A Melissa Data White Paper

Data Quality

Six Steps to Managing Data Quality with SQL Server Integration Services

MELISSA DATA®

Six Steps to Managing Data Quality with SQL Server Integration Services (SSIS)

Introduction

A company's database is its most important asset. It is a collection of information on customers, suppliers, partners, employees, products, inventory, locations, and more. This data is the foundation on which your business operations and decisions are made; it is used in everything from booking sales, analyzing summary reports, managing inventory, generating invoices and forecasting. To be of greatest value, this data needs to be up-to-date, relevant, consistent and accurate – only then can it be managed effectively and aggressively to create strategic advantage.

Unfortunately, the problem of bad data is something all organizations have to contend with and protect against. Industry experts estimate that up to 60 percent or more of the average database is outdated, flawed, or contains one or more errors. And, in the typical enterprise setting, customer and transactional data enters the database in varying formats, from various sources (call centers, web forms, customer service reps, etc.) with an unknown degree of accuracy. This can foul up sound decision-making and impair effective customer relationship management (CRM). And, poor source data quality that leads to CRM project failures is one of the leading obstacles for the successful implementation of Master Data Management (MDM) – where the aim is to create, maintain and deliver the most complete and consolidated view from disparate enterprise data.

The other major obstacle to creating a successful MDM application is the difficulty in integrating data from a variety of internal data sources, such as enterprise resource planning (ERP), business intelligence (BI) and legacy systems, as well as external data from partners, suppliers, and/or syndicators. Fortunately, there is a solution that can help organizations overcome the complex and expensive challenges associated with MDM – a solution that can handle a variety of data quality issues including data deduplication; while leveraging the integration capabilities inherent in Microsoft's SQL Server Integration Services (SSIS 2005/2008) to facilitate the assembly of data from one or more data sources. This solution is called Total Data Quality.

The Six Steps to Total Data Quality

The primary goal of an MDM or Data Quality solution is to assemble data from one or more data sources. However, the process of bringing data together usually results in a broad range of data quality issues that need to be addressed. For instance, incomplete or missing customer profile information may be uncovered, such as blank phone numbers or addresses. Or certain data may be incorrect, such as a record of a customer indicating he/she lives in the city of Wisconsin, in the state of Green Bay.

Setting in place a process to fix these data quality issues is important for the success of MDM, and involves six key tasks: **profiling, cleansing, parsing/standardization, matching, enrichment,** and **monitoring.** The end result – a process that delivers clean, consistent data that can be distributed and confidently used across the enterprise, regardless of business application and system.

1. Profiling

As the first line of defense for your data integration solution, profiling data helps you examine whether your existing data sources meet the quality standards of your solution. Properly profiling your data saves execution time because you identify issues that require immediate attention from the start – and avoid the unnecessary processing of unacceptable data sources. Data profiling becomes even more critical when working with raw data sources that do not have referential integrity or quality controls.

There are several data profiling tasks: column statistics, value distribution and pattern distribution. These tasks analyze individual and multiple columns to determine relationships between columns and tables. The purpose of these data profiling tasks is to develop a clearer picture of the content of your data.

- **Column Statistics –** This task identifies problems in your data, such as invalid dates. It reports average, minimum, maximum statistics for numeric columns.
- **Value Distribution –** Identifies all values in each selected column and reports normal and outlier values in a column.
- **Pattern Distribution –** Identifies invalid strings or irregular expressions in your data.

2. Cleansing

After a data set successfully meets profiling standards, it still requires data cleansing and de-duplication to ensure that all business rules are properly met. Successful data cleansing requires the use of flexible, efficient techniques capable of handling complex quality issues hidden in the depths of large data sets. Data cleansing corrects errors and standardizes information that can ultimately be leveraged for MDM applications.

3. Parsing and Standardization

This technique parses and restructures data into a common format to help build more consistent data. For instance, the process can standardize addresses to a desired format, or to USPS® specifications, which are needed to enable CASS Certified™ processing. This phase is designed to identify, correct and standardize patterns of data across various data sets including tables, columns and rows, etc.

Here is one scenario:

A customer of a hotel and casino makes a reservation to stay at the property using his full name, Johnathan Smith. So, as part of its customer loyalty-building initiative, the hotel's marketing department sends him an email with a free night's stay promotion, believing he is a new customer – unaware that the customer is already listed under the hotel's casino/gaming department as a VIP client – under a similar name John Smith.

The problem:

The hotel did not have a data quality process in place to standardize, clean and merge duplicate records to provide a complete view of the customer. As a result, the hotel was not able to leverage the true value of its data in delivering relevant marketing to a high value customer.

4. Matching

Data matching consolidates data records into identifiable groups and links/merges related records within or across data sets. This process locates matches in any combination of over 35 different components – from common ones like address, city, state, ZIP®, name, and phone – to other not-so-common elements like email address, company, gender and social security number. You can select from exact matching, Soundex, or Phonetics matching which recognizes phonemes like “ph” and “sh.” Data matching also recognizes nicknames (Liz, Beth, Betty, Betsy, Elizabeth) and alternate spellings (Gene, Jean, Jeanne).

5. Enrichment

Data enrichment enhances the value of customer data by attaching additional pieces of data from other sources, including geocoding, demographic data, full-name parsing and genderizing, phone number verification, and email validation. The process provides a better understanding of your customer data because it reveals buyer behavior and loyalty potential.

- **Address Verification –** Verify U.S. and Canadian addresses to highest level of accuracy – the physical delivery point using DPV® and LACS$^{\text{Link®}}$, which are now mandatory for CASS Certified processing and postal discounts.

- **Phone Validation –** Fill in missing area codes, and update and correct area code/prefix. Also append lat/long, time zone, city, state, ZIP, and county.

- **Email Validation –** Validate, correct and clean up email addresses using three levels of verification: Syntax; Local Database; and MXlookup. Check for general format syntax errors, domain name changes, improper email format for common domains (i.e. Hotmail, AOL, Yahoo) and validate the domain against a database of good and bad addresses, as well as verify the domain name exists through the MaileXchange (MX) Lookup, and parse email addresses into various components.

- **Name Parsing and Genderizing –** Parse full names into components and determine the gender of the first name.

- **Residential Business Delivery Indicator –** Identify the delivery type as residential or business.

- **Geocoding –** Add latitude/longitude coordinates to the postal codes of an address.

6. Monitoring

This real-time monitoring phase puts automated processes into place to detect when data exceeds pre-set limits. Data monitoring is designed to help organizations immediately recognize and correct issues before the quality of data declines. This approach also empowers businesses to enforce data governance and compliance measures.

Supporting MDM

Along with setting up a Total Data Quality solution, you will need to deal with the other challenge of MDM – mainly, the deduplication of data from disparate sources with the integration provided by SSIS.

An MDM application that combines data from multiple data sources might hit a roadblock merging the data if there isn't a 'unique identifier' that is shared across the enterprise. This typically occurs when each data source system (i.e. an organization's sales division, customer service department, or call center) identifies a business entity differently.

There are three general categories or ways to organize your data so that it can ultimately be merged for MDM solutions – they are unique identifiers, attributes, and transactions.

Unique Identifiers – These identifiers define a business entity's master system of record. As you bring together data from various data sources, an organization must have a consistent mechanism to uniquely identify, match, and link customer information across different business functions. While data connectivity provides the mechanism to access master data from various source systems, it is the Total Data Quality process that ensures integration with a high level of data quality and consistency. Once an organization's data is cleansed, its unique identifiers can be shared among multiple sources. In essence, a business can develop a 'single customer view' – it can consolidate its data into a single customer view to provide data to its existing sources. This ensures accurate, consistent data across the enterprise.

Attributes – Once a unique identifier is determined for an entity, you can organize your data by adding attributes that provide meaningful business context, categorize the business entity into one or more groups, and provide more detail on the entity's relationship to other business entities. These attributes may be directly obtained from source systems.

While managing unique identifiers can help you cleanse duplicate records, you will likely need to cleanse your data attributes. In many situations, you will still need to perform data cleansing to manage conflicting attributes across different data sources.

Transactions – Creating a master business entity typically involves consolidating data from multiple source systems. Once you have identified a mechanism to bridge and cleanse the data, you can begin to categorize the entity based on the types of transactions or activities that the entity is involved in. When you work with transaction data, you will often need to collect and merge your data before building it into your MDM solution.

Building Support for Compliance and Data Governance

MDM applications help organizations manage compliance and data governance initiatives. Recent compliance regulations, such as Sarbanes-Oxley and HIPAA, have increased the need for organizations to establish and improve their data quality methodologies. Without a solid MDM program in place, it would be difficult to make sense of the data residing in multiple business systems. Having well-integrated and accurate data gives organizations a "central system of record" – allowing them to comply with government regulations as a result of gaining a better understanding of their customer information.

Conclusion

A business can't function on bad, faulty data. Without data that is reliable, accurate and updated, organizations can't confidently distribute their data across the enterprise – which could potentially lead to bad business decisions. Bad data also hinders the successful integration of data from a variety of data sources.

But developing a strategy to integrate data while improving its quality doesn't have to be costly or troublesome. With a solid Total Data Quality methodology in place – which entails a comprehensive process of data profiling, cleansing, parsing and standardization, matching, enrichment, and monitoring – an organization can successfully facilitate an MDM application. Total Data Quality helps expand the meaning between data sets, consolidates information, and synchronizes business processes. It gives organizations a more complete view of customer information – unlocking the true value of its data, creating a competitive advantage and more opportunities for growth.

About Melissa Data Corp.

For more than 23 years Melissa Data has empowered direct marketers, developers and database professionals with tools to validate, standardize, de-dupe, geocode and enrich contact data for custom, Web and enterprise data applications. The company's flagship products include: Data Quality Suite, MatchUp®, and MAILERS+4®. For more information and free trials, visit www.MelissaData.com or call 1-800-MELISSA.

About Total Data Quality Integration Toolkit (TDQ-IT)

TDQ-IT is a full-featured enterprise data integration platform that leverages SQL Server Integration Services (SSIS) to provide a flexible, affordable solution for total data quality and master data management (MDM) initiatives. For a free trial visit www.MelissaData.com/tdq

About Microsoft

Founded in 1975, Microsoft (Nasdaq "MSFT") is the worldwide leader in software, services and solutions that help people and businesses realize their full potential.

About Microsoft Integration Services and SQL Server

Microsoft Integration Services is a platform for building enterprise-level data integration and data transformations solutions. For SQL Server 2008 product information, visit Microsoft SQL Server 2008.