Taming a Complex Data Landscape

Carl Gerber
Principal Architect & Independent Consultant

Moving Towards MDM & SOA
Southwest Ohio Data Management Association
February 27, 2009
1. MDM as Pretext to SOA
2. Business Context & Motivation
3. Data Landscape
4. Approach & Technology
5. Data Quality & Stewardship
MDM as a Pretext to SOA

• SOA without maturity is a Silver Bullet

• Services publish & consume Data

• Must Get Data House In Order to publish data services

• Therefore: Comprehensive Master Data Management Builds a SOA Foundation
Master Data Management (MDM)

Business Context
Business Context

• Support Revenue initiatives
• Enhance business decision making
• Deliver operational efficiencies and enhancements

Bad times, are the best times, to prepare for good times
Business Context

Build Transparency & Trust Across Core Data Objects in the Landscape

Value Chain

Custom Application  | ERP  | 3rd Party Hosted App | 3rd Party Hosted App | POS Application | ERP  | Data Warehouse
Custom Application | Custom Application | Spreadsheets | Custom Application | ERP  | Spreadsheets | Spreadsheets

Enterprise Systems

2/27/2009
Motivation
Motivation - Key Questions

• Where is the business critical information?

• How can I relate the business critical information together?

• Where are the inconsistencies that lead us to lower efficiencies?
Motivation - Architecture

### Implications of Architecture Maturity Stages

<table>
<thead>
<tr>
<th>Strategic Implications of IT</th>
<th>Local/Functional Optimization</th>
<th>IT Efficiency</th>
<th>Operational Efficiency</th>
<th>Strategic Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>36%</td>
<td>25%</td>
<td>16%</td>
<td>15%</td>
</tr>
<tr>
<td>85%</td>
<td>21%</td>
<td>32%</td>
<td>34%</td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td>14%</td>
<td>35%</td>
<td>33%</td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>11%</td>
<td>17%</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Local Applications**
- **Enterprise Systems**
- **Shared Infrastructure**
- **Shared Data**

### Architecture Maturity

- **IT Budget**
  - 100%
  - 85%
  - 75%
  - 120%

*Center for Information Systems Research
© 2006 MIT Sloan CiSR – Ross*

Copyright MIT Sloan Center for Information Systems Research. Used with permission.
Motivation - Benefits

Cost Reduction
- Rationalization of Redundant data
- Reduced use of Storage
- Reduction in Effort for Data Maintenance
- Reduce Duplicate purchase of 3rd party data

Improved Efficiency
- Reduction of Data Inconsistencies (in Silos)
- Multiple checks for Data Errors
- Single “trusted” source of
  - Customer Data
  - Product Data

Improved Customer Service
- Reduce Exposure of incorrect data
- Reduce Customer complaints / penalties
- Systematic improvement of Data Quality
Motivation - Benefits

Improved Reporting
- Reduce Data Inconsistencies in Reports
- Provide Enterprise Level Reporting
- Better understanding of Risks and Exposure

Productivity
- Gain access to “Cross Silo” View of Data
- Promote Knowledge Sharing
- Establish mechanism for Data Ownership

Compliance
- Improve Regulatory Reporting
- Ensure Traceability and Auditability of Data
- Assist in SOX Compliance
Data Landscape
## Data Landscape - *Typical Challenges*

<table>
<thead>
<tr>
<th>Complex Data Landscape</th>
<th>Point-to-Point Data Flows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data is Understood Locally</td>
<td>Custom Interfaces</td>
</tr>
<tr>
<td>Metadata is Not Documented</td>
<td>Transformations Hard-coded</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Redundant Data</td>
<td>Organic Data Environment</td>
</tr>
<tr>
<td>Difficult to Synchronize</td>
<td>Un-planned Growth</td>
</tr>
<tr>
<td>Data Ownership is Unknown</td>
<td>Constantly Changing</td>
</tr>
<tr>
<td>System of Record is Uncertain</td>
<td></td>
</tr>
</tbody>
</table>

2/27/2009
CarlGerberJr@gmail.com    614.783.1442
Data Landscape - As Is

“In large established organizations, the IT infrastructure can sometimes appear to have been constructed in a series of weekend handyman jobs. It does the job it was designed to do but is apt to create problems whenever it is applied to another purpose.”

*Competing on Analytics*, Thomas Davenport, Jeanne Harris.

2/27/2009
Approach

- Inventory Important Data Objects
- Document Current Business Data Owner Roles
- Apply Consistent Data Quality Checks
- Implement an Enterprise Data Model
- De-couple Systems via Hubs
- Leverage Existing Tools
Approach - *Technique*

- **Engage Subject Matter Experts**
  - Inventory & Data Quality Requirements
  - Source-to-Target Data Mappings
  - Document Semantic Meaning

- **Engage Data Architects**
  - Model Data Hubs
  - Publish Semantic Meaning
  - Establish ETL Patterns
    - Error Handling
    - Data Quality Checks

- **Create Enterprise Data Assets**
  - Re-used Across Systems and Projects
Approach - **Technologies**

- Data Modeling CASE Tool
- Data Profiling & Cleansing Tools
- Universal Metadata Tool
  - MS Excel
- *Metadata Repository*
  - Intranet
- Extract Processes
  - ETL Tool Direct Connectors
  - Custom Code & Flat Files
- Transform & Load
  - RDBMS-based Data Hubs
  - ETL Tool
Data Landscape - To Be

Data Hubs

- Item & Merchandise
- Inventory
- Location
- Sales
- Promotions
- Others

Merchandise
Inventory
Point of Sale
Others
Financials
Stores
Business Intelligence Data Warehouse
Data Integration Hubs

• Built Using Existing Tools
• New Mind Set
• Leverages Your Data Assets
# Data Hub Anatomy

<table>
<thead>
<tr>
<th></th>
<th>Master</th>
<th>Transaction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Enterprise Data Model</strong></td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td><strong>Error Handling</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Insert &amp; Notify</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Write to Error Table &amp; Notify</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>Data Cleanse &amp; Notify</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>Threshold Reached - Stop</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td><strong>History</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week of Daily Snapshots</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Cyclic (Deltas)</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>Full (Snapshot)</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Update/Destructive Overwrites (Type 1)</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>Insert Only – No Update</td>
<td></td>
<td>●</td>
</tr>
</tbody>
</table>

2/27/2009

CarlGerberJr@gmail.com  614.783.1442
# Data Hub Anatomy

<table>
<thead>
<tr>
<th>Process Control Dimensions</th>
<th>Master</th>
<th>Transaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run Dates, Batch &amp; Audit IDs</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Error Identifiers</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Source System</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Subject Area &amp; Data Stewards</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Transaction, Processed, Post Dates</td>
<td></td>
<td>●</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transfer Architecture</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Publish &amp; Subscribe</td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>Push</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structure</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Key</td>
<td>Surrogate Key + Date</td>
<td>Surrogate Key</td>
</tr>
<tr>
<td>Alternate Key</td>
<td>Business Keys</td>
<td>Biz Key + Trans_ID + Date</td>
</tr>
</tbody>
</table>
Methodology
Methodology

1. Understand what exists
   - *Create an inventory*

2. Organize the information
   - *Classify the data into relevant categories*

3. Understand what it means
   - *Get the semantic metadata*

4. Determine the distribution of the shared data
   - *Understand the relationships between the data categories*
Methodology

5. Map the information
   • Attribute level mapping
   • Understand the data flows
   • Understand the hierarchies

6. Publish the understanding
   • Provide enterprise level knowledge-base

7. Create a Data Quality Framework
   • Leverage data as an asset
   • Resolve inconsistency & inaccuracy in the data environment
1. Understand what exists

- **Inventory Databases & Data Stores**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Seq</td>
<td>Application</td>
<td>Description</td>
<td>Database Instance Name</td>
<td>Semantic Metadata Available</td>
<td>Database Type</td>
<td>Environment Type</td>
<td>Server</td>
<td>Internal/External</td>
<td>Owner</td>
<td>Organization</td>
<td>Contributor</td>
<td>Last Update Date</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Catalog Core Business Data Objects**
Criteria for Determining Core Business Objects

QUESTION:
What data objects are important to the business?

Technical Criteria
1. Object must have an entity associated with it (unique ID)
2. Object must be present in multiple databases
3. Object must be associated with multiple attributes.

Business Criteria
1. Knowledgeable individuals in multiple areas must acknowledge importance
2. Multiple individuals must agree that objects are “core” to business.
2. Organize the information

- **Product**
  - Product (Item)
  - Catalog
  - Price
  - Product Grp
  - Dimensions

- **Customer**
  - Customer Account Type

- **Location**
  - City
  - State
  - Province
  - Country
  - District
  - Region

- **Vendor**
  - Vendor Type
  - Distributor

- **Market**
  - Market Segment
  - Promotion Channel

- **Financial**
  - Chart Accnts
  - Accnts Payble
  - Currency

- **Financial**
  - Chart Accnts
  - Accnts Payble
  - Currency

- **Financial**
  - Chart Accnts
  - Accnts Payble
  - Currency

- **Organization**
  - Company
  - Division
  - Group
  - Employee
  - HR Org Unit

- **Assets**
  - Fixed
  - Digital Infrastructure

- **General**
  - Address
  - Email
  - Language

- **Time**
  - Calendar Yr
  - Fiscal Year
  - Quarter
  - Month
  - Week

- **Transaction Objects**
  - Sales
  - Invoices
  - Shipments
  - Purchase Orders

*Shared Data Object Classification*
3. Understand what it means
3. Understand what it means
4. Determine the distribution of the shared data

Criteria for Understanding Commonalities Across Databases

**Technical Criteria**

1. Object ID must be present in multiple databases
2. Metadata should show common attributes
3. Attribute data should show strong overlaps.

**Business Criteria**

1. Current Data Stewards should confirm the object distribution.
2. Current Data Stewards should confirm differences in data structures.
## Shared Data Object Distribution

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>L</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DataObject Matrix</td>
<td>System of Record</td>
<td>Business Owner</td>
<td>Data Steward</td>
<td>Product_App</td>
<td>ERP</td>
<td>Data Warehouse</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Product</td>
<td>Product / Item</td>
<td>Product_App</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PRODUCT</td>
<td>MASTER ITEMS</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Customer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Vendor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Click

2/27/2009  CarlGerberJr@gmail.com  614.783.1442
Shared Data Object Distribution

<table>
<thead>
<tr>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Table Name</strong></td>
<td><strong>Column Name</strong></td>
<td><strong>Datatype</strong></td>
<td><strong>Length</strong></td>
<td><strong>Mandatory</strong></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. Map the information

Source: Global IDs, Used with Permission
6. Publish the understanding

• Remember When You Couldn’t Find the Metadata, Glossary, Data Dictionary?
  - Create a Table in Core Databases for Definitions

<table>
<thead>
<tr>
<th>ID</th>
<th>Table Name</th>
<th>Column Name</th>
<th>Definition</th>
</tr>
</thead>
</table>

• A Page on the Intranet
• Data Portal
7. Create a Data Quality Framework

- Ownership
- System of Record
- Data Validation Rules
  - Completeness
  - Timeliness
  - Accuracy
- Stewardship
  - Roles & Responsibilities
  - Process
Determining System of Record, Ownership, Stewardship

How should System of Record be determined?

1. Data Origination
2. Data Linkages (Inbound and Outbound)
3. Trust Level

Who should be the Data Owner?

1. Stakeholder from Business side
2. Stakeholder with most interest in data accuracy
3. Stakeholder with most to lose, with incorrect data

ACCOUNTABILITY : Data Accuracy from Business Perspective

Who should be the Data Steward?

1. Stakeholder from Technology side
2. Stakeholder with most knowledge of data quality
3. Stakeholder with ability to measure, monitor and fix inconsistencies.

ACCOUNTABILITY : Day to Day Monitoring and Error Correction
Criteria for Understanding Data Inconsistencies

QUESTION:
How are data inconsistencies identified?

Technical Criteria

1. Data Profiling Analysis
   • Valid Values
   • Valid Formats
   • Valid Patterns
   • Valid Lengths

2. Data Quality Analysis.
   Create Data Quality Metrics for
   • Data Accuracy
   • Data Completeness
   • Data Timeliness
   • Data Conformity

Business Criteria

1. Current Data Stewards should confirm inconsistencies found in technical analysis.

2. Data Stewards should provide mechanisms for fixing inconsistency.
Data Governance - *Goal*

- Business teams are doing marketing, manufacturing, sales...
- Data governance is implicitly practiced across the business teams today

Goal: **Formalize Data Governance**
Data Governance - *Style*

- Take cue from corporate culture
- Top-down with a burning imperative
- Bottom-up rally the troops, lead by example
Data Stewardship Process

Data Steward

- Fix Inconsistencies.
- Implements Business Data Rules.
- Makes process & system changes, e.g. screen edits.
- Performs data modeling and facilitates data definition.

Act

Business Data Owner

- Establishes Data Quality Culture.
- Creates Business Data Definitions.
- Empowers Subject Matter Experts & Data Stewards to correct data defects.
- Establishes Data Quality standards, tolerances & metrics.

Plan

Data Steward

- Raises awareness
- Facilitates cross-functional teams.
- Tracks & Reports Quality Metrics.
- Facilitates Root Cause Analysis
- Measures effectiveness of improvements.

Check

Information Quality Leader

- Performs Root Cause Analysis.
- Gains approval for improvement plans.

Do

Adapted from the PDCA Cycle. Originally conceived by Walter Shewhart in 1930's, and later adopted by W. Edwards Deming. The model provides a framework for the improvement of a process or system.
Closing Thoughts
Closing Thoughts - *Change Isn’t Easy*

[Diagram showing the process of change with stages like Begin Executing Change Plan, Change Plan?, Train Key Personnel in Change Leadership Skills, Severity of Reaction to Change, Valley of Despair, Pity City, End of Old Way, Climb to Productivity, and Duration of disruption due to change.]

*Brigdes, Enhancements From Val Larson 2002
iSixSigma LLC 2002*
Closing Thoughts

• Obtain a License to Practice
  - Top-Down Program Charter
• Gain Buy-In from Your Peers
• Stand on the Shoulders of Others
• Deliver Value Every 90 Days
• Good Places to Start
  - Customer-Facing Data
  - Link to an Initiative
  - Service Oriented Architecture Pre-work