Data Engineering

204426

Data Quality and Data Auditing

Outline

- Data Quality
  - Completeness
  - Uniqueness
  - Consistency
  - Validity
  - Accuracy
  - Timeliness
- Data Auditing
  - Quality Assessment Process
  - Data Quality – A Simple 6 Step Process
  - Data Quality Assessment

Data Quality

Data Quality

The challenges of data quality

- The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.
- Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.
- Data change very fast and the “timeliness” of data is very short, which necessitates higher requirements for processing technology.
- No unified and approved data quality standards.
  - ISO 8000 data quality standards

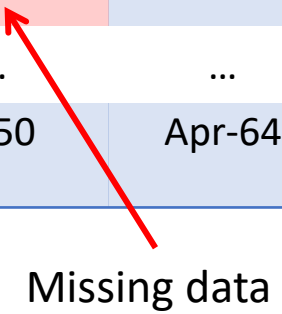
Data Quality

Completeness

- The values of all components of a single datum are valid.
- For example, for image color, RGB can be used to describe red, green, and blue, and RGB represents all parts of the color data. If the color value of a certain component is missing, the image cannot show the real color and its completeness is destroyed
- The customer address includes an *optional* landmark attribute, data can be considered complete even when the landmark information is missing.

Data Quality

name	id	align	eye	hair	gender	alive	appearances	first_appear	publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Black Hair	Male	Living Characters	NA	Aug-62	marvel
…	…	…	…	…	…	…	…	…	…
Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel



Data Quality

Uniqueness

- No duplication or overlaps.
- Data uniqueness also improves data governance and speeds up compliance.

Data Quality

Duplicate data

name	id	align	eye	hair	gender	alive	appearances	first_appear	publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Black Hair	Male	Living Characters	NA	Aug-62	marvel
…	…	…	…	…	…	…	…	…	…
Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

Data Quality

Consistency

- Refers to whether the *logical relationship* between correlated data is correct and complete.
- The same data that are located in different storage areas should be considered to be equivalent.
- Equivalency means that the data have equal value and the same meaning or are essentially the same.

Data Quality

1 – Living Characters
0 – Deceased Characters

name	id	align	eye	hair	gender	alive	appearances	first_appear	publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	1	4043	Aug-62	marvel
Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
…	…	…	…	…	…	…	…	…	…
Natalia Romanova (Earth-616)	Public	1	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

1 – Good
0 – Bad

Data Quality

Validity

- The value attributes are available for aligning with the specific domain or requirement.
- For example, ZIP codes are valid if they contain the correct characters for the region.
- In a calendar, months are valid if they match the standard global names.
- Using business rules is a systematic approach to assess the validity of data.

Data Quality

month	started_at	ended_at	DURATION	start_location_name	end_location_name
May	5/21/2019 18:33	5/21/2019 18:40	0:17:03	1901 Roma Ave NE, Albuquerque, NM 87106, USA	1899 Roma Ave NE, Albuquerque, NM 87106, USA
May	5/21/2019 19:07	5/21/2019 19:12	0:04:57	1 Domenici Center en Domenici Center, Albuquerque, NM 87106, USA	1111 Stanford Dr NE, Albuquerque, NM 87106, USA
May	5/21/2019 19:13	5/21/2019 19:15	0:01:14	1 Domenici Center en Domenici Center, Albuquerque, NM 87106, USA	1 Domenici Center en Domenici Center, Albuquerque, NM 87106, USA
…	…	…	…	…	…
July	7/21/2019 23:55	7/22/2019 1:46	1:51:00	105 Stanford Dr SE, Albuquerque, NM 87106, USA	3339 Central Ave NE, Albuquerque, NM 87106, USA

Data Quality

Accuracy

- Accuracy can be easily measured, such as gender, which has only two definite values: male and female.
- But in other cases, there is no known reference value, making it difficult to measure accuracy.
- Because accuracy is correlated with context to some extent, data accuracy should be decided by the application situation.

Data Quality

Timeliness

- The time delay from data generation and acquisition to utilization.
- Data should be available within this delay to allow for meaningful analysis. In the age of big data, data content changes quickly so timeliness is very important.

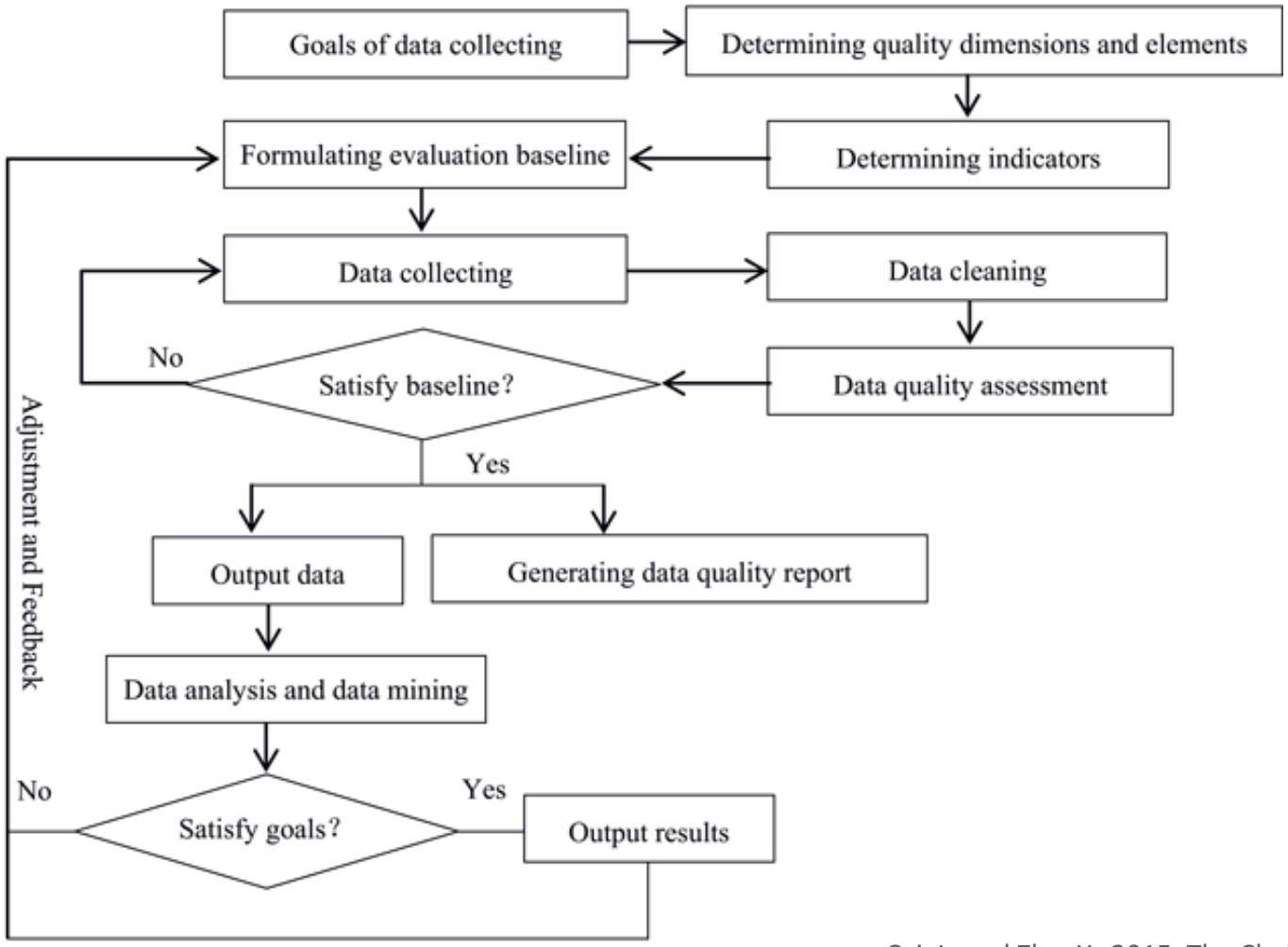
Data Auditing

The auditing of data to assess its quality or utility for a specific purpose.

3 Key Data Auditing Functions:

- **Data Quality**: Identifies inaccurate data and root causes—allowing organizations to implement processes to remediate issues.
- **Regulatory Compliance**: Helps organizations adhere to corporate, industry, and government regulations by providing deep visibility into the location, usage, and data security.
- **Improve Operations**: From sales and marketing to customer service and human resources, data auditing up-levels data quality, making operations run more smoothly and effectively.

Quality Assessment Process for Big Data



In different business environments, the selection of data quality elements will differ.

The formulation of assessment indicators also depends on the actual business environment.

Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002

Data Quality – A Simple 6 Step Process

Step 1 – Definition

- Define the business goals for Data Quality improvement, data owners / stakeholders, impacted business processes, and data rules.
- Examples for customer data:
  - Goal: Ensure all customer records are unique, accurate information (ex: address, phone numbers etc.), consistent data across multiple systems, etc.
  - Data owner: Sales Vice President
  - Stakeholders: Finance, Marketing, and Production
  - Impacted business processes: Order entry, Invoicing, Fulfillment etc.
  - Data Rules: Rule 1 – Customer name and Address together should be unique; Rule 2: All addresses should be verified against an approved address reference database etc.

Data Quality – A Simple 6 Step Process

Step 2 – Assessment

- Assess the existing data against rules specified in Definition Step.
- Assess data against multiple dimensions such as accuracy of key attributes, completeness of all required attributes, consistency of attributes across multiple data sets, timeliness of data etc.
- Depending on the volume and variety of data and the scope of Data Quality project in each enterprise, we might perform qualitative and/or quantitative assessment using some profiling tools.
- This is the stage to assess existing policies (data access, data security, adherence to specific industry standards/guidelines etc.) as well.
- Examples:
  - Assess %of customer records that are unique (with name and address together); % of non-null values in key attributes etc.

Data Quality – A Simple 6 Step Process

Step 3 – Analysis

- Analyze the assessment results on multiple fronts.
- One area to analyze is the gap between data quality business goals and current data.
- Another area to analyze is the root causes for inferior data quality (if that is the case).
- Examples:
  - If customer addresses are inaccurate by more than the business defined goal, what is the root cause? Is the order entry application data validations the problem? Or the reference address data inaccurate?

Data Quality – A Simple 6 Step Process

Step 4 – Improvement

- Design and develop improvement plans based on prior analysis.
- The plans should comprehend timeframes, resources, and costs involved.
- Examples:
  - All applications modifying addresses must validate against selected address reference database;
  - Customer name can only be modified via order entry application
  - The intended changes to systems will take 6 months to implement and requires XYZ resources and $$$.

Data Quality – A Simple 6 Step Process

Step 5 – Implementation

- Implement solutions determined in the Improve stage.
- Comprehend both technical as well as any business process related changes.
- Implement a comprehensive ‘Change Management’ plan to ensure that all stakeholders are appropriately trained.

Data Quality – A Simple 6 Step Process

Step 6 – Control

- Verify at periodic intervals that the data is consistent with the business goals and the data rules specified in the Definition Step.
- Communicate the Data Quality metrics and current status to all stakeholders on a regular basis to ensure that Data Quality discipline is maintained on an ongoing basis across the organization.
- Data Quality is not a onetime project but a continuous process and requires the entire organization to be data-driven and data-focused.

Data Quality Assessment

Functional Forms

- **Simple Ratio**
  - The ratio of desired outcomes to total outcomes.
  - **Free-of-error dimension** (represents data correctness - accuracy):

$$1 - \frac{\text{the number of data units in error}}{\text{the total number of data units}}$$

    - Clearly defined criteria
      - what constitutes a data unit
      - what is an error

Data Quality Assessment

Functional Forms

- **Simple Ratio**
  - **Completeness dimension**
    - Schema completeness, the degree to which entities and attributes are not missing
    - Column completeness: a function of the missing values in a column of a table
    - Population completeness: For example, if a column should contain at least one occurrence of all 50 states, but it only contains 43 states. ← *population incompleteness*
    - Each of the three types can be measured by

$$1 - \frac{\text{the number of incomplete items}}{\text{the total number of items}}$$

Data Quality Assessment

Functional Forms

- **Simple Ratio**
  - **Consistency dimension**
    - Can be viewed from a number of perspectives, one being consistency of the same (redundant) data values across tables.
    - A metric measuring consistency is

$$1 - \frac{\text{violations of a specific consistency type}}{\text{the total number of consistency checks}}$$

Data Quality Assessment

Functional Forms

- **Min or Max Operation**
  - The aggregation of multiple data quality indicators (variables), the minimum or maximum operation can be applied.
  - One computes the minimum (or maximum) value from among the normalized values of the individual data quality indicators.

Data Quality Assessment

Functional Forms

- **Min or Max Operation**
  - **Believability dimension**
    - It reflect an individual's assessment of the credibility of the data source, comparison to a commonly accepted standard, and previous experience.
    - Example,
      - Believability of the data source is rated as 0.6
      - Believability against a common standard is 0.8
      - Believability based on experience is 0.7
      - So, the overall believability rating is then 0.6
      - * An alternative is to compute the believability as a weighted average of the individual components.

Data Quality Assessment

Functional Forms

- **Min or Max Operation**
  - **Timeliness dimension**
    - How up-to-date the data is with respect to the task it's used for
    - The maximum of one of two terms
      - 0 and one minus the ratio of currency to volatility
        - Currency - the age plus the delivery time minus the input time
        - Volatility - the length of time data remains valid
        - Delivery time - when data is delivered to the user
        - Input time - when data is received by the system
        - Age - the age of the data when first received by the system

Data Quality Assessment

Functional Forms

- **Weighted Average**
  - For the multivariate case, an alternative to the min operator is a weighted average of variables.
  - If a company has a good understanding of the importance of each variable to the overall evaluation of a dimension, then a weighted average of the variables is appropriate.

Data Quality Assessment

Example: Measure Data Quality

Metric	Definition	How to calculate
Ratio of Data to Errors	How many errors do you have relative to the size of your data set?	Divide the total number of errors by the total number of items.
Number of Empty Values	Empty values indicate information is missing from a data set.	Count the number of fields that are empty within a data set.
Data Transformation Error Rates	How many errors arise as you convert information into a different format?	How often does data fail to convert successfully?
Amounts of Dark Data	How much information is unusable due to data quality problems?	Look at how much of your data has data quality problems.

Data Quality Assessment

Example: Measure Data Quality

Metric	Definition	How to calculate
Email Bounce Rates	What percentage of recipients didn't receive your email because it went to the wrong address?	Divide the total number of emails that bounced by the total number of emails sent, then multiply by 100.
Data Storage Costs	How much does it cost to store your data?	What is your data storage provider charging you to store information?
Data Time-to-Value	How long does it take for your firm to get value from its information?	Decide what "value" means to your firm, then measure how long it takes to achieve that value.

Christopher Tozzi (2021). How to Measure Data Quality – 7 Metrics to Assess the Quality of Your Data.
url: https://www.precisely.com/blog/data-quality/how-to-measure-data-quality-7-metrics

References

- Wang, R., & Storey, V. (1995) Framework for Analysis of Quality Research. IEEE Transactions on Knowledge and Data Engineering 1(4), pp 623–637.
- McGilvray, D. (2010) Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, Beijing: Publishing House of Electronics Industry.
- Silberschatz, A., Korth, H., & Sudarshan, S. (2006) Database System Concepts, Beijing: Higher Education Press.
- Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002
- Leo L. Pipino, Yang W. Lee, and Richard Y. Wang (2002). Data Quality Assessment. COMMUNICATIONS OF THE ACM 45(4), pp 211-218.
- Ramesh Dontha (2017). Data Quality – A Simple 6 Step Process. url: https://www.dataversity.net/data-quality-simple-6-step-process/
- Ankur Gupta (2021). The 6 dimensions of data quality. url: https://www.collibra.com/blog/the-6-dimensions-of-data-quality