# Data Preparation: A Survey of Commercial Tools

Article *in* ACM SIGMOD Record · December 2020

DOI: 10.1145/3444831.3444835

2 authors:

Mazhar Hameed
Hasso Plattner Institute

4 PUBLICATIONS 39 CITATIONS

SEE PROFILE

Felix Naumann
Hasso Plattner Institute

302 PUBLICATIONS 8,612 CITATIONS

SEE PROFILE

# Data Preparation: A Survey of Commercial Tools

Mazhar Hameed
Hasso Plattner Institute
University of Potsdam, Germany
mazhar.hameed@hpi.de

Felix Naumann
Hasso Plattner Institute
University of Potsdam, Germany
felix.naumann@hpi.de

## ABSTRACT

Raw data are often messy: they follow different encodings, records are not well structured, values do not adhere to patterns, etc. Such data are in general not fit to be ingested by downstream applications, such as data analytics tools, or even by data management systems. The act of obtaining information from raw data relies on some *data preparation* process. Data preparation is integral to advanced data analysis and data management, not only for data science but for any data-driven applications. Existing data preparation tools are operational and useful, but there is still room for improvement and optimization. With increasing data volume and its messy nature, the demand for prepared data increases day by day.

To cater to this demand, companies and researchers are developing techniques and tools for data preparation. To better understand the available data preparation systems, we have conducted a survey to investigate (1) prominent data preparation tools, (2) distinctive tool features, (3) the need for preliminary data processing even for these tools and, (4) features and abilities that are still lacking. We conclude with an argument in support of automatic and intelligent data preparation beyond traditional and simplistic techniques.

## Keywords

data quality, data cleaning, data wrangling

## 1. THE NEED FOR DATA PREPARATION

Raw data appears in many situations: logs, sensor output, government data, medical research data, climate data, geospatial data, etc. It accumulates in many places, such as file systems, data lakes or online repositories. In typical scenarios, raw data from various sources is accrued without any standardized formats or structure and with no specific target use-case; thus, it can appear messy, contain invalid characters, use different encodings, lack necessary columns, contain unwanted rows, have missing values, not follow valid patterns, etc.

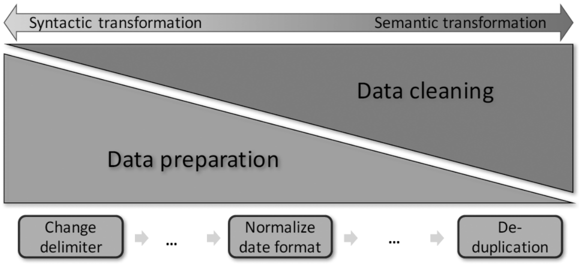


Figure 1: Data preparation vs. data cleaning

We define data preparation as the set of preprocessing operations performed in early stages of a data processing pipeline, i.e., data transformations at the structural and syntactical levels. We provide many examples of such transformations throughout the article. In contrast, data cleaning concerns subsequent data transformations and corrections at the semantic level (Figure 1).

One example scenario in need of data preparation technology are data lakes to store heterogeneous raw data [20]. They can turn into vast repositories or collections of unstructured, semi-structured, unformulated, messy, and unclean data. The large volumes of data in data lakes are compelling and can generate valuable information, provided they are thoroughly pre-processed, cleaned, and prepared [29]. With the ever-increasing amount of raw data, the need for data preparation has become more apparent.

Preparing data yields many advantages, such as prompt error detection, improved analytics, improved data quality, enhanced scalability, accelerated data usage and more easy data collaboration [9, 12].

### 1.1 The data-to-application process

To understand data processing in the data to application life-cycle, it is important to identify the phases required to create data that is valid for the consuming application. Data creation occurs typically in raw format, possibly to be stored in data lakes. Before these raw data are sent to applications it is crucial to enhance its structure and, if needed, its content: To make data readable and machine understandable, a number of steps are typically performed, such as (1) data exploration [5, 14, 26, 27], (2) data collection [18, 31], (3) data profiling [8, 21, 22], (4) data preparation [3, 12, 25], (5) data integration [7, 17, 28], and (6) data cleaning [2, 4, 24] in various orders and iterations.

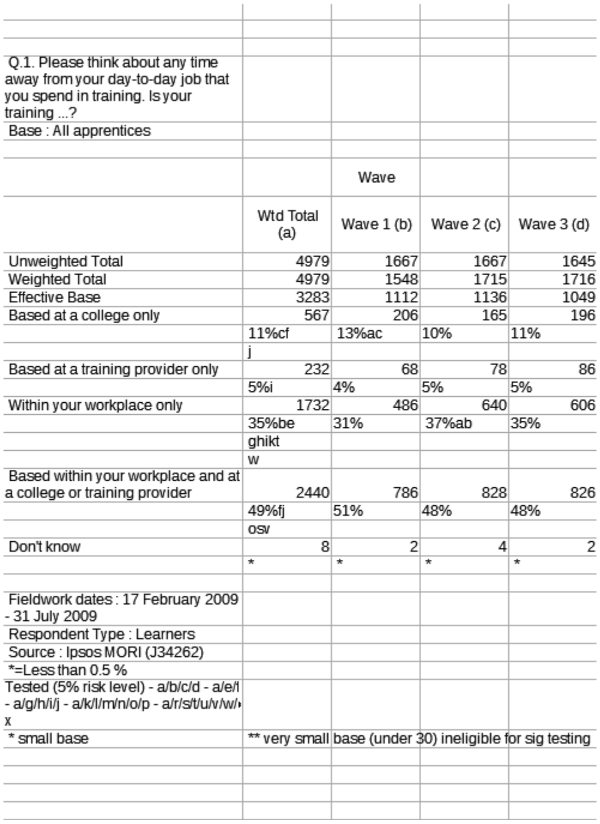
These aforementioned steps are applied to originally 'raw data', before they are sent to the main application for further processing. In our research focus, and based on evidence from noted surveys, a critical and important step is data preparation. Trifacta's data preparation study shows that 72% of respondents indicated that data preparation by data users is critical, while 88% indicated at least its importance, and only 4% indicated that it is not important for the user [1]. Data scientists spend approximately 80% of the time on preparing the data and about 20% on actual model implementation and deployment [12,23,29]. Clearly, these numbers cannot be reduced to 0%, due to the semantic difficulties of understanding and interpreting data. However, the time spent on data preparation can be decreased to a significant amount using sophisticated data preparation techniques, and, in turn, data scientists attain more time for model implementation and deployment.

Keeping in mind the importance and impact of data preparation, developers and researchers have contributed various techniques that facilitate the data preparation process [9–11, 15, 29, 32].

To address the aforementioned challenges and the general importance of data preparation, many tools have been designed by not only the industry, but also by research and academia to address varying use cases. In light of that, we have surveyed commercial data preparation tools to analyze available features and methods. Our survey is not comparative, nor do we explicitly evaluate the individual tools. Rather, we want to show the current state of the art and identify research and development opportunities for the data preparation community at large.

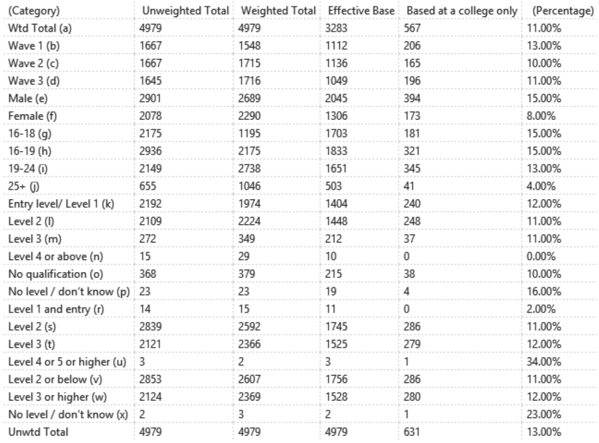
## 1.2 A preparation example

Let us explain, with the help of an example, the usefulness and importance of data preparation. For instance a data scientist has been handed a csv-formatted file as shown in Figure 2a, from a government data portal[1] to examine and answer how much time each employee is spending on training besides their 9-5 job. Although the csv format defines rows of records, once the file is opened it is clear that there is no coherent relational structure: data is laid out in a somewhat human-readable format, making automated analysis impossible. More-



| | | | | |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| Q.1. Please think about any time away from your day-to-day job that you spend in training. Is your training ...? | | | | |
| Base : All apprentices | | | | |
| | | | | |
| | | Wave | | |
| | Wtd Total (a) | Wave 1 (b) | Wave 2 (c) | Wave 3 (d) |
| Unweighted Total | 4979 | 1667 | 1667 | 1645 |
| Weighted Total | 4979 | 1548 | 1715 | 1716 |
| Effective Base | 3283 | 1112 | 1136 | 1049 |
| Based at a college only | 567 | 206 | 165 | 196 |
| | 11%cf | 13%ac | 10% | 11% |
| | j | | | |
| Based at a training provider only | 232 | 68 | 78 | 86 |
| | 5%i | 4% | 5% | 5% |
| Within your workplace only | 1732 | 486 | 640 | 606 |
| | 35%be | 31% | 37%ab | 35% |
| | ghikt | | | |
| | w | | | |
| Based within your workplace and at a college or training provider | 2440 | 786 | 828 | 826 |
| | 49%fj | 51% | 48% | 48% |
| | osv | | | |
| Don't know | 8 | 2 | 4 | 2 |
| | * | * | * | * |
| | | | | |
| Fieldwork dates : 17 February 2009 - 31 July 2009 | | | | |
| Respondent Type : Learners | | | | |
| Source : Ipsos MORI (J34262) | | | | |
| *=Less than 0.5 % | | | | |
| Tested (5% risk level) - a/b/c/d - a/e/f - a/g/h/i/j - a/k/l/m/n/o/p - a/r/s/t/u/v/w/ x | | | | |
| * small base | ** very small base (under 30) ineligible for sig testing | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

(a) Unprepared data



| (Category) | Unweighted Total | Weighted Total | Effective Base | Based at a college only | (Percentage) |
|---|---|---|---|---|---|
| Wtd Total (a) | 4979 | 4979 | 3283 | 567 | 11.00% |
| Wave 1 (b) | 1667 | 1548 | 1112 | 206 | 13.00% |
| Wave 2 (c) | 1667 | 1715 | 1136 | 165 | 10.00% |
| Wave 3 (d) | 1645 | 1716 | 1049 | 196 | 11.00% |
| Male (e) | 2901 | 2689 | 2045 | 394 | 15.00% |
| Female (f) | 2078 | 2290 | 1306 | 173 | 8.00% |
| 16-18 (g) | 2175 | 1195 | 1703 | 181 | 15.00% |
| 16-19 (h) | 2936 | 2175 | 1833 | 321 | 15.00% |
| 19-24 (i) | 2149 | 2738 | 1651 | 345 | 13.00% |
| 25+ (j) | 655 | 1046 | 503 | 41 | 4.00% |
| Entry level/ Level 1 (k) | 2192 | 1974 | 1404 | 240 | 12.00% |
| Level 2 (l) | 2109 | 2224 | 1448 | 248 | 11.00% |
| Level 3 (m) | 272 | 349 | 212 | 37 | 11.00% |
| Level 4 or above (n) | 15 | 29 | 10 | 0 | 0.00% |
| No qualification (o) | 368 | 379 | 215 | 38 | 10.00% |
| No level / don't know (p) | 23 | 23 | 19 | 4 | 16.00% |
| Level 1 and entry (r) | 14 | 15 | 11 | 0 | 2.00% |
| Level 2 (s) | 2839 | 2592 | 1745 | 286 | 11.00% |
| Level 3 (t) | 2121 | 2366 | 1525 | 279 | 12.00% |
| Level 4 or 5 or higher (u) | 3 | 2 | 3 | 1 | 34.00% |
| Level 2 or below (v) | 2853 | 2607 | 1756 | 286 | 11.00% |
| Level 3 or higher (w) | 2124 | 2369 | 1528 | 280 | 12.00% |
| No level / don't know (x) | 2 | 3 | 2 | 1 | 23.00% |
| Unwtd Total | 4979 | 4979 | 4979 | 631 | 13.00% |

(b) Prepared data

Figure 2: Example of data preparation

[1] http://webarchive.nationalarchives.gov.uk/+/http://www.bis.gov.uk/assets/biscore/further-education-skills/docs/n/11-708-data-nlss-2009.csv (February, 2019)

over in this case, almost 1,000 tables are stacked one below each other (not shown), interleaved by metadata information in the form of preambles and comments that more often than not repeat themselves without meaningful addition. On top of that, inside the actual data tables, alphanumeric characters appear in what seem to be other-wise numeric records, and there are apparently inconsistent representations for zeros/null values (e.g., '*','-' or empty cell). The data scientist might perform the sequence of steps listed below to each file making their data more comprehensive, prepared, structured and machine-readable as depicted in Figure 2b., before feeding them to the analysis tool. In this way, the data scientist avoids the cumbersome and time-consuming manual execution of these tasks and could use this sequence again for future use cases and tasks.

1. Split the file to isolate one data table at a time. For each obtained file:
2. Remove preamble and comment rows.
3. Unify null-value representations.
4. Remove rows with no meaningful information, e.g., empty rows or rows with only null-values.
5. Clean numeric data rows by removing special characters.
6. Fill missing values, e.g., by value imputation or using functional dependencies.
7. Transpose table.
8. Add missing header.

It is evident from the aforementioned example that with the help of various data preparation steps and tools we were able to target messy data and convert it into clean and machine-readable data, highlighting the significance of data preparation in the market for both industry and academia. The application of simple data preparation tasks on raw data files improves their usability, readability, interpretability, etc. Software vendors have identified the importance and need of data preparation and offer dedicated tools. To provide a snapshot of the current state of development, we have conducted a detailed survey of seven commercial data preparation tools. Our paper makes the following contributions:

1. **Organisation**: We propose six broad categories of data preparation and identify 40 common data preparation steps, which we classify into those categories (Section 2).
2. **Documentation**: We validate the availability of these features and broader categories for seven selected tools and document them in a feature matrix (Section 3).
3. **Evaluation**: We evaluate the selected features of surveyed tools to identify whether the tool offers the stated functionalities or not (Section 4).
4. **Recommendation**: We identify shortcomings of commercial data preparation tools in general and encourage researchers to explore further in the field of data preparation (Section 5).

## 2. DATA PREPARATION TASKS

Data preparation is not a single step process. Rather, it usually comprises many individual preparation steps, implemented by what we call *preparators*, and which we have organized anew into six broader categories, defined here.

**Data discovery** is the process of analyzing and collecting data from different sources, for instance to match data patterns, find missing data, and locate outliers.

**Data validation** comprises rules and constraints to inspect the data, for instance for correctness, completeness, and other data quality constraints.

**Data structuring** encompasses tasks for the creation, representation and structuring of information. Examples include updating schema, detecting & changing encoding and, transform data by example [13].

**Data enrichment** adds value or supplementary information to existing data from separate sources [30]. Typically, it involves augmenting existing data with new or derived data values using data lookups, primary key generation, and inserting metadata.

**Data filtering** generates a subset of the data under consideration, facilitating manual inspection and removing irregular data rows or values. Examples include extracting text parts, and keeping or deleting filtered rows.

**Data cleaning** refers to removal, addition, or replacement of less accurate or inaccurate data values with more suitable, accurate or representative values. Typical examples are deduplication, fill missing values, and removing whitespace.

Despite our definition, which distinguishes data preparation and cleaning, we include data cleaning steps here as well, as most data preparation tools also venture into this area.

Our set of 40 individual preparators is shown and categorized in Table 2, which is introduced in the next section.

## 3. PREPARATION TOOLS AND TASKS

Data preparation tools are vital to any data preparation process. They usually provide implementations of various preparators and a frontend to sequentially apply preparations or to specify data preparation pipelines. The flexibility, robustness and intelligence of these tools contribute significantly towards the data analysis and data management tasks. In this section, we discuss in detail a selection of tools for our research study that are supported by supplementary documentation for experimentation and guidance. Section 3.1 discusses the selected data preparation tools (see Table 1 for an overview) and Section 3.2 highlights our approach to populate the preparator matrix (Table 2), organized by data preparator categories with selected preparation tasks.

### 3.1 Available data preparation tools

In general, data preparation is an expensive and time-consuming activity, especially without automated and mature data preparation tools. Traditionally, data scientists write specific preparation scripts to accomplish the project-specific goals. Recently, the market has answered to some of the general needs of data preparation by providing commercial preparation tools that can lower the burden of data scientists.

To better understand commercial tools and their capabilities, we initiated our study with a discovery phase. We collected notable commercial data preparation tools gathered from business reports and analyses, company portals, and online demonstration videos. Our preliminary investigation resulted in 42 initial commercial tools (shown in Table 3 in the appendix), which we then examined for the extent of their *data preparation* capabilities.

Not all collected tools were dedicated to data preparation. Rather, many tools were primarily targeting data visualization, data analysis, and business intelligence applications, with only some added data preparation features. To focus on the topic of our survey, we established, necessarily soft, criteria for tool selection.

- Domain specificity: tools that specifically address the data preparation task.
- Comprehensiveness: the extent and sophistication to which tools adequately covered preparation features listed in Section 2.
- Guides and documentation: the availability of proper documentation for the tools, i.e., useful, up-to-date documentation with listings of features and how-to guides
- Trial availability: the availability of a trial version, giving us the opportunity to test the tools and validate their features
- GUI: the availability of a comprehensive and intuitive graphical user interface to select and apply preparations.
- Customer assistance: compliant support teams that assisted users with generic and specific tool queries, when needed.

Finally, we selected seven tools for detailed investigation (shown in Table 1). We now discuss (in alphabetical order) the seven qualifying tools for our data preparation survey. In the appendix we have collected additional functional and non-functional features that are not specific data preparation tasks.

**Altair Monarch Data Preparation**, called Datawatch until the company's merger with Altair, provides common data preparators for structured data but also transforms tables from within PDF and text files to tabular data. The extracted files from Altair's table extractor feature can be used independently as a table or they can be merged with other tables or files using a variety of join and union operations.

**Paxata Self-Service Data Preparation** offers many features to organize and prepare structured data and also deals efficiently with semi-structured data. In addition to common data preparation features, Paxata offers so-called data filtergrams, which allow various visual interactions to perform filter operations on data, such as, text filtergrams, numeric filtergrams, Boolean filtergrams, and source filtergrams. The user experience is emphasized in this tool, which is designed to support also non-experts.

**SAP Agile Data Preparation** runs on top of SAP's HANA database system. It offers many common data preparators with some specific system features, such as Schedule Snapshot, which allows the user to take periodic snapshots and retrieve data from a remote source on demand. It offers interactive suggestions to help users navigate and prepare data efficiently. Multi-user access allows to prepare data in collaboration.

**SAS Data Preparation** is part of SAS Viya System Management, which runs its operations with

Table 1: Selected data preparation tools

| Tool name | URL |
|---|---|
| Altair Monarch Data Preparation | https://www.datawatch.com/in-action/monarch-draft/ |
| Paxata Self Service Data Preparation | https://www.paxata.com/self-service-data-prep/ |
| SAP Agile Data Preparation | https://www.sap.com/germany/products/data-preparation.html |
| SAS Data Preparation | https://www.sas.com/en_us/software/data-preparation.html |
| Tableau Prep | https://www.tableau.com/products/prep |
| Talend Data Preparation | https://www.talend.com/products/data-preparation/ |
| Trifacta Wrangler | https://www.trifacta.com/products/wrangler-editions/ |

distributed in-memory processing. In addition to common features, SAS offers code-based transformations for users to write and share custom code to transform data, supporting re-usability of preparation pipelines.

**Tableau Prep** implements a workflow approach to organize and prepare messy data. With its interactive interface and workspace plans, users have the freedom to perform multiple operations simultaneously. Tableau prep comprises two parts, namely Tableau Prep Builder, which is designed to develop so-called flows, manage data and apply operations on data, and Tableau Prep Conductor to share, schedule, and monitor the flows.

**Talend Data Preparation** offers many specific data preparation functionalities tailored to the task at hand. For instance, for data cleaning, different functions exist for cleaning numeric data values, strings and date inputs. One of its main features is "selective sampling" of data for insights and operations that can be later deployed on the entire dataset. Talend actively contributes to solving system-level challenges, e.g., one of its intelligent system features is pipeline automation, to save and reuse data preparation tasks or steps.

**Trifacta Wrangler** prepares data using multiple data preparation functions and intelligently predicts patterns to provide suggestions that help users to transform data. Apart from common preparation tasks, it offers additional interesting features, such as primary key generation, transform data by example, and permitted character checks. Wrangler uses regular expressions for most of its pattern-based features. The significance of Wrangler preparators is their degree of sophistication. For example, the locate outlier not only identifies the outliers, but also plots a histogram of the entire column. The tool was spun out of the Wrangler project [16].

## 3.2 Preparator matrix

Table 2 provides a feature matrix showing which preparator is supported by which tool in each of the six categories. We evaluated each of these preparators on three datasets downloaded from public data repositories: (i) Kaggle – 120 years of Olympic history (athletes and results)[2], (ii) IMDb – data about movies[3], and (iii) UK government web archive, as mentioned in Section 1.2.

The population of this preparator matrix was not a trivial task. Initially, we analyzed the tool's documentation to gather all available preparators. We then downloaded trial versions of all tools and (generously) evaluated for each of the seven tools and each of the 40 preparators whether they offer this functionality. Section 4 describes in more detail how we populated the feature matrix. All tools and their corresponding documented preparators were gathered before September 2, 2019.

The basic functionality of most preparators is self-explanatory by their name – their precise implementation and parameterization might differ from tool to tool and it would be beyond the scope of this article to describe each. Instead, we have selected three exemplary preparators to illustrate their function and the intricacies involved in even simple data preparation tasks. We use the same three preparators in Section 4 to highlight some capabilities of individual tools.

Keep or Delete Filtered Rows: Filtering operations customize data views and provide output based on specified predicates, for instance to filter data that can be deleted, extracted or altered for further analysis. In its basic form, filtering allows simple predicates, akin to SQL conditions. A more intelligent approach would be to use a richer language, such as regular expressions, for filtering.

Value Standardization: A typical preparation operation is to change the values of a column to follow some standard. That standard could be a frequent pattern derived from the data itself or taken from an external authority. A more sophisticated preparator could help in automatically detecting relevant data clusters for standardization. Popular techniques include fuzzy matching for clustering to

[2] https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results
[3] ftp://ftp.fu-berlin.de/pub/misc/movies/database/frozendata/

provide a better representation of data.

Split Column: Messy data can include values that consist of multiple atomic parts. Split column can separate (split) data into multiple columns based on defined criteria (e.g., split after ',' or at last whitespace in string, etc). A more sophisticated preparator could identify split column cases by using existing patterns in data, and be able to handle splits into more than two columns.

## 4. EVALUATION OF EXISTING TOOLS

To better explain how we evaluated the preparators, we provide an example for each of the three preparators discussed in the previous section. In general, even the simplest versions of the respective preparators earned the tool a checkmark in our matrix (Table 2). More sophisticated versions could incorporate preparators that intelligently detect relevant problems and actively provide suggestions for their configuration, e.g., suitable regular expressions or standard formats.

Keep or Delete Filtered Rows: Data filtering techniques improve data quality using predefined criteria, such as removing records that contain empty values or that do not conform to some user-defined pattern. The majority of data preparation tools offers various types of filters. For instance, Talend Data Preparation offers filters based on patterns using pre-defined syntactic data types:

EXAMPLE 1. *Using pattern filtering, a user might want to keep only official email addresses. Using Talend's syntax, corresponding patterns might be:*
`[word]@ibm.[word],[char].[word]@ibm.[word]`
*Thus, private addresses such as* *`bob1992@gmail.com`* *or* *`alice25@yahoo.com`* *would be filtered, while* *`a.peter@ibm.com`* *would be retained.*

Value Standardization: A typical step in case of heterogeneously formatted values is standardization using patterns, e.g., phone number patterns, datetime patterns, patterns by example, etc. For instance, Trifacta Wrangler provides suggestions for applicable patterns and transforms data to the suggested or a selected standard.

Also, in case of different representations of the same real-world value within a column, value standardization groups those values and transforms them to a single, common representation.

EXAMPLE 2. *Trifacta might group records with city values* `NY`*,* `NYC`*,* `New York` *and* `New York City` *and standardize all occurrences to* `New York City`*. Alternatively, users can review the cluster and manually choose the correct standard value.*

Split Column: Multi-valued columns reduce flexibility in handling data (and also their readability). Split column splits such columns based on some criterion. For instance, SAS Data Preparation implements this technique in several ways, e.g., split based *on*, *before*, or *after* a delimiter, on a fixed length, and "quick split", which intelligently identifies a split criterion.

EXAMPLE 3. *Using comma as a delimiter, the user wants to split the location column (and implicitly trim accrued whitespace). In addition, the user specified headers for the output columns. As can be seen in the example, due to a missing value in the original data, the value "USA" is misplaced; a later validation step might identify this error.*

*Input:*

| ***Location*** | ... |
|---|---|
| *Melbourne, Victoria, Australia* | |
| *San Francisco, USA* | |
| *Potsdam, Brandenburg, Germany* | |

*Output:*

| ***City*** | ***State*** | ***Country*** | ... |
|---|---|---|---|
| *Melbourne* | *Victoria* | *Australia* | |
| *San Francisco* | *USA* | | |
| *Potsdam* | *Brandenburg* | *Germany* | |

## 5. CHALLENGES AND FUTURE WORK

Some of the most prominent challenges that we came across during our research and survey are the following:

**Dataset pre-processing**: Interestingly, despite being data preparation tools, all tools that we have surveyed and explored require a pre-prepared or cleaned dataset as their input. For example, if the file had comment-lines, additional header or footer information, or poorly placed quotation marks, it was misinterpreted and loaded improperly. In fact, most tools make the following broad assumptions:

- Single table file (no multi-table files)
- Specific file encoding
- No preambles, comments, footnotes, etc.
- No intermediate headers
- Specific line-ending symbol
- Homogeneous delimiters
- Homogeneous escape symbols
- Same number of fields per row

Table 2: Data preparation tool feature matrix

| **Categories** | **Available features** | **Data preparation tools** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Altair | Paxata | SAP | SAS | Tableau | Talend | Trifacta |
| Data discovery | Locate missing values (nulls) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Locate outliers | | ✓ | | ✓ | | | ✓ |
| | Search by pattern | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Sort data | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Data validation | Compare values (selection and join) | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Check data range | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Check permitted characters | | | | | | | ✓ |
| | Check column uniqueness | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Find type-mismatched data | | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Find data-mismatched datatypes | | ✓ | | | | ✓ | ✓ |
| Data structuring | Change column data type | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Delete column | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Detect & change encoding | | | | | | ✓ | ✓ |
| | Pivot / unpivot | ✓ | ✓ | ✓ | | ✓ | | ✓ |
| | Rename column | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Split column | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Transform by example [13] | | | | | | ✓ | ✓ |
| Data enrichment | Assign semantic data type | | | | ✓ | ✓ | ✓ | |
| | Calculate column using expressions | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Discover & merge external data | ✓ | ✓ | ✓ | | | ✓ | ✓ |
| | Duplicate column | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Generate primary key column | | | ✓ | | | | ✓ |
| | Join & union | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Merge columns | ✓ | | ✓ | | ✓ | ✓ | ✓ |
| | Normalize numeric values | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Data filtering | Delete/keep filtered rows | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Delete empty and invalid rows | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Extract value parts | ✓ | | | ✓ | | ✓ | ✓ |
| | Filter with regular expressions | | | | | | | ✓ |
| Data cleaning | Change date & time format | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Change letter case | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Change number format | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Deduplicate data | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ |
| | Delete by pattern | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ |
| | Edit & replace cell data | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Fill empty cells | ✓ | ✓ | | | | ✓ | ✓ |
| | Remove extra whitespace | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Remove diacritics | | | ✓ | | | | |
| | Standardize strings by pattern | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | Standardize values in clusters | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

- Relational data (no nested or graph-structured data, such as XML, JSON or RDF)

Some of the aforementioned assumptions pose interesting research problems in themselves, which have been addressed in isolation by other researchers, such as detecting tables in complex spreadsheets [6] or converting HTML tables to relations [19].

**User expertise needed**: Another challenge we experienced was the need of the combination of domain knowledge and IT-knowledge for tool usability. Most tools require the user to be an expert in the dataset domain and have prior knowledge and understanding of the datasets and of the data preparation goal.

Moreover, beyond simple predicates, most tools allow the use of regular expressions to match, split, or delete data. A typical domain-expert cannot be expected to formulate often intricate regular expressions.

**Lack of intelligent solutions**: All surveyed tools offer useful data preparation functions. However, most tools and most preparators lack intelligent solutions for more automated data preparation tasks. For example Deduplicate data removes duplicate records from a source. The surveyed tools deduplicate data only on exact match conditions, a more sophisticated version would involve deduplication based on similarity measures. Another problem for many tools is column heterogeneity, i.e., if columns contain data in multiple formats. Currently, users need to manually filter those different groups and prepare them separately. An automatic homogenization would be helpful but also poses a challenging research problem.

**Unstructured data**: The scope of our survey is that of preparing structured data. However, many datasets include some textual component, such as product descriptions, plot synopses, etc. Such textual data can also benefit from basic preparation steps, such as stopword-removal, lemmatization, or sentence breaking, to then e.g. perform named entity extraction or sentiment analysis.

One outlook is to include such capabilities in the existing tools for structured data preparation. Another is to develop a dedicated framework and toolset for the case of unstructured data preparation (or text preparation), similar to the tools survey in this article.

**Preparation pipelining**: Data preparation is not a one-step process. Rather, it involves many subsequent steps, organized in a preparation pipeline to gradually transform a dataset towards the desired output.

Creating and managing pipelines yields many system-level challenges and opportunities. For instance, preparation suggestion, pipeline adaption, and pipeline optimization, that need to be addressed accordingly. Such systematic data preparation can benefit from a comprehensive and well-defined yet extensible set of operators. By incorporating the ability to create and manage preparation pipelines, data preparation tools can be massively improved and generalized for more intelligent and self-service techniques. After a pipeline has been established, optimization and customization policies can be designed according to needs of the problem at hand or business use cases under considerations.

To summarize, existing tools already cover basic data preparation needs by implementing simple and obvious preparators. In few cases we observed more sophisticated abilities, such as automatic suggestion of patterns or even of preparators for the data at hand. All of these tools are excellent platforms for further development in several dimensions, as outlined above. In our opinion, the need for self-service data preparation and tool capabilities goes beyond current technology and we encourage research in this emerging field.

# 6. CONCLUSION

In this paper, we have discussed and surveyed major commercial tools for data preparation. We have gathered and organized their capabilities in the form of “preparators”, organized in six categories.

As more and more data are produced there is more and more opportunity to create value by integrating and analyzing them. Thus, the need for data preparation and data cleaning grows: Data have many types of syntactic and semantic issues that can be bridged by careful automated or manual preparation and cleaning steps. Current technology is still far from enabling a fully automatic transformation of data from their raw form to a shape and quality that can be readily consumed by downstream applications. Commercial (and academic) tools provide good user-support and tooling for a wide range of preparation needs. Nevertheless, data preparation remains a largely manual task to be performed by data experts or by domain experts with data engineering skills.

## Acknowledgments

This research was funded by the HPI research school on Data Science and Engineering.

# 7. REFERENCES

[1] Trifacta end user data preparation. `https://www.trifacta.com/wp-content/uploads/2018/02/End-User-Data-Preparation-Market-Study-2018.pdf`. Accessed: 2019-09-19.

[2] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. Detecting data errors: Where are we and what needs to be done? *PVLDB*, 9(12):993–1004, 2016.

[3] Gregorio Convertino and Andy Echenique. Self-service data preparation and analysis by business users: New needs, skills, and tools. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1075–1083. ACM, 2017.

[4] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F Ilyas, Mourad Ouzzani, and Nan Tang. Nadeef: a commodity data cleaning system. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 541–552. ACM, 2013.

[5] Yanlei Diao, Kyriaki Dimitriadou, Zhan Li, Wenzhao Liu, Olga Papaemmanouil, Kemi Peng, and Liping Peng. Aide: an automatic user navigation system for interactive data exploration. *PVLDB*, 8(12):1964–1967, 2015.

[6] Haoyu Dong, Shijie Liu, Shi Han, Zhouyu Fu, and Dongmei Zhang. TableSense: Spreadsheet table detection with convolutional neural networks. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, volume 33, pages 69–76, 2019.

[7] Xin Luna Dong and Divesh Srivastava. Big data integration. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1245–1248. IEEE, 2013.

[8] Jens Ehrlich, Mandy Roick, Lukas Schulze, Jakob Zwiener, Thorsten Papenbrock, and Felix Naumann. Holistic data profiling: Simultaneous discovery of various metadata. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, pages 305–316, 2016.

[9] Florian Endel and Harald Piringer. Data wrangling: Making data useful again. *IFAC-PapersOnLine*, 48(1):111–112, 2015.

[10] Tim Furche, Georg Gottlob, Leonid Libkin, Giorgio Orsi, and Norman W Paton. Data wrangling for big data: Challenges and opportunities. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, pages 473–478, 2016.

[11] Anders Haug, Frederik Zachariassen, and Dennis Van Liempd. The costs of poor data quality. *Journal of Industrial Engineering and Management (JIEM)*, 4(2):168–193, 2011.

[12] Joseph M Hellerstein, Jeffrey Heer, and Sean Kandel. Self-service data preparation: Research to practice. *IEEE Data Engineering Bulletin*, 41(2):23–34, 2018.

[13] Zhongjun Jin, Michael R Anderson, Michael Cafarella, and HV Jagadish. Foofah: Transforming data by example. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 683–698. ACM, 2017.

[14] Manas Joglekar, Hector Garcia-Molina, and Aditya G Parameswaran. Interactive data exploration with smart drill-down (extended version). *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, (1):1–1, 2017.

[15] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank Van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[16] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. Wrangler: interactive visual specification of data transformation scripts. In *Proceedings of the International Conference on Human Factors in Computing Systems (CHI)*, pages 3363–3372, 2011.

[17] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the Symposium on Principles of Database Systems (PODS)*, pages 233–246. ACM, 2002.

[18] Yanying Li, Haipei Sun, Boxiang Dong, and Hui Wendy Wang. Cost-efficient data acquisition on online data marketplaces for correlation analysis. *PVLDB*, 12(4):362–375, 2018.

[19] George Nagy, Sharad Seth, and David W. Embley. End-to-end conversion of HTML tables for populating a relational database. In *Proceedings of the IAPR International Workshop on Document Analysis Systems*, pages 222–226, 2014.

[20] Fatemeh Nargesian, Erkang Zhu, Renée J Miller, Ken Q Pu, and Patricia C Arocena. Data lake management: challenges and opportunities. *PVLDB*, 12(12):1986–1989, 2019.

[21] Felix Naumann. Data profiling revisited. *SIGMOD Record*, 42(4):40–49, 2014.

[22] Thorsten Papenbrock, Tanja Bergmann, Moritz Finke, Jakob Zwiener, and Felix Naumann. Data profiling with Metanome. *PVLDB*, 8(12):1860–1863, 2015.

[23] Gil Press. Cleaning data: Most time-consuming, least enjoyable data science task. *Forbes*, March 2016.

[24] Vijayshankar Raman and Joseph M Hellerstein. Potter's wheel: An interactive data cleaning system. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, 2001.

[25] Tye Rattenbury, Joseph M Hellerstein, Jeffrey Heer, Sean Kandel, and Connor Carreras. *Principles of data wrangling: Practical techniques for data preparation*. O'Reilly Media, Inc., 2017.

[26] Thibault Sellam and Martin Kersten. Ziggy: Characterizing query results for data explorers. *PVLDB*, 9(13):1473–1476, 2016.

[27] Tarique Siddiqui, Albert Kim, John Lee, Karrie Karahalios, and Aditya Parameswaran. Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. *PVLDB*, 10(4):457–468, 2016.

[28] Michael Stonebraker and Ihab F Ilyas. Data integration: The current status and the way forward. *IEEE Data Engineering Bulletin*, 41(2):3–9, 2018.

[29] Ignacio G Terrizzano, Peter M Schwarz, Mary Roth, and John E Colino. Data wrangling: The challenging journey from the wild to the

lake. In *Proceedings of the Conference on Innovative Data Systems Research (CIDR)*, 2015.

[30] Pei Wang, Yongjun He, Ryan Shea, Jiannan Wang, and Eugene Wu. Deeper: A data enrichment system powered by deep web. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 1801–1804. ACM, 2018.

[31] Susan C Weller and A Kimball Romney. *Systematic data collection*, volume 10. Sage publications, 1988.

[32] Shichao Zhang, Chengqi Zhang, and Qiang Yang. Data preparation for data mining. *Applied artificial intelligence*, 17(5-6):375–381, 2003.

## Appendix

Our initial survey found 42 software tools that asserted some for of data preparation functionality. These tools are listed in Table 3. Section 3.1 describes our selection process to reach the seven tools (in bold) that we analyzed more closely.

In our survey of commercial tools, we came across many functional and non-functional system features that did not cater to our data preparation focus. Nonetheless, these features are important and interesting when explored and utilized. Thus, we have gathered them in Table 4.

Table 3: Discovered tools with asserted data preparation capabilities

| Tool name | URL |
|---|---|
| **Altair Monarch Data Preparation** | https://www.datawatch.com/in-action/monarch-draft/ |
| Alteryx Data Preparation | https://www.alteryx.com/solutions/analytics-need/data-preparation |
| BigGorilla Data Preparation | https://www.biggorilla.org/ |
| Cambridge Semantics Anzo | https://www.cambridgesemantics.com/ |
| Datameer | https://www.datameer.com/ |
| EasyMorph Data Preparation and Automation | https://easymorph.com/ |
| Erwin | https://erwin.com/ |
| FICO | https://www.fico.com/ |
| Google Cloud Data Prep by Trifacta | https://cloud.google.com/dataprep/ |
| Hitachi-Pentaho Business Analytics | https://www.hitachivantara.com/en-us/products/data-management-analytics.html |
| IBM Data Refinery | https://www.ibm.com/cloud/data-refinery |
| INFOGIX | https://www.infogix.com/data3sixty/analyze/ |
| Informatica Enterprise Data Preparation | https://www.informatica.com/products/data-catalog/enterprise-data-prep.html |
| Looker | https://looker.com/ |
| Lore IO | https://www.getlore.io/ |
| Microsoft Power BI | https://powerbi.microsoft.com/en-us/ |
| MicroStrategy | https://www.microstrategy.com/us/product/analytics/data-visualization |
| Modak-nabu | https://modakanalytics.com/nabu.html |
| OpenRefine | http://openrefine.org/ |
| Oracle Analytics Cloud | https://www.oracle.com/business-analytics/analytics-cloud.html |
| **Paxata Self Service Data Preparation** | https://www.paxata.com/self-service-data-prep/ |
| Qlik Data Catalyst | https://www.qlik.com/us/products/qlik-data-catalyst |
| Quest Toad Data Point | https://www.quest.com/products/toad-data-point/ |
| Rapid Insight | https://www.rapidinsight.com/solutions/data-preparation/ |
| RapidMiner Turbo Prep | https://rapidminer.com/products/turbo-prep/ |
| **SAP Agile Data Preparation** | https://www.sap.com/germany/products/data-preparation.html |
| **SAS Data Preparation** | https://www.sas.com/en_us/software/data-preparation.html |
| Smarten Advanced Data Discovery | https://www.smarten.com/self-serve-data-preparation.html |
| Solix Common Data Platform | https://www.solix.com/products/solix-common-data-platform/ |
| Sparkflows | https://www.sparkflows.io/data-science |
| **Tableau Prep** | https://www.tableau.com/products/prep |
| **Talend Data Preparation** | https://www.talend.com/products/data-preparation/ |
| Tamr | https://www.tamr.com/ |
| Teradata Vantage | https://www.teradata.com/Products/Software/Vantage |
| TIBCO Spotfire Analytics | https://www.tibco.com/products/tibco-spotfire |
| TMMData | https://www.tmmdata.com/ |
| **Trifacta Wrangler** | https://www.trifacta.com/products/wrangler-editions/ |
| Unifi Data Platform | https://unifisoftware.com/platform/ |
| Waterline Data | https://www.waterlinedata.com/ |
| Workday-Prism Analytics | https://www.workday.com/en-us/applications/analytics/prism-analytics.html |
| Yellowfin Data Prep | https://www.yellowfinbi.com/suite/data-prep |
| Zoho Analytics | https://www.zoho.com/analytics/ |

Table 4: Further features of data preparation tools

| **Altair** | **Paxata** | **SAP** | **SAS** | **Tableau** | **Talend** | **Trifcata** |
|---|---|---|---|---|---|---|
| Advanced Filtering | Add Comments | Action History | Create Custom Code | Adjust Sample Size | Advanced Filtering | Add Comments |
| Audit User Actions | Advanced Filtering | Advanced Filtering | Data Lineage | Advanced Filtering | Aggregation Using Charts | Advanced Filtering |
| Comparison Functions | Aggregation | Aggregation | Data Sampling | Aggregation | Audit User Actions | Aggregation |
| Copy and Paste Columns | Check Spelling | Comparison Functions | Job Monitoring | Change Color Scheme | Calendar Formats | Comparison Functions |
| Create Summaries | Cluster Data Prep Steps | Copy and Paste Columns | Job Scheduling | Check Spelling | Country Name into Codes | Copy and Paste Columns |
| Data Histogram | Comparison Functions | Data Quality Statistics | Maintain Log | Data Size Details | Data Masks | Data Histogram |
| Data Lineage | Copy and Paste Columns | Date & Time Formats | Math functions | External Data Use | Date & Time Formats | Data Profiling |
| Data Sampling | Data Histogram | Deduplication Statistics | Multi User Access | Group and Replace | External Data Use | Date & Time Formats |
| Data Size Details | Data Lineage | External Data Use | Preparation Versions | Group Tasks | Extract Quarter from Date | Diagnose Failed Jobs |
| Edit Field Values | Data Profiling | Hide Column | Refresh Data from Source | Intelligent Bar | Find and Group | External Data Use |
| External Data Use | Data Sampling | Intelligent Bar | Reorder Preparation Steps | Maintain Log | Intelligent Bar | Fix Dependency Issues |
| Hide Column | Data Size Details | Maintain Log | Search and Replace | Math Functions | Maintain Log | Group and Replace |
| Job Scheduling | Date & Time Formats | Math Functions | Share Dataset | Mini Maps | Math Functions | Initial Parsing Steps |
| Maintain Log | External Data Use | Multi User Access | String Functions | Multi Language Support | Multi Language Support | Intelligent Bar |
| Math Functions | Find and Group | Multiple Graphs for Visuals | Transpose Data | Preparation Versions | Preparation Versions | Logical Functions |
| Move Column | Group and Replace | Preparation Versions | View Table Properties | Publish Flows | Reorder Preparation Steps | Maintain Log |
| Preparation Versions | Intelligent Bar | Refresh Data from Source | Visual Feedback | Refresh Data from Source | Search and Replace | Manage Flows with Folders |
| Refresh Data from Source | Intelligent Ingest | Reorder Preparation Steps | Work with Plans | Reorder Preparation Steps | Share Dataset | Manage String Lengths |
| Row and Column Counts | Maintain Log | Search and Replace | | Schedule Flows | String Functions | Math Functions |
| Search and Replace | Math Functions | Share Dataset | | Search and Replace | Suggestions | Multi Language Support |
| String Functions | Multi Language Support | String Functions | | Share Dataset | Swap Column Content | Preparation Versions |
| Transpose Data | Preparation Versions | Suggestions | | String Functions | Use Metric Symbols | Reorder Preparation Steps |
| Visual Feedback | Publish Dataset | | | Suggestions | Visual Feedback | Row and Column Counts |
| | Reorder Preparation Steps | | | Visual Feedback | | Search and Replace |
| | Search and Replace | | | | | Sequence Datasets |
| | Send Notifications | | | | | Share Dataset |
| | Share Dataset | | | | | String Functions |
| | String Functions | | | | | Suggestions |
| | Suggestions | | | | | Target-driven preparation |
| | Version History | | | | | Track Data Changes |
| | Visual Feedback | | | | | Visual Feedback |
| | | | | | | Workflow Automation |