



Building Bridges the IT Way — Creating a Solid Data Foundation

A Step-by-Step Guide to Obtaining Information Consolidation

Jon Billig, President, Information Consolidation, LLC

If your company has ever invested or has considered investing any money into a data warehouse, federated data warehouse, business intelligence (BI) application, or customer relationship management (CRM) project, you've probably done the research and you know that the foundation of any data warehouse or similar business initiative is accurate, well-analyzed information.

You may have also read the horror stories about companies throwing away hundreds of millions of dollars on data integration platforms that never really materialized. They're now sitting in some software graveyard in the bowels of the organization. Of course, you want to avoid that from happening to your company, but it seems like such a daunting task to consolidate all of the information coming into an organization and to ensure that this data remains usable and uncorrupt. Is there a way to effectively consolidate the information while reducing the risk of the platform becoming another multi-million dollar dust collector? There is, but first it's necessary to understand what information consolidation is and what the steps are to attain it.

Information Consolidation, or as it's sometimes called, Data Integration, is the act of gathering of information from across an organization and storing it while maintaining high quality data. Information Consolidation provides a means to retrieve and update information quickly and efficiently. Once data is consolidated, it is then transformed into usable information, which converts into knowledge. How you use this knowledge becomes the key to success.

It may not be necessary to go out and purchase or build a new Information Consolidation tool in order to be successful in your project. It's very possible your organization already has the tools and they just need to be retrofitted for your project. An independent organization that does not have any ties to any particular product line can recommend a specific solution to fit your company's individual needs.

It Takes Blue-Collar Work to Make it Happen

There is so much theory swirling around Information Consolidation, and volumes have been written on the subject. But this isn't academia, and there is no time to have meeting after meeting about it. You have a job that must be done, and there are deliverables that must be created. You just need to roll up your sleeves and do it.

With experience (or the help of someone with that experience), Information Consolidation is quite simple. There are hundreds of proven success stories. The

companies that have failed usually cannot blame a technical problem; rather the lack of a proper plan is usually the culprit. There are many consultants out there who talk about data integration and the things a company must do in order to succeed in the endeavor, but many of these consultants merely add to the confusion (in order to maintain their own job security). The bottom line is, in order to succeed, you need to move away from the theory of business intelligence and dive in to the practical steps of getting it done.

Inspect, Inspect, Inspect!

Information Consolidation is like building a bridge. Both require a well laid-out plan, an engineer, high quality construction, and a great deal of review and testing. You can hire the best architect to design your structure (data model), but without the foundation (data integrity), your bridge will be swept away.

Like building a bridge haphazardly, it's dangerous to throw together a data warehouse. If a Vice President at your company tells you to build a data warehouse, you most likely will go out and build one. Within a matter of weeks you have your new data warehouse, but it's a rush job, so you skip steps and forgo inspecting as you go along. The inspection phase is critical! Just as the law states that the concrete pilings on a bridge must be inspected, so should every stage of building a data warehouse end with an inspection.

Users should sign off after each of the following phases:

- Design
- Inventory
- Adjustment
- Organization
- Link
- System Test

Ten Steps of Information Consolidation

There is a direct proportion between how much time you spend with data and the success of your project. Like solving an algebra equation, Information Consolidation requires you to complete several steps, and, if you skip any, the final product will simply be incorrect. The steps to a successful Information Consolidation project are below:

1. Inventory (a.k.a. Data Profiling or Investigation)

The first step in a successful data consolidation effort is to take inventory of all of the information that will be going into your warehouse. Study the source data thoroughly to understand its content, structure, quality and integrity. Just as a nuts and bolts company would take inventory of all of the nuts and bolts going into its warehouse, a company building a data warehouse should inventory all of the information going into it. Without proper inventory, data quality issues will begin to corrode your database, before you have begun.

It may be tempting to search for anomalies in the data by looking at a sample rather than the whole data set, but it's not sufficient. **All** of the information going into the warehouse needs to be correct, a situation impossible to attain through sampling.

Once you've identified invalid or incorrect information, it needs to be corrected or specified as a missing value. The missing information can be corrected by using enhancement procedures, and there are many good tools available for inventory and reporting.

Don't overlook the meta data. Make sure to report all anomalies you find, and obtain review and signoff of the results — the reports are worthless if only IT looks at them.

2. Plot (a.k.a. Mapping)

Plotting is the process of determining where the data will go in the target system. By plotting your course to show the relationship between the columns of data in your source and target files, you can capture the transfer-and-load requirements to successfully move your data from the source database to the target.

Tips for a successful data plot:

- Use inventory results and specify your conversion rules — there are a few tools on the market that allow you to connect the inventory results to the conversion rules.
- Leave no column undefined. A good rule of thumb is that the last 15% of the plot will take as long as the first 85%
- Just as in the inventory phase, make sure you document all of the rules in a meta data set.
- Get review — the project sponsor must assign someone to review.
- Get signoff to assure that the documentation is reviewed.

3. Unload (a.k.a. Extract)

Unloading is the act of obtaining in usable form the legacy data that is to undergo Information Consolidation. When unloading data, simplicity is the key. It is usually best to employ the database utility because it is available and efficient. Some extract, transfer, and load (ETL) tools also provide unload utilities. You should pick one that allows the input of legacy information into your meta data repository.

If you use coded programs to unload then you run the risk of those programs masking some of the data issues that you will face in production.

When the source fields are embedded in a hierarchical database with their keys implied by the record relationships the process of unloading must explicitly supply all those keys which will ultimately be required to successfully load the target fields.

4. Adjust (a.k.a. Data Cleansing, Conditioning, or Transfer)

Once the data has been unloaded, you will inevitably need to make changes or add information. If you are dealing with customer information, you will find that name and address fields must be parsed and standardized so that any aliases or variations on a name (Doctor William H. Smith 3rd and Dr. Bill H. Smith III may be the same person) are fixed and any abbreviations or non-standard address formats are changed

to standard formats. Some data will be adjusted because links have been established, however it may not be passed to the target database.

Again, the meta data component is extremely important during the Adjust phase — an auditor may review your reports and want to know how the data has been changed. Keep any data adjustments documented and then get review and sign off from the project sponsor.

5. Organize (a.k.a. Data Cleansing, Conditioning, or Transfer)

By the time you get to the Organize phase, you will not be changing the actual data values, just moving it around. This is your opportunity to make sure that the data is in the correct fields and it's in the correct format. For example, components of parsed names should each be in a separate field (separate prefix, first, middle, and last name), dates of birth should be in a consistent format (mm/dd/yyyy may need to change to yyyy/mm/dd), and address components should each be in a separate field (separate building number, street name, street type, etc.)

By now, you're probably getting the point that you need to document each step of your process in your meta data files. The Organize phase is no exception. Again, someone conducting a future query must know how the data has been reformatted just in case it is audited.

A business sponsor doesn't need to review or signoff on the Organize phase because there are no changes to the information, but it should have internal IT approval. At this point you should implement change control.

6. Link (a.k.a. Matching, or Data Cleansing)

Linking is part of the data cleansing process. It identifies duplicate customers and links them to others who share a common name, address, account number, or any other user-specified fields (e.g., SSN, driver's license, eye color, etc). When conducting the Link phase, you are using free form text as the key — as long as the same algorithm (scores, weights, etc) is used you will always get the same link. When you change the algorithm after the Sponsor begins using the data then you must prepare to perform updates to the keys they depend on for analysis and down stream systems.

You should start by conducting a probability query, or a fuzzy match, where weights are commonly used. Next, take a look at the domain, or the exact match fields, taking into account that those fields will also affect your match efficiency. Then review, review, and review! First conduct a loose review and then two or three iterative reviews, until you're sure anything that should be linked, is linked.

You probably know what comes next — document — but this time make sure that your meta data is in non technical terms because this information must be approved by the Sponsor. Scenarios of each rule must be included for clarification.

Obtaining signoff on the Link phase is critical for the success of the entire project, so make sure not to skip it.

7. Load

Loading the data is simply the act of creating a file that can be uploaded to your target Data Base. In order to successfully get through the Load phase, you should first adjust data to make sure it adheres to the Data Base edit criteria (null values, missing values, etc.).

Next you need to organize the data so it is in the format expected by the Data Base load.

You know what comes next...document!

8. DB Load

When conducting a DB Load, you will use a load utility or tool — many are readily available on the market — to take all of your information and load the Data Base table by table. This usually a DBA assisted job or may be handled entirely by the DBA.

9. Survivor

The Survivor phase is precisely what it sounds like: it is the one customer record that survives (replaces) all the duplicates. Remember, one size does not fit all — each company will have different requirements defining what rules will be used to determine the best survivor, so if you haven't done so already, you will need to determine what rules best fit your company's requirements. Here are some suggestions on determining Survivors:

- Best of best: some companies pick the best field from all records
- Latest: some companies will pick the last record updated
- Best one: some companies will choose one out of all the records
- Non-name/address criteria: other elements such as birth date, policy role, checking account signee, etc. can also be used.

Once you've chosen a survivor you must have the results reviewed and signed off by the business user. It's important that the user(s) agree that their requirements are satisfied based on the rules you've chosen.

Document.

10. Household

The process of householding is primarily for your organization's marketing group. Householding provides your company with a full-view of your clients' households. It allows for cross selling opportunities because you have linked up members of one household (or one company). It also has operational benefits such as consolidated statements and it can help eliminate duplicate mailings to different members of the same household.

Please realize, however, that like the Survivor phase, each company will have different requirements of how the rules will be setup to define a household. A good rule of thumb is to try to determine the primary contact. There is usually one person in the household that will be the person to which communication from your company will be made. The rules for determining that person will vary from company to company.

The user of the household information must review and signoff on the results of the rules implemented and make sure that they result in satisfying all their requirements.

Document.

Consultants Are Meant To Be Temporary

When you work with consultants on an information consolidation effort, remember that they should be working themselves out of a job by training your team to understand the methodology and tools used to create the information consolidation project. They should make formal training available as the final process delivered to your company.

System Documentation is Key

You probably have figured out that documentation in your meta data is a must on the project work plan. This is one item that cannot be cut back when the project is looking to save time or money. If it is it will result in your project deliverables becoming useless in a matter of months, I promise.

Don't Forget About Ongoing Data Quality

Just because you ran a data quality tool at the beginning of your process doesn't mean that you can check data quality off your project plan (that is, unless your company wants the same situation that started this project in the first place — unusable information.) It's essential that you create base metrics and ongoing monitoring of the information quality in your new system.

Follow Your Company's Method

During your project make sure that you follow a system development methodology. This method probably already exists in your company and any outside consultant should be flexible enough to follow those standards. If there are not standards that apply to the Information Consolidation project cycle, or if a system development method hasn't been developed for your company, the consultant must be prepared to help with developing those standards.

About Information Consolidation, LLC

Information Consolidation, LLC is an independent consulting firm that takes operational information and inventories, plots, unloads, adjusts, organizes, links and loads it to a target system designed by you or by us. That system could be a Data Base, Data Mart, Federated Data Warehouse, Corporate Information Factory or any other data store that then can become the object of Business Intelligence tools. We have expertise in

Ascential's Quality Stage (formally Vality's Integrity), Trillium, First Logic, Innovative System's Ilytics, Avellino's Discovery, Informatica's Power Center, SAS, SAS Data Flux and many other tools used in Information Consolidation. We are the blue-collar workers that enable the information in these target databases to be accurate, complete, reliable and accessible. For more information about how you can utilize the consulting services for Information Consolidation, contact Information ConsolidationSM, LLC at (704) 533 3097.

