Excerpted from CIO Story, September 2014

Big Data in Action

By Greg Hussey, CGI, and Jeff McTygue, California Franchise Tax Board

According to the hype around Big Data, we can find answers to almost any question by combining large volumes of data. Data is becoming more about what it can tell us, and less about recording transactions. As emerging technology makes it easier to combine data sets, new questions are being asked—to help improve services, detect fraud and increase security, for example—that require comparing data across systems.

Most large organizations have multiple systems designed for specific programs with unique data models that need to share data with other systems. Master Data Management (MDM) is helping such organizations get a handle on large data volumes by linking data to a "single source of the truth."

Often when people discuss MDM, they focus on technology. The important thing to remember is that MDM is much more than technology. It is also about the business processes and practices that include how to organize the data and govern it.

Let's look at an example of an organization dealing with massive data volumes and what they have learned from their MDM experience so far.

The State of California's Franchise Tax Board (FTB), which administers both personal and corporate income taxes, is in the midst of its Enterprise-Data-to-Revenue (EDR) project. This 66 month project seeks to increase FTB's ability to use data to improve both tax compliance and customer service.

FTB receives, consumes and generates an enormous amount of data, keeping track of 60 million individuals and business entities, as well as 50 million tax returns and payments, 500 million tax related schedules and documents, and 1 billion income information returns each year. Five primary applications, and more than a dozen secondary ones, support their tax account and compliance activities. Each application has its own data extraction and matching rules, so there are dozens of matching routines.

FTB's new, proactive MDM approach identifies and organizes data at the front-end instead of the back end. It objectives are to:

- Develop and maintain an enterprise customer data store
- Have a consolidated process for matching and searching
- Use probabilistic vs. deterministic matching routines to better accommodate data irregularities such as character transpositions, misspellings, etc., so less data is lost.

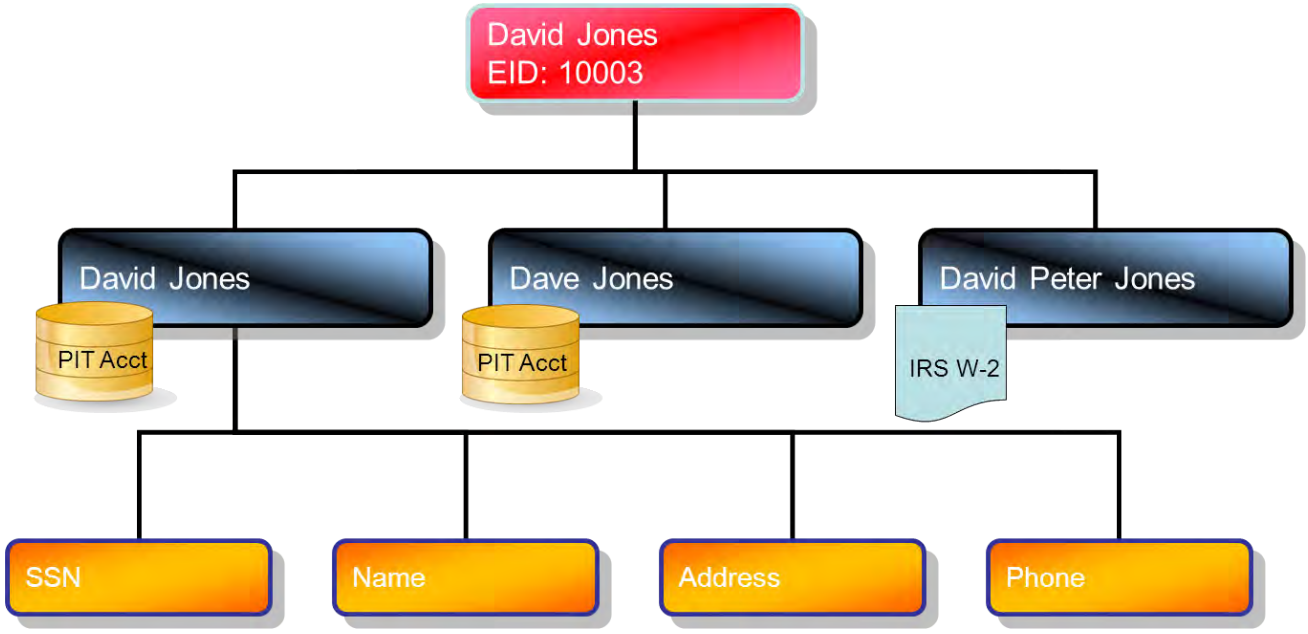
Linking data using Entities

EDR receives demographic data from multiple "Sources"—its five primary systems plus 35-to-40 external feeds. In the MDM, records from Sources are called "Members," which can be unique or repeated. Multiple Members are linked together to create an "Entity," which represents a unique person or organization. The Entity is what ties the data together.

What is so powerful about keeping Members separate is that the data is not co-mingled, so each Source retains its identity and provides an audit trail. In Figure 1, the Entity, "David Jones", gets data records (Members) from three Sources. As the Entity continues to be built, all variations are retained in all Source files to enrich the data matching over time. The end state is a centralized view of the customer, and the ability to use the best data from each Source by applying business rules.

About the authors: Greg Hussey is CGI's Project Manager for the EDR project, one of the largest legacy systems integration projects of its kind for a U.S. state tax and revenue agency. Jeff McTygue is the Client Data Manager for the FTB, who, in partnership with CGI, helps to implement data management activities for EDR.

Figure 1: Entity Example



The EDR MDM methodology

In the MDM world, having and following a methodology is critical to success. Following is the six-step approach being used by EDR, which is software-agnostic:

1. Analyze the demographic data elements that are captured and received to understand the relationships between them
2. Profile the data to identify the trends
3. Standardize certain aspects of the data (such as address and name) and use software programs to enrich addresses
4. Load the data into the MDM via different methods, such as batch, SOA, etc.
5. Build the Entities using the data loaded in the MDM, and then analyze the Entities until an agreeable Entity-to-Member threshold is reached for various risk factors. A threshold is a number on a scale developed to make a decision, such as Link, Ignore, or assign to a Task
6. Expose the data to the consuming applications via the MDM Services.

Lessons learned

From their MDM experiences to date, the EDR project team has identified several keys to success:

- Get early business involvement and buy in. Business champions are often the tiebreakers who can escalate issues to help resolve inevitable disagreements
- Recognize that data quality and cleanup are "lifetime activities," and not just pre-conversion work
- Since data governance has many meanings, be specific about what issues need to be resolved and in what priority
- While standardization and data cleanliness are critical, understand that cleanliness of data pulled from outside systems may be out of your control
- Accept that there are some limitations on the ability to transform data; retaining "information as submitted" may be required by policy or law
- Use analytics to better understand characteristics of the data to be rationalized. A much better MDM system can be designed with greater knowledge of the nature of the data itself.

Big Data has matured over the last several years, moving from using MDM as a coherent methodology for data storage, to recording transactions, to data analysis, and now to using that analysis to generate positive business results. CGI's partnership with FTB is a demonstration of the results that can be achieved when MDM is applied correctly. FTB is using MDM to improve its revenue collection with a goal of achieving more than $4.9 billion over a 7 year period. As of July 2014, the project already had generated more than $1 billion in recovered revenue for the State.