MODULE 2: DATA ANALYSIS, DATA SCIENCE, AND BUSINESS INTELLIGENCE

CT ACADEMY | DATA ACTION LAB

5. DATA QUALITY

DATA ANALYSIS, DATA SCIENCE, AND BUSINESS INTELLIGENCE

DRIVERS FOR DATA QUALITY

GoC data governance policies/directives

Departmental data policies

Departmental data directives

Facilitates move towards GoC Open Data

But, it's mostly just **good business practice**!

- Ability to utilize data more efficiently and effectively
- More timely access to data
- Increased confidence in decision making
- Preparation for advanced analytics such as AI and Machine Learning

“Improving data quality leads to improved decision-making throughout the enterprise. The more high-quality data you have, the more confidence you will have in your decision-making.”

WHAT IS DATA QUALITY?

Data quality (DQ) is the degree to which data **meets** or **exceeds** business requirements (i.e., the extent to which data is “fit for purpose”).

It involves the **planning** and **implementation** of quality management techniques to **measure**, **assess**, and **improve** the fitness of data for use within an organization.

EXERCISE

In groups, list issues you've encountered related to poor data quality.

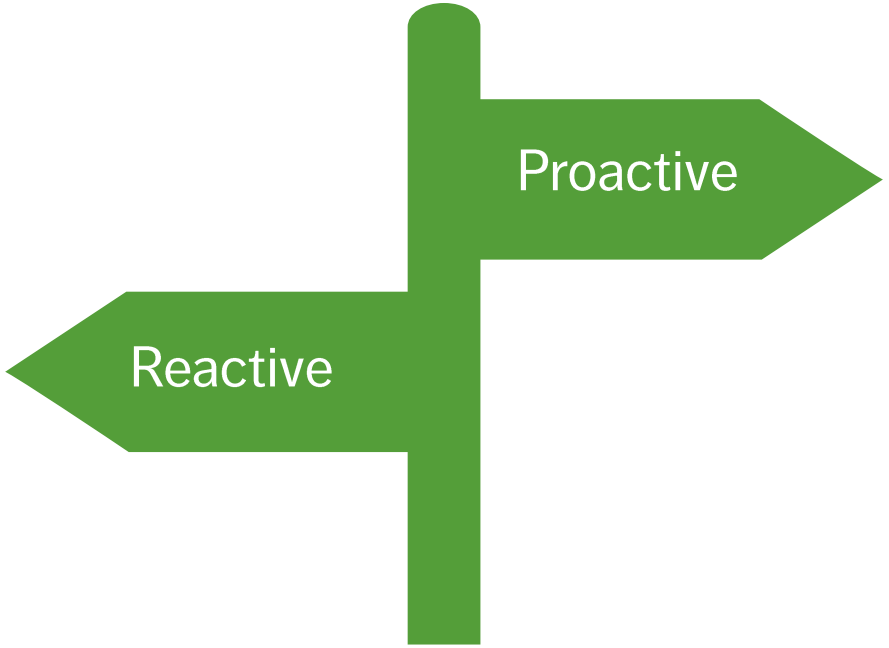
WHY IS DATA QUALITY IMPORTANT?

Data has **intrinsic value** due to its **information content**. The impacts of poor-quality data can include:

- Bad decisions, wasted resources, lowered performance chasing after issues.
- Escalating costs to remediate if issues are not caught early.
- Lack of trust in data for decision-making.
- Outdated and meaningless data bloat.
- Challenges with data sharing, integration and access to historical data.
- Poor-quality data is detrimental to analytics.

Data is the currency of **control:** we can't control what we don't know.

DATA QUALITY IS REACTIVE AND PROACTIVE



Quality Assurance

Preventing data quality issues ***before*** they impact operations and programs

Quality Control

Identifying and fixing data quality issues ***after*** they occur

A data quality program needs elements of both **control** and **assurance** to be fully effective.

DATA QUALITY DIMENSIONS

To implement data quality, we typically introduce categories called "**dimensions**".

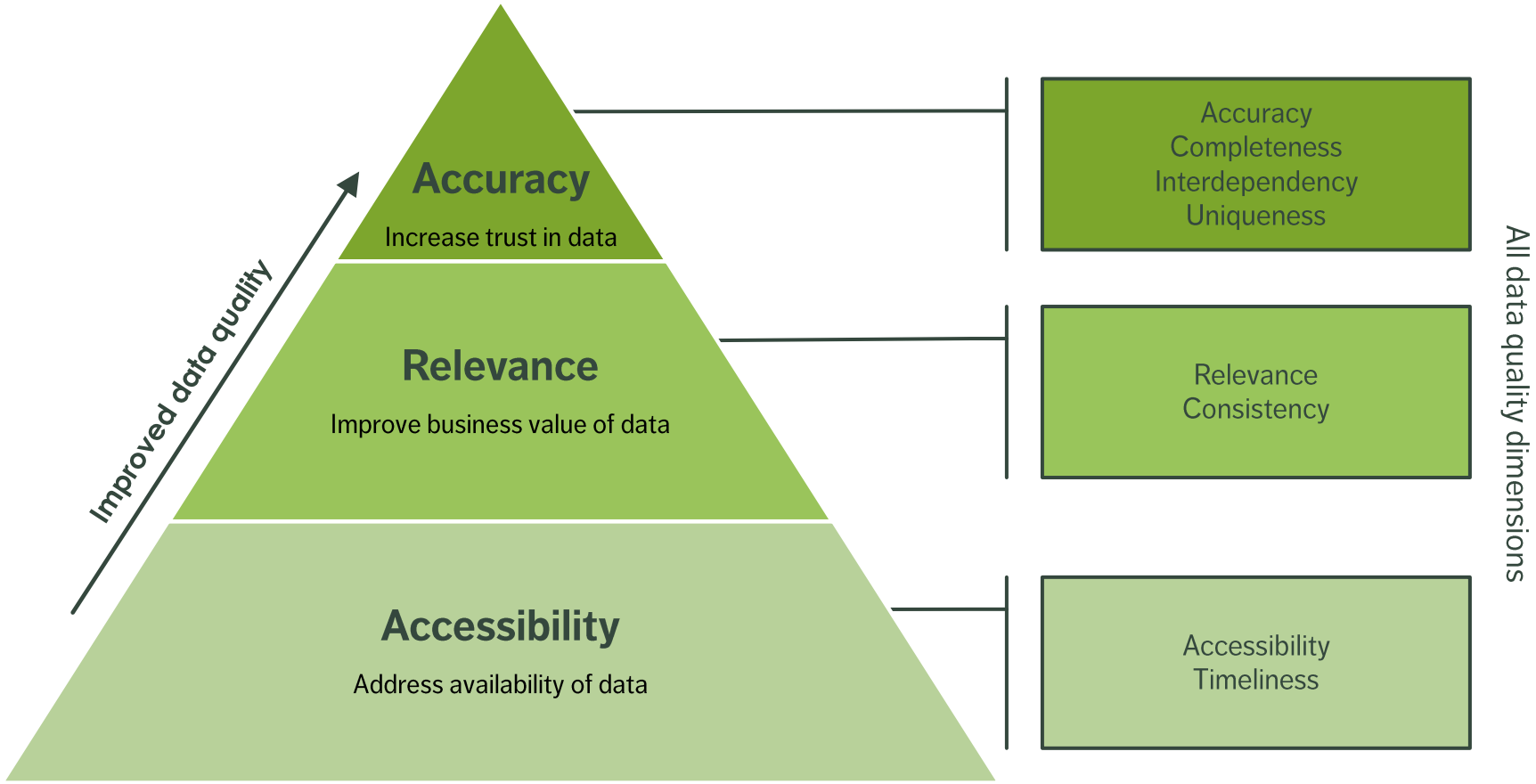
These dimensions are **measurement attributes** of data, which can be individually **assessed**, **interpreted**, and **improved**.

We can use them to help build **data quality dashboards**.

They can also help us group **issues** and **risks**, helping us:

- perform **trend analyses**;
- identify underlying, **systemic issues**;
- set-up **data quality testing programs**, etc.

DATA QUALITY DIMENSIONS

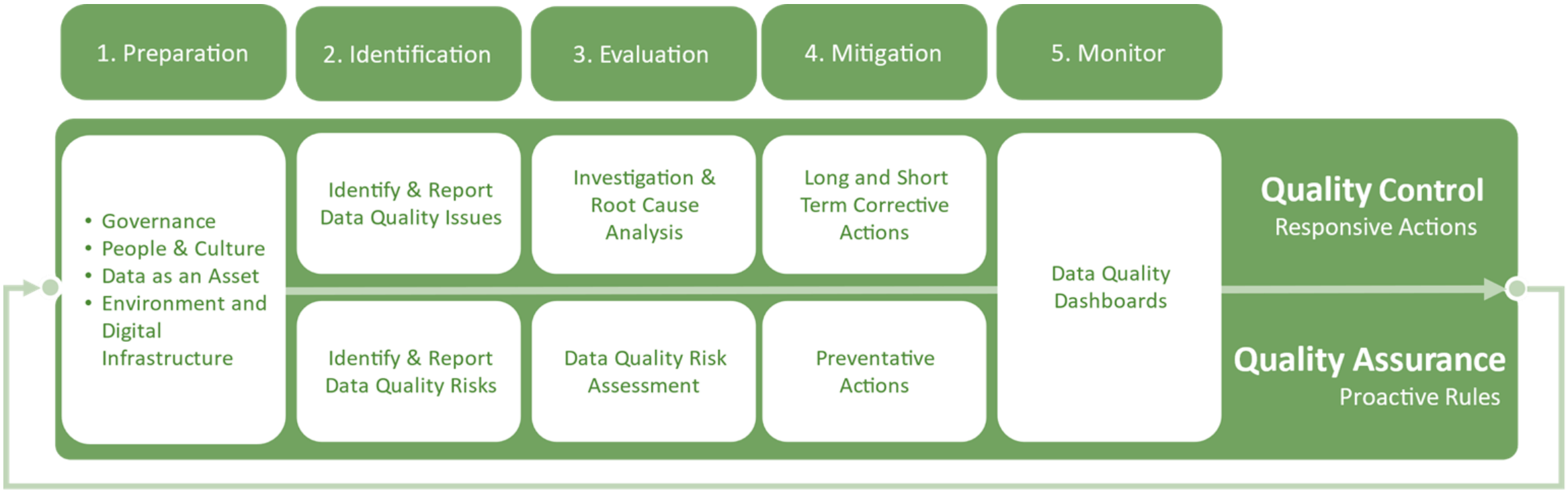


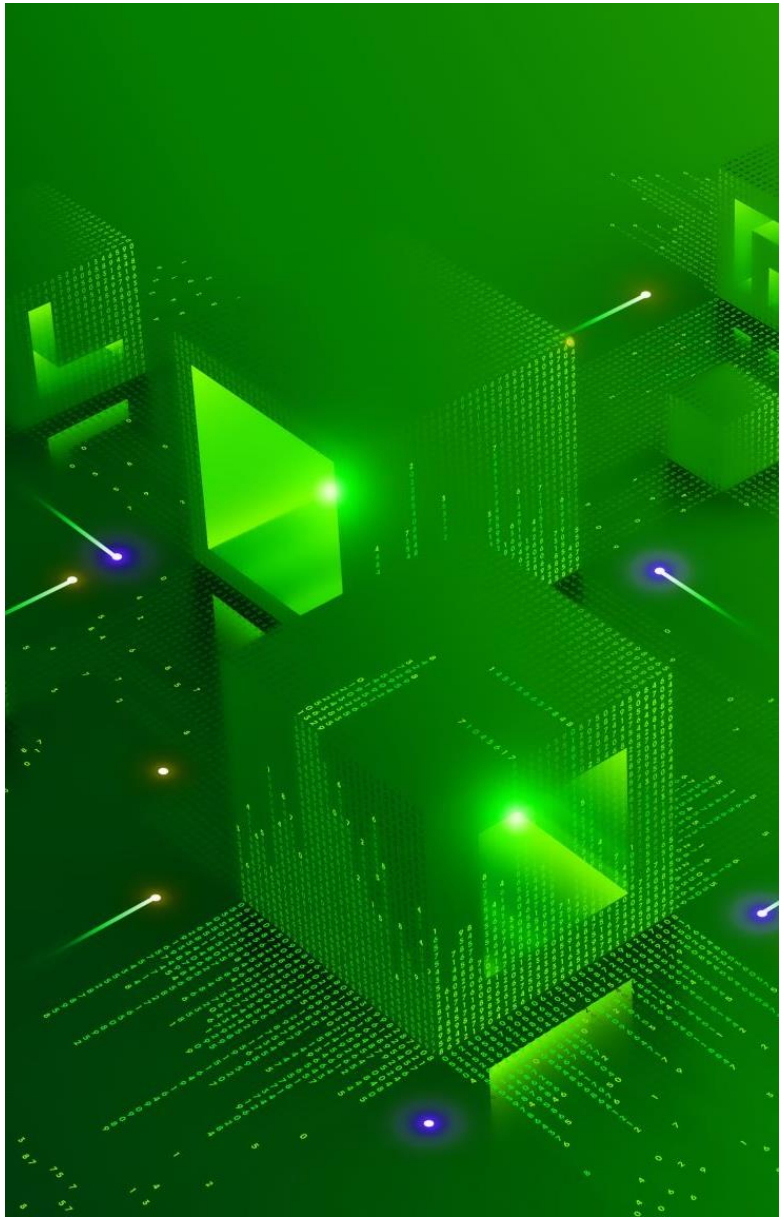
DATA QUALITY IS A PROCESS

There are three focus areas in addressing data quality:

1. identify and mitigate **existing** data quality issues through quality control (e.g., testing a database to identify incorrect values then replacing them);
2. identify sources of high risk that could **create** quality issues and mitigate those risks through quality assurance (e.g., replacing a free form text field on a new software app with a dropdown list), and
3. track and **monitor** all known data quality issues and report them on a regular basis through quality monitoring (e.g., creating a list of all known issues and monitoring when they get fixed).

DATA QUALITY IS A PROCESS





STAGE 1: PREPARATION

We can improve data quality by implementing programs such as:

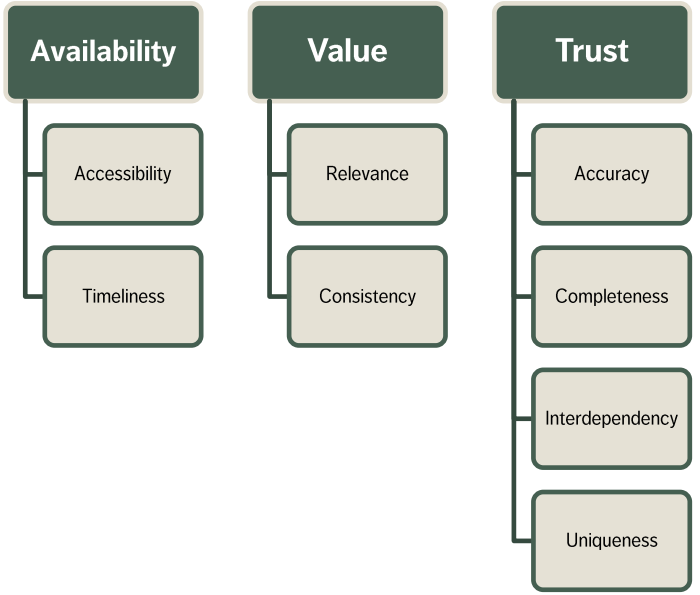
- **People & Culture** (data literacy, culture, and communication)
- **Environment & Digital Infrastructure** (tools, data asset catalogue)
- **Data Management** (metadata, reference/master data, dimensions & rules)
- **Governance** (roles & responsibilities, DQ planning, process definition)

Although it isn't critical to have all the above activities in place before starting on Data Quality, they do make a significant impact on the effectiveness of all DQ activities.

STEP 2: IDENTIFICATION

The second step in the process is to identify **data quality issues**. Currently-existing issues are called **data quality non-conformances**; issues that are yet to appear are known as **data quality risks**.

- DQ dimensions identify data attributes that can be used to measure data quality (often defined in an organization's **Data Quality Framework**).
- Business rules define the **business requirements** for data and how **data quality tests** are performed.
- Data quality metrics track the results of these tests over time.



STAGE 2: IDENTIFICATION METHODS

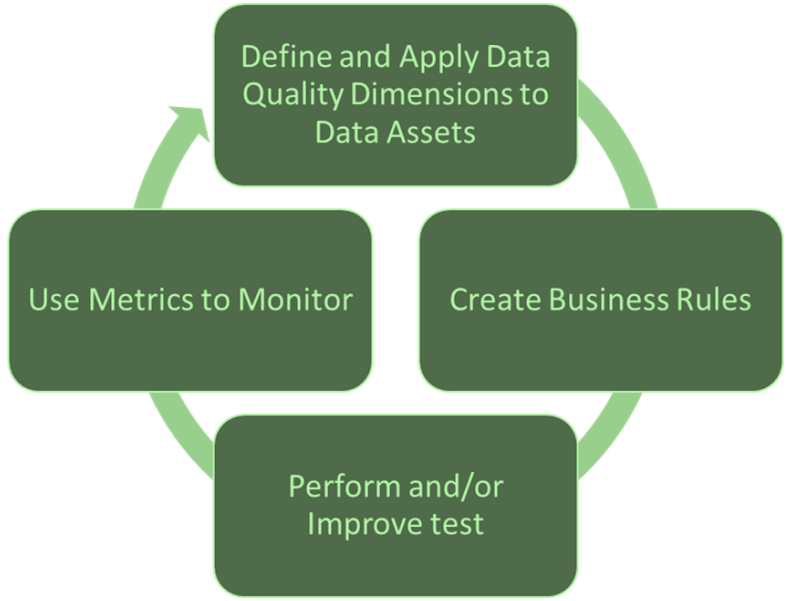
Methods for identifying **DQ non-conformances** and **DQ risks** include:

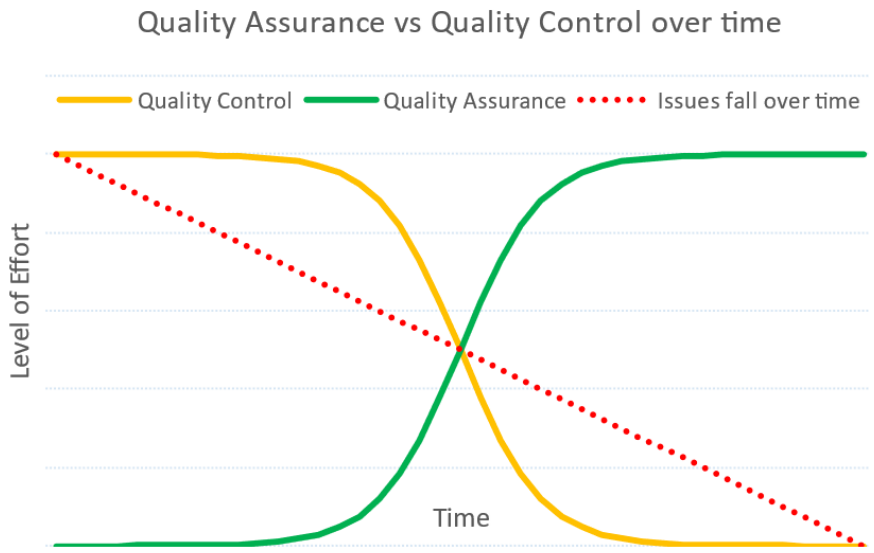
Quality Control

- DQ testing using software
- systems, process, and procedure auditing
- data consumer feedback

Quality Assurance

- creation of risk register

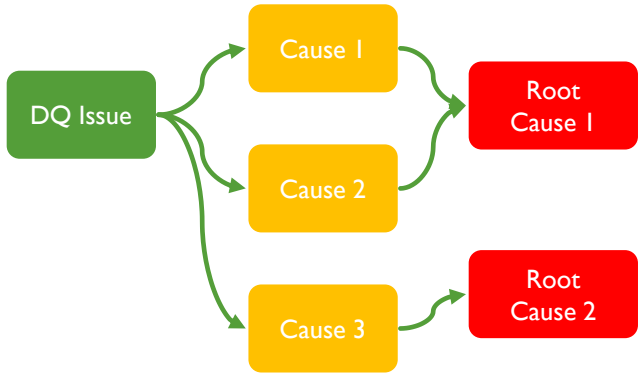


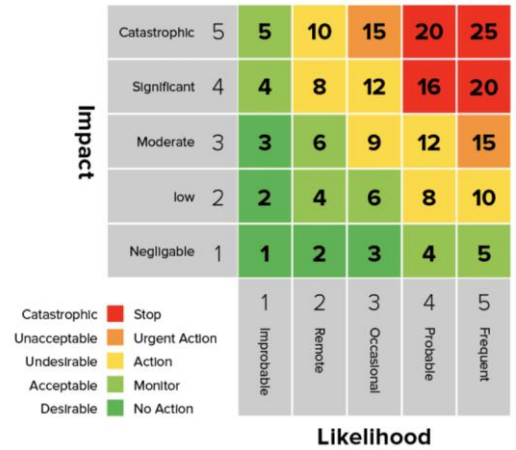


STAGE 2: IDENTIFICATION EXAMPLE

We perform data quality tests by applying business rules and dimensions, for example:

1. HR has identified that an employee surname is a critical pieces of data.
2. We use **completeness** as a dimension (missing values for this field are **important**).
3. The **business rule** that we define is the "**surname**" column in the corresponding data table should be 100% complete (no missing values).
4. The corresponding **metric** is implemented in Power BI; it counts the total number of rows in the column and the total number of non-missing entries. We then divide the entries by the total rows to calculate the **percentage of completion**.
5. The table fails the **DQ test** as only 97.2% of the records have a surname value.
6. This is **reported** to the right group, and we move to the next DQ process phase.





STAGE 3: EVALUATION

Once a DQ issue has been identified, we must **evaluate** it to prioritize **mitigation activities**. Evaluation methods typically include:

Quality Control

- formally investigate DQ Issues
- perform a **root cause analysis** to evaluate causes not symptoms

Quality Assurance

- evaluate and prioritize risks regarding impact

STAGE 3: EVALUATION EXAMPLE

1. From the previous example, the failed DQ test for **completeness** of the "surname" column has a metric value of 97.2%.
2. Next, the appropriate line of business **evaluates** the situation to find how critical it is to fix the problem.
3. As the database in question is related to pay, the 2.8% of missing values is seen as a **HIGH priority**, to be fixed as soon as practicable.
4. An investigation and **root cause analysis** is then carried out to find out what happed and what were the root causes of the issue.
5. It was determined that the root cause was an automated process used to copy surnames from another database, which failed because of a software update; we can now move to the next DQ process phase.

STAGE 4: MITIGATION

Once a DQ issue has been evaluated it may then require **mitigation** (fixing the issue). The actual fix will vary depending on the issue itself, but they fall into different categories.

Quality Control

- short term corrective actions to immediately fix the issue
- long term corrective actions ensure the issue does not recur

Quality Assurance

- preventative actions ensure that high risks do not turn into DQ issues

STAGE 4: MITIGATION EXAMPLE

From the previous example the following mitigation activities may be carried out.

1. A **short-term corrective action** is to re-run the data transfer process manually to ensure that the column is updated and 100% complete.
2. A **long-term corrective action** is to implement an automated DQ test once the transfer is complete to ensure that all records had been transferred between data assets.
3. We can now move to the next DQ process phase.

STAGE 5: MONITOR

It is best practice to **monitor** DQ on an **ongoing** basis.

Quality Control & Quality Assurance

- status of DQ issues (non-conformances and risks)
- status of DQ investigations and root cause analysis
- status of internal quality audits
- status of corrective actions and preventative actions

STAGE 5: MONITOR EXAMPLE

As an issue had occurred in the “surname” column, the following items were recorded and included in the **ongoing monitoring** of DQ:

1. the date the issue was identified;
2. the criticality of the issue (high priority);
3. the type of issue (“completeness” non-conformance);
4. the actual test result (97.2%);
5. the date the short-term corrective action was completed;
6. the date the long-term corrective action was completed, and
7. a measure of the effectiveness of the long-term corrective action (“highly effective”).

As the line of business continues to monitor non-conformances and risks, a profile of the health of the data asset is created which shows **improvement over time** (?)

This is replicated for **all onboarded assets** for aggregated monitoring & reporting.

"Any substantial improvement must come from action on the system and is the responsibility of management. Wishing and pleading and begging the workers to do better is totally futile."

(W. Edwards Deming, *Out of the Crises*)

EXERCISE

For the data quality issues identified in the previous exercise, identify or create:

1. the root cause(s);
2. short term corrective action(s), and
3. long term corrective action(s).

Use the 5 stages to inform your answers.

SUPPLEMENTAL MATERIAL

5. DATA QUALITY

DATA QUALITY DIMENSIONS

Typical set of data quality dimensions

Accessibility	Business processes and consumers can access and use the data asset
Timeliness	Data values are sufficiently up-to-date for business processes and consumers needs
Relevance	Data asset is of value to and used by business processes and data consumers
Consistency	Data representations are the same within and across data assets and repositories
Accuracy	Data accurately represents a real-world entity, object, concept, etc.
Completeness	Data asset has no missing data values
Interdependency	Relationships between data elements are preserved within or across data assets
Uniqueness	Data representations are not duplicated within or across data assets

DATA QUALITY IS ENACTED BY PEOPLE

Role	Description	Focus on Data Quality
Chief Data Officer	Responsible for all data related activities at department	Accountable for DQ program
Branch/Regional/ Program Heads	Responsible for Branch/Regional/Programs	Participation in the DQ program
Data Trustee	Strategically manages data assets and ensures compliance with data-related strategies, regulations, policies, directives	Creates a proactive, risk-based approach to the reduction of DQ issues
Data Steward	Advises, enacts, and enforces data policies and standards	Performs testing, risk assessment, investigations and auditing. Implements ongoing monitoring of DQ
Data Custodian	Ensures safe custody and integrity of hosted data	Provides technical support for corrective and preventative actions
Data Contributor	Ensures the data they provide aligns with technical and business policies, procedures, and standards	Ensures quality of data prior to inclusion in data asset
Data Consumer	Ensures usage of data supports all objectives and mandates	Reports on data issues found when using data