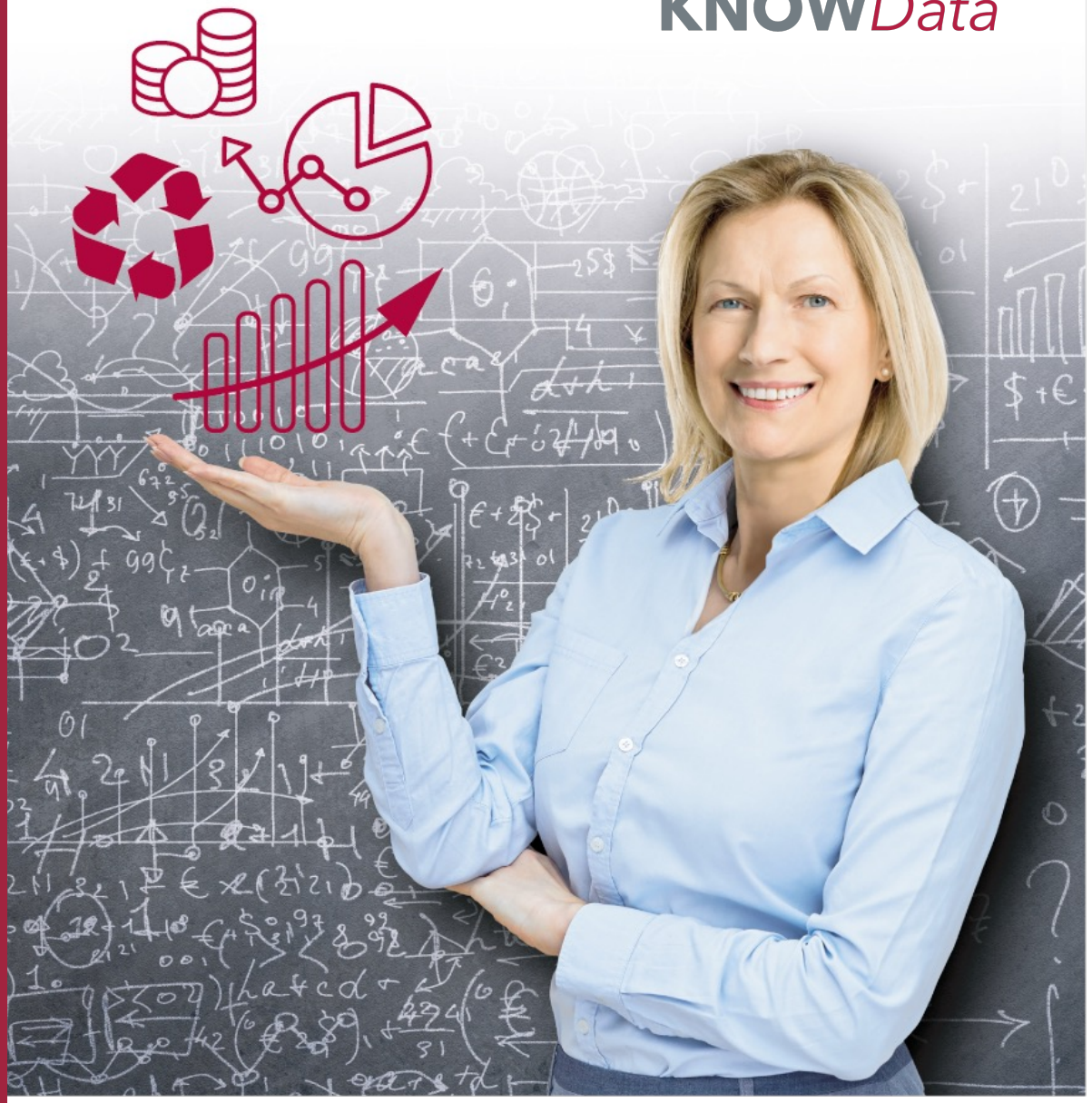
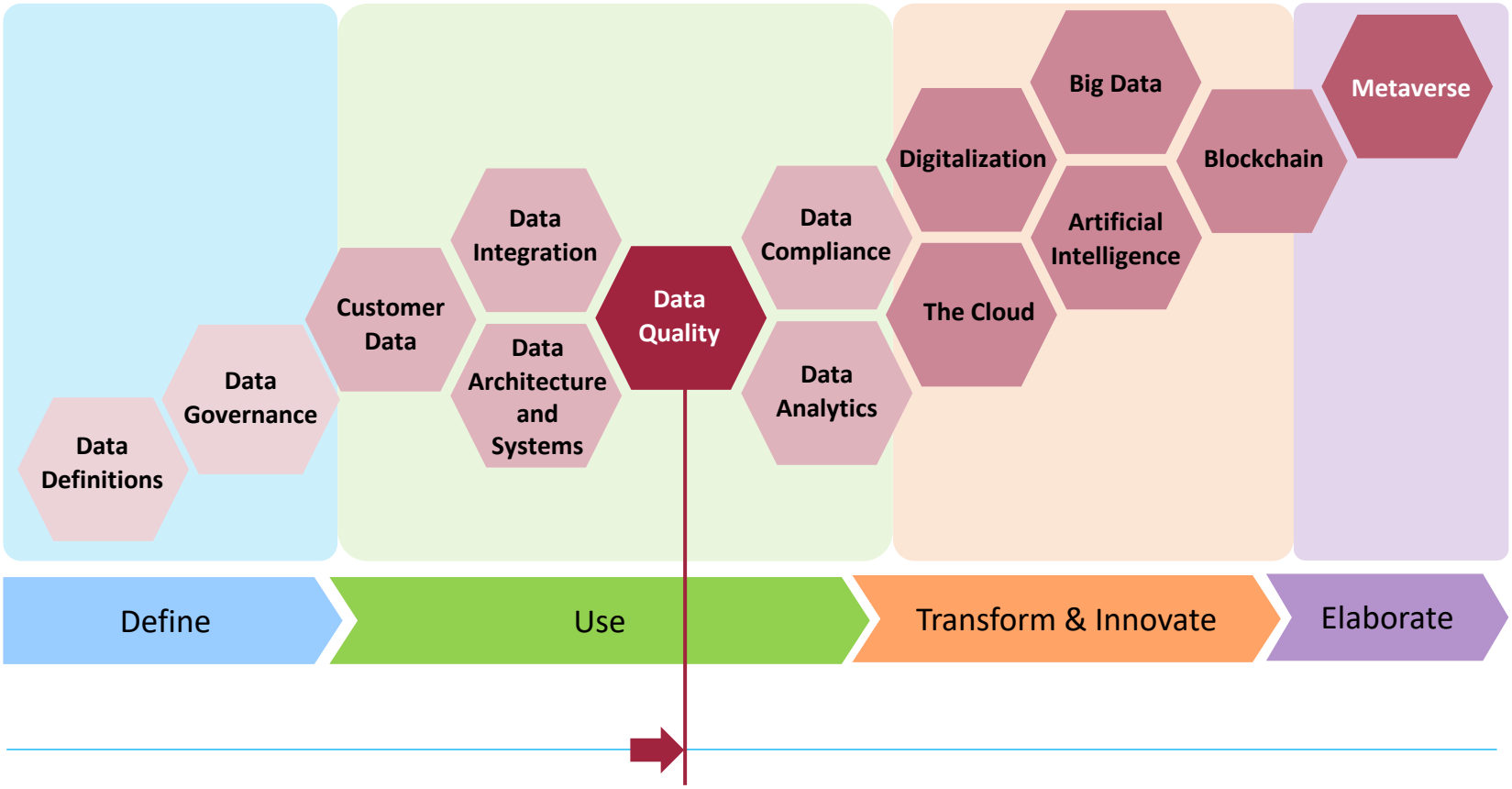
Data Quality

KNOWData

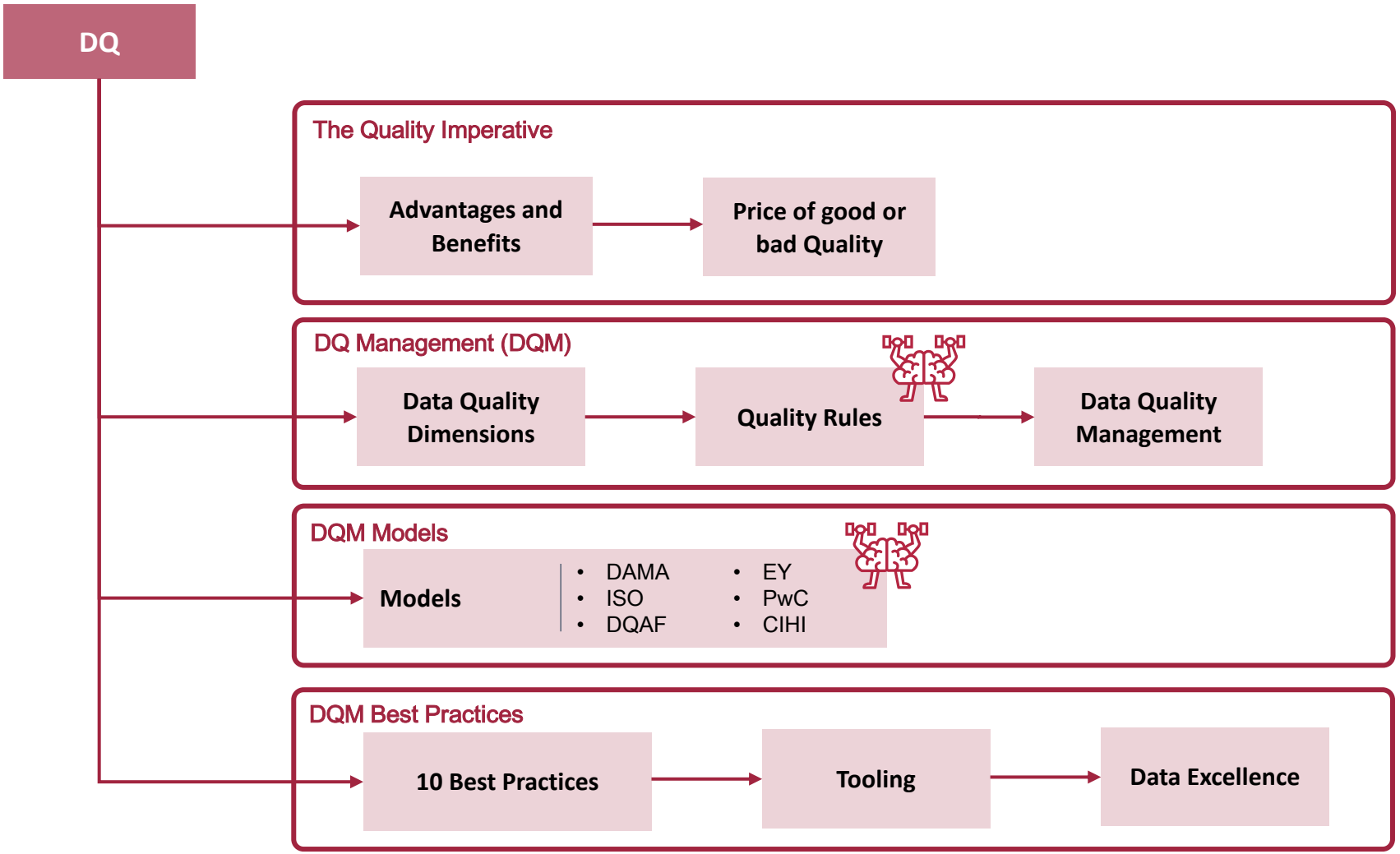


June 15, 2026

Lionel Pilorget


Data Definitions
Data Governance
Customer Data
Data Integration
Data Architecture and Systems
Data Quality
Data Compliance
Data Analytics
Digitalization
The Cloud
Artificial Intelligence
Big Data
Blockchain
Metaverse
Define
Use
Transform & Innovate
Elaborate

Structure of the presentation



Uber's $45M Driver Payment Miscalculation

In 2017, reports emerged that Uber had been miscalculating its commission and costing New York drivers a percentage of their rightful earnings. Instead of calculating its commission based on its net fare, minus sales tax and other fees, Uber took their cut based on the gross fare. This meant that for two and a half years, the company took 2.6% more from drivers than its own terms and conditions allowed.

Uber leaders reported they were paying back those earnings, plus 9% annual interest, to every driver impacted — an average payment of around $900 each. Given that "tens of thousands" of drivers were underpaid, the Wall Street Journal estimated the cost would be **at least $45 million**.

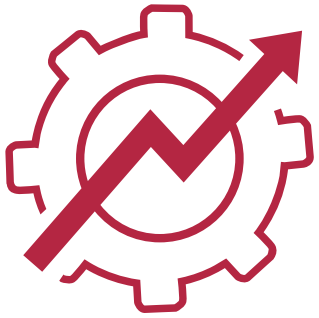
Like Equifax, Uber was already dealing with dissatisfied drivers at the time this incident became public. In 2017 alone, they settled for $20 million with the FTC for overinflating estimated driver earnings, and their then-CEO was caught on a dashcam video in an unflattering argument with his own Uber driver about decreasing earnings.

In this case, Uber's data wasn't inaccurate in and of itself, but the company was basing its calculations on the wrong numbers altogether. It's crucial for all teams to maintain transparency and visibility into data to maintain fair business practices — and avoid costly, public embarrassments.

=> Drivers were underpaid because the system incorrectly processed payment data

Source: www.montecarlodata.com/blog-bad-data-quality-examples/

Which Advantages of Good Data?



Benefits for Uber (1/2)

Accurate Decision-Making	Better decisions about pricing, promotions, or driver allocation by accurately analysing ride demand, driver availability, and fare structures
Operational Efficiency	Reduce miscalculations, such as incorrect driver payments, if fare data, tax information, and time tracking are accurate from the start
Regulatory Compliance	Comply with minimum wage laws, avoid legal battles and compensation payouts
Improved Customer Experience	Optimize ride-matching algorithms with better data, ensuring faster pickups and more efficient routes, improving both rider and driver satisfaction
Cost Savings	Better data accuracy in fare and wage calculation would have prevented a costly mistake, causing millions in refunds
Enhanced Reputation and Trust	Avoid negative press and preserve its reputation with both drivers and riders

Benefits for Uber (2/2)

Better Forecasting and Planning	Use high-quality data to predict peak hours, driver demand, and geographic trends, allowing to allocate resources more efficiently and enhance revenue opportunities
Risk Mitigation	Early detection of issues or compliance concerns to avoid costly lawsuits and public relations crises
Innovation and Competitiveness	Innovations in dynamic pricing, ride-sharing features, and further delivery services
Scalability	To expand into new regions or to bring services (like UberEats), good data ensure that pricing, local regulations, and customer preferences are managed effectively

What Are Bad Data?

Inaccurate, incomplete, inconsistent, irrelevant, duplicate

CustomerId	Name	LastName	SocialNumber	Mobile Number
1001	Brayan	Smith	12000101	79781033
1002	Lukas	Saberi	27128402	(076)-7511224
1003	Tom	Cruise	330248101	(076)-7788224
1004	Anna	Hanse	330246478	0
1005	Sarah	Baden	450767214	
1006	Thomas	Cruise	330248101	(076)-7788224
1007	Barbara	Streisand	670456203	(076)-7498224
Product ABC	Card	-	1%	-

inconsistent

duplicate

incomplete

inaccurate

irrelevant

The Price of Bad and Good Data

The price of bad data is often paid in hidden and long-term costs, while good data delivers measurable and sustainable returns on investment

Increased Manual Work

Reputation Damage

Lost Opportunities

Direct Financial Costs

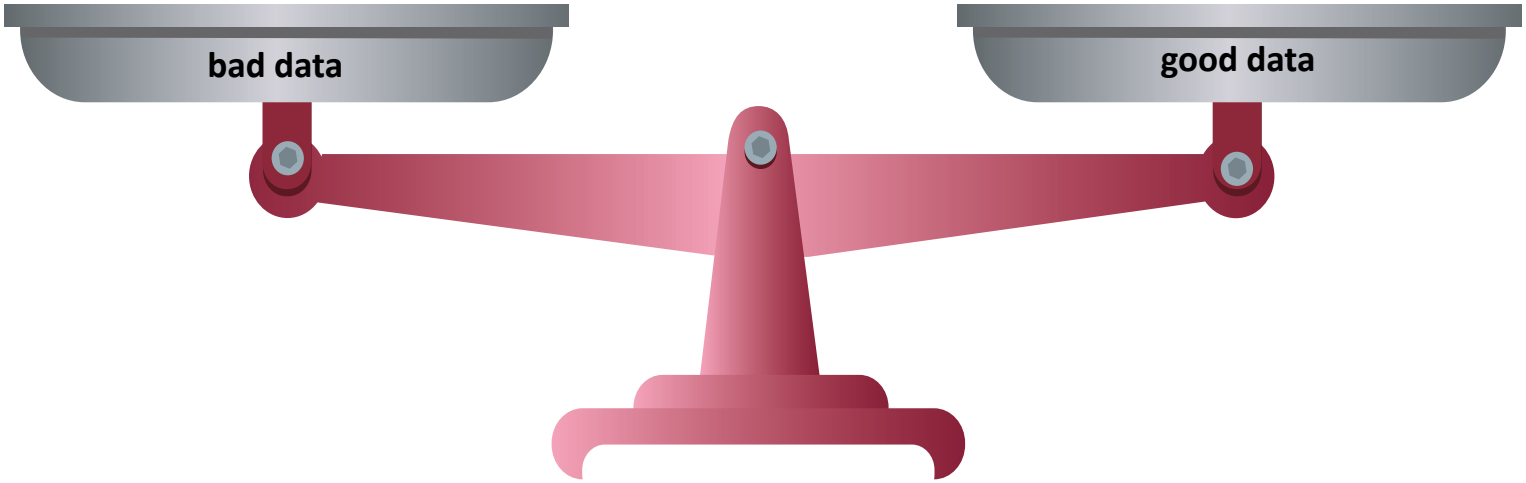
Customer Loyalty

Competitive Advantage

Compliance and Risk Mitigation

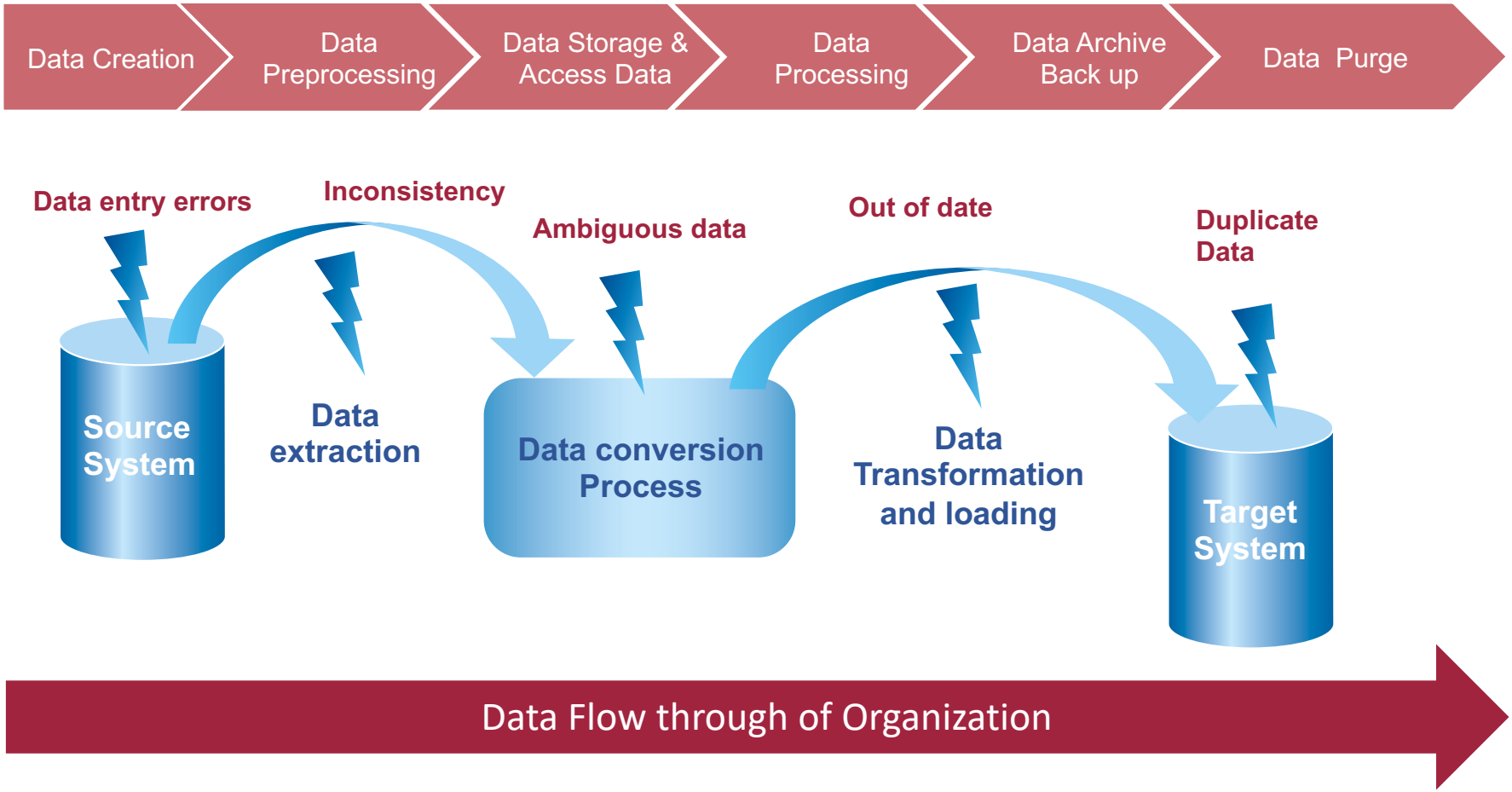
Operational Efficiency

Initial Investment



Data Quality Tracing

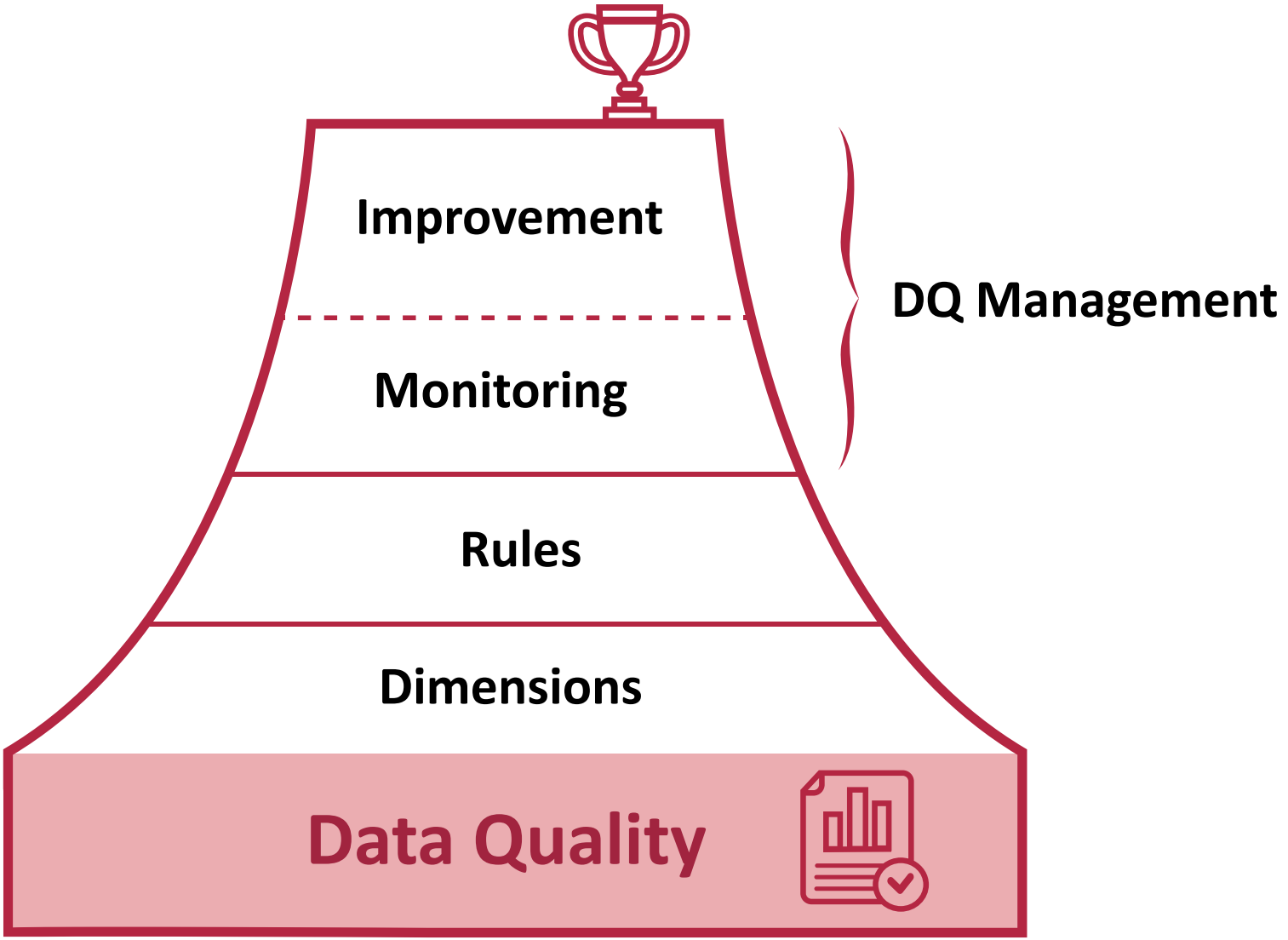
Bad data quality causes pain points in all phases of data cycle



Is it worth having good data?

How to get good data?

Dimensions > Rules > DQ Management



Different Types of Quality Dimensions

Source: R. Mahanti, 2019

6 main Dimensions of Data Quality

Examples of the DQ Core Dimensions

Core dimension	Purpose	Method	Implementation
Completeness	Ensuring that all required data is captured and available, e.g., first name, last name, email address, and phone number	Data validation in the input forms that enforce mandatory fields before submission	Ensure that mandatory fields cannot be left blank
Uniqueness	No entity is represented more than once in a dataset (no duplicates)	Automated duplicate detection algorithms to flag duplicate records based on matching fields like email or address	duplicates should trigger alerts
Timeliness	Data is up-to-date and available when needed	Automated data refresh or batch processes to ensure data is updated regularly and in real-time when necessary	Critical data points (like inventory levels) should be updated within a predefined time frame
Validity	Data conforms to the required format and business rules	Pattern matching and regular expressions to validate data formats at the point of entry	e.g. phone numbers should follow a specific format
Accuracy	Data correctly describes the real-world object or event	Integrate third-party data validation tools during data entry to ensure correct information	e.g. address verification services
Consistency	Ensuring that data is the same across multiple systems or datasets	Use master data management (MDM) tools to synchronize data across multiple systems	Ensure that customer records in different systems (e.g., CRM, Billing) match and have consistent values for shared fields (like phone number and address)

DQ generic Rules (DQR)

DQRs are the antidote to potential errors

1. Missing Data

- Mandatory Fields Rule
- Conditional Completeness Rule

2. Duplicate Data

- Unique Identifier Enforcement Rule
- De-duplication Check Rule

3. Inconsistent Data

- Cross-System Consistency Rule
- Cross-Field Consistency Rule

4. Invalid Data

- Format Validation Rule
- Value Range Validation Rule
- Allowed Values Rule

5. Outdated Data

- Data Refresh Timeliness Rule
- Timestamp Validation Rule

6. Incorrect Data

- Data Verification Against External Sources Rule
- Business Logic Validation Rule

7. Unstandardized Data

- Standard Format Enforcement Rule
- Unit Conversion Standardization Rule

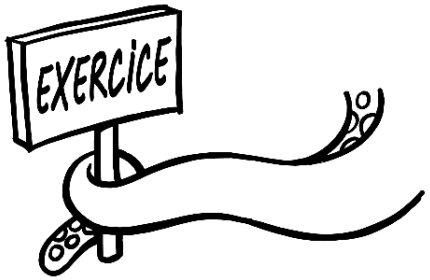
8. Misaligned Data

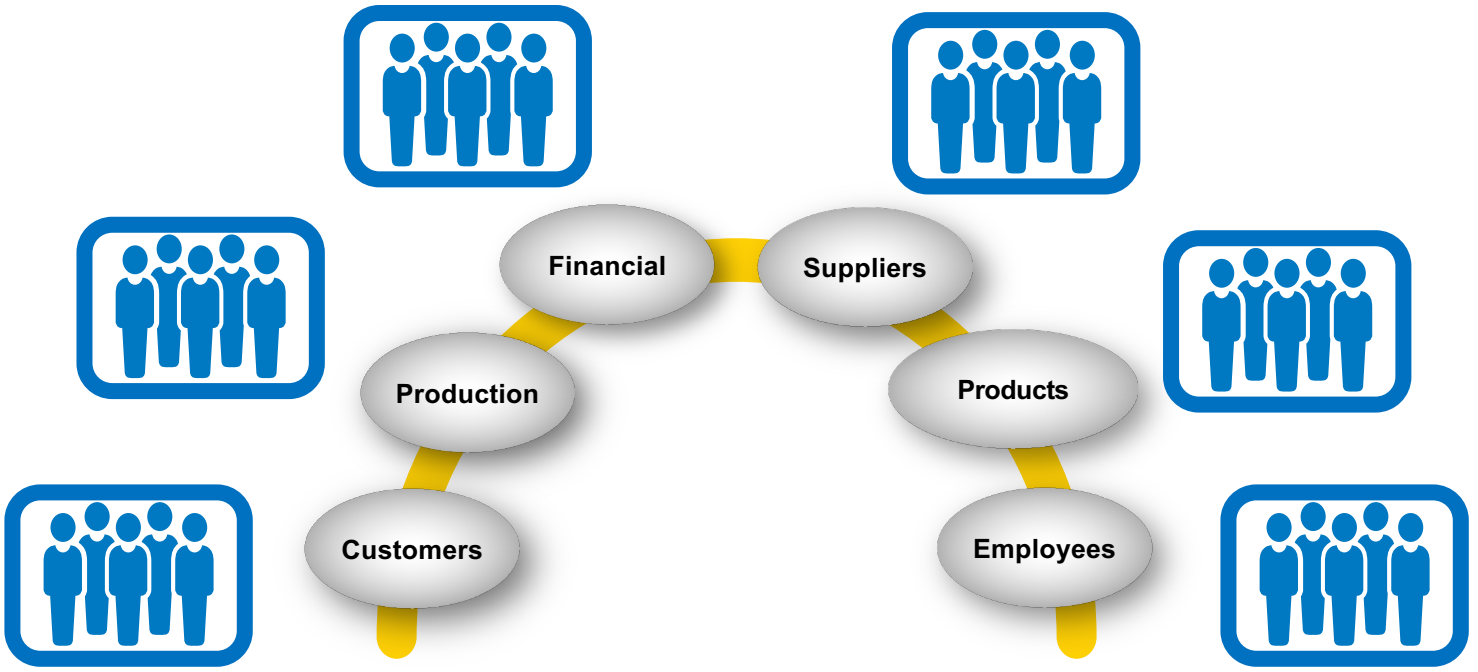
- Contextual Validation Rule
- Business Rule Validation (delivery date)

Which Data Quality Rules?

Which rules to maintain high data quality?

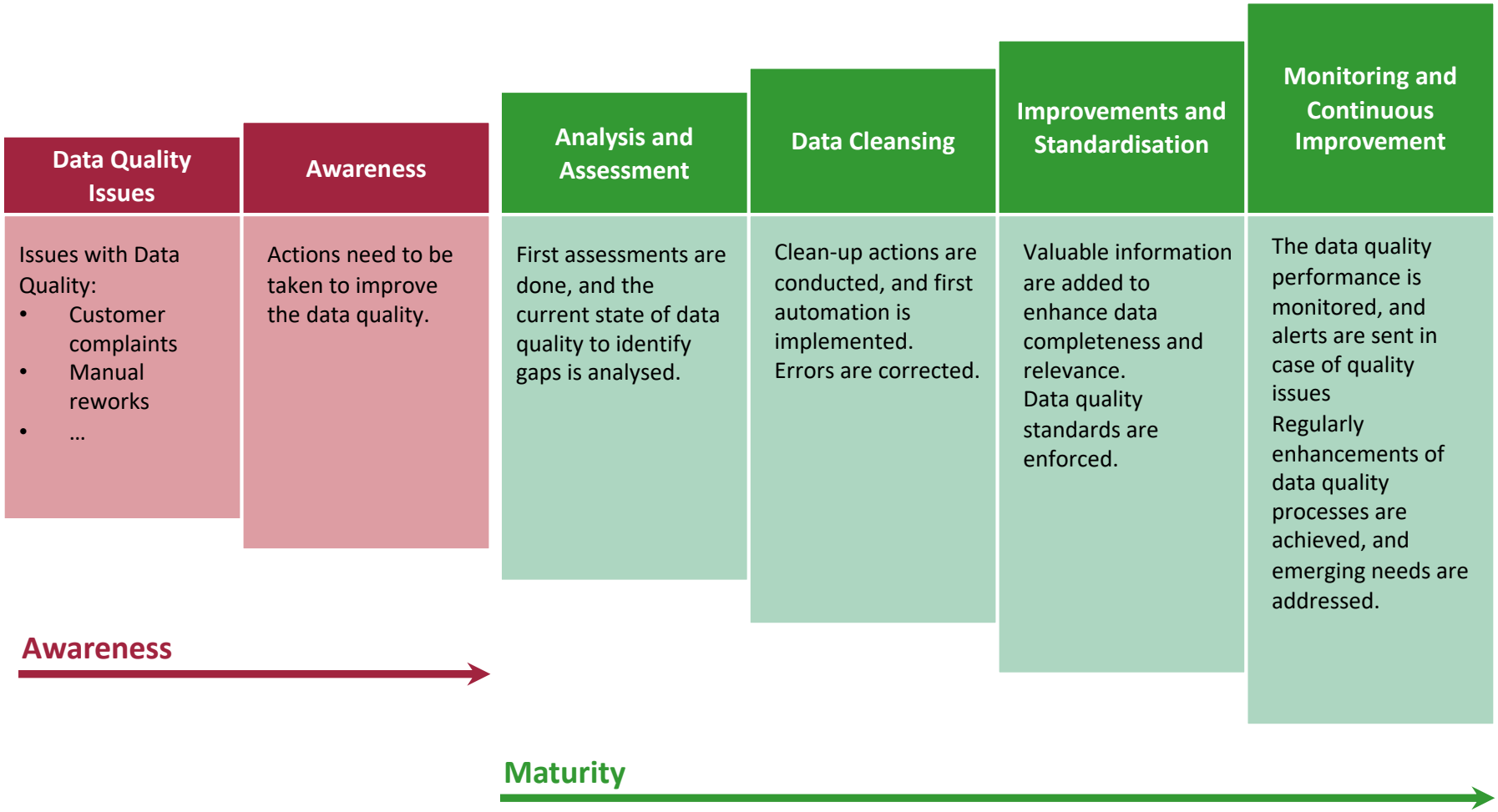
- Customer Data
- Production Data
- Financial Data
- Supplier Data
- Product Data
- Employee Data





Data Quality Management Process

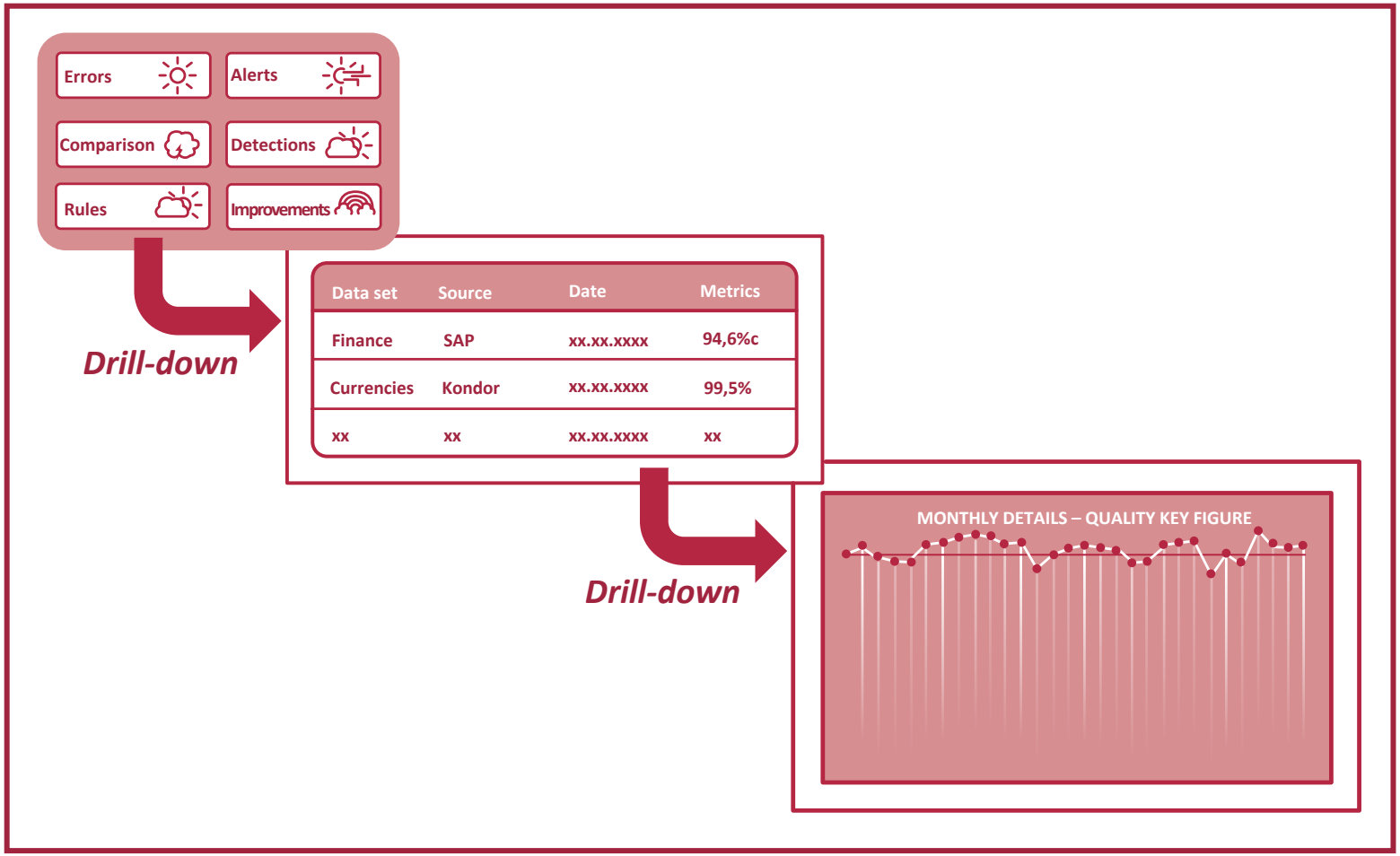
From the Awareness to a high Maturity



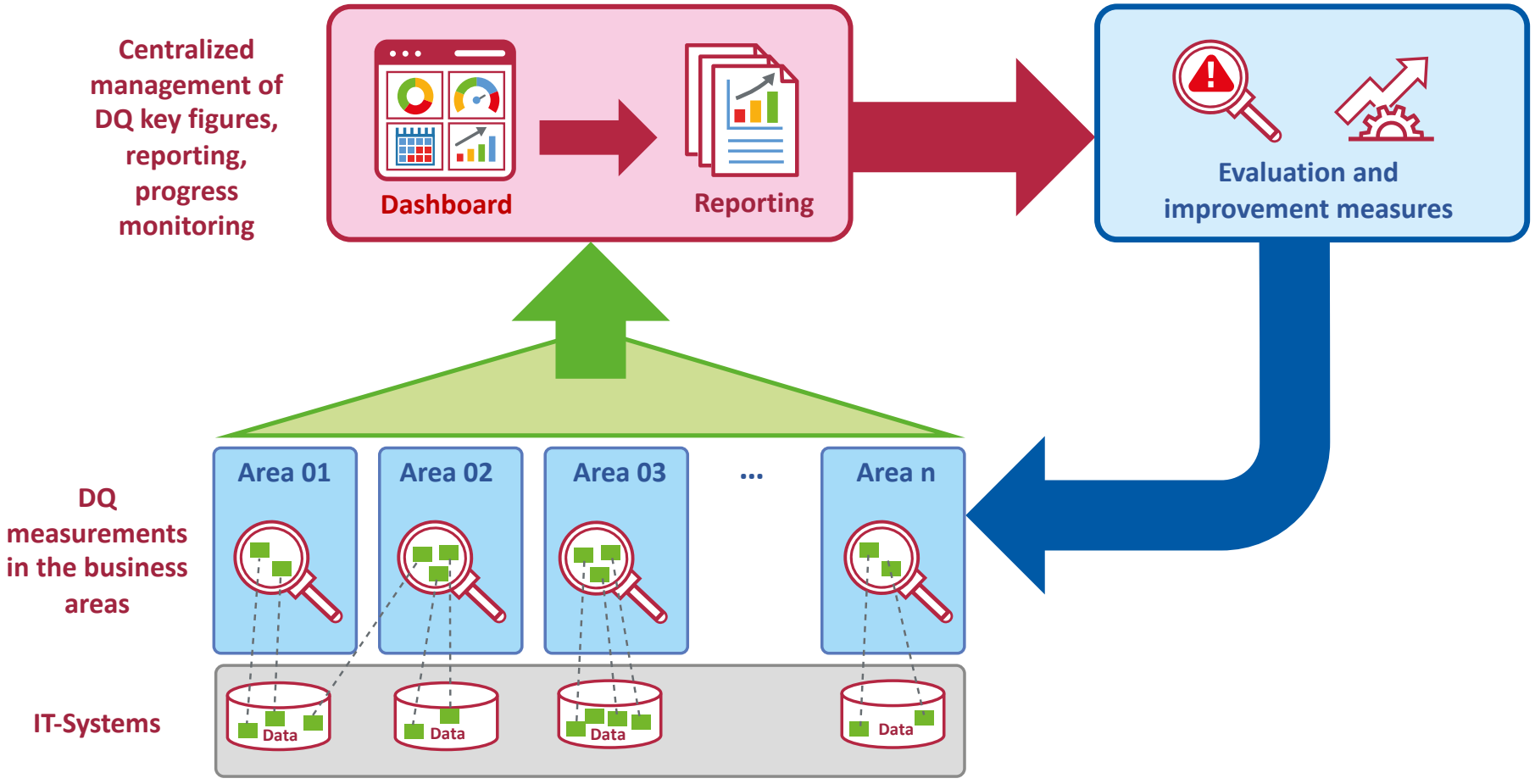
DQ Metrics

Core dimension	Example	Formula	Usage
Completeness	The proportion of missing or null fields in a dataset	$\frac{\text{Number of Missing Fields}}{\text{Total Number of Records}} \times 100$	Missing customer email addresses
Uniqueness	percentage of records that are duplicates	$\frac{\text{Number of Duplicate Records}}{\text{Total Number of Records}} \times 100$	Customer records appearing multiple times in the CRM system
Timeliness	Time it takes for data to be updated or made available after an event occurs	{Time of Event}–{Time Data is re corded}	Average delay between a transaction occurring and being reflected in the system
Validity	Percentage of data that fails to meet the defined format or criteria	$\frac{\text{Number of Invalid Records}}{\text{Total Number of Records}} \times 100$	Email addresses not conforming to user@domain.com
Accuracy	Percentage of data that is correct and matches reality	$\frac{\text{Number of Correct Records}}{\text{Total Number of Records}} \times 100$	Postal addresses checked against a valid database are correct
Consistency	Percentage of records where data is inconsistent between two systems or within the same dataset	$\frac{\text{Number of Inconsistent Records}}{\text{Total Number of Records}} \times 100$	Inconsistencies between systems if for instance a customer's address is recorded differently in two databases

Dashboard for Data Quality Monitoring

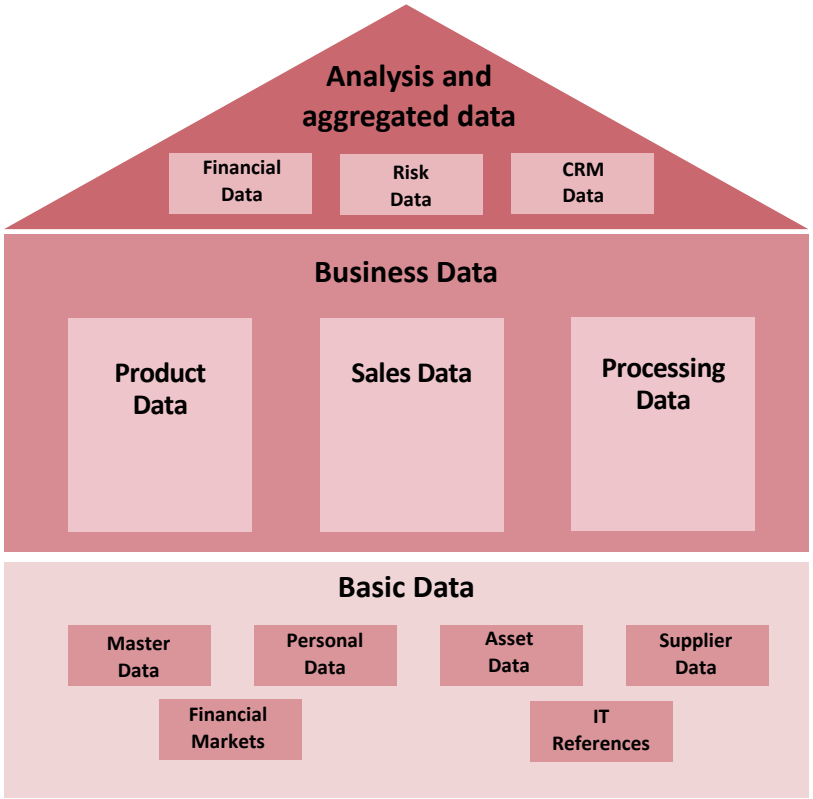


DQ Cycle


Centralized management of DQ key figures, reporting, progress monitoring
Dashboard
Reporting
Evaluation and improvement measures
DQ measurements in the business areas
Area 01
Area 02
Area 03
...
Area n
IT-Systems
Data
Data
Data
Data

Assignment of Data Responsible

Enterprise data model



Data Responsible

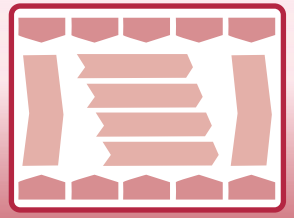
Number	Data Area	Responsible	Comments
Basic Data			
1.1	Master Data	xx	
1.1.1	Adresses		
1.1.2	Contact Information		
...			
1.2	Personal Data	HR	
1.2.1	Employee function		
1.2.2	Organigramm		
...			
1.3	Asset Data		
1.4	Supplier Data		
1.4.1	Contract		
1.4.2	SLAs		
...			
1.5	Financial Markets		
1.6	IT References		
1.6.1	Applications		
1.6.2	SW Licences		
...			
Business Data			
2.1	Product Data		
2.1.1	Product Description		
2.1.2	Pricing		
...			
2.2	Sales Data		
2.2.1	Product terms		
2.2.2	Revenues generated		
...			
2.3	Processing Data		
2.3.1	Transaction types		
2.3.2	Processing fees		
...			
Analysis and aggregated data			
3.1	Financial Data	CFO	
3.2	Risk Data	Compliance	
3.3	CRM Data	Business development	

The Cause-Effect of DQ

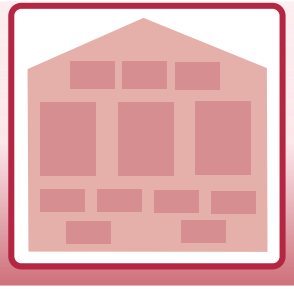
Strategy - Processes - Data

Business goals

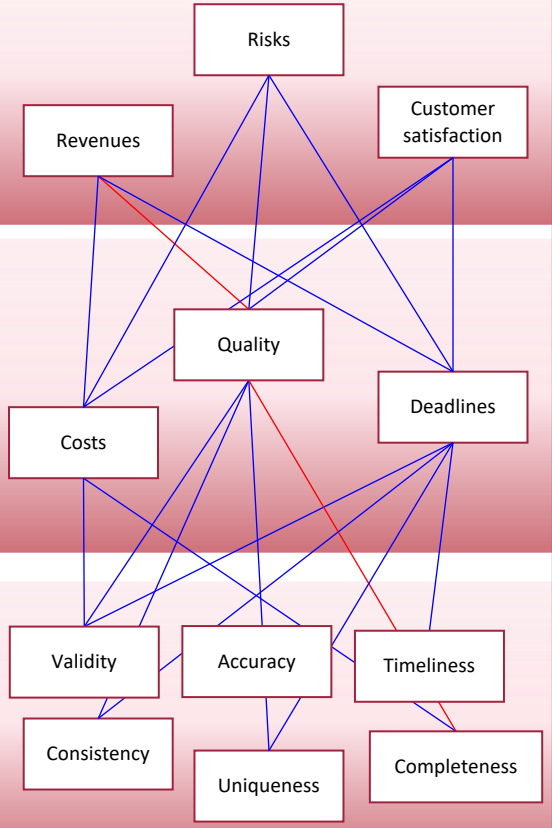
Processes



Data



Management systems



Causal chain

Management objectives

Efficient automated processes without manual intervention

Processing of payments

Straight Through Processing (STP)

Customer data

Completeness

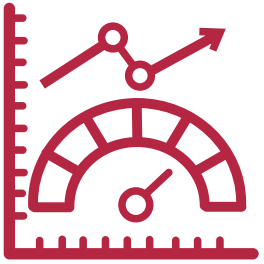
Example

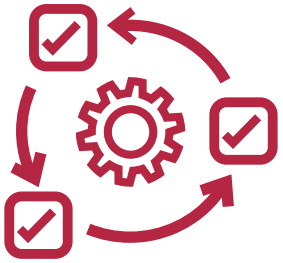
Too many manual adjustments have a negative impact on the STP rate, prevent the achievement of targets and cause additional costs

Payment orders can only be processed automatically with domicile addresses. If these are missing, the orders must be processed manually.

Domicile addresses must be available.

Which Critical Success Factors for DQM?





DQM Model

Framework or structured approach that organizations use to ensure the quality of their data throughout its lifecycle.
DQM models focus on establishing processes, tools, and metrics to assess, manage, and improve data quality.

DQM Model

Model	Focus	Key components	Usage
DAMA-DMBOK	Comprehensive data management across multiple domains, including data quality	Comprehensive data management across multiple domains, including data quality	Comprehensive data management across multiple domains, including data quality
ISO 8000-61 Data Quality Management Standard	Standardized approach for data quality management and data exchange	15 characteristics grouped into two categories: inherent data quality (e.g., accuracy, completeness) and system-dependent data quality (e.g., accessibility, confidentiality)	Provides guidelines for organizations to assess, improve, and maintain the quality of their data
DQAF Data Quality Assessment Framework	Framework for evaluating and improving the quality of statistical data	6 dimensions: Integrity, Methodological Soundness, Accuracy, Reliability, Serviceability, Accessibility	Commonly used by statistical offices and agencies for ensuring the quality and reliability of statistical data
EY Data Quality Maturity Model	Assessment of the maturity level of data quality practices within an organization, identifying strengths and gaps	5 maturity levels (aware, reactive, proactive, managed, optimized)	Helps organizations evaluate and benchmark their current state of data quality management and maturity
PwC Data Excellence Framework	Focuses on achieving data excellence by linking data quality with business objectives	From Data strategy to systems and technologies	Used to ensure data aligns with business goals and supports decision-making and operational efficiency
CIHI's Information Quality Framework	Quality of health information and data management practices	Process-oriented model including foundation, activities, outputs and outcomes	Tailored to health organizations for improving the quality and usability of health data

DQM-01: DAMA

10. Data Quality Management

Definition: Planning, implementation, and control activities that apply quality management techniques to measure, assess, improve, and ensure the fitness of data for use.

Goals:

- To measurably improve the quality of data in relation to defined business expectations.
- To define requirements and specifications for integrating data quality control into the system development lifecycle.
- To provide defined processes for measuring, monitoring, and reporting conformance to acceptable levels of data quality.

Inputs:

- Business Requirements
- Data Requirements
- Data Quality Expectations
- Data Policies and Standards
- Business Meta-data
- Technical Meta-data
- Data Sources and Data Stores

Suppliers:

- External Sources
- Regulatory Bodies
- Business Subject Matter Experts
- Information Consumers
- Data Producers
- Data Architects
- Data Modelers
- Data Stewards

Activities:

1. **Develop and Promote Data Quality Awareness (O)**
2. **Define Data Quality Requirements (D)**
3. **Profile, Analyze, and Assess Data Quality (D)**
4. **Define Data Quality Metrics (P)**
5. **Define Data Quality Business Rules (P)**
6. **Test and Validate Data Quality Requirements (D)**
7. **Set and Evaluate Data Quality Service Levels (P)**
8. **Continuously Measure and Monitor Data Quality (C)**
9. **Manage Data Quality Issues (C)**
10. **Clean and Correct Data Quality Defects (O)**
11. **Design and Implement Operational DQM Procedures (D)**
12. **Monitor Operational DQM Procedures and Performance (C)**

Participants:

- Data Quality Analysts
- Data Analysts
- Database Administrators
- Data Stewards
- Other Data Professionals
- DRM Director
- Data Stewardship Council

Tools:

- Data Profiling Tools
- Statistical Analysis Tools
- Data Cleansing Tools
- Data Integration Tools
- Issue and Event Management Tools

Primary Deliverables:

- Improved Quality Data
- Data Management Operational Analysis
- Data Profiles
- Data Quality Certification Reports
- Data Quality Service Level Agreements

Consumers:

- Data Stewards
- Data Professionals
- Other IT Professionals
- Knowledge Workers
- Managers and Executives
- Customers

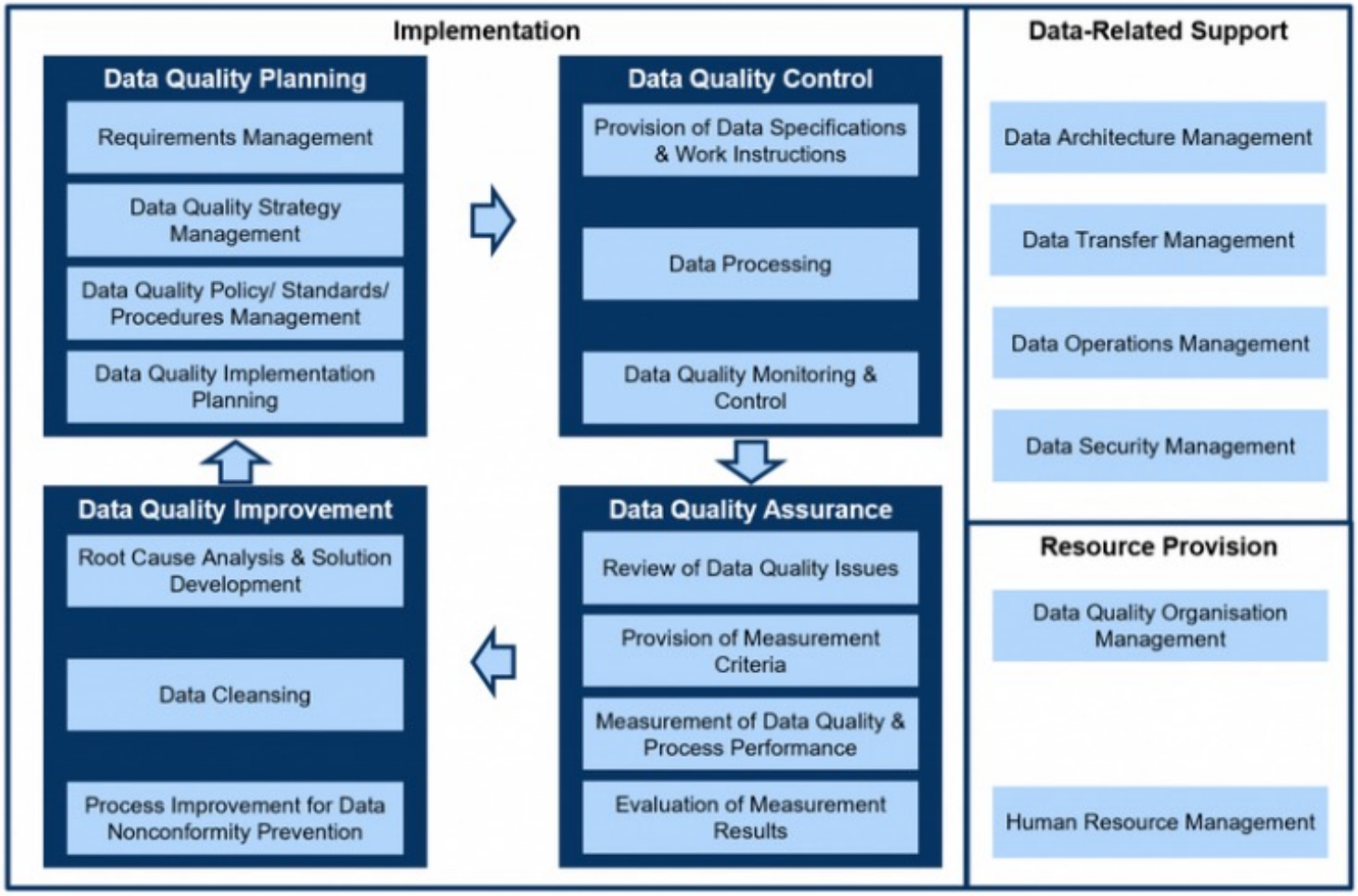
Metrics:

- Data Value Statistics
- Errors / Requirement Violations
- Conformance to Expectations
- Conformance to Service Levels

Activities: (P) – Planning (C) – Control (D) – Development (O) - Operational

Source: www.dama.org/cpages/dmbok-images

DQM-02: ISO 8000-61 DQ Process Reference Model



Source: www.iso.org/obp/ui/#iso:std:iso:8000:-61:ed-1:v1:en

DQM-03: Data Quality Assessment Framework (DQAF)

Quality Dimensions	Elements	Indicators
0. Prerequisites of quality	**0.1 Legal and institutional environment**—*The environment is supportive of statistics.*	0.1.1 The responsibility for collecting, processing, and disseminating the statistics is clearly specified. 0.1.2 Data sharing and coordination among data-producing agencies are adequate. 0.1.3 Individual reporters' data are to be kept confidential and used for statistical purposes only. 0.1.4 Statistical reporting is ensured through legal mandate and/or measures to encourage response.
	0.2 Resources—*Resources are commensurate with needs of statistical programs.*	0.2.1 Staff, facilities, computing resources, and financing are commensurate with statistical programs. 0.2.2 Measures to ensure efficient use of resources are implemented.
	0.3 Relevance—*Statistics cover relevant information on the subject field.*	0.3.1 The relevance and practical utility of existing statistics in meeting users' needs are monitored.
	0.4 Other quality management—*Quality is a cornerstone of statistical work.*	0.4.1 Processes are in place to focus on quality. 0.4.2 Processes are in place to monitor the quality of the statistical program. 0.4.3 Processes are in place to deal with quality considerations in planning the statistical program.
1. Assurances of integrity *The principle of objectivity in the collection, processing, and dissemination of statistics is firmly adhered to.*	**1.1 Professionalism**—*Statistical policies and practices are guided by professional principles.*	1.1.1 Statistics are produced on an impartial basis. 1.1.2 Choices of sources and statistical techniques as well as decisions about dissemination are informed solely by statistical considerations. 1.1.3 The appropriate statistical entity is entitled to comment on erroneous interpretation and misuse of statistics.
	1.2 Transparency—*Statistical policies and practices are transparent.*	1.2.1 The terms and conditions under which statistics are collected, processed, and disseminated are available to the public. 1.2.2 Internal governmental access to statistics prior to their release is publicly identified. 1.2.3 Products of statistical agencies/units are clearly identified as such. 1.2.4 Advance notice is given of major changes in methodology, source data, and statistical techniques.
	1.3 Ethical standards—*Policies and practices are guided by ethical standards.*	1.3.1 Guidelines for staff behavior are in place and are well known to the staff.
2. Methodological soundness *The methodological basis for the statistics follows internationally accepted standards, guidelines, or good practices.*	**2.1 Concepts and definitions**—*Concepts and definitions used are in accord with internationally accepted statistical frameworks.*	2.1.1 The overall structure in terms of concepts and definitions follows internationally accepted standards, guidelines, or good practices.
	2.2 Scope—*The scope is in accord with internationally accepted standards, guidelines, or good practices.*	2.2.1 The scope is broadly consistent with internationally accepted standards, guidelines, or good practices.
	2.3 Classification/ sectorization—*Classification and sectorization systems are in accord with internationally accepted standards, guidelines, or good practices.*	2.3.1 Classification/sectorization systems used are broadly consistent with internationally accepted standards, guidelines, or good practices.
	2.4 Basis for recording—*Flows and stocks are valued and recorded according to internationally accepted standards, guidelines, or good practices.*	2.4.1 Market prices are used to value flows and stocks. 2.4.2 Recording is done on an accrual basis. 2.4.3 Grossing/netting procedures are broadly consistent with internationally accepted standards, guidelines, or good practices.
3. Accuracy and reliability *Source data and statistical techniques are sound and statistical outputs sufficiently portray reality.*	**3.1 Source data**—*Source data available provide an adequate basis to compile statistics.*	3.1.1 Source data are obtained from comprehensive data collection programs that take into account country-specific conditions. 3.1.2 Source data reasonably approximate the definitions, scope, classifications, valuation, and time of recording required. 3.1.3 Source data are timely.

Quality Dimensions	Elements	Indicators
	3.2 Assessment of source data—*Source data are regularly assessed.*	3.2.1 Source data—including censuses, sample surveys, and administrative records—are routinely assessed, e.g., for coverage, sample error, response error, and nonsampling error; the results of the assessments are monitored and made available to guide statistical processes.
	3.3 Statistical techniques—*Statistical techniques employed conform to sound statistical procedures.*	3.3.1 Data compilation employs sound statistical techniques to deal with data sources. 3.3.2 Other statistical procedures (e.g., data adjustments and transformations, and statistical analysis) employ sound statistical techniques.
	3.4 Assessment and validation of intermediate data and statistical outputs—*Intermediate results and statistical outputs are regularly assessed and validated.*	3.4.1 Intermediate results are validated against other information where applicable. 3.4.2 Statistical discrepancies in intermediate data are assessed and investigated. 3.4.3 Statistical discrepancies and other potential indicators of problems in statistical outputs are investigated.
	3.5 Revision studies—*Revisions, as a gauge of reliability, are tracked and mined for the information they may provide.*	3.5.1 Studies and analyses of revisions are carried out routinely and used internally to inform statistical processes (see also 4.3.3).
4. Serviceability *Statistics, with adequate periodicity and timeliness, are consistent and follow a predictable revisions policy.*	**4.1 Periodicity and timeliness**—*Periodicity and timeliness follow internationally accepted dissemination standards.*	4.1.1 Periodicity follows dissemination standards. 4.1.2 Timeliness follows dissemination standards.
	4.2 Consistency—*Statistics are consistent within the dataset, over time, and with major datasets.*	4.2.1 Statistics are consistent within the dataset. 4.2.2 Statistics are consistent or reconcilable over a reasonable period of time. 4.2.3 Statistics are consistent or reconcilable with those obtained through other data sources and/or statistical frameworks.
	4.3 Revision policy and practice—*Data revisions follow a regular and publicized procedure.*	4.3.1 Revisions follow a regular and transparent schedule. 4.3.2 Preliminary and/or revised data are clearly identified. 4.3.3 Studies and analyses of revisions are made public (see also 3.5.1).
5. Accessibility *Data and metadata are easily available and assistance to users is adequate.*	**5.1 Data accessibility**—*Statistics are presented in a clear and understandable manner, forms of dissemination are adequate, and statistics are made available on an impartial basis.*	5.1.1 Statistics are presented in a way that facilitates proper interpretation and meaningful comparisons (layout and clarity of text, tables, and charts). 5.1.2 Dissemination media and format are adequate. 5.1.3 Statistics are released on a preannounced schedule. 5.1.4 Statistics are made available to all users at the same time. 5.1.5 Statistics not routinely disseminated are made available upon request.
	5.2 Metadata accessibility—*Up-to-date and pertinent metadata are made available.*	5.2.1 Documentation on concepts, scope, classifications, basis of recording, data sources, and statistical techniques is available, and differences from internationally accepted standards, guidelines, or good practices are annotated. 5.2.2 Levels of detail are adapted to the needs of the intended audience.
	5.3 Assistance to users—*Prompt and knowledgeable support service is available.*	5.3.1 Contact points for each subject field are publicized. 5.3.2 Catalogs of publications, documents, and other services, including information on any changes, are widely available.

Source: www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm

DQM-04: EY Data Quality Maturity Model (1/2)

★ = Where we see the industry

	Stage 1: aware	Stage 2: reactive	Stage 3: proactive	Stage 4: managed	Stage 5: optimized
Data quality program	A DQ program was set-up; however, it does not interface with the business and operates in a vacuum.	The DQ program exists but only is concerned with limited areas of the data or business or the program does not consider prioritization in its planning.	★ Program with full-time employees exists and there are regular meetings with the business. Executive sponsorship to alleviate barriers has not been gained at all levels.	A DQ program has been set up with executive sponsorship and regular interaction with the business; however, limited areas of the data and the enterprise have not been involved.	DQ program is mature and understood throughout the business. Executive sponsorship has been gained at all levels and the appropriate representatives from the business are involved to prioritize and make decisions about the direction of clean-up.
Data quality people	Limited business sponsors have been identified, but they are not actively involved in data quality efforts and setting of direction of future data clean-up projects.	Certain pockets of the organization have identified business sponsors for data quality efforts but no business analysts or technical analysts.	★ Business sponsors have been identified and participate in the program; areas still need sponsorship. business analysts and technical analysts are assigned.	Business sponsors have largely been identified across the organization and buy-in has been gained, minor improvements need to be made in a few areas.	DG has identified sponsors for all areas of the business. Sponsors are involved in ownership, accountability and buy-in for DQ improvements. Business/technical analysts have roles and responsibilities.
Data quality metrics	Some use of metrics is in place; however, they are not regularly updated or are not areas of interest to the DQ program.	★ Creation and generation of metrics is performed for some DQ initiatives. Trouble determining the progress against improvement efforts.	Data-driven projects are actively defining metrics, but they are not consistently updated and socialized with the project stakeholders.	Data-driven projects widely generate metrics and consistently update them; however, there is room for improvement in their comprehensiveness.	DQ metrics are generated and regularly updated across the swathe of data improvement initiatives. Their definitions are agreed upon, widely understood and are indicative of progress against DQ efforts.
Data field standardization	Efforts to ensure data field standardization are sporadic and limited to systems and fields. Approval process is non-existent for changing the definition of fields.	Some efforts to standardize data fields have taken place; however, the effort has not been prioritized and significant issues still exist.	★ Fields across systems are largely standardized; however, use of a data definition dictionary is not commonplace.	No major issues exist with data field standardization and efforts are performed during system development to ensure that fields adhere to definitions.	Data fields have standard definitions and calculations across systems and businesses. Definition of new fields is reviewed against existing fields to ensure meaning. Changes to fields and calculations are widely socialized.
Business rule validation	Limited data standards are defined; however, they may not adequately reflect the business model.	Data standards are defined for certain processes and systems; however, there are significant gaps.	★ Data standards are defined based upon business rules and systematically enforced. Users are informally made aware of data standards. Training has not been deployed to all parts of the organization. Documentation of data standards is mostly complete.	Data standards are defined based upon business rules and systematically enforced. Users receive training on most standards for data entry and standards have been documented in a single repository. Change control process is not followed.	Data standards are defined based upon business rules and systematically enforced, where applicable. Users receive training on standards for data entry and standards have been documented in a single repository. A change control process is followed upon request for a change to a data standard.

Source: EY

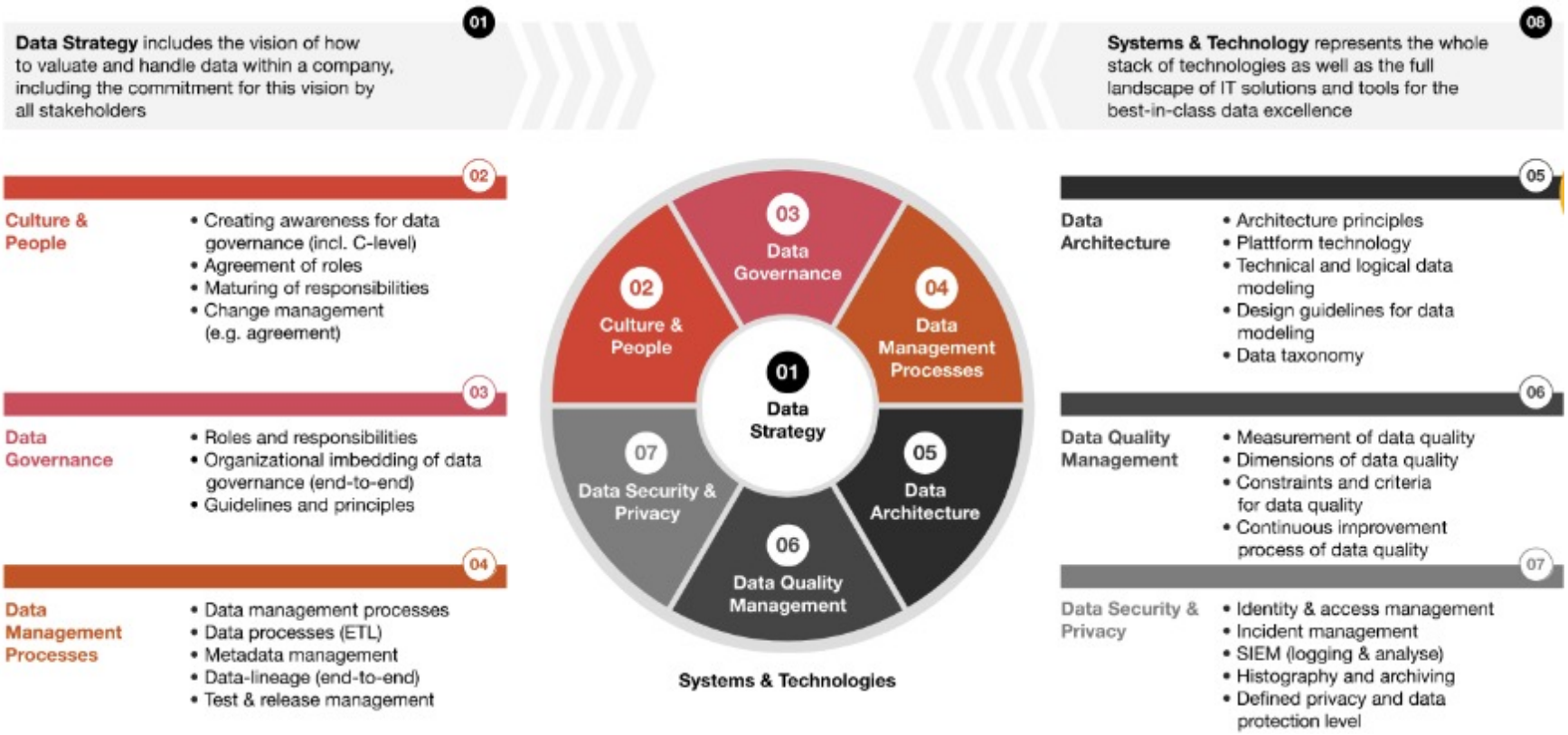
DQM-04: EY Data Quality Maturity Model (2/2)

★ = Where we see the industry

	Stage 1: aware	Stage 2: reactive	Stage 3: proactive	Stage 4: managed	Stage 5: optimized
Data cleansing	Limited data standards are defined; however, they may not adequately reflect the business model.	Data standards are defined for certain processes and systems; however, there are significant gaps.	Data standards are defined based upon business rules and systematically enforced. Users are informally made aware of data standards Training has not been deployed to all parts of the organization. Documentation of data standards is mostly complete.	Data standards are defined based upon business rules and systematically enforced. Users receive training on most standards for data entry and standards have been documented in a single repository. Change control process is not followed.	Data standards are defined based upon business rules and systematically enforced, where applicable. Users receive training on standards for data entry and standards have been documented in a single repository. A change control process is followed upon request for a change to a data standard.
Data monitoring	The business uses data quality related tools in places, but does not exploit the capabilities offered to a sufficient degree as to derive real quality improvements.	The client is using tools but only in parts of the organization – tools may be felt to cause more problems than they solve.	Most data fixes are done using applications or custom developed cleansing modules. Gaps are filled in relation to manipulations on data to ensure fixes deployed.	Monitoring notifications are in place for most processes. Notifications provide user of reason why it's incorrect, business impact and instructions on how to fix the problem.	Monitoring notifications are in place for most processes. Notifications provide user of reason why it's incorrect, business impact and instructions on how to fix the problem. A variety of mechanisms are employed in order to best meet business requirements.
Metadata	Metadata is not centrally managed or maintained.	BU focused information and metadata repositories. Metadata is limited in scope and usage.	Centralized metadata process and tools established.	There is an active use of data lineage and impact analysis across metadata repositories.	Fully integrated metadata with wide array of operational and analytical platforms.
Master data management	Multiple copies of reference data. No master data models in place. Lack of ownership and stewardship roles for common master data.	Few or project driven data and content models exist. Data warehouse used as a core repository for master data; limited attempts to consolidate reference data.	Shared master data repository architecture in place. Golden source of data established. Key master data services defined.	Governance roles are implemented. Integration of business rules with master data operations.	Harmonization of master data across consuming and providing systems. Service components implemented for establishing new data object types.

Source: EY

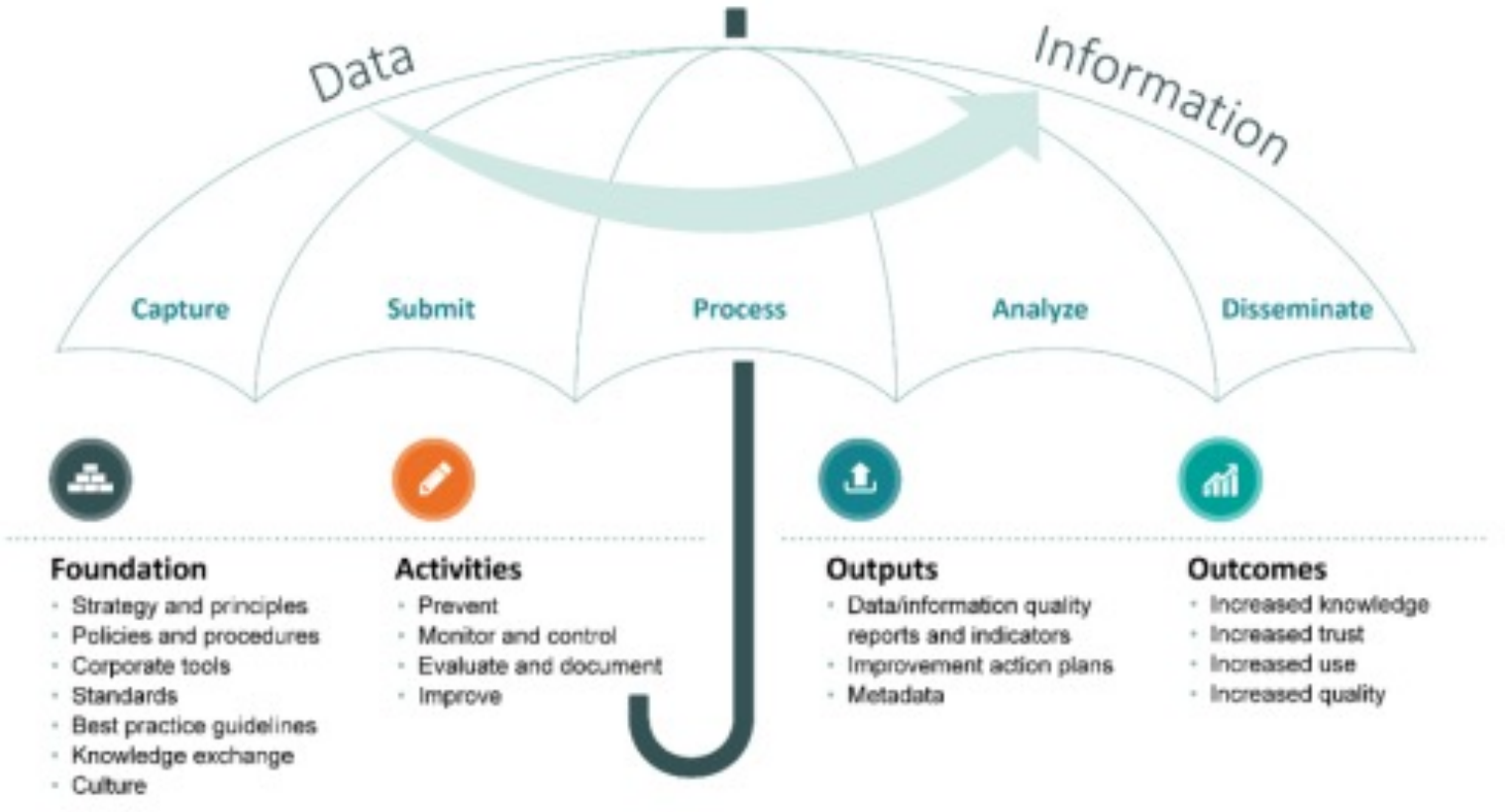
DQM-05: PwC Data Excellence Framework



Source: www.pwc.de/en/customer-centric-transformation/data-and-quality-management.html

DQM-06: CIHI's Information Quality Framework

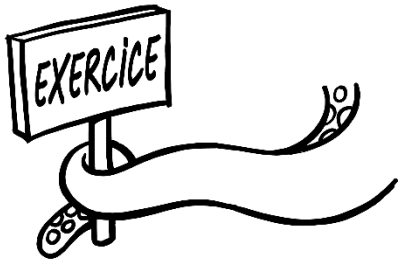
CIHI: Canadian Institute for Health Information

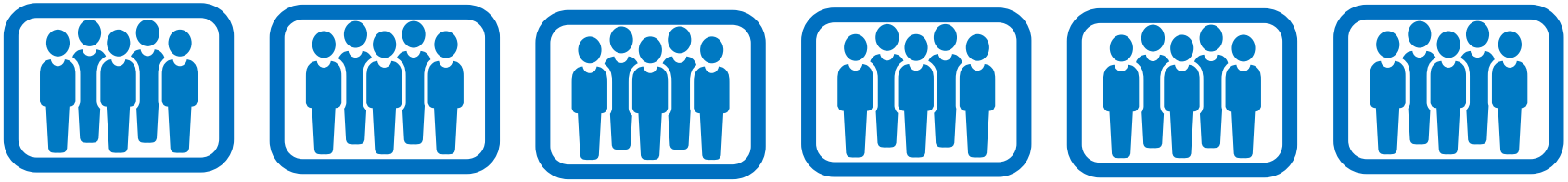


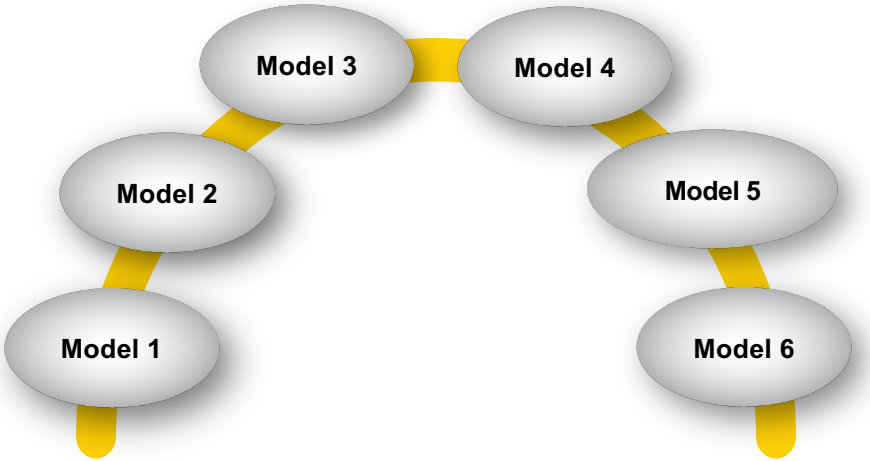
Source: www.cihi.ca/sites/default/files/document/iqf-summary-july-26-2017-en-web_0.pdf

Develop your own DQM Model

- Give the common elements of the 6 models presented
- Show the differences
- Design your own DQM Model







A Data Quality Framework in six Steps

01

Define data quality requirements

- ✓ Meta data and Data sources
- ✓ Type, Format, Range, character set
- ✓ Cardinality

02

Data Analysis

- ✓ Statistics
- ✓ First summary of the data quality reported
- ✓ DQD: completeness, validation, currency, accuracy

03

Quality Rules Proposals

- ✓ Based on some elementary DQDs that fit all varieties and data types.

04

Quality Evaluation

- ✓ Apply quality rules (step3) on each attributes
- ✓ Mean of primary quality score

05

Quality Remidiation

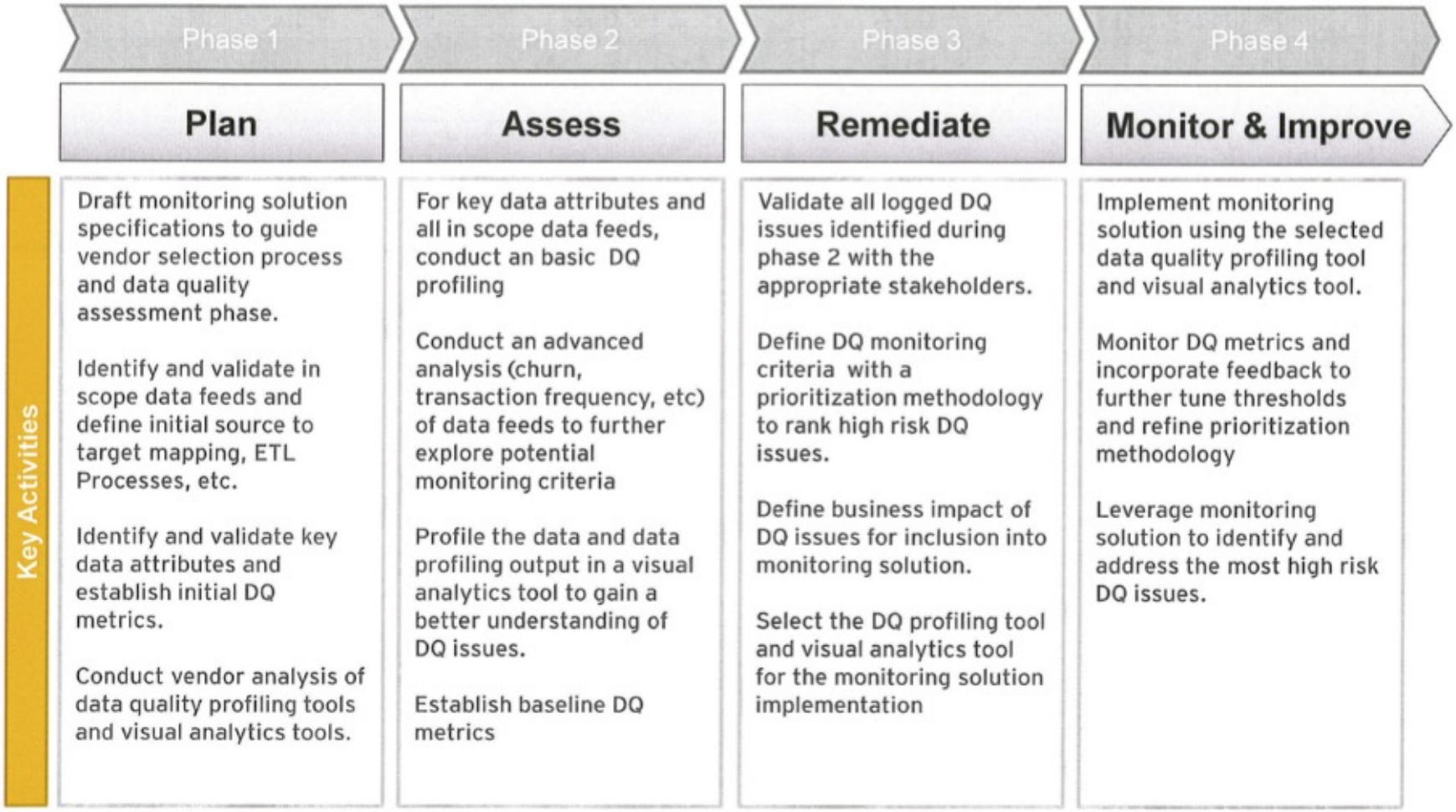
- ✓ Data Cleansing
- ✓ Repeat Step(1-4)
- ✓ Update quality rules
- ✓ Update quality scores

06

Tracing

- ✓ Monthly report
- ✓ Monitoring of the overall quality score

Data Quality Monitoring according to EY



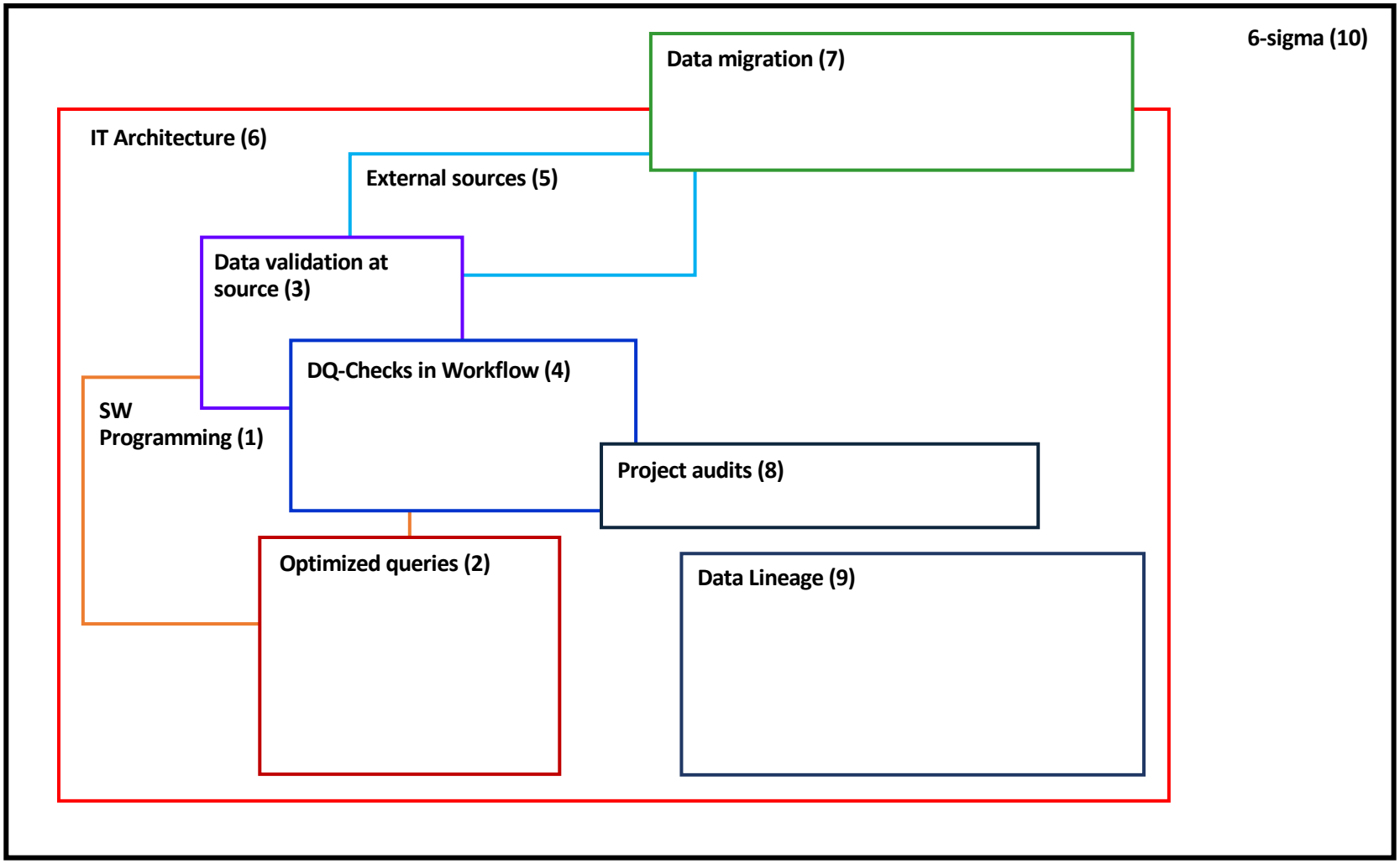
Source: www.ey.com/en_ch/

Which Chance of Success?

"If at first you don't succeed, change your definition of success."

10 Best Practices (BPs)

From “quick wins” to more sophisticated approaches



Unicode: Universal Encoding

Unicode is a universal character encoding standard that allows computers to represent and manipulate text from any writing system, including letters, numbers, and symbols from nearly every language. It was designed to overcome the limitations of older character encodings, which were often specific to a single language or group of languages.

0x0041 = A
U+00C7 = Ç
U+00FC = ü
U+4F60 = 你 (you)

Optimized Queries (BP-02)

These queries should be efficient in execution and comprehensive in their scope to identify issues such as duplicates, null values, outliers, or inconsistencies.

Null values often indicate incomplete or missing data. An optimized query to find rows where a column contains `NULL`:

sql — Code kopieren

```sql
SELECT *
FROM table_name
WHERE column_name IS NULL;
```

Adding the proper index on `column_name` will make this query more efficient by avoiding full table scans.

You can check if numeric columns have values outside of a valid range (outliers):

sql — Code kopieren

```sql
SELECT *
FROM table_name
WHERE column_name < valid_min_value
   OR column_name > valid_max_value;
```

This query checks for outliers and can be further optimized by having an index on `column_name`.

For data consistency, you might want to check for orphaned rows in child tables (foreign key relationships):

sql — Code kopieren

```sql
SELECT child_table.*
FROM child_table
LEFT JOIN parent_table ON child_table.parent_id = parent_table.id
WHERE parent_table.id IS NULL;
```

This query identifies rows in the child table that do not have a corresponding record in the parent table.

You can check if a column has an even distribution or unexpected skew:

sql — Code kopieren

```sql
SELECT column_name, COUNT(*)
FROM table_name
GROUP BY column_name
ORDER BY COUNT(*) DESC;
```

This query will provide the distribution of values, useful for detecting anomalies in categorical data.

Data Validation at Sources (BP-03)

Ensuring that data is accurate, complete, and valid at the point of entry into a system or database and inaccuracies are caught before the data propagates further into the system

- **Example:** Ensuring that a user input in an "email" field follows the correct format (e.g., `example@domain.com`).

sql — Code kopieren

```sql
SELECT email
FROM users
WHERE email NOT LIKE '%_@__%.__%';
```

- **Example:** Validating that a user's age falls between 18 and 65.

sql — Code kopieren

```sql
SELECT age
FROM users
WHERE age < 18 OR age > 65;
```

- **Example:** Making sure that a customer record includes mandatory fields like "first name" and "email".

sql — Code kopieren

```sql
SELECT *
FROM customers
WHERE first_name IS NULL OR email IS NULL;
```

- **Example:** Validating that a username does not exceed 20 characters.

sql — Code kopieren

```sql
SELECT username
FROM users
WHERE LEN(username) > 20;
```

Incorporate DQ-Checks in Workflow Automation (BP-04)

You can create a trigger that checks for invalid `order_amount` values when an insert or update occurs. If the amount is invalid, it will log the issue in a `data_quality_issues` table and send an alert via an email (or other alert mechanism).

```sql
-- Table to log data quality issues
CREATE TABLE data_quality_issues (
    issue_id INT PRIMARY KEY AUTO_INCREMENT,
    table_name VARCHAR(100),
    issue_description VARCHAR(255),
    detected_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP
);

-- Trigger to check order amount validity
DELIMITER $$

CREATE TRIGGER trg_check_order_amount
BEFORE INSERT ON orders
FOR EACH ROW
BEGIN
    IF NEW.order_amount <= 0 THEN
        -- Insert into the data quality issues table
        INSERT INTO data_quality_issues (table_name, issue_description)
        VALUES ('orders', CONCAT('Invalid order amount detected for order_id: ', NEW.order

        -- You can also send an email alert using a stored procedure or external system he
        -- Example: CALL send_alert('Data quality issue detected for order_id: ', NEW.orde

        -- Prevent the invalid data from being inserted
        SIGNAL SQLSTATE '45000'
        SET MESSAGE_TEXT = 'Order amount must be greater than zero';
    END IF;
END$$

DELIMITER ;
```

1) **Trigger Activation**
   The trigger is activated before an insert operation on the orders table
2) **Validation**
   It checks if the new order's order_amount is less than or equal to 0.
3) **Log Issue**
   If the condition is true (i.e., the data is invalid), the trigger inserts a record into the data_quality_issues table, logging the issue.
4) **Alert Mechanism**
   A procedure sends an email or alert to the data quality team.
5) **Abort the Insert**
   The SIGNAL SQLSTATE statement prevents the invalid data from being inserted into the orders table and provides a clear error message to the user

After a failed attempt to insert invalid data, the `data_quality_issues` table might look like this:

issue_id	table_name	issue_description	detected_at
1	orders	Invalid order amount detected for order_id: 10	2024-09-15 10:15:00

External Sources (BP-05)

Some useful sources

Reference Data

- ISO Standards (country codes, currency codes,…)
- Postal Services for postal code validation and address standardization

Third-Party Data

- NOGA Codes
- Contact information (addresses, emails, phone numbers)
- Demographic and geographic data

Government Data and Regulatory Sources

- Financial records
- Validation of business data
- Demographic data

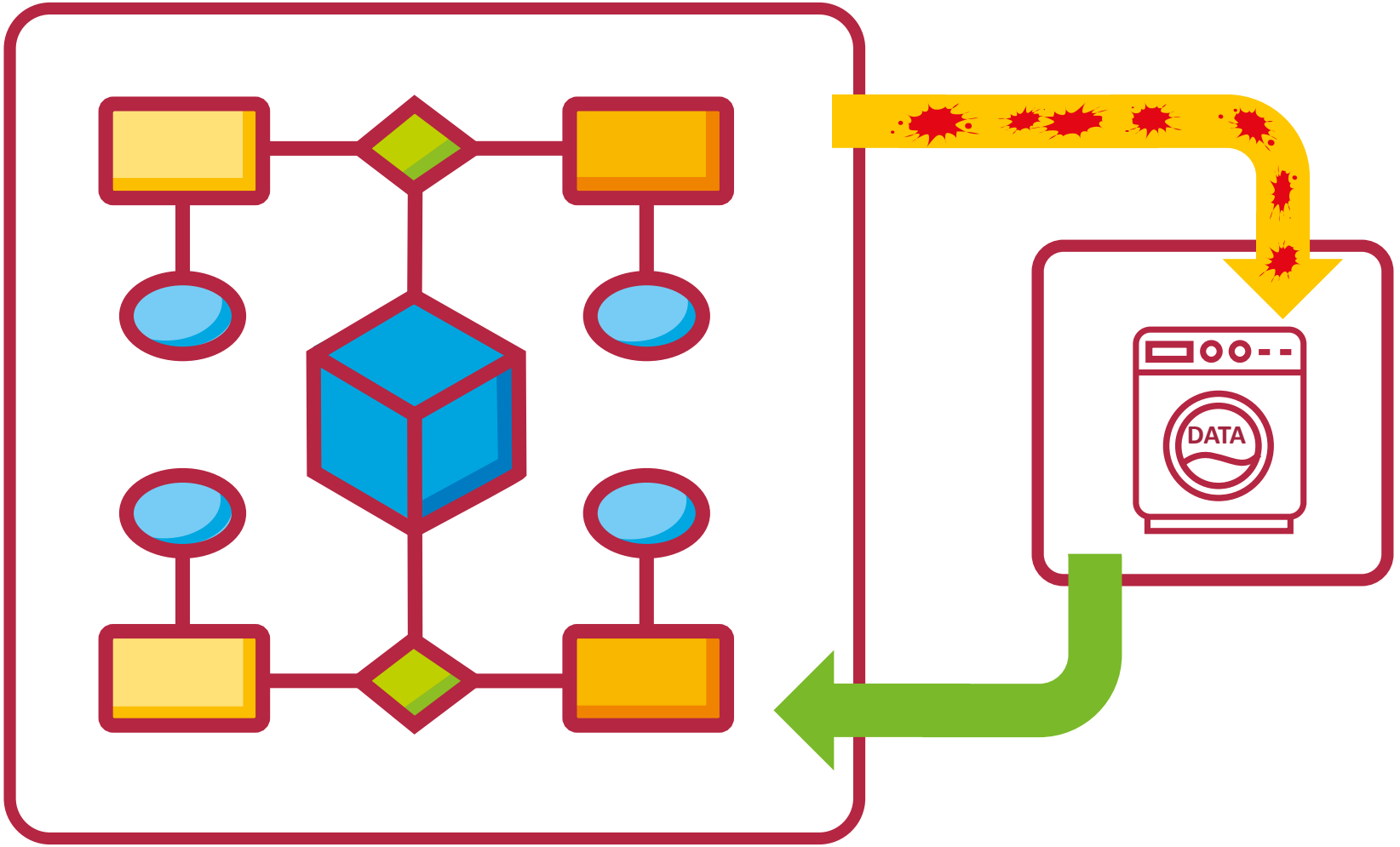
Financial and Credit Data

- Creditworthiness checks
- Financial data validation
- Cross-referencing customer data with sanction and anti-money laundering

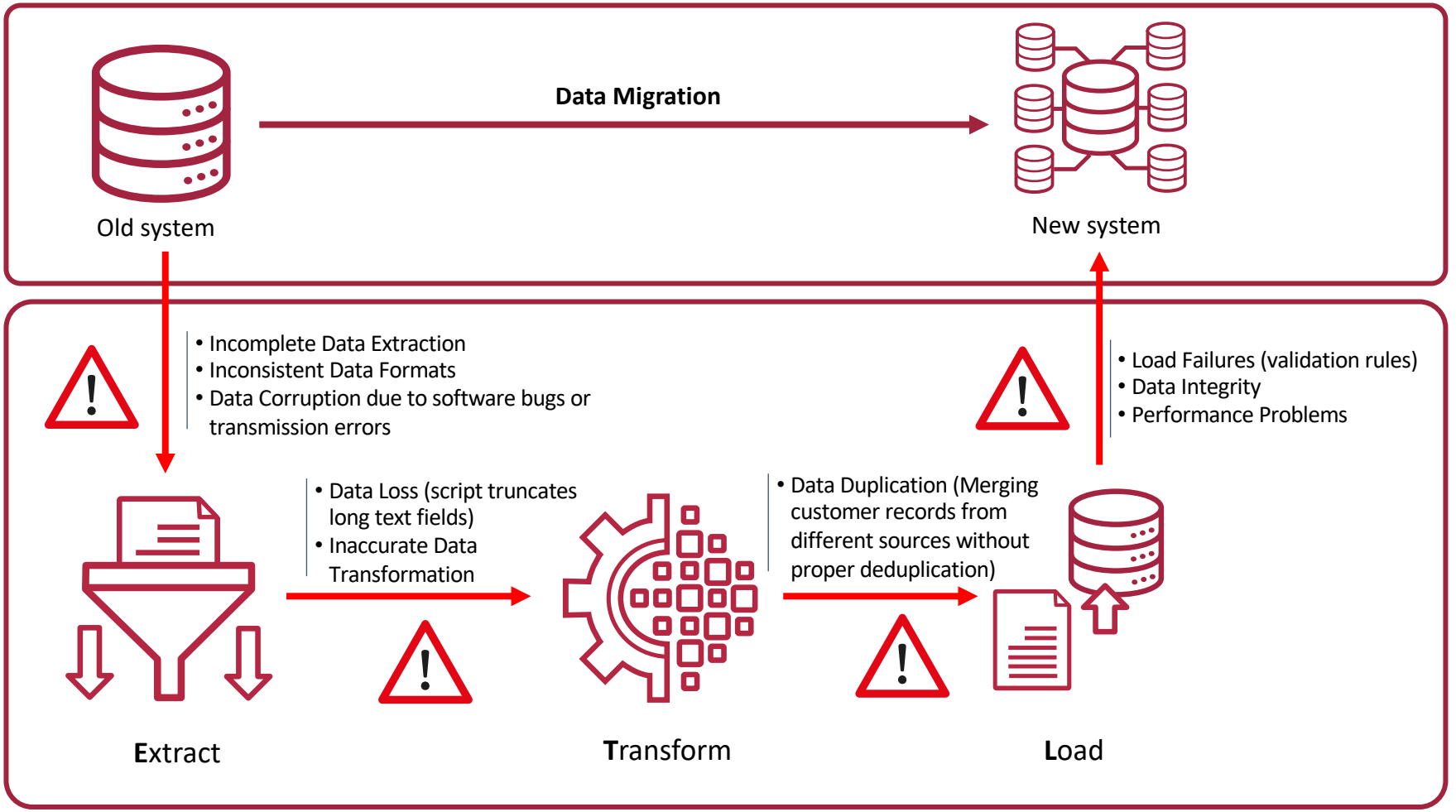
Fraud Detection and Anti-Money Laundering (AML)

- Verification and fraud detection
- Identity resolution and risk detection
- List of politically exposed persons (PEPs)

IT Architecture: the Data Washing Machine (BP-06)



ETL: Extract, Transform and Load



Data Quality / Compliance Audit in Projects (BP-08)

Reference Data

- ✓ Country Codes
- ✓ Product Categories
- ✓ Vendor Codes

Master Data

- ✓ Customer Data
- ✓ Supplier Data
- ✓ Product Data

Transactional Data

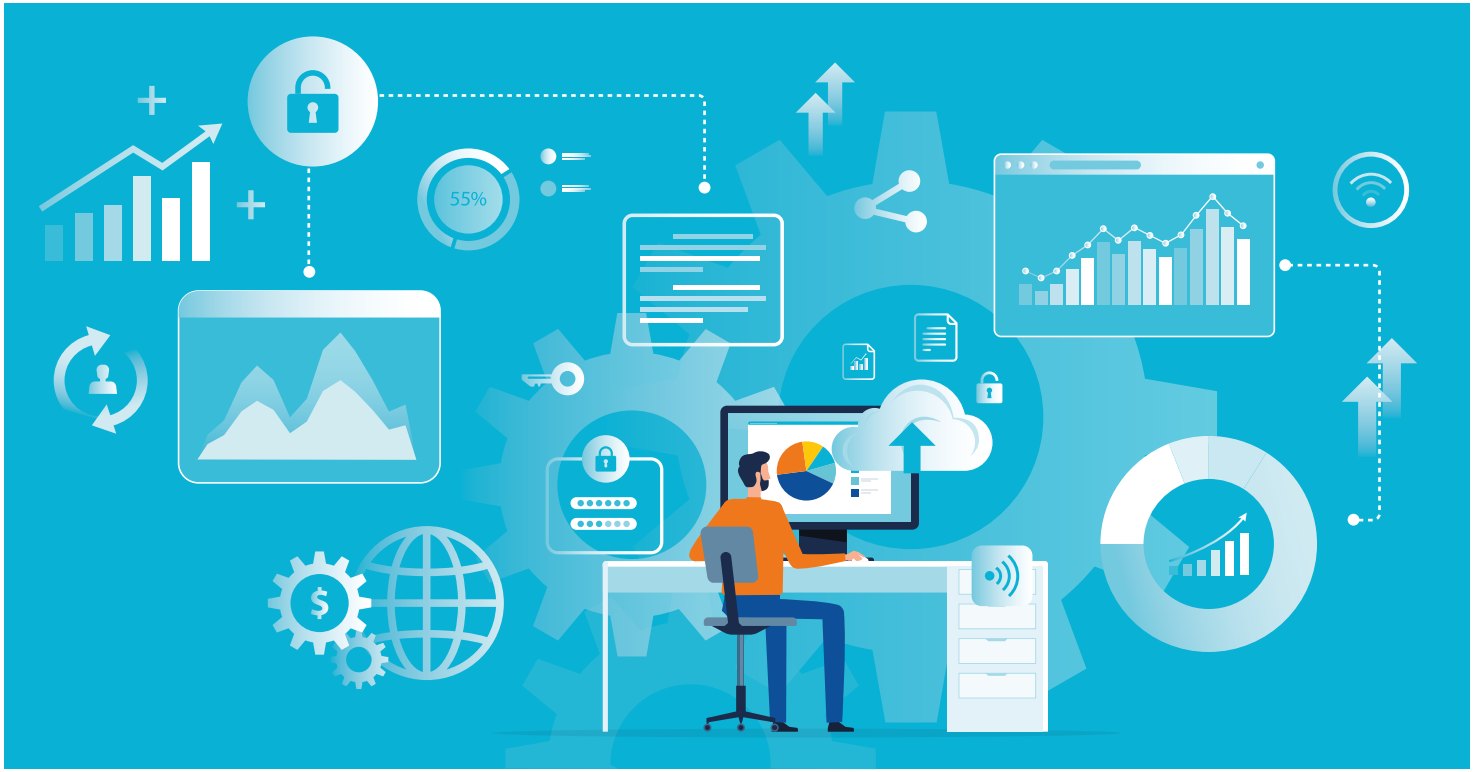
- ✓ Sales Orders
- ✓ Purchase Orders
- ✓ Invoices

Financial Data

- ✓ General Ledger Entries
- ✓ Financial Statements
- ✓ Account Balances

Data Lineage (BP-09)

Tracking and visualizing the flow of data through various stages in its lifecycle, from its origin to its destination



- Transparency and Understanding
- Impact Analysis
- Error Tracing and Root Cause Analysis
- Efficient Troubleshooting
- Data Provenance
- Audit Trails
- Data Ownership and Stewardship
- Enhanced Data Integration
- Data Integrity and Accuracy

6 - Sigma for Data Quality (BP-10)

Example for DQ issues

D	M	A	I	C
Define Identify the problem and define the scope of the improvement project	**Measure** Collect and measure data to understand the current state and quantify the problem	**Analyze** Analyze the data to identify root causes of defects or issues	**Improve** Develop and implement solutions to address the root causes and improve the process	**Control** Establish controls and monitoring systems to sustain the improvements and ensure long-term success
Customer data has a high rate of inaccuracies, resulting in frequent errors in marketing campaigns and customer service	Measurement of the accuracy of customer data by comparing it to verified sources and calculates the defect rate in terms of incorrect or missing information	Inaccuracies in customer data are due to inconsistent data entry practices across different departments and lack of standardized validation rules	Implementation of standardized data entry procedures, introduction of data validation rules at the point of entry, and training to employees on data quality best practices	Regular data quality audits, monitoring of data accuracy through dashboards, and feedback loop to continuously refine data entry practices

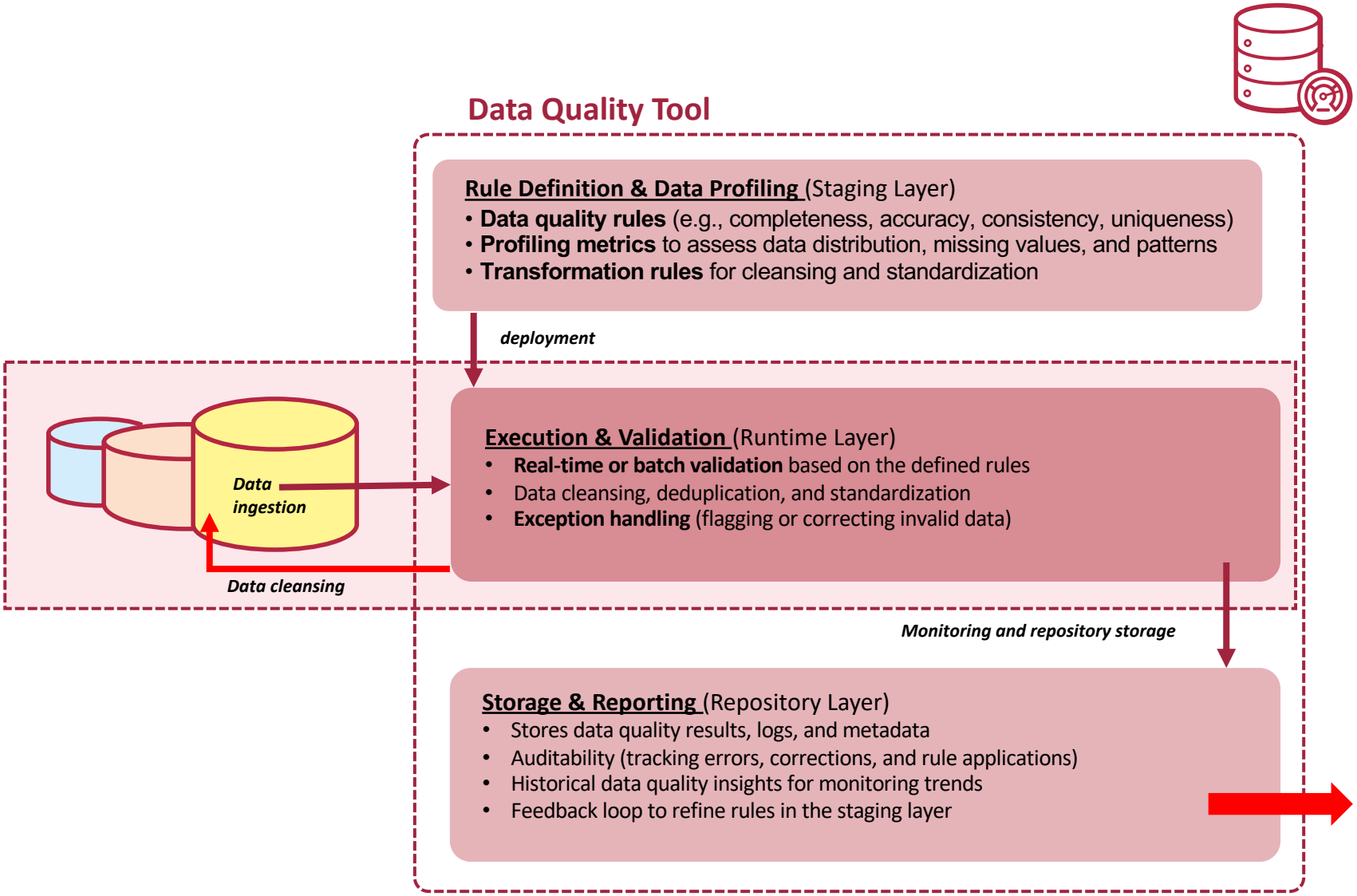
Data Quality Tool (DQT)

DQT refers to a **Data Quality Tool**, which is a software used to assess, manage, and improve the quality of data within a system.

These tools help organizations ensure their data is accurate, consistent, and reliable by performing tasks such as:

- **Data Cleansing**
  Identifying and correcting errors or inconsistencies in data
- **Data Profiling**
  Analyzing data to understand its structure, completeness, and quality
- **Data Validation**
  Ensuring data meets predefined rules or standards
- **Data Deduplication**
  Removing duplicate records to maintain uniqueness
- **Data Enrichment**
  Enhancing existing data with additional, relevant information

How does a DQT work?



Example of Reporting from a DQT

Rules

Rule designer to check contact information of customers
(Address, Phone Number, Email)

1 Completeness Rules

Rule 1.1: All customer records must have a non-null and non-empty value for the Address field
Rule 1.2: All customer records must have a non-null and non-empty value for the Phone Number field
Rule 1.3: If the address is provided, it must include at least the following components: Street, City, Postal Code, and Country
Rule 1.4: If a phone number is provided, it must include the country code

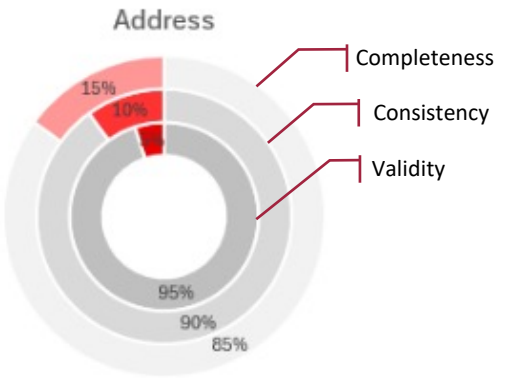
2. Consistency Rules

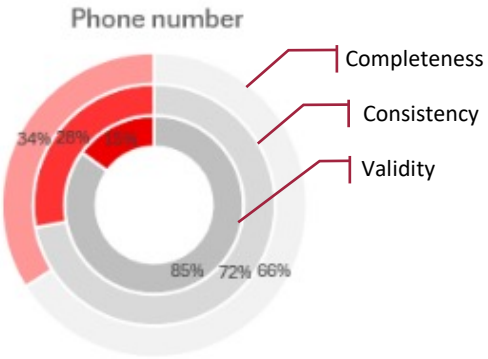
Rule 2.1: The Phone Number field must follow a consistent format (e.g., international format: +[Country Code][Area Code][Local Number])
Rule 2.2: The Address field must use consistent abbreviations (e.g., "St." for "Street", "Ave" for "Avenue")
Rule 2.3: The Postal Code field must match the expected format for the respective country (e.g., 5 digits for the US, 6 alphanumeric characters for Canada)
Rule 2.4: The Country field must use a standardized naming convention (e.g., "United States" instead of "USA" or "U.S.A.")

3. Validity Rules

Rule 3.1: The Phone Number field must contain only numeric characters (except for the '+' sign for country codes) and be of a valid length for the respective country
Rule 3.2: The Postal Code field must be valid for the specified country (e.g., validate against a list of known postal codes for that country)
Rule 3.3: The Address field must be verifiable against a known address database or geolocation service
Rule 3.4: The Country field must match a valid country name or code from a predefined list of countries

Reporting



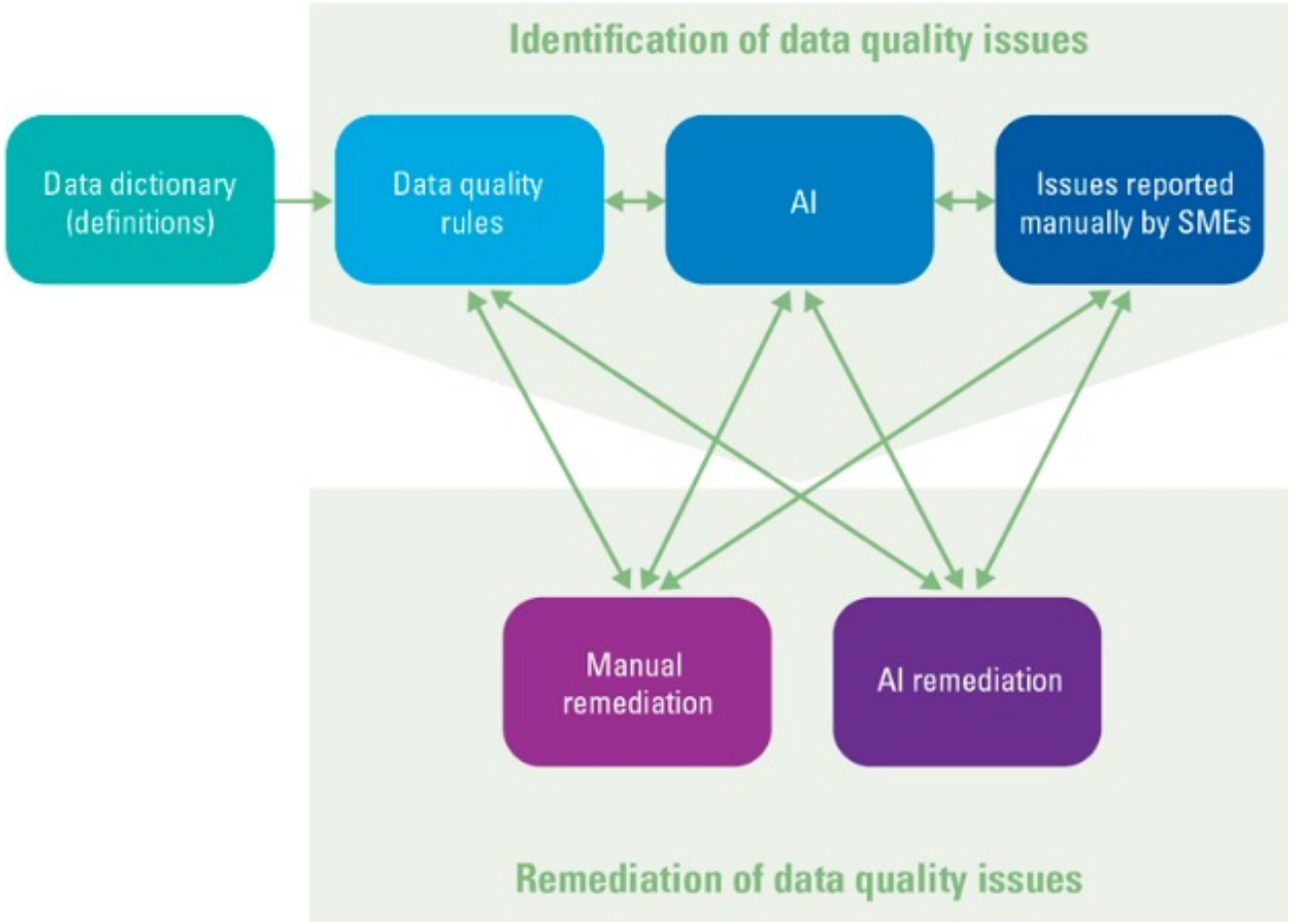


List of DQTs and their Applications

Name	Application	Use Case
Informatica Data Quality	Data profiling, cleansing, matching, and monitoring	• Customer data cleansing • Master Data Management (MDM) • Regulatory compliance (e.g., GDPR)
Talend Data Quality	Open-source data integration and quality management	• Data migration • Real-time data validation • Data enrichment
IBM InfoSphere QualityStage	Data cleansing, standardization, and matching	• Fraud detection • Customer 360 views • Data warehousing
Ataccama ONE	Unified data management platform with data quality capabilities	• Data governance • Customer data integration • Data cataloguing
Data Ladder	Data matching, deduplication, and cleansing	• Marketing data cleansing • Supply chain data standardization • Data migration
Precisely Trillium DQ	Data quality and integrity management	• Address validation • Customer data enrichment • Data governance
Experian Pandora	Data quality monitoring and management	• Financial data validation • Customer data profiling • Risk management

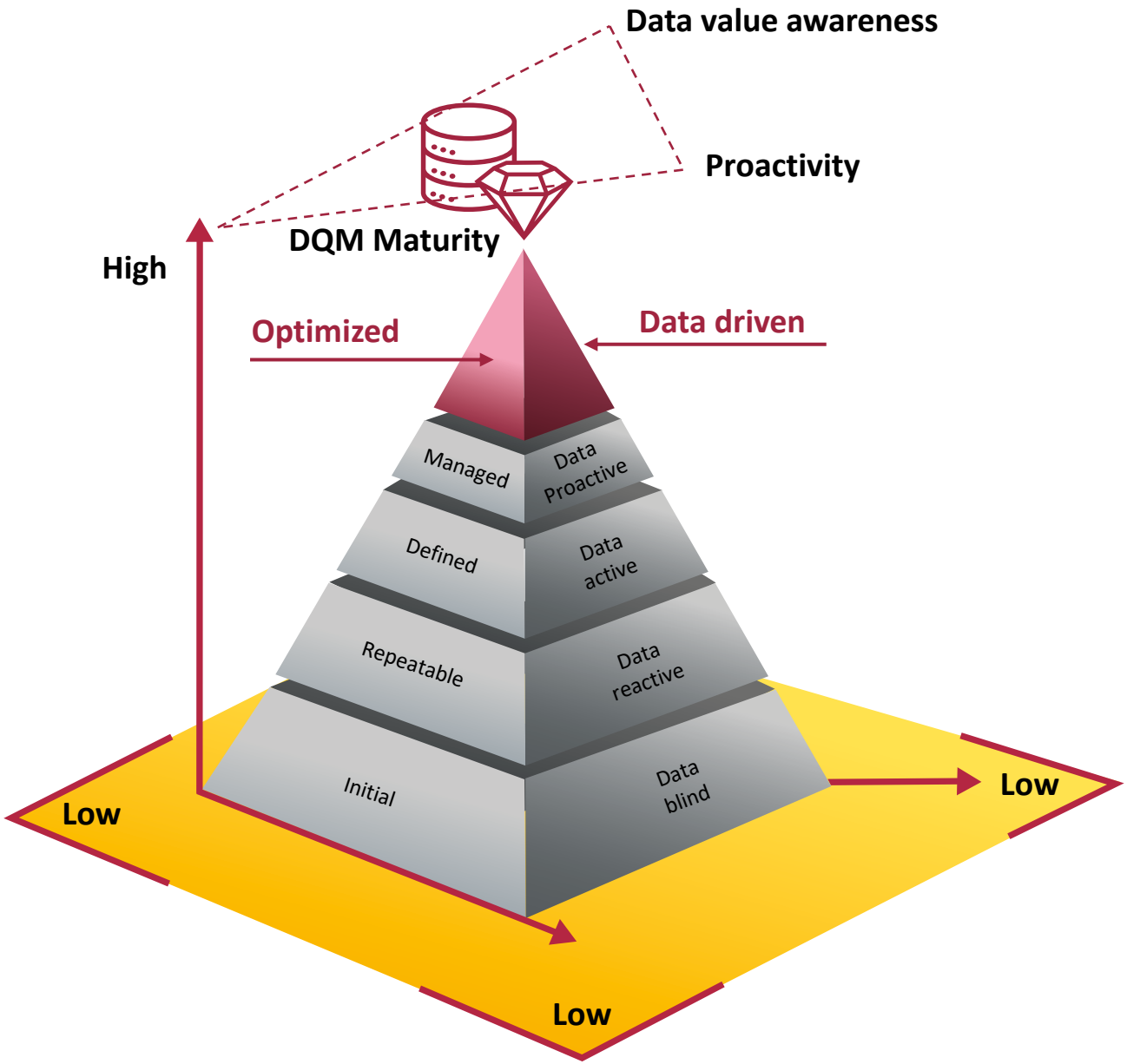
AI for DQM

Good support to identify and detect data quality issues
Less for correction…



Source: www.compact.nl/articles/ai-for-data-quality-management/

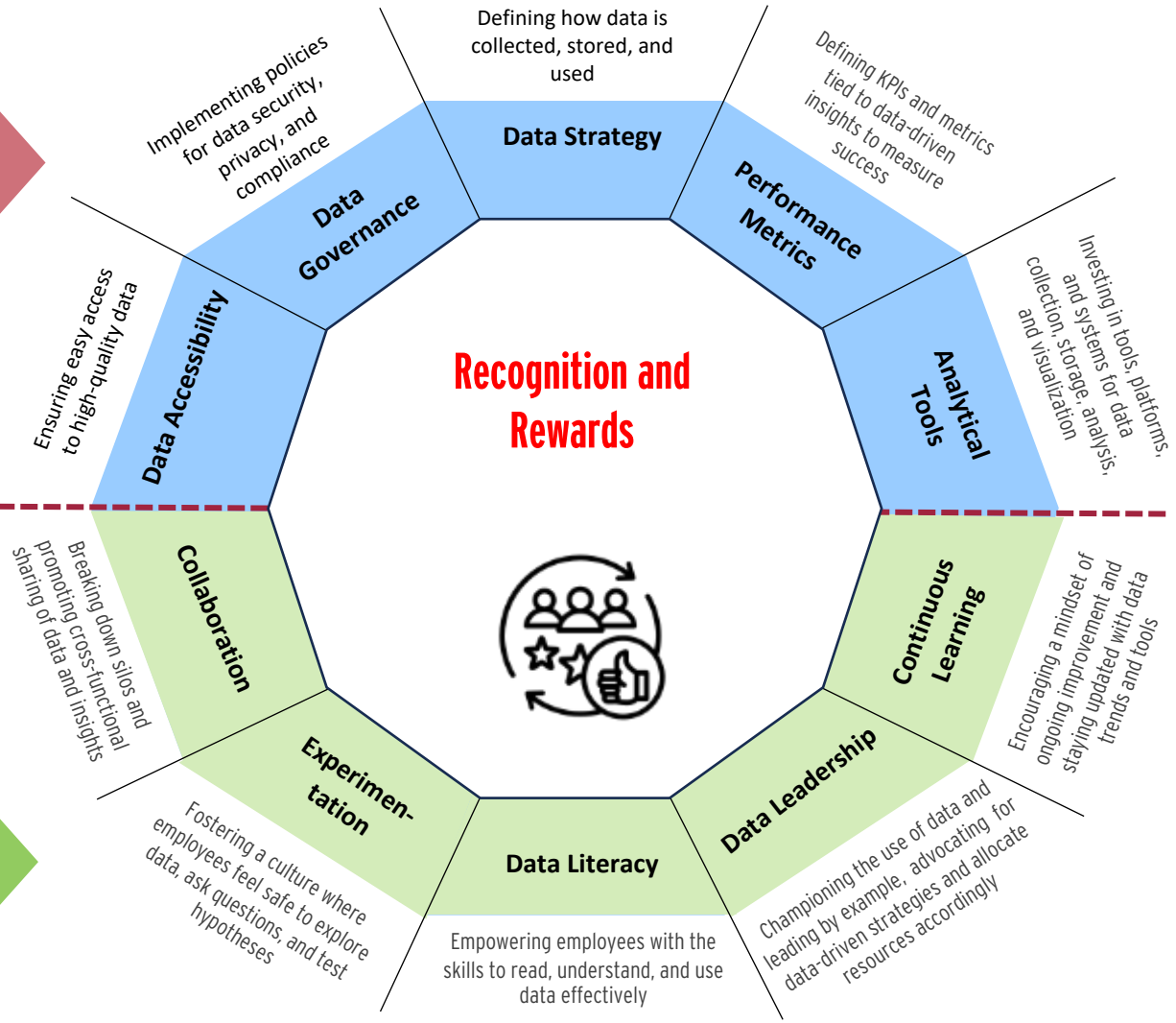
Data Excellence



Data-driven Culture

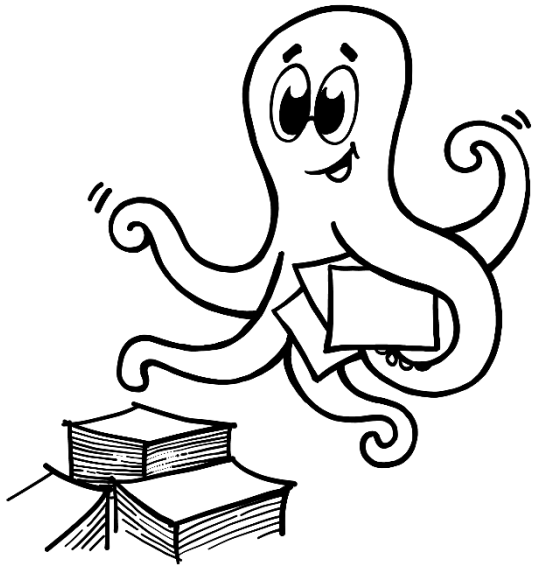
Management and Infrastructure (Hard Facts)
Tangible, operational, and technical elements that form the foundation of a data-driven culture

Leadership and Culture (Soft Facts)
Intangible, people-focused, and cultural elements that drive the adoption and sustainability of a data-driven culture



Take away and must know!

- The benefits for organizations of good data
- The DQ core dimensions
- The challenges to achieve good data quality
- The key components of a DQM
- Some best practices for an effective DQM



Literature

- DAMA International (2009) The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide). Technics Publications, LLC, Bradley Beach, NJ 07720 U.S.A.
- Mahanti (2019) Data Quality Dimensions, Measurement, Strategy, Management, and Governance. ASQ Quality Press Milwaukee,Wisconsin
- Wang R. Y., Mostapha Ziad M., Yang W. Lee Y. W. (2001) Data Quality. Springer

KNOWLEDGE

www. know-ledge.ch