

Redundancy profiling compares data between tables of the same or different data sources, determining which columns contain overlapping or identical sets of values. It looks for repeating patterns among an organization's "islands of information" - billing systems, sales force automation systems, post-sales support systems, etc. Redundancy profiling identifies attributes containing the same information but with different names (synonyms) and attributes that have the same name but different business meaning (homonyms). It also helps determine which columns are redundant and can be eliminated and which are necessary to connect information between tables. Redundancy profiling eliminates processing overhead and reduces the probability of error in the target database. As with dependency profiling, this process cannot be accomplished manually.

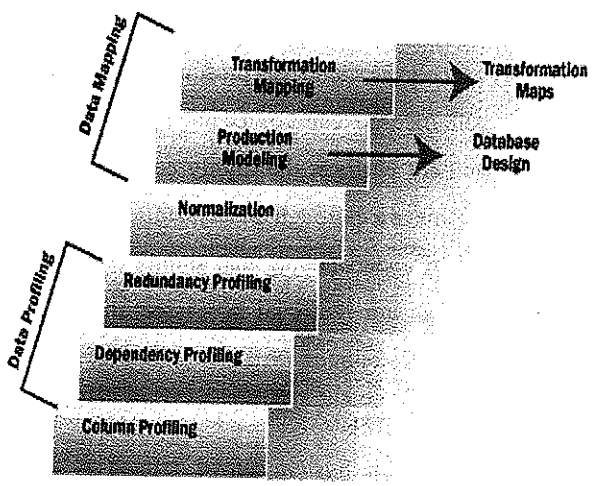

Figure 1: Key Steps In Data Profiling and Mapping

Once the data profiling process is finished, the profile results can be used to complete the remaining three data mapping steps of a migration project: normalization, model enhancement and transformation mapping.

Normalization

By building a fully normalized relational model based on and fully supported by the consolidation of all the data, the data model will not fail.

Model Enhancement

This process involves modifying the normalized model by adding structures to support new requirements or by adding indexes and denormalizing the structures to enhance performance.

Transformation Mapping

Once the data model modifications are complete, a set of transformation maps can be created to show the relationships between columns in the source files and tables in the enhanced model, including attribute-to-attribute flows. Ideally, these transformation maps facilitate the capture of scrubbing and transformation requirements and provide essential information to the programmers creating conversion routines to move data from the source to the target database. [7][8][9].

Developing an accurate profile of existing data sources is the essential first step in any successful data migration project. By executing a sound data profiling and mapping strategy, small, focused teams of technical and business users can quickly perform the highly complex tasks necessary to achieve a thorough understanding of source data - a level of understanding that simply cannot be achieved through conventional processes and semi-automated query techniques.

4. Reasons for failure

A Legacy Information System can be defined as "any information system that significantly resists modification and evolution" [4]. Legacy information systems typically form the backbone of the information flow within an organisation and are the main vehicle for consolidating information about its business. A failure of these systems may have a serious business impact [5].

At any given time, according to industry analyst estimates, roughly two-thirds of the projects are data conversion projects including migration from legacy systems to packaged applications, data consolidations, data quality improvements and creation of data warehouses and data marts. These projects are driven by internal financial pressures, increasing competition, ongoing deregulation, industry consolidation due to mergers and acquisitions.

Unfortunately, most data migration projects don't go as smoothly as anticipated. According to The Standish Group, 74 percent of all IT projects either overran or failed, resulting in unexpected costs. The result of the recent survey statistics says that 88 percent of the data migration projects will either overrun or fail.

One of the primary reasons for this extraordinary failure rate is the lack of a thorough understanding of the source data early on in these projects. Conventional approaches to data profiling and migration can create nearly as many problems as they resolve - data not loading properly, poor quality data and compounded inaccuracies, time and cost overruns and, in extreme cases, late- stage project cancellations.

1. Understanding Your Data

Before undertaking large scale legacy-to-x application data migrations, data analysts need to learn as much as possible about the data they plan to move. Considering the magnitude of an organization's data, the task of the data analyst to obtain this knowledge is overwhelming,

to say the least. IT organizations can begin their data analysis by implementing a two-step process: ***data profiling*** and ***mapping***. Data profiling involves studying the source data thoroughly to understand its content, structure, quality and integrity. Once the data has been profiled, an accurate set of mapping specifications can be developed based on this profile - a process called data mapping. The combination of data profiling and mapping comprises the essential first step in any successful data migration project and should be completed prior to attempting to extract, scrub, transform and load the data into the target database.

2. Conventional Techniques in Data Migration: Problems and Pitfalls

The conventional approach to data profiling and mapping starts with a large team of people (data and business analysts, data administrators, database administrators, system designers, subject matter experts, etc.). These people meet in a series of joint application development (JAD) sessions and attempt to extract useful information about the content and structure of the legacy data sources by examining outdated documentation, COBOL copy books, inaccurate meta data and, in some cases, the physical data itself. Typically, this is a very labor-intensive process supplemented, in some cases, by semi-automated query techniques. Profiling legacy data in this way is extremely complex, time- intensive and error-prone. Once the process is complete, only a limited understanding of the source data is achieved.

At that point, according to the project flow chart, the data analyst moves on to the mapping phase. However, since the source data is so poorly understood and inferences about it are largely based on assumptions rather than facts, this phase typically results in an inaccurate data model and set of mapping specifications. Based on this information, the data is extracted, scrubbed, transformed and loaded into the new database.

Not surprisingly, in almost all cases, the new system doesn't work correctly the first time. Then the rework process begins: redesigning, recoding, reloading and retesting. At best, the project incurs significant time and cost overruns. At worst, faced with runaway costs and no clear end in sight, senior management cancels the project, preferring to live with an inefficient but partially functional information system rather than incur the ongoing costs of an "endless" data migration project.

5. Keys To Successful Data Migration

In order to improve the data migration success rate, organizations need to replace a purely tactical approach with a strategic one. This means addressing data migration, not just as a one time move, but as an ongoing process of making the data work no matter what changes occur in the company's systems. A data migration strategy should address the challenges of identifying source data, interacting with continuously changing targets, meeting data quality requirements, creating appropriate project methodologies, and developing general migration expertise.

To create an effective strategy companies need to dedicate substantial up-front effort to understanding sources. Simply profiling and sampling the data are not comprehensive enough required. Instead, organizations need to devote a significant amount of time to making a full and accurate identification of source data, including:

- Does it exist?
- Where is it?
- Can disparate data be related?
- What is the focus of each source?
- What about standardization?
- Is sufficient detail available?

• What about unstructured data?

• Is data orphaned anywhere?

Once the source data has been properly identified, the strategy team can then test that the data will support the required functionality of the target application. The team should begin by identifying quality problems in the source data such as syntax and semantic errors, format problems, and integrity issues and plan how to correct outstanding issues. The team should also identify and prepare to correct problems accessing source data.

Informatica Research Reveals Keys To Data Migration Success

Informatica Corporation (NASDAQ: INFA), a leading provider of data integration software announced findings from the research study

6. Data migration Tools

The tasks connected to a migration workflow are diverse and complicated. Doing all of them manually requires plenty of time and a migration team highly experienced in the source as well as the target system. As both factors are not available in most situations, migration tools may come handy and should be considered to ease the migration workload.

Some advantages: [10]

- ★ Reduce the migration time by 50% to 90%.
- ★ Automate repetitive tasks.
- ★ Re-use scripts from other migrations.
- ★ Automated validation of pre and post states.
- ★ Documentation of all migration steps.

We have several Data Migration Tools available in the market as open source tool, some of the popular tools are listed here[10].

1. **SwisSQL Data Migration Tool:**
   Tool to migrate tables (along with indexes, constraints etc) and data from/to all popular databases. Supports Oracle, SQL Server, MySQL, DB2, PostgreSQL and Sybase. Supports migration of data from files - Excel and CSV (Comma Separated Value) into databases
2. **Clover.ETL :** open source ETL tool, engine-based, embeddable, with pipeline parallelism.
3. **Data Migration Toolkit (DMT) :** freely available, GUI-based Java utility for migrating files and database data.
4. **ETL Integrator :** JBI-enabled open source ETL tool for data migration in SOA environments.
5. **Scriptella :** open source ETL and script execution tool used typically for database migration.
6. **Talend :** open source code generation and script execution application for data transformation.
7. **Monarca Enterprise 2.0, database migration tool:**
8. **Endian Software has released Monarca Enterprise 2.0,** a commercial database migration tool. With Monarca you can import, integrate, transform, validate and migrate data from any-to-any existing databases.Monarca takes advantage of JDBC to allow migrating data from or to any database that is JDBC compatible, including ODBC databases. You can migrate from FoxPro to Oracle, from Microsoft Access to IBM DB2;, from Informix to PostgreSQL, from Microsoft Excel to MySQL, from Sybase to Oracle, etc. Just plug-in your preferred JDBC Driver for your source and target databases and let Monarca do the entire job for you.
9. **SQL Script Builder :** Multiple Platform Database Migration Tool SQL Script Builder is powerful software that can create a database migration sql script (or dump file) or database files from any ODBC supported database. The script produced will migrate the database (multiple tables selection) or

only one table and the records. Scripts are available in 5 output formats ; MySql, MS SQL, Oracle, Pervasive and PostgreSQL and files comes in Access mdb, Excel csv, MS xml. SQL Script Builder is very simple to use, you just have to choose the database and the table from the list. SQL Script Builder scripts can be used on your DBMS (*database management system*) or uploaded on a server.

10. **IBM Informix Dynamic Server,** Version 11.50 Dynamic Server provides tools, utilities, and SQL statements that you can use to move data from one IBM® Informix® database to another.

7. Tips for Choosing data migration tools

The data migration problem growing, the temptation can be to keep migrating data the same old way. However, native operating system utilities like "move" and "copy" are rarely sufficient when migrating hundreds of gigabytes or terabytes of data that may require constant monitoring. So here are some tips on how to pick a data migration tool [13].

First, determine how the product migrates the data, since not every product does it the same way. Some copy the data from the source to the target volume and then mirror writes on both volumes until the cut-over occurs. Others move the data to the target volume deleting the data on the source volume making it more difficult to fall back.

Second, give preference to data migration tools that are host-based. They are almost always storage agnostic, can be configured to migrate data in the background and give users more choices when picking storage vendors in the future.

Third, give preference to block-level migration tools but keep a file-level tool in your back pocket. Since storage is often over-allocated to servers, a data migration provides an excellent time to correct that situation. However, if a Windows server is only using 200GB of a 500GB volume, a block-level migration tool only recognizes the 500GB volume and can only move those blocks to another volume of the same size or larger. A file-level data migration tool allows users to move just the 200GB of files to a smaller volume.

With most shops not looking to hire people to simply migrate data, having the right data migration tools can go a long way to maintaining application availability, not living in the office while data migrations occur and keeping your environment reasonably well managed.

8. Preventive measures and actions to achieve Data Migration Success

- Implement data governance across the enterprise so that the location, nature and condition of enterprise data are always properly understood. This includes extending data governance to any acquired organization prior to undertaking a data migration.
- Implement data quality initiatives so that data is maintained at a high level of migration-readiness, in addition to the numerous other benefits of ensured data quality.
- Profile and analyze all data sources in advance using a proper tool in order to fully understand the scope of data issues that will be encountered including how they may impact project timelines and costs.
- Ensure full familiarity with, and training in the use of, any tools, prior to commencing the project.

9. Conclusion

Data migration is not a painful process, but it can be. To ensure that it does not, The Organization have to plan it well. Well-organized IT Company will take data migration in its stride and plan it. By following proper data profiling , mapping, and using the right data migration tool, companies can take their data migration projects to completion successfully and extensive design rework and late-stage project cancellations can also be avoided. Data profiling and mapping, if done correctly, can dramatically lower project risk, enabling valuable

resources to be redirected to other, more fruitful projects. So Planning and using the right set of methodologies is key to a successful migration.

References

[1] Bisbal. J, Lawless. D, Wu. B and Grimson. J, *"Legacy Information System Migration: A Brief Review of Problems"*, Solutions and Research Issue, IEEE Software, 6(5), 1999.

[2] Youn. C, Ku. C.S, *"Data migration"* in Systems, Man and Cybernetics, IEEE International Conference on Vol. 2, Issue. 18-21, PP .1255 - 1258, Oct. 1992.

[3] John. B. Shepherd, *"Data Migration Strategies"*, DM Review Magazine, June 1999.

[4] M. Brodie, M. Stonebraker, *"Migrating Legacy Systems: Gateways, Interfaces and the Incremental Approach"*, Morgan Kaufmann Publishers, Inc. USA, 1995.

[5] K. Bennett, *"Legacy Systems: Coping with Success"*, IEEE Software, 12(1), PP.19-23, Jan 1995.

[6] Jean Henrard and Jean-Luc Hainaut, *"Data dependency elicitation in database reverse engineering"*, In: Proc. of the 5th European Conference on Software Maintenance and Reengineering (CSMR 2001), collection IEEE Computer society, PP. 11–19, 2001.

[7] Manfred A. Jeusfeld and Uwe A. Johnen, *"An Executable Meta Model for Re-Engineering of Database Schemas"*, International Journal on Cooperative Information Systems, 4: 237– 258, 1995.

[8] Rateb Abu-Hamdeh, James Cordy and Patrick Martin, *"Schema Translation Using Structural Transformation"*, In: Proceedings of the 1994 conference of the Centre for Advanced Studies on Collaborative Research, PP. 123–143, 1994.

[9] Wie Ming Lim and John Harrison, *"An Integrated Database Reengineering Architecture - A Generic Approach"*, In: Software Engineering Conference, 1996, Proceedings of 1996 Australian, PP. 146–154, 1996.

[10] Jutta Horstmann, Mat. No. 203551, Thesis on *"Migration to Open Source Databases (Migrationsziel Open Source Datenbank)"*, in Technical University Berlin ,Electrical Engineering and Computer Science,Computation and Information Structures (CIS), Einsteinufer 17 D-10587 Berlin Germany, September 29, 2005.

[11] Practical Data Migration by John Morris

[12] Oracle 8i Data Migration Handbook by Paul Dorsey, Joseph R. Hudicka, and Jeremy, Techniques for Analysis of Migration-History Data by Julie Davanzo.

[13] Tips for choosing Data Migration Tools www.dmreview.com/white_papers/

[14] www.swissql.com

[15] www.ouzounov.com

[16] www.endiansoft.com

Author's Biography

A.Mahendiran studied Master of Computer Applications (MCA) at Thanthai Hans Roever College, Bharathidasan University. He has completed M.Tech Computer Science and Engineering from SASTRA University, India. He has also completed MBA and M.Phil in both Management and Computer Science.His Specialization area includes Database Tuning and Moving Object Databases. He has got 12 years of teaching experience at PG level. Currently he is working as Lecturer in the Department of MCA at SASTRA University, Tanjore, India.