Data Quality

Linus Rundberg Streuli

Table of contents

1 Introduction

- When organizations are using data for getting insights and making decisions, it is assumed that the data is reliable and trustworthy - or in other words, of high *quality*.
- This assumption is however not always true.
- There are several factors that may lead to poor data quality, among them:
  - lack of understanding about the effects of poor data quality,
  - incomplete documentation,
  - a lack of standards, and
  - a lack of governance.
- Organizations that formally manage the quality of the data have fewer problems than those that leave data quality to chance.

1.1 Risks and Costs Related to Poor Quality Data

- Poor quality data might damage an organization’s reputation, resulting in fines, lost revenue and customers, and negative media exposure.
- Some examples of costs related to poor quality data include:
  - Inability to invoice correctly,
  - Increased customer service calls and decreased ability to resolve them,
  - Revenue loss due to missed business opportunities,
  - Increased exposure to fraud, and
  - Loss due to bad business decisions driven by bad data.

1.2 Business Drivers

- Business drivers for establishing a data quality management program include:
  - Increasing the value of organizational data and the opportunities to use it,
  - Reducing risks and costs associated with poor quality data,
  - Improving organizational efficiency and productivity, and
  - Protecting and enhancing the organization’s reputation.

1.3 Goals and Principles

1.3.1 Goals

- The general goals of a data quality program are:
  - Developing a governed approach to make data fit for purpose based on data consumers' requirements,
  - Defining standards and specifications for data quality controls as part of the data life cycle,
  - Defining and implementing processes to measure, monitor, and report on data quality levels, and
  - Identifying and advocating for opportunities to improve the quality of data, through changes to processes and systems and engaging in activities that measurably improve the quality of data based on data consumer requirements.

1.3.2 Principles

- Data quality programs should be guided by the following principles:
  - **Criticality**: Focus on the data most critical to the enterprise.
  - **Life cycle management**: The data quality should be managed across the data life cycle.
  - **Prevention**: Focus should be on preventing data errors, not simply fixing them.
  - **Root cause remediation**: Problems with data quality should be understood and addressed at their root cause. This often requires changes to processes and the systems that support them.
  - **Governance**: Data governance activities must support the development of high quality data, and vice versa.
  - **Standards-driven**: Data quality requirements should be defined as measurable standards.
  - **Objective measurement and transparency**: Data quality levels need to be measured objectively and consistently, and the methodology should be shared with stakeholders.
  - **Embedded in business processes**: Business process owners must enforce data quality standards in their processes.
  - **Systematically enforced**
  - **Connected to service levels**: Data quality reporting and issues management should be incorporated into Service Level Agreements (SLA).

1.4 Data Quality

- Data is of high quality to the degree that it meets the expectations and needs of the data consumers.
- These expectations are not always known or articulated.
- An ongoing discussion between the data consumers and the people managing the data regarding these expectations and needs is an important part of a data quality program.

1.5 Data Quality Dimensions

- A data quality *dimension* is a measurable feature or characteristic of data, analogous to the dimensions (length, width, height) of physical objects.
- Dimensions provide a basis for measurable rules, such as “98% of customer email addresses are useable”.
- Several authors have published sets of data quality dimensions, such as the Strong-Wang framework (1996) and the dimensions described in Thomas Redman’s *Data Quality for the Information Age* (1996).
- These formulations contain common ideas: dimensions include some characteristics that can be objectively measured, and others that depend heavily on context.

1.5.1 The DAMA UK Core Dimensions (Objective)

- In 2013, DAMA UK published a white paper describing six core dimensions:
  - **Completeness**: The proportion of data stored against the potential for 100%.
  - **Uniqueness**: No entity instance (thing) will be recorded more than once based upon how that thing is identified.
  - **Timeliness**: The degree to which data represent reality from the required point in time.
  - **Validity**: Data is valid if it conforms to the syntax (format, type, range) of its definition.
  - **Accuracy**: The degree to which data correctly describes the “real world” object or event being described.
  - **Consistency**: The abscence of difference, when comparing two or more representations of a thing against a definition.

1.5.2 Further Characteristics (Subjective/Context Dependent)

- The DAMA UK white paper further describes a number of characteristics not defined as dimensions:
  - **Usability**: Is the data understandable, simple, relevant, maintainable and at the right level of precision?
  - **Timing issues**: Is it stable yet responsive to legitimate change requests?
  - **Flexibility**: Is the data comparable and compatible with other data? Does it have useful groupings and classifications? Can it be repurposed? Is it easy to manipulate?
  - **Confidence**: Are data governance, data protection, and data security processes in place? What is the reputation of the data, and is it verified or verifiable?
  - **Value**: Is there a good cost/benefit case for the data? Is it being optimally used? Does it endanger people's safety or privacy, or the legal responsibilities of the enterprise? Does it support or contradict the corporate image or the corporate message?
- The DAMA DMBOK Table 29 (p. 432-433) contains more elaborate descriptions of data quality dimensions.

1.6 Data Quality and Metadata

- Data quality is about meeting expectations.
- Metadata is a primary means of clarifying expectations.
- A metadata repository can contain results of data quality measurements, so that they are shared across the organization.

1.7 Data Quality Improvement Life Cycle

- During its life cycle, data is subjected to a set of processes.
- A process can be defined as a series of steps that turns inputs into outputs.
- At any step during its life cycle, data can be negatively affected: collected incorrectly, dropped or duplicated on its way between systems, aggregated incorrectly, etc.
- Improving data quality requires the ability to assess the relationship between inputs and outputs, in order to ensure that the inputs meet the requirements of the process and that outputs conform to expectations.

1.7.1 Plan-Do-Check-Act

- One general approach to data quality improvement is the Shewhart/Deming cycle, also known as “plan-do-check-act” (see Figure 1 on the next slide).
- The data quality must be measured against standards, and if it does not meet them, the root cause(s) must be identified and remediated, after which the data quality is measured once again.

1.7.2 Plan-Do-Check-Act Figure

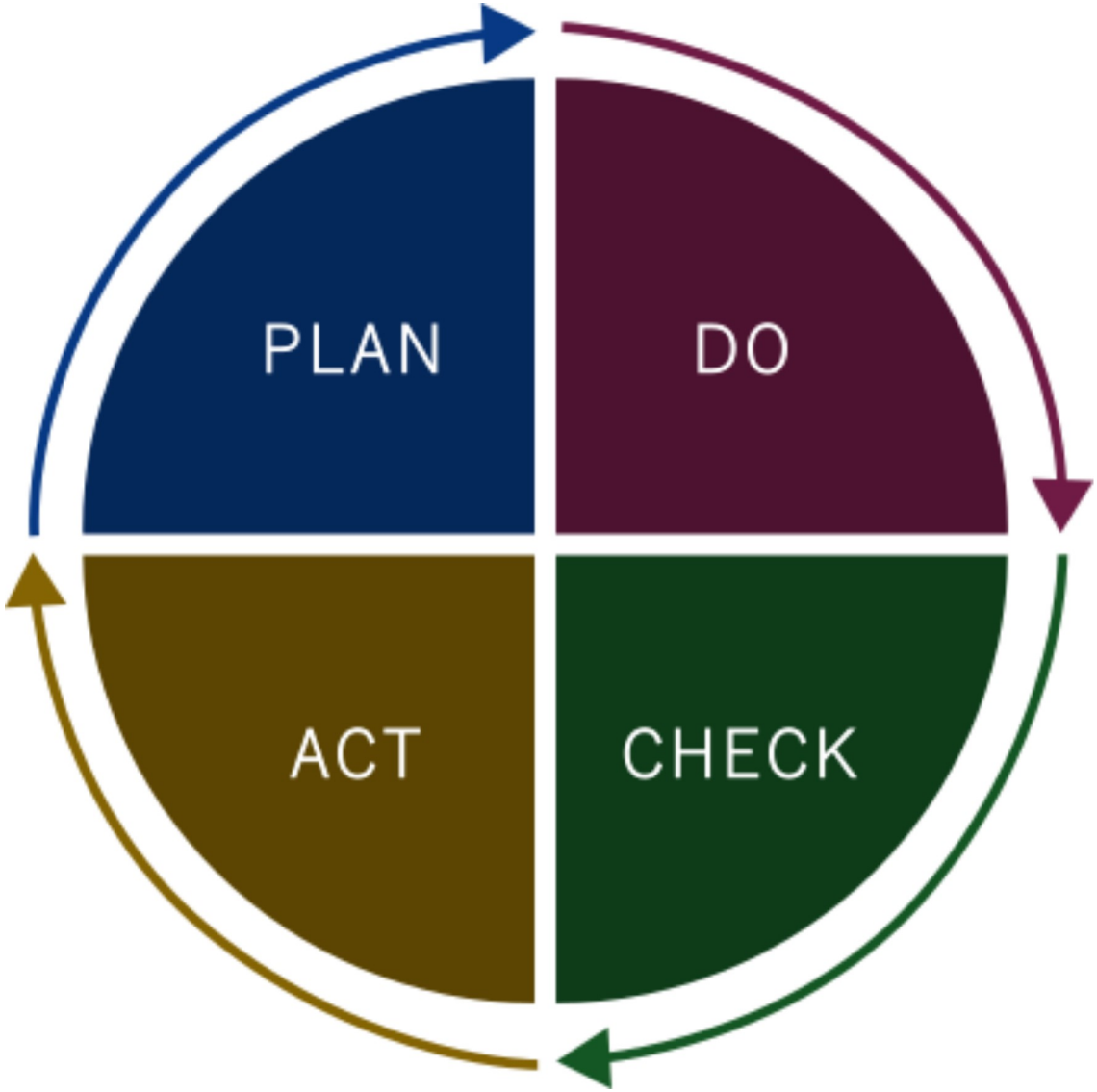


Figure 1: The Shewhart/Deming Cycle

1.7.3 Plan-Do-Check-Act Explained

- In the *Plan* stage, the data quality team assesses the scope, impact and priority of known issues, and evaluates alternatives to address them.
- In the *Do* stage, the data quality team leads efforts to address the root causes of issues and plan for ongoing monitoring of data. The DQ team cooperates with different other teams depending on whether the root causes are technical or non-technical.
- In the *Check* stage, the data quality is actively monitored against requirements. As long as the data quality meets the defined thresholds, no additional actions are required.
- The *Act* stage is for activities to address and resolve emerging data quality issues. As the issues are assessed and solutions are proposed, the cycle begins anew.
- New cycles begin as:
  - Existing measurements fall below thresholds,
  - New data sets come under investigation,
  - New data quality requirements emerge for existing data sets, and
  - Business rules, standards or expectations change.

1.8 Data Quality Business Rules

- Data quality business rules describe how data should exist in order to be useful and usable within an organization.
- Some common simple business rule types are:
  - **Definitional comformance**: Confirm that the same understanding of data definitions is implemented and used properly in processes across the organization.
  - **Value presence and record completeness**: Rules defining the conditions under which missing values are acceptable or unacceptable.
  - **Format compliance**: One or more patterns to specify values assigned to a data element, such as standards for formatting phone numbers.
  - **Value domain membership**: Specify that the value in a data field is included in the accepted values for that field.
  - **Range conformance**: Ensure that a value is within a specified numeric, lexicographic or time range.
  - **Mapping conformance**: Ensure that the value corresponds to an equivalent value from another value domain, such as different ways to express states or counties.
  - **Consistency rules**: Conditional assertions that refer to maintaining a relationship between two or more attributes based on the values of those attributes, such as postal codes and states or cities.
  - **Accuracy verification**: Compare a value against a value from a verified source to verify that they match.
  - **Uniqueness verification**: Ensure that values that must be unique are not duplicated.
  - **Timeliness verification**: Rules that indicate the characteristics associated with expectations for accessibility and availability of data.

1.9 Common Causes of Data Quality Issues

- Common causes behind data quality issues include:
  - Lack of leadership,
  - Data entry processes,
  - Data processing functions, and
  - System design
- One might also find issues caused by fixing issues.

1.9.1 Issues Caused by Lack of Leadership

- Research indicates that many data quality problems are caused by a lack of organizational commitment to high quality data. This in turn stems from a lack of leadership, both in forms of governance and management.
- Within most organizations, differences in data structure, format and use of values is a larger problem than just simple errors.
- Defining terms and creating a common “data language” across an organization is a starting point for getting more consistent data.
- Barriers to effective management of data quality include:
  - Lack of awareness on the part of leadership and staff,
  - Lack of business governance,
  - Lack of leadership and management,
  - Difficulty in justification of improvements, and
  - Inappropriate or ineffecive instruments to measure value.
- The above is adapted from The Leader’s Data Manifesto (https://dataleaders.org/manifesto/)

1.9.2 Issues Caused by Data Entry Processes

- **Data entry interface issues**: Poorly defined data entry interfaces can contribute to data quality issues.
- **Field overloading**: Re-using fields for different business purposes rather than making changes to the data model and user interface results in inconsistent and confusing population of the fields.
- **Training issues**: Lack of process knowledge can lead to data quality issues even if controls and edits are in place. Data processors being incented for speed rather than accuracy likely has impact on the data quality.
- **Changes to business processes**: Data errors will result if an interface is not upgraded to accomodate new or changed requirements. Changes to business rules must be propagated throughout the entire system.
- **Inconsistent business process execution**: Data created through processes that are executed inconsistently is likely to be inconsistent.
    - Inconsistent execution may be due to lack of training or documentations as well as to changing requirements.

1.9.3 Issues Caused by Data Processing Functions

- **Incorrect assumptions about data sources**: Productions issues can occur due to errors or changes, inadequate or obsolete system documentation, or inadequate knowledge transfer.
- **Stale business rules**: Business rules should be periodically reviewed and updated, along with the technical processes for measuring data quality.
- **Changed data structure**: Source systems may change structures without informing downstream consumers (human and system), or without providing enough time to account for the changes.

1.9.4 Issues Caused by System Design 1

- **Failure to enforce referential integrity**: If referential integrity is not enforced, various data quality issues can arise:
  - Duplicate data,
  - Orphan rows that may or may not be included in reports,
  - Inability to upgrade due to restored or upgraded referential integrity requirements, and
  - Inaccurate data due to missing data being assigned default values.
- **Failure to enforce uniqueness constraints**: If there are insufficient checks for uniqueness, data aggregation results can be overstated.
- **Coding inaccuracies and gaps**: Incorrect data mappings and inaccurate rules for data processing leads to data quality issues, ranging from incorrect calculations to data being assigned to, or linked to, improper fields, keys, and relationships.
- **Data model inaccuracies**: If the actual data does not support assumptions in the data model, quality issues such as data loss due to values exceeding field lengths can arise.

1.9.5 Issues Caused by System Design 2

- **Field overloading**: Just like issues caused by data entry processes, re-use of fields for different purposes can result in data quality issues.
- **Temporal data mismatches**: In the abscence of a central data dictionary, multiple systems could implement disparate date formats or timings, which in turn can lead to data mismatch and data loss during data synchronization.
- **Weak master data management**: Immature master data management can lead to choosing unreliable sources for data.
- **Data duplication**: There are two main types of undesirable duplication issues:
  - **Single source - multiple local instances**: For example, instances of the same customer in multiple tables in the same database.
  - **Multiple sources - single instance**: Data instances with multiple authoritative sources or systems of record.

1.9.6 Issues Caused by Fixing Issues

- Manual data patches are made directly on the data in the database, generally written and executed in a hurry to “fix” data in an emergency.
- Such data patches have a high risk of causing further data quality issues and are strongly discouraged.

1.10 Data Profiling

- Data profiling is a form of data analysis used to inspect data and assess quality, using statistical techniques such as:
  - **Counts of nulls**: Identifies if nulls exist and allows for inspection of whether they are allowable or not,
  - **Max/min value**: Identifies outliers such as negative values,
  - **Max/min length**: Identifies outliers or invalids for columns with specific length requirements,
  - **Frequency distribution of values for individual columns**: Enables assessment of reasonability, and
  - **Data type and format**: Identifies level of non-conformance to format requirements, as well as identification of unexpected formats.
- Results from the profiling engine must be assessed by an analyst, who can use the results to confirm known relationships and uncover hidden characteristics and patterns within and between data sets.

1.11 Data Quality and Data Processing

- While data quality improvement efforts often focus on the prevention of errors, data quality can also be improved through some forms of data processing, such as:
  - Data cleansing,
  - Data enhancement,
  - Data parsing and formatting, and
  - Data transformation and standardization

1.11.1 Data Cleansing

- Data cleansing transforms the data to make it conform to quality standards and domain rules.
- It includes detecting and correcting data errors to bring the quality to an acceptable level.
- Ideally, the need for data cleansing should decrease over time, as the underlying root causes of data issues are resolved.
- Such root causes can be addressed by:
  - Implementing controls to prevent data entry errors,
  - Correcting the data in the source system, and
  - Improving the business processes that create the data.

1.11.2 Data Enhancement

- Data enhancement (or enrichment) is the process of adding attributes to a data set to increase its quality and usability.
- Examples of data enhancement include:
  - **Time/date stamps**: Documenting the time the record is created, modified, or retired, can help to track historical events and to isolate time frames during root cause analysis.
  - **Audit data**: Auditing can document data lineage, which is important for historical tracking as well as validation.
  - **Reference vocabularies**: Business specific terminology, ontologies (what the data represents and how it relates to other pieces of data), and glossaries enhance understanding and control.
  - **Contextual information**: Adding context such as location, environment, or access methods and tagging data for review and analysis.
  - **Geographic information**: Enhancement through address standardization and geocoding such as regional coding, coordinate pairs and other kinds of location-based data.
  - **Demographic information**: Customer data can be enhanced by adding age, marital status, gender, income and ethnic coding. Business records can be enhanced by adding annual revenue, number of employees, etc.
  - **Psychografic information**: Data used to segment the target population by specific behaviors, habits, or preferences.

1.11.3 Data Parsing and Formatting

- Data parsing is the process of analyzing data using pre-determined rules to define its content and value.
- Many data quality issues involve situations where variation in data values representing similar concepts introduces ambiguity.
- One example could be customer names, where a good standardization tools would be able to parse the different components of a name (first name, surname, initials, etc.) and rearrange them into a standardized format recognized by other data services.

2 Activities

- As previously stated,
  - Data is of high quality to the degree that it meets the expectations and needs of the data consumers.
- But what are those expectations and needs?

2.1 Define High Quality Data

- Before launching a data quality program, we ask a set of questions to understand the current state and assess organizational readiness for data quality improvement:
  - What do stakeholders mean by “high quality data”?
  - What is the impact of low quality data on business operations and strategy?
  - How will higher quality data enable business strategy?
  - What priorities drive the need for data quality improvement?
  - What is the tolerance for poor quality data?
  - What governance is in place to support data quality improvement?
  - What additional governance structures will be needed?

2.2 Define a Data Quality Strategy

- Improving data quality requires a strategy that accounts for the work that needs to be done and the way people will execute it.
- To ensure that data quality priorities align with business strategy, a framework can be adapted or developed. Such a framework should include methods to:
  - Understand and prioritize business needs,
  - Identify the data critical to meeting business needs,
  - Define business rules and data quality standards based on business requirements,
  - Assess data against expectations,
  - Share findings and get feedback from stakeholders,
  - Prioritize and manage issues,
  - Identify and prioritize opportunities for improvement,
  - Measure, monitor, and report on data quality,
  - Manage metadata produced through data quality processes, and
  - Integrate data quality controls into business and technical processes

2.3 Identify Critical Data

- Any data quality program should start with an importance analysis.
- Critical data can be defined as data that, if it were of higher quality, would provide greater value to the organization and its customers.
- Data can be prioritized based on factors such as regulatory requirements, financial value, and direct impact on customers.
- The importance analysis results in a ranked list of data that can be used to focus the efforts of the data quality team.

2.4 Identify Business Rules

- Business rules describe expectations about the quality characteristics of data.
- For example, a company that wishes to target customers in a certain demographic will want to define business rules stating acceptable standards of quality in demographic fields like age, gender and household income.
- Most people are not used to thinking about data in terms of rules.
- When identifying expectations and needs, it might be necessary to get at the rules indirectly, by asking stakeholders questions about requirements, pain points, how they recognize bad data, etc.
- It is not necessary to know all the rules before assessing data quality. It is an ongoing process.
- Sharing the results of assessments is one of the best ways to get at the rules, as they give stakeholders a new perspective on the data.

2.5 Initial Data Quality Assessment 1

- The goal of an initial data quality assessment is to learn about the data in order to define an actionable plan for improvement.
- It is usually best to start with a small, focused effort.
- Steps include:
  - Define the goals of the assessment,
  - Identify the data to be assessed; focus should be on a small data set, even a single data element, or a specific data quality problem,
  - Identify uses of the data and the consumers of the data,
  - Identify known risks with the data to be assessed, including the potential impact of data issues on organizational processes,
  - Inspect the data based on known and proposed rules, and
  - Document levels of non-conformance and types of issues.

2.6 Initial Data Quality Assessment 2

- Further steps include:
  - Perform additional, in-depth analysis based on initial findings in order to
    - Quantify findings,
    - Prioritize issues based on business impact, and
    - Develop hypotheses about root causes of data issues.
  - Meet with Data Stewards, SMEs, and data consumers to confirm issues and priorities.
  - Use findings as a foundation for planning:
    - Remediation of issues, ideally at their root causes,
    - Controls and process improvements to prevent issues from recurring, and
    - Ongoing controls and reporting.

2.7 Identify and Prioritize Potential Improvements

- The next goal is to apply the improvement process strategically.
- Prioritizing potential improvements requires a combination of data analysis and discussions with stakeholders.
- The steps involved are essentially the same as in the small-scale assessment mentioned in the previous slides: define goals, understand data uses and risks, measure against rules, document and confirm findings with SMEs, and use these findings for prioritizing.

2.8 Define Goals for Data Quality Improvement

- Improvement can take different forms, from simple corrections of the data, to root cause analysis and remediation.
- The strategic focus of a data quality improvement plan should be on addressing root causes and implementing preventive mechanisms to stop bad data from entering the organization in the first place.
- Obstacles in the way of improvement efforts include:
  - System constraints,
  - Age of data,
  - Ongoing projects using questionable data, and
  - Cultural resistance to change.
- To avoid these obstacles, set specific, achievable goals based on consistent quantification of the business value of the improvements to data quality.

2.8.1 Measuring the Value of Quality Improvements

- Showing improvement is done by comparing initial measurements with improved results.
- However, few people care about data quality improvements unless there is a business impact. There must be a positive return of investment for the quality improvements.
- When issues are found, determine ROI of fixes based on:
  - The importance ranking of the data affected,
  - The amount of data affected,
  - The age of the data,
  - The number and type of business processes impacted by the issue,
  - The number of customers, clients, vendors, or employees impacted by the issue,
  - The risks associated with the issue,
  - The costs of remediating root causes, and
  - The costs of potential work-arounds.

2.9 Develop and Deploy Data Quality Operations

- In order to sustain data quality, a DQ program should put in place a plan that allows the team to:
  - Manage data quality rules and standards,
  - Monitor data's ongoing conformance with rules,
  - Identify and manage data quality issues, and
  - Report on quality levels.
- Other activities of the DQ team involves documenting data standards and business rules.

2.9.1 Manage Data Quality Rules

- Data quality rules and standards are a critical form of metadata.
- Rules should be:
  - **Documented consistently**: Establish standards and templates for documenting rules.
  - **Defined in terms of data quality dimensions**: Dimensions of quality help people understand what is being measured.
  - **Tied to business impact**: Standards and rules should be connected directly to their impact on organizational success.
  - **Backed by data analysis**: Rules should be tested against actual data.
  - **Confirmed by SMEs**: Often, it takes knowledge of organizational processes to confirm that rules correctly describe the data. Subject matter experts need to confirm or explain the results of the data quality analysis.
  - **Accessible to all data consumers**: To allow understanding and improvement of the rules, all data consumers must have access to them, as well as a means to ask questions and provide feedback on them.

2.9.2 Measure and Monitor Data Quality

- There are two equally important reasons to implement operational data quality measurements:
  - To inform data users about levels of quality, and
  - To manage risks that may be introduced through changes to business or technical processes.
- Measures intended to inform data consumers will focus on critical data elements and relations that, if they are not sound, will directly impact business processes.
- Measurements related to managing risks should focus on relationships that have gone wrong in the past and may go wrong in the future.

Measurement Results

- Measurement results can be described both at the level of individual rules, and as overall aggregates.
- Each rule should have a standard, target, or threshold index for comparison. This function most often reflects the percentage of correct data, or percentage of exeptions, as calculated by the following equations.
- $\mathrm{ValidDQL(r)} = \dfrac{(\mathrm{TestExecutions(r)} - \mathrm{ExeptionsFound(r)})}{\mathrm{TestExecutions(r)}}$
- $\mathrm{InvalidDQL(r)} = \dfrac{\mathrm{ExeptionsFound(r)}}{\mathrm{TestExecutions(r)}}$
- So if 10000 tests for business rule r found 560 exceptions, then
  $\mathrm{ValidDQ} = \dfrac{9440}{10000} = 94.4\%$, and
- $\mathrm{InvalidDQ} = \dfrac{560}{10000} = 5.6\%$
- Table 30 (p. 452-453) in the DAMA DM-BOK shows examples of how metrics can be organized and presented.
- Table 1 on the next slide shows a simplified example.

Measurements Results Example

Column 1[1]	Column 1 quality dimensions	Column 2[2]	Column 2 quality dimensions	Column 3[3]	Column 3 quality dimensions
555-123-4567		17 King Way		*555-123-4567*	Duplicate value
555-456-1234		22 B Street		555-123-4568	
4567	Rule violation	45 H Lane		555-123-4569	
555-236-8596		4 Parker Road		*555-123-4567*	Duplicate value
555-897-5632		NULL	Missing value	555-123-4530	
NULL	Missing value Rule violation		Missing value	555-123-4545	
3	Rule violation	09876	Suspect value	555-123-4555	
$4/7 = 57.1\%$		$4/7 = 57.1\%$		$5/7 = 71.4\%$	

Table 1: Simple measurement results table (Adapted from https://t.ly/0qR0d)

1. Must contain 7 digits
2. Street address
3. No duplicate values

2.9.3 Develop Operational Procedures for Managing Data Issues

- When issues are found, the DQ team must respond to findings in a timely and effective manner.
- There must be detailed operational procedures for:
  - Diagnosing issues,
  - Formulating options for remediation, and
  - Resolving issues.

Diagnosing Issues

- The objective is to review the symptoms of the data quality incident, trace the lineage of the data in question, identify the problem and where it originated, and pinpoint potential root causes.
- The prodecure should describe how the DQ team would:
  - Review the data issues in the context of the appropriate data flows, and isolate the location in the process where the flaw is introduced,
  - Evaluate whether there have been any environmental changes that would cause errors entering into the system,
  - Evaluate whether or not any other process issues contributed to the error, and
  - Determine whether the data quality has been affected by issues with external data.
- The success of this procedure requires collaboration between DQ analysts, SMEs and other stakeholders.

Formulating Options for Remediation

- Based on the diagnosis, evaluate alternatives for addressing the issue. These may include:
  - Addressing non-technical root causes such as lack of training, lack of leadership support, unclear accountability and ownership, etc.,
  - Modification of the systems to eliminate technical root causes,
  - Developing controls to prevent the issue,
  - Introducing additional inspection and monitoring,
  - Directly correcting flawed data, or
  - Taking no action based on the cost and impact of correction versus the value of the data correction.

Resolving Issues

- Having identified the options, the DQ team must confer with the business data owners to determine which way to best resolve the issue.
- These procedures should detail how the analysts:
  - Assess the relative costs and merits of the alternatives,
  - Recommend one of the planned alternatives,
  - Provide a plan for developing and implementing the resolution, and
  - Implement the resolution.

2.9.4 Documenting Issue Management

- The decisions made during the issue management process should be documented and tracked using an incident tracking system.
- This documentation can provide valuable insights about the causes and costs of data issues.
- The incident tracking system provide important metrics describing the effectiveness of the quality management workflow.
- The incident tracking data also helps data consumers understand the data they are analyzing. It is important to record the modifications made to the data, as well as the reasons behind those modifications.
- Data quality incident tracking requires staff to be trained on how issues should be classified, logged, and tracked.

Supporting Effective Tracking

- **Standardize data quality issues and activities**: Define a standard vocabulary for the concepts used in data quality management. Standardization also helps with measurements, pattern identification, and reporting.
- **Provide an assignment process for data issues**: Drive the assignment process by suggesting which individuals to assign the issue to, based on their specific areas of expertise.
- **Manage issue escalation prodecures**: Specify the sequence of escalation within the data quality Service Level Agreement.
- **Manage data quality resolution workflow**: The data quality SLA specifies objectives for monitoring, control, and resolution of data quality issues. The incident tracking system can support workflow management to track progress with issue diagnosis and resolution.

2.9.5 Establish Data Quality Service Level Agreements

- A data quality SLA specifies an organization's expectations for response and remediation for data issues in each system.
- Data quality inspections as scheduled in the SLA help to identify issues to fix, and over time, reduce the number of issues.
- The data quality SLA also defines the roles and responsibilities associated with performance of operational data quality procedures.

2.9.6 Develop Data Quality Reporting

- The work of assessing data quality and managing data quality issues needs to be shared with the organization.
- Reporting should focus around:
  - Data quality scorecards, which provides a high-level view of the scores associated with various metrics,
  - Data quality trends, which show over time how the quality of the data is measured, and whether trending is up or down,
  - SLA metrics, showing whether data quality operational staff respond to reported quality issues in a timely manner,
  - Data quality issue management, which monitors the status of issues and resolutions,
  - Conformance of the DQ team with data governance policies,
  - Conformance of IT and business teams to data quality policies, and
  - Positive effects of improvement projects.
- Figure 2, Figure 3 and Figure 4 on the following slides show examples of data quality scorecards.

Data Quality Scorecard Example 1

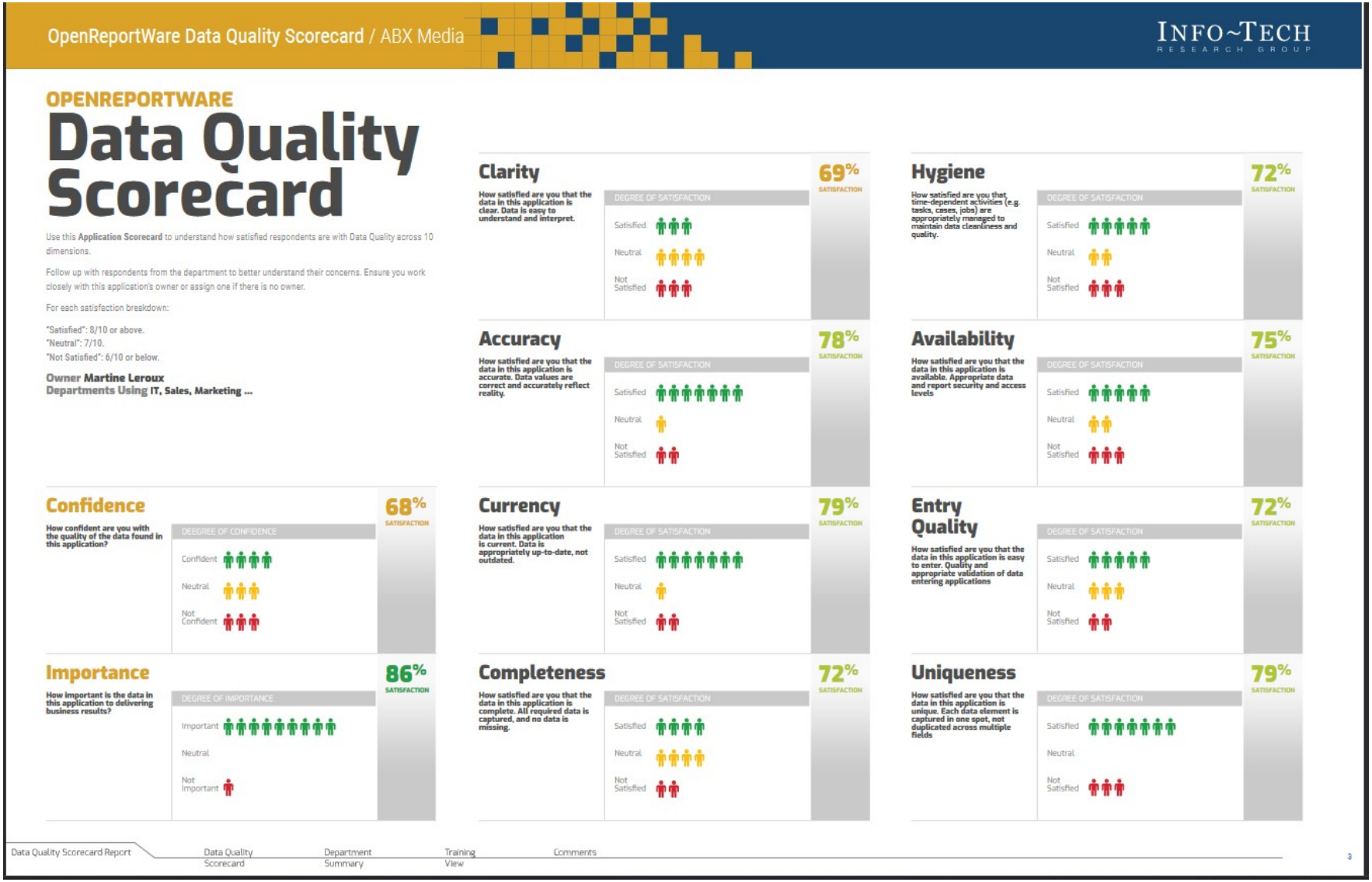
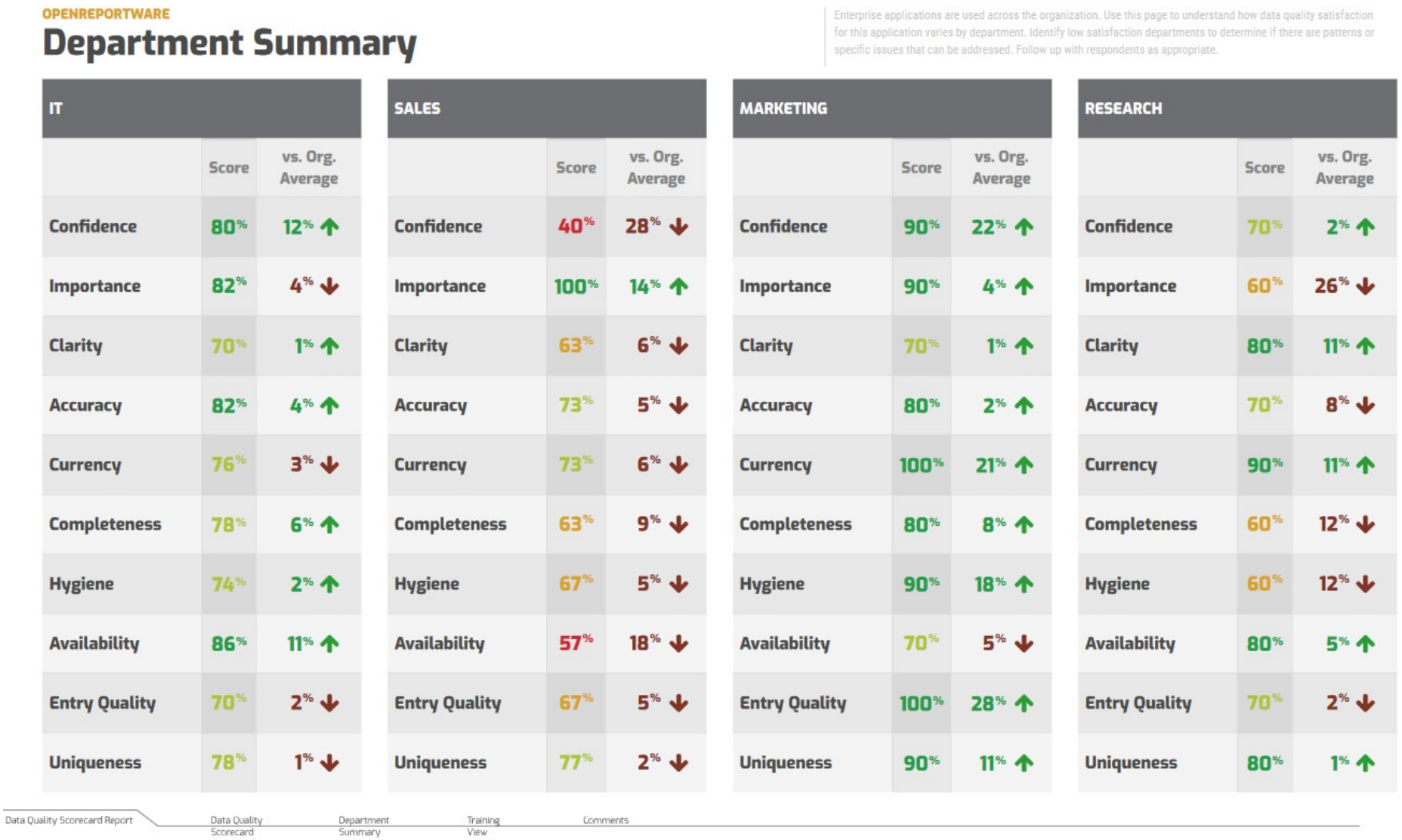


Figure 2: Example of Data Quality Scorecard from Info-Tech Research Group, https://infotech.com

Data Quality Scorecard Example 2



OPENREPORTWARE

Department Summary

Enterprise applications are used across the organization. Use this page to understand how data quality satisfaction for this application varies by department. Identify low satisfaction departments to determine if there are patterns or specific issues that can be addressed. Follow up with respondents as appropriate.

IT	Score	vs. Org. Average
Confidence	80%	12% ↑
Importance	82%	4% ↓
Clarity	70%	1% ↑
Accuracy	82%	4% ↑
Currency	76%	3% ↓
Completeness	78%	6% ↑
Hygiene	74%	2% ↑
Availability	86%	11% ↑
Entry Quality	70%	2% ↓
Uniqueness	78%	1% ↓

SALES	Score	vs. Org. Average
Confidence	40%	28% ↓
Importance	100%	14% ↑
Clarity	63%	6% ↓
Accuracy	73%	5% ↓
Currency	73%	6% ↓
Completeness	63%	9% ↓
Hygiene	67%	5% ↓
Availability	57%	18% ↓
Entry Quality	67%	5% ↓
Uniqueness	77%	2% ↓

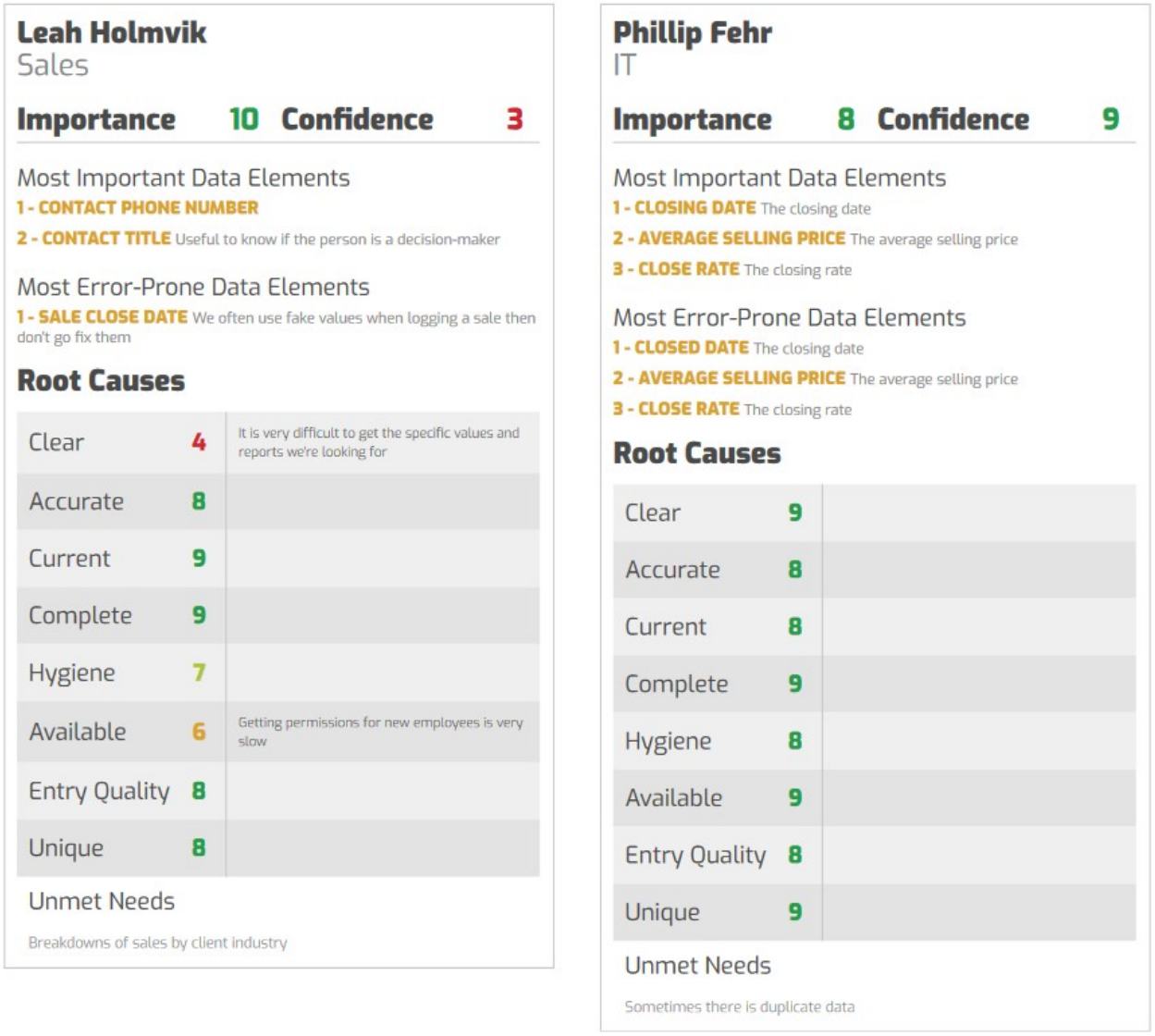
MARKETING	Score	vs. Org. Average
Confidence	90%	22% ↑
Importance	90%	4% ↑
Clarity	70%	1% ↑
Accuracy	80%	2% ↑
Currency	100%	21% ↑
Completeness	80%	8% ↑
Hygiene	90%	18% ↑
Availability	70%	5% ↓
Entry Quality	100%	28% ↑
Uniqueness	90%	11% ↑

RESEARCH	Score	vs. Org. Average
Confidence	70%	2% ↑
Importance	60%	26% ↓
Clarity	80%	11% ↑
Accuracy	70%	8% ↓
Currency	90%	11% ↑
Completeness	60%	12% ↓
Hygiene	60%	12% ↓
Availability	80%	5% ↑
Entry Quality	70%	2% ↓
Uniqueness	80%	1% ↑

Data Quality Scorecard Report | Data Quality Scorecard | Department Summary | Training View | Comments

Figure 3: Example of Data Quality Scorecard from Info-Tech Research Group, https://infotech.com

Data Quality Scorecard Example 3



Leah Holmvik
Sales

Importance 10 **Confidence** 3

Most Important Data Elements

1 - CONTACT PHONE NUMBER

2 - CONTACT TITLE Useful to know if the person is a decision-maker

Most Error-Prone Data Elements

1 - SALE CLOSE DATE We often use fake values when logging a sale then don't go fix them

Root Causes

Clear	4	It is very difficult to get the specific values and reports we're looking for
Accurate	8	
Current	9	
Complete	9	
Hygiene	7	
Available	6	Getting permissions for new employees is very slow
Entry Quality	8	
Unique	8	

Unmet Needs

Breakdowns of sales by client industry

Phillip Fehr
IT

Importance 8 **Confidence** 9

Most Important Data Elements

1 - CLOSING DATE The closing date

2 - AVERAGE SELLING PRICE The average selling price

3 - CLOSE RATE The closing rate

Most Error-Prone Data Elements

1 - CLOSED DATE The closing date

2 - AVERAGE SELLING PRICE The average selling price

3 - CLOSE RATE The closing rate

Root Causes

Clear	9
Accurate	8
Current	8
Complete	9
Hygiene	8
Available	9
Entry Quality	8
Unique	9

Unmet Needs

Sometimes there is duplicate data

Figure 4: Example of Data Quality Scorecard from Info-Tech Research Group, https://infotech.com

More on Data Quality Scorecards

- More on how to craft a data quality scorecard can be found on https://www.datafold.com/blog/crafting-a-data-quality-scorecard.

3 Tools

- Tools should be selected during the planning phase of the data quality program.
- Types of tools include:
  - Data profiling tools,
  - Data querying tools,
  - Modeling and ETL tools,
  - Data quality rule templates, and
  - Metadata repositories.

3.1 Data Profiling Tools

- Data profiling tools produce high-level statistics that enable analysis to identify patterns in data and perform initial assessment of quality characteristics.
- There are many tools on the market. One open source alternative is the Kylo data lake management software platform. (kylo.io)

3.2 Data Querying Tools

- After data profiling, DQ team members also need to query data more deeply to answer questions raised by the profiling results.
- The tools used depend on how the data is stored and managed.

3.3 Modeling and ETL Tools

- The tools used to model the data and create ETL processes can have a direct impact on the quality of the data.
- If they are used with the data in mind, they can enable high quality data.
- On the other hand, if they are used without knowledge of the data, the effects may be the opposite.
- DQ teams should work with development teams to ensure that data quality risks are addressed and data quality policies are being followed.

3.4 Data Quality Rule Templates

- Rule templates allow analysts to capture expectations for data.
- Templates also help to bridge the communications between business and technical teams and can be a pivotal part in creating a common “data language” across an organization.

3.5 Metadata Repositories

- DQ teams should work closely with teams that manage metadata to ensure that data quality requirements, rules, measurement results and documentation of issues are made available to data consumers.

4 Techniques

4.1 Preventive Actions

- The best way to create high quality data is to prevent poor quality data from entering the organization in the first place.
- Approaches include:
  - **Establishing data entry controls**: Create data entry rules that prevent invalid or inaccurate data from entering a system.
  - **Train data producers**: Ensure staff in upstream systems understand the impact of their data on downstream users. Create incentives based on data quality, not just speed.
  - **Define and enforce rules**: Create a “data firewall” with a table of all the data quality rules. This firewall can inspect the quality of the data used in a process and notify analysts if it falls below a certain threshold.
  - **Demand high quality data from data suppliers**: Examine external data provider’s processes to assess how well their data will integrate.
  - **Implement data governance and stewardship**: Ensure roles and responsibilities are defined, and work with data stewards in forming processes and mechanisms for managing data quality.
  - **Institute formal change control**: Ensure all changes to stored data are defined and tested before being implemented.

4.2 Corrective Actions

- Corrective actions are implemented after a problem has occured and has been detected.
- There are three general ways of performing data correction:
  - **Automated correction**: Techniques include rule based standardization, normalization, and correction. Modified values are obtained or generated and committed without manual intervention.
  - **Manually-directed correction**: Automated tools generates corrected values but manual review is required before committing data to storage.
  - **Manual correction**: May be the only option, if automated tools are unavailable or it is determined that changes need to be handled through human oversight. Manual corrections should be made through an interface with data entry controls and edits which leave an audit trail. This method should be avoided.

4.3 Quality Checks and Audit Code Modules

- Create shareable, linkable, and reusable code modules that execute repeated data quality checks and audit processes. Make these modules available to users from a library.
- If a module needs to be updated, any code using the module will be using the updated module.

4.4 Effective Data Quality Metrics

- When developing metrics, DQ teams should account for these characteristics:
  - **Measurability**: Define clear criteria if the metric is subjective, such as “relevancy”. Expected results should be quantifiable within a discrete range.
  - **Business relevance**: Every data quality metric should correlate with the influence of the data on key business expectations.
  - **Acceptability**: Determine whether the data meets business expectations based on specified acceptability thresholds.
  - **Accountability/stewardship**: Metrics should be understood and approved by key stakeholders such as business owners and data stewards, who are notified when the data does not meet quality standards. The business owner is accountable while the data steward takes appropriate corrective action.
  - **Controllability**: If there is no way to respond to a metric being out of range, the metric is probably not useful.
  - **Trending**: Metrics enables an organization to measure data quality improvement over time.

4.5 Statistical Process Control

- Statistical Process Control (SPC) is a method to manage processes by analyzing measurements of variation in process inputs, outputs and steps. It is based on the assumption that when a process with consistent inputs is executed consistently, it will produce consistent outputs.
- It uses measures of central tendency, such as mean, median or mode, and of variability, such as range, variance or standard deviation, to establish tolerances within a process.
- Figure 5 in the next slide shows the primary tool used for SPC - the control chart. It is a time series graph that includes a central line for the average as well as upper and lower control limits.
- In a stable process, values outside of the control limits indicate a special case.
- When a process is in statistical control, a baseline can be established against which changes can be detected.
- SPC is used for control, detection, and improvement.

4.5.1 SPC Chart

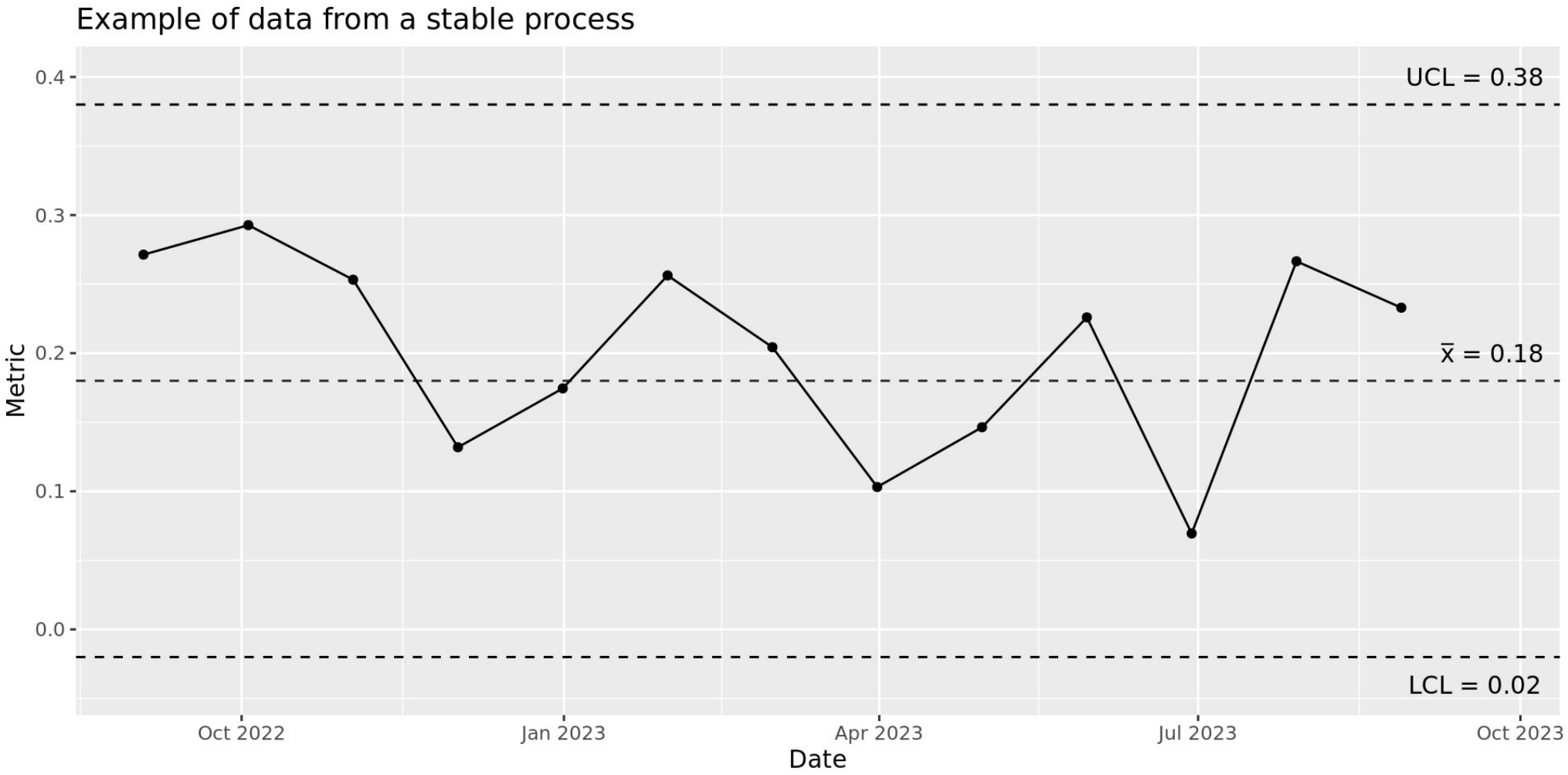


Figure 5: Control Chart of a Process in Statistical Control

4.6 Root Cause Analysis

- A root cause of a problem is a factor that, if eliminated, would remove the problem itself.
- Root cause analysis is a process of understanding factors that contribute to problems and the way that they contribute.
- Common techniques for root cause analysis include Pareto analysis (the 80/20 rule) or Ishikawa (or fishbone) diagram analysis.

5 Implementation Guidelines

- Improving data quality requires changes in how people think and behave toward data. Most data quality program implementations need to plan for:
  - **Metrics on the value of data and the cost of poor quality data**
  - **Operating model for IT/Business interactions**
  - **Changes in how projects are executed**: Project funding must include steps related to data quality.
  - **Changes to business procedures**: The DQ team needs to be able to assess and recommend changes to non-technical, as well as technical, processes that impact the quality of the data.
  - **Funding for remediation and improvement projects**: Data will not fix itself. Document the costs and benefits of remediation projects to help with prioritizing.
  - **Funding for data quality operations**: Sustaining data quality requires ongoing operations to monitor data quality, report on findings, and continue to manage issues as they are discovered.

5.1 Readiness Assessment/Risk Assessment

- Organizational readiness to adopt data quality practices can be assessed by considering the following characteristics:
  - Management commitment to managing data as a strategic asset,
  - The organization’s current understanding of the quality of its data,
  - The actual state of the data,
  - Risks associated with data creation, processing, or use, and
  - Cultural and technical readiness for scalable data quality monitoring.
- Findings from a readiness assessment will help determine where to start and how quickly to proceed.

5.2 Organizational and Cultural Change

- The quality of data will not be improved through a collection of tools and concepts, but through a common mindset focused on the quality of data and what the business and its customers need.
- Employees need to think and act differently if they are to produce better quality data. This require training.
- Training should focus on:
  - Common causes of data problems,
  - Relationships within the organization's data ecosystem and why improving data quality requires an enterprise approach,
  - Consequences for poor quality data,
  - Necessity for ongoing improvement, and
  - Becoming "data-lingual", able to articulate the impact of data on organizational strategy.

6 Data Quality and Data Governance

- A data quality program is more effective when part of a data governance program.
- A governance organization can accelerate the work of a data quality program by:
  - Setting priorities,
  - Identifying and coordinating access to those who should be involved in various data quality-related decisions and activities,
  - Developing and maintaining standards for data quality,
  - Reporting relevant measurements of enterprise-wide data quality,
  - Providing guidance that facilitates staff involvement,
  - Establishing communications mechanisms for knowledge-sharing,
  - Developing and applying data quality and compliance policies,
  - Monitoring and reporting on performance,
  - Sharing data quality inspection results, and
  - Resolving variations and conflicts.

6.1 Data Quality Policy

- Data quality efforts should be supported by and should support data governance policies. Each policy should include:
  - Purpose, scope and applicability of the policy,
  - Definition of terms,
  - Responsibilities of the data quality program,
  - Responsibilities of other stakeholders,
  - Reporting, and
  - Implementation of the policy.

6.2 Metrics

- Measuring and reporting on data quality is a large part of the DQ team's work.
- High-level categories of data quality metrics include:
  - Return of investment,
  - Levels of quality,
  - Data quality trends,
  - Data issue management metrics,
  - Conformance to service levels, and
  - Data quality plan rollout

7 Sources

- DAMA-DMBOK, Chapter 13
- https://t.ly/0qR0d
- https://infotech.com