How to calculate a Data Quality score?

Earlier this year I worked for an organisation that was looking to improve its data quality measurement capabilities. The existing data quality measurement system was really only measuring the effectiveness of their processes, not the data collected. So, for example, a "data quality" measurement was how long new data took to arrive in their central registry from the original source of the data, in comparison to the standard benchmark time. Data that took less than the benchmark was rated as good, whilst data that took longer was marked as "bad".

I worked with them to identify alternative measures that were more closely aligned to quality measures of the actual data. Whilst doing so, I found that the Netherlands branch of the Data Management Association (DAMA) had compiled a great list of data quality measures. This document defined more than 60 different measures of data quality and provided definitions and units of measurement for each of them. The different units of measurement defined by the Dutch team were:

- Percentage
- Number
- Grade (subjective assessment on a defined scale by a person)
- Boolean (yes/no, true/false)
- Duration (expressed as a measure of time)
- Story (value cannot be expressed as a number, but only as a description)

As an example, here is an example of a data quality measure using each of the units of measure:

- Percentage : Attribute-level Completeness = "the degree to which all the required attributes in the dataset are present"
- Number : Data Redundancy = "the degree to which logically identical data are stored more than once"
- Grade : Reputation = "The degree to which data are trusted or highly regarded in terms of their source or content"
- Boolean : Ability to represent null values = "The degree to which a format allows null values in an attribute"
- Duration : Timeliness = "The degree to which the period between the time of creation of the real value and the time that the dataset is available is appropriate"

- Story : Recoverability = "The degree to which datasets are preserved in the event of incident"

This work by the Dutch team is excellent, but leaves me with an outstanding question: How do you aggregate data quality measures with different units of measure into a single aggregated total that represents, at a summary level, the overall quality metric for the relevant data set? Whether or not you use this foundational work from DAMA-NL, or use your own metrics and measures, it is very likely that you will end up with a range of different units of measurement on the one hand, and on the other a reasonable request from management to provide summary data quality metrics.

The options I've identified to solve this problem are:

1. Don't do it.
2. Aggregate only valid numeric measures, ignoring the qualitative metrics
3. Convert qualitative metrics into numeric values and aggregate the entire set of measures

Let's discuss these in more detail.

Option 1: Don't do it.

Is it really too much to ask for executive team members to review a small number of individual measures instead of a single aggregated measure? The level of abstraction necessary to create an aggregated measure introduces distortions to the data that make interpretation very prone to error and makes it impossible to accurately identify root cause and plan remediation action. The first point of call is to see if the most critical measures can be presented individually to avoid these issues.

Option 2: Only aggregate the numeric values

If a single metric of data quality must be presented, then one alternative is simply to aggregate the numeric values only. This does weaken the accuracy of the measure because it excludes quantitative information, but can be justified as not introducing human bias into the process and making it more sustainable over time. The method of aggregation could be simple addition, or a more complicated weighted average could be used, however the key difference with this solution is that the numeric values are included in the end result, without any attempt to include qualitative data.

Option 3: Convert qualitative data into numeric values

This option is clearly the most complicated. It requires a manual interpretation of text and the generation (from that interpretation) of a numeric score on the basis of the reader's assessment against a pre-defined numeric scale. For example, if half the data could be recovered in the event of a disaster, a score of 50% may be assigned as the equivalent numeric value. However, what happens if the data that could be recovered was all of last year, but none of this year. Is 50% still a valid response in that scenario? What happens if the 50% that could be recovered was all the text values for each record, but none of the numeric values? Is that still a 50% result? Inevitably, conversion of a

text-based description of a complex situation will require the reader to use their own experiences to determine the appropriate numeric value.

None of these, in my view, are ideal solutions.

How do you deal with aggregating quality metrics with different units of measure? If you or your business needs help with your Data Quality, please contact a Fujitsu Data & AI specialist now.

Contact
Fujitsu Data & AI
+61 3 9924 3000