A MELISSA DATA PUBLICATION
05/11
Volume 6, Issue 3

Data Quality INSIDER

Strategic Data Quality Solutions for Developers

< IN THE SPOTLIGHT >

The Dupe Detective Part 1

An overview of Fuzzy Matching in SSIS

The process of finding approximate matching records in your data to get rid of duplicates is precisely that – fuzzy. It raises as many questions as answers. Am I using a good matching algorithm? Am I matching on the right fields? Is it a true match or a false one?

The problem begins when inconsistent data enters from multiple sources. The meticulous process of finding these similar records and comparing to see if they are actually the same is a daunting challenge.

But with the release of the Melissa Data Fuzzy Matching Component for SQL Server Integration Services (SSIS), you now have a tool that will make this all elementary. With this component, you become a Sherlock Holmes – easily cracking the case of the data doppelgangers.

Finding the Culprits

Matching duplicate records are identified through a percent score. Compared records will be given a match score ranging from 0% (non-matching) to 100% (exact match). So what about records that score in between?

continued on page ***3***

DATA QUALITY TOOLS

- Address Object
- Contact Verification Server
- < Data Quality Components for SSIS >
- Data Quality Suite
- Email Object
- GeoCoder Object
- IP Locator Object
- MatchUp Object
- Name Object
- Phone Object
- Presort Object
- RightFielder Object
- SmartMover Web Service
- Telco SmartSearch Web Service
- WebSmart Services
- WebSmart Global

The Insider Pro

We recently featured a webinar on how to use Melissa Data's Data Quality Components for SSIS to perform data profiling and fuzzy matching. Our BI architect Ira Whiteside demonstrates how you can gain an aggregate understanding of your contacts by profiling, consolidating and linking data from all sources of your organization. Also, we featured a new video detailing how you can create a metadata mart to develop a strong MDM program. These programs are now available on-demand. To view, go to:
www.MelissaData.com/shows

Bud Walker
Director,
Data Quality Solutions

Case Study >> PCI Utilizes WebSmart to Streamline Operations

Company

Headquartered in Dallas, Texas, PCI publishes printed alumni directories and offers ENGAGE – the company's best-in-class online community software and online marketing and management services.

Challenge

A key component of PCI's services is the collection and management of its clients' contact data, which includes mailing addresses, phone numbers, email addresses, and other information such as names of spouses, children, and educational and professional degrees. So every name, number, letter and digit of its clients' contact data has to be correct when PCI reaches out to them.

Solution

WebSmart is a suite of Web services that verifies addresses, phone numbers, email addresses. WebSmart also geocodes addresses, identifies IP location, and performs change-of-address updates.

Scaling its Business Operations to Improve Efficiency

In the past, PCI conducted its contact verification by calling individuals listed on its clients' contact lists, and also by sending postcards to membership lists, with subsequent inbound calls going to PCI's call centers for confirmation and correction of data. Each updated record then went to PCI's editing team to manually key the updated data into the company's database. This proved to be a very labor-intensive process, especially as PCI began experiencing significant growth.

How PCI Uses WebSmart

PCI utilizes Melissa Data's WebSmart for real-time verification of contact data to support its clients' campaigns for alumni engagement and outreach.

Rebecca Cap Mertens, PCI's vice president of operations, said the decision to select Melissa Data's WebSmart was a natural one. WebSmart is seamlessly integrated into PCI's in-house application. The company uses a Microsoft SQL 2000 database and Microsoft operating system.

"We were looking for a tool that was CASS Certified™ and offered a Web service component to allow us to validate submissions in real-time," Mertens said. "Melissa Data has proven itself as a top-tier provider of address validation for high volume mail and call-center environments."

Real-Time Data Verification

While PCI's staff knows whom they're contacting when making outbound calls, that information isn't at their fingertips when calls come in to PCI. But, by utilizing WebSmart, callers can provide an ID number from their postcard to the PCI agent – and by entering that number in the system, the PCI agent can pull up the records to confirm and update the information in real-time.

The Results

By using WebSmart, 75 percent of all address updates obtained by PCI's agents are verified, validated and standardized at the point of entry in real-time. On a weekly basis, PCI validates the contact data of 15,000 to 18,000 individuals who respond to marketing outreach campaigns, and cleanses data in new files streaming into the company – for a total of 55,000 names per week.

"We were able to double the number of address updates on a weekly basis and reduce staffing needs from 5.5 people down to three, and increase the number of records handled by 40 percent in the last six months," Mertens said.

< IN THE SPOTLIGHT >

continued from ***Page 1***

The Dupe Detective...

By leveraging the ETL capabilities of SSIS, the Fuzzy Matching Component allows you to send the results through three different output destinations: Match, Non-Match; and Possible Match. Based on how strict or loose you set our thresholds to be, records will be redirected to the output tables accordingly – making the job of keeping the bad records out much easier.

More Brains are Better Than One

So how exactly does the Fuzzy Matching Component determine match percentages? Similarity computation is done through built-in fuzzy matching algorithms. But why settle with just one algorithm for similarity computation when you can have 16!

Available algorithms include common algorithms like the Jaro and Levenstein, but also includes other more advanced algorithms such as Smith-Waterman-Gotoh and PhonetEx. Why so many you might ask? Each algorithm has its own strengths and weaknesses.

Some algorithms are more accurate when it comes to company and peoples' names, while some are more effective when it comes to company names. But with the wide array of algorithms to choose from, you have the flexibility to choose the logic that works for your data.

The Seven Features of the Fuzzy Matching Component

The SSIS component includes these features:

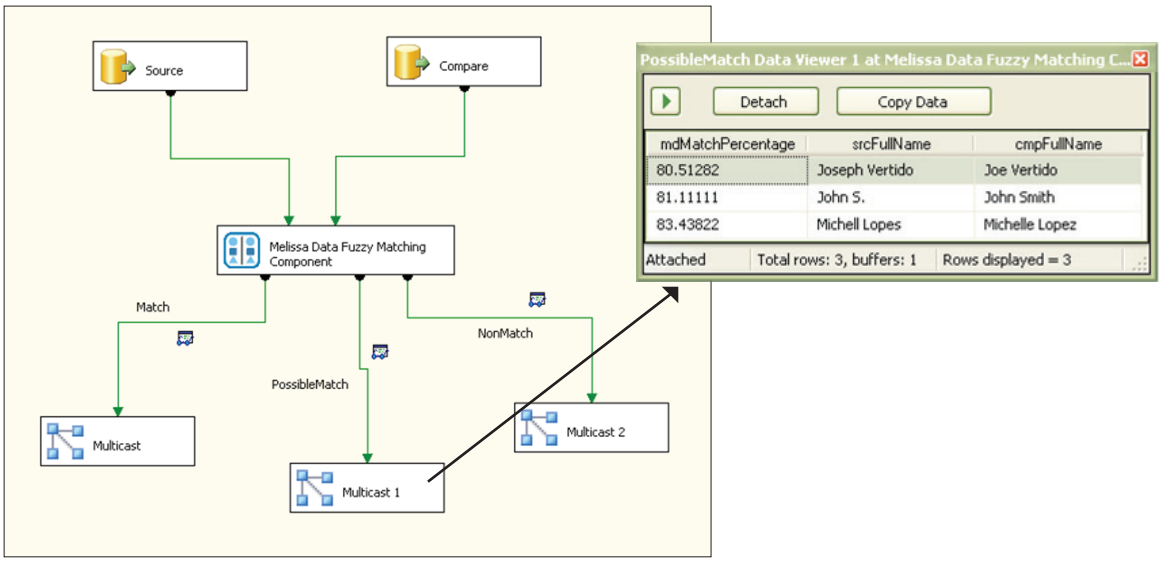
1) Match based on several columns in your data
2) Get actual Match Score Percentages
3) Multiple fuzzy matching algorithms
4) Automatic filtering of matching, non-matching, and possible matching records
5) Built-in pre-cleansing through search and replace patterns
6) Data Driven Model for easy migration to production
7) Matching Metadata for data driven decision making

It's Elementary, My Dear Watson

In what seems to be an impossible task of finding and blocking dupe records, Melissa Data aims to help you put the puzzle pieces together and help solve the mysteries of fuzzy matching. With the Melissa Data Fuzzy Matching Component for SSIS, you're now one step closer to making your data problem-free.

To be continued:
In the next issue, we'll take a deep dive into Matching examples.

For more info, go to: www.MelissaData.com/dqi-ssis



News Bytes

>> New Simpler Smart Mover

Melissa Data's SmartMover is now simpler and faster. You will no longer have to use the Ticket system to request a ticket before processing records. Now, you can simply call the SmartMover Web service with a batch of records, just like our current WebSmart line of Web services.

Additionally, we have made several improvements and optimizations under the hood to speed up processing time. We have enabled gzip compression – a simple, effective way to save bandwidth by making the transmission size smaller. This means SmartMover will now be able to process larger files faster.

If you are currently using the older SmartMover, we encourage you to upgrade to the new version right away. Converting your code to use the new version is quick and easy and will give you immediate benefits.

For more information, go to: www.MelissaData.com/tech/tech.html

>> Understanding RightFielder Object Functionality

RightFielder Object standardizes inconsistently entered data, run-on lines of text containing multiple data types, or data with no structure at all using two types of sophisticated parsing engines. Here are more details on how each engine works:

Fielded Data – This parsing logic takes data already in certain fields, and places the data in the right fields for better accuracy and standardization. You have the flexibility to specify, identify and re-field custom data types such as SSNs, IPs, email addresses, URLs, dates or proprietary account number formats.

	Column 1	Column 2	Column 3	Column 4	Column 5
1	Joy Pars	12345 Main St	joe@aol.com	949-121-3323	91745
2	9823 Collette Ave	10098	232-132-2345	Kim Jones	kim@yahoo.com
3	bill@gmail.com	800-800-1234	2238 Viam Way	92688	Bill Warner

	Name	Address	Zip Code	Phone	Email
1	Joy Pars	12345 Main St	91745	949-121-3323	joe@aol.com
2	Kim Jones	9823 Collette Ave	10098	232-132-2345	kim@yahoo.com
3	Bill Warner	2238 Viam Way	92688	800-800-1234	bill@gmail.com

FreeForm Data – This function analyzes free form or unfielded textual data. It will identify which parts of the input data are a name, address, company, phone number. etc. - and then place them in separate correct output properties.

Bud M Walker 123 Main St Mission Viejo CA 92691 bud@melissadata.com 9495895200 Melissa Data

Name 1:	Bud M Walker
Company 1:	Melissa Data
Address:	123 Main St
Address2:	
Address3:	
City:	Mission Viejo
State:	CA
Postal Code:	92691
Country:	
Dept1:	
Phone1:	9495895200
Email1:	bud@melissadata.com
URL1:	
SSN1:	

For more information, go to: www.MelissaData.com/dqi-right

Helpful Links

Melissa DATA Resources
Link to our white papers, case studies, and insightful articles to help you get the most out of your direct mail, marketing, and data management initiatives.
www.MelissaData.com/resources

Discussion Forums
Ask questions, talk product, and get fast answers in our discussion forums.
http://forum.MelissaData.com

Free Trials

Identify geographic location of an IP address. Get a free trial of IP Locator.
www.MelissaData.com/dqi-ip

Verify U.S./Canadian phone numbers. Request a free trial of Phone Object.
www.MelissaData.com/dqi-phone

Identify the gender makeup of your contact data. Download a free trial of Name Object.
www.MelissaData.com/dqi-name

Director, Data Quality Solutions
Bud Walker
bud@melissadata.com
1-800-635-4772 x159

Melissa Data Technical Support
tech@melissadata.com
1-800-635-4772 x4 (6 am to 5 pm PST)

Editor Abby Telleria
Writers Bud Walker, Admound Chou, Tim Sidor, Joseph Vertido
Art Director Melody Yen
Graphic Designer Timothy Magoun

Contact the editor at insider@melissadata.com

Melissa Data Corp.
22382 Avenida Empresa
RSM California, 92688-2112
Ph 1-800-MELISSA (635-4772)
Fax 949-589-5211
www.MelissaData.com

© 2011 Melissa Data Corp. All rights reserved.

MELISSA DATA®
Your Partner in Data Quality