September 2021

Data Quality Coverage

W. Max Lees

Yang Liu

Steven Lee

Mingyang Li

Keyu He

See next page for additional authors

Inventor(s)

W. Max Lees, Yang Liu, Steven Lee, Mingyang Li, Keyu He, Eric Wu, Emmett Cunningham, David Rissato Cruz, and Chioma Ezete

Data Quality Coverage

ABSTRACT

The quality of data is typically measured on a subset of all available data. It is of interest to know if such a measurement, performed on a subset of data, is representative of the entire corpus of data. This disclosure describes techniques that use historical data and metadata of a given time series to determine the set of useful data quality checks that can exist. The set of useful data quality checks is compared to the actual set of data quality checks to provide a percentage of data quality coverage that a given data set has.

KEYWORDS

- Data quality
- Data coverage
- Data subset
- Data stream
- Data analytics
- Data warehousing
- Dimensional values
- Code coverage

BACKGROUND

The quality of data is typically measured on a subset of all available data. It is of interest to know if such a measurement, performed on a subset of data, is representative of the entire corpus of data. Examples of this can be found throughout industry.

- *Example 1*: A taxi company estimates revenue along several dimensions, e.g., time-of-day, driver, district, etc. The revenue is determined only in a subset of all available times-of-days, drivers, districts, etc. The measurement shows a dip. Is the dip observed in the subset of data also present in the full data set?
- *Example 2*: A video sharing site measures hours watched along several dimensions, e.g., based on viewer attributes such as gender, age, country, etc. The hours-watched data is

measured only in a subset of all available genders, ages, countries, etc. The measurement shows a spike. Is the spike observed in the subset of data also present in the full data set? Is the spike observed in one country also observed in other countries? The answers to such questions can point to the causes and the nature of the spike, e.g., is it local, is it global, is it due to the popularity of a newly released video, is it due to malicious activity, etc., which in turn may lead to follow-on action, e.g., better identification and surfacing of popular videos, defensive action by computer security teams, etc. Conversely, an observed dip in a subset of data is better verified for quality, as it could point to a true decline in viewership, servers being unable to handle load, etc.

Another important question is to know the likelihood of the set of existing data quality checks catching any data quality issue of import. Data quality checks on time series data thus provide a way to test in an automated fashion whether a given stream of data is of high quality. Generally, the problem is the measurement of how well a set of data quality checks cover a given data set.

Although the problem of data quality coverage bears some resemblance to test coverage on code, e.g., how well unit tests discover bugs in code, there is no clear map from test coverage to data quality coverage. While tools exist that perform time series analysis on data, such tools don’t indicate how well the analysis covers a given data set.

DESCRIPTION

This disclosure describes techniques that use historical data and metadata of a given time series to determine the set of useful data quality checks that can exist. The set of useful data quality checks is compared to the actual set of data quality checks to provide a percentage of data quality coverage that a given data set has.

Some types of data quality coverages include metric coverage, dimensional values coverage, dimensions coverage, row-count coverage, etc., described in greater detail below.

Metric coverage

A metric is a quantitative measurement of data related to the success of a business or operation. In the example of the taxi business, a metric can be the distance traveled per trip; in the example of the video-sharing site, a metric can be session duration. Metric coverage answers questions asked by a producer of data, such as "how well does the data quality check detect data quality problems on the metric?" or questions asked by a consumer of data, such as "how likely are data quality problems detected on this metric?" Per the described techniques, for a given metric, if a metric sum is defined, it is deemed to have 100% coverage. If no metric sum is defined, its coverage is deferred to the dimensions coverage check described below. Thus, a single coverage value indicates how well a metric is covered as a whole.

Dimensional values (d-value) coverage

A dimension is an attribute or metadata about data. In the example of the taxi business, a dimension can be the date of the ride; in the example of the video-sharing site, a dimension can be the country of viewership; etc. D-value coverage answers questions posed by a producer of data, such as "how well does the data quality check detect data quality problems on this metric in the presence of some d-value?" or questions posed by a consumer of data, such as "how likely are data quality problems detected on this metric in the presence of some d-value?"

Per the techniques, if the metric has a dimensional breakdown check that includes the target d-value (either explicitly or implicitly), the d-value is given a 100% coverage for that metric. If the metric does not have a dimensional breakdown but still has a top-level metric sum, an effect-size equation is applied to determine whether the d-value is covered by the top-level metric. The d-value gets 100% coverage for the metric if the effect size equation is true. Thus, a single coverage value indicates whether the d-value is covered for a given metric.

The effect-size equation evaluates to true if the following inequality holds:

$$n > \frac{2.58\sigma^2 m}{\mu q},$$

where:

q is the percent change that triggers an anomaly detection;

m is the total number of rows in a data set;

σ is the standard deviation of the metric;

μ is the mean of the metric; and

n is the total number of rows corresponding to the d-value.

Dimensions coverage

In certain cases, the number of d-values for a given metric is so large that covering every d-value is infeasible or unhelpful. In such situations, where guarantees cannot be provided about the coverage of a specific d-value, a data producer can pose the following question. Since all d-values cannot be covered, for a user querying a metric with some randomly selected d-value, “how likely is it that existing data quality checks would detect data quality problems on that metric with that d-value?”

Per the techniques, for each possible d-value for a target dimension, coverage is calculated as follows.

$$c_D = \sum_{\square=1}^{\square} I_d c_d,$$

where *D* is the total number of possible d-values for the dimension, I_d is the importance of a d-value, and c_d is the coverage of the d-value. Thus, a single coverage number describes how likely any particular query that includes dimensional breakdown on that metric is covered by data quality checks. The sum of the importances over all possible d-values *d* equals one. Importance can be determined in various ways, such as:

- Row-count importance can be defined as the fraction of rows associated with a d-value as compared to the whole metric:

  $$I = \frac{r}{R},$$

  where *r* is the number of rows associated with the field of interest and *R* is the total number of rows.
- Contribution importance can be defined as the fraction of the total metric sum produced from the d-value:

  $$I = \frac{v}{V},$$

  where *v* is the sum of the metric of interest when some dimension equals a specific d-value and *V* is the sum of the whole metric.
- Popularity importance is the number of times a specific metric is queried compared to all queries made to the data set:

  $$I = \frac{\lambda}{\Lambda},$$

  where λ is the number of times a specific d-value was queried in some time window and $\Lambda$is the total number of times the data set was queried in the same time window.

Because each dimension has multiple possible d-values, not all necessarily included in a single check, a determination is made of the number of d-values that are covered. Although it is possible to simply look at the number of d-values covered by the check and divide that by the total number of d-values that exist for the dim, such a technique doesn’t work for dimensions with high-cardinality.

Instead, the percentage of the total metric that each d-value covers, and the d-values actually included in the check are determined, as follows. The weighted importance w_{dj} of a d-value for dimension d and index j is found by summing a metric wherever the dimension equals some d-value and by dividing that sum by the total sum of the metric. Let u_{dj} be one if the d-value j is covered by a check and 0 if not. Then

$$c_d = \sum_j \quad w_{dj} u_{dj}.$$

Row-count coverage

The row-count coverage is set to 100% if a check on the number of rows in the data exists and is zero otherwise.

Various types of coverage values, some examples of which are described above, can be further consolidated into a single number using a weighted linear combination, where the weights depend on the relative importances of each coverage to the business or organization. For example, a full coverage figure can be defined as

$$\begin{aligned} \text{full coverage} \quad = \quad & 0.3 \times \text{metric coverage} + 0.3 \times \text{row-count coverage} + \\ & 0.2 \times \text{dimensions coverage} + 0.2 \times \text{d-value coverage}. \end{aligned}$$

where the weights{0.3, 0.3, 0.2, 0.2} which are tunable parameters that sum to one and that represent the relative importance of the respective type of coverage value.

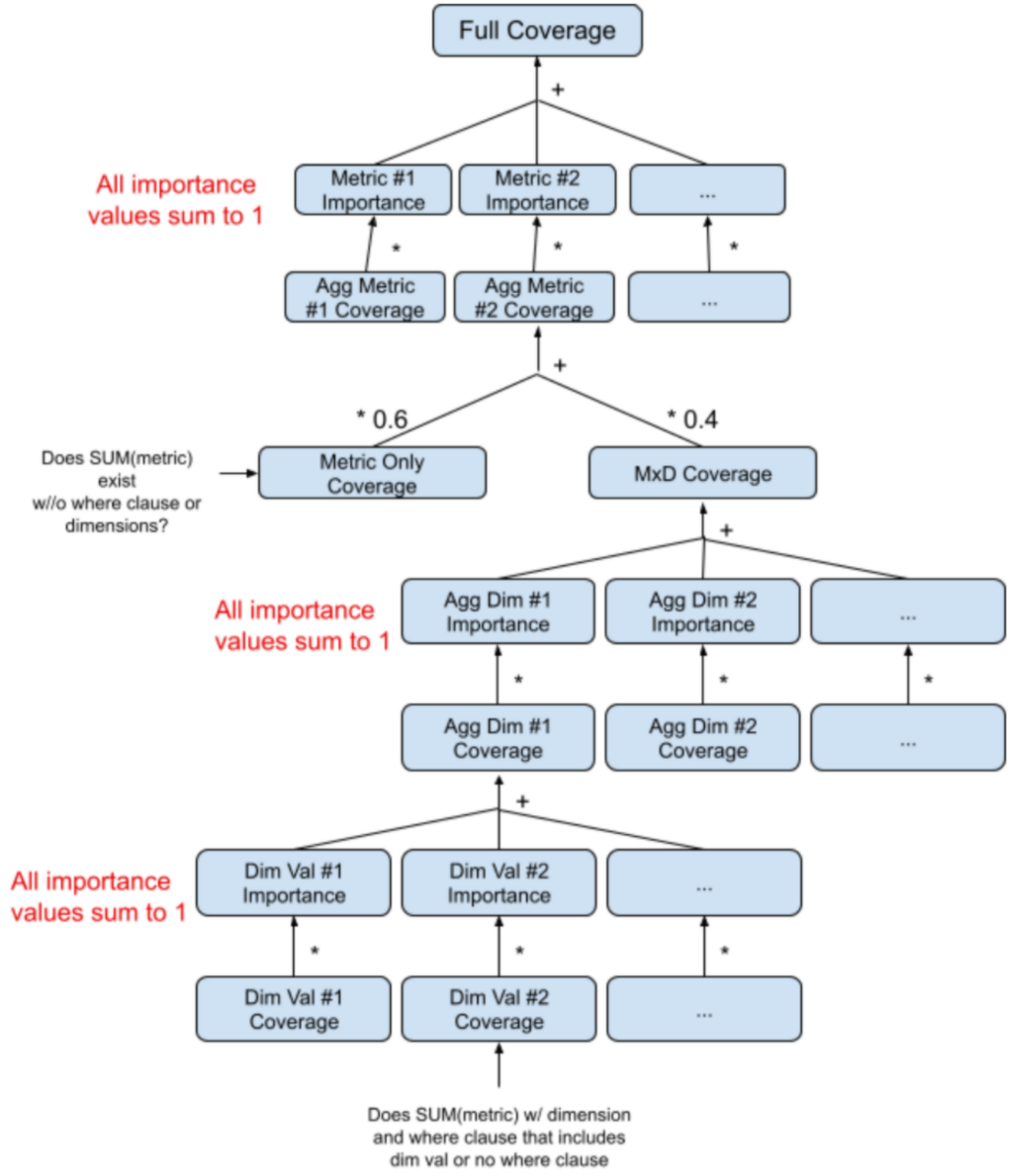

Fig. 1: Computing the full coverage from d-value, metric, dimensional, etc. coverages.

Fig. 1 illustrates the computation of the full coverage from coverage types such as d-value, metric, etc., each of which, as explained above, themselves can comprise importance-weighted sums of dimensional component coverages.

In this manner, the techniques of this disclosure provide an actual measurement of how well a set of data quality checks covers a data set.

CONCLUSION

This disclosure describes techniques that use historical data and metadata of a given time series to determine the set of useful data quality checks that can exist. The set of useful data quality checks is compared to the actual set of data quality checks to provide a percentage of data quality coverage that a given data set has.