Cleaning up Very Large Databases and Keeping Them Clean

Priscilla Broberg
U.S. Caden (a division of Manpower)
currently working at Agilent Technologies

Executive Summary

This presentation shows a real-world example of how a very large Customer database was cleansed and de-duplicated to shrink it down to a manageable size. The techniques used to do this are shown, as well as the processes that were implemented to maintain the new level of data cleanliness. The tricks and techniques are applicable to customer files or databases of any size in any business. Actual before and after data examples are shown.

Topics covered include:
- Typical customer data flows, from data entry to reporting
- Proper placement of data cleansing and merging in the data flow
- Techniques to maximize effectiveness of merge/purge (de-duplication)
- Ideas for maintaining a higher level of data cleanliness, and minimizing data duplication.
Cleaning up Very Large Databases and Keeping Them Clean

The story of how a customer database got very large and very messy, then got small and clean again.

The Consolidation Point Provides Clean, De-Dupped Data to the Warehouse

- Cleanses data
- Standardizes data
- Enhances data (e.g. zip+4)
- Eliminates duplicates (merge/purge)
- Communicates back to transaction systems
  - rejected transactions
  - successfully loaded transactions

WHY DO WE NEED TO MERGE/PURGE CUSTOMER DATA?

- Data from separate transaction systems is entered and identified differently
- Need for company-wide view of customers ("Master list")
- Need to consolidate customer information Worldwide
  - avoid double counting
  - save on database storage
  - able to identify one customer with one unique identifier (cross-referenced to source systems)

Standardization Process Flow

MERGE/PURGE PROCESS FLOW
How Did We Get Into This Mess?

- ODS Database designed in late 1980’s to cleanse and load a single type of customer data - Order Processing Customers. Data only went to one application for reporting. ALL records were required to be loaded, regardless of data quality!
- Later, additional sources of data, as well as receiving applications were added for Direct Marketing. These were allowed to be rejected, if they did not meet data quality standards.
- Merge/Purge rules changed.
- Moved from Mainframe to Unix platform, and changed cleansing and merge/purge tools.
- ODS Database had no delete capability. All data was added or updated, then remained there forever!
- Only incoming transactions were cleansed and merge/purged against the database.
- Once data was loaded, it was never re-cleansed or re-merge/purged.

Other Contributing Factors

- Many records were coded with the wrong country code. Only those with US country codes (US, and affiliates such as Puerto Rico, Guam, etc) went through standardization!
- We have no edit for verifying the country code against the address. We just accepted what was input.
- Once a record is loaded as non-standard, it NEVER participates in the merge/purge.
- Non-standardized records contributed to a lot of data duplication.
- Split from Hewlett-Packard caused us to inherit a database full of HP customers as well as Agilent customers. There was no attribute of the customer data to tell them apart.

Preparing for The BIG CLEAN-UP

Step #1: Pre-clean up

- Removed data associated with Direct Marketing
  - Identified by Source Number
  - Had to make sure data was not also associated with active sources.
- Documented current Merge/Purge rules and reviewed with users
- Using a Marketing Reporting Tool (the Data Warehouse recipient of our customer data), we were able to identify which customers belonged to Agilent by reporting customer numbers on orders with Agilent product lines.
- Identified customers who had been active in the past two years, and deleted all others.
- Number of site (address records) after clean-up went from approximately 11 million rows to 1.1 million rows.
- This became the starting point for our re-standardization and re-merge/ purge.

Preparing for the BIG CLEAN-UP

Step #2: Analyze remaining data

- Determine how much data is US, how much Canada, and how much non-US.
- Country code not unreliable. However, we used ACE to discover this, and locate the incorrectly coded records!
- Perform test merge/ purge runs on non-US/non-Canada data, using line1, line2, etc. method.
- Adjust merge/purge parameters based on results of test runs.

Clean-up Steps

Data-Cleansing

- Country Code clean-up must be done first. Since this is part of the match-key, re-calculate match-key.
- Re-standardize US and Canada. After re-standardization, re-calculate match key again. (Postal code is also a component of match-key)
- Update database with new country codes and match-keys, as well as newly-standardized addresses.
- Our match-key algorithm: First letter of Business Name, followed by first four numbers of address, followed by first 3 bytes of postal code, followed by 3-byte country code. '000' used as filler where no data exists.
- Example: IBM 123 Main Street, Anytown, Anystate, 99999 would be coded as: 1123@999000 (‘000’ is our country code of US).

Example - Before Standardization

<table>
<thead>
<tr>
<th>ADDRESS</th>
<th>CITY</th>
<th>ZIP</th>
<th>STATE</th>
<th>PHONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>123 Main Street</td>
<td>Anytown</td>
<td>99999</td>
<td>Anystate</td>
<td></td>
</tr>
</tbody>
</table>

Proceedings of the Sixth International Conference on Information Quality
Country Codes

201 MEXICO
223 COSTA RICA
301 COLOMBIA
333 PERU
351 BRAZIL
359 URUGUAY
357 ARGENTINA
405 FINLAND
412 UNITED KINGDOM
427 FRANCE
428 GERMANY
489 TURKEY
549 KINGDOM OF THAILAND
583 TAIWAN

Example - After Country Code and Match Key Updates

Attn: Tims14 Bob Smith 1305 E Algonquin Rd Schaumburg IL 60196-4041 000 M1305601000
Accounts Payable 2732 E Miraloma Ave Anaheim CA 92806-1701 000 I2732928000
Ste 118 990 Richard Ave Santa Clara CA 95050-2828 000 W990@950000
Ste 625 10777 Westheimer Rd Houston TX 77042-3478 000 P1077770000
Attn: Iliana P. Velazquez 3009 NW 75th Ave Miami FL 33122-1439 000 M3009331000

Example - After Standardization

Attn: Tims14 Bob Smith 1305 E Algonquin Rd Schaumburg IL 60196-4041 000 M1305601000
Accounts Payable 2732 E Miraloma Ave Anaheim CA 92806-1701 000 I2732928000
Ste 118 990 Richard Ave Santa Clara CA 95050-2828 000 W990@950000
Ste 625 10777 Westheimer Rd Houston TX 77042-3478 000 P1077770000
Attn: Iliana P. Velazquez 3009 NW 75th Ave Miami FL 33122-1439 000 M3009331000

Clean-Up Steps

De-Dupping

• Using the updated match-codes from the prior clean-up steps, group data for merge/purging.
• Merge/purge US separately from Canada, and separately from non-US. This is because different merge/purge job files (rule sets) will be used for each.
• Manually inspect dup groups created by merge/purge prior to eliminating any data.
• The “mail” file from the merge/purge runs will represent the surviving sites. Use the “dups” file to create elimination transactions (if required by your system).

Excerpt From International Merge/Purge Job File

Accounts Payable 1301 E Algonquin Rd Schaumburg IL 60196-1078 428 M6019130428
Attn: Tims14 Bob Smith 1305 E Algonquin Rd Schaumburg IL 60196-4041 428 M1305130428
Accounts Payable 2732 E Miraloma Ave Anaheim CA 92806-1701 549 I2732273549
Ste 118 990 Richard Ave Santa Clara CA 95050-2828 583 W9505950583
Ste 625 10777 Westheimer Rd Houston TX 77042-3478 351 P@@@@107351
Ste 1050 5200 Blue Lagoon Dr Miami FL 33126-7008 333 A@@@@105333
Attn: Iliana P. Velazquez 3009 NW 75th Ave Miami FL 33122-1439 351 M@@@@331351

Sample Dups File (International)

Example - After Standardization

Attn: Tims14 Bob Smith 1305 E Algonquin Rd Schaumburg IL 60196-4041 000 M1305601000
Accounts Payable 2732 E Miraloma Ave Anaheim CA 92806-1701 000 I2732928000
Ste 118 990 Richard Ave Santa Clara CA 95050-2828 000 W9505950583
Ste 625 10777 Westheimer Rd Houston TX 77042-3478 000 P1077770000
Attn: Iliana P. Velazquez 3009 NW 75th Ave Miami FL 33122-1439 000 M3009331000
Clean-up Steps
Eliminating Duplicate Data

• If required to maintain record of eliminated data, use the dup groups to create elimination transactions. An elimination transaction is basically like a “change of address” transaction. All that is needed is the old address identifier and the new (surviving) address identifier.

• If this is NOT required, delete any addresses not in your “mail” file, and you are done!

Clean-up Steps
Eliminating Duplicate Data (cont)

• Steps to performing eliminations:
  – 1) Create elimination transactions from dup groups
  – 2) Apply eliminations to all tables in which the address identifier is used. For example, our database uses this identifier in X tables. Change old identifier to new identifier, based on transaction.
  – 3) Once all tables have been updated, create a row in an “elimination table” to keep track of this change (old ID --> new ID)
  – 4) Finally, delete old (eliminated) address record

Sample Elimination Transactions
(Created from Sample Dups File)

<table>
<thead>
<tr>
<th>Old ID</th>
<th>New ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>017588008</td>
<td>017508378</td>
</tr>
<tr>
<td>018679713</td>
<td>018660122</td>
</tr>
<tr>
<td>017626543</td>
<td>017037130</td>
</tr>
<tr>
<td>017784498</td>
<td>017037130</td>
</tr>
<tr>
<td>017818072</td>
<td>017037130</td>
</tr>
<tr>
<td>017907210</td>
<td>017037130</td>
</tr>
<tr>
<td>007083284</td>
<td>003488990</td>
</tr>
<tr>
<td>017023274</td>
<td>010279043</td>
</tr>
<tr>
<td>017095742</td>
<td>010279043</td>
</tr>
<tr>
<td>017819768</td>
<td>010279043</td>
</tr>
<tr>
<td>018408512</td>
<td>018180551</td>
</tr>
<tr>
<td>007083034</td>
<td>002223280</td>
</tr>
<tr>
<td>015338690</td>
<td>002223280</td>
</tr>
<tr>
<td>017140834</td>
<td>017038804</td>
</tr>
</tbody>
</table>

Lessons Learned

• It is important to understand your current data flow and processing. If you haven’t documented it thoroughly, start now!

• Make sure your users understand the data-cleansing and merge/purge rules. They own these!

• Know what data you have control over, and what data you do not. For example, we can clean-up data, but we cannot force the source systems to send us clean data.

• For best results, re-standardize all addresses in database whenever you get a new zip+4 update file from Firstlogic.

• Re-merge/purge entire database at least 4 times a year.

• Use Firstlogic tools to analyze your data, as well as to cleanse it in production.

• Don’t assume you cannot merge/purge non-US data. It can be done quite effectively using the user-definable fields (Merg_Purg1, Merg_Purg2, etc).

• Read the Firstlogic Software Update Bulletins and Customer Care Bulletins that come with your upgrades. There may be new features you can take advantage of!

Improvements/Benefits

• Reduced address rows in database from 11 million to < 2 million
  – Benefits:
    • Less disk space usage
    • Easier database administration
    • Faster processing times, as data merge/purges against fewer rows
    • Improved data quality, as duplicates are eliminated
    • Better decision making, as user confidence in data improved
    • Improved processing times on downstream systems, as less data is passed to them

Cost Savings

• Support went from 3 full-time programmers rotating on-call duty (24/7), to Call-center, with 1 on-call “deep support” programmer.

• Call-center support much less expensive

• Support programmers became available to work on new projects.

• Went from one full-time DBA to one part-time DBA.

• Lowered disk space costs

• Lowered processing (machine time) costs

• Estimated total annual savings: $500,000