Data Quality Analysis: Profiling and beyond

DAMA-NL Data Quality Working Group

Key words: **Data Profiling, Data Quality Analysis**

Introduction

This document provides a short introduction in Data Quality Analysis. We coin this term as sub-area of Data Quality Management which covers the process of discovering Data Quality Rules and Data Quality Issues.

Profiling is most known as a method to perform such analysis. However, this is just one of many possible approaches to analyse existing data from a Data Quality perspective. Therefore, we have determined this broader more generalized definition.

Data Quality Analysis is the process of analysing the content of Critical Data Elements in its current known data model and architecture in order to discover common patterns that will provide information on the current state of the data quality and its underlying requirements.

Data Quality Analysis should not be confused with the prior process of Data Discovery, which constitutes the identification of existing data sourcing, its structure and architecture and Critical Data Elements within them. Usually, Data Quality Analysis is performed as a follow up to this.

Purpose of Data Quality Analysis

Data Quality Analysis can be employed at multiple moments and for various purposes during the deployment of a Data Quality Management System. Its application depends on the organisations' needs as described in their Data Quality Objectives and Data Quality Policy as well as its progression towards implementing a Data Quality Management System.

Common applications of Data Quality Analysis:

- Data Quality Assessment: Gaining overall insight in current Data Quality status. Finding general problem areas in order to determine focus areas and further shape Data Quality strategy and policy.
- Data Quality Problem Analysis: Gaining deep insight in various types of existing errors and its root causes. Identify occurrences, frequencies and impact in order to define corrective and preventive actions.
- DQ Requirements Definition: Gaining deep insight in Data Quality requirements in order to set up a set of Data Quality Rules and install complete and relevant DQ Measurement System.

Regardless of its context specific application, Data Quality Analysis usually has the following generalised purpose:

- Discover and understand patterns in existing data
- Discover existing errors and its root causes
- Discover and operationalize existing Data Quality requirements

General procedure

Executing Data Quality Analysis can be done on various levels of depth using many methods, however there is always a common procedure following a number of steps:

- Step 1: Gather data and metadata: Gain access to data sources, information on architecture and modelling, reference and metadata, data definitions and information on data owners. This information might not be complete or accurate early on in Data Management initiatives
- Step 2: Prepare data: Convert data into a proper format for analysis. Usually this is a table/view form for using various SQL or ETL tools. This is usually easy for data in relational database, but might be problematic with more complex data structures, such as graph data, geo data, unstructured data and raw machine generated logs.
- Step 3: Execute analysis: Load a recent snapshot of the data into analysis tooling. When data is large use a representative amount of random sampled records. Apply the specific chosen analysis methods. More on this in the next chapter.
- Step 4: Interpret outcomes: Data Stewards, Data Quality Analysts and/or Data Scientists will examine the outcomes and prepare presentation to business users, Data Engineers and IT people. They will discuss their findings to arrive at conclusions: Found errors, Data Quality rules and identified problem areas.
- Step 5: Follow-up: Assess outcomes on relevance and impact and select qualified results. Then dependent on the purpose of the Data Quality Analysis, either compile Data Quality Rules and enter them into the monitor and/or compile errors and problem areas and work this out into an assessment or action plan.

In practise these steps are rarely performed in a waterfall approach. Often intermediate insights require going back, correct and redoing previous steps. This often happens during step 3 and 4.

Also, iterative approaches happen often where the same procedure is repeated multiple times to deepen prior analyses or update them to recent context changes.

Finally, another common approach is to divide the data into chunks and use an incremental approach where pieces of the data are analysed and results are added to a general collection of errors, business rules or assessment results.

Data Quality Analysis levels and methods

Data Quality Analysis can be divided into four consecutive levels of analysis depth, each adding additional insights:

- Level 1: Profiling Analysis: Basic independent analysis of single data fields.
- Level 2: Dependency Analysis: Identify generalized relations between data fields.
- Level 3: Bivariate Correlation Analysis: Specify the influence between two fields.
- Level 4: Pattern Analysis: Specify complex correlation between multiple fields.

Each level of analysis provides Data Quality insights on its own, however they reinforce each other. It is recommended to choose a desired analysis depth and start from level 1 and work forwards. We will walk through these levels now and make clear how they are logically stacked.

Level 1: Profiling Analysis

The first step is to perform a profiling analysis on all columns in the table. This analysis involves assessing frequencies of value occurrences in the column. For scale type data, often numeric and datetime, we also assess statistical distributions (average, variance and skewness). Finally, we assess the occurrences of empty/missing data and proxies for these, such as 0, -1, or 1900-01-01.

The main goal of the profiling analysis is to gain knowledge on the nature of each data field and its meaning:

- Discover the role of the data field: Is it an ID, listed category, free text, description, numeric scale value or a numeric code for categories or ordinal levels?
- Discover the relevance of the field. Which are all empty or the same? What is actively used and contains business information and what is auto-filled or auto-generated?
- Discover Data Quality requirements within the field. Discover (in)valid categories and ranges. Discover general completeness requirements (Is the field allowed to be empty?) and uniqueness requirements (Are duplicates allowed?)

Level 2: Dependency Analysis

The most relevant and impactful Data Quality requirements often lie within the relation between data fields. Therefore, the next step is to identify which fields are correlated. This is done using a Dependency Analysis where a N-N confrontation matrix is formed that contains all relations between two of the N columns in the table. This visualises which fields are related.

There are roughly two methods to do this:

- Simple correlation analysis: Here we use statistical methods to quantify the correlation. Dependent on the nature of the two data fields we use Chi-Square (Two categorical fields), Annova/T-Test (Categorical versus Scale data) or regression (Two Scale fields).
- Association Analysis: Here we calculate the degree of uncertainty of Entropy to determine to which extent a variable can be explained/predicted by another variable, such as Theil's, U. Association Analysis has the added advantage of identifying the direction of the dependency between two variables in addition to its correlation.

The main goal of Dependency Analysis is to identify on a high level how fields are related to each other:

- Discover related fields in order to reduce the search field for further analysis. Single out groups of related data. Narrow down search field for further in-depth analysis
- Discover the direction of dependency: Which fields dictate other fields? Are some fields calculated using others? What are source fields and which are derived?
- Identify important fields for further analysis and requirement definition. Relevant fields with high impact are often strongly related to many other data fields.

Level 3: Bivariate Correlation Analysis:

After the most relevant fields and relations have been identified, the next step is to analyse the exact nature of these correlations. This will lead to concrete insights in Data Quality problems and identify Data Quality Rules and errors.

This is most commonly done by visualising the relation between fields, dependent on the type of data:

- Frequency tables: For the comparison of two categorical fields, a frequency table is often most useful. This shows the number of occurrences combinations of two values occur. Often you will find that many combinations never occur. For example, the city name Rotterdam will mostly coincide with the province field Zuid-Holland, indicating an obvious relation between city and province.
- Boxplot visualisations: This is used to compare categorical data with scale/numeric data. This visualisation shows the ranges of the scale data for various categories, indicating possible differences between the groups. For example, apartments usually have a smaller net floor space than condos. 30 m2 is plausible for an apartment, however unlikely for a condo.
- Scatter plot visualisations: For visualising two scale/numeric variable scatter plots are useful. Additionally, regression lines can be fitted when there is a clear correlation. Otherwise, clustering analyses can be applied to identify groups of corresponding values.

The main goal of the Bivariate Correlation Analysis is to substantiate the abstract correlations between fields:

- Identify valid combinations and improbable outliers.
- Define Consistency related Data Quality Rules using strong, trivial dependencies.
- Define Accuracy related Data Quality rules using moderate dependencies.

Level 4: Pattern analysis

Often the business logic behind Data Quality Rules is more complex in nature and comparing two variables at the time does not lead to a clear dependency. Often there is a more complex if-then logic in the process that leads to the data. Comparing multiple fields becomes too computational since there are too many possible combinations.

This is where machine learning methods become useful. We can apply these techniques to our data set to train models that predict values using multiple other fields. We then extract the logic of these models and translate them into readable and understandable if-then-logic.

Useful algorithms for these are K-Nearest Neighbour, Decision Trees/Random forests, Support Vector Machines and simple Neural Nets with fixed network structures. Also, for finding common text structures Natural Language Processing techniques and regex fitting techniques can be applied.

The main goal of this final step is to materialise complex data requirement into clear Data Quality Rules:

- Identify more complex multivariate dependencies between variables
- Specify the most relevant Data Quality rules that have most impact on business processes.
- Discover hidden outliers and errors that would go unnoticed otherwise and jeopardise the progress towards a data driven organisation.