iykra Data Fellowship

IYKRA
Data Fellowship Program

Data Quality Techniques and Anomaly Detection

by Fariz Wakan

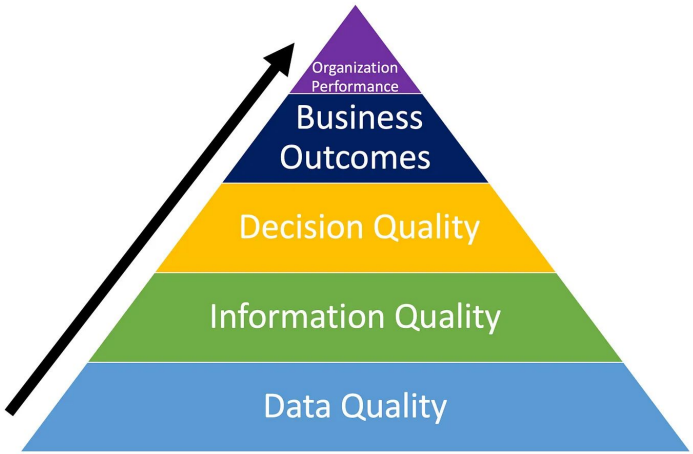
April 19th, 2024

Data Fellowship

Intro to Data Quality

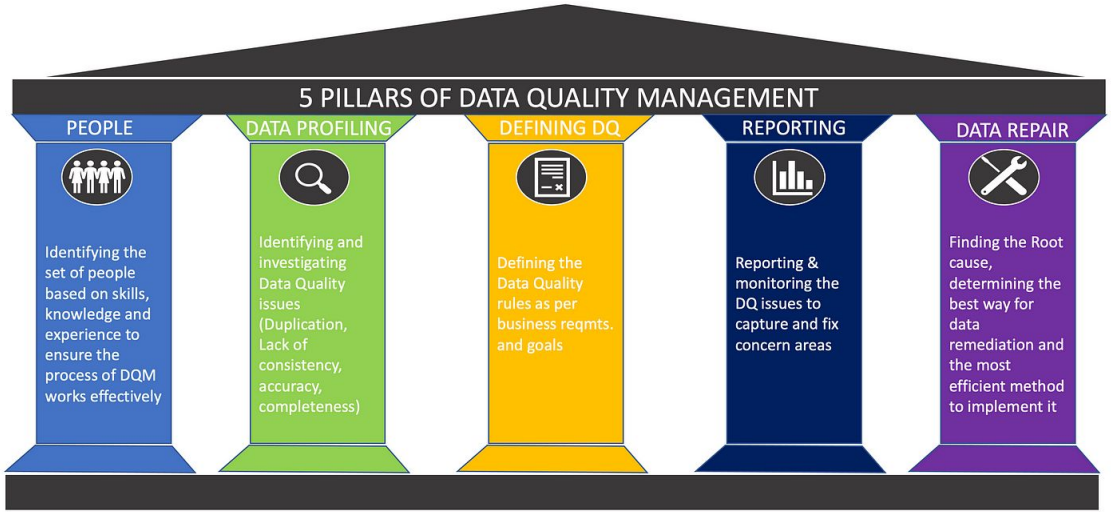
Data Quality

Data quality is defined by how well a given dataset meets a user's needs. Data quality is an important criteria for ensuring that data-driven decisions are made as accurately as possible.



Data Quality Management

Set of practices that aim at improving and maintaining a high quality of information within the organization



Impact of poor data quality

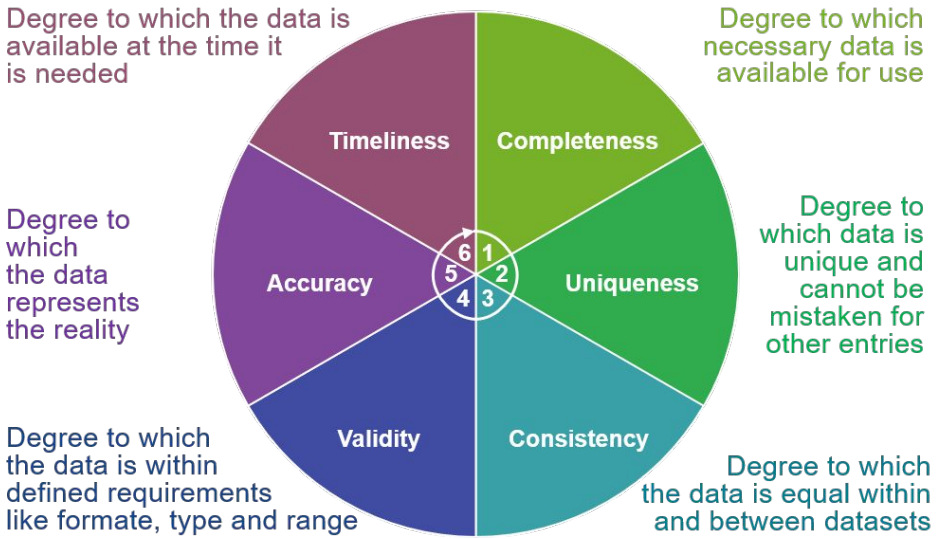
Cost of poor data quality

- According to IBM's estimate, the US **lost $3.1 trillion yearly** due to bad data.
- Gartner.com suggests that organizations **lose between $10 to $14 Million USD annually** due to poor data.
- Integrate reported that around 40% of all leads have inaccurate data.
- Cio.com identified that around 80% of companies believe they lost revenue due to data challenges.
- MIT Sloan reported that employees spend half of their time coping with managing data quality tasks.
- Pragmaticworks states **20 to 30 percent of operating expenses** are due to bad data.
- Econsultancy.com reported that due to poor data, companies having mail delivery issues **lost about 30% of their revenue**, in addition to the **21% of businesses experienced reputation damages**.
- Gartner also reported that data scientists **spend around 80% of their time cleaning and organizing data**.

Why do we have bad data?

- Intuition is more important to management than data
- Manual data entry errors
- Data Silos
- Data migration and conversion projects
- Scaling of the business and its datasets
- No Data Governance rules

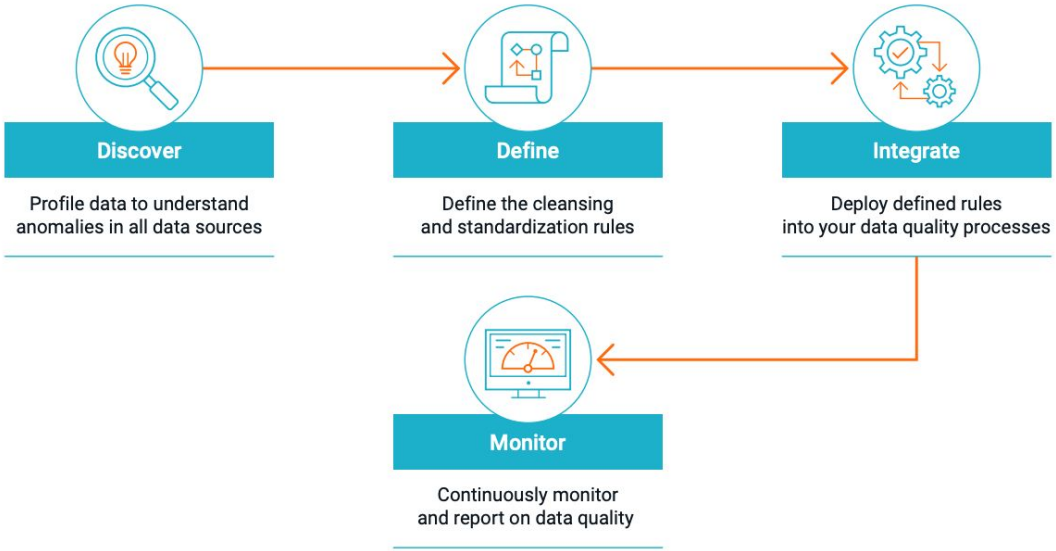
Data Quality Dimensions



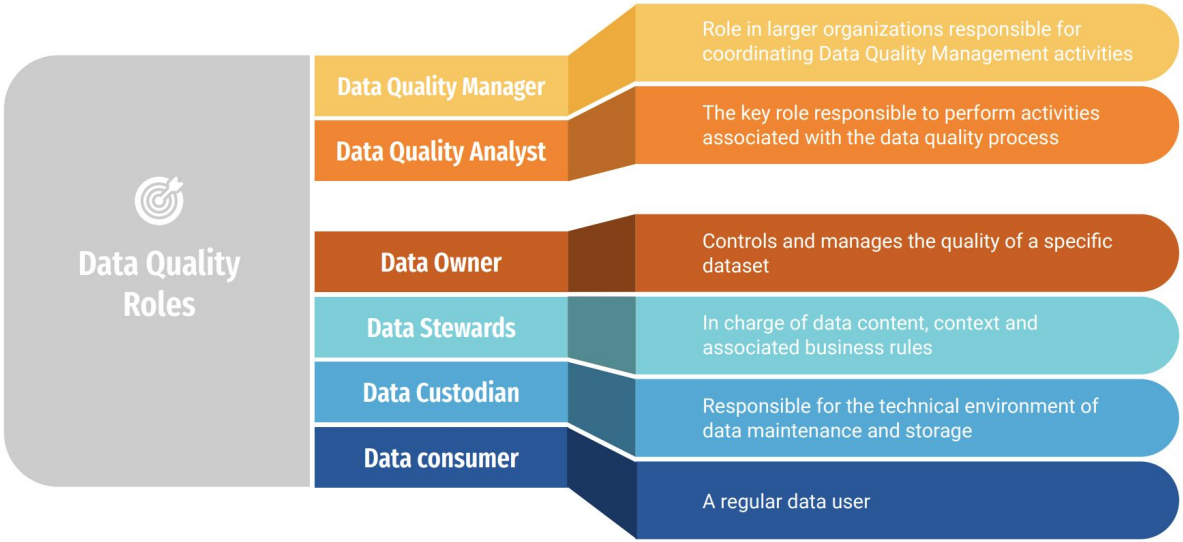
Data Quality Rules

- Also known as data validation rules
- Business rules for data
- Automated data quality checks
- Primary tool for determining data quality

Defining your Data Quality Rules



Data Quality Roles



Data Fellowship

Data Quality Techniques

Data Profiling

Data Profiling is a technique for discovering and investigating data quality issues, such as duplication, lack of consistency, and lack of accuracy and completeness.

- Data Profiling involves analyzing one or multiple data sources and collecting metadata
- The data steward uses the results to investigate the origin of the data errors
- Previously done manually, now using Data Quality Tools
- The tools provide data statistics, such as degree of duplication, etc

Data Parsing

Data Parsing is the process of separating complex data entries into separate fields. It can also mean converting data into a different data format.

Example - converting full name or address into separate fields
Example - converting HTML into plain text

Data Parsing Benefits

Extract pieces of data to validate if it follows a specific pattern!

Use Cases for Data Parsing:

- Check phone number
- Check mail addresses
- Check names
- Check addresses

Data Standardization

Data Standardization is the process of converting data to a common format

Identity Resolution

Identity Resolution is a process that checks, validates and appends information across datasets to create a single, data-rich profile for a person, business or other entry

Identity Resolution Process

1. **Identify** - channels, platforms and devices
2. **Connect** - connect the dots between the different channels, platforms, devices
3. **Match** - based on a defined set of attributes (same household, IP, wi-fi network, timing patterns, etc)
4. **Validate** - validate that it is the same identity
5. **Activate** - create a single, data-rich profile

Identity Resolution Benefits

- Single Customer View
- Personalized Customer Experience
- Happy Employees
- Understand customer’s network
- Focused view for each function
- Contextual Marketing
- Governance

Data Linkage

Data linkage, also known as record linkage, is the process of identifying, matching and merging records that correspond to the same person from several datasets or even within one dataset

Data Cleansing

Data cleansing is the process of resolving corrupt, inaccurate, incomplete or irrelevant data

Data Enhancement

Data enhancement is a data improvement process that adds information from third-party or internal datasets to increase the value that the organization derives from the data

Key Points:

- Data enhancements builds on parsing, standardization and record linkage
- Sometimes called data enrichment

Data Inspection and Monitoring

The process of Data Inspection and Monitoring:

- Data Profiling exposes potential business rules
- Data Quality Analyst documents the rules and confirms their criticality
- The rules describe the end-user data expectations
- The rules are used to measure and monitor the data
- Monitoring of the rules provide a proactive assessment of compliance
- The results are used to populate Data Quality Metric scorecard for Leadership

References

- Data Quality Management by George Smarts
- Data Quality by Yatin Jaisingh
- https://dataladder.com/the-impact-of-poor-data-quality-risks-challenges-and-solutions/
- https://www.lean-data.nl/data-quality/

iykra Data Fellowship

Thank you!