Data Engineering

204426

Data Quality and Data Auditing

Outline

- Data Quality
  - Completeness
  - Uniqueness
  - Consistency
  - Validity
  - Accuracy
  - Timeliness
- Quality Assessment Process
- Data Auditing

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

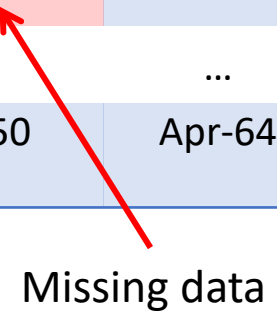
Data Quality

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

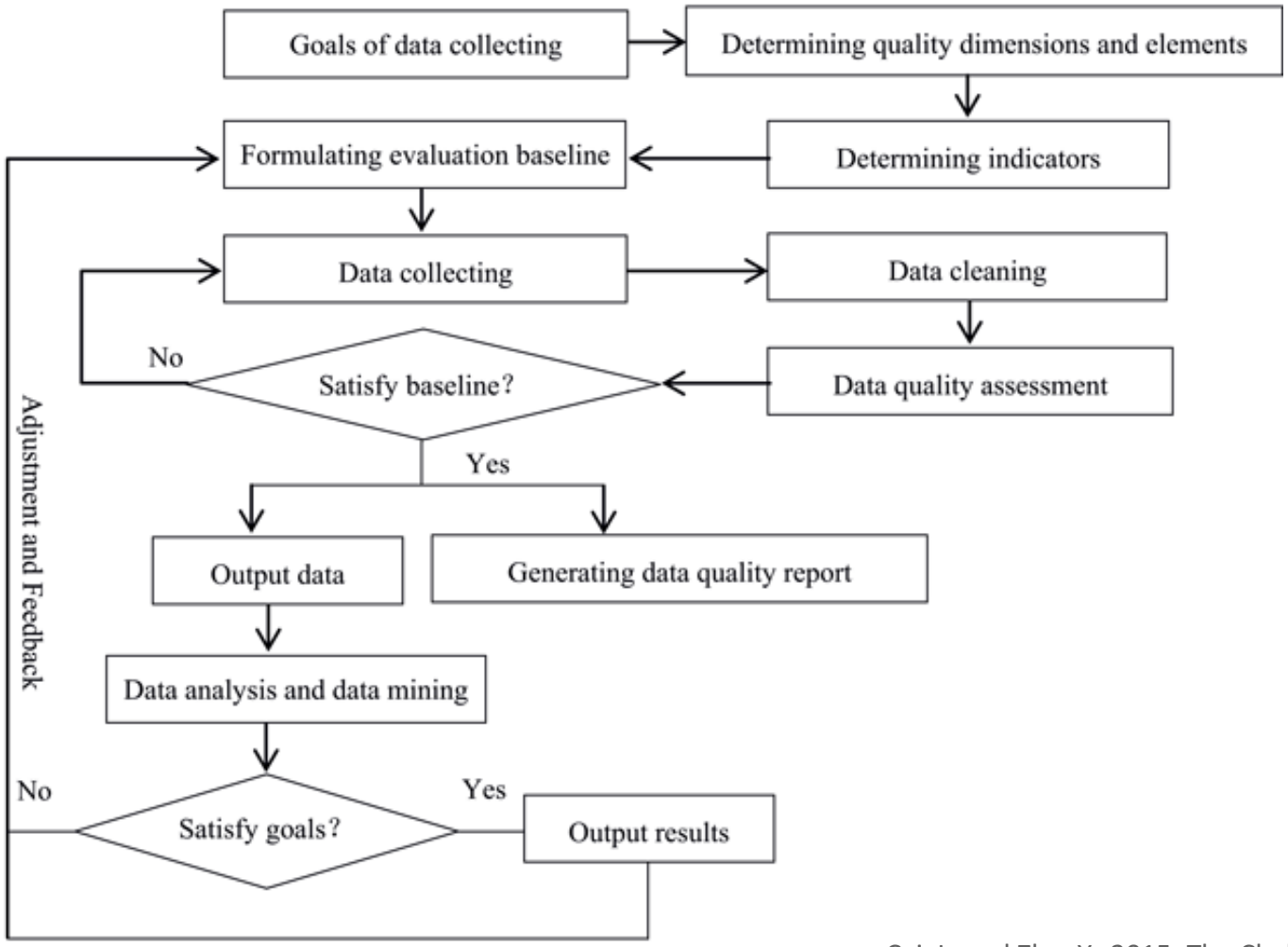
- Accuracy can be easily measured, such as gender, which has only two definite values: male and female.
- But in other cases, there is no known reference value, making it difficult to measure accuracy.
- Because accuracy is correlated with context to some extent, data accuracy should be decided by the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

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

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
- https://www.collibra.com/blog/the-6-dimensions-of-data-quality