Data cleansing

Definition

Data cleansing is the process of detecting and correcting data issues to improve the quality of data to an acceptable level.

Notes

An organization should define an acceptable data quality level for each data quality dimension.

Dimensions of data quality that can be improved by data cleansing are:

- Accuracy of data values
- Completeness of records
- Completeness of data values
- Compliance of data with laws, regulations, and standards]]
- Consistency of data values
- Currency of data values
- Integrity of data values
- Linkability of data files
- Metadata compliance of data values
- Precisions of data values
- Referential integrity of data values
- Uniqueness of records
- Validity of data values

Synonyms

- Data Cleaning
- Data Remediation
- Data Scrubbing
- Error Correction

Purpose

To detect and correct data issues and inconsistencies.

Life cycle

Phase	Activity
Plan	* Detect unexpected, incorrect, and inconsistent data * Select data elements or data files to be cleaned * Select data cleansing methods * Develop this data cleansing process (manual/automated)

Phase	Activity
Do	* Import data * Merge data sets * Rebuild Missing data * Standardize * Normalize * De-duplicate * Verify & enrich * Export Data * Document data cleansing results
Check	* Verify the results * Evaluate the data cleansing process
Act	* Adapt the data cleansing process for further use

Note 3: Data cleansing can be executed manually or automatically or in mixed mode.

Methods

The next methods of correcting data issues can be distinguished:

Method	Example or explanation
Abbreviation expansion	Abbreviation expansion transforms abbreviations into their full form. There are different kinds of abbreviation. One type shortens each of a set of words to a smaller form, where the abbreviation consists of a prefix of the original data value. E.g. “USA” stands for “United States of America.”
Clustering	Clustering is one of the statistical methods that can be used to find values that are unexpected and thus erroneous. Clustering is a classic data mining technique based on machine learning that divides groups of abstract objects into classes of similar objects. Clustering helps to split data into several subsets. Each of these clusters consists of data objects with high inter-similarity and low intra-similarity.
Cross-checking with a validated data set	Some data cleansing solutions will clean data by cross-checking with a validated data set. E.g. addresses checked against the BAG data set as part of the Dutch government system of basic registration and contains basic municipal data of all addresses and buildings in Dutch municipalities, the Dutch postal code check with the data set of PostNL and check the data set of the chamber of commerce.
Remove duplicates	Duplicates are data points that are repeated in your dataset. Every duplicate detection method proposed requires an algorithm for determine whether two or more tuples are duplicate representations of the same entity. Classification is a method to remove duplicate data.
Data enhancement	Enhancement is the process that expands existing data with data from other sources (enrichment). Here, additional data is added to close existing information gaps.
Data harmonization	Data harmonization is the process of bringing together your data of varying file formats, naming conventions, and columns, and transforming it into one cohesive data set.

Method	Example or explanation
Remove inconsistency	Data inconsistency is a condition that occurs between tables when we keep similar data in different formats in two different tables. Data inconsistency creates unreliable information, because it will be difficult to determine which version of the information is correct.
Remove irrelevant data	Irrelevant data are the data that are not actually needed, and don't fit under the context of the problem we're trying to solve.
Merging	The merging of two or more databases will both identify errors (where there are differences between the two databases) and create new errors (i.e. duplicate records).
Drop or impute missing values	Missing values are data or data points of a variable that are missing. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.
Normalization	Normalization is a formal technique that eliminates the data redundancy in a number of steps (= normal forms) by splitting the data according to fixed rules.
Remove outliers	Outliers are values that are significantly different from all other observations. Outliers are innocent until proven guilty. With that being said, they should not be removed unless there is a good reason for that.
Parsing	Parsing is a method where one string of data gets converted into a different type of data. Parsing in data cleansing is performed for the detection of syntax errors.
Patterns	Patterns give a generalized view of how the data is formatted; it involves parsing the data and classifying all tokens into appropriate classes and replacing those classes with pre-defined labels.
Removing typographical errors	A typographical error (often shortened to typo), also called misprint, is a mistake (such as a spelling mistake) made in the typing of printed (or electronic) material. The term includes errors due to mechanical failure or slips of the hand or finger, but excludes errors of ignorance, such as spelling errors, or changing and misuse of words such as "than" and "then".
Standardization	Standardization transforms data into a standard form. Standardization is used to extract entity information (e.g., person, company, telephone number, location) and to assign some semantic value for subsequent manipulation. Standardization will incorporate information reduction transformations during a consolidation or summarization application.
Statistical methods	Statistics is the science and technique of collecting, processing, interpreting and presenting data based on rules of mathematics and the laws of logic. Statistical methods are used to identify data issues. Statistical methods include regression, correlation, min, max, standard deviation, mean and clustering.
Correct syntax errors	A syntax error is an error in the syntax of a sequence of characters or tokens that is intended to be written in compile-time. A program will not compile until all syntax errors are corrected. A syntax error gives you important clues on how to correct your code. Types of syntax errors are typo's, pad strings and white spaces.
Transformation	Transformation involves transforming data according to rules and lookup tables or making combinations of data from different sources. Data selection, mapping and data cleansing are some basic transformation techniques. Advanced data transformation techniques include: standardization, character set conversion and encoding handling, field splitting and merging, summary and de-duplication.
Type conversion	Type conversion (also called casting) is an operation that converts a piece of data of one data type to another data type. Type conversion can be used to make sure that numbers are stored as numerical data types and that a date should be stored as a date object.

Method	Example or explanation
Edit rules	Edit Rules, a new class of data quality rules, are rules that tells how to fix errors, i.e. which attributes are wrong and what values they should take.
Data lifecycle management	Data Lifecycle Management can be defined as the different stages that the data traverses throughout its life from the time of inception to destruction. Data lifecycle stages encompass creation, utilization, sharing, storage, and deletion.

Note 4: Data issue prevention is far superior to data issue detection and cleansing, as it is cheaper and more efficient to prevent issues than to try and find them and correct them later.

It is also important that when issues are detected that feedback mechanisms ensure that the issue doesn't occur again to both the collection of the data and the entry of the data, or that there is a much lower likelihood of it re-occurring.

Data issue prevention prevents data cleansing and make sure that no choice have to be made for a data cleansing method.

Principles of data cleansing

Chapman (2005) states that many of the principles of data cleansing overlap with general data quality principles covered in the associated document on Principles of Data Quality (Chapman 2005a). The data cleansing key principles include:

- Planning is Essential (Developing a Vision, Policy and Strategy)
- Organizing data improves efficiency
- Prevention is better than cure
- Responsibility belongs to everyone (collector, custodian and user)
- Partnerships improve Efficiency
- Prioritisation reduces Duplication
- Setting of Targets and Performance Measures
- Minimise duplication and reworking of data
- Feedback is a two-way street
- Education and Training improves techniques
- Accountability, Transparency and Auditability
- Documentation is the key to good data quality

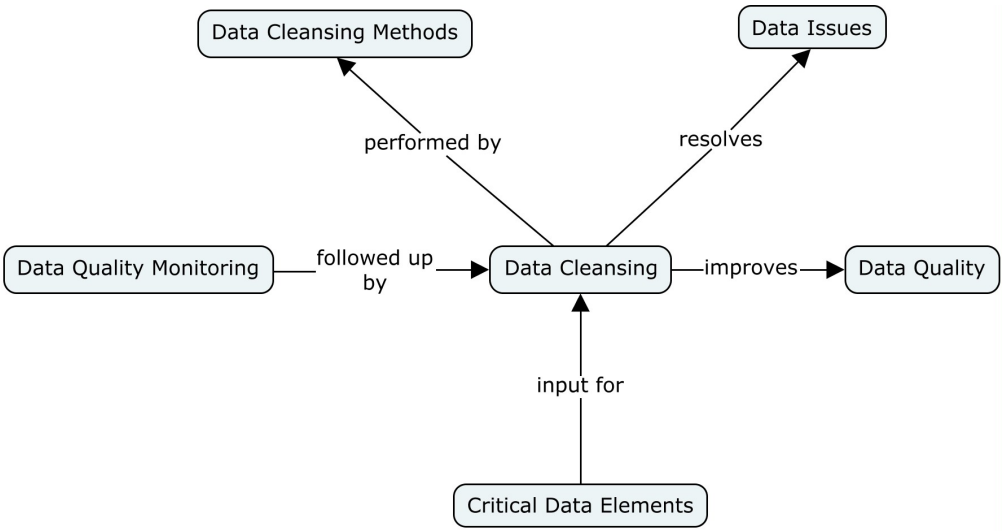
Characteristics

Characteristic	Requirements
Effectiveness of data cleansing	Data Cleansing improve the data quality and meets the norm of the data quality dimension(s).
Cost-effectiveness of data cleansing	Data cleansing must lead to a positive business case, i.e. the benefits must be bigger than the costs.

Relations

- Data cleansing resolves data issues.
- Data quality monitoring is follow-up by data cleansing.

- Purpose of data cleansing is to improve data quality and make it ‘fit for use’
- A data cleansing process can be performed using one of the different data cleansing methods.
- Critical data elements are input for the procedure to data cleansing.



Story

In this story we look at filling missing data and erasing incomplete data as an example of data cleansing. It concerns a Polish Business to Business contractors addresses database of architect Bolek, which are saved in CRM in the following format: voivodship, district, postal code, city, and street.

Let us assume that Bolek wants to have only complete company addresses, i.e., complete data sets (incomplete data does not contribute anything to the business process). We can approach this topic in two ways:

- delete all records that have an empty value in any field (which is not an ideal solution, because we lose a lot of information),
- complete incomplete records (which is a much better choice, considering that a voivodeship or a commune can be easily completed based on the name of the city or postal code), and only what cannot be retrieved with a supplement (in this case, e.g., sets with empty street info) remove.

We decide to clean the database in the second way.

In order to facilitate this task and perform it fully professionally, it is necessary to define some repetitive and exhaustive rules that will apply to this data set in turn. They take the following form:

- If the voivodship field is empty, we complete it based on the city.
- If the city field is empty, we check whether we can determine the city name based on the postcode field (we will not always be able to do this - there are many common postal codes for various smaller towns and villages).
- If the district field is empty, we complete it based on the city and postal code.
- We are introducing a few rules for clearing the data in the street column, such as clearing null strings or removing values where there are no letters other than street.

- In the last step, we get rid of the records that are still left with empty values in any of the fields of a single dataset.

After applying the above set of rules, our cleaned database of company addresses looks like this:

References

Antkowiak, M., & Nowaczyk, M. (2021, 26 februari). *Data cleansing examples - Blog Transparent*

Chapman, A.D. (2005). *Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data, version 1.0.* Report for the Global Biodiversity Information Facility, Copenhagen. Available online at https://www.gbif.org/document/80528.

Chapman A.D. (2005a). *Principles of Data Quality.* Global Biodiversity Information Facility. https://doi.org/10.15468/doc.jrgg-a190

CXO's Guide to Marketing and Sales Data Cleansing and Enrichment | DEO Blog. (2018, 5 juni). Dataentryoutsourced.Com/Blog/. https://www.dataentryoutsourced.com/blog/cxos-guide-to-marketing-and-sales-data-cleansing-and-enrichment/

DAMA (2017). DAMA-DMBOK. *Data Management Body of Knowledge.* 2nd Edition. Technics PublicationsLlc. August 2017.

DAMA Dictionary of Data Management. 2nd Editioin 2011. Techniscs Publications, LLC, New Jersey.

DAMA NL Foundation, Black, A., & van Nederpelt, P. (2020, november). *Dictionary of dimensions of data quality(3DQ) - Dictionary of 60 Standardized Definitions* (v1.2). DAMA NL Foundation. http://www.dama-nl.org/wp-content/uploads/2020/11/3DQ-Dictionary-of-Dimensions-of-Data-Quality-version-1.2-d.d.-14-Nov-2020.pdf

Data cleaning: The benefits and steps to creating and using clean data. (n.d.). Tableau Software. https://www.tableau.com/learn/articles/what-is-data-cleaning

Data ENG. Medium.

Diallo, T., Petit, J. M., & Servigne, S. (2012). Editing Rules: Discovery and Application to Data Cleaning. *Conference: 10th International Workshop on Quality in Databases In conjunction with VLDB*. Published.

Elgabry, O. (2019, March 2). The ultimate guide to data cleaning. Medium. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4#6058

Freytag, J., & Mueller, H. (2005). *Problems, Methods, and Challenges in Comprehensive Data Cleansing*. Computer Science.

Gimenez, L. (2020, 20 November). *6 steps for data cleaning and why it matters*. GEOTAB. https://www.geotab.com/blog/data-cleaning/

Kumar, S. (z.d.). What is Data Lifecycle Management? stealthbits.com. Geraadpleegd op 6 maart

2022, van https://stealthbits.com/blog/what-is-data-lifecycle-management/

https://medium.com/transparent-data-eng/data-cleansing-examples-24581c3d14f1

ISO 9000:2015. Quality Management Systems – Requirements.

ISO 9001:2015. Quality Management Systems – Fundamentals and vocabulary.

Species-Occurrence Data (1.0 ed.). Global Biodiversity Information Facility.

What is data cleansing? Guide to data cleansing tools, services and strategy. (2020, August 13). Talend Real-Time Open Source Data Integration Software. https://www.talend.com/resources/what-is-data-cleansing/