Business Impacts of Poor Data Quality: Building the Business Case

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

Data Quality Challenges

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

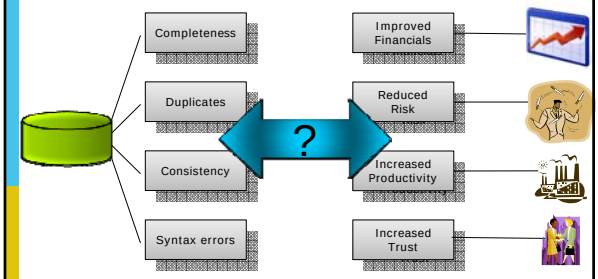
Data Quality Processes

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
- Business process performance



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

Productivity Impact Classification

Risk Impact Classification

Trust Impact Classification

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

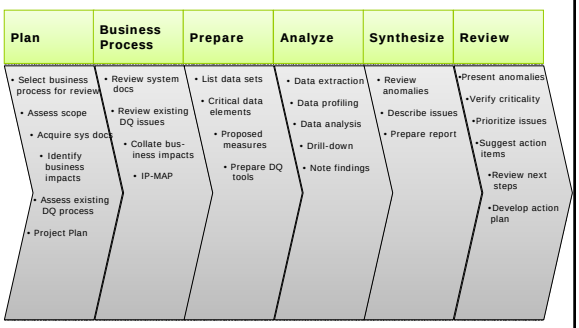
Assessment and Building the Business Case

- 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

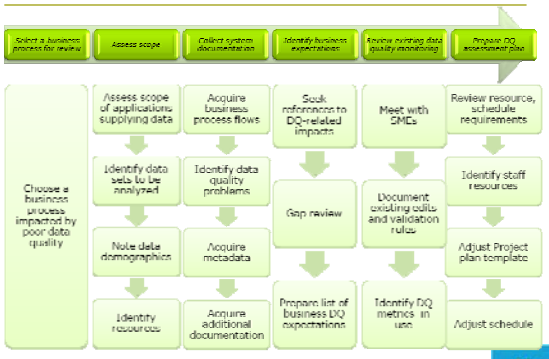
Data Quality Assessment – Goals & Objectives

- 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

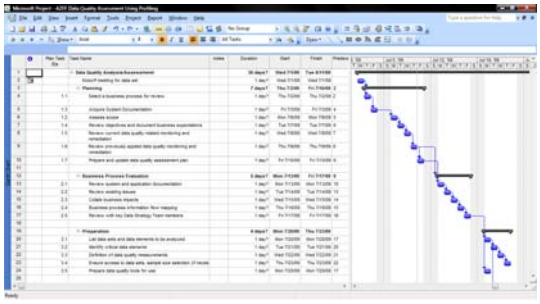
Data Quality Assessment – Process



Phase 1: Plan



Adjusting the Template Plan



Phase 2: Business Process Evaluation

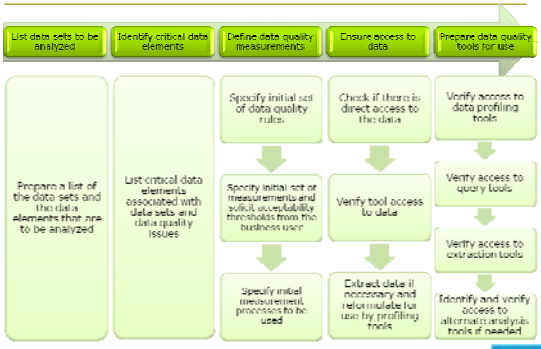
Business Impacts

Impact Category	Examples of issues for review
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

Issue ID	Data Issue	Business Impact	Measure	Severity
Assigned identifier for the issue	*Description of the issue*	*Description of the business impact attributable to the data issue; there may be more than one impact for each data issue*	*A means for measuring the degree of impact*	*An estimate of the quantification of the cumulative impacts*

Prepare for Data Quality Assessment

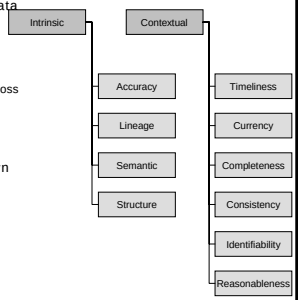


Documenting Data Elements

Data Set	Data Element	Comment
Table name	Data element name	• Description of data element • Specifics of metadata implying business rules or potential data quality issues

Classifying Data Quality Rules

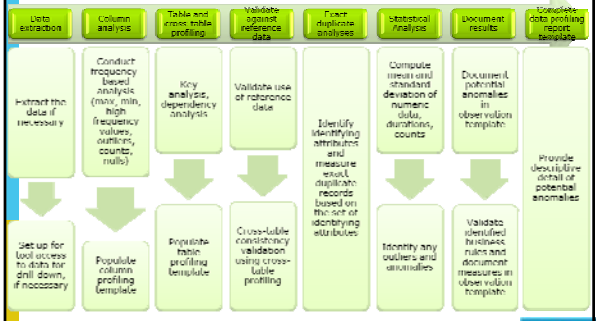
- Standardizing classes of rules for data quality simplifies measurement
- These categories are intended to represent different measurable aspects of data quality
  - Used in characterizing relevance across different source data sets
  - Measurements are taken to review compliance with data quality rules
- Each group within the organization has the freedom to introduce its own data quality rules with their own priorities



Measuring Data Quality

Data Set	Data Element	Rule Category	Measurement process	Acceptability Threshold
Table name	Data element name	The class of data quality rule being measured	Method used for measurement, one of: •Data profiling statistics •Data profiling, validation rule •SQL query •Other tools •Combination of techniques •Manual measurement process	Quantified level that demonstrates data meets business expectations

Data Profiling and Analysis



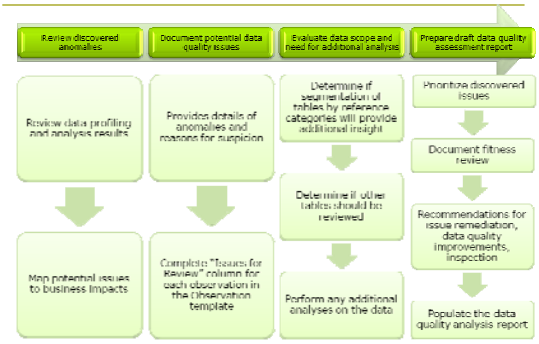
Column Analysis Template

Table and Column Name	Record Count	Inferred data type	# Distinct	# Null	% null	Max	Min	Number of patterns	Mean	Median	Standard Deviation

Observation Template

ID	Table/Column Name	Inspection	Reported items	Issues for Review	Fitness Assessment
Assigned identifier for issue	*Table name and column name(s)*	*What measure was reviewed*	*Result of measurement*	*What needs to be reviewed, next steps*	*Characterized based on business impact and severity*

Synthesis of Results



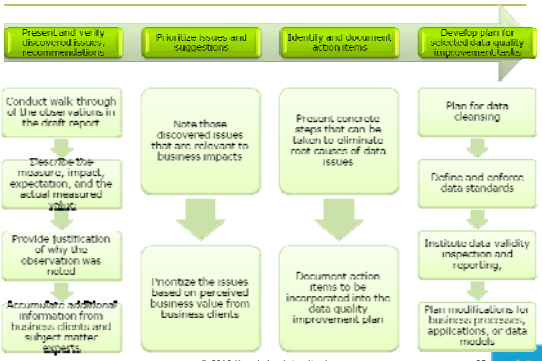
Recommendation Template

ID	Priority	Recommendation
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



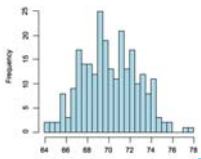
Data Profiling Concepts

- 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

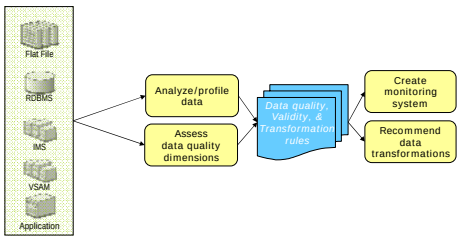
Cross-Table Analysis

- Foreign key analysis
- Synonyms
- Reference data coordination
- Business rule discovery

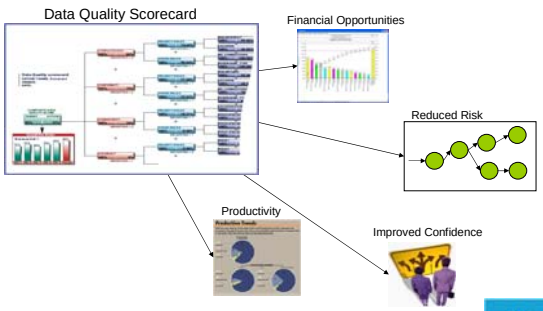
Ongoing Monitoring Using Data Profiling

- 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



Creating a Data Quality Scorecard



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