MASTER DATA MANAGEMENT IN THE AGE OF BIG DATA

PRESENTED TO IRMAC – MAY 15, 2013
STEVE PAPAGIANNIS
STEVE.PAPAGIANNIS@SAS.COM
416 307 4620
**Master data** is information that is key to the operation of a business. It is the primary focus of the Information Technology (IT) discipline of Master Data Management (MDM), and can include reference data. This key business information may include data about customers, products, employees, materials, suppliers, and the like.

*Source: Wikipedia*

**Big Data.** When the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making.

*Source: SAS Website*
Master Data is the high value information that is used most frequently across an enterprise.

- Example: Customers, Members, Groups, Organizations, Patients, Vendors, Providers, Citizens, Employees, Services, Accounts, Materials, Locations, Assets, etc.

Big Data is the low value density information generated by systems

- Example: Security Logs, Call Detail Records, Sensor Information
WHY NOW???

• Master Data is getting richer
• Big Data is being heavily explored
• Costs to achieve this are dropping
• Hype cycle
THIRVING IN THE BIG DATA ERA

DATA SIZE

VOLUME

VARIETY

VELOCITY

VALUE

BIG DATA

INFORMATION OVERLOAD

RELEVANT DATA

TODAY

THE FUTURE

THRIVING IN THE BIG DATA ERA

DATA SIZE

VOLUME

VARIETY

VELOCITY

VALUE

BIG DATA

INFORMATION OVERLOAD

RELEVANT DATA

TODAY

THE FUTURE

THRIVING IN THE BIG DATA ERA

DATA SIZE

VOLUME

VARIETY

VELOCITY

VALUE

BIG DATA

INFORMATION OVERLOAD

RELEVANT DATA

TODAY

THE FUTURE

THRIVING IN THE BIG DATA ERA

DATA SIZE

VOLUME

VARIETY

VELOCITY

VALUE

BIG DATA

INFORMATION OVERLOAD

RELEVANT DATA

TODAY

THE FUTURE

THRIVING IN THE BIG DATA ERA

DATA SIZE

VOLUME

VARIETY

VELOCITY

VALUE

BIG DATA

INFORMATION OVERLOAD

RELEVANT DATA

TODAY

THE FUTURE

THRIVING IN THE BIG DATA ERA

DATA SIZE

VOLUME

VARIETY

VELOCITY

VALUE

BIG DATA

INFORMATION OVERLOAD

RELEVANT DATA

TODAY

THE FUTURE

THRIVING IN THE BIG DATA ERA

DATA SIZE

VOLUME

VARIETY

VELOCITY

VALUE

BIG DATA

INFORMATION OVERLOAD

RELEVANT DATA

TODAY

THE FUTURE

THRIVING IN THE BIG DATA ERA

DATA SIZE

VOLUME

VARIETY

VELOCITY

VALUE

BIG DATA

INFORMATION OVERLOAD

RELEVANT DATA

TODAY

THE FUTURE

THRIVING IN THE BIG DATA ERA

DATA SIZE

VOLUME

VARIETY

VELOCITY

VALUE

BIG DATA

INFORMATION OVERLOAD

RELEVANT DATA

TODAY

THE FUTURE

THRIVING IN THE BIG DATA ERA
Hwy 410, Brampton, Acme Construction

Data Profiling
Metadata Discovery
Business Rule Definition
Entity Definition

Contract Center

Batch | Real-Time

Budgeting

Highway 410, Brampton, Peel Region, Kilometer 10, January, $75MM, Acme Construction, resurfacing

Scattered Data

Highway 410, Peel Region

Batch | Real-Time

Engineering

Batch | Real-Time

Hwy 410, January, Kilometer 10-12, resurfacing

Data Integration
Data Quality
Data Model
Business Services
Stewardship Console
Data Governance
Identity Management
Reporting

Hwy 410, $75MM
## SURVIVING CONTRIBUTORS INTO MASTER RECORD

<table>
<thead>
<tr>
<th>CRM</th>
<th>1001</th>
<th>30391-244</th>
<th>30391-244</th>
<th>14239</th>
<th>3721B</th>
<th>30391-244</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ConsumerID</td>
<td>First Name</td>
<td>Middle</td>
<td>Last Name</td>
<td>DOB</td>
<td>SSN</td>
<td>Address</td>
</tr>
<tr>
<td>30391-244</td>
<td>William</td>
<td>James</td>
<td>Crown</td>
<td>04/12/1939</td>
<td>563-49-123</td>
<td>123 Oak St., Eves, IL 30319</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ERP</th>
<th>ERP</th>
<th>ERP</th>
<th>ERP</th>
<th>ERP</th>
<th>ERP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Member ID</td>
<td>First Name</td>
<td>Middle</td>
<td>Last Name</td>
<td>DOB</td>
<td>SSN</td>
</tr>
<tr>
<td>30391244</td>
<td>William</td>
<td>J.</td>
<td>Crowne</td>
<td>04-12-39</td>
<td>563491234</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ONLINE</th>
<th>ONLINE</th>
<th>ONLINE</th>
<th>ONLINE</th>
<th>ONLINE</th>
<th>ONLINE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person ID</td>
<td>First Name</td>
<td>Middle</td>
<td>Last Name</td>
<td>DOB</td>
<td>SSN</td>
</tr>
<tr>
<td>14239</td>
<td>Bubba</td>
<td>J.</td>
<td></td>
<td>April 12</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DW</th>
<th>DW</th>
<th>DW</th>
<th>DW</th>
<th>DW</th>
<th>DW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>First Name</td>
<td>Middle</td>
<td>Last Name</td>
<td>DOB</td>
<td>SSN</td>
</tr>
<tr>
<td>3721B</td>
<td>William</td>
<td>James</td>
<td>Corp.</td>
<td>April 12</td>
<td>563491234</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SFA</th>
<th>SFA</th>
<th>SFA</th>
<th>SFA</th>
<th>SFA</th>
<th>SFA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>First Name</td>
<td>Middle</td>
<td>Last Name</td>
<td>DOB</td>
<td>SSN</td>
</tr>
<tr>
<td>30391-244</td>
<td>William</td>
<td>James</td>
<td>Crown</td>
<td>04/12/1939</td>
<td>563-49-123</td>
</tr>
</tbody>
</table>

**Source Keys**
- CRM 30391-244
- ERP 30391244
- ONLINE 14239
- DW 3721B
- SFA 30391-244

**Survived Fields**
- William
- James
- Crowne
- 04/12/1939
- 563-49-123
- 123 Oak Street
- Eves
- CA
- 91403
Evolving From DM to MDM

Data Management

Master Data Management

Data Integration and/or Data Quality Initiative

Enrich: Data Augmentation and Enrichment

Survive: Entity Resolution & Surviving Record Analysis

Persist: Create a “Master” data record

Synchronize: Tie master data record into existing source systems

Surface: Provide access to other systems in real time via SOA

Full-blown MDM Initiative
Phase 1: Define / Discover

- Project Definition, Requirements Assessment
- Installation
- Source Analysis/Profiling
- Define Entities, ETL, Relationships

Phase 2: Apply Data Quality

- Data Connectivity
- Define and Implement DQ, Matching, Survivorship Rules
- Data Enrichment, Verification, Validation

Phase 3: Perform Initial Load

- Stage Data, Load Hub
- Integrate with Source Systems
- Define and Implement Services
- Performance Testing

Phase 4: QA / Knowledge Transfer

- UAT
- Build Reports
- UI Enhancements
- Support/ Knowledge Transfer
WHAT IS THE INTERSECTION?

Big Data

MDM

Governance
SAS® INFORMATION MANAGEMENT

SUPPORT FOR ENTIRE INFORMATION MANAGEMENT CONTINUUM

Strategy

STRATEGY & VISION TO EXECUTE

Governance

INFORMATION GOVERNANCE

Capabilities

DATA MANAGEMENT

Provides unified data management capabilities that include data governance, data integration, data quality and MDM

DECISION MANAGEMENT

Provides decision services that include business rules and workflow that facilitates integration of the information services into the business systems

ANALYTICS MANAGEMENT

Provides complete analytics management that includes model management, deployment, monitoring and governance of the analytics information asset
WHERE IS THE DATA
MORE VARIETY THAN BEFORE
HOW IS IT PROCESSED
MORE VARIETY THAN BEFORE

- Good ol’ SQL
- ETL tools & Custom Processes
- Data Quality
- Information Management
- MDM Solutions
- Event Stream Processing
BUSINESS DATA NETWORK

• Enables data governance
  ▪ Common language improves communication & supports compliance regulations
  ▪ Represent and expose business relationships
  ▪ Track history of changes

• Accountability and responsibility
  ▪ Document and communicate ownership
  ▪ Notify interested parties on changes

• Supports better collaboration
  ▪ Capture and share annotations between team members
  ▪ Greater understanding of the context of information
  ▪ Use and reuse of trusted information
Database = ORACLE
Schema = NAACCT
Table = DLYTRANS
Column = TAXVL
data type =
Decimal(14,2)
Derivation:
SUM(TRNTXAMT)

Category: Costs
Term: Tax Value
Description: Tax to be paid on Gross Income.
(John Walsh is responsible for updates. 90% reliable source)
Status: UNDER REVIEW

Achieve a common vocabulary between business & technical users

Business Data Network
ENABLES DATA GOVERNANCE

- Common language improves communication & supports compliance regulations
ENABLES DATA GOVERNANCE

• Represent and expose business relationships
ENABLES DATA GOVERNANCE

- Track history of changes
ACCOUNTABILITY AND RESPONSIBILITY

- Document and communicate ownership
- Notify interested parties on changes
COLLABORATION

- Capture and share **Notes** between team members
COLLABORATION (CON’T)

• Understand the context of information
EVENT STREAM PROCESSING
EVENT STREAM PROCESSING (ESP)

ESP is a subcategory of Complex Event Processing (CEP) focused on analyzing/processing ‘events in motion’ called Event Streams.*

The SAS ESP is an embeddable engine that can be integrated into or front-end solutions.

* This is the definition provided by the Event Processing Technical Society
TYPICAL CHARACTERISTICS OF EVENT STREAM PROCESSING APPLICATIONS:

- Continuous queries on data in motion (with incrementally updated results)
- Very low (max) event processing latencies (i.e., Usecs-msecs)
- High volumes (>100k events/sec)
- Derived event windows with retention policies
- Memory constrained for performance (i.e., Bounded state)
- Predetermined data mining, decision making, alerting, position management, scoring, profiling, …
- Event out-of-order handling to ensure ordered source streams
EVENT STREAM PROCESSING (ESP) VS. RELATIONAL DATABASE MANAGEMENT SYSTEM (RDBMS)

ESP

- ESPs store the queries and continuously stream data through the queries

RDBMS

- Databases store the data and periodically run queries against the stored data
WHAT PROBLEMS ARE WE TRYING TO SOLVE?

- Capture value otherwise lost through information lag
- Enable new opportunities through producing actionable intelligence with lower latencies
- Continuously analyze events as they occur
- Eliminate storage latencies
- Incrementally update intelligence as new events occur
- Enable new analysis & processing models to be developed and modified quickly to increase the opportunity windows and reduce costs.
**Portfolio Equity Position Management**

Portfolio positions of interest, e.g., market equity positions by account, trader, department, location, region, and venue.

**Continuous Change Data Management**

Get master data changes as they occur and provide a consolidated up-to-the-moment view across silo systems. Example is a consolidated customer view for telco services across home services, wireless, …

**Capital Markets Liquidity**

Consolidated L1 & L2 order books for selected instruments across selected venues. This can be used for algorithmic trading, trade execution, dark pool analysis.

**Credit Card Fraud Prevention**

Maintain account-based usage signatures and known fraud signatures to enable credit card purchase scoring for authorization requests.

**Telco Prepaid Call Authorization**

Correlate call authorization requests to account & account call plans to determine the call duration based on current balance. Maintain account balances upon call completions.
Data Flow Model:

1. Customer services information is captured from the various DBs of silo systems as a query snapshot followed by change log deltas.
2. Customer account maps are used to normalize the various account numbers across disparate systems.
3. The normalized service data is cleansed for accuracy and completeness.
4. All customer services are consolidated into one holistic view of each customer, which is used by Customer Care.
DataFlux qMDM provides multi-entity support. Many teams (admins, stewards, business users) can work together to define various elements related to each entity like attribute lists, match rules, data quality rules, security roles, and so on.

Entities are not “live” in the system until they have been published in Master Data Manager and can’t accept data from other systems until the jobs/services to support each entity type have been generated using Master Data Manager.
Entity types support inheritance. This allows entities derived from others to inherit attributes, clustering rules, roles, and other properties from their parent entity. Inheritance will in some cases render properties read-only on the child entity if the property was inherited from a parent. Modifications would need to be made at the higher level.

Entity types can also be designated as abstract, which means they can carry attributes, roles, and so on but they won't be used to store actual instance data.

Administrator Activity

This entity type inherits from the "Party" entity type.
In this administrative area, clustering conditions can be designed for each entity type. By virtue of inheritance, clustering rules can be accessed by entities that take others as their parent.

Job generation functionality will take these rules and other properties defined for each entity and will create ready-to-use data jobs for batch loads and updates, real-time queries, and real-time add, modify, and retire services.

Entities will not be active until published by an administrator.
Entities can be related to other entities through defined relationships. Here a relationship between companies and parts has been established. Notice the labels and descriptions. These will appear throughout Master Data Manager to show the kind of relationship that is being displayed.

Relationships are defined by a set of match rules that link the entity instance data. Here, if the match code for the company name in one entity matches the supplier match code in the other, a relationship will be made.
Master Data Manager provides several ways to find the specific instance data you are looking for. This can be done through workflow tasks or reports but typically searching is done in the Master Data area. Searches can be done on entities and hierarchies. You can use any number of search fields or use an advanced search feature for more granular control.

Entities that inherit from a common parent will have search results displayed together on the same screen.

Fields available for searching are configured by an administrator and can differ across entity types.
New entity instance data can be created directly from Master Data Manager. The fields available in the edit form are configured by an administrator for each entity type.

Fields can be marked as required or read-only and they can be constrained by regular expressions to enforce data standards.
At the heart of DataFlux qMDM is the notion of match clusters. These represent the group of records from different sources that have been deemed to be representations of the same person, place or thing.

Every match cluster has a best record that is constructed through business rules from the values held by one or more contributing records. Best records can be authored and edited using Master Data Manager.
Cluster compare functionality allows data stewards to look across contributing records in a cluster and quickly see, through highlighted fields, what is different.

Like the main Cluster Members view, you can tag this entity for a workflow, create a new best record, retire the cluster, or split off contributors into new clusters.

You can double-click field values on the right to populate your potential new best record on the left.
Every change to entity data is captured in the qMDM database. Modifications to best records result in regenerated best records and older versions are retired. Entire match clusters can also be retired, making the data inactive but available for queries of the system’s history. In this view, previous versions of the best record are shown in gray while active information is shown normally. Users can toggle the history view by choosing to show or hide retired records.

Use the Show Retired menu item to view the history of changes to the match cluster.
Every relationship definition that involves an entity type will be shown as new windows in the entity Properties area. Master Data Manager will query for discovered relationships for the first relationship type but you’ll have to expand each additional relationship window to see all of them.

Discovered relationships will appear automatically; however, you can also manually add and retire relationships here.
The relationship diagram is a visual representation of an entity and all of its relationships with other entities. The area at the right sets the action for expanding related items (it’s configurable since some queries can take more than a few second to run).

A double-click will expand the relationship diagram to the next set of related entities. The window at the bottom either shows attributes for the selected entity or, if a group-type node is selected, all of the entities that comprise the group. These can be selectively added to the diagram.
DataFlux qMDM supports the concept of entity hierarchies. There is no limit on the number of unique hierarchies or entity types per hierarchy.

Hierarchies relate entities at the “best record” level and respond accordingly to splits and merges of match clusters. Entities can participate in several hierarchies simultaneously and appear in the same hierarchy more than once outside of the same level.
The master data dashboard dynamically updates as new entity types are added to DataFlux qMDM. It shows summary statistics by source and entity type for:

- Record counts
- Volume growth
- Number of contributors and survivors
- Data analysis for sparsity and max/min values
- Source system contribution ratios
- Batch load history
Dynamic and batch reports are supported through the Reports area. Both are specially designed DataFlux data jobs interpreted as reports by Master Data Manager.

Dynamic reports can accept user input as parameters and the output can be linked directly to entity data elsewhere in Master Data Manager for easy access.

Batch jobs can be initiated from this location and if their output contains an HTML file, that data can be displayed here.
Master Data Manager supports the concept of workflows. Two are available with a basic installation:

- **Tag** – this is an ad hoc workflow that can be used to raise awareness of issues to other qMDM users. It is enabled by default.
- **Entity Lifecycle** – this can be used to route new records and edits to existing clusters to a group of users for review. It is available but not enabled by default.

Here an entity has been tagged for review.
WORKFLOW TASKS

Once a user chooses to act on a workflow task (they have already chosen to “accept” the task in the Actions item), the area below the Status information will contain entity attributes that they can view and modify.

The entity attributes values shown are what was captured at the time the workflow was created. The actual entity attributes in the system may have changed after the workflow task was initiated. If so, any submission of changes from this location will be integrated with changes made in the interim.

The Notes field can be used to communicate issues to other team members.
The workflows dashboard provides a snapshot of workflow activity generated in DataFlux qMDM. Among the statistics it reports are the following:

- Workflows by status
- Workflows by priority
- Workflows by entity type
- Workflows by workflow type
- Workflows by submitter
- Workflows by modification date
- Workflows by start date
DATA EXPORT

Data from most panels in Master Data Manager can be exported to CSV, PDF, or Excel formats.