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.

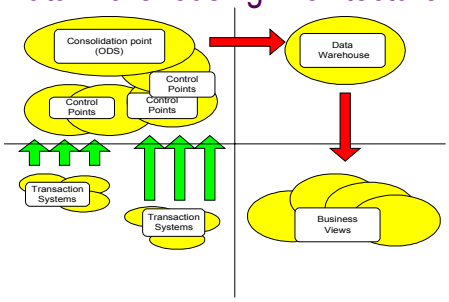
Topics covered include:

-
- 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.

Data Warehousing Architecture



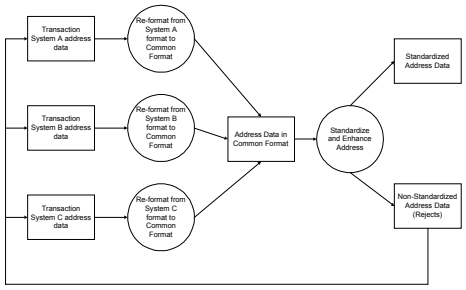
The Consolidation Point Provides Clean, De-Dupped Data to the Warehouse

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

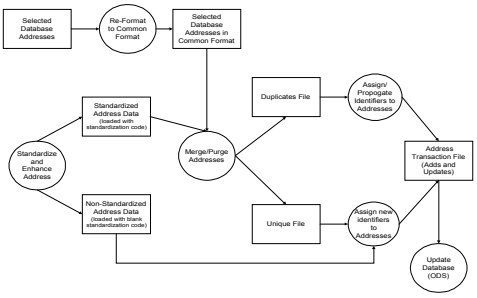
WHY DO WE NEED TO MERGE/PURGE CUSTOMER DATA?

- 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-calculcate 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. '@' used as filler where no data exists.
- Example: IBM 123 Main Street, Anytown, Anystate, 99999 would be coded as: I123@999000 ('000' is our country code of US).

Example -Before Standardization

3009 NW 75 AVENUE	MIAMI FLORIDA 33122	ATTN: ILIANA P. VELAZQUEZ	351 M0000331351
7200 NW 7 STREET 2ND FLOOR	MIAMI FLORIDA 33126		351 M0000331351
C/O TRANEXCO 105295	7801 NW 37 STREET	MIAMI, FL. 33166-6559	301 I3316@@@301
14413 IMPORT DR.	LAREDO,TX 78041 USA		201 N7804144201
C/O MIAMI PANALPINA INT	3505 N.W. 107TH AVE.	MIAMI, FL 33178	355 C3317331355
5200 BLUE LAGOON DRIVE SUITE 1050	MIAMI FLORIDA 33126		333 A@@@@105333
10777 WESTHEIMER STE 625	HOUSTON TX 77042		351 P@@@@107351
1900 CONCOURSE DRIVE	SAN JOSE, CA. 95131		223 R@@@@190223
420-B PANAMERICAN DRIVE	EL PASO, TX 79907		201 M7990799201
2629 TERMINAL BLVD.	MOUNTAIN VIEW, CA 94043		357 E@@@@262357
400 REIMANN AVENUE	SANDWICH IL 60548-0900		412 C6054@@@412
1330 MEMOREX DRIVE	SANTA CLARA, CA 95050		583 W@@@@133583
SUITE 118	990 RICHARD AVENUE	SANTA CLARA CA 95050	583 W9505950583
	6900 MAIN STREET	STRATFORD, CT 06615-9129	489 S6900@@@489
2732 EAST MIRALOMA AVE	ANAHEIM, CA 92806		549 I@@@@273549
ACCOUNTS PAYABLE	2732 E MIRALOMA AVENUE	ANAHEIM CA 92806	549 I2732273549
1101 CYPRESS CREEK ROAD	CEDAR PARK TX 78613		405 E7861786405
6781-R SIERRA COURT	DUBLIN, CA 94568		427 T@@@@945427
6780-R SIERRA COURT	DUBLIN CA 94568		427 T@@@@945427
ATTN: TIM814 BOB SMITH	1305 E. ALGONQUIN ROAD	SCHAUMBURG IL 60196	428 M1305130428
ACCOUNTS PAYABLE	1301 E ALGONQUIN ROAD	SCHAUMBURG IL 60196	428 M6019130428

Country Codes

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

Example - After Standardization

Attn: Iliana P. Velazquez	3009 NW 75th Ave	Miami	FL	33122-1439	351	M0000331351
	7200 NW 7th St	Miami	FL	33126-2941	351	M0000331351
	7801 NW 37th St	Miami	FL	33166-6503	301	I3316000301
	14413 Import Rd	Laredo	TX	78045-7947	201	N7804144201
C/O Miami Panalpina Int	3505 NW 107th Ave	Miami	FL	33178-1889	355	C3317331355
Ste 1050	5200 Blue Lagoon Dr	Miami	FL	33126-7008	333	A0000105333
Ste 625	10777 Westheimer Rd	Houston	TX	77042-3478	351	P0000107351
	1900 Concourse Dr	San Jose	CA	95131-1719	223	R0000190223
	420 Pan American Dr	El Paso	TX	79907-5637	201	M7990799201
	2629 Terminal Blvd	Mountain View	CA	94043-1131	357	E0000262357
	400 Reimann Ave	Sandwich	IL	60548-1866	412	C6054000412
	1330 Memorex Dr	Santa Clara	CA	95050-2853	583	W0000133583
Ste 118	990 Richard Ave	Santa Clara	CA	95050-2828	583	W9505950583
	6900 Main St	Stratford	CT	06614-1385	489	S6900000489
	2732 E Miraloma Ave	Anaheim	CA	92806-1701	549	I0000273549
Accounts Payable	2732 E Miraloma Ave	Anaheim	CA	92806-1701	549	I2732273549
	1101 Cypress Creek Rd	Cedar Park	TX	78613-3615	405	E7861786405
	67818 Sierra Ct	Dublin	CA	94568-2611	427	T0000945427
Ste R	6780 Sierra Ct	Dublin	CA	94568-2600	427	T0000945427
Attn: Tims14 Bob Smith	1305 E Algonquin Rd	Schaumburg	IL	60196-4041	428	M1305130428
Accounts Payable	1301 E Algonquin Rd	Schaumburg	IL	60196-1078	428	M6019130428

Example - After Country Code and Match Key Updates

Attn: Iliana P. Velazquez	3009 NW 75th Ave	Miami	FL	33122-1439	000	M3009331000
	7200 NW 7th St	Miami	FL	33126-2941	000	M7200331000
	7801 NW 37th St	Miami	FL	33166-6503	000	I7801331000
	14413 Import Rd	Laredo	TX	78045-7947	000	N1441780000
C/O Miami Panalpina Int	3505 NW 107th Ave	Miami	FL	33178-1889	000	C3505331000
Ste 1050	5200 Blue Lagoon Dr	Miami	FL	33126-7008	000	A5200331000
Ste 625	10777 Westheimer Rd	Houston	TX	77042-3478	000	P1077770000
	1900 Concourse Dr	San Jose	CA	95131-1719	000	R1900951000
	420 Pan American Dr	El Paso	TX	79907-5637	000	M4209799000
	2629 Terminal Blvd	Mountain View	CA	94043-1131	000	E2629940000
	400 Reimann Ave	Sandwich	IL	60548-1866	000	C4006605000
	1330 Memorex Dr	Santa Clara	CA	95050-2853	000	W1330950000
Ste 118	990 Richard Ave	Santa Clara	CA	95050-2828	000	W9909950000
	6900 Main St	Stratford	CT	06614-1385	000	S6900066000
	2732 E Miraloma Ave	Anaheim	CA	92806-1701	000	I2732928000
Accounts Payable	2732 E Miraloma Ave	Anaheim	CA	92806-1701	000	I2732928000
	1101 Cypress Creek Rd	Cedar Park	TX	78613-3615	000	E1101786000
	67818 Sierra Ct	Dublin	CA	94568-2611	000	T6781945000
Ste R	6780 Sierra Ct	Dublin	CA	94568-2600	000	T6780945000
Attn: Tims14 Bob Smith	1305 E Algonquin Rd	Schaumburg	IL	60196-4041	000	M1305601000
Accounts Payable	1301 E Algonquin Rd	Schaumburg	IL	60196-1078	000	M1301601000

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

```
BEGIN  Match Criteria:        KLen BPos BLen MPos MLen Match BlnkP BlnkM
Pre_Name..................... =    0    1    0    1    0    n    0   no
First_Name................... =    0    1    0    1    0    n    0   no
Mid_Name..................... =    0    1    0    1    0    n    0   no
Last_Name.................... =    0    1    0    1    0    n    0   no
Post_Name.................... =    0    1    0    1    0    n    0   no
Title........................ =    0    1    0    1    0    n    0   no
Firm......................... =    0    1    0    1    0    n    0   no
Street Range................. =    0    1    0    1    0    n    0   no
Street Pre-directional ...... =    0    1    0    1    0    n    0   no
Street Primary Name.......... =    0    1    0    1    0    n    0   no
Street Suffix................ =    0    1    0    1    0    n    0   no
Street Post-directional ..... =    0    1    0    1    0    n    0   no
Street Secondary Range ...... =    0    1    0    1    0    n    0   no
PO Box....................... =    0    1    0    1    0    n    0   no
Rural Route Number........... =    0    1    0    1    0    n    0   no
Rural Route Box.............. =    0    1    0    1    0    n    0   no
City......................... =    0    1    0    1    0    n    0   no
State........................ =    0    1    0    1    0    n    0   no
ZIP.......................... =    0    1    0    1    0    n    0   no
ZIP+4........................ =    0    1    0    1    0    n    0   no
Country...................... =    0    1    0    1    0    n    0   no
Merg_Purg1................... =   40    1    5    1   40    e    0   no
Merg_Purg2................... =   40    1    5    1   40    1    0   no
Merg_Purg3................... =   40    1    5    1   40    t    0   no
Merg_Purg4................... =   40    1    5    1   40    t    0   no
Merg_Purg5................... =   40    1    5    1   40    t    0   no
Merg_Purg6................... =   11    1   11    1   11    e    0   no
```

Sample Dups File (International)

```
MAST0ZUR METTERMARTE 50                 01109 DRESDEN                    017508378 428 A0110####428
MAST ZUR METTERMARTE 50                 01109 DRESDEN                    017588008 428 A0110####428
MAST0VITERBO                            I 01100 VITERBO VT               018660122 475 A0110####475
MAST VITERBO                            I 01100 VITERBO VT               018679713 475 A0110####475
MAST0WILSCHDORFER LANDSTR. 101          01139 DRESDEN                    014268700 428 A0113####428
MAST WILSCHDORFER LANDSTR. 101          01139 DRESDEN                    014707877 428 A0113####428
MAST0NOTIFY KYOTO ELECTRO SINGAPORE PTE LTD 00 12 TANJONG PENJURU CRESCENT UBTS 017037130 559 A0120####559
MAST NOTIFY KYOTO ELECTRO SINGAPORE PTE LTD 00 12 TANJONG PENJURU CRESCENT UBTS 017626543 559 A0120####559
MAST NOTIFY KYOTO ELECTRO SINGAPORE PTE LTD 00 12 TANJONG PENJURU CRESCENT UBTS 017784498 559 A0120####559
MAST NOTIFY KYOTO ELECTRO SINGAPORE PTE LTD 00 12 TANJONG PENJURU CRESCENT UBTS 017818073 559 A0120####559
MAST NOTIFY KYOTO ELECTRO SINGAPORE PTE LTD 00 12 TANJONG PENJURU CRESCENT UBTS 017907210 559 A0120####559
MAST0MARSCHNERSTR. 26                   01307 DRESDEN                    003488990 428 A0130####428
MAST MARSCHNERSTR. 26                   01307 DRESDEN                    007083284 428 A0130####428
MAST0BERTOLD-BRECHT-ALLEE 22            01309 DRESDEN                    010279043 428 A0130####428
MAST BERTOLD-BRECHT-ALLEE 22            01309 DRESDEN                    017023274 428 A0130####428
MAST BERTOLD-BRECHT-ALLEE 22            01309 DRESDEN                    017095742 428 A0130####428
MAST BERTOLD-BRECHT-ALLEE 22            01309 DRESDEN                    017819768 428 A0130####428
MAST0Berthold-Brecht-Allee 22           01309 Dresden                    018188551 428 A0130####428
MAST Berthold-Brecht-Allee 22           01309 Dresden                    018408512 428 A0130####428
MAST0KYFFHAEUSERSTR. 27                 01309 DRESDEN                    002223280 428 A0130####428
MAST KYFFHAEUSERSTR. 27                 01309 DRESDEN                    007083034 428 A0130####428
MAST KYFFHAEUSERSTR. 27                 01309 DRESDEN                    015338690 428 A0130####428
MAST0POSTFACH 11 01 10                  01330 DRESDEN                    017038804 428 A0133####428
MAST POSTFACH 11 01 10                  01330 DRESDEN                    017140834 428 A0133####428
```

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)

Old ID	New ID
017588008	017508378
018679713	018660122
014707877	014268700
017626543	017037130
017784498	017037130
017818072	017037130
017907210	017037130
007083284	003488990
017023274	010279043
017095742	010279043
017819768	010279043
018408512	018180551
007083034	002223280
015338690	002223280
017140834	017038804

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