Applied A.I. Solutions

Foundations of Data Management

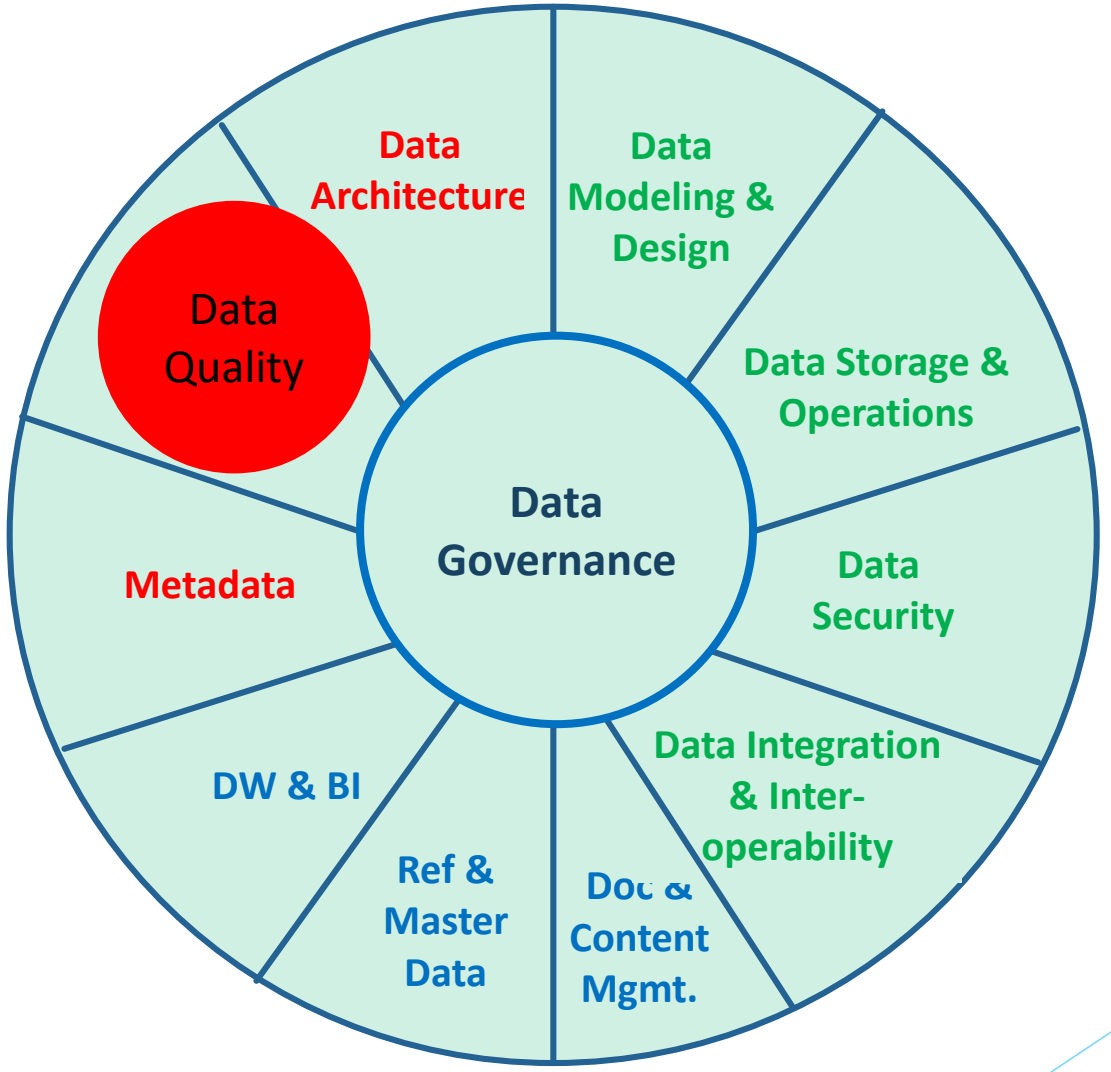
Professor

Daniel Vitaver-Bronstein, B.Sc., EMBA

daniel.vitaver-bronstein@georgebrown.ca

DATA QUALITY

The DAMA Wheel



[1] Main source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

1. INTRODUCTION

- DM involves a set of complex, interrelated processes that enable an organization to use its data to achieve strategic goals

- DQ within the DMF includes managing data through its entire lifecycle

- DQ programs have data-related operational responsibilities and accountabilities

[1] Main source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Definition

The planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meets the needs of data consumers

[1] Main source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Goals

- Develop a governed approach to meet consumers' DQ requirements
- Define standards, requirements, specifications, processes, metrics for DQ Monitoring and DQ level control
- Identify, advocate for opportunities to improve data quality

Business Drivers

[1] Main source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

DQ programs are guided by the following principles:

- Criticality
- Lifecycle Management
- Prevention
- Root cause remediation
- Governance
- Standards-driven
- Objective measurement and transparency
- Embedded in business processes
- Connected to service levels

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Inputs

- Data Policies and Standards
- DQ expectations, Business requirements and rules
- Data requirements
- Metadata (business, technical)
- Data sources and Data stores
- Data lineage

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Activities

1. Define high-quality data
2. Define a data-quality strategy
3. Define scope of initial assessment
4. Perform initial DQ assessment
5. Identify and prioritize improvements
6. Develop and deploy DQ operations

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Deliverables

1. DQ Strategy and Framework
2. DQ Program organization
3. Data profiling analyzes
4. Root-cause analysis of issues and recommendations
5. DQM procedures and DQ reports
6. DQ Governance reports
7. DQ Policies, Guidelines and SLA

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Suppliers	Participants	Consumers
• Business Management • SME • Data Architects • Data Modelers • System specialist • Data Stewards • BP Analysts	• CDO • DQ Mangers, DQ Analysts • Data Stewards, Data Analysts, Data Owners, DBA, Data Team • Data Integration Managers • IT Operations • Compliance Team	• Business Data Consumers • Data Stewards • Data and IT Professionals • Knowledge Workers • DG Bodies • Partner Organizations • Centres of Excellence

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Technical Drivers

Techniques

- Spot-checking using multiple subsets
- Tags and notes to mark data issues
- Root-cause analysis
- Statistical process control

Tools

- Profiling engines, query tools
- DQ rule templates
- Quality check and Audit Code modules

Metrics

- Governance and conformance metrics
- DQ measurement results
- Improvement trends
- Issue management metrics

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Drivers

- Increase the value of organizational data
- Reduce risks and costs associated with poor quality data
- Improve organizational efficiency and productivity
- Protect and enhance the organization's reputation

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

1. CONCEPTS

- Data Quality
- Critical Data
- Data Quality Dimensions
  - Intrinsic
  - Contextual
  - Representational
  - Accessibility

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Common Dimensions of DQ

- Accuracy
- Completeness
- Consistency
- Integrity
- Reasonability
- Timeliness
- Uniqueness
- Validity

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Data Quality ISO 8000 Standard

"The ability to create, collect, store, maintain, transfer, process and present data to support business process in a timely and cost-effective manner requires both the understanding of the characteristics of the data that determine its quality, and the ability to measure, mange and report on data quality"

Note: refer to the original complete ISO 8000 standard for more details (ISO - International Organization for Standardization) https://www.iso.org/home.html

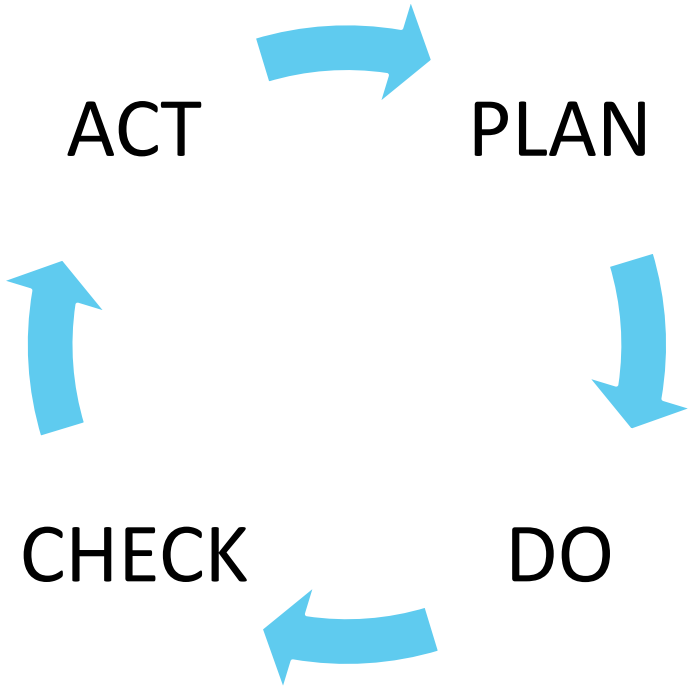
[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

- Goal: to help define what is and is not quality data, enable to ask for quality data using standards conventions, and verify the reception of quality data using those same standards

- DQ Management include
  - DQ Planning
  - DQ Control
  - DQ Assurance
  - DQ Improvement

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Data Quality Improvement Lifecycle (Shewhart / Deming cycle)



[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Common simple DQ Business Rules:

- Definitional conformance
- Value presence and Record completeness
- Format compliance
- Value domain membership
- Range conformance
- Mapping conformance
- Consistency rules
- Accuracy verification
- Uniqueness verification
- Timeliness validation

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Data Quality and Data Profiling

- Data inspection and data quality assessment
- Statistical techniques use to identify patterns, rules, constraint validity
- Identify possible duplicates, missing keys, faulty referential integrity
- Analyze conformity to rules and requirements
- Help to isolate the root causes of issues

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Data Quality and Data Processing

- Data Cleansing
- Data Enhancement
- Data parsing and Formatting
- Data Transformation and Standardization

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

2. ACTIVITIES

a) Define High Quality Data

b) Define Data Quality Strategy

c) Identify Critical Data and Business Rules

d) Perform an Initial Data Quality Assessment

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

e) Identify and Prioritize Potential Improvements

f) Define Goals for Data Quality Improvement

g) Develop and Deploy Data Quality Operations

h) Manage Data Quality Rules

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

i) Measure and Monitor Data Quality at

- Level of detail related to the execution of individual rules
- Level of aggregated data from the rules

$$\text{Valid Data (r)} = \frac{\text{Test Executions (r)} - \text{Exceptions Found (r)}}{\text{Test Executions (r)}}$$

$$\text{Invalid Data I (r)} = \frac{\text{Exceptions Found (r)}}{\text{Test Executions (r)}}$$

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

Dimensions & Business Rule	Metrics	Status Indicator
Completeness	# recs populated/ total # recs * 100	Unacceptable <80% populated
Uniqueness	Duplicate # recs/ total # recs * 100	Unacceptable > 0%
Timeliness	Incomplete trans/ total # trans in a time period * 100	Unacceptable <99% completed on time
Validity	# recs condition met/total # recs * 100	

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

j) Develop Operational Procedures for Managing Data Issues

- Diagnosing issues
- Formulating options for remediation
- Resolving issues
- Implementing incident tracking systems

k) Develop Data Quality Reporting

l) Establish Data Quality Service Level Agreement

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

3. TOOLS

- Data Profiling
- Data Querying Tools
- Modeling and ETL Tools
- DQ Rule Templates
- Metadata Repositories

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

4. TECHNIQUES

- **Preventive Actions**
  - Data entry controls
  - Training
  - Rules
  - DQ SLA
  - DG and Stewardship
  - Formal change control
- **Corrective Actions**
  - Automated correction including rule-based standardization, normalization and correction
  - Manually-directed correction
  - Manual correction

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

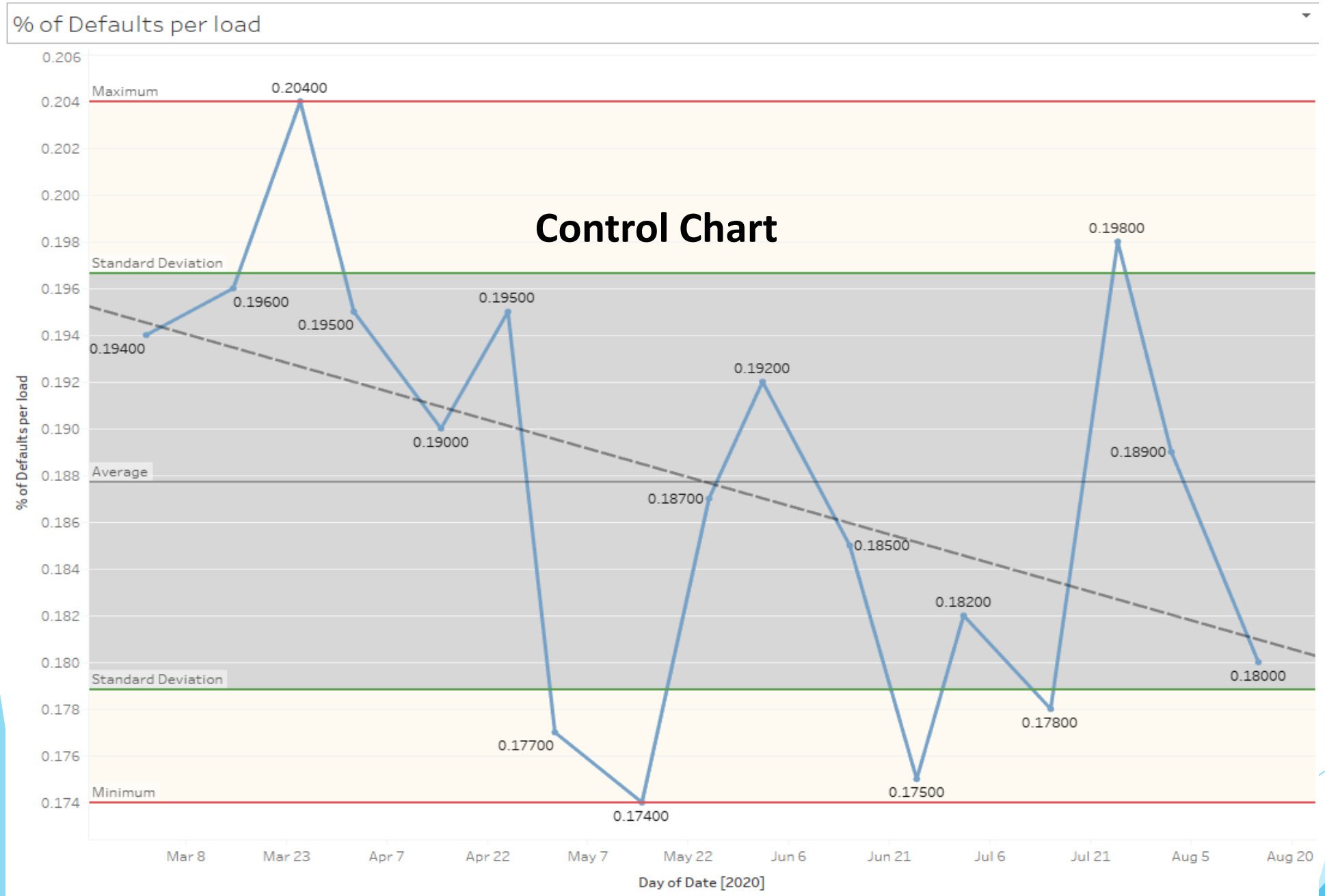
- **Effective Data Quality Metrics**
  - Measurability
  - Business Relevance
  - Acceptability
  - Accountability / Stewardship
  - Controllability
  - Trending

- **Quality Check and Audit Code Modules**
  - Shareable, re-usable code modules that execute repeated DQ checks and audit processes

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

- **Statistical Process Control**
  - Uses measures of central tendency (mean, median, or mode) and of variability around a central value (range, variance, standard deviation) to establish tolerances for variation within a process
  - SPC measures the predictability of process outcomes by identifying variation within a process
  - SPC is used for control, detection, and improvement

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA


% of Defaults per load
Control Chart
Maximum
Standard Deviation
Average
Standard Deviation
Minimum
0.20400
0.19600
0.19500
0.19400
0.19000
0.19500
0.17700
0.17400
0.18700
0.19200
0.18500
0.17500
0.18200
0.17800
0.19800
0.18900
0.18000
0.206
0.204
0.202
0.200
0.198
0.196
0.194
0.192
0.190
0.188
0.186
0.184
0.182
0.180
0.178
0.176
0.174
% of Defaults per load
Mar 8
Mar 23
Apr 7
Apr 22
May 7
May 22
Jun 6
Jun 21
Jul 6
Jul 21
Aug 5
Aug 20
Day of Date [2020]

- **Root Cause Analysis**
  - It is a process of understanding factors that contribute to problems and the ways they impact the business processes
  - Common techniques:
    - Pareto analysis (the 80/20 rule)
    - Fishbone diagram analysis
    - Track and trace
    - Process analysis
    - The Five Whys (McGilvray, 2008)

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

5. IMPLEMENTATION GUIDELINES

- Metrics on the value of data and the cost of poor data quality
- Operating model for IT/Business interactions
- Changes in how projects are executed
- Changes to business processes
- Funding for remediation and improvement projects
- Funding for data quality operations

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

- Readiness Assessment / Risk Assessment
  - Management commitment
  - Organization's understanding of the quality of data
  - Actual state of data
  - Risk associated with data lifecycle
  - Cultural and technical readiness
- Organizational Change
  - Enhance corporate communications

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA

6. DATA QUALITY AND DATA GOVERNANCE

- DQ is more effective when it is part of a DG program
- DQ contributes and supports DG policies

[1] Source: Copyright © 2017 DAMA International – DMBOK2 - Technics Publications, Basking Ridge, New Jersey, USA