D.G. Ruparel College of Arts, Science and Commerce
DATA WRANGLING
Roll no: 6003

Objectives

This case study mainly focuses on the process of data wrangling and the challenges faced by the analyst while performing it. While writing this case study I was able to:

- Get a deeper knowledge about what data wrangling actually is.
- Read different research papers and learn the actual process.
- Gather knowledge on the various steps and tools involved in the process of performing data wrangling.
- Learn how the need of data wrangling arise.
- Get to know about the challenges analyst face while performing data wrangling and their causes.
- Figure out solutions for the existing drawbacks using the data from the research papers.

And all of the above points are briefly reported in this case study.

Introduction

Data wrangling or data munging is defined as the process of taking disorganized or incomplete raw data and standardizing it so that you can easily access, consolidate, and analyse it. It also involves mapping data fields from source to destination. A data wrangling example could be targeting a field, row, or column in a dataset and implementing an action like joining, parsing, cleaning, consolidating, or filtering to produce the required output.

You can then use this wrangled data to process it further for business intelligence (BI), reporting, or improving business processes. Therefore, the process ensures that the data is ready for automation and further analysis.

Just like most data analytics processes, it is an iterative process in which you have to perform the five data wrangling steps recurrently to get the results you want. The five broad data wrangling steps are:

- **Understanding Data**

The first data wrangling step is to understand the data in great depth. Before applying procedures to clean it, you must have a clear idea of what the data is about. This will help you find the best approach for productive analytic explorations.

- **Structuring**

In most cases, you'll have raw data in a disorganized manner. There won't be any structure to it. In this second step of data wrangling, you have to restructure it for easy accessibility.

- **Cleansing**

Almost every dataset includes some outliers that can skew the outcomes of the analysis. You'll have to clean the data for optimum results. In the third data wrangling step, the data is cleansed exhaustively for superior analysis. You'll have to change null values, remove duplicates and special characters, and standardize the formatting to improve the consistency of the data.

- **Enriching**

After cleansing, your data must be enriched, which means taking stock of what's in the dataset and strategizing how to make it better by adding supplementary data.

- **Validating**

Validation rules in data wrangling denote some repetitive programming steps that are used to authenticate the reliability, quality, and safety of the data you have.

- **Publishing**

For an organization to use the data after the wrangling process has been completed, you have to publish and share the information. This could come in the form of uploading the data to an automation software or storing the file in a location where the organization knows it is ready to be used. It's also a good idea to document the steps taken and logic used in the data wrangling process for future reference.

Literature Review

1) **Data Wrangling: Making data useful again.**
   Author names: Florian Endel and Harald Piringer.
   Date of publication: 19/05/2015
   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.

2) **Data Wrangling: The Challenging Journey from the Wild to the Lake .**
   Author names: Ignacio Terrizzano, Peter Schwarz, Mary Roth, John E. Colino.
   Date of publication: 04/01/2015
   Much has been written about the explosion of data, also known as the "data deluge". The term data lake has been coined to convey the concept of a centralized repository containing virtually inexhaustible amounts of raw (or minimally curated) data that is readily made available anytime to anyone authorized to perform analytical activities.

3) **Research directions in data wrangling: Visualizations and transformations for usable and credible data.**
   Author names: Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche,Chris Weaver, Bongshin Lee, Dominique Brodbeck and Paolo Buono.
   Date of publication: 10/08/2011
   In this article, authors have reviewed the challenges and opportunities associated with addressing data quality issues and also the point that analysts might more effectively wrangle data through new interactive systems that integrate data verification, transformation, and visualization.

4) **Data Wrangling for Big Data: Towards a Lingua Franca for Data Wrangling**
   Author names: Tim Furche, Georg Gottlob, Bernd Neumayr and Emanuel Sallinger.
   Date of publication: 01/01/2016
   The 4 V's of big data: volume (the scale of data), velocity (speed of change), variety (different forms and formats) and veracity (uncertainty) are discussed. These pose challenges for each component of data wrangling itself, but also for the whole system and how sharing knowledge between these components has the most potential of improving the data wrangling process. In this paper the vision and design principles of a data log based language for data wrangling that facilitates such sharing of knowledge is described.

5) **Data Wrangling for Big Data: Challenges and Opportunities**
   Author names: Tim Furche, George Gottlob, Leonid Libkin, Giorgio Orsi and Norman Paton.
   Date of publication: 15/03/16
   This paper argues that providing cost-effective, highly-automated approaches to data wrangling involves significant research challenges, requiring fundamental changes to established areas such as data extraction, integration and cleaning, and to the ways in which these areas are brought together.

Findings

Data wrangling has been recognised as a recurring feature of big data life cycles. Data wrangling has been defined as: **a process of iterative data exploration and transformation that enables analysis.**

In some cases, definitions capture the assumption that there is significant manual effort in the process: **the process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data with the help of semi-automated tools.**

Data wrangling is also defined as a process of iterative data exploration and transformation that enables analysis. One goal is to make data **usable**, to put them in a form that can be parsed and manipulated by analysis tools. Data usability is determined relative to the tools by which the data will be processed; such tools might include spreadsheets, statistics packages, and visualization tools. We say data is **credible** if, according to an analyst's assessment, they are suitably representative of a phenomenon to enable productive analysis.

Ultimately, data is useful if it is usable, credible, and responsive to one's inquiry. In other words, data wrangling is the process of making data useful. Ideally, the outcome of wrangling is not simply data; it is an editable and auditable transcript of transformations coupled with a nuanced understanding of data organization and data quality issues.

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.

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.

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.

Data quality does not only occur in discrete statuses, i.e. clean and faulty. 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 can help to intuitively present uncertainty.

Typical research papers tend to showcase the result of visualizing previously cleaned data, but more often than not neglect to mention how data errors were found and fixed. Ideally, the output of a wrangling session should be more than a clean data set; it should also encompass the raw data coupled with a well-defined set of data operations and potentially some metadata indicating why these operations were performed. These operations should be **auditable** and **editable** by the user. Secondary benefits of a high-level data transformation language include easier reuse of previous formatting efforts and an increased potential for social, distributed collaboration around data wrangling.

The 4 V's of big data refer to some recurring characteristics: Volume represents scale either in terms of the size or number of data sources; Velocity represents either data arrival rates or the rate at which sources or their contents may change; Variety captures the diversity of sources of data, including sensors, databases, files and the deep web; and Veracity represents the uncertainty that is inevitable in such a complex environment. When all 4 V's are present, the use of ETL (extract, transform, load) processes involving manual intervention at some stage may lead to the sacrifice of one or more of the V's to comply with resource and budget constraint.

Data wrangling has long been an elephant in the room of data analysis. Extraordinary amounts of time are spent getting a data set into a shape that is suitable for downstream analysis tools, often exceeding the amount of time spent on the analysis itself. At the same time, all this effort is wasted when the data set changes and may be duplicated by many other analysts looking at the same data. Data cleaning cannot be done by computers alone, as they lack the domain knowledge to make informed decisions on what changes are important. On the other hand, manual approaches are time consuming and tedious. The principled coupling of visual interfaces with automated algorithms provides a promising solution, and we hope visual analytics research will contribute a lasting impact to this critical challenge.

Importance of Good Data Wrangling:

In the simplest of terms, data wrangling is so crucial because it's the only way to make raw data usable. Many a times in practical business setting, customer information or financial information, comes in different pieces from different departments. Sometimes, this information gets stored on various computers across different spreadsheets, and on different systems including legacy systems leading to data duplication, incorrect data or data that can't be found to be used. To create a whole picture of what is happening within a business, it's best to have all data in a centralized location so it can be used. This is just one way in which data automation tools help the data wrangling process along.

Existing Problems

The greatest challenge in data wrangling is its time-intensive nature. This is in part due to the detailed structural nature of the analytic data sets that are produced from the process. As you can see below, each of these structures hold a great amount of information, specific to an individual use case.

Analysis Base Table (ABT): When machine learning is required, ABT is used to help analyse tables of data for patterns, as well as to predict outcomes. Data is sorted into rows with columns that identify information about each entity, like characteristics, history, or how it relates to other entries.

Making sense of the end result: The questions asked of a piece of data depends on its end use. A data wrangler must consider things like which entities are important, whether the data is needed for immediate analysis, or if it will be used to reflect trends over time. If it will be used historically, what is the time period involved? These are the types of questions a data wrangler must clearly define.

Data access: A data wrangler should have direct permission to access the data, if not, more instructions will be required when data is needed. Working around these policy rules can be time-consuming.

Clean data: Different entities that are actually the same must be cleaned, or de-duplicated. For example, a customer named Bob Smith, could be B. Smith. That same customer could have more than one account or a shared account with family members. All these things need to be reconciled before using.

Source data relationships over time: Understanding how data entities are related to one another takes a significant amount of time, effort, and verification. Using a data warehousing model can help sort this out more quickly.

Manual data integration: Some data comes from sources, like hard-copy documents or spreadsheets, which need to be manually added and organized in the system.

Proposed Solution

CREATING A PLAYPEN TO ENABLE PLAY TO PRIORITIZE DATA NEEDED:

This is where a free-format data lake or playpen can really add value. They should be used to enable IT to dump data there with minimal effort, or for insight teams to access potential data sources for one-off extracts to the playpen. Here, analysts or data scientists can have opportunity to play with the data. However, this capability is far more valuable than that sounds. Better language is perhaps "data lab'." Here, the business experts have the opportunity to try use of different potential data feeds and variables within them and to learn which are actually useful/predictive/used for analysis or modelling that will add value. The great benefit of this approach is to enable a lower cost and more flexible way of de-scoping the data variables and data feeds actually required in live systems. Reducing those can radically increase the speed of delivery for new data warehouses or releases of changes/upgrades.

CLEAN, BUT DON'T OVER-SCRUB:

One way to make sure you don't over-scrub is by not over-aggregating data. Retaining a more granular level, at least as an initial pass, can give you enough information to start doing analytics and machine learning. If you're making a dashboard or report, you want that aggregated data; you're looking at broader trends over time. But to do something more predictive, you want more details.

AUTOMATE, BUT AUDIT:

Once the data is clean, and you're trying to match what you know in your database against new information, you can do some measures on which rows exist that also existed in the previous one. If the rows were updated, did they gain cells? Did they lose cells? Are you overall gaining information by this update? And if not, something's probably broken.

Conclusion

In current practice, wrangling often consists of manual editing and reshaping using a general-purpose tool such as Microsoft Excel. Although this approach is feasible for smaller data sets, a great deal of effort is wasted on relatively menial tasks and no audit trails are stored. We expect that this way of working will become increasingly rare in the near future for two reasons. First, in an increasingly data-driven society we need auditable information on the data sets on which we intend to base our decisions. Second, as the number and size of data sets continues to grow, a completely manual approach will become infeasible.

Extraordinary amounts of time are spent getting a data set into a shape that is suitable for downstream analysis tools, often exceeding the amount of time spent on the analysis itself. At the same time, all this effort is wasted when the data set changes and may be duplicated by many other analysts looking at the same data. Data cleaning cannot be done by computers alone, as they lack the domain knowledge to make informed decisions on what changes are important. On the other hand, manual approaches are time consuming and tedious.

Data wrangling is a problem and an opportunity:

• A problem because the 4 V's of big data may all be present together, undermining manual approaches to ETL.

• An opportunity because if we can make data wrangling much more cost effective, all sorts of hitherto impractical tasks come into reach.

References

https://www.astera.com/type/blog/data-wrangling/

https://www.solvexia.com/blog/what-is-data-wrangling-why-its-so-important

https://www.talend.com/resources/data-wrangling/

https://www.insurancethoughtleadership.com/the-challenges-of-data-wrangling/

https://builtin.com/data-science/data-wrangling

Research papers:

https://publik.tuwien.ac.at/files/publik_253619.pdf

http://people.cs.uchicago.edu/~aelmore/class/topics17/wrangling-wild.pdf

https://idl.cs.washington.edu/files/2011-DataWrangling-IVJ.pdf

https://ora.ox.ac.uk/objects/uuid:d061cac9-0b57-4424-bcb0-de053e618a8e/download_file?file_format=pdf&safe_filename=paper20.pdf&type_of_work=Conference

https://www.pure.ed.ac.uk/ws/files/25070478/paper_94.pdf