The Process of Data Mapping for Data Integration Projects Data Mapping -A Key Work Product for Data Warehouse, Data Integration, and Data Migration Projects

Method · October 2019

DOI: 10.13140/RG.2.2.10352.81925

CITATIONS
0

READS
2,860

1 author:

Wayne Yaddow
Datagaps
6 PUBLICATIONS **0** CITATIONS

SEE PROFILE

The Process of Data Mapping for Data Integration Projects

Data Mapping - A Key Work Product for Data Warehouse, Data Integration, and Data Migration Projects

Wayne Yaddow
Data Quality Analyst - Consultant
wyaddow@gmail.com

Contents

Introduction

Data mapping is among the most important design steps in data migration, data integration, and business intelligence projects. Mapping source to target data greatly impacts project success – perhaps more than any other task. The outcome of the mapping process is a primary tool for communications between project architects, developers, and testers.

We've come a long way from the time when "data mapping" was a dirty word in e-discovery. But as data becomes more dispersed and voluminous across organizations, having a centralized resource for quickly identifying where certain electronically stored information (ESI) resides is extremely valuable.

Data mapping*: the process of creating data element mappings between source and target data models. Data mapping is used as the first step for a variety of data movement tasks including:*

- *Data transformation or data mediation between a data source and a destination*
- *Consolidation of multiple databases into a single database and identifying redundant columns of data for consolidation or elimination*
- *Mappings that document the origins of data, the processing paths through which data flows, and the descriptions of the transformations applied to the data along those different paths.*
- *Specifying business transformation/conversion rules to be applied to source data*
- *Identification of data relationships as part of data lineage analysis*

Data mapping bridges the differences between two systems, or data models, so that when data is moved from a source, it is accurate and usable at the target destination.

Data mapping is the first step in a range of data integration tasks, one of them being data transformation between the source and destination. A data mapping tool or data mapper connects the distinct applications and governs the way the data from source application will look like when it is mapped to the destination application. It also supports the application of multiple data manipulation functions that are applied to data when it is transformed from source to destination. Along with data, a data mapper should handle multiple structured and unstructured files and formats to map the corresponding fields, creating the output in the desired schemas. Thus, it should support complex data integration tasks.

Data transformations are among the most common problems facing systems integrators as source data is often in an inconsistent format or structure for target systems needing to use that data. This requires integrators and migrators to design and implement code for the mapping operations required to convert the data from one form to another (e.g., from one relational database format to another).

A simple example of data mapping is moving the value from a source 'address' field in a customer database to a target 'client address' field in a sales department database – and changing the target field length and "cleaning" those addresses at the same time.

Data mapping is required at many stages of data integration, data migration, and data warehouse life-cycles. Consequently, data integration professionals must learn data mapping in order to move

and test data; often using an ETL (extract, transform, and load) process.

When and How Data Maps Are Used

Data mapping is used for many types of data movement projects. However, all of the tasks fall into one of two categories.

> **Data migration projects -** the process of selecting, preparing, extracting, and transforming data and permanently transferring it from one IT storage system to another
>
> **Data integration and conversion projects** - combining data residing in different sources and providing users with a unified view in a target system

For data mapping success, an important heuristic is a relationship between the source and the data target - it could be one source to many targets, many sources to one target, or many sources to many targets.

Combined with a well-documented use case describing the need for a map and its intended purpose, heuristics are essential for mapping success. It is imperative that decisions made regarding business rules and map heuristics are clearly documented so the evidence is available to describe why each decision was made and by whom. This serves as an audit trail of decisions made during the mapping process.

Data Mappings Described

It should be assumed that source to target mappings are key for any ETL solution. In addition to containing the mapping of fields from sources to targets, data mappings should define the following important basic information. See Figure 1 for a high-level example of information commonly documented as a source to target data mapping.

Data modelers, data and business analysts, ETL developers, and testers have a keen interest in

- Database connections for source and target tables
- Source and target data descriptions – what each data set represents
- Source and target field descriptions – what each field represents
- Examples of field/attribute contents

Mapping development rules must be verifiable in order to validate mapping(s), regardless of whether it was accomplished by automated or manual means. Assumptions must be used with caution while mapping data; instead full documentation should be used.

Common Development Phases for Data Mapping

Step 1: Discover and define data to be moved — including data sets, the fields within each table, and the format of each field after movement. For data integrations, the frequency of data transfer is also defined.

Step 2: Map the data — map source fields to target destination fields

Step 3: Transformation data — when fields require transformations/conversions, formulas or rules are designed and coded

Step 4: Test — using a test system and sample data from sources, run the transfers to see how it all works and make adjustments as necessary

Step 5: Deploy — once it's determined that data transformations are working as planned, schedule a migration or integration go-live event

Step 6: Maintain and update — for ongoing data integration, data maps will require updates and changes as new data sources are added, as data sources change, or as requirements at the destination change

- Source and target data types, dates, and times (metadata)
- Null, not Null, default indicators
- Transformation, aggregation, enrichment description rules for each field
- Error handling conditions and logic for each record, each field
- Columns participating in referential integrity
- Primary / foreign key columns that assure source records are unique
- How tables are joined (the type of SQL join)
- Slowly changing dimension (SCD) and change data capture (CDC) attributes and logic
- Change and version log entries to describe additions and changes to the mappings

Figure 1: Sample data mapping template

ETL Package Name	*Load_Dim_Product.dtsx*
Target Table	*ADDW.Dim.Product*
Source Tables	*Production.Product* *Production.ProductSubcategory*
Load Frequency	*Daily*
Target Type	*SCD Type 1*
Mapping Description	AD.Production.Product is the lowest level of source data. Join the source tables in the following manner: FROM AD.Production.Product INNER JOIN AD.Production.ProductSubcategory ON Product.SubcategoryID = ProductSubcategory.SubcategoryID INNER JOIN AD.Production.ProductCategory ON ProductSubcategory.ProductCategoryID = ProductCategory.ProductCategoryID
Error Handling	On an error occuring any inserted records or updates needs to roll back and leave the system in the previous correct state.

	TARGET					SOURCE					
Mapping Change Date	**Target Table**	**Target Column**	**Nullable**	**PK**	**Data-type, Length**	**Source Table**	**Source Column**	**Data-type, Length**	**Expression, Transformation**	**Default Value**	**Error Types and Handling**
Y/M/D	Name of target table	Target table column name	Whether a field can be null	Primary key field for target	Data type & length for target column	Name of source table	Column in source table from which data is extracted	Data type and length for this source column	Decodes, aggregates, conversions, if statements, lookup functions	Value to use in target field when source field is null	Used to document Not null, value if looked up, pk, fk, etc…comments, issues

Planning for Data Mapping Projects

Planning is arguably the most important stage of the entire data mapping project.

Documentation needed for a project's sources should include data maps and data dictionaries to deliver a complete definition and the intended use of each source data component. Data dictionaries help to assure that the interpreted meaning of each source data element is correct. For example, a field for "Provider ID" could have many different definitions over several elements such as billing identification number, national provider identifier, social security number, etc. Data definitions with similar or exact names could differ considerably in meaning.

Analogous to any complex project, planning for data maps requires:

- Define objectives for the mapping project
- Gain IT and business management buy-in
- Define specific mapping deliverables
- Assign mapping roles and responsibilities

Meeting the Challenges of Data Mapping Projects

Source to target mappings describe how one or more attributes in source data sets are related to one or more attributes in a target data set. These source-to-target mappings are derived from the ETL transformation rules described in a requirements specification document, comments inside the transformation scripts, spreadsheets, ER diagrams, or SQL scripts.

Data mapping is complex and challenging. So what makes data mapping so difficult? The following are common challenges and shortcomings associated with data mapping and how they can be mitigated.

- **The time, people, and tools needed to build data maps can be substantial**

The process of connecting data sources, building mappings for data transformation and integration, and validating the transformed data often require significant resources, particularly when the entire process is done manually.

There are several ways to ease the data mapping burden significantly. It starts by defining the process for gathering information to be documented for each source and target. In most cases, systematic interviews with data stewards are the most efficient way to collect info for a data map. Interviews with subject matter experts (SME's) should be direct using data mapping templates. Meta data and data mapping tools should be used to automate as much as possible.

- **The information needed is not always available for building data maps**

A common mistake organizations make with data maps is that they omit important information and therefore render the data map far less useful than it should be. Before data mapping initiatives get off the ground, project organizers should assemble key stakeholders and gather feedback on what information needs to be included in mappings for sources and targets. For example, retention schedules, litigation risk profiles, and accessibility constraints of particular data sources. Privacy officers will want to know which data sources contain sensitive customer information that must be carefully protected.

- **Substantial efforts needed to maintain data maps**

As with all important project documents, data maps should be constantly evaluated, updated and assessed for quality. One method to ensure the data mappings are maintained is to make sure the process is fully integrated into the organization's master data management program. With every change to requirements, data maps should be reviewed to assess the impact.

- **Data mapping with spreadsheets can pose long-term issues**

Creating manual mappings using spreadsheets is often difficult and time-consuming.:

Mappings specifications built using spreadsheets cannot be easily managed

Data mappings cannot be easily versioned and auditability of what and who has changed mappings remains a constant issue.

Creating maps internally and using unqualified personnel for map development compromises the integrity of results. Use skilled personnel familiar with data mapping requirements, limitations, and pitfalls to ensure reliable results.

Steps to Success in the Data Mapping Planning Process

1. Determine which data sources are needed to meet requirements for the target system

General steps to source data discovery:

I. Identify the data needed to meet required business tasks
II. Identify potential internal and external sources of that data
III. Assure that each source meets the privacy and regulatory requirements
IV. Assure that each source will be adequately available and accessible according to required frequencies

2. Identify tools for data analysis, data preparation, and data mapping

It will be necessary to load (i.e., frequently samples) data sources into an environment of data preparation (DP) tools where the data can be analyzed and manipulated. It's important to get the data into an environment where it can be examined and readied for the next steps.

3. Conduct data profiling on potential and selected source data

This is the vital (but often discounted) step in DP. The project team must analyze source data before it can be properly prepared for downstream consumption. Beyond simple visual examination, projects often need to profile data, detect outliers, and find null values and other unwanted data among sources. This can provide an insight into the state of data quality.

4. Cleanse and screen source data

Based on the knowledge of the business goals, experiment with different data cleansing strategies that will get the relevant data into a usable format. Start with a small, statistically-valid sample to iteratively experiment with different data prep strategies, refine data record filters, and discuss with business stakeholders.

Planning Considerations for Data Mapping Projects

- *A typical plan begins with meeting IT to gain an understanding of systems, assets, retention policy and practice, employee separation procedures, archives, backup system and outsourcing of data storage or management. IT is the primary authority on sources such as corporate email and backups.*
- *The next step is to consult with business unit leaders about needs and general data practices. They will flag the seemingly inevitable data repositories and associated software programs IT doesn't know about. And, don't' forget to meet with records managers about specific document management systems, databases and file rooms.*
- *At this point in the planning process enough information has been gathered to build the data map in outline form. Request information about the format, volume, security information, etc. from IT - continue to narrow the focus by gradually filling in gaps and resolving inconsistencies.*
- *Enterprise data mapping software solutions automate some parts of the process and can be used to generate the map instead of manually creating a spreadsheet. Large corporations with complex IT systems and companies in highly regulated industries should evaluate investing in data mapping software.*
- *Processes and procedures must be clearly defined and documentation prepared to explain how the map was developed, and tested, to work correctly for its intended purpose.*
- *All data maps of any kind must be identified, inventoried, maintained with schema changes, and verified. Poorly designed and out-of-date mappings create significant data integrity problems. Undetected errors in data maps have the potential to introduce many problems, including current and "up the line" as data is propagated down-stream to other systems.*
- *When evaluating data integrity issues that involve mapping, it's critical to understand the elements of all the code sets or data sets that will be mapped. The characteristics of each source or code set, their intended use, and how the map is created are all important to building a successful data map. Using a map for a purpose not intended or misunderstanding the construct of the source and target can lead to incomplete, incorrect, and inappropriate maps.*

Planning for data mappings should begin after project requirements are "ready" and after all data sources have been identified and data appropriately "prepared" to meet the needs of requirements and target data.

Plan enough time to evaluate the source data, to compare the available data to the needed data, and to drill down to the detail needed for source-to-target mapping. One goal is to ensure that data

migration development is the shortest task in the project plan. The source data and target data should be well-understood before coding begins. Project leaders don't want to experience the very costly surprise of learning that the source data is not "fit to use" at integration testing or implementation.

Best Practices for Data Mapping Projects

All data maps require an investment of time and resources, some more than others. When the source data and the target data are similar in structure with a high percentage of exact matches in content and meaning, the time needed for data mapping validation will be minimal.

However, when the source and target do not result in an exact match, time must be spent to determine which of the mapping choices are appropriate based on the map's use case. Unless the mapping is a one-to-one match from source to target, decisions must be made to meet the intent of the map.

In order to optimize the use of data maps, the following practices are recommended:

- Document the map heuristics and business rules surrounding the development of each mapping. Include use cases for each data mapping; identify applications that use the maps; document how mapping rule is created and deployed in the workflow.

  Mapping heuristics represent a "rule of thumb" guidance that provides rules for how to map from source to target in a consistent manner for a specific project. Detailed instructions should be provided so that consistency is ensured between map developers throughout the project. Every mapping project must have clear instructions to assure map results are "understandable, reproducible, and reliable."

- Perform a Data Mapping Assessment - What are we moving, and what are the transformation rules? How much time will it take the team to complete data mapping? It's often necessary to have a general idea of what it will take to create the final design and implementation. The captured business needs, the source and target system metadata, and the data profiling results from the data quality assessment all create the information needed to understand the mapping effort.

- Prepare a process to test the validity and reproducibility of the mapped data. A verification process should represent the data mapping development process to include tools used from map development to end-user acceptance testing and approval.

- There is no one-size-fits-all data map template. IT professionals should select an appropriate data mapping template to manage their data integration or migration.

- Authoritative maps save development costs. Data mapping templates supported by standards development organizations or mandated by government agencies usually have been validated and tested to ensure they work for the purposes for which they were developed. This saves the cost of creating and testing locally developed maps.

- An identified organization, department, or individual should be in charge of implementing, maintaining, and updating each data map.

- There is an increasing range of data mapping tools and software solutions available in the market and among open source. Commercial and open source tools should be assessed to aid the mapping process.

- Mapping should be reviewed then revised when source and targets are updated. This may require updating maps multiple times per year. Each update must be clearly identified as a different version, and documentation should detail the revisions for both the source and target.

Data Mapping Automation and Associated Tools

Data mapping is complex and can be accomplished in a variety of ways. Many software providers offer data mapping software. However, these various solutions do not each provide comparable or comprehensive features. When comparing different types of data mapping software programs, particular key factors that should be considered:

- Tools should offer advanced data visualization capabilities for selecting functions, selecting sources and targets, and reports for review of mapping results

- Tools should be customizable. Such features allow users to adapt the software to fit the particulars of their technologies and business needs better

- Tools should be easy to use, and not require extensive training in order to implement. This improves adoption rates, and saves on costs associated with the implementation

- Tools should support a large variety of data sets (ex., RDBMS, JMS, SOAP), and formats (csv, XML, etc.). This assures that users are able to retrieve and map their variety of data easily

Some organizations continue documenting data mappings on spreadsheets. However, modern data integrations and migrations are too complex and varied for manual efforts to be effective. With more data, more mappings, and constant changes, such "manual" processes should be reconsidered. They often lack transparency and don't easily allow tracking the inevitable changes that occur in project requirements, data models, and schemas.

Choosing the Best Data-Mapping Solutions for the Project

A key to choosing the correct data-mapping solution is product research. Software providers who offer free trail periods make it easier to understand what kind of value is offered. Online reviews may be useful for determining which data mapping programs to investigate further, but business leaders should remember that not every solution is a perfect fit for every user, and some negative reviews may be a result of incompatibility between the users and the software for mapping data. The best data mapping software should be customizable and adaptable enough to provide value to businesses of all kinds.

Conclusion

Data mapping is always resource-intensive requiring hands-on development, review, and knowledge about all sources and targets. Human intervention is necessary for mapping design and validation of map outcomes. Commercial and open source mapping tools can assist in the process by providing varying degrees of automation. Manual review is required, to a varying extent, to map the portions that failed automated mapping and to validate the results of automated mapping.

Data mapping can make a difference when it comes to getting data under control. It makes it easier to generate reports and to figure out how the data coming into the organization's workplace is organized. This can make a real difference in terms of getting information ready for data integrations or migrations.

References

"10 Best Data Mapping Tools Useful in an ETL Process", https://www.softwaretestinghelp.com/data-mapping-tools/

"5 Best Data Mapping Software Tools", http://intellspot.com/data-mapping-tools/

"Data Mapping Tools", https://www.alooma.com/blog/data-mapping-tools

"Understanding Data Mapping and Its Techniques", https://www.astera.com/type/blog/understanding-data-mapping-and-its-techniques/

Mohammad Azad, "*Secret to Data Migration: Put Data Mapping First*", https://www.linkedin.com/pulse/secret-successful-data-migration-put-mapping-first-mohammad-azad/