November 2021

Automated Generation of Data Quality Checks by Identifying Key Dimensions and Values

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, Emmett Cunningham, David Rissato Cruz, Chioma Ezete, and Eric Wu

Automated Generation of Data Quality Checks by Identifying Key Dimensions and Values

ABSTRACT

Monitoring the quality and integrity of the data stored in a data warehouse is necessary to ensure correct and reliable operation. Such checks can include detecting anomalous and/or impermissible values for the metrics and dimensions present in the database tables. Determining useful and effective checks and bounds is difficult and tedious, and requires high levels of expertise and familiarity with the data. This disclosure describes techniques for automating the creation of data quality checks based on examining database schema and contents to identify important dimensions and values for data quality checks. The techniques utilize the observation that, in practice, a subset of the values of a database field are likely of operational importance. These are automatically identified based on calculating importance-adjusted data quality coverage by assigning importance to metrics, dimensions, and dimension values. Data quality checks are automatically generated for effective coverage of the key dimensions and values. The generation of checks can involve selecting from a repository of historically effective checks generated by experts and/or applying time series anomaly detection to metrics in entirety or sliced by key dimension values.

KEYWORDS

- Data quality check
- Data integrity
- Data validation
- Data warehouse
- Database dimensions
- Anomaly detection
- Invariant detection
- Data quality coverage

BACKGROUND

Websites and services that operate on large volumes of data, metadata, and analytics utilize backend databases that serve as the warehouse for the data. Ensuring correct and reliable operation requires continually monitoring the quality and integrity of the stored data. Such checks usually involve detecting anomalous and/or impermissible values for the metrics (i.e., numeric) and dimensions (i.e., non-numeric) present in the database tables.

Ideally, the specified quality checks need to cover the data such that all issues that are present are detected. In practice, higher coverage generates more true positives and fewer false negatives, thus catching more data quality errors caught. Detecting all issues present requires catching data quality problems at all levels: low (e.g., a single row), intermediate (e.g., a single metric/dimension value combination), and high (e.g., an entire metric). For instance, the value for a cell being 500% higher than the highest value for other cells of that type indicates an anomaly at a low level, all values of a metric for a given time being 50% lower than those for the same metric historically is a high-level problem, and all values of metric being 100% higher than previous for a given time for a specific device type is an intermediate-level issue.

The kinds of checks that are feasible and/or appropriate for a dimension field depend on the cardinality (the number of unique possible values for that dimension, with checking all values possible for low (e.g., < 10) cardinality. For medium (e.g., >=10 but < 1000) cardinality dimensions, checks are often restricted to a select few values, such as the most “popular” values. At high (e.g., >=1000) cardinality, any metric for a single value is likely to be highly unstable over time since new values are added constantly. Therefore, checks for dimensions with high cardinality typically involve aggregate measures rather than those connected to individual values.

Manually specifying and performing quality checks on the data is infeasible at large scales, which thus requires automation. The person specifying the automated checks to be performed need to choose the fields (dimensions and metrics) that need to be checked for anomalies and invariance, and configure the bounds of these checks (sensitivity for anomaly checks and static thresholds for invariant checks). Moreover, the number of checks increases with the number of metrics, dimensions, and dimension values, which makes specifying all the required checks a tedious task that is repetitive and error prone. For instance, creating an anomaly check for a metric sliced by several different dimensions requires similar but separate check specifications for each metric-dimension combination. Further, each check requires separate testing and debugging.

Therefore, determining the most useful, effective, and optimal checks and bounds is difficult and tedious, and requires high levels of expertise and familiarity with the data. Without such expertise and familiarity, the checks specified for monitoring data quality can yield a low signal-to-noise ratio and may lack balance between coverage and noise. As a result, such checks are likely to generate a large number of alerts that tend to be ignored because the receiver is overwhelmed and/or habituated.

Currently, specifying data quality checks typically involves a data producer picking the most important dimensions and adding anomaly detection on all values. Such an approach often results in an overwhelming number of anomaly alerts, thus leading the data producer to refine the initial specification by reducing the dimensions and/or values covered by the checks. The operation involves large amounts of repetitive work that yields low, inconsistent, and noisy coverage of the data that creates uncertainty regarding whether the checks catch issues to a sufficient extent.

DESCRIPTION

This disclosure describes techniques for automating the creation of data quality checks based on examining a database schema and contents. The automatically created checks are designed to balance maximization of coverage while minimizing the burden on the user monitoring data quality based on the outputs produced by the quality monitoring resulting from continually running the checks.

The techniques are based on the observation that only a few values of a database field are of operational importance in practice. Such values are typically stable over time and represent most of the data for that field barring a few exceptions (e.g., UNKNOWN or NULL values). Therefore, automated creation of data quality checks is based on automatically obtaining and classifying metadata about the database tables and identifying such key fields and values as focus of the data quality checks. The identification is based on calculating importance-adjusted data quality coverage by assigning importance to metrics, dimensions, and dimension values based on the number of metrics, dimensions, and values present in the database such that the sum of all importance values adds up to 1. The user tasked with specifying and monitoring data quality can be provided the option to inspect the output of the automated classification and make necessary edits.

Data quality checks are automatically generated for effective coverage of the identified key fields and values as determined above and optionally confirmed by the user. Generating the checks involves applying one or more of the following approaches:

- Selecting an applicable check from a repository of checks specified by experts in data quality checking, depending on whether the check has historically provided actionable alerts with low false positive and false negative rates;

- Applying time series anomaly detection to the sum of a given metric for a given slice;
- For a subset of the dimensions and their respective values, summing the same metrics sliced by these values while applying the same anomaly detection; etc.

The generated checks are then applied to generate example data for the selected metrics and dimensions to illustrate their application and operation to the user tasked with data quality monitoring. The user can examine the example results and comparisons with existing and historical quality check specifications, and choose to accept, reject, or edit the underlying check specification as appropriate.

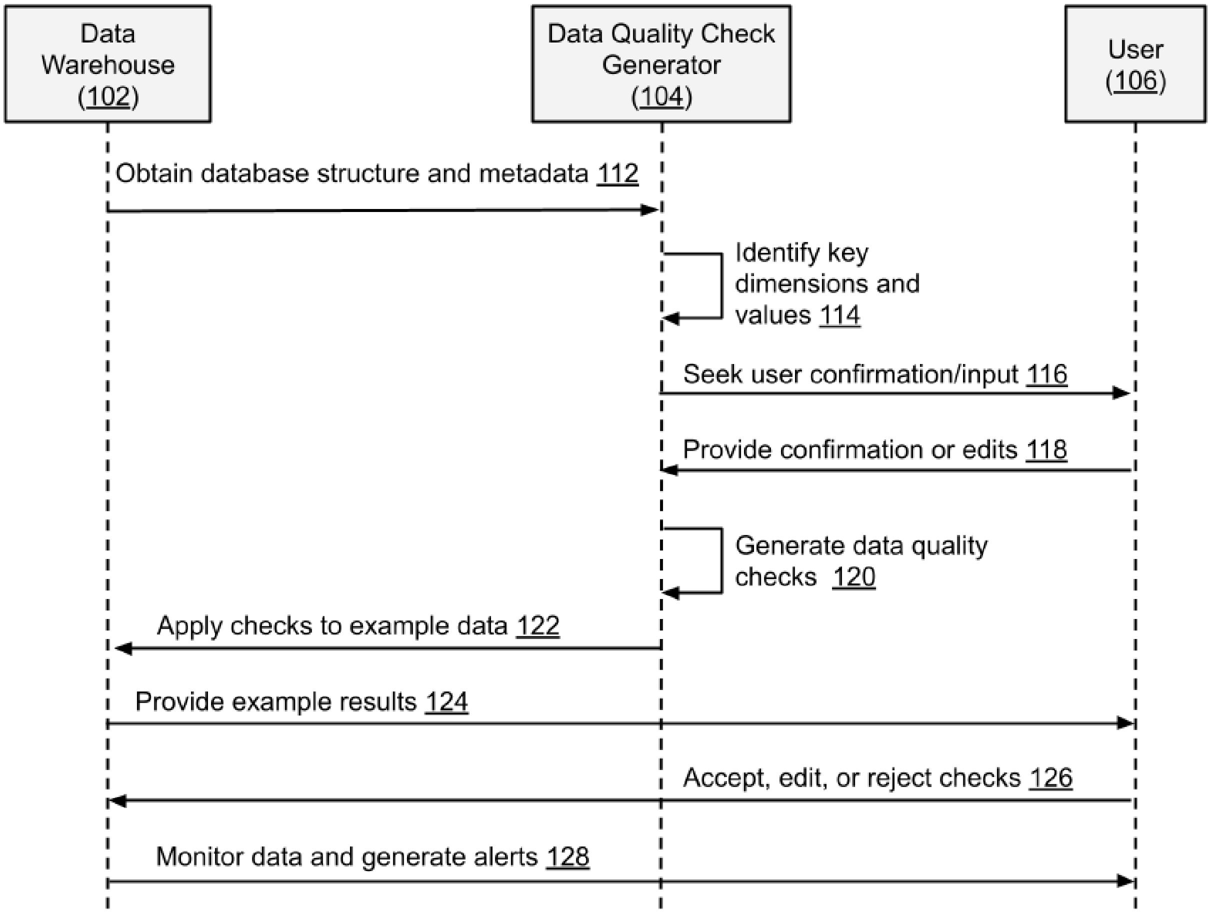
Data Warehouse (102)
Data Quality Check Generator (104)
User (106)
Obtain database structure and metadata 112
Identify key dimensions and values 114
Seek user confirmation/input 116
Provide confirmation or edits 118
Generate data quality checks 120
Apply checks to example data 122
Provide example results 124
Accept, edit, or reject checks 126
Monitor data and generate alerts 128

Fig. 1: Automated generation of data quality checks for a data warehouse

Fig. 1 shows an example operational flow for automated generation of data quality checks for a data warehouse per techniques of this disclosure. Relevant structural information and metadata from a data warehouse (102) is obtained (112) from a data quality check generator (104). The database information is used to identify key dimensions and values (114) within the data that are the most important as the target for data quality checks. The user (106) tasked with monitoring data quality is shown the automated classification of dimensions and values (116) to seek confirmation after making any edits, if needed (118). Upon user confirmation, appropriate data quality checks are generated (120) for maximizing coverage based on the identified key dimensions and values. The generated checks are then applied to example data (122) to provide the user with example results (124). The user can view the results to determine whether to accept, edit, or reject the automatically generated checks (126). The checks that are accepted by the user are applied for continual monitoring of the data to generate alerts when a check indicates a potential problem (128).

The generated automated data quality checks are simple to understand and do not require domain expertise in data quality. The checks produce a high signal-to-noise ratio and avoid undue burden on the user monitoring the data quality. The generated checks provide high coverage of key metrics and dimensions as appropriate for the given database table and its place in the broader data warehouse.

The backend operations involved in the implementation of the techniques can be connected to any suitable frontend, such as a user interface (UI) accessible from a web browser. Compared to manual specification, using the suggested data quality checks automatically generated using the described techniques can help those producing and/or monitoring the data gain a more refined understanding of the quality of the underlying data and quickly generate effective data quality checks. The resultant checks can provide substantially higher coverage while generating fewer alerts and reducing setup time from several hours to a few minutes. Moreover, setting up effective data quality checks using the described techniques requires no specialized data expertise and/or familiarity with the underlying data.

The techniques described in this disclosure can be applied to generate automated data quality checks for any type of database or data warehouse, irrespective of the type of data. Application of the techniques can be particularly useful for any entity such as a website, application, or online service that makes use of large data warehouses that require automated quality monitoring.

CONCLUSION

This disclosure describes techniques for automating the creation of data quality checks based on examining database schema and contents to identify important dimensions and values for data quality checks. The techniques utilize the observation that, in practice, a subset of the values of a database field are likely of operational importance. These are automatically identified based on calculating importance-adjusted data quality coverage by assigning importance to metrics, dimensions, and dimension values. Data quality checks are automatically generated for effective coverage of the key dimensions and values. The generation of checks can involve selecting from a repository of historically effective checks generated by experts and/or applying time series anomaly detection to metrics in entirety or sliced by key dimension values.