Business Impacts of Poor Data Quality: Building the Business Case

David Loshin
Knowledge Integrity, Inc.
www.knowledge-integrity.com
Data Quality Challenges

- Multiple sources
- Process Failures
- Inconsistency
- Repurposing
- Duplication
- Ambiguity
Addressing the Problem

- To effectively ultimately address data quality, we must be able to manage the
  - Identification of customer data quality expectations
  - Definition of contextual metrics
  - Assessment of levels of data quality
  - Track issues for process management
  - Determination of best opportunities for improvement
  - Elimination of the sources of problems
  - Continuous measurement of improvement against baseline
Data Quality Processes

- DQ Inspection
- Identify the Problem
- Assess the Size and Scope
- Act on What is Learned
- Measure the Improvement

- DQ Issues Tracking
- Performance Monitoring
- Resolution Workflow

- DQ Issue Reporting
- DQ Assessment
- DQ Rules & Metrics

- Data Standards
- Metadata Management
- Acceptability Thresholds
- Service Level Agreements
- Remediation actions
Understanding Business Impacts of Data Flaws

- Consider costs and risks related to data use
- Understand data quality expectations
- Defining data validity rules
- Measuring and reporting business-related data quality
Aligning Data Quality with Business Expectations

- Validity of data
  - Completeness
  - Duplicates
  - Consistency
  - Syntax errors

- Business process performance
  - Improved Financials
  - Reduced Risk
  - Increased Productivity
  - Increased Trust
Understanding Business Process Impacts

- For each perceived business problem:
  - What makes this a critical business problem?
  - What are the measurable impacts?
  - How is each impact classified?
  - How is the impact measured?

- Assess the relationship to flawed data:
  - How is the business problem related to an application data issue?
  - How often does the data issue occur?
  - When the data issue occurs, how is it identified?
  - How often is the data issue identified before the business impact is incurred?
Financial Impact Classification

- Overhead
- Depreciation
- Direct
- Financial
- Cash Flow
- Fees
- Combined Ratio
Productivity Impact Classification

- Workloads
- Throughput
- Output Quality
- Supply
- Volume
- Staffing

Productivity

© 2010 Knowledge Integrity, Inc.
www.knowledge-integrity.com
(301)754-6350
Risk Impact Classification

Regulatory compliance

Risk

Fraud

Bureau

CAT Management

Diversity

Model/Parameter risk
Trust Impact Classification

- Forecasting
- Reporting
- Credibility
- Customer satisfaction
- Decisions
- Employee satisfaction
Successive Refinement – Drilling Down

- Risk
  - Diversity
    - Property Locations/CAT Management
    - Variety of Core Products
    - Agency/Broker Distribution
    - Market Dislocations
    - Growth Potential/Risk
    - International Diversification
    - Cross-Industry Diversification
    - Rent
Examples - Insurance

- **Health Insurance company:**
  - Incomplete diagnostic codes skews calculation of premiums, leading to significant decrease in profitability

- **Health Insurance company:**
  - Missing and invalid data impacts ability to calculate amounts of reserves for risk assurance

- **Property & Casualty Insurance company:**
  - Inconsistency of location data impacts assessment of potential expenses involved in insuring the client (regional/local taxes and fees)

- **Property & Casualty Insurance company:**
  - Inconsistent data affects determination of changes in capacity based on exposure in a given geographic area

- **Property & Casualty Insurance company:**
  - Difficulty in resolving unique customer identities impacts evaluation of overall corporate risk
Examples – Financial

- Energy Services Company:
  - Inconsistent supplier data results in early (and incorrect) payments
  - Increased effort for entering the same data multiple times

- DoD Guidelines on Data Quality:
  - “... the inability to match payroll records to the official employment record can cost millions in payroll overpayments to deserters, prisoners, and “ghost” soldiers.”
  - “... the inability to correlate purchase orders to invoices is a major problem in unmatched disbursements.”

- Telecommunications company:
  - Applied revenue assurance to detect underbilling indicated revenue leakage of just over 3 percent of total revenue due to poor data quality
  - Identified 49 misconfigured (but assumed to be unusable) high-bandwidth circuits that could be returned to productive use
Examples – Risk

- Pharmaceutical/Medical Device company
  - Party database used to manage grantees
  - Grantees may also be providers
  - Inability to properly track grantees exposed company to risk of violating Federal Anti-Kickback statute

- Banking industry, credit risk:
  - Low-documentation and no-documentation loans
  - Risk models with vague/incorrect assumptions
Examples – Trust

- Pharmaceutical company:
  - Large investment made in creating front-end sales application fed by back-end database
  - Application clients refused to use new application due to mistrust of back-end database

- Agriculture company:
  - Multiple sales databases conflicted with accounting databases
  - Sales staff did not trust that their commissions were being properly calculated
Assessment and Building the Business Case

- Identify key business performance criteria related to data quality assurance
- Review how data problems contribute to each business impact
- Determine the frequency that each impact occurs
- Sum the measurable impacts/costs associated with each impact incurred by a data quality issue
- Assign an average cost to each occurrence of the problem
- Validate the evaluation with subject matter experts
Data Quality Assessment – Goals & Objectives

- Data quality assessment using data profiling and other analyses to:
  - Identify specific data issues related to known business impacts
  - Introduce a process for assessing objective data quality
  - Support the process of defining data quality dimensions and corresponding data quality validations and measures
  - Correlate discovered issues to business impacts
# Data Quality Assessment – Process

<table>
<thead>
<tr>
<th>Plan</th>
<th>Business Process</th>
<th>Prepare</th>
<th>Analyze</th>
<th>Synthesize</th>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Select business process for review</td>
<td>• Review system docs</td>
<td>• List data sets</td>
<td>• Data extraction</td>
<td>• Review anomalies</td>
<td>• Present anomalies</td>
</tr>
<tr>
<td>• Assess scope</td>
<td>• Review existing DQ issues</td>
<td>• Critical data elements</td>
<td>• Data profiling</td>
<td>• Verify criticality</td>
<td>• Verify criticality</td>
</tr>
<tr>
<td>• Acquire sys docs</td>
<td>• Collate business impacts</td>
<td>• Proposed measures</td>
<td>• Data analysis</td>
<td>• Prioritize issues</td>
<td>• Prioritize issues</td>
</tr>
<tr>
<td>• Identify business impacts</td>
<td>• IP-MAP</td>
<td>• Prepare DQ tools</td>
<td>• Drill-down</td>
<td>• Suggest action items</td>
<td>• Suggest action items</td>
</tr>
<tr>
<td>• Assess existing DQ process</td>
<td></td>
<td>• Note findings</td>
<td>• Prepare report</td>
<td>• Review next steps</td>
<td>• Review next steps</td>
</tr>
<tr>
<td>• Project Plan</td>
<td></td>
<td></td>
<td></td>
<td>• Develop action plan</td>
<td>• Develop action plan</td>
</tr>
</tbody>
</table>

© 2010 Knowledge Integrity, Inc.  
www.knowledge-integrity.com  
(301)754-6350
Phase 1: Plan

Choose a business process impacted by poor data quality

- Select a business process for review
- Assess scope
- Collect system documentation
- Identify business expectations
- Review existing data quality monitoring
- Prepare DQ assessment plan

- Assess scope of applications supplying data
- Acquire business process flows
- Seek references to DQ-related impacts
- Meet with SMEs
- Review resource, schedule requirements

- Identify data sets to be analyzed
- Identify data quality problems
- Gap review
- Document existing edits and validation rules
- Identify staff resources

- Note data demographics
- Acquire metadata
- Identify DQ metrics in use
- Adjust Project plan template
- Adjust schedule

- Acquire additional documentation
- Prepare list of business DQ expectations
- Identify additional documentation
- Adjust schedule

© 2010 Knowledge Integrity, Inc.
www.knowledge-integrity.com
(301)754-6350
Adjusting the Template Plan

<table>
<thead>
<tr>
<th>Task ID</th>
<th>Task Name</th>
<th>Notes</th>
<th>Duration</th>
<th>Start</th>
<th>Finish</th>
<th>Precedent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data Quality Analysis/Assessment</td>
<td></td>
<td>30 days?</td>
<td>Wed 7/1/09</td>
<td>Tue 8/11/09</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Kickoff meeting for data set</td>
<td></td>
<td>1 day</td>
<td>Wed 7/1/09</td>
<td>Wed 7/1/09</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Planning</td>
<td></td>
<td>7 days?</td>
<td>Thu 7/2/09</td>
<td>Fri 7/10/09</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Select a business process for review</td>
<td></td>
<td>1 day</td>
<td>Thu 7/2/09</td>
<td>Thu 7/2/09</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Acquire System Documentation</td>
<td></td>
<td>1 day</td>
<td>Fri 7/3/09</td>
<td>Fri 7/3/09</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Assess scope</td>
<td></td>
<td>1 day</td>
<td>Mon 7/6/09</td>
<td>Mon 7/6/09</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Review objectives and document business</td>
<td></td>
<td>1 day</td>
<td>Tue 7/7/09</td>
<td>Tue 7/7/09</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Review current data quality related monitoring</td>
<td></td>
<td>1 day</td>
<td>Wed 7/8/09</td>
<td>Wed 7/8/09</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Review previously applied data quality</td>
<td></td>
<td>1 day</td>
<td>Thu 7/9/09</td>
<td>Thu 7/9/09</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Prepare and update data quality assessment plan</td>
<td></td>
<td>1 day</td>
<td>Fri 7/10/09</td>
<td>Fri 7/10/09</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Business Process Evaluation</td>
<td></td>
<td>5 days?</td>
<td>Mon 7/13/09</td>
<td>Fri 7/17/09</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Review system and application documentation</td>
<td></td>
<td>1 day</td>
<td>Mon 7/13/09</td>
<td>Mon 7/13/09</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Review existing issues</td>
<td></td>
<td>1 day</td>
<td>Tue 7/14/09</td>
<td>Tue 7/14/09</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Collate business impacts</td>
<td></td>
<td>1 day</td>
<td>Wed 7/15/09</td>
<td>Wed 7/15/09</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Business process information flow mapping</td>
<td></td>
<td>1 day</td>
<td>Thu 7/16/09</td>
<td>Thu 7/16/09</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Review with key Data Strategy Team members</td>
<td></td>
<td>1 day</td>
<td>Fri 7/17/09</td>
<td>Fri 7/17/09</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Preparation</td>
<td></td>
<td>4 days?</td>
<td>Mon 7/20/09</td>
<td>Thu 7/23/09</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>List data sets and data elements to be analyzed</td>
<td></td>
<td>1 day</td>
<td>Mon 7/20/09</td>
<td>Mon 7/20/09</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Identify critical data elements</td>
<td></td>
<td>1 day</td>
<td>Tue 7/21/09</td>
<td>Tue 7/21/09</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Definition of data quality</td>
<td></td>
<td>1 day</td>
<td>Wed 7/22/09</td>
<td>Wed 7/22/09</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Ensure access to data sets, sample size selection (if needed)</td>
<td></td>
<td>1 day</td>
<td>Thu 7/23/09</td>
<td>Thu 7/23/09</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Prepare data quality tools for use</td>
<td></td>
<td>1 day</td>
<td>Mon 7/20/09</td>
<td>Mon 7/20/09</td>
<td></td>
</tr>
</tbody>
</table>
Phase 2: Business Process Evaluation

- Review application documentation
- Review existing issues
- Collate business impacts
- Develop IP-Map
- Review with Data Governance team

Research application architecture and information flows

- Review application documentation
  - Review data dictionary, metadata, and reference data

Review reported incidents, interview SMEs to identify business impacts

Review reported data issues by business impact, and prioritize by impact

Organize reported data issues by business impact, and prioritize by impact

Construct Information-Product map (IP-Map)

- Present developed artifacts (data elements, business impact template, IP-Map)
  - Prioritize issues for review
  - Update plan
  - Obtain resources

Obtain resources
## Business Impacts

<table>
<thead>
<tr>
<th>Impact Category</th>
<th>Examples of issues for review</th>
</tr>
</thead>
</table>
| **Operational Efficiency** | • Time and costs of cleansing data or processing corrections  
                            • Inaccurate performance measurements for employees  
                            • Inability to identify suppliers for spend analysis |
| **Risk/Compliance** | • Missing credit data leads to inaccurate credit risk  
                        • Regulatory compliance violations  
                        • Privacy violations |
| **Revenue** | • Lost opportunity cost  
                     • Identification of high net worth customers  
                     • Increased value from matching against master customer database |
| **Productivity** | • Decreased ability for straight-through processing via automated services |
| **Satisfaction** | • Reduced ease-of-use for staff  
                        • Inability to provide unified billing to customers |
| **Performance** | • Impaired decision-making for setting prices |
## Business Impact Template

<table>
<thead>
<tr>
<th>Issue ID</th>
<th>Data Issue</th>
<th>Business Impact</th>
<th>Measure</th>
<th>Severity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assigned identifier for the issue</td>
<td>Description of the issue</td>
<td>Description of the business impact attributable to the data issue; there may be more than one impact for each data issue</td>
<td>A means for measuring the degree of impact</td>
<td>An estimate of the quantification of the cumulative impacts</td>
</tr>
</tbody>
</table>
Prepare for Data Quality Assessment

- List data sets to be analyzed
- Identify critical data elements
- Define data quality measurements
- Ensure access to data
- Prepare data quality tools for use

Prepare a list of the data sets and the data elements that are to be analyzed.

List critical data elements associated with data sets and data quality issues.

Specify initial set of data quality rules.

Check if there is direct access to the data.

Verify access to data profiling tools.

Verify access to query tools.

Verify tool access to data.

Verify access to extraction tools.

Specify initial set of measurements and solicit acceptability thresholds from the business user.

Specify initial measurement processes to be used.

Extract data if necessary and reformulate for use by profiling tools.

Identify and verify access to alternate analysis tools if needed.
# Documenting Data Elements

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Data Element</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table name</td>
<td>Data element name</td>
<td>• Description of data element</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Specifics of metadata implying business rules or potential data quality issues</td>
</tr>
</tbody>
</table>
Classifying Data Quality Rules

- Standardizing classes of rules for data quality simplifies measurement

- These categories are intended to represent different measurable aspects of data quality
  - Used in characterizing relevance across different source data sets
  - Measurements are taken to review compliance with data quality rules

- Each group within the organization has the freedom to introduce its own data quality rules with their own priorities
# Measuring Data Quality

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Data Element</th>
<th>Rule Category</th>
<th>Measurement process</th>
<th>Acceptability Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table name</td>
<td>Data element name</td>
<td>The class of data quality rule being measured</td>
<td>Method used for measurement, one of: • Data profiling statistics • Data profiling, validation rule • SQL query • Other tools • Combination of techniques • Manual measurement process</td>
<td>Quantified level that demonstrates data meets business expectations</td>
</tr>
</tbody>
</table>
Data Profiling and Analysis

- Extract the data if necessary
- Conduct frequency-based analysis (max, min, high frequency values, outliers, counts, nulls)
- Key analysis, dependency analysis
- Validate use of reference data
- Validate against reference data
- Exact duplicate analyses
- Statistical Analysis
- Table and cross-table profiling
- Cross-table consistency validation using cross-table profiling
- Identify identifying attributes and measure exact duplicate records based on the set of identifying attributes
- Compute mean and standard deviation of numeric data, durations, counts
- Document potential anomalies in observation template
- Document results
- Populate column profiling template
- Populate table profiling template
- Complete data profiling report template
- Provide descriptive detail of potential anomalies
- Identify any outliers and anomalies
- Validate identified business rules and document measures in observation template

© 2010 Knowledge Integrity, Inc.
www.knowledge-integrity.com
(301)754-6350
## Column Analysis Template

<table>
<thead>
<tr>
<th>Table and Column Name</th>
<th>Record Count</th>
<th>Inferred data type</th>
<th># Distinct</th>
<th># Null</th>
<th>% null</th>
<th>Max</th>
<th>Min</th>
<th>Number of patterns</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<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></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>

© 2010 Knowledge Integrity, Inc.
www.knowledge-integrity.com
(301)754-6350
## Observation Template

<table>
<thead>
<tr>
<th>ID</th>
<th>Table/Column Name</th>
<th>Inspection</th>
<th>Reported items</th>
<th>Issues for Review</th>
<th>Fitness Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assigned identifier for issue</td>
<td>Table name and column name(s)</td>
<td>What measure was reviewed</td>
<td>Result of measurement</td>
<td>What needs to be reviewed, next steps</td>
<td>Characterized based on business impact and severity</td>
</tr>
</tbody>
</table>
Synthesis of Results

Review discovered anomalies

Review data profiling and analysis results

Provides details of anomalies and reasons for suspicion

Evaluate data scope and need for additional analysis

Determine if segmentation of tables by reference categories will provide additional insight

Prepare draft data quality assessment report

Prioritize discovered issues

Document fitness review

Determine if other tables should be reviewed

Recommendations for issue remediation, data quality improvements, inspection

Perform any additional analyses on the data

Populate the data quality analysis report

Map potential issues to business impacts

Complete “Issues for Review” column for each observation in the Observation template

Document potential data quality issues

© 2010 Knowledge Integrity, Inc.
www.knowledge-integrity.com
(301)754-6350
## Recommendation Template

<table>
<thead>
<tr>
<th>ID</th>
<th>Priority</th>
<th>Recommendation</th>
</tr>
</thead>
</table>
| Unique identifier | As assigned by business partner | • Driver for the recommendation,  
• Reason for assigned priority, and  
• Specific actions to take |
Data Quality Assessment Report

1. Executive Summary, provides high level overview of the task and the results.
2. Introduction, describes how data profiling and additional analyses were used to assess the quality of selected data sets.
3. Goals, enumerating the specific goals of the analysis, such as “reviewing the quality of data prior to integration in a data warehouse.”
4. Scope, detailing the results of task 1.2 and the business impacts identified tasks under phase 2.
5. Approach, describing the details of the outputs of phase 3, namely profiling and analyses to be performed, identified critical data elements, proposed measurements, and the techniques applied.
6. Data Analysis Results, providing the observations listed in the reasonableness template completed during phase 4.
7. Recommendations, detailing the suggestions resulting from the synthesis of phase 5.
8. Open Issues, in which any unresolved questions are listed.
9. Next Steps, providing the action items resulting from the recommendations review and any requirements to resolve any of the open issues.
10. Additional Supporting Material, such as raw statistics from the column, table, and cross-table templates and any other (non-profiling) analyses to support the recommendations.
Client Review

- Conduct walk-through of the observations in the draft report
- Describe the measure, impact, expectation, and the actual measured value
- Provide justification of why the observation was noted
- Accumulate additional information from business clients and subject matter experts
- Note those discovered issues that are relevant to business impacts
- Prioritize the issues based on perceived business value from business clients
- Present concrete steps that can be taken to eliminate root causes of data issues
- Document action items to be incorporated into the data quality improvement plan
- Develop plan for selected data quality improvement tasks
- Plan for data cleansing
- Define and enforce data standards
- Institute data validity inspection and reporting
- Plan modifications for business processes, applications, or data models
Data Profiling Concepts

- Column analysis
  - Review statistical aspects of values within a column
- Cross-column dependency analysis
  - Review relationships across sets of columns within a single view
- Cross-table redundancy analysis
  - Review overlapping data across columns in different tables

Frequency Distribution
Range Analysis
Distinction
Sparseness, Value Absence, Nulls
Format Evaluation
Cardinality and Uniqueness
Abstract Type Recognition
Overloading
Column Profiling Techniques

- Range Analysis
- Sparseness
- Format Evaluation
- Cardinality and Uniqueness
- Frequency Distribution
- Value Absence
- Abstract Type Recognition
- Overloading
Cross-Column Analysis

- Key discovery
- Normalization & structure analysis
- Derived-value columns
- Business rule discovery
Cross-Table Analysis

- Foreign key analysis
- Synonyms
- Reference data coordination
- Business rule discovery
Ongoing Monitoring Using Data Profiling

- Rule validation can be used to assert data quality expectations throughout the processing flow.
- Use profiling jobs as “probes” across the information flow graph to identify where flaws are introduced.
- Correlate occurrences of errors to documented business impact for prioritization.
Finding Hidden Value with Data Profiling

- Analyze/profile data
- Assess data quality dimensions
- Data quality, Validity, & Transformation rules
- Create monitoring system
- Recommend data transformations

**Application**
- Flat File
- RDBMS
- IMS
- VSAM
Creating a Data Quality Scorecard

Data Quality Scorecard

- Reduced Risk
- Financial Opportunities
- Productivity
- Improved Confidence

© 2010 Knowledge Integrity, Inc.  
www.knowledge-integrity.com  
(301)754-6350
Summary

- Standardized process for performing data quality assessment
- Can be adjusted to support operational and analytical business process consumers
- Allows for identification of key data quality metrics that can feed data stewardship activities, data monitoring, and a data quality scorecard
Questions and Open Discussion

- www.knowledge-integrity.com

- If you have questions, comments, or suggestions, please contact me
  
  David Loshin
  
  301-754-6350
  
  loshin@knowledge-integrity.com