



Previews of TDWI course books are provided as an opportunity to see the quality of our material and help you to select the courses that best fit your needs. The previews cannot be printed.

TDWI strives to provide course books that are content-rich and that serve as useful reference documents after a class has ended.

This preview shows selected pages that are representative of the entire course book. The pages shown are not consecutive. The page numbers as they appear in the actual course material are shown at the bottom of each page. All table-of-contents pages are included to illustrate all of the topics covered by a course.



# TDWI Data Quality Fundamentals

---

---

All rights reserved. No part of this document may be reproduced in any form, or by any means, without written permission from The Data Warehousing Institute.

# TABLE OF CONTENTS

<b>Module 1</b>	<b><i>Data Quality Concepts .....</i></b>	<b>1-1</b>
<b>Module 2</b>	<b><i>Data Quality Practices and Processes .....</i></b>	<b>2-1</b>
<b>Module 3</b>	<b><i>Data Quality Assessment .....</i></b>	<b>3-1</b>
<b>Module 4</b>	<b><i>Data Quality Improvement .....</i></b>	<b>4-1</b>
<b>Module 5</b>	<b><i>Summary and Conclusion .....</i></b>	<b>5-1</b>
<b>Appendix A</b>	<b><i>Integrity Rules and Data Models .....</i></b>	<b>A-1</b>
<b>Appendix B</b>	<b><i>Bibliography and References .....</i></b>	<b>B-1</b>
<b>Appendix C</b>	<b><i>Exercises .....</i></b>	<b>C-1</b>



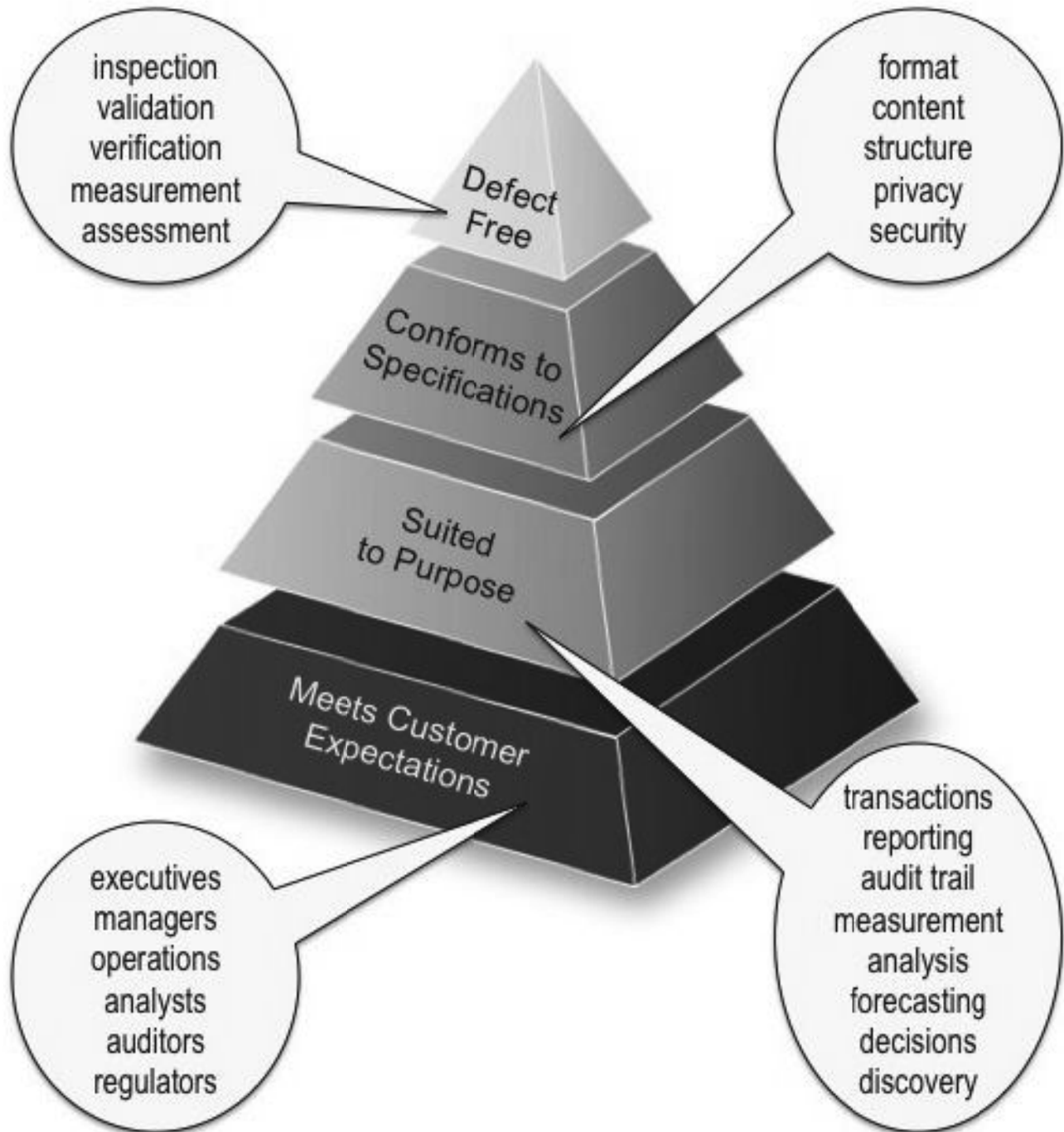
# Module 1

## Data Quality Concepts

Topic	Page
Defining Data Quality	1-2
Dimensions of Data Quality	1-10
Common Causes of DQ Problems	1-26

# Defining Data Quality

## Applying Quality Definitions to Data



# Defining Data Quality

## Applying Quality Definitions to Data

### DATA QUALITY DEFINITIONS

Data quality is a hot topic and data quality management a concern of nearly every enterprise. But little consensus exists about the definition of data quality. Wikipedia states that “data is of good quality when it is complete.”<sup>1</sup> But isn’t there more to data quality than completeness? SearchDataManagement says, “Data quality is the reliability and effectiveness of data, particularly in a data warehouse.”<sup>2</sup> Is data quality less important for non-warehouse data? These are but two examples of the many, incomplete, and sometimes-conflicting definitions of data quality. To manage data quality you must first define it. And the definitions that you use must align with the management actions that you anticipate.

### DATA AND DEFECTS

Defect-free data requires identification of the things that are data defects (more about this later). Then you can manage by inspecting data to find defects, validating and verifying data as free of defects, and measuring defects as part of data quality assessment.

### DATA AND SPECIFICATIONS

Conformance to specifications requires formal data specifications, which may address any or all of data format, content and structure as well as usage-oriented specifications such as those for data privacy and security. Data quality management will test data against specifications.

### DATA AND PURPOSE

Suitability to purpose must consider all of the purposes for which data is used, ranging from business transactions and operational reporting to business intelligence and analytics. Expect the quality criteria to vary widely among the different uses. Variations in quality criteria increase the level of difficulty in data quality management, but attention to them makes quality management efforts more effective and far-reaching.

### DATA AND EXPECTATIONS

Data quality as meeting customer expectations must consider the wide range of data and information consumers. Expect wide variation in the expectations through the range of consumers, both internal and external. Quality management implications of varied expectations are much like those for varied purpose – greater complexity and greater impact.

### MULTIPLE DEFINITIONS

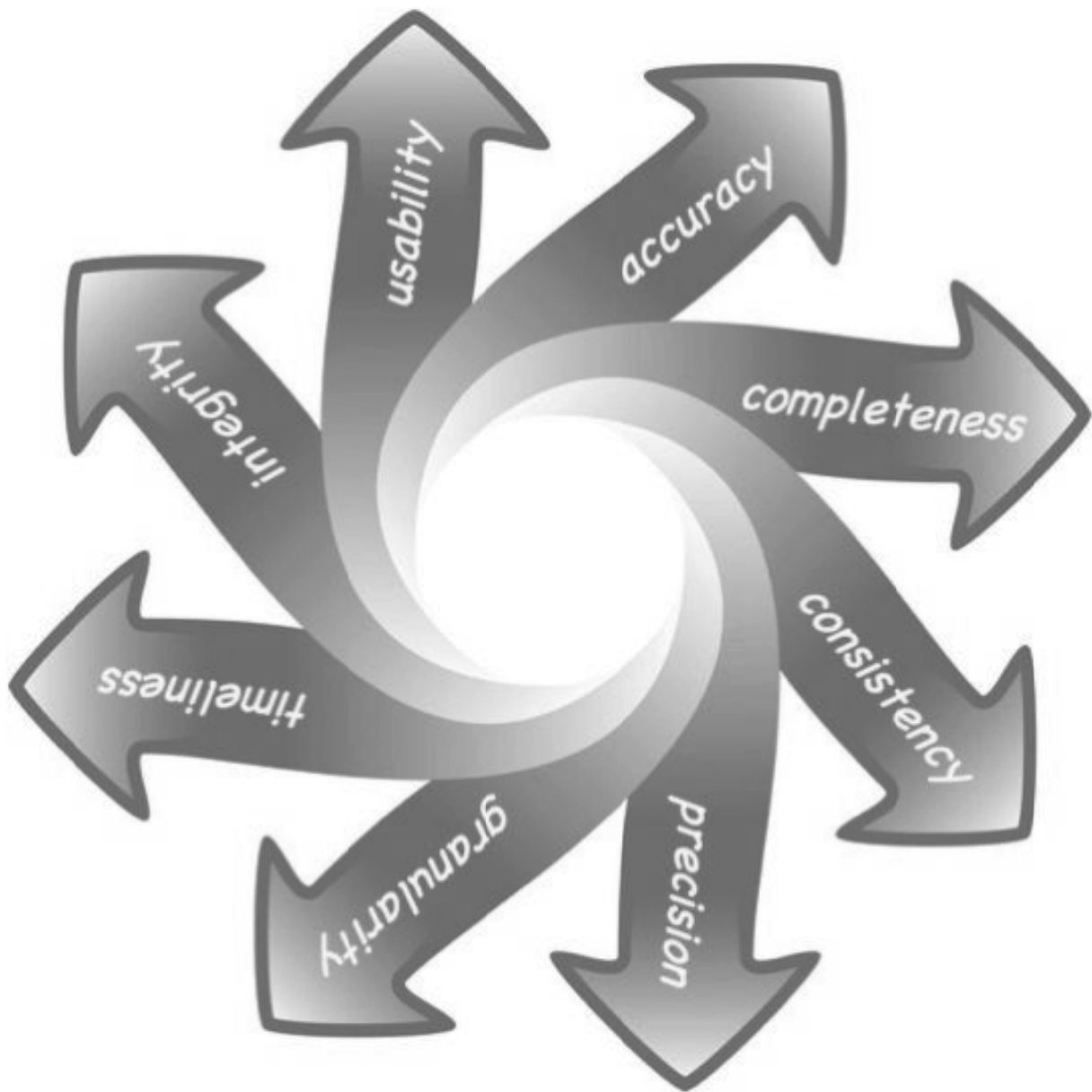
It is likely that every data quality program will include elements of each definition. The four definitions can be applied as a logical sequence where expectations establish purpose, purpose drives specifications, and specifications are the criteria to identify defects.

<sup>1</sup> [http://en.wikipedia.org/wiki/Data\\_quality](http://en.wikipedia.org/wiki/Data_quality) (1/3/2011)

<sup>2</sup> <http://searchdatamanagement.techtarget.com/definition/data-quality>

# Dimensions of Data Quality

## Many Facets of Data Quality



---

# Dimensions of Data Quality

---

## Many Facets of Data Quality

### **CATEGORIES OF QUALITY CRITERIA**

Defining data quality criteria is the first step of the quality management cycle. It is a critical step because these criteria drive all of the steps that follow. But an unstructured approach to quality definition is complex, confusing, and difficult. It helps to begin with a classification system for grouping of criteria.

### **EIGHT DIMENSIONS OF QUALITY**

This course uses a classification system of eight quality dimensions:

- Accuracy – The data represents reality.
- Completeness – All needed data is available.
- Consistency – The data is free of internal conflicts.
- Precision – The data is as exact as is needed.
- Granularity – The data is kept and presented at the right level of detail to meet the needs.
- Timeliness – The data is as current as needed and is retained until no longer needed.
- Integrity – The data is structurally sound.
- Usability – The data is accessible, understandable, and navigable.

Many variations exist in the categories that are used by data quality practitioners. Some combine precision and granularity into a “detail” classification. Some identify auditability (ability to trace back to an original business transaction) as a quality dimension. In some industries, conformity to industry standards for recording or data exchange is important.

A classification system is important. It is equally important that the system fits your business and meets your needs for quality management.

# Common Causes of DQ Problems

## Data Definition

	UNIQUE MEANINGFUL STANDARDIZED GLOBAL VIEW	CLEAR UNAMBIGUOUS CONSISTENT WITH EXAMPLES
	NAME	DESCRIBE
ENTITIES	business language fully qualified standard abbreviations	complete sentences written for the novice non-circular
ATTRIBUTES	business language describes properties use of class words	complete sentences written for the novice illustrated by example
RELATIONSHIPS	business language bi-directional verb form in context	complete sentences written for the novice including cardinality
TABLES	descriptive technical format applied standards	purpose database relationships system relationships
COLUMNS	descriptive technical format applied standards	data typing keys and indexing constraints

# Common Causes of DQ Problems

---

## Data Definition

### COMPREHENSIVE DATA DEFINITIONS

Every data object – entities, attributes, data elements, files, tables, and columns – needs to have a complete definition. Good data definition practices ensure consistency, eliminate confusion, enhance communication, enable data consolidation, and improve overall data quality.

### DATA DEFINITION PRACTICES

In *Data Resource Quality*, Michael Brackett identifies bad habits and good practices for data definition<sup>1</sup>:

#### Bad Habits

non-existent definitions  
 unavailable definitions  
 short definitions  
 meaningless definitions  
 outdated definitions  
 unrelated (nonsensical) definitions

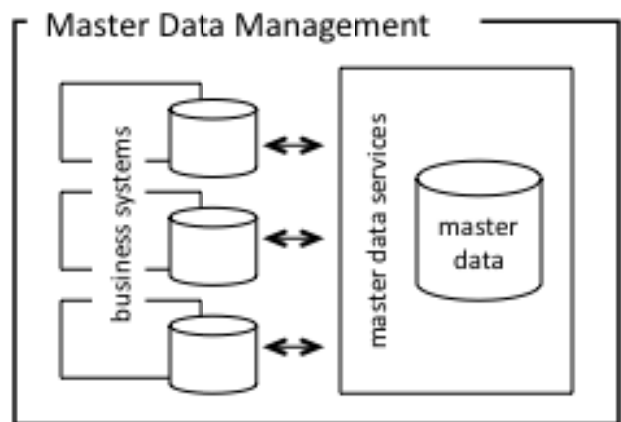
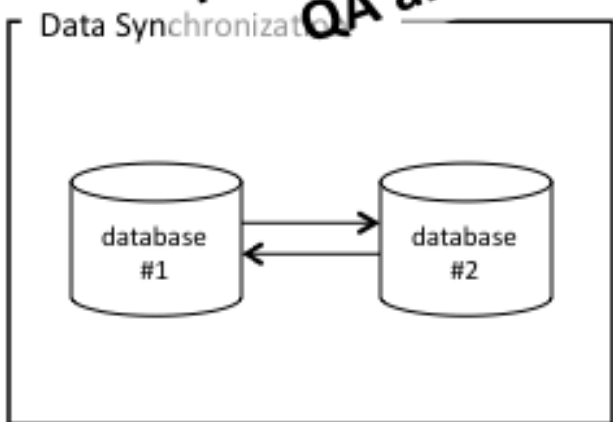
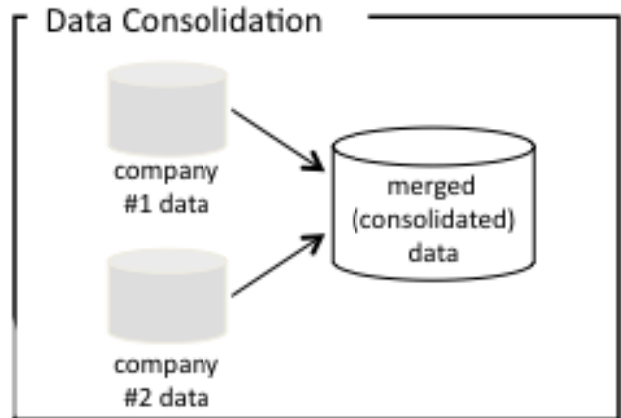
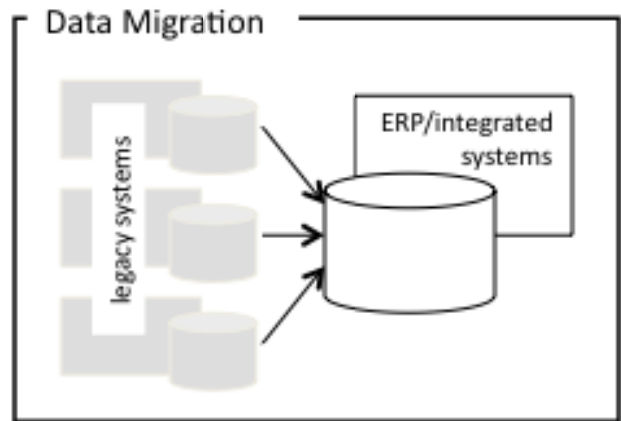
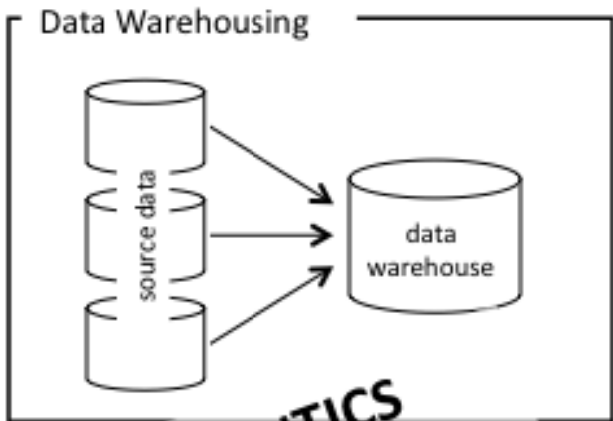
#### Good Practices

business-based meaning  
 thorough with no size limit  
 current & time-independent  
 fundamental (base) definitions  
 inheritance of definitions

<sup>1</sup> *Data Resource Quality: Turning Bad Habits into Good Practices*, pp 51-71, Brackett

# Common Causes of DQ Problems

## Conversion, Consolidation, and Integration



**SEMANTICS**  
**BUSINESS RULES**  
**DATA DEFINITIONS**  
**ENTERPRISE VIEW**  
**TIME VARIANCE**  
**TRANSFORMATION**  
**PROCESS QUALITY**  
**QA and QC**



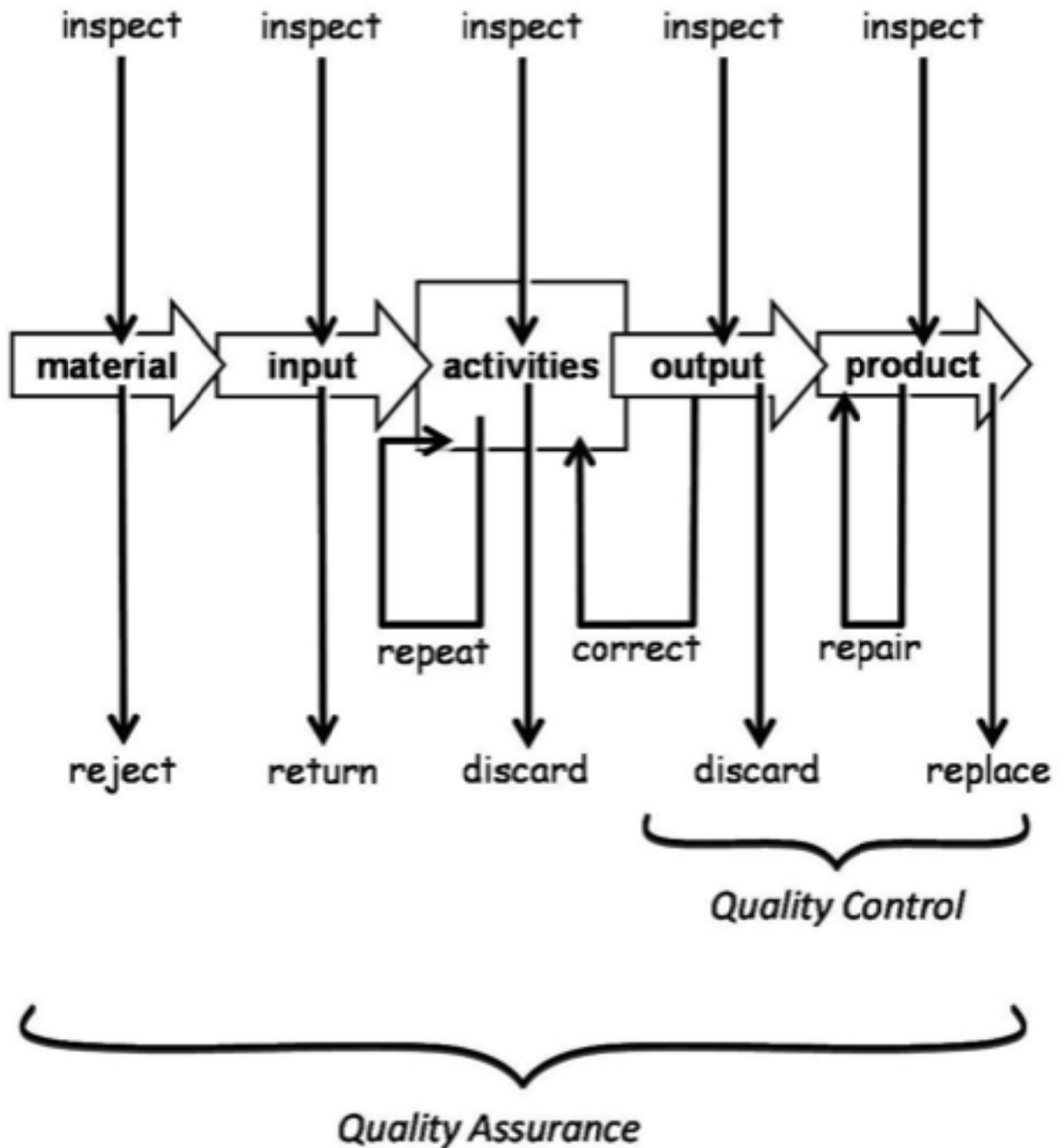
# Module 2

## Data Quality Practices and Processes

Topic	Page
Quality Management Practices	2-2
Quality Management and Data	2-10
Data Quality Organizations	2-16
Data Quality Processes	2-30
Data Quality Tools	2-40

# Quality Management Practices

## Correction and Prevention



# Quality Management Practices

## Correction and Prevention

### RESPONDING TO DEFECTS

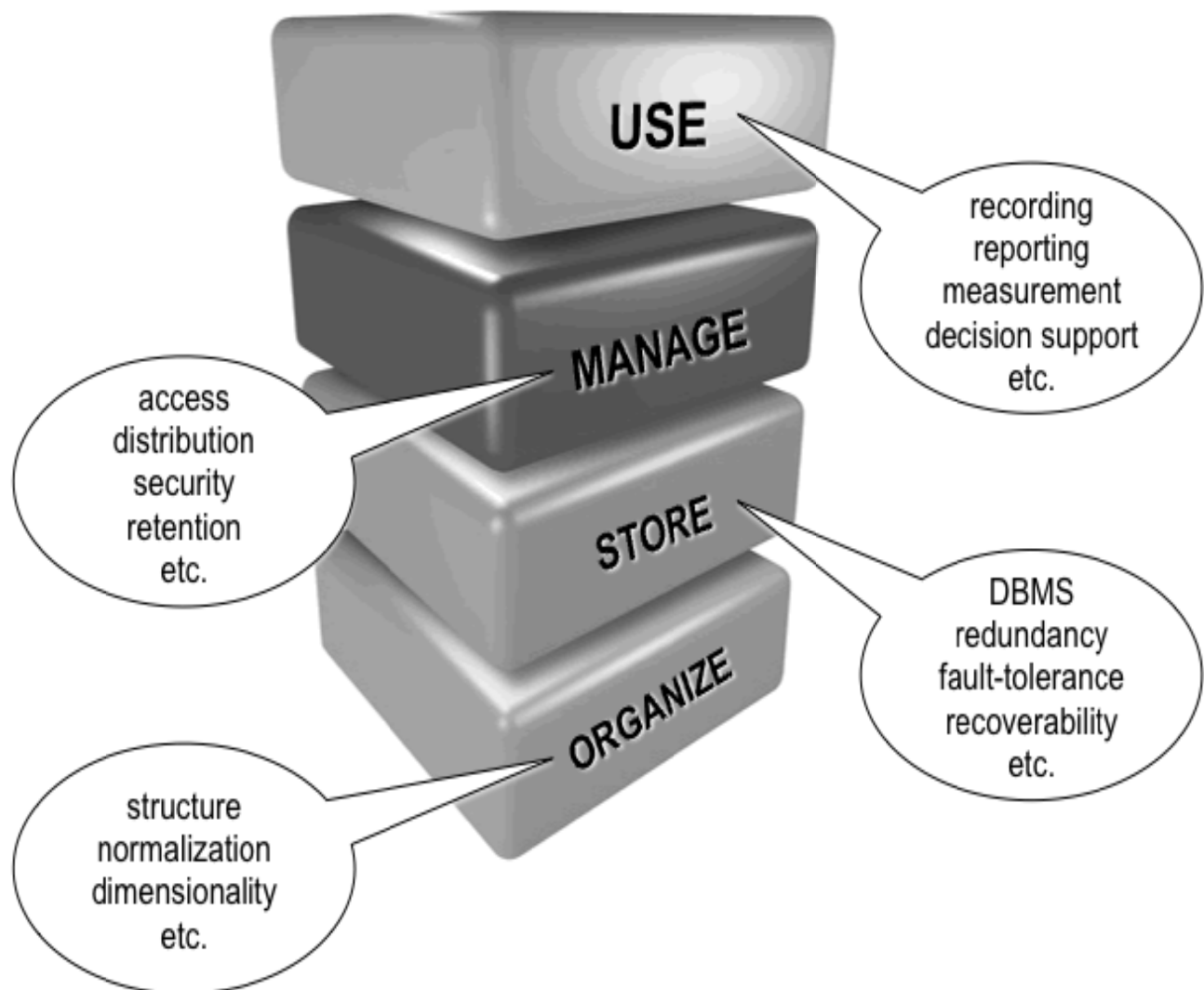
Inspection simply finds defects; it does nothing to respond to them. Response can take many forms depending on the point at which a defect is detected. The range of responses is illustrated on the facing page and described in the table below.

Point of Inspection	Response to Defect
Inspect the product at time of delivery.	Replace the product. <i>Take back the defective product and substitute another in its place. Cost of waste is incurred. Customer dissatisfaction is a risk.</i>
	Repair the product. <i>Take the steps necessary to remove defects from the product. Cost of rework is incurred.</i>
Inspect the product as the last production activity.	Discard the product. <i>Remove the product from the line before delivery. Cost of waste is incurred.</i>
	Correct the defect. <i>Take the steps necessary to remove defects from the product. Cost of rework is incurred.</i>
Inspect work-in-process at defined stages of the production process.	Discard work in process. <i>Remove the defective component or assembly from the line. Cost of waste is incurred.</i>
	Repair work in process. <i>Take the steps necessary to remove defects from defective work. Cost of rework is incurred.</i>
Inspect materials prior to use as the first production activity.	Discard the material (not shown in diagram.) <i>Remove defective material before it enters the line. Cost of waste is incurred.</i>
	Return the material. <i>Send defective material back to the supplier. Cost of work (return processing) is incurred.</i>
Inspect materials upon receipt from suppliers.	Reject the material. <i>Refuse to accept delivery of defective material. Cost is incurred by the supplier.</i>

# Data Quality Organizations

## Data Architecture

Data Architecture describes how data is organized, stored, managed, and used in an organization. Architecture makes it practical to predict, model, and control the lifecycle and flow of data through an enterprise and its systems.



---

# Data Quality Organizations

---

## Data Architecture

### ARCHITECTURE

Data architecture defines how data is organized, stored, managed, and used in an organization or enterprise. It establishes standards of data management that make it practical to predict, model, gauge, and control the flow of data in the system. This is especially important for data quality when data and system components are not fully integrated and are used to serve many audiences and purposes.

### ARCHITECTS

Data architects are responsible to establish the standards, structures, and guidelines to plan and control the collection, storage, and distribution of data throughout an enterprise. Architects define classes of system and data components and describe the roles of and the relationships among those component types. An architect may, for example, define standard roles and relationships of data warehouses and data marts. They may also define the standards for and relationships among subject areas such as customer and product. Regardless of the kinds of components, architectural standards address data organization, storage, management, and use. Thus the data architect's skill set is quite broad, encompassing data definition, data modeling, metadata management, database management systems, data warehousing, business processes, and much more.

### THE IMPACTS

The impact of architecture on data management is suggested by this very concise definition of architecture from Philippe Kruchten: "Architecture encompasses the set of decisions about the system structure."<sup>1</sup> For data systems, then, architecture encompasses decision-making about data structure.

For data developers, stewards, owners, custodians, integrators, and consumers data architecture enables data quality. Architecture provides standards, guidelines, and expectations about how data is managed. Standards, guidelines, and expectations guide good decisions about data organization, storage, management, and use. Good decisions lead to improved data quality.

---

<sup>1</sup> "An Ontology of Architectural Design Decisions in Software-Intensive Systems," Philippe Kruchten (University of British Columbia 2004)

# Data Quality Tools

## DQ Technology

	COTS-DQ	Open Source DQ	DQ Services	DW & BI Tools	Build Your Own
Data Profiling	✓	✓	✓	☐	☺
Address Verification	✓	✓	✓	☐	☹
Data Standardization	✓	✓	✓	☐	☹
Geo-Coding	✓		✓	☐	☹
Matching & Grouping	✓	✓	✓	☐	☺
De-Duplication	✓	✓	✓	☐	☹
Data Transformation	✓	✓		☐	☹
Measurement & Monitoring					☺
Metadata Management	✓	✓		☐	☹
Data Quality Scorecard			✓		☯
Rule-Based Audit			✓		☯
Rule-Based Cleansing			✓		☯

✓ many choices

✓ a few choices

☐ often embedded

☐ sometimes embedded

☺ good approach

☹ bad choice

☺ OK sometimes

☯ best available

---

# Data Quality Tools

---

## DQ Technology

### THE ROLE OF TECHNOLOGY

Tools don't manage data quality. That is something that people must do. But technology fills several important needs including:

- Helping to understand the data
- Documenting, modeling, and metadata management
- Profiling and rule discovery
- Applying data standards (e.g., mailing address standards)
- Automating data quality audits
- Procedural and rule-based data cleansing

### TECHNOLOGY SOURCES

DQ technology exists in many forms from a variety of sources. The most common sources include:

- Commercial off-the-shelf (COTS) tools specifically designed to perform DQ functions.
- Open Source data quality tools.
- DQ Service providers with proprietary tools and technology.
- Data warehousing and BI tools with embedded DQ functions.
- Internally developed data quality tools and processes.

### TECHNOLOGY FUNCTIONS

Common DQ technology functions include:

- Data profiling
- Address verification, geocoding, and phone/address verification
- Matching, grouping, and de-duplication
- Data transformation and cleansing
- Metadata management

Common needs where technology is lacking include:

- DQ measurement and monitoring
- Data quality scorecards
- Rule-based data audit and cleansing

### CHOOSING DQ TECHNOLOGY

The matrix on the facing page maps technology sources and functions as a guide and recommendations for finding and choosing DQ technology.





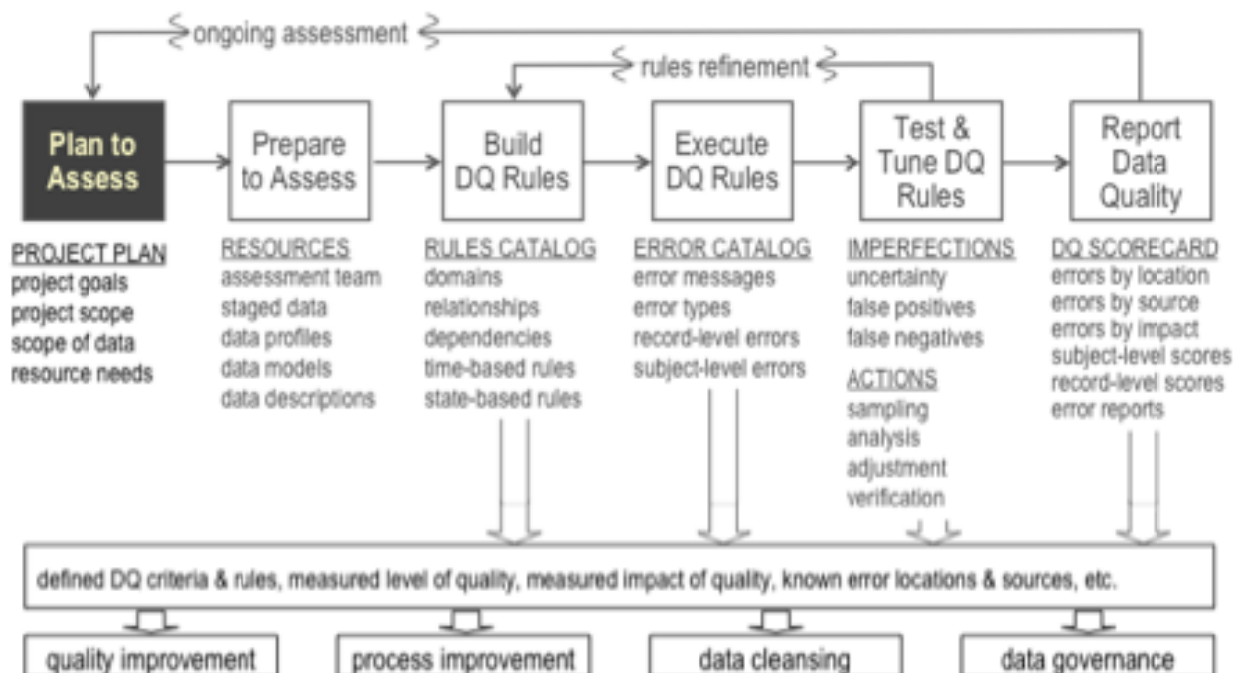
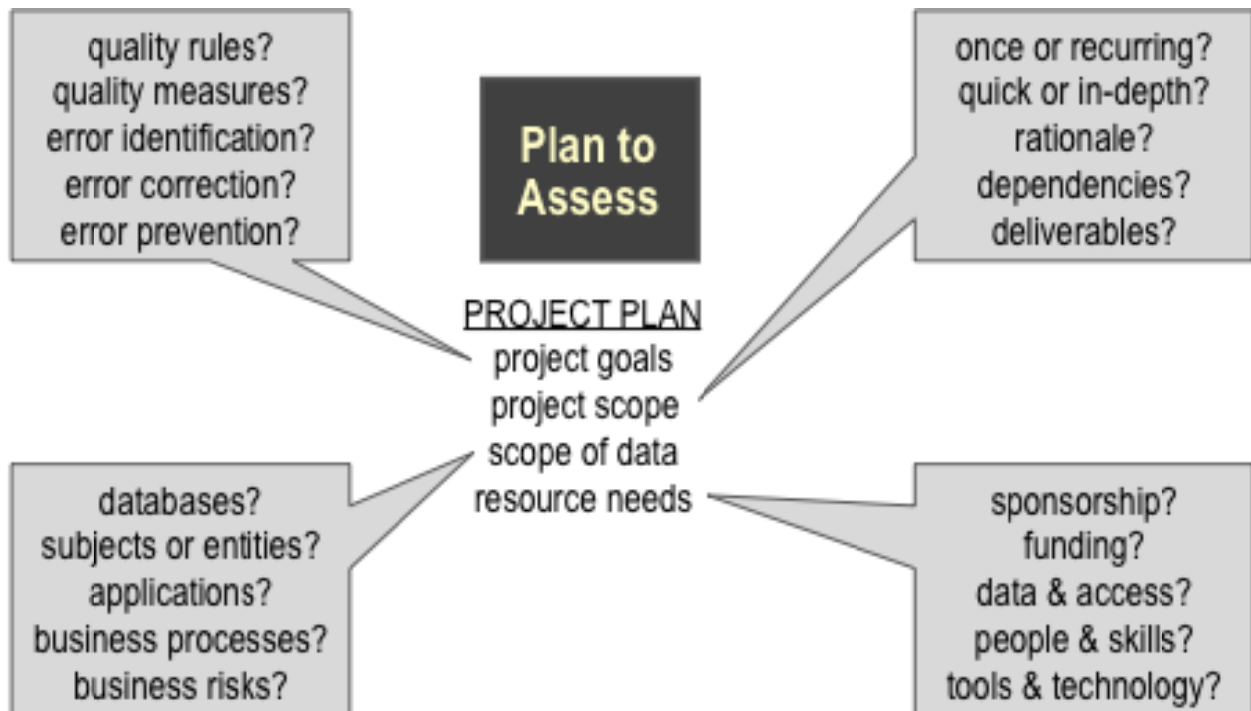
# Module 3

## Data Quality Assessment

Topic	Page
Planning and Preparation	3-2
Conducting the Assessment	3-8
Assessment Results	3-14
Applied Results	3-16

# Planning and Preparation

## Project Planning



# Planning and Preparation

---

## Project Planning

### THE DISCIPLINE OF PROJECTS

Data quality assessment is a project – or more correctly, each iteration of data quality assessment is a project. As with all projects, getting the right results begins with project planning. An assessment project plan needs to address four areas: goals, project scope, data scope, and resource needs.

### GOALS

Typical assessment projects include goals of rule identification, error identification, and quality measurement. It is practical, however, to conduct an assessment project that does more, or does less, than the norm. A project, for example, may need to go beyond quality measurement to include error correction or prevention. And if the additional goals are value expectations, they must be identified. Conversely, there may be value in doing less – a project, for example, simply to identify and define DQ rules.

### PROJECT SCOPE

It is important to know the scope and constraints of an assessment project. Why is the assessment needed? Is it a one-time or recurring assessment? What are the trade-offs between speed and thoroughness? Do other projects depend on the assessment, or the assessment on other projects? What are the must-have deliverables?

### SCOPE OF DATA

Data scope describes which data is to be included in the assessment? Too broad a scope is impractical, and too narrow produces no value. Jill Dyché defines five levels of DQ delivery<sup>1</sup>:

- Project – DQ efforts linked with a specific project.
- Application – DQ connected with implementing a new application or replacing an existing application.
- Subject Area – DQ for a specific problematic set of data.
- Process – DQ efforts targeting a particular business process.
- System – DQ for an individual system that is known to have bad data.

### RESOURCE NEEDS

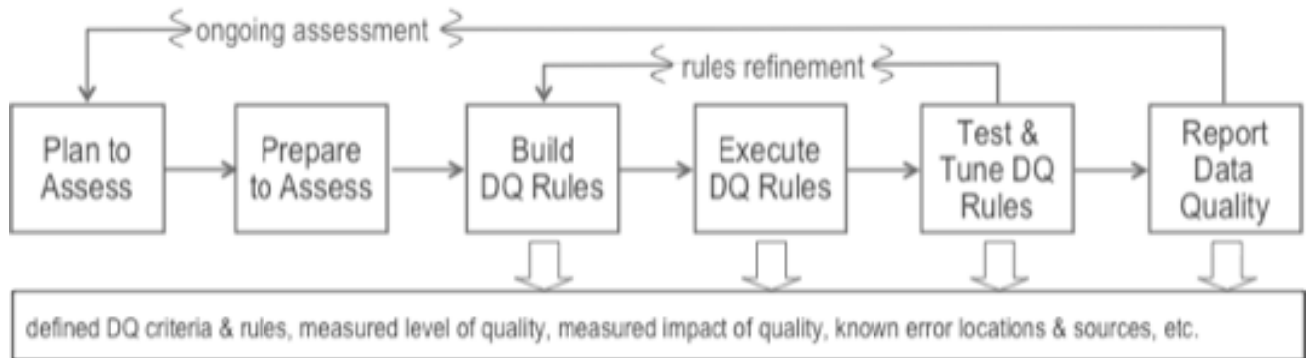
Project planning isn't complete without identifying the resources needed to do the work and produce the deliverables. All of the common project resources – sponsorship, funding, staffing, and technology – are needed. DQ assessment project planning should give special attention to needs for data and access to data.

---

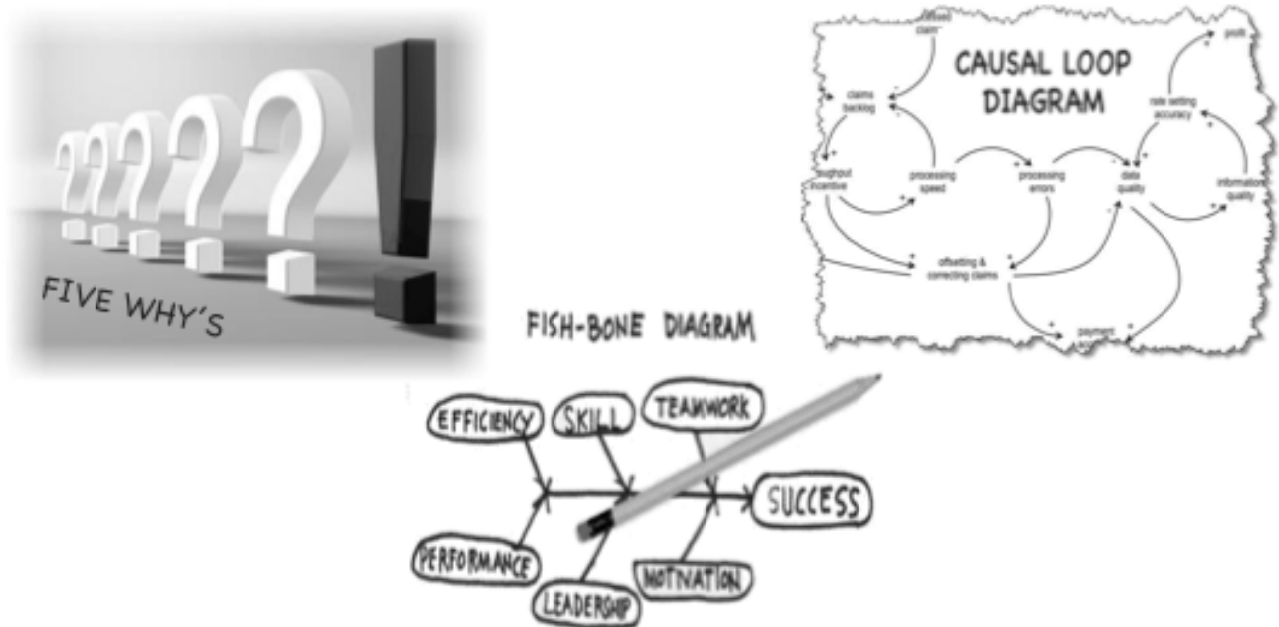
<sup>1</sup> *The Importance of Scope in Data Quality Efforts*, Dyché, Information Management Blogs, June 29, 2010

# Applied Results

## Root Cause Analysis



Knowing **WHAT** defects occur ...



Knowing **WHY** defects occur ...

Knowing **HOW** to correct & prevent defects ...

# Applied Results

---

## Root Cause Analysis

### **FROM WHAT TO WHY**

Data quality assessment identifies defects and quantifies data quality. It describes *what* is the state of data quality. With scoring, assessment describes *how much* data quality, and with dimensionality it tells *where* defects exist. All of this – *what, how much, and where* – is helpful for QC/QA activities of replacement and repair. But it is difficult to address the full scope of quality management – to prevent defects – without knowing *why* defects occur.

Root cause analysis (RCA) is an approach to problem solving that focuses on finding fundamental causes of problems to avoid the ineffective and incomplete solutions that result from addressing symptoms. A root cause is the initial influence in a cause-and-effect chain. RCA is the means to know why defects occur.

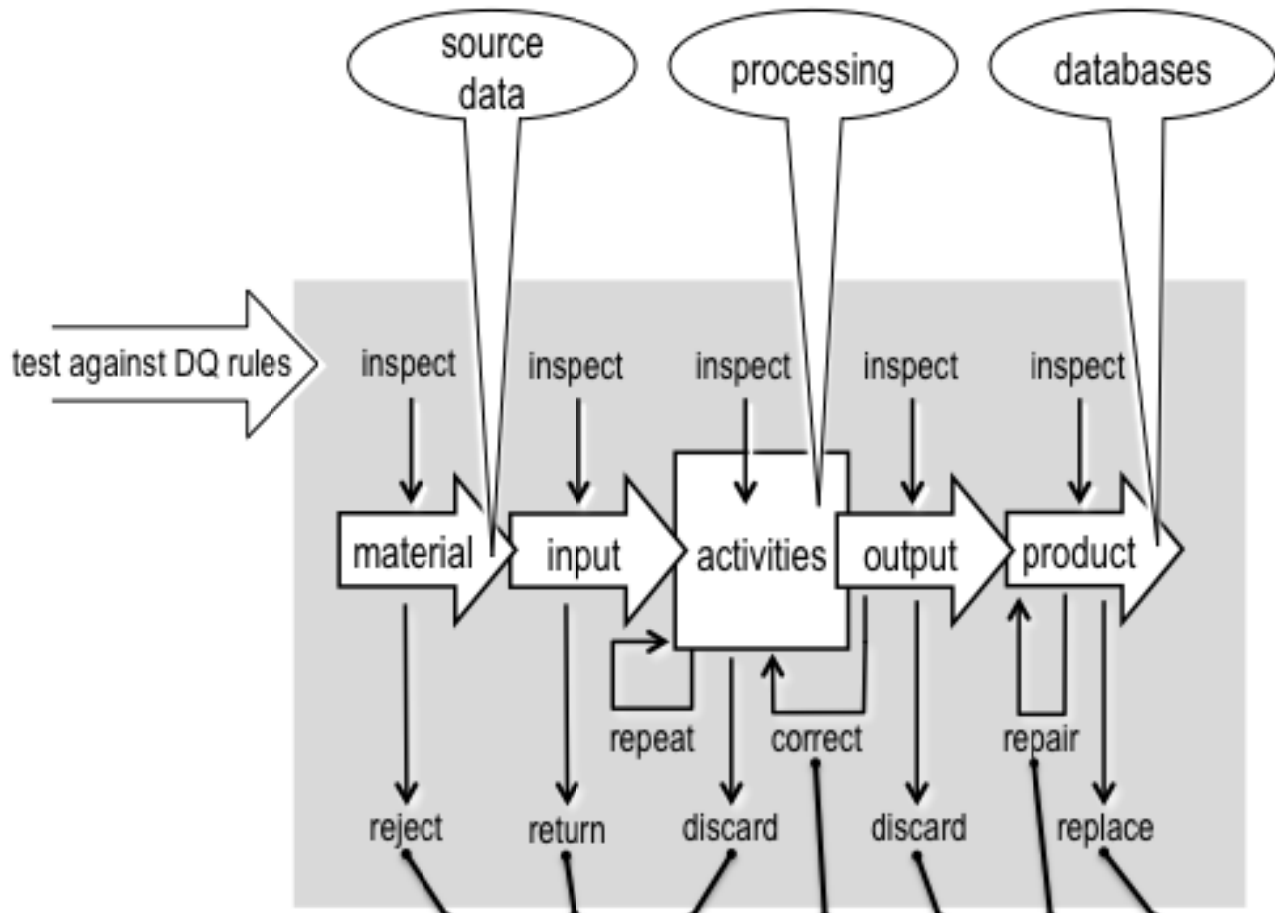
Three common techniques of RCA are:

- Five Why's – A fast and easy technique for simple problems.
- Fishbone Diagramming – An intermediate level technique for more complex problems.
- Causal Loop Modeling – An advanced technique for the most vexing problems.

Each technique is described on the following pages.

# Applied Results

## Data Cleansing – How to Cleanse?



	reject	return	discard	correct	discard	repair	replace
refuse the data	✓	✓					
remove the data			✓		✓		
use an alternate source				✓		✓	✓
derive probable values				✓		✓	✓
use default values				✓		✓	✓
manual data change							✓

---

# Applied Results

---

## Data Cleansing – How to Cleanse?

### REMOVING DEFECTS

The choices for removing DQ defects are really very few. There is no magic process that you can apply to turn bad data into good data. The goal must be to turn bad data into useful data that is suited to the purposes for which it will be used. Automated data cleansing choices include:

- Refuse to accept incoming data when defects are detected.
- Remove defective data (null out values or delete rows) when defects are detected.
- Use an alternate data source.
- Derive probable values based on surrounding data.
- Use default values.

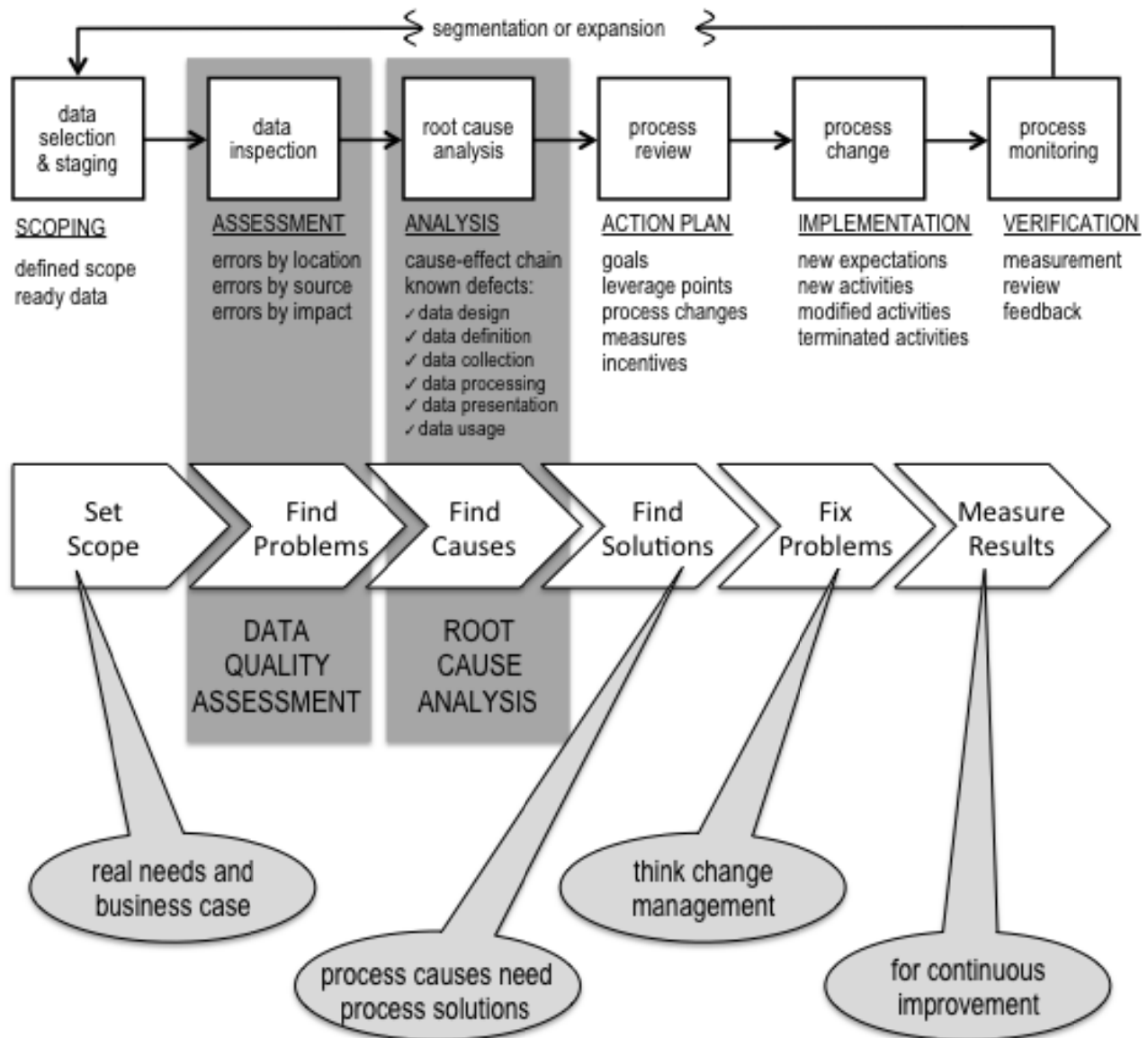
The remaining choice is manual data change, which is sometimes accomplished by placing defective data in a “suspense” file pending manual review and correction.

### DATA CLEANSING AND QA/QC

The diagram on the facing page maps data cleansing choices to quality assurance and quality control actions.

# Applied Results

## Process Improvement



---

# Applied Results

---

## Process Improvement

### **ANALYSIS, CHANGE, AND FEEDBACK**

Process improvement is the work of preventing future defects by finding and fixing the root causes of those defects. Process improvement includes two previously described processes – data quality assessment and root cause analysis. These predecessors to process review help to determine which processes are subjects of review.

Process review establishes goals and identifies the specific process changes to be made. Implementation puts the changes into effect. Every change should be linked to goals and driven by root cause analysis.

- When root causes are linked to material deficiencies, processes are adjusted to prevent defective materials from getting into the production line. The adjustments may involve changing inspection methods and changing supplier relationships. In DQ management these adjustments apply to data sourcing, data acquisition, and data entry.
- When root causes are linked to quality of work, adjustments are made to the activities that are performed, to the resources that are used, and to the workforce performing the activities. In DQ management these adjustments apply to data creation and updating, data storage, data maintenance, and data integration.
- When root causes are linked to product delivery, adjustments may be made to delivery methods and to customer relationships. In DQ management these adjustments apply to data presentation and visualization, reporting, research, analysis, and consumer choices and expectations for data usage.

Finally, every process change needs to be monitored. Measurement and feedback are essential to learn if the change is effective and to determine how effective. Monitoring also helps with early discovery of unintended side effects of change.

Expect process improvement to be iterative. Process evolution through a series of changes over an extended time period is often more effective, and certainly less risky than radical process reengineering.





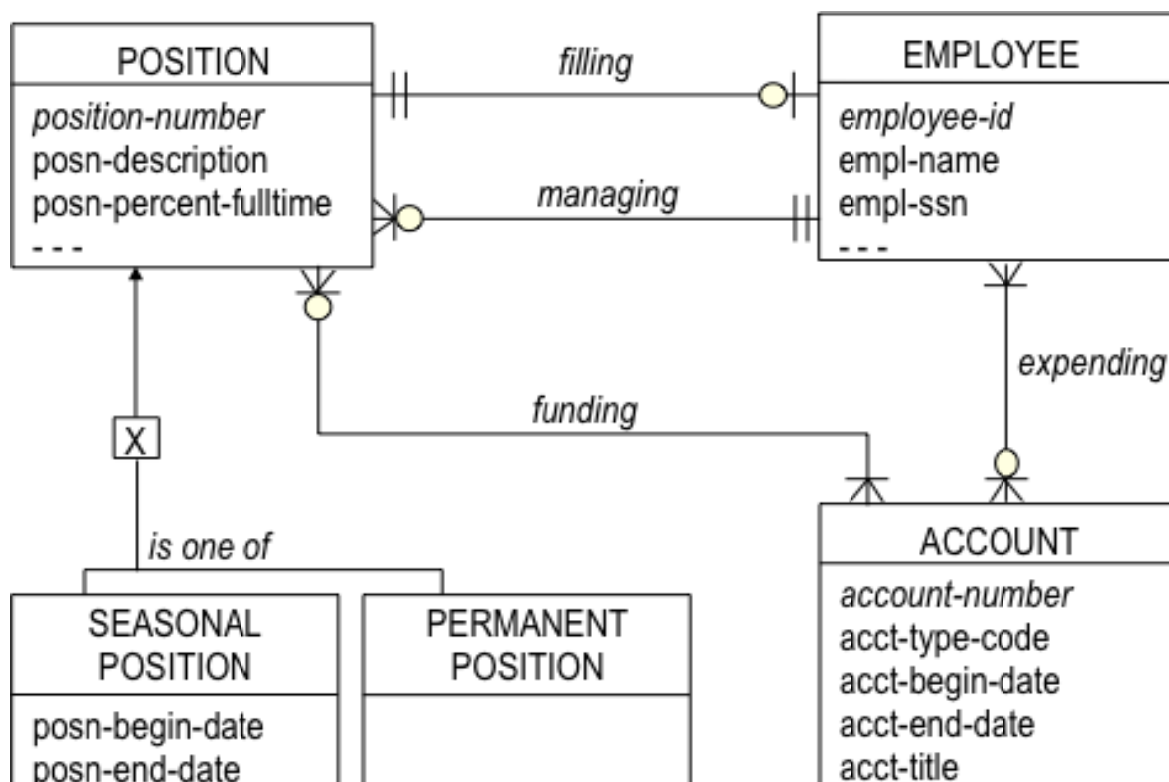
# Module 4

## Data Quality Improvement

Topic	Page
Procedural Data Quality	4-2
Rule-Based Data Quality	4-14
IT Processes and Data Quality	4-36
Business Processes and Data Quality	4-46

# Rule-Based Data Quality

## Data Models and Integrity Rules



### Explicit Rules in Data Models:

- Identity – *position-number* must be unique, *employee-id* must be unique, etc.
- Reference – employee must fill a position that exists, position may not be filled by non-existing employee, etc.
- Cardinal – employee must fill exactly one position, position must be funded by one or more accounts, etc.
- Inheritance – every position is either seasonal or permanent, no position is both seasonal and permanent, only seasonal positions have end-date, etc.

---

# Rule-Based Data Quality

---

## Data Models and Integrity Rules

---

### MODEL-BASED RULES

For data modelers, a model-based approach to integrity rules is effective. Each of the integrity rule types is based upon entity/relationship data modeling concepts. All data integrity rules are either expressed (explicit) in data models, or implied by those models. Explicit rules can be directly extracted from data models. Implicit rules are identified through analysis that uses data models as one source of information.

### EXPLICIT RULES

The explicit rule types are:

- Identity rules
- Reference rules
- Cardinal rules
- Inheritance rules

### IMPLICIT RULES

The implicit rule types are:

- Value set rules
- Relationship dependency rules
- Attribute dependency rules

### RULE IDENTIFICATION

Each of the rule types (and sub-types of dependency rules) can be identified through analysis using data models. The process of extracting integrity rules from data models is discussed in *Appendix A: Integrity Rules and Data Models*.

# Business Processes and Data Quality

## Defining Data



### Data Entities:

- clear, unambiguous business definitions
- consistent use throughout the business
- consistent use across computer systems
- consistent data naming standards

### Data Elements:

- clear, unambiguous business definitions
- finite and defined domain of allowed values
- defined meaning for every code value
- consistent use throughout the business
- consistent use across computer systems
- controlled and purposeful redundancy
- declared system of record
- documented origin, users, and usage
- defined privacy and security requirements
- consistent data naming standards

**Metadata:** The data and information that is needed by an organization to manage its data and information resources. Metadata serves four purposes: to classify, describe, guide, and control.

classify describe guide control

Group data by similar characteristics & manage by categories

Common standards, practices, processes, and technologies

Classify by several characteristics including

- subject (finance, workforce, product, customer, etc.)
- usage (transactions, analysis, prediction and forecasting, etc.)
- time (current, historical, predictive, etc.)
- content (structured, geographic, text, image, etc.)
- scope (global, local, departmental, master data, etc)

classify describe guide control

Define the data, itemize properties, and document

Understand the nature of the data both logically and physically

Describe data properties that include

- meaning (business definition and description)
- structure (objects, identifiers, groups, relationships, etc.)
- content (text, codes, numbers, currencies, dates, etc.)
- values (lists, patterns, ranges, character sets, etc.)
- lineage (source, derivation, calculation, processing, etc.)

classify describe guide control

Help people to find the data that they need,  
and to make decisions about utility and veracity of the data

Find, evaluate, and access data using

- keywords and search terms
- qualitative indicators (origin, verification, certification, etc.)
- synonyms and aliases
- domain taxonomies and ontologies
- entry points, access paths, and navigation paths

classify describe guide control

Know and document regulatory and service level requirements that are associated with the data

Describe data properties and constraints including

- data privacy regulations, policies, and procedures
- data security requirements and practices
- distribution, availability, and timeliness requirements
- retention and archiving requirements and practices
- recovery and business resumption level of need

---

# Business Processes and Data Quality

---

## Defining Data

### DEFINITIONS

Early in the course we discussed the need for data definitions, and described the good practices of data definition. As a quick review, some of the good practices are business-based meaning, and thorough, current, and time-independent definitions without size limits.<sup>1</sup>

### METADATA

Data definition is really a process of collecting metadata. Data names and definitions are among the most fundamental business metadata items. Metadata encompasses all of the data that is needed by an organization to effectively and efficiently manage its data and information resources. This includes data to classify information, to describe information, to guide users of information, and to control information.

### CLASSIFY

Classification provides the means to organize and manage data by category – to group together similar kinds of data to which we can apply common standards, practices, processes, and technologies. Data is typically classified by several categories or criteria which may include any or all of:

- subject – financial data, product data, customer data, etc.
- usage – transactional, analytical, regulatory, etc.
- time – current data, historical data, predictive data, etc.
- scope – global data, local data, master data, etc.
- sensitivity – classified data, restricted data, public data, etc.

### DESCRIBE

Description is the act of defining the data, itemizing properties of the data, and documenting the definitions and properties. The skills discussed earlier – architecture and definition – are metadata collection skills. Well described data includes descriptions of:

- data meaning – business definitions of data including examples and important distinctions between similar kinds of things .
- data structure – descriptions of data objects (entities, tables, files, records, etc.) their identifiers, logical groupings, and relationships
- data content – description of the kinds of data to be collected including text, codes, numbers, currencies, dates, etc.
- data values – description of constraints on the allowable values of data – lists, patterns, ranges, allowed and disallowed characters, etc.
- data lineage – data sources, derivations, calculations, cleansing, and other processes that are part of data propagation through a business.

---

<sup>1</sup> *Data Resource Quality: Turning Bad Habits into Good Practices*, pp 51-71, Brackett





# Module 5

---

## Summary and Conclusion

Topic	Page
Summary of Key Points	5- 2
References and Resources	5- 4

# Summary of Key Points

---

## A Quick Review

- To manage quality you must first define quality.
- Four common definitions are defect free, conforming to specification, suited to purpose, and meeting customer expectations.
- Data quality involves both content correctness and structural integrity.
- Dimensions of quality include accuracy, completeness, consistency, precision, granularity, timeliness, integrity, and usability.
- Common causes of DQ problems are poor definition, poor design, errors in data entry/collection processes, and errors in data manipulation processes.
- Proven practices for QA, QC, and QM can be directly applied to data quality management. We don't need to reinvent them.
- Data quality management involves people. Identification of stakeholders and designation of responsibilities is important.
- Profiling, assessment, cleansing, and process improvement are core data quality processes.
- Data quality assessment is a goal-driven, rule-based, and repeatable process to evaluate, measure, and report the state of data quality.
- Assessment only finds and reports defects. Changing data quality involves data cleansing, root cause analysis, and process improvement.
- Procedural data quality is an algorithmic approach to cleansing or standardization of data where predictable and consistent patterns exist.
- Rule-based data quality applies inspection and logic to data where the patterns and criteria are not predictable, but are specialized to the enterprise or to the database.
- Data models are a good source of DQ rules – both explicit and implicit.
- Building in quality is a superior approach to fixing defects later. IT processes offer many opportunities to build in quality.
- Business processes should have designated responsibility for data quality in data definition, data creation, data updates, and data use.

# Summary of Key Points

---

## A Quick Review

**SUMMARY**

The facing page summarizes many of the key points from this course. It can be a useful quick reference as you plan your approach to data quality management.



# Appendix A

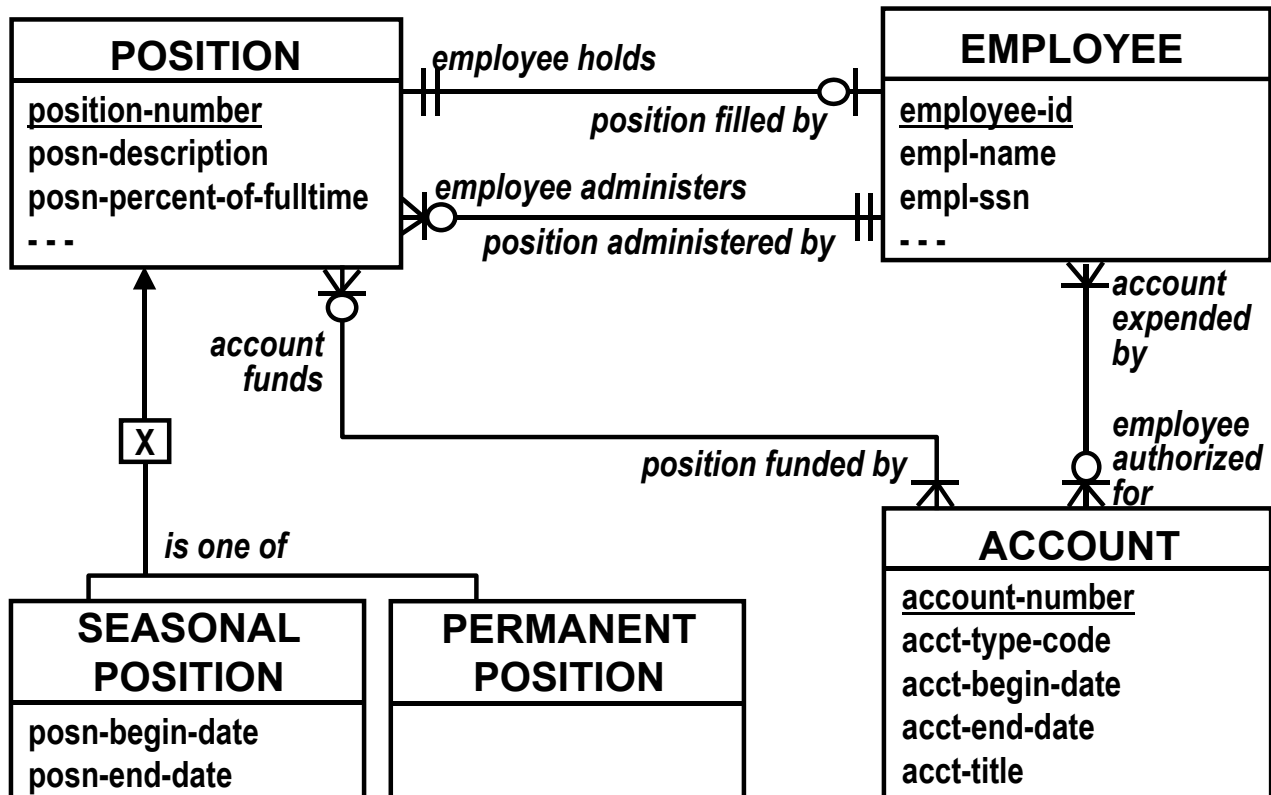
---

## Integrity Rules and Data Models

Topic	Page
Model-Based Integrity Rules	A-2
Identity Rules	A-4
Reference Rules	A-8
Cardinal Rules	A-12
Value Set Rules	A-16
Inheritance Rules	A-20
Relationship Dependency Rules	A-24
Attribute Dependency Rules	A-32
Data Integrity Rules Summary	A-44

# Model-Based Integrity Rules

## A Data Model Example



---

# Model-Based Integrity Rules

---

## A Data Model Example

### A BASIS FOR ALL INTEGRITY RULE EXAMPLES

The sample data model shown on the facing page is used to illustrate each of the rule types by example. This model is intentionally *not* in the third normal form, as it is meant to represent a range of real possibilities that may occur in business models, application development projects, and legacy data examination.

This small data model illustrates the following business rules:

- The entity types of interest are *POSITION*, *EMPLOYEE*, and *ACCOUNT*.
- Every *POSITION* is either a *SEASONAL POSITION* or a *PERMANENT POSITION*.
- One *POSITION* is filled by zero or one *EMPLOYEE*.
- One *EMPLOYEE* holds exactly one *POSITION*.
- One *POSITION* is administered by exactly one *EMPLOYEE*.
- One *EMPLOYEE* administers zero, one, or many *POSITIONS*.
- One *POSITION* is funded by one or many *ACCOUNTS*.
- One *ACCOUNT* funds zero, one, or many *POSITIONS*.
- One *ACCOUNT* is expended by one or many *EMPLOYEEs*.
- One *EMPLOYEE* is authorized for zero, one, or many *ACCOUNTS*.
- *EMPLOYEEs* are uniquely identified by *Employee-Id*.
- *POSITIONS* are uniquely identified by *Position-Number*.

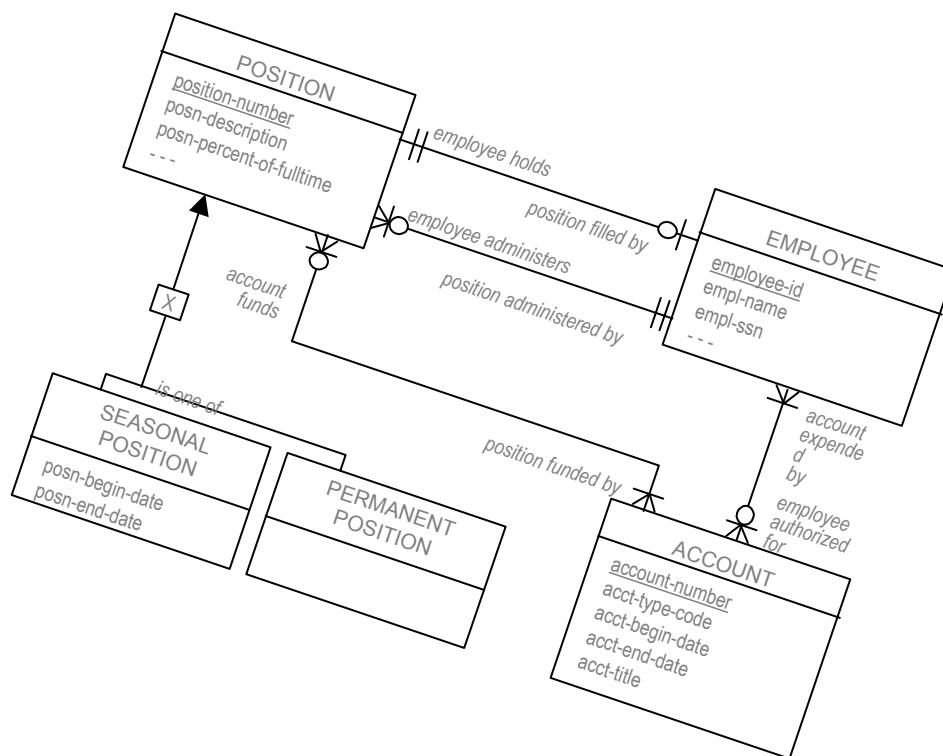
### BUSINESS RULES AND INTEGRITY RULES

Note that the model above is described as illustrating **business rules**. This important concept helps to align data integrity with the concept of quality through “meeting customer expectations.” Business people are the customers who use the data warehouse. Data quality rules with a business basis are most likely to meet business expectations.

# Data Integrity Rules Summary

## The Number of Possible Rules

**This small data model  
yields at least 32 distinct  
data integrity rules**



# Data Integrity Rules Summary

---

## The Number of Possible Rules

### RULES FROM THE EXAMPLE

This small data model yields a surprisingly large number of rules. The model consists of three entities, four relationships, and one specialization into two subtypes. Yet it produces more than 32 integrity rules.

The explicit rules are:

- 3 identity rules – 1 for each non-subtype entity
- 6 reference rules – 1 for each one-to-many relationship, and 2 for each many-to-many relationship
- 4 cardinal rules – 1 per cardinality (except zero-many)
- 6 inheritance rules – 2 per super-type and 2 per sub-type

Implicit rules will include at least:

- 13 value set rules – a minimum of one per attribute.
- more rules are possible from relationship dependencies and from attribute dependencies.

### HOW MANY RULES?

Consider the size of the data models – source and target – in specific data warehousing environments and begin to think about how many integrity rules are possible. The number of integrity rules inherent in a data model can be roughly estimated using this formula:

$$\begin{aligned} \text{Number of rules} = & (\text{number of entity types}) \\ & + (\text{number of relationships} * 3) \\ & + (\text{number of attributes}) \\ & + (\text{number of sub-types} * 2) \\ & + (\text{number of super-types} * 2) \end{aligned}$$

The formula does not account for dependency rules, typically a relatively small part of the total rule set.

### WHAT IT MEANS FOR SOURCE DATA

Not surprisingly, source data often lacks integrity. How many edits are possible in the source environment when integrity rules are factored by the circumstances of adding, changing, and deleting data? In how many programs should those edits be executed? How much data integrity is missing from source systems?

### WHAT IT MEANS FOR THE DATA WAREHOUSE

Although add, change, and delete aren't considerations, the warehousing environment has similar data integrity risks. How many integrity rules are possible in your warehousing environment? How many are identified? How many are applied?





---

# Appendix C

## Exercises

---

<b>Exercise</b>	<b>Page</b>
Exercise 1: Defining Quality	C-2
Exercise 2: Data Quality Perceptions	C-3
Exercise 3: Data Quality Defects	C-4
Exercise 4: Understanding Data Quality Rules	C-6
Exercise 5: Finding Data Integrity Rules	C-8

# Exercise 5: Finding Data Integrity Rules

## Integrity Rules in Data Models

### INSTRUCTIONS

Pages C-10 and C-11 show a data model related to project management. Study that data model to answer the questions below and on the facing page. Refer to *Appendix A: Integrity Rules and Data Models* for rule descriptions, examples, and guidelines for rule identification.

1. How many identity rules are expressed in this data model?

---

---

2. What are some examples of identity rules?

---

---

---

---

---

---

---

---

---

---

3. What are the identity rules for the entity named Activity?

---

---

---

---

---

---

---

---

---

---

---

## Exercise 5: Finding Data Integrity Rules

---

### Integrity Rules in Data Models

4. What are some examples of reference rules?

---

---

---

---

---

---

---

5. What are some examples of cardinal rules?

---

---

---

---

---

---

---

6. Can you find examples of state-dependent attributes? If yes, list a few below.

---

---

---

---

7. How many explicit data integrity rules are possible in this model? (Hint – see page A-45.)

---

---