Contact Zone®

"Given the wide spectrum of quality associated with the provided location addresses, [Contact Zone's] 92% verification [rate] was impressive and sufficient for our analysis."

- David Loshin
President, Knowledge Integrity

Massive Address Reconciliation Using Contact Zone®

Introduction

One might think that the explosion of electronic commerce would have reduced the demand for information about physical locations. On the contrary, increased e-commerce has increased the need for precise and accurate location data. This need is particularly acute in the context of parcel delivery as the fulfillment process for items ordered online can only be completed once the package is consigned to the proper recipient.

In general, item delivery is complicated by reliance on the interested parties (either the sender, the recipient, or both) to provide accurate delivery point information. However, there are natural tendencies for variance in presentation of location data, and a byproduct is inconsistency in capture of standardized representations of locations and addresses. This article is a case study of a project involving pattern analysis of deliveries to specific locations, and demonstrates how Melissa Data's Contact Zone product was used in support of this analysis.

Challenge

The business client presented a large volume (on the order of 500,000,000) of business transaction records, each of which associated with a specific location within the U.S. Without loss of generality, we can abstract the analysis to counting the numbers of times that particular types of transactions occurred at each unique location.

These key aspects of the challenge had a cumulative impact on the ability to generate the expected analytical results:

- Data volume: The sheer volume of transaction records was the most obvious barrier to success, necessitating a platform with sufficient storage capacity and computational power to preprocess the transaction records and reorganize them in a way that was suited to the aggregate analysis.

- Data variation: The variety of representations associated with customer-presented location address information increased in a way that was proportional to the data volumes. At the same time, the breadth of delivery locations across

About Knowledge Integrity

Since 1999, Knowledge Integrity, Inc. has developed technical and management methodologies for instituting Data Quality, Master Data Management, Data Standards, and Data Governance programs within organizations to enable the analysis, assessment, and improvement of data quality for transactional systems, business intelligence, operational, and reporting purposes. They have provided services to many different organizations, both public and private sector, in many different industries, including Finance, Banking, Insurance, Health Care, Manufacturing, Pharmaceuticals, and Government agencies.

www.knowledge-integrity.com

the U.S. created the scenario for a long-tail effect, in which a majority of the locations had a relatively small number of transaction occurrences. That created need for address location reconciliation as a means of resolving variant address spellings and representations to get accurate counts for each address.

- Location Precision: There were many differences in the precision of provided addresses. For example, records for deliveries to suites or apartments in large buildings in urban locations often had different designations for the unit (including “APT,” “UNIT,” “SUITE,” “STE,” among others), sometimes just annotated with the floor (such as “7th FL”), or sometimes the unit designation was completely missing. Other precision issues included missing or incorrect postal codes, the use of vanity names, and state code mismatches.

Accuracy in the analytical results clearly depended on a method for standardizing massive numbers of addresses consistently, while executing at a level of performance that scales in the presence of massive amounts of data.

“Contact Zone blends the Melissa Data technology for address parsing, standardization, and cleansing of addresses with an intuitive user interface for specifying sources, outputs, fields to include in the parsing and standardization process.”

Approach

The computing platform was a high-end desktop computer running Windows 8, containing a 4-core i7-3770 CPU with a maximum speed of 3.7 GHz, 4TB disk space, and a 60 GB Solid State Disk (SSD) drive. We augmented the system by adding additional disk drives (one 3TB and one 4TB) for data and backup space, as well as an additional 256 GB SSD drive to improve runtime data access speed.

While we could have used one of the newer “big data” NoSQL data management frameworks, we opted for the simplicity of MySQL, as it remains suited for both data capture and for devising a data mart model suited to analysis. Our data tables were partitioned and then indexed by state, designed to speed analytical queries via a “divide and conquer” methodology.

Most importantly, in accordance with industry best practices, our approach was to employ a tool certified by the United States Postal Service® (USPS®) for address standardization and cleansing. Because the aggregation was by unique location, it was important to use a product that both standardized addresses and provided enough additional information to link records that share a representation of the same actual location and reconcile them together.

Solution – Contact Zone

We reviewed a number of candidate vendor products and selected Contact Zone by Melissa Data. Contact Zone blends the Melissa Data technology for address parsing, standardization, and cleansing of addresses with an intuitive user interface for specifying sources, outputs, fields to include in the parsing and standardization process. Contact Zone has a rich and diverse connectivity capability and can access a wide variety of data sources ranging from text files to a broad inventory of database management systems.

Contact Zone uses data from the USPS for address verification, and either validates that a standardized address is deliverable or that it is flawed and does not verify. In the cases where the address does not validate, Contact Zone attempts to use a knowledge base to adjust the address and manipulate it into a “correct” form. Some examples include appending apartment or suite numbers, modifying the city name, or correcting street directional such as “N” or “W.”

Contact Zone: Performance and Throughput

Contact Zone is coupled with a data integration capability that streams source data in a rapid and efficient manner. The considerations for accessing source data from disk suggest a potential performance lag while waiting for the data to be streamed into memory.

Our use of an additional SSD certainly helped in providing an additional level of caching for data. At the same time, Contact Zone performance is tuned to use all available processing capability, which proved to especially take advantage of the eight virtual processors provided by the i7-3770 CPU.

Our data set was broken out into approximately one hundred files, each containing between 4 and 5 million records. In one example, Contact Zone processed a file containing 4,577,000 records in 55 minutes, for an effective speed of approximately 1390 records per second, which is an effective rate of 5,000,000 records per hour, a noticeably respectable throughput on our selected desktop computer.

“Contact Zone processed a file containing 4,577,000 records in 55 minutes, for an effective speed of approximately 1390 records per second, which is an effective rate of 5,000,000 records per hour, a noticeably respectable throughput on our selected desktop computer.”

Because of Contact Zone’s ability to take advantage of the multi-core CPU’s multithreading capability, and because of the data independence across different source files, we can presume linear scalability simply by adding additional multi-core CPUs. In fact, at some point during the project we replicated our production system with a second desktop platform, and by running both machines simultaneously, we achieved an effective rate of 10,000,000 address records processed per hour.

From a qualitative perspective, Contact Zone was generally effective in parsing, standardizing, and validating address data. With the out-of-the-box configuration for address standardization, and looking at the same sample file referenced above, 87% of the addresses were accurately resolved to a known deliverable USPS address.

An additional 5% of the total records were “partially verified” in that they could be uniquely identified, although they may have missed one address component such as an apartment or suite number.

These percentages remained generally consistent across the set of source data files. Given the wide spectrum of quality associated with the provided location addresses, this 92% verification was impressive and sufficient for our analysis.

Considerations for Linkage and Reconciliation

As described, the objective of this aggregate analysis critically depended on unique identification of addresses, and while Contact Zone’s output provided numerous data element values, we were able to select a minimal set of data elements for unique address resolution.

First, for each address, Contact Zone generates an Address Key; for verified addresses, the Address Key uniquely identifies the deliverable address. In some cases, though, this Address Key must be augmented with additional information, such as when there are multiple suites at the same address.

In our process, we used the combination of Address Key, Suite Number, and Private Mailbox Number. The Private Mailbox Number is associated with a location that collects mail on behalf of multiple individuals, such as the front desk for a virtual office space that provides office services to many customers; that mail reception location itself may be a suite in a multi-tenant building, necessitating additional precision for unique address resolution.

Results

The measured throughput performance demonstrated that the application is finely tuned to take advantage of available computing resources, and the precision and accuracy of the product was sufficient to address the specific analytical requirements for addressing the business problem.

By applying Contact Zone's address parsing, standardization, and enhancement, we were able to utilize its output for resolving multiple address representations associated with unique real locations. This enabled greater precision and level of trust in the results of our aggregate analysis.

About Melissa Data

Melissa Data is a leading provider of data quality, marketing and mailing solutions. Melissa Data helps companies acquire and retain customers, validate and enhance data, improve marketing ROI, and save money on postage and mail processing. Since 1985, Melissa Data has helped companies like Mercury Insurance, Xerox, Disney, AAA, and Nestle improve customer communications.

Melissa DATA
22382 Avenida Empresa
Rancho Santa Margarita
California 92688

1-800-MELISSA
P 949-858-3000
F 949-589-5211

www.MelissaData.com