Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online
CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-1 (2015) 111–112

Data Wrangling: Making data useful again

Florian Endel * **Harald Piringer** **

* *University of Technology Vienna (florian.endel@tuwien.ac.at)*
** *VRVis Research Center, Vienna, Austria*

Abstract: Data analysis has become an everyday business and advancements of data management routines open up new opportunities. Nevertheless, transforming and assembling newly acquired data into a suitable form remains tedious. It is often stated, that data cleaning is a critical part of the overall process, but also consumes sublime amounts of time and resources. Data Wrangling is not only about transforming and cleaning procedures. Many other aspects like data quality, merging of different sources, reproducible processes, and managing data provenance have to be considered. Although various tools designed for specific tasks are available, software solutions accompanying the whole process are still rare.
In this paper, some aspects of this first phase of most data driven projects, also known as data wrangling, data munging or janitorial work are described. Beginning with an overview on the topic and current problems, concrete common tasks as well as selected software solutions and techniques are discussed.

Keywords: Data acquisition, Databases, Bad data identification, Data wrangling.

1. INTRODUCTION

The terms “data science”, “datafication”, “business analytics” and “big data” (Ayankoya et al. (2014); Cukier and Mayer-Schoenberger (2013); Larson (2013); Lohr (2012)) were coined based on many different developments in data retrieval, storage and analysis during the last years. Although tools and technologies evolve constantly, understanding and preparing a newly acquired dataset for further usage still requires much time and effort. This initial and very fundamental process of examining and transforming data into a usable form is known as “data wrangling”, “data munging” or even “janitorial work”.

As foundation of the data wrangling process, a broad and deep understanding of the content, structure, quality issues and necessary transformations as well as appropriate tools and technological resources are needed. The whole wrangling procedure needs to be very efficient, especially for small projects or unique datasets, where the effort to automate and document does not seem to be achievable, although necessary.

Altogether, data cleaning accounts for 50 percent to 80 percent (Kandel et al. (2011); Dasu and Johnson (2003); Lohr (2014)) of the time and costs in analytic or data warehousing projects respectively.

Different challenges and possible solutions are described as basis for further discussion. The focus is set primarily on short-term projects, directed by a tight schedule and (naturally) merely adequate resources.

Recognising the practical significance and describing applicable workflows for data wrangling, while keeping in mind the necessity to focus real-world problems and constraints, is the main scope of this discussion.

2. MOTIVATION

Collecting data from more or less structured sources and preparing it for visualization, modelling or permanent storage is a necessity of our projects. As described above, this work can be tedious and consumes an inordinate amount of time. An overview of common topics, routines and challenges alongside with achievable solutions is collected to speed up and improve structuring of future work.

3. APPROACH

Based on published literature and gathered experiences, several key aspects concerning data wrangling are identified and discussed. Understanding data preparation as iterative, multidisciplinary process, the focus is set on realistically applicable routines and solutions for most common sources of error.

3.1 Example scenarios

Two diverse, real world exemplary scenarios are presented. By reference to problems, applied solutions, and lessons learned during these projects, several aspects of the wrangling process are described.

First, a huge collection of several undocumented data files have to be prepared and cleaned. In the end a highly structured database storing the foundation for further analysis as well as a reproducible process of the whole transformation is needed.

Second, complex but well known data is extracted from another database and put into a format that can easily be used for visualization. Additionally, auxiliary information from additional sources have to be merged into the exported file.

10.1016/j.ifacol.2015.05.197

3.2 Challenges

A collection of main challenges appearing in nearly every data driven project are summarized. Next, solutions for each objective and ideally, for a model process covering at least typical requirements have to be discussed.

Basic Features Approaching a new dataset for the first time, many basic aspects like size, encoding, and format have to be explored.

Data Quality is a central aspect of almost every data driven project. Several dimensions of data quality and metrics to quantify them are defined in literature. Ideally, data quality is monitored and documented throughout the whole lifetime of a dataset.

Merging & Linking Integration of data is often needed to complete the picture. Therefore, merging and linking of divergent structures, coding conventions, formats and data models are an important task when preparing data for further utilization. Varying scenarios can be identified based on the variety of sources, type of data, linkage method, objective and reason for data linkage.

Reproducibility & Documentation are important but often neglected components of every scientific research. Handling changes of data and methods over time as well as the recreation of previously acquired results are common demands which can be hard to fulfil, especially in connection with manual interaction.

Big Data Generally, large amounts of data can be complicated to analyse. Although the volume of data which is labelled as being big depends on the application at hand, managing and analysing millions or billions of datasets or several gigabytes / terabytes respectively, does require special treatment and technologies (e.g. Hu et al. (2014)).

Dirty Data Despite all efforts conducted during data quality assessment processes, it is not always clear how knowledge about quality issues can be handled appropriately. In some cases further feedback loops with data sources and providers are viable and cleaning up flaws is possible. Otherwise, strategies to work with dirty data have to be implemented. Depending on the severity of the errors, the resulting effects can range from minor disturbances to the necessity to completely re-engineer analytical processes.

Uncertainty Data quality does not only occur in discrete statuses, i.e. clean and faulty. In Skeels et al. (2008) five different types of uncertainty are classified: measurement precision, completeness, inference, disagreement, and credibility. Reasons for uncertainty range from measurement errors, processing errors as far as intentionally introduced inaccuracies e.g. due to privacy concerns. Visualization (e.g. Correa et al. (2009)) can help to intuitively present uncertainty.

Error tolerance Known and unknown errors as well as uncertainties should be expected when using real world data (e.g. Haug et al. (2011)). Therefore, data storage (e.g. Benjelloun et al. (2007)), analytic routines, and algorithms need to tolerate different kinds of errors.

Transformation & Editing Although the whole data wrangling process is about transforming some kind of input into a usable form, the concrete implementation of these transformations including all additional constraints (e.g. quality, reproducibility) is still an open issue. Many software solutions exists with advantages and drawbacks.

4. RESULTS

Summarizing, it can be shown that data wrangling is still a defiant process. Although a variety of solutions for single problems and special situations exist, hardly any of them cover all raised considerations.

Concepts and tools facilitating the struggle of data wrangling require additional research. Additionally, further requirements as well as the meaningfulness of presented objectives have to be discussed and refined.

REFERENCES

Ayankoya, K., Calitz, A., and Greyling, J. (2014). Intrinsic relations between data science, big data, business analytics and datafication. In *Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference 2014 on SAICSIT 2014 Empowered by Technology*, SAICSIT '14, 192:192–192:198. ACM, New York, NY, USA. doi: 10.1145/2664591.2664619.

Benjelloun, O., Sarma, A.D., Halevy, A., Theobald, M., and Widom, J. (2007). Databases with uncertainty and lineage. Technical Report 2007-26, Stanford InfoLab. URL `http://ilpubs.stanford.edu:8090/811/`.

Correa, C., Chan, Y.H., and Ma, K.L. (2009). A framework for uncertainty-aware visual analytics. In *IEEE Symposium on Visual Analytics Science and Technology, 2009. VAST 2009*, 51–58. doi:10.1109/VAST.2009.5332611.

Cukier, K.N. and Mayer-Schoenberger, V. (2013). The rise of big data. *Foreign Affairs*, (May/June 2013).

Dasu, T. and Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, Inc., New York, NY, USA, 1 edition.

Haug, A., Zachariassen, F., and Liempd, D.v. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, 4(2), 168–193. doi: 10.3926/jiem..v4n2.p168-193.

Hu, H., Wen, Y., Chua, T.S., and Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access*, 2, 652–687. doi: 10.1109/ACCESS.2014.2332453.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N.H., Weaver, C., Lee, B., Brodbeck, D., and Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4), 271–288.

Larson, E.B. (2013). Building trust in the power of big data research to serve the public good. *JAMA*, 309(23), 2443–2444. doi:10.1001/jama.2013.5914.

Lohr, S. (2012). Big data's impact in the world. *The New York Times*.

Lohr, S. (2014). For big-data scientists, janitor work is key hurdle to insights. *The New York Times*.

Skeels, M., Lee, B., Smith, G., and Robertson, G. (2008). Revealing uncertainty for information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '08, 376–379. ACM, New York, NY, USA. doi:10.1145/1385569.1385637.