Data Quality & MDM
Costs of making decisions on bad data

Vincent Lam
Marketing Director
What was the data issue?
A universal and growing problem

Cost Of Bad Data – a few examples

• Poor quality customer data costs U.S. businesses $611 billion ($750 Billion in 2013 dollars) a year (5% of GDP) - *TDWI Report Series, 2002*

• Poor data quality costs the typical company at least 10% of revenue; 20% is probably a better estimate. *DM Review, 2004*

• Gartner estimates that more than 25% of critical data within large businesses is somehow inaccurate or incomplete. *InformationWeek 2006*

• For insurance the estimated the cost (of bad data) is between 15-20% of operating revenues. *Insurance Data Management Association, 2010*

• Bad data cost around 20% revenues/operating budget. *Australian IT news, 2011*

• A 1% error rate can more than double the cost of all transactions
Beyond Operational Inefficiencies

- **Identities**
  - Duplicate IDs
  - Wrong Addresses
  - Multiple Names/Aliases

- **Risk Management**
  - Auditing & Reporting
  - Privacy Concerns (FERPA Legislation)
  - Fraud Prevention (e.g. Purchasing Card)

- **Finance & Accounting**
  - Inaccurate Statements (duplicates, multiple systems)
  - Inconsistent View of Data

- **Science**
  - NASA lost the $125M Mars Climate Orbiter because one group of engineers used kilograms and meters, while the other group used pounds and feet.
How do we solve this problem?
River Streams and Data Streams

1st order streams
2nd order streams
3rd order streams

* (s-cool revision website)
Observations

- Data originates from different sources
  - Not uniform
  - Different latencies

- Data is joined with other data
  - Potential contamination

- Processes depend on data and create data
  - Bad data breeds bad data

- Output is combination of data and processes
  - Archived data is impacted and unreliable
Our 3 Step Approach to a Complex Problem

- **Information Access**
  - Real-Time and Batch Information Movement
  - Access any system or data on any platform

- **Data Quality**
  - Comprehensive Data Profiling
  - Correct Data Quality issues before they propagate

- **Master Data Management**
  - Centralize the management of information
  - Control the information throughout to organization
# Sources, Latencies, Formats, and Types of Data

<table>
<thead>
<tr>
<th>ERP/Financials</th>
<th>Industry</th>
<th>Legacy Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ariba</td>
<td>HIPAA</td>
<td>CICS</td>
</tr>
<tr>
<td>I2</td>
<td>CIDX</td>
<td>IMS</td>
</tr>
<tr>
<td>JD Edwards</td>
<td>HL7</td>
<td>VSAM</td>
</tr>
<tr>
<td>Lawson</td>
<td>RNIF</td>
<td>.NET</td>
</tr>
<tr>
<td>Manugistics</td>
<td>SWIFT</td>
<td>Java</td>
</tr>
<tr>
<td>Microsoft</td>
<td>1Sync</td>
<td>TUXEDO</td>
</tr>
<tr>
<td>Oracle</td>
<td></td>
<td>etc</td>
</tr>
<tr>
<td>SAP</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFA/CRM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amdocs/Clarify</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMC/Remedy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSDynamics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oracle/Siebel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salesforce.com</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Warehouse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DB2</td>
<td></td>
<td>B2B</td>
</tr>
<tr>
<td>ETL</td>
<td></td>
<td>Internet EDI</td>
</tr>
<tr>
<td>Oracle/Essbase</td>
<td></td>
<td>Legacy EDI</td>
</tr>
<tr>
<td>MS SSAS/OLAP</td>
<td></td>
<td>MFT</td>
</tr>
<tr>
<td>Netezza</td>
<td></td>
<td>Online B2B</td>
</tr>
<tr>
<td>SAP BW</td>
<td></td>
<td>XML</td>
</tr>
<tr>
<td>Teradata</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our 3 Step Approach to a Complex Problem

1. **Information Access**
   - Real-Time and Batch Information Movement
   - Access any system or data on any platform

2. **Data Quality**
   - Comprehensive Data Profiling
   - Correct Data Quality issues before they propagate

3. **Master Data Management**
   - Centralize the management of information
   - Control the information throughout to organization

Data Governance
Data Quality Management

- KPI Definition
- Ongoing Monitoring
- Deviance Identification

Monitoring & Reporting

- Profiling
- Data Impact Analysis
- Improvement Design

Data Understanding

- KPI Definition
- Ongoing Monitoring
- Deviance Identification

Data Cleansing

- Parsing
- Format Standardization
- Content Validation

Data Enhancement

- Content Enrichment
- Match & Merge
- Relationship Association

IT Professional

Business User
Insight from Profiling

- How good or bad is the data?

- Analysis
  - Statistical Analysis
    - Extremes, distribution, frequency
  - Domain Analysis
    - Data types
  - Mask and Group Analysis
    - Formats, groups and dimensions
  - Foreign Key and Dependency Analyses

- Business Rules
  - User-defined business rules adherence

- Compliance over Time
  - Reporting and analysis across multiple data set analyses
  - Web and/or hardcopy report viewing and distribution
Implementing Data Quality in Real-Time

Data Quality
- Pre-Processing
- Transformations and Enrichment
- Data Cleansing
- Matching and Consolidation

Data Integration Process

Data Resources

Metadata Management

Information Sources

Protocols

Apps
- EDI
- Text
- XML
- File
- MQ
- MSMQ
- ODBC
- JDBC

Data Loading
- Master Data Management
- Master Data
- Data Warehouse
- Other Apps
- Partners, etc.
Capabilities to Improve Data Quality

- **Parsing**: Decomposition of fields into component parts.

- **Cleansing**: Modification of data values to meet domain restrictions, integrity constraints or other business rules that define sufficient data quality for the organization.

- **Standardization**: Formatting of values into consistent layouts based on industry standards, local standards, user-defined business rules and knowledge bases of values and patterns.

- **Validation**: Formatting of values into consistent layouts based on industry standards, local standards, user-defined business rules and knowledge bases of values and patterns.

- **Enrichment**: Enhancing the value of internally held data by appending related attributes from external sources.

- **Matching**: Identification, linking or merging related entries within or across sets of data.
iWay Enterprise Information Management Approach

### Information Access
- Real-Time and Batch Information Movement
- Access any system or data on any platform

### Data Quality
- Comprehensive Data Profiling
- Correct Data Quality issues before they propagate

### Master Data Management
- Centralize the management of information
- Control the information throughout to organization
Master Data Management

- Smith, John
- John R. Smith
- J. R. Smith
Master Data

- What is Master Data?
  - Data describing your main business entities
  - Data duplicated in multiple systems
  - Data reused by multiple business processes

- Common Domains of Master Data
  - Customer
    - Customer Data Integration
  - Product
    - Product Information Management
  - Employee
  - Vendor

- Multi-Domain
  - MDM in more than 1 domain
Master Data Management Architectures

Consolidated
- Master is Single Version of Truth
- Data Quality at Master
- Updates occur at Sources
- Updates propagated to Master

Coexistence
- Master is Single Version of Truth
- Data Quality is Ongoing
- Updates occur at Sources or Master
- Updates propagated to other Sources

Registry
- Multiple Versions of Truth
- Data Quality is Ongoing
- Updates occur at Sources
- Keys and Metadata in Registry
- Updates propagated to other Sources (Optional)

Centralized
- Master is Single Version of Truth
- Data Quality at Master
- Updates occur at Master
- Updates propagated to Sources
Data Stewardship – Issue Resolution

- Workflow and issue assignments
- Issue editor allowing to change the data
- Manual or Automated handling
- Monitoring and Reporting of Stewardship processes and results
- Real-Time monitoring
Real-Time + Batch Data Quality with Data Governance

1. Poor quality data are extracted to the DQ issues DB.
2. DQ stewards resolve the issues by specifying the remedy to be done in the source systems.
3. The data in the source systems are updated as suggested by the stewards.

Web Services are used to validate the proposed changes against the rules, and reference data or get the proposed values.

Source Systems

DQ Engine

DQ Plan

Web Services / ISM

Validator Plan / Workflows

DQ Issues DB

Issue Tracking Application

LDAP

Data Guardians

Data Stewards
Bad Data is Expensive
- Impacts processes, customers, and requires correction

One piece of bad data joined with any other data = bad data (Data Streams)

Data comes in many formats, types, and latencies

Utilize real-time upstream data quality capabilities to maximize benefit

Comprehensive Data Quality and MDM technologies complement one another with Data Governance overseeing entire process

Interfaces and tools available for business and IT personnel make improving data quality much easier