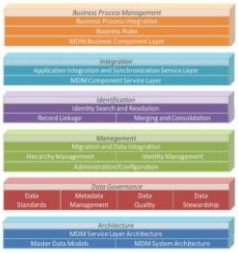
Rapid Data Quality Assessment Using Data Profiling

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

David Loshin, Knowledge Integrity Inc.



MDM Component Model

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

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

Assessing the Quality of Data

Unified master customer data, matching/linkage, centralized analytics, quality data, eliminate redundancy

Data quality, semantic consistency across business processes, consistency, availability

STP, eliminate redundant data, functionality, licenses, rules/policy-driven

Matching/linkage, vendor management, 3rd party data integration

Unified product data, matching/linkage, centralized analytics

Centralized analytics, unified employee data, inspection, monitoring, control

- 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

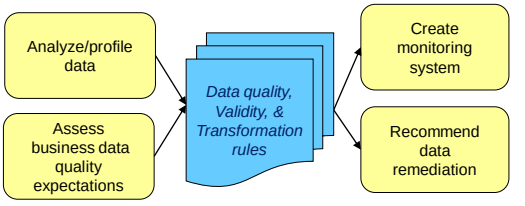
- To effectively ultimately address data quality, we must be able to manage the
  - Identification of business client data quality expectations
  - Definition of contextual metrics
  - Assessment of levels of data quality
  - Track issues for process management
  - Determination of best opportunities 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
- Engage business subject matter experts
- *Empirical Analysis*
  - Statistical analysis of actual existing data
  - Identification of potential anomalies
  - Validation of known expectations
- Data profiling



Tasks

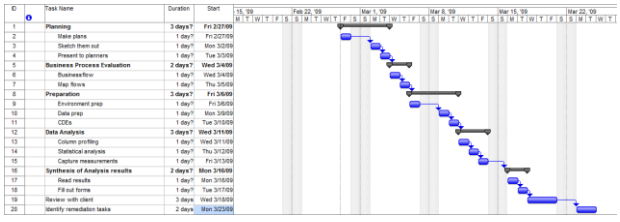
- Review business process, use cases, data dependence and issues with customers' data
- Isolate and quantify business impacts
- Identify critical mandatory elements and data quality expectations
  - Examples: customers appear only once in data set
- Information product mapping
- Identify source tables
- 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")

Data Quality Assessment – Process

Plan	Business Process	Prepare	Analyze	Synthesize	Review
• Select business process for review • Assess scope • Acquire sys docs • Identify business impacts • Assess existing DQ process • Project Plan	• Review system docs • Review existing DQ issues • Collate business impacts • Information production flow	• List data sets • Critical data elements • Proposed measures • Prepare DQ tools	• Data extraction • Data profiling • Data analysis • Drill-down • Note findings	• Review anomalies • Describe issues • Prepare report	•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



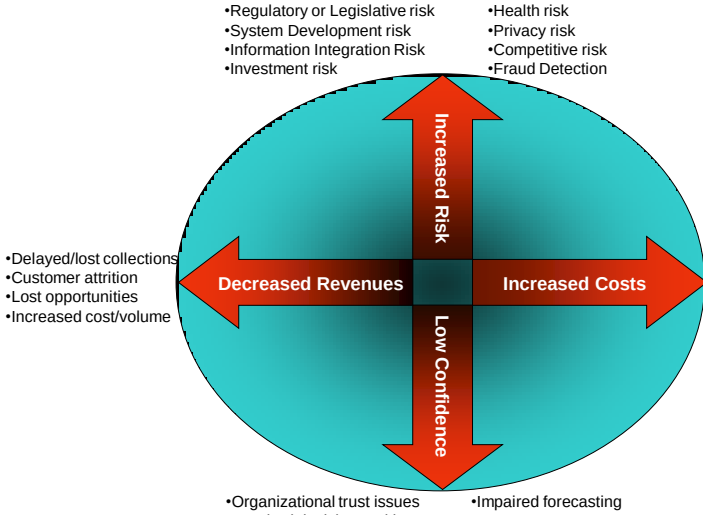
Business Process Evaluation

- 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"

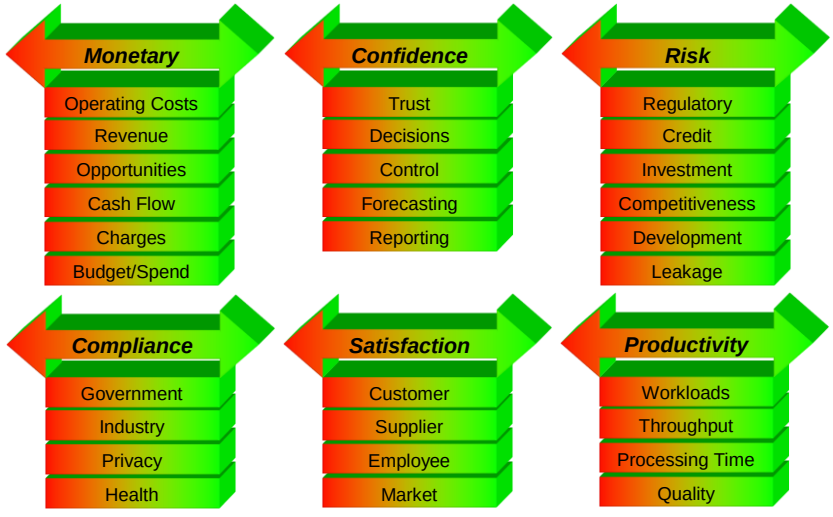
Soliciting Business Data Quality Expectations

- Very straightforward interview questions:
  - How do you use {*customer, supplier, product*, …} 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



Categorizing Data Quality Impacts



Classifying Business Impacts

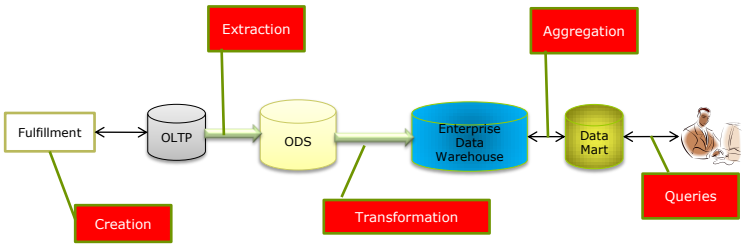
Impact Category	Examples of issues for review
Operational Efficiency	• Time and costs of cleansing data or processing corrections • Inaccurate performance measurements for employees • Inability to identify suppliers for spend analysis
Risk/Compliance	• Missing data leads to inaccurate credit risk • Regulatory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

Using the 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*

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



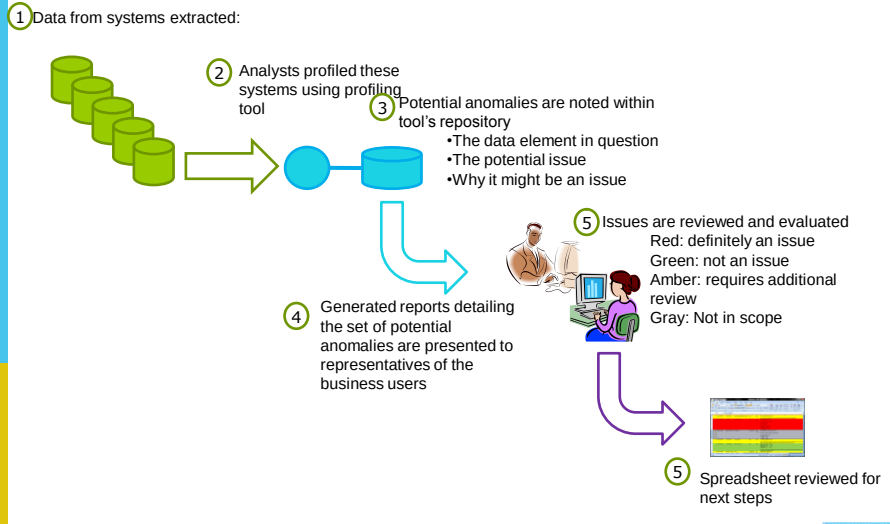
Scoping: Identify Critical Data Elements (CDEs)

- 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

Asset	Preparation Steps
ETL	*•Identify data sources* *•Develop data extraction scripts*
Data profiling	*•Install tool* *•Training as needed* *•Verify connectivity to data sources as necessary* *•Provide data extracts*
Query access	*•Provide direct access to source data*
Data mining	*•Install tool(s)* *•Training as needed* *•Verify connectivity to data sources as necessary* *•Provide data extracts*
Desktop productivity	*•Acquire templates for capturing results* *•Acquire reporting templates*
Data	*•Extract data*

Using Profiling for Data Quality Assessment



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

Column Name	Number of Records	Inferred data type	Number Distinct	Number Null	% null	Maximum	Minimum	Number of patterns	Mean	Median	Standard Deviation

Observation and Synthesis

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

- 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

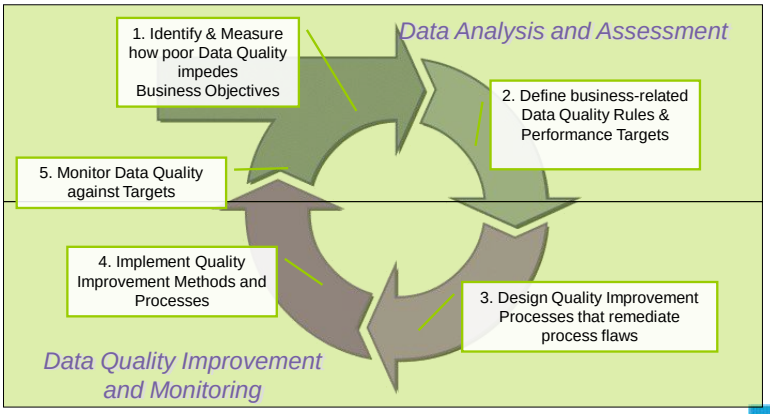
- Present data analysis report and associated measurements
- Review and prioritize discovered issues
- Correlate with business impacts
- Select opportunities for
  - Further investigation
  - Mitigation
  - Remediation (elimination of root cause)
  - Improvement

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

- Download pdf:

http://knowledge-integrity.com/Assets/DQAssessmentDataProflingMIT.pdf

- If you have questions, comments, or suggestions, please contact me

David Loshin
301-754-6350
loshin@knowledge-integrity.com