firstlogic White Papers

Power of Matching: Minimizing Duplicate Data Overhead

Published August 18th, 2021

In this whitepaper, we look at the expanded data quality applications for well-known data manipulation techniques: matching, consolidating, merging, and de-duping.

When corporate stakeholders think of data matching and consolidation software, most focus on tasks like eliminating duplicates or running a data file past a suppression list. The object of this process is to trim production and postage expenses wasted on duplicate materials or unqualified prospects.

But matching and consolidation software can do much more. Some companies recognize this potential and use their software in very sophisticated ways. Their applications are a long way from what one might describe as “merge/purge”. Their matching and consolidation benefits extend far beyond postage savings.

Especially today, companies are looking for ways to create a 360-degree view of their customers, or what they refer to as a “single customer view” (SCV). They understand that customer experience is an important competitive differentiator. Companies want to be sure they deliver a better customer experience than their rivals. Using information they know about each customer to treat them as individuals is a key to enabling superior customer experiences.
Because enterprises store customer data in a myriad of unconnected databases, duplicate and redundant data abounds. Companies understand they must first resolve that duplication and then connect those disparate data silos before they can achieve the SCV. Data matching and consolidation software is the only way to do that.

In this whitepaper we’ll tell you about ways software like Firstlogic’s Match IQ, part of our Data Quality Platform, can benefit companies and help them achieve their business objectives. We’ll cover some features this class of software offers and explain how to use them to gain a competitive advantage.

Why Match and Consolidate?

Organizations waste time and money working with duplicate, inconsistent, or contradictory data. As data becomes more sophisticated and available, the problem worsens. Although Big Data expands opportunities for identifying markets and prospects, taking advantage of those opportunities requires companies to correlate new data sources with their own databases. Comprehensive data matching and consolidation practices are essential. The benefits of such efforts impact several areas of the business.

Lower Costs

The traditional benefit of duplicate removal is lower printing and mailing costs for direct mail marketing, correspondence, shareholder communications, etc. Those benefits still exist, but companies that merge databases and consolidate information also save money by making better business decisions. Duplicates skew statistics about geographical customer distribution, product interest, or other factors. Cleaning up the data so corporate databases record information for each customer only once will help companies make informed judgements about capital investments, promotional campaigns, store locations, and many other decisions.

Improve Campaign ROI

Organizations that send two direct mail pieces to the same person on the same day double the cost and lower the response rates. Eliminating duplicates through householding improves campaign performance. Data merged from multiple sources provides the “single customer view” companies crave. With a consolidated view of all the interactions and relationships each customer has with a company, the organization can make their communications more relevant and interesting. Consolidated data allows companies to present each customer with messages more likely to elicit a positive response.

Lower cost, plus targeted messaging, equals a greater return-on-investment for a marketing campaign.

Improve Customer Experience

When companies recognize individual customer relationships, the customer experience improves. By matching records and consolidating data into a single customer view, companies enable all customer communications to acknowledge customer loyalty, refer to past purchases, etc. The 360-degree customer view made possible by matching and merging previously unconnected data allows organizations to tailor communications and pair them with the relationships and interactions the company has with each customer.

Uncover New Opportunities

Techniques such as multi-buyer identification allow companies to test new products with their best customers or identify new markets otherwise invisible in environments where data silos keep information separate. Multi-buyer identification is useful even when the objects of the analysis are not yet customers. A company seeking prospects to purchase professional-grade golf clubs, for example, might examine subscriber lists from several golf magazines. Subscribers appearing on three or more lists are more likely to be serious about the game. These individuals will be better prospects for high-priced equipment than golfers who read only one golf publication.

Shorten Run Times

Processing duplicate or irrelevant data makes batch jobs run longer than necessary and requires extra bandwidth to move data throughout the organization or to offsite archives. In time-sensitive situations, duplicate data may force companies to add servers, switches, or other networking hardware to meet deadlines. Duplicates dropped from data files can speed up processes and make expensive equipment upgrades unnecessary.

Decrease Storage and Archive Space

Disk space is relatively inexpensive, but it is not free. Excess data records occupy space more valuable for accommodating positive growth. Companies can delay or avoid hardware acquisition costs, maintenance, and overhead associated with adding more physical or cloud storage by merging data or ridding themselves of low-value duplicate information.

Big Data

Clean, complete data has a huge impact on performance and results. Most companies use data from several sources, some from within the organization and some from without. Consistency and quality are always issues. Formats and data structures typically differ.

Input File Analysis

Before attempting matching and consolidation projects, users must clearly understand the data with which they will be working. Data scattered across an enterprise exists in multiple formats. Software used by different departments store data in proprietary layouts. Data files from outside the organization compounds data structure diversity. Time spent on research, documentation, and verification at this stage of a data consolidation project will pay off as work progresses.

Standardization

Tools like Firstlogic's DataRight IQ® can handle the most complex cases of multi-format data files, translating them into a common, reliable file structure. Matching then becomes a simple matter, with fewer provisions for exceptions. Without software like Firstlogic DataRight IQ, users may be forced to reformat files with custom programs or text editors – a tedious and time-consuming task.

Include/Exclude

If a project calls for only a subset of the entire database, processing every data record in large files increases run times unnecessarily. Include and exclude logic built into the matching and consolidating software limits operations to only the pertinent data. Without access to commercial-grade matching software, users turn to filters in general purpose software like Excel or text editors to extract desired records from the main files manually. These extra manual steps take time and increase opportunities for mistakes.

Building Match Keys

Match keys can be simple, such as an account number or customer name, or they can be complex, involving dozens of data fields. Advanced software allows users to define match keys from within the program. Those attempting to remove duplicates or match data records with tools such as Microsoft Excel or Access must build the match keys themselves in a separate step.

Sequencing

Input file sequence directly influences the matching software's performance. With some tools, matching will only work if duplicate records are adjacent in the file. Data file sequence is also critical for operations, such as selecting records with the most recent transaction date from a group of duplicates or choosing the record with the highest transaction dollar amount.

Suppression Lists

Organizations must define suppression lists before running match/consolidation jobs. Common suppression lists include current customers, deceased individuals, DMA Choice Do Not Mail list, prisons, nursing homes, and bankruptcies.

List Preference/Priority

The best matching software allows users to specify a priority if a customer appears on more than one list. Data accuracy may vary among lists, prompting users to designate records from a particular list as the master record, rather than

allowing the software to make a random decision. Some simple matching software products assign the first record it reads as the master record, regardless of data origin.

Rules

Rules are the most important part of any matching or merging effort. The criteria that define a duplicate can change from job to job – even when using the same data sources! Carefully consider the conditions when the matching software should deem records a match and when it should not. Special circumstances may override the normal criteria, so users must integrate exceptions into the rules as well.

Rules for Matching

The rules applied to a data matching project will vary according to the intended use of the results. The first decisions for users are identifying the data fields used to define match keys. What information in the data is necessary to match two or more records? How loose or tight must the match be? Deciding what data to inspect and how to rank possible matches can make a big difference in the results.

Rules for Not-Matching

Just as important as describing matching rules is identifying conditions that disqualify records from matching. Consider a situation where home address is the data field used to identify matches. Does finding two or more records with the same home address automatically constitute a match? Can you drop the duplicates, assuming all records are members of the same family? What about unrelated roommates? Users may have to include fields such as names or account numbers in the match keys to create the most accurate results, depending on the objectives of each project.

Exceptions

Data conditions can cause matching software to match records that should be separate or fail to identify matches as intended. Recognizing these exceptions, and adding logic to handle them, improves the validity of match criteria and software settings. Blank fields, for instance, can cause two unrelated records to match. Conversely, blank fields can also disqualify matching records the software should group, depending on user-set criteria.

Multi-Buyers

One type of matching analysis is identifying buyers or other entities that appear more than once. Rather than eliminating duplicates, users may find value in locating this data and assigning them a priority. Sometimes, this analysis involves comparing data from different lists. Users rank individuals that appear on multiple lists higher (or lower) according to the number of different lists on which the data is found. In other cases, the number of times an individual appears, regardless of how many lists include them, is the determining factor.

Iterative Process

Matching and consolidation projects can require multiple runs where users adjust settings and criteria, fine tuning the process to produce desired results. When dealing with very large databases, unexpected exceptions and anomalies hidden in the data can skew the outcome. Users often run match or consolidate jobs, examine reports and output files, and then add or modify the rules before running the jobs again.

Data Matching Science

You might think matching and duplicate data detection is a straightforward exercise, and many times it is. Want to see if a data file contains duplicate account numbers? That's easy. A middle-schooler could write a routine that accomplishes this task. It involves no variables other than simple issues, like formatting or leading-zero suppression.

But what if you're analyzing data files of customer contact information constructed in different time periods, by different organizations, or with different rules? This data may have names, customer behavior, history, and other contact information in various formats. The data values are probably inconsistent from file to file. A standard match routine will not recognize that James Arnold, Jim Arnold, JR Arnold, Ross Arnold, Junior Arnold, and Arnold James could be the same person. For more sophisticated matching, you'll need software developed by data scientists. The routines may use the deterministic or probabilistic methods of match detection-or both!

Deterministic Matching

Deterministic matching seeks equal values for data fields from one data record to another. This may sound like the simple account number matching example we mentioned above, but sophisticated deterministic matching uses scoring to decide how strong a match it has made. The software will also account for the presence or absence of data values. This is more sophisticated than a simple byte-by-byte comparison.

One hundred percent positive matches occur when the values of all inspected data fields are the same in both data records. When data fields exist in both compared records but the values are different, the software will decide the records do not-match exactly and will assign a weighted score value depending on the strength of the match.

Combined matching and non-matching data fields ultimately control the score for a pair of data records with field-to-field scoring. Field-to-field scoring uses word or phrase similarity, noise word removal, cross-field comparisons, and weighted scoring of fields, which all contribute to the overall record score.

Users decide the thresholds for taking action. If the score falls below the threshold the matching software will not merge the data. High scores may be considered positive matches and cause the data to be combined. Scores between the high and low thresholds may be tagged for manual review.

Probabilistic Matching

With probabilistic matching, the software computes a matching score that determines the probability of a match. To use our example from above, matching “James Arnold” with “Jim Arnold” would yield a higher score than matching “James Arnold” with “Junior Arnold”. “Jim” is a common nickname for “James”, but “Junior” is not. We can’t rule out the possibility that they are the same person, however, without additional data.

If the social security numbers for James and Junior are different, the software won’t make the match. Contrarily, if supporting information such as matching birthdates, spouse names, or street addresses exist, the match score for “Junior Arnold” could rise.

To be most effective, the probabilistic method considers many data fields. The more pieces of data the software compares, the more accurate the results. Probabilistic matching is sometimes referred to as “fuzzy matching” because it includes educated guesses, not exact matches. A scoring system helps software avoid matching records where the ambiguity is too high.

Great Matching Takes Both

In most complex matching scenarios, data scientists combine deterministic and probabilistic matching to make data merging decisions. The two methods complement one another. Laypersons may believe they should rely only on deterministic matches because it’s more of a sure thing, but they do not understand that probabilistic matching methods can add value to a deterministic-based task. Adding probabilistic methods expands the scope of the matching or consolidation project.

Consider a case where the primary match criterion is a data field well-suited to deterministic matching, such as an email address. If some data sources do not include email addresses for all records, deterministic- only routines might skip valuable information from that data source. Consequently, an organization might lose data such as internet browsing patterns or customer buying history, simply because the data records containing this information lacked an email address matching the master record.

By adding probabilistic matching, the software can compare several data elements, even if the data values vary, and match the records with an acceptable level of certainty. Important customer data will be retained, allowing the organization to use this information to enhance future customer experiences and run more effective marketing campaigns.

Best Practice Match/Consolidate Methods

The best match/consolidate software combines deterministic and probabilistic methods to provide the best possible performance while allowing users complete control over the matching process. Controls within the software allow for deterministic settings, such as:

1. Create simple match/no match rules
2. Define rules for what to do when data fields in one or both compared records are blank
3. Rank the records in a match group based on completeness of fields in the record
4. Specify a weighted score for each field
5. Set match vs. no match thresholds

Firstlogic uses probabilistic matching most often when comparing names of people or firms. Our Match/Consolidate software uses name aliases and cross-compares known alternate names. We also cross- compare first and middle names, knowing that Hubert James Smith is likely to call himself James Smith. We can even unscramble some names. If a data record listed a misspelled customer name as “James Anrold” the Firstlogic software would recognize the likely character transposition and score the match to “James Arnold” appropriately. Our probabilistic approach also removes noise words that can detract from the matching process, allowing us to identify all the viable matches.
Understanding how matching works is important in evaluating data quality requirements and selecting the right tools for the job. Employing both

deterministic and probabilistic matching methods ensures consistently superior results for data matching, duplicate recognition, and data consolidation.

Results

Often the result of a matching or consolidation effort is an output database cleansed of duplicates, but not always. Sometimes reports and statistics about the data are the desired outcome. On some occasions, users want to combine and consolidate data collected from multiple sources. In other cases, only a single data source is involved. Service providers often require two results—a clean data file, for instance, along with reports they can share with their clients describing the records they matched, dropped, or combined.

Output Files

Data files intended as input for further processes are often the product of match/merge projects. It is important to examine output data records and compare record counts to expected results before moving along to the next step in production. Deduped or consolidated record output files will contain a single occurrence of each record processed by the match/consolidate software. The software will exclude all the duplicates from the output file.

All-duplicate output files include master and subordinate records identified by the software as matches. Suppression records, if used in the job, are normally also included in all-duplicate output files.

Multi-occurrence output files include only those records found more than once in the input data, according to the multi-buyer or multi-occurrence criteria specified for the job. Only one record per individual appears in a multi-occurrence output file.

Duplicate Records Report

A common report produced by deduping software shows all the records the system identified as duplicates. Typically, the matching software prints each data record flagged as a duplicate, with the master record appearing as the first in each

group. These reports can validate the settings used to run the job or cause users to modify settings and try again.

Executive Summary Report

Some matching/consolidation software products produce summary reports that provide details about the number of input records, list distribution, data parsing statistics, matching, suppression lists, etc. Service providers often deliver summary reports to their customers rather than burdening them with complete details.

Job Summary Report

Job summary reports display the settings and parameters used during job execution, identifying fields used to construct match keys. Service providers should keep job summary reports as references for trouble-shooting, or to allow them to recreate settings for a future job. Job summary reports may also include logs showing each step in the process as it executed, and the resources the job accessed.

Other Reports

The matching software and the attributes of individual jobs determine the reports the software generates.

Matching/consolidation software may create reports helpful in validating results or making adjustments to settings. Matching software that does not produce enough reports leaves users in the dark when something goes wrong. Some reports, especially for complex processes using multiple lists, priorities, and criteria, are invaluable as tools to tweak the settings and produce the best results for the application.

Conclusion

Matching and consolidation activities go way beyond the “merge/purge” approaches companies have used to manipulate mailing lists for decades. Since data is driving your business, it makes sense to use your data in the most efficient

and effective ways. Sometimes that means you've got to clear out the deadwood that is making every data operation more difficult.

Besides cleaning up your data and ridding yourself of duplicate or conflicting information, matching and consolidation software is essential for any organization attempting to achieve a single customer view (SCV) and boost the customer experience. The obstacle to SCV has always been dealing with unconnected data silos.

Before they can make any progress on SCV, companies have to connect and simplify customer information stored in billing, sales, marketing, operations, customer service, and other isolated systems. This simply cannot be done without combining the data with matching and consolidation software tools. SCV is a big project.

Fortunately, matching and consolidation software offers companies interim benefits while they pursue their ultimate SCV goals. Resolved duplicates will benefit organizations in many ways, from lower document production and distribution costs to faster processing times. They won't need to invest in more data storage or network transmission bandwidth when they reduce the workload by eliminating duplicate or redundant data.

As we described in this paper, data matching effectiveness depends on data quality. We recommend organizations use tools like those included in Firstlogic's Data Quality Platform to standardize and enhance their data first with DataRight IQ, before embarking on complicated matching or data consolidation projects with Match IQ. But that doesn't mean you can't take advantage of the matching software's comprehensive array of features for simpler tasks like deduping files or processing suppression lists.

Firstlogic DQ10 includes all the tools an organization needs to handle any data situation, and the software has a long track record of success.

No-Fee Assessment

Firstlogic specializes in delivering data services solutions to data-driven companies. Firstlogic's products set the standard for address and data quality

software when first introduced in 1984. Many users of these products have been customers for more than 30 years, with good reason. Firstlogic's development and support professionals are highly acclaimed and are continuously innovating enhancements to the products, building on their stellar data parsing engine. This engine is acknowledged by many as the best in the business. To find out how software from Firstlogic can help your organization be more productive, accurate, and competitive, schedule a discovery call.

Gain insight into the health of your organization's data quality! We will process a sample set of your data using the latest data quality tools. Our system will find anomalies in your data and Identify opportunities for improvements.

When you request your no-fee data quality assessment you'll get a data profiling report showing data strengths and weaknesses, along with a custom assessment prepared by Firstlogic data experts.

Data is driving your business. Make sure you're getting the maximum benefit from the customer information you've collected and stored. Contact us today.

© Copyright Firstlogic Solutions, LLC All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of Firstlogic Solutions, LLC.

The information contained herein may be changed without prior notice.

Mover IQ, Sequence IQ and Workflow IQ are trademarks and Firstlogic, Firstlogic Solutions, FirstPrep, ACE, DataRight IQ, Match/Consolidate, PAF Manager and Data Quality. Delivered. are registered trademarks of Firstlogic Solutions, LLC.

The following trademarks are owned by the United States Postal Service: CASS, CASS Certified, DPV, RDI, eLOT, First-Class, DSF2, LACSLink, NCOALink, SuiteLink, USPS, U.S. Postal Service, United States Postal Service, United States Post Office, ZIP, ZIP+4, ZIP Code.