Rapid Data Quality Assessment Using Data Profiling

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

David Loshin, Knowledge Integrity Inc.

David Loshin, president of Knowledge Integrity, Inc. (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of data governance, data quality methods, tools, and techniques, master data management, and business intelligence.

David is a prolific author regarding BI best practices, either via the expert channel at www.b-eye-network.com, "Ask The Expert" at Searchdatamanagement.techtarget.com, as well as numerous books on BI and data quality.

His most recent book, "Master Data Management," has been endorsed by data management industry leaders, and his valuable MDM insights can be reviewed at www.mdmbook.com.

David can be reached at loshin@knowledge-integrity.com.
## Business-Driven Information Requirements

<table>
<thead>
<tr>
<th>Driver</th>
<th>Benefit</th>
<th>Information Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Intelligence</td>
<td>Increased revenue, increased share, cross-sell/up-sell, segmentation, targeting, retention, customer satisfaction, ease of doing business</td>
<td>Unified master customer data, matching/linkage, centralized analytics, quality data, eliminate redundancy</td>
</tr>
<tr>
<td>Risk &amp; Compliance</td>
<td>Compliance, privacy, risk management, accurate response to audits, prevent fraud</td>
<td>Data quality, semantic consistency across business processes, consistency, availability</td>
</tr>
<tr>
<td>Operational Efficiency</td>
<td>Reduced M&amp;A costs, lowered costs, streamlined processes, increased volumes, increased throughput, optimized promotions</td>
<td>STP, eliminate redundant data, functionality, licenses, rules/policy-driven</td>
</tr>
<tr>
<td>Supplier Management</td>
<td>Faster onboarding, reduced vendor count, spend management, improved supply chain management</td>
<td>Matching/linkage, vendor management, 3rd party data integration</td>
</tr>
<tr>
<td>Product Performance</td>
<td>Product design, improved product and brand management time to market, product performance, better manufacturing processes</td>
<td>Unified product data, matching/linkage, centralized analytics</td>
</tr>
<tr>
<td>Organizational Performance</td>
<td>Increased employee productivity, reduced reconciliations</td>
<td>Centralized analytics, unified employee data, inspection, monitoring, control</td>
</tr>
</tbody>
</table>

### Assessing the Quality of Data

**Important questions:**

- What are the most critical business issues attributable to poor data quality?
- What constitutes “poor” data quality?
- How is data quality measured?
- What are the levels of acceptability?
- How are data issues managed?
- What remediation and correction actions are feasible?
- How can we know when the data has been improved?
- How is data quality improvement related to business process performance?
Addressing the Problem

- To effectively ultimately address data quality, we must be able to manage the
  - Identification of business client 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 Assessment for Business Improvement

- Identify key business objectives and corresponding metrics
  - Identify specific data issues related to known business impacts
  - Correlate discovered issues to business impacts
- Data profiling and analysis
  - Understand what you are working with, provide quantified metrics
- Improve automated matching/linkage
  - Reduce false positives, expand universe of identifying attributes, reduce need for manual intervention
- Institute managed data quality
  - Collect organizational data requirements, data inspection and control, incident management, data quality scorecards
Analysis Process: Correlating Business and Data Issues

- **Business Impact Analysis**
  - Identify business data issues
  - Prioritize impacts
  - Identify critical data elements
  - Correlate data dependencies and business impacts

- **Empirical Analysis**
  - Statistical analysis of actual existing data
  - Identification of potential anomalies
  - Validation of known expectations

- **Data profiling**

Tasks

- Review business process, use cases, data dependence and issues with customers’ data
- Isolate and quantify business impacts
- Identify critical mandatory elements and data quality expectations
  - Examples: customers appear only once in data set
- Information product mapping
- Identify source tables
- Profile critical attributes from source data
- Report potential anomalies
- Review potential anomalies with clients to
  - De-emphasize criticality (“Low priority”)
  - Isolate for further review and analysis (“potential problem”)
  - Select for remediation (“definite problem”)

© 2009 Knowledge Integrity, Inc.
www.knowledge-integrity.com
(301) 754-6350
### 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></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></td>
</tr>
<tr>
<td>Identify business impacts</td>
<td>Information production flow</td>
<td>Prepare DQ tools</td>
<td>Drill-down</td>
<td>Suggest action items</td>
<td></td>
</tr>
<tr>
<td>Assess existing DQ process</td>
<td></td>
<td></td>
<td>Note findings</td>
<td>Prepare report</td>
<td></td>
</tr>
<tr>
<td>Project Plan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Present anomalies
- Verify criticality
- Prioritize issues
- Suggest action items
- Review next steps
- Develop action plan

### Planning

- Identify team, tasks, resources, level of effort
- Create plan for data quality assessment process
Business Process Evaluation

- Evaluate business impacts attributable to data flaws
- Select the specific business process(es) associated with those business impacts
- Review application system documentation
- Map business flows and production of information “products”

Soliciting Business Data Quality Expectations

- Very straightforward interview questions:
  - How do you use \{customer, supplier, product, \ldots\} data?
  - What are the biggest data issues impeding business success?
  - Why are those the most critical problems?
  - What do you do when you come across a data problem?
  - What are your expectations when you report a problem?
  - Provide any other perceptions about data quality
Potential Business Impacts

- Regulatory or Legislative risk
- System Development risk
- Information Integration Risk
- Investment risk
- Health risk
- Privacy risk
- Competitive risk
- Fraud Detection

- Detection and correction
- Prevention
- Spin control
- Scrap and rework
- Penalties
- Overpayments
- Increased resource costs
- System delays
- Increased workloads
- Increased process times

- Organizational trust issues
- Impaired decision-making
- Lowered predictability
- Impaired forecasting
- Inconsistent management reporting

Categorizing Data Quality Impacts

Monetary
- Operating Costs
- Revenue
- Opportunities
- Cash Flow
- Charges
- Budget/Spend

Confidence
- Trust
- Decisions
- Control
- Forecasting
- Reporting

Risk
- Regulatory
- Credit
- Investment
- Competitiveness
- Development
- Leakage

Compliance
- Government
- Industry
- Privacy
- Health

Satisfaction
- Customer
- Supplier
- Employee
- Market

Productivity
- Workloads
- Throughput
- Processing Time
- Quality
Classifying Business Impacts

<table>
<thead>
<tr>
<th>Impact Category</th>
<th>Examples of issues for review</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Operational Efficiency</strong></td>
<td>• Time and costs of cleansing data or processing corrections</td>
</tr>
<tr>
<td></td>
<td>• Inaccurate performance measurements for employees</td>
</tr>
<tr>
<td></td>
<td>• Inability to identify suppliers for spend analysis</td>
</tr>
<tr>
<td><strong>Risk/Compliance</strong></td>
<td>• Missing data leads to inaccurate credit risk</td>
</tr>
<tr>
<td></td>
<td>• Regulatory compliance violations</td>
</tr>
<tr>
<td><strong>Revenue</strong></td>
<td>• Lost opportunity cost</td>
</tr>
<tr>
<td></td>
<td>• Identification of high value opportunities</td>
</tr>
<tr>
<td><strong>Productivity</strong></td>
<td>• Decreased ability for straight-through processing via automated services</td>
</tr>
<tr>
<td><strong>Procurement Efficiency</strong></td>
<td>• Improved ease-of-use for staff (sales, call center, etc.)</td>
</tr>
<tr>
<td></td>
<td>• Improved ease of interaction for requestors and approver</td>
</tr>
<tr>
<td></td>
<td>• Reduced time from order to delivery</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td>• Impaired decision-making</td>
</tr>
</tbody>
</table>

Using the 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>
Information Production Flow

- Identify key locations in business process flow where there are data dependencies
- Select location(s) for inspection
- Qualify business expectations for data quality at inspection points

Scoping: Identify Critical Data Elements (CDEs)

- Business facts that are deemed critical to the organization
- Example criteria:
  - Contributes to successful completion of operational business activities
  - Is used to support part of a published business policy
  - Is used by one or more external reports
  - Is used to support regulatory compliance
  - Is designated as Protected Personal information (PPI)
  - Is designated critical employee information
  - Is recognized as critical supplier information
  - Is designated as critical product information
  - Is designated as critical for operational decision-making
  - Is designated as critical for scorecard performance

- *Poor quality of Critical Data Elements will negatively impact achieving business objectives*
### Data & Tools: Preparation

<table>
<thead>
<tr>
<th>Asset</th>
<th>Preparation Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETL</td>
<td>• Identify data sources&lt;br&gt;• Develop data extraction scripts</td>
</tr>
<tr>
<td>Data profiling</td>
<td>• Install tool&lt;br&gt;• Training as needed&lt;br&gt;• Verify connectivity to data sources as necessary&lt;br&gt;• Provide data extracts</td>
</tr>
<tr>
<td>Query access</td>
<td>• Provide direct access to source data</td>
</tr>
<tr>
<td>Data mining</td>
<td>• Install tool(s)&lt;br&gt;• Training as needed&lt;br&gt;• Verify connectivity to data sources as necessary&lt;br&gt;• Provide data extracts</td>
</tr>
<tr>
<td>Desktop productivity</td>
<td>• Acquire templates for capturing results&lt;br&gt;• Acquire reporting templates</td>
</tr>
<tr>
<td>Data</td>
<td>• Extract data</td>
</tr>
</tbody>
</table>

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

---

### Using Profiling for Data Quality Assessment

1. Data from systems extracted:

2. Analysts profiled these systems using profiling tool

3. Potential anomalies are noted within tool's repository:
   - The data element in question
   - The potential issue
   - Why it might be an issue

4. Generated reports detailing the set of potential anomalies are presented to representatives of the business users

5. Issues are reviewed and evaluated:
   - Red: definitely an issue
   - Green: not an issue
   - Amber: requires additional review
   - Gray: Not in scope

6. Spreadsheet reviewed for next steps
Data Profiling and Data Analysis

- Column profiling
  - Frequent values, outliers, maximum, minimum, nulls, patterns, overloaded use
- Table and cross-table profiling
  - Dependencies, candidate primary keys, candidate foreign keys, cardinality of relationships, referential integrity
- Additional analysis
  - Mean, median, standard deviations, uniqueness, ranges, reasonableness

```
<table>
<thead>
<tr>
<th>Column Name</th>
<th>Number of Records</th>
<th>Inferred data type</th>
<th>Number Distinct</th>
<th>% null</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Number of patterns</th>
<th>Mean</th>
<th>Median</th>
<th>Standard Deviation</th>
</tr>
</thead>
</table>
```

Observation and Synthesis

- Review potential anomalies
- Describe issues and determine fitness for uses
- Prioritize by severity and opportunity
- Prepare profiling report:
  - Detail inspected item, reported results, issue for review, possible reasons, business implications, business activities affected, etc.
- Determine requirements for deeper analysis
- Provide recommendations for remediation, correction, validation, and other approaches for improvement
Review with Business Owner

- Present data analysis report and associated measurements
- Review and prioritize discovered issues
- Correlate with business impacts
- Select opportunities for
  - Further investigation
  - Mitigation
  - Remediation (elimination of root cause)
  - Improvement

Summary

- Document key business issues that are attributable to poor data quality
- Perform empirical assessment to identify potential anomalies
- Prioritize based on
  - Correlation to business impact(s)
  - Severity of impact
  - Opportunity for improvement
- Scope focus to areas that can feasibly provide tactical improvements and strategic value
- Specify data inspection rules to quantify levels of acceptability
- Institute inspection, monitoring, and reporting
- Provide continuous process for assessment, remediation, reporting of measurable improvement
Next Steps

- Evaluate alternatives for improvement based on findings
- Perform rapid data quality assessment on different data sets, business processes
- Determine business justification for continuous data quality management

Case Study: Insurance (1)

- Large regional property & casualty carrier
- Focus on specialty and high net worth client policies
- Target business process: *client clearance*
- Objective: lower business acquisition risks
- Sample findings of business-critical issues:
  - Misidentification using alternate customer identifiers (tax identifiers, D&B DUNS numbers, internal customer identifiers)
  - Legacy inconsistencies from previous migrations or flawed business processes
  - Multiple names associated with the same client in records sourced through a variety of data creation channels
  - Invalid or missing location data
Case Study: Insurance (2)

- Large regional property & casualty carrier
- Largely focused on auto, then home & life
- Objective: improve quality of customer data for marketing and cross-selling
- Sample findings of business-critical issues:
  - Last name values have extremely low uniqueness
  - Many different patterns (other than the presumed valid ones) for customer number
  - Customer number is 61% unique
  - Instances of records with non-individual names in first/last name fields
  - Birth date is 12/31 or 01/01 unreasonable number of times
  - Records from specific source are missing critical location values (area code, ZIP code, city)
  - Invalid location information

Case Study: Energy Services

- Objective: Understand failures in supply management processes
- Sample findings of business-critical issues:
  - Supplier names marked with “DO NOT USE” in name field
  - Location information is missing
  - Duplication in contact information
  - Records missing identifying attribute values
  - Inconsistency between supplier acquisition system and accounts payable system
Questions?

- www.knowledge-integrity.com
- www.mdmbook.com


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