Legacy Data Migration: DIY Might Leave You DOA

White Paper

In any application migration/renewal project, data migration is usually a relatively minor component in terms of overall project effort. Yet failure of the data migration process can and does cause failure of the entire project. So why do project managers fail to take essential steps to mitigate this risk?

Often, data migration is perceived as "the customer's problem", or a "problem to be considered later", the expectation being that someone will write a suite of extraction programs that will dump the legacy data into flat files to be loaded into the new relational database management system (RDBMS). However, anytime a particular data implementation is expressed in a foreign representation, there is great opportunity for mistranslation, so such a simplistic point of view opens the door to potential catastrophe.

Given the maturity, wealth of functionality and relative low cost of tools like tcVISION, as compared to the effort, complexity and risk entailed in a "Do-It-Yourself" solution, there is no reason why a legacy renewal project should run aground on data migration.

A comprehensive legacy data migration strategy must promote a number of important tactics and features:

1. Discovery and analysis of legacy source data structures, including both logical and physical definitions;
2. Data modeling and design of the RDBMS tables, often with the goal of not only reflecting the structure and content of the legacy source, but also its behavior;
3. Specification and implementation of appropriate mappings and transformations from source to target;
4. Capture of all source and target metadata into a repository;
5. Accommodation of source data structure changes over time;
6. Native interfaces to both the source database and the target database to maximize throughput and ensure correct interpretation and formatting;
7. An evolutionary methodology for refining the mappings and process, allowing for repeatable and comparative testing;
8. Minimized extraction impact on production systems;
9. Support for phased project implementation, including Change Data Capture (CDC) and possibly bidirectional replication capabilities.

It would be a daunting challenge for a migration project team to design, develop, test and deploy tools to support these tactics and features. Furthermore, to reinvent the wheel for each new project is simply a waste of resources. Fortunately, this is unnecessary, since mature, proven, feature-rich and cost-effective products that contribute substantially to project success are commercially available.

As a leading provider of legacy data migration, replication and integration products and services, Treehouse Software leverages production experience gained in the field since the mid-1990s. Treehouse data replication solutions have been used to support many data transfer and migration projects. Each customer's project has led to enhancements and refinements such that today it is possible to handle virtually any conceivable legacy data migration requirement.

As an illustration of the potential pitfalls of legacy data migration, consider some issues that Treehouse has successfully addressed with respect to Software AG's ADABAS DBMS:

- ADABAS has no concept of "transaction isolation", in that a program may read a record that another program has updated, in its updated state, even though the update has not been committed. This means that programmatically reading a live ADABAS database—one that is available to update users—will almost inevitably lead to erroneous extraction of data. Record modifications (updates, inserts and deletes) that are extracted, and subsequently backed out, will be represented incorrectly—or not at all—in the target. Because of this, at Treehouse we say "the only safe data source is a static data source"—not the live database.
- Many legacy ADABAS applications make use of "record typing", i.e., multiple logical tables stored in a single ADABAS file. Often, each must be extracted to a separate table in the target RDBMS. The classic example is that of the "code-lookup file". Most shops have a single file containing state codes, employee codes, product-type codes, etc. Records belonging to a given "code table" may be distinguished by the presence of a value in a particular index (descriptor or superdescriptor in ADABAS parlance), or by a range of specific values. Thus, the extraction process must be able to dynamically assign data content from a given record to different target tables depending on the data content itself.
- ADABAS is most often used in conjunction with Software AG's NATURAL 4GL, and "conveniently" provides for unique datatypes ("D" and "T") that appear to be merely packed-decimal integers on the surface, but that represent date or date-time values when interpreted using Software AG's proprietary NATURAL-oriented algorithm. The most appropriate way to migrate such datatypes is to recognize them and map them to the corresponding native RDBMS datatype (e.g., Oracle DATE) in conjunction with a transformation that decodes the NATURAL value and formats it to match the target datatype.

tcVISION, a comprehensive data migration and replication solution offered by Treehouse, provides high-productivity, high-efficiency automated capabilities in dealing with these and other issues (Figure 1). While the discussion below will focus on ADABAS, in fact tcVISION provides similar capabilities for a wide range of legacy sources, including CA-IDMS, CA-Datacom, IMS/DB, DL/I, DB2, SQL/DS, VSAM and even sequential files, as well as "non-legacy" sources such as Oracle Database, Microsoft SQL Server, DB2 LUW and ODBC data sources. Moreover, any of the aforementioned sources can also be a target, making tcVISION the most comprehensive and flexible data migration and replication solution anywhere.

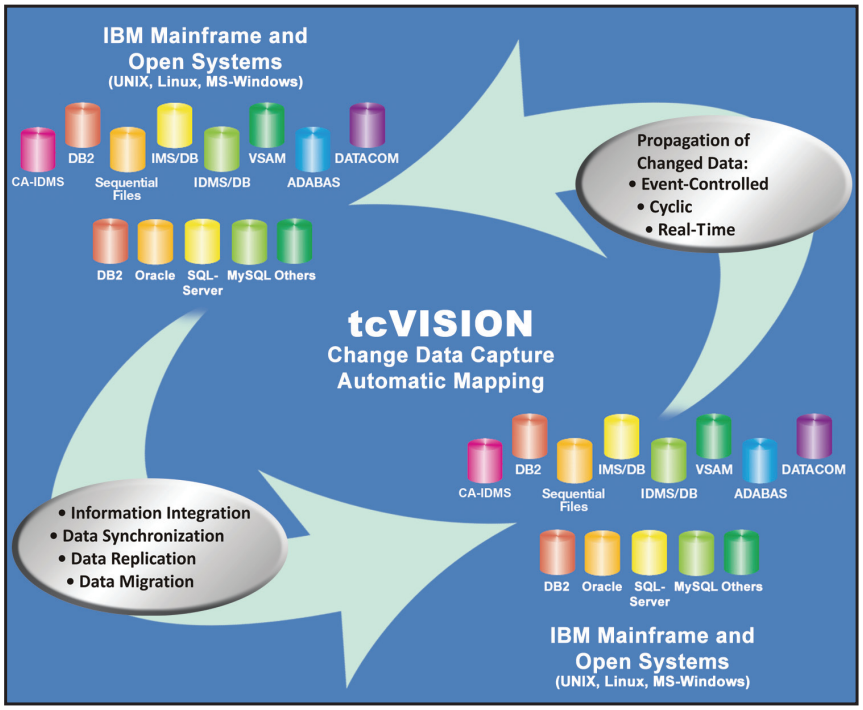

*Figure 1: The **tcVISION** data migration and replication solution*

For discovery and analysis of legacy source data structures, tcVISION includes modeling and mapping facilities to view and capture logical ADABAS structures, as documented in Software AG's PREDICT data dictionary, as well as physical structures as described in ADABAS Field Definition Tables (FDTs). Note that PREDICT is a "passive" data dictionary—there is neither requirement nor enforcement that the logical and physical representations agree, so it is necessary to scrutinize both to ensure that the source structures are accurately modeled.

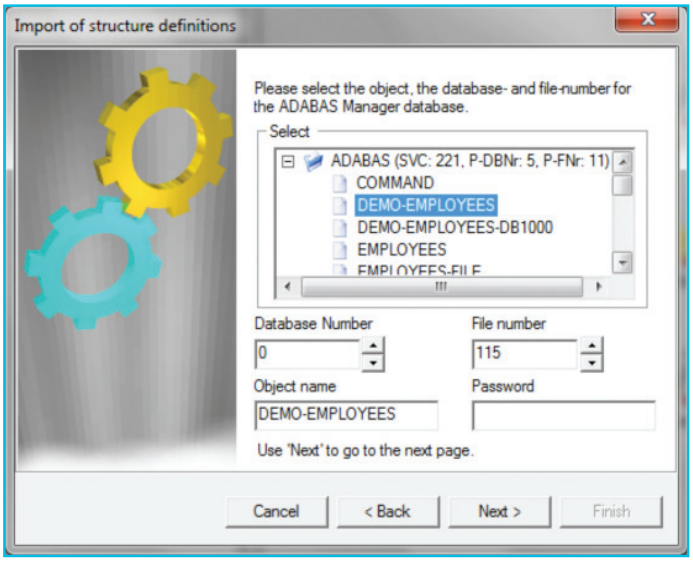

*Figure 2: **tcVISION** capture of source structure metadata*

Productivity in the data modeling and design of the RDBMS tables is greatly enhanced with tcVISION. In seconds, the product automatically generates a complete, native, fully-functional yet completely-customizable RDBMS (IBM DB2, Oracle, Microsoft SQL Server, etc.) schema based on a specified ADABAS file (Figure 3).

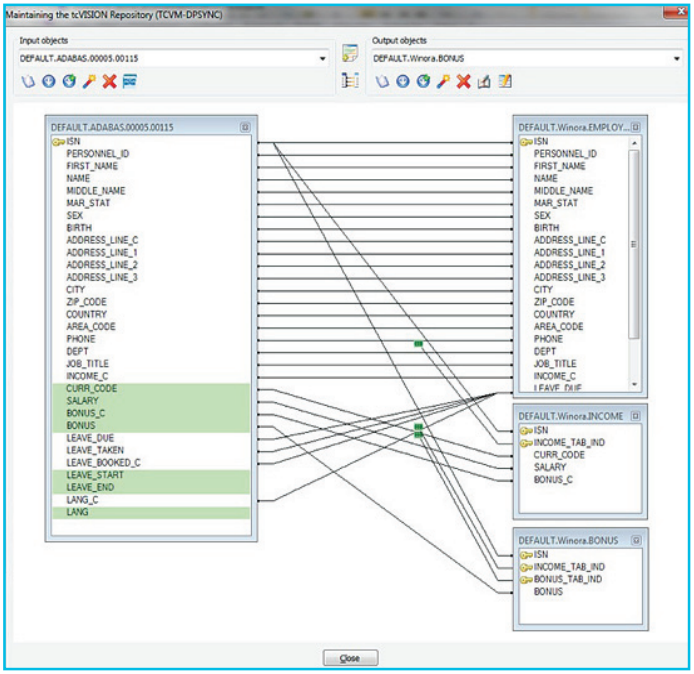

*Figure 3: **tcVISION** target schema auto-generation and mapping*

tcVISION even generates the RDBMS Data Definition Language (DDL) to instantiate the schema (Figure 4).

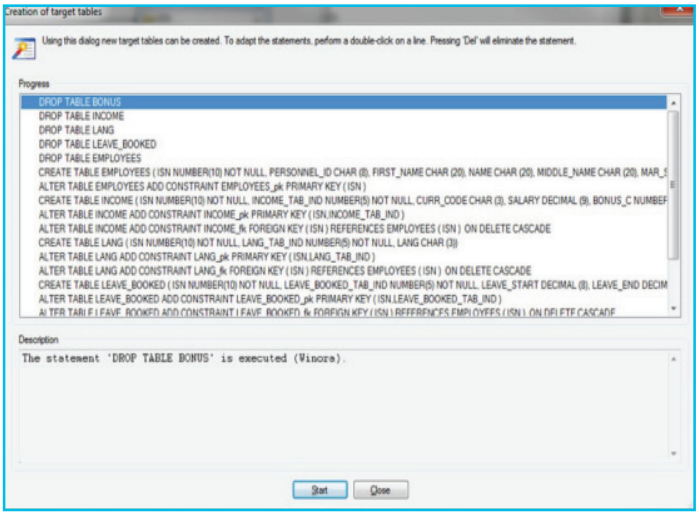

*Figure 4: **tcVISION** target RDBMS DDL generation*

Furthermore, tcVISION generates specification and implementation of appropriate mappings and transformations for converting ADABAS datatypes and structures to corresponding RDBMS datatypes and structures, including automatic handling of the proprietary “D” and “T” source datatypes.

It is important to note that the schema, mappings and transformations that result from auto-generation can be tailored to any specific requirements after the fact. It is even possible to import an existing RDBMS schema and retrofit it, via drag-and-drop, to the source ADABAS elements.

Using the tcVISION GUI, the most complex transformations can be specified. Source fields can be combined into a single column, decomposed into separate columns, and be subject to calculations, database lookups, string operations and even programmatic manipulation. Furthermore, mapping rules can be implemented to specify that data content from a source ADABAS record be mapped to one or more target RDBMS tables—each with its own different structure, as desired—based on the data content itself. Target tables can even be populated from more than one source file.

Over the course of a migration project, it is nearly inevitable that source data structure changes will be required and must be accommodated, which almost certainly will affect the target and the data migration process. tcVISION source-to-target mappings can be specified with a “Valid From” date/time and “Valid To” date/time so that the structure change can be anticipated, scheduled and implemented smoothly (Figure 5).

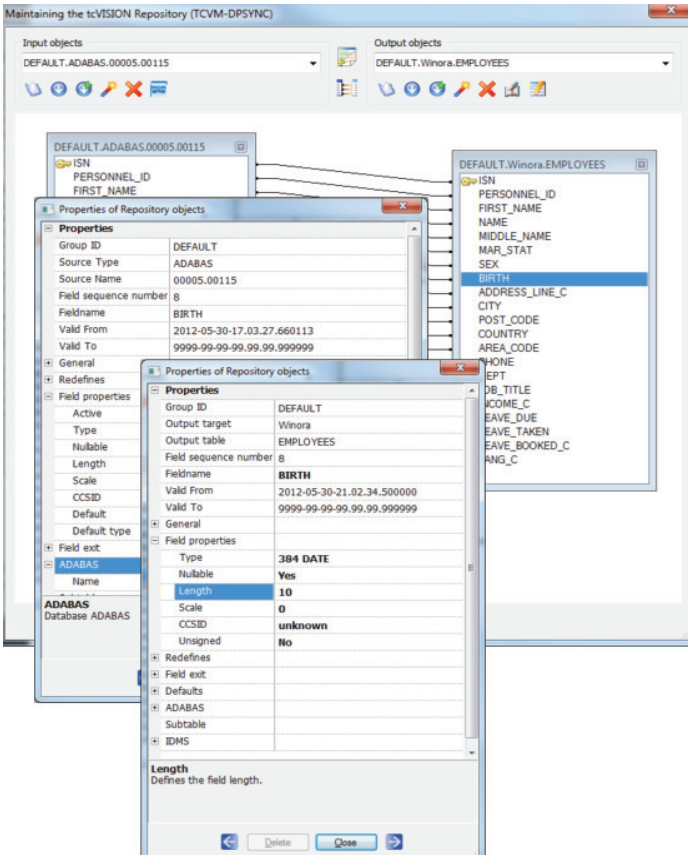

*Figure 5: **tcVISION** source-to-target mapping*

tcVISION captures all source and target metadata into its own RDBMS-based repository, the schema of which is documented to facilitate customized querying, reporting or even integration with other applications. At any time, all metadata and specifications for source-to-target data migration are at the project team's fingertips, and where applicable can be exposed to or integrated with application-migration toolsets.

tcVISION offers highly-efficient batch Extract-Transform-Load (ETL) processing, known as Bulk Loading in tcVISION terminology. Bulk Loading is optimized by use of native interfaces to both the source database and the target database: a static native ADABAS data source (a utility unload of the pertinent database files) can be used as a data source instead of accessing the live database itself—thus avoiding contention and transaction-isolation issues—and the transformed data is natively formatted to be RDBMS-loader-ready, including automatic generation of the loader utility control instructions, with a high-throughput transformation engine in the middle.

tcVISION ETL executes as a single continuous process from source to target, without the need for intermediate storage or middleware. The process can be scheduled and monitored from the tcVISION Control Board (Figure 6).

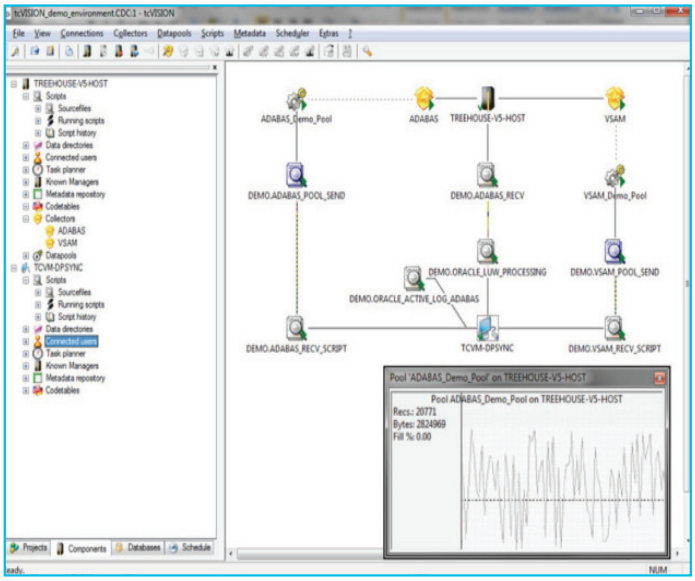

*Figure 6: **tcVISION** Control Board*

With these productivity aids, and an optimized ETL process, an evolutionary methodology for refining the mappings and process is enabled. Treehouse advises data migration customers to "migrate early and often". The sooner that data can be populated into the prospective RDBMS, the sooner that data quality and assumption issues can be discovered, confronted and resolved, and capacity and performance planning commenced. Moreover, very early on in the process a test bed can be generated for modeling and prototyping the new application.

Using a static data source allows ETL processing to be repeated and compared between various mapping scenarios and data samples. Thus, the RDBMS schema and mappings can be iteratively refined over time as requirements are developed and discoveries made.

Customers also benefit by minimized extraction impact on the operational ADABAS database and applications. ETL can be resource-intensive and costly, particularly in scenarios where database calls are charged back to the user and batch windows are tight. When configured to use a static data source, tcVISION places no workload whatsoever on ADABAS itself. Moreover, it is possible to simply transmit the unloaded source data in its raw format off the mainframe and have tcVISION process on a platform with available low-TCO capacity (Windows, Linux, Unix)—incurring no mainframe workload at all.

Due to pure logistics, most migration projects require support for phased project implementation. Only the smallest projects can risk a "big bang" approach. tcVISION users benefit from Replication, tcVISION's term for CDC, in that it enables implementation of the new RDBMS for certain tables, to be used in a read-only manner (generally speaking) in the new system, while maintenance of the source ADABAS data (the "system of record") continues in the legacy system for a period of time. tcVISION's Log Processing mode enables changes made to ADABAS (reflected in the ADABAS protection log, or PLOG) to be periodically extracted and transformed to appropriate SQL statements, according to the established mapping rules—the same ones as are used for ETL—to update the target RDBMS tables. This can even be done in real time using tcVISION's DBMS Extension mode, where changes are captured live from the operational ADABAS database. And for very complex project implementations, tcVISION can be configured to capture changes from both the ADABAS and the RDBMS for batch or real-time bidirectional Replication.

In projects where phased implementation may or may not be contemplated, but where data volumes to be migrated are large, tcVISION Replication enables the data migration to be staged over a period of time—as long a period as is needed to accomplish the migration according to the available processing capacity or other constraints. ADABAS files can be sequenced and grouped as desired, and each file or group of files separately Bulk Loaded to the RDBMS. Thereafter, Replication can be activated for each Bulk Loaded file. This process can continue until all files have been migrated through Bulk Loading and then maintained in sync via Replication—so that all migrated data "arrives at the goal line" at the same time. In fact, the go-live can be scheduled whenever convenient by simply continuing Replication until the appropriate time. This capability offers enormous planning and scheduling flexibility for the migration team.

It is impractical to discuss all the features and capabilities of tcVISION within an overview discussion, so suffice it to say that we have only touched on those most critical ones here. Given the maturity, wealth of functionality and relative low cost of tools like tcVISION, as compared to the effort, complexity and risk entailed in a "Do-It-Yourself", solution there is no reason why a legacy renewal project should run aground on data migration

Treehouse SOFTWARE

About Treehouse Software, Inc.

Since 1982, Treehouse Software has been serving enterprises worldwide with industry-leading software products and outstanding technical support. Today, Treehouse is a global leader in providing data migration, replication and integration solutions for the most complex and demanding heterogeneous environments, as well as feature-rich, accelerated-ROI offerings for information delivery, business intelligence and analytics and application modernization. Treehouse Software customers are able to:

- REPLICATE Data Anywhere
- INTEGRATE Data Everywhere
- MODERNIZE Data from Legacy Applications
- ANALYZE Data for Business Advantage

With unmatched comprehensiveness of tools and depth of experience, Treehouse Software applies proven approaches to help customers and partners mitigate risk and profit sooner from modernization benefits.

For more information, contact us:

Email: sales@treehouse.com

Phone: 1.724.759.7070

Treehouse SOFTWARE

2605 Nicholson Road, Suite 1230
Sewickley, PA 15143