I.J. Information Technology and Computer Science, **2018, 1, 32-39**
Published Online January 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2018.01.04

A Systematic Study of Data Wrangling

Malini M. Patil
Associate Professor, Dept. of Information Science and Engineering.
J.S.S Academy of Technical Education, Bengaluru, Karnataka
E-mail: drmalinimpatil@gmail.com

Basavaraj N. Hiremath
Research Scholar, Dept. of Computer Science and Engineering,
JSSATE Research Centre, J.S.S Academy of Technical Education, Bengaluru, Karnataka
E-mail: basavaraj@ieee.org

Received: 26 September 2017; Accepted: 07 November 2017; Published: 08 January 2018

***Abstract*—**The paper presents the theory, design, usage aspects of data wrangling process used in data ware housing and business intelligence. Data wrangling is defined as an art of data transformation or data preparation. It is a method adapted for basic data management which is to be properly processed, shaped, and is made available for most convenient consumption of data by the potential future users. A large historical data is either aggregated or stored as facts or dimensions in data warehouses to accommodate large adhoc queries. Data wrangling enables fast processing of business queries with right solutions to both analysts and end users. The wrangler provides interactive language and recommends predictive transformation scripts. This helps the user to have an insight of reduction of manual iterative processes. Decision support systems are the best examples here. The methodologies associated in preparing data for mining insights are highly influenced by the impact of big data concepts in the data source layer to self-service analytics and visualization tools.

***Index Terms*—**Business Intelligence, wrangler, prescriptive analytics, data integration, predictive transformation.

I. Introduction

The evolution of data warehouse (DWH) and business intelligence (BI) started with basic framework of maintaining the wide variety of data sources. In traditional systems, the data warehouse is built to achieve compliance auditing, data analysis, reporting and data mining. A large historical data is either aggregated or stored in facts to accommodate ad hoc queries. In building these dimensional models the basic feature focused is ‘clean data’ and ‘integrate data’ to have an interaction, when a query is requested from the downstream applications, to envisage meaningful analysis and decision making. This process of cleansing not only relieves the computational related complexities at the business intelligence layer but also on the context of performance. The two key processes involved are detecting discrepancy and transforming them to standard content to carry out to the next level of data warehouse architecture. The wrangler provides interactive transformative language with power of learning and recommending predictive transformational scripts to the users to have an insight into data by reduction in manual iterative processes. There are also tools for learning methodologies to give predictive scripts. Finally, publishing the results of summary data set in a compatible format for a data visualization tool. In metadata management, data integration and cleansing play a vital role. Their role is to understand how to utilize the automated suggestions in changing patterns, be it data type or mismatched values in standardizing data or in identifying missing values.

Data wrangling term is derived and defined as a process to prepare the data for analysis with data visualization aids that accelerates the faster process [1]. It allows reformatting, validating, standardizing, enriching and integration with varieties of data sources which also provides space for self-service by allowing iterative discovery of patterns in the datasets. Wrangling is not dynamic data cleaning [2]. In this process it manages to reduce inconsistency in cleaning incomplete data objects, deleting outliers by identifying abnormal objects. All these methods involve distance metrics to make a consistent dataset. Techniques of data profiling [3] are involved in few of the data quality tools, which consists of iterative processes to clean the corrupted data. These techniques have limitations in profiling the data of multiple relations with dependencies. The optimization techniques and algorithms need to be defined in the tools.

Data wrangling refers to ‘Data preparation’ which is associated with business user savvy i.e. self-service capability and enables, faster time to business insights and faster time to action into business solutions to the business end users and analysts in today’s analytics space. As per the recent best practices report from Transforming Data with Intelligence (TDWI) [4], the percentage of time spent in preparing data compared to the time spent in performing analysis is considerable to the tune of 61 percent to 80 percentage. The report emphasizes on the challenges like limited data access, poor data quality and delayed data preparation tasks. The best practices fol

lowed are as follows:

1. Evolve as an independent entity without the involvement of information technology experts, as it's not efficient and time consuming.
2. Data preparation should involve all types of corporate data sets like data warehouses/data lakes, BI data, log data, web data and historical data in documents and reports.
3. To create a hub of data community which eases collaboration for individuals and organization, more informed as agile and productive.

The advent of more statistical and analytical methods to arrive at a decision-making solution of business needs, prompted the use of intensive tools and graphical user interface based reports. This advancement paved the way for in-memory and 'data visualization' tools in the industry, like Qlik Sense, Tableau, SiSense, and SAS. But their growth was very fast and limited to 'BI reporting area'. On the other hand, managing data quality, integration of multiple data sources in the upstream side of data warehouse was inevitable. Because of the generous sized data, as described by authors [5], a framework of bigdata ecosystem has emerged to cater the challenges posed by five dimensions of big data viz., volume, velocity, veracity, value and variety. The trend has triggered growth in technology to move towards, designing of self-service, high throughput systems with reduction in deployment duration in all units of building end to end solution ecosystem. The data preparation layer which focuses on "WRANGLING" of data by data wrangler tools, A data management processes to increase the quality and completeness of overall data, to keep as simple and flexible as possible. The features of data wrangler tools involve natural language based suggestive scripts, transformation predictions, reuse of history scripts, formatting, data lineage and quality, profiling and cleaning of data. The Big data stack includes the following processes to be applied for data [6], data platforms integration, data preparation, analytics, advanced analytics.

In all these activities data preparation consumes more time and effort when compared to other tasks. So, the data preparation work requires self-service tools for data transformation. The basic process of building and working with the data warehouse is managed by extract, transform and loading (ETL) tools. They have focused on primary tasks of removing inconsistencies and errors of data termed as data cleansing [7] in decision support systems. Authors specify need of 'potters wheel' approach on cleansing processes in parallel with detecting discrepancy and transforming them to an optimized program in [8].

II. Related Work

Wrangle the data into a dataset that provides meaningful insights to carryout cleansing process, it requires writing codes in idiosyncratic characters in languages of Perl, R and editing manually with tools like MS-Excel [9]. There are processes, that user might use iterative framework to cleanse the data at all stages like analysis in visualization layer and on downstream side. The understanding of sources of data problems can give us the level of quality of data for example sensor data which is collected has automated edits in a standard format, whereas manually edited data can give relatively higher errors which follows different formats.

In the later years, many research works are carried out in developing query and transformational languages. The authors [10] suggests to focus future of research should be on data source formats, split transforms, complex transforms and format transforms which extracts semi automatically restricted text documents. The research works continued along with advancement in structure and architecture of functioning of ETL tools. These structures helped in building warehousing for data mining [11]. The team of authors [12] from Berkeley and Stanford university made a breakthrough work in discovering the methodology and processes of wrangler, an interactive visual specification of data transformation scripts by embedding learning algorithms and self-service recommending scripts which aid validation of data i.e. created by automatic inference of relevant transforms. According to description of data transformation in [13], specify exploratory way of data cleansing and mining which evolved into interesting statistical results. The results were published in economic education commission report, as this leads to statutory regulations of any organization in Europe. The authors [14] provide an insight of specification for cleansing data for large databases. In publication [15], authors stipulated the need of intelligently creating and recommending rules for reformatting. The algorithm automatically generates reformatting rules. This deals with expert system called 'tope' that can recognize and reformat. The concepts of Stanford team [7] became a visualized tool and it is patented for future development of enterprise level, scalable and commercial tools The paper discusses about the overview of evolution of data integrator and data preparation tools used in the process of data wrangling. A broad description of what strategy is made to evolve from basic data cleansing tools to automation and self-service methods in data warehousing and business intelligence space is discussed. To understand this the desktop version of Trifacta software, the global leader in wrangling technology is used as an example [16]. The tool creates a unique experience and partnership with user and machine. A data flow is also shown with a case study of a sample insurance data, in the later part of the paper. The design and usage of data wrangling has been dealt in this paper. The paper is organized as, in section III, detailed and the latest wrangling processes are defined which made Datawarehousing space to identify self-service methods to be carried out by Business analysts, in order to understand how their own data gives insights to be derived from this dataset. The section III A, illustrates a case study for data wrangling process with a sample dataset followed with conclusions.

A. Overview of Data Wrangling Process

For building and designing different DWH and BI roadmap for any business domain the following factors are followed. namely,

- Advances in technology landscape
- Performance of complete process
- Cost of resources
- Ease of design and deployment
- Complexity of data and business needs
- Challenges in Integration and data visualization.
- Rapid development of Data Analytics solution frameworks

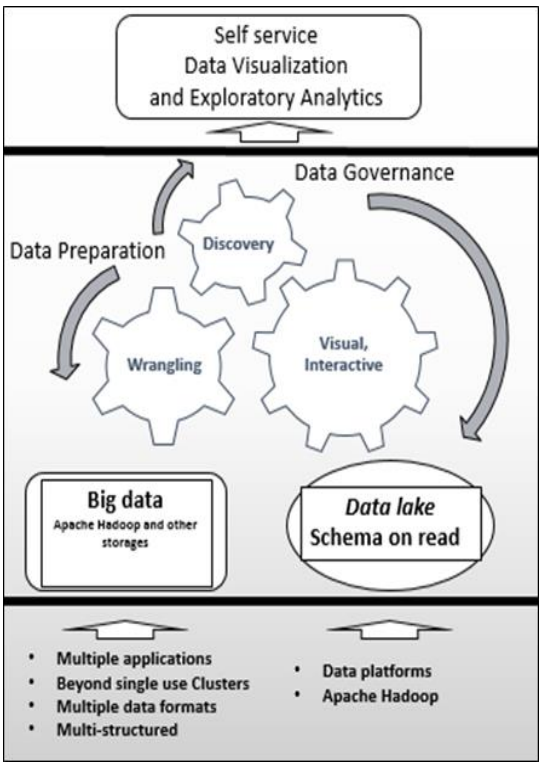

Fig.1. Data Flow diagram for data preparation

This trend has fuelled the fast growth of ETL process of enterprise [8] or data integration tools i.e. building the methodologies, processing frameworks and layers of data storages. In the process of evolution of transformation rules, for a specific raw data, various data profiling and data quality tools are plugged in. In handling, various data sources to build an enterprise data warehouse relationship, a "Data Integrator" tool is used which forms the main framework of ETL tools. At the other end, the robust reporting tools, the front end of DWH, built on the downstream of ETL framework also transformed to cater both canned and real-time reports under the umbrella of business intelligence, which helps in decision making. Fig. 1., is a typical data flow which needs to be carried out for preparing data, subjected to exploratory analytics. In this data flow, the discussion on big data is carried out in two cases one, cluster with Hadoop platform and other with multiple application and multi structured environment called as "data lake". In both the cases the data preparation happens with data wrangling tools which has all the transformation language features. The data obtained from data lake platform has lot of challenging transformations i.e. not only reshaping or resizing but data governance tasks [17] which is like "Governance on need" [6] and the schema of the data is "Schema on read' which is beyond the predefined silos structure. This situation demands self-service tasks in transformations beyond IT driven tasks. So, on hand prescriptive scripts, visual and interactive data profiling methods, supports guided data wrangling tasks steps to reshape and load for the consumption of exploratory analytics. The analysis of evolution of enterprise analytics is shown in Fig. 2. Earlier, business intelligence was giving solutions for business analysis, now predictive and prescriptive analytics guides and influence the business for 'what happens' scenarios. These principles guide the data management decisions about the data formats and data interpretation, data presentation and documentation. Some of the Jargons associated are 'Data munging' and 'Janitor works'.

The purpose and usage of Data Lake will not yield results as expected unless the information processing is fast and the underlying data governance is built with a well-structured form. The business analysis team could not deal with data stored in Hadoop systems and they could not patch with techno functional communication to get what business team wants to accomplish as most of the data is behind the frame, built and designed by IT team. These situations end in the need of self-service data preparation processes.

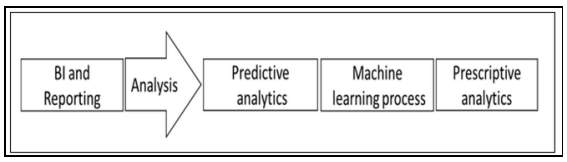

Fig.2. Evolution of Enterprise Analytics

The self-service tool will help to identify typical transformation data patterns in 'easy to view visualize forms. Four years back there were few frameworks with open source tools released as a data quality tools to execute by IT teams. For example, TALEND DQ is emerged as an open source tool for building solutions to analyze and design transformations by IT teams, but they lack automation scripts and predictions for probable transformations. on the other side, advancement of learning algorithms and pattern classifications, fuzzy algorithms helped to design transformational language for enterprise level solution providers. The complexity of data size for identifying business analytical solutions is one of the important challenges of five dimensions of big data.

The authors [18] have the opinion that, wrangling is an opportunity and a problem with context to big data. The unstructured form of big data makes the approaches of manual ETL processes a problem. If the methods and techniques of wrangling becomes cost effective, the

impractical tasks are made practical. Therefore, the traditional ETL tools, struggled to support for the analytical requirements and adopted the processes by framing tightly governed mappings and released as builds, with a considerable usage of BI and DWH methodologies. The data wrangling solutions were delivered fast by processing the 'raw' data into the datasets required for analytical solutions. Traditionally the ETL technologies and data wrangling solutions are data preparation tools. but for gearing up any analytics initiatives they require cross functional and spanning various business verticals of organizations.

There is a need of analysis of data at the beginning, which ETL tools lacks, this can be facilitated by wrangling processes. The combination of visualization, pattern classification of data using algorithms of learning and the interaction of analysts makes process easy and faster, which will be self-serviced data preparation methods. This transformation language is [19] built with visualization interfaces to display profile and to interact with user by recommending predictive transformation for the specific data in wrangling processes. Earlier the programming [11] was used for automating transformations. It's difficult to program the scripts for complex transformations using a programming language, spread sheets and schema mapping jobs. So, an outcome of research was design of a transformation language, the following basic transformations were considered in this regard.

Sizing: Data type and length of data value matters a lot in cleansing, rounding, truncate the integer and text values respectively.

Reshape: The data to consider different shapes like pivot/un-pivot makes rows and columns to have relevant grid frame.

Look up: To enrich data with more information like addition of attributes from different data sources provided Key ID present.

Joins: Use of joins helps to join two different datasets with join keys.

Semantics: To create semantic mapping of meanings and relationships with a learning tools.

Map positional transforms: To enrich with time and geographical reference maps for the available a data.

Aggregations and sorting: The aggregations (count, mean, max, list, min etc.) by group of attributes to facilitate drill down feature and sorting on specific ranking for attribute values.

These transformation activities can be done by framework of architecture which includes connectivity to all the data sources of all sizes of data. This process supports task of metadata management as it is more effective for enrichment of data. An embedded knowledge learning system i.e. using machine learning algorithms are built to give user, a data lineage and matching data sets. The algorithm efficiently identifies categories of attributes like geospatial [20]. A frame of transformational language in Fig. 3. depicts sample script) is used by few wrangling tools which uses their own enterprise level 'wrangling languages' [15] like Refine tool uses Google refine expression language [14]. This framework must have an in-built structure to deal with data governance and security, whereas, ETL tool frames an extra module built in enterprise level. Though the current self-service BI tools give limitless visual analytics to showcase and understand data, but they are not managed to give us end to end data governance lineage among data sets. This enables expert users to perform expressive transformations with less difficulty and tedium. They can concentrate on analysis and not being bogged down in preparing data. The design is built to suggest transforms without programming. The wrangler language uses input data type as semantic role to reduce tasks arise out of execution and evaluation by providing natural language scripts and virtual transform previews [8].

```
split('data').on(NEWLINE).max_splits(NO_MAX)
split('split').on(COMMA).max_splits(NO_MAX)
columnName().row(0)
delete(isEmpty())
extract('Year').on(/.*/).after(/in /)
columnName('extract').to('State')
fill('State').method(COPY).direction(DOWN)
delete('Year starts with "Reported"')
unfold('Year').above('Property_crime_rate')
```

Fig.3. Sample script of transformation language

B. Tools for Data Wrangling

The data wrangling technology is evolving at a very faster rate to prove and get placed as enterprise level industry standard software (self-service data preparation tool) [14] To quote few, clear story data, Trifacta, Tamr, Paxata and Google refine, other tools Global ID's, IBM Data works, Informatica Springbok. These tools are basically used in industry for data transformation, data harmonization, data preparation, data integration, data refinery and data governance tools [15] Currently Trifacta and Paxata are the strong performer tools [16]. In some organizations, the deployment is done both with ETL solutions and Wrangling tools, as ETL solutions do primary data integration and load into enterprise data warehouse. From this system, the business users can experience data analysis by exploring wrangling solutions. Most of data wrangling tools uses predictive transformation scripts with underlying machine learning algorithms, visualizations methods and scalable data technologies, namely., Hadoop based infrastructure framed by Cloudera, Hortonworks and IBM. In earlier phase, most of the "traditional" enterprise ETL / BI tools were doing cleansing of structured data and semi structured data to fit in RDBMS (Relational database management system) frame. But the advent of big data challenged the growth of ETL tools to be reshaped with various add-ons as plugins into the existing framework. So, the enterprise ETL tools were reformed as individual bundled tools to carry out the specific purpose of action as per the industry needs. They are data quality, data

integrator, big data plugins, cloud integrators. In parallel, the BI space, as discussed earlier, because of fast growth of data visualization and in-memory tools and their ease of use, they evolved into self-service BI tools to leverage purpose of saving ease of design and development i.e. lesser intervention of IT resources, more ease of use of domain expertise to be called as 'self-service tools'.

III. Data Wrangling Process Using Trifacta

This section presents the data wrangling process using the desktop version of Trifacta tool. The description of the dataset used for the wrangling process is also discussed.

A benchmark dataset is used to investigate the features of the wrangling tool is taken from catalog USA government dataset [21]. The data set is from insurance domain and is publicly available, which emphasizes economic growth by the USA federal forum on open data. This case study has three different files to be imported with scenarios of cleansing and standardizing [22]. The dataset has variations of data values with different data types like integer for policy number, date format for policy date, travel time in minutes, education profile in text and has a primary key to process any table joins. The dataset has the scope to enrich, merge, identify missing values, rounding to decimals and to visualize in the wrangling tool with perfect variations in attribute histogram. (Observed in Fig.4).

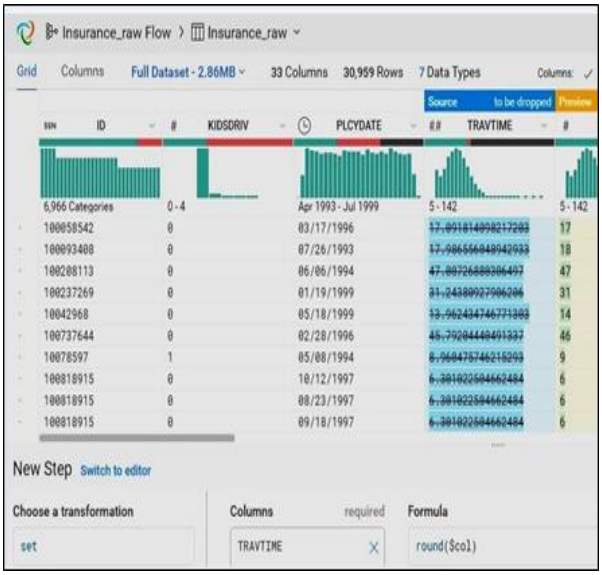

Fig.4. Visualization of distribution of each data attributes

An example of insurance company is considered to investigate the business insights of it. A sample data is selected to wrangle. This case study has three different files to be imported with scenarios of cleansing and standardizing [22]. They are main transaction file, enrich the data file with merged customer profile and the add on data from different branches. Trifacta enables the connectivity to various data sources. The data set is imported using drag and drop features available in Trifacta and saved as a separate file. The data can be previewed to have a first look of attributes and to know the relevance after importing. The data analysis is done in grid view displayed each attribute, data type patterns, in order to give a clear data format quality tips as shown in Fig. 4.

The user can get a predictive transformation tips on the processed data values for standardization, cleaning and formatting that will be displayed at the lower pane of the grid like rounding of integer values for e.g., travel time of each insured vehicle is represented in minutes as whole number and not as decimals. If a missing value is found, Trifacta intelligently suggests the profile of data values in each attribute, that is easy for analysis. To have an insight on valuable customers, the user can concentrate on data enrichment, that must be a formatting suggestion for date format.

All the features can be viewed in Fig. 5. The data enrichment is done by bringing profile data of the nature of personal information to be appended with lookup attributes. The lookup parameters are set and mapped to the other files as shown in Fig. 6.

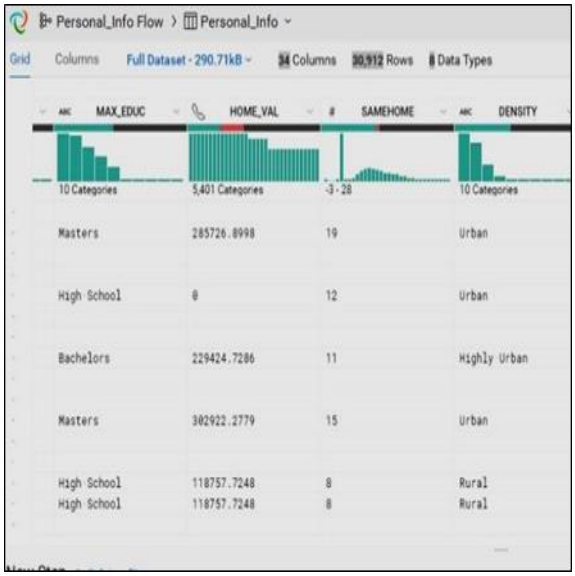

Fig.5. Data format script of transformation language

Trifacta provides an option of retaining or deleting the information by selecting number of attributes in the file as shown in Fig. 6. Trifacta also supports editing the user defined functions and can be programmed using other languages like Java and Python. This feature can be used for doing sentiment analysis using social media data as Trifacta is able to import Json data and other text data. Finally, merging of the data files as received from other branches with existing records is done by opening tool menu using union tool which prompts for right keys and attributes. All the modified or processed data transformation steps are stored as stepwise procedures which can be reused and modified for future use. A graphical representation of flow diagram is shown in Fig. 7. Once the job is run, the result summary parameters can be seen in Fig. 8. Data enrich, data cleansing and data merge is done with simple steps of visualizing and analyzing the data, finally the summarized result data set is stored in the

native format of data visualization tool like tableau or QlikView or in csv and Json format.

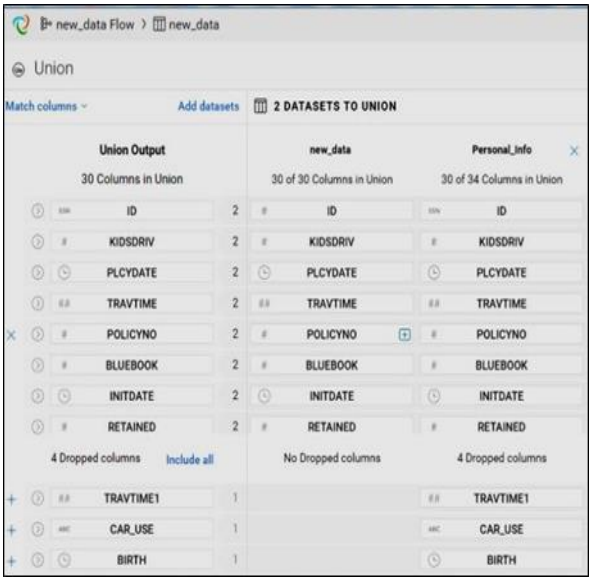

Fig.6. Display of Lookup attributes

Thus, it is found that the wrangler tool allows human machine interactive transformation of real world data. It enables business analysts to iteratively explore predictive transformation scripts with the help of highly trained learning algorithms. Its inefficient to wrangle data with Excel which limits the data volume. The wrangling tools come with seamless cloud deployment, big data and with scalable capabilities with wide variety of data source connectivity which has evolved from preliminary version displayed in Fig. 9 [19].

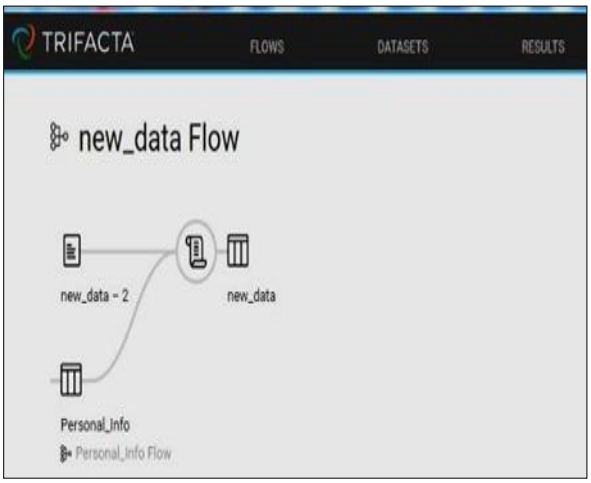

Fig.7. Graphical representation of flow diagram

The future of Data science leverages data sources reserved for data scientists which are now made available for business analysis, interactive exploration, predictive transformation, intelligent execution, collaborative data governance are the drivers of the future to build advanced analytics framework with prescriptive analytics to beat the race of analytical agility. The future insights of the data science need a scalable, faster and transparent processes. The [16] (2017 Q1) report cites that the future data preparation tools must evolve urgently to identify customer insights and must be a self-service tool equipped with machine learning approaches and move towards independence from technology management.

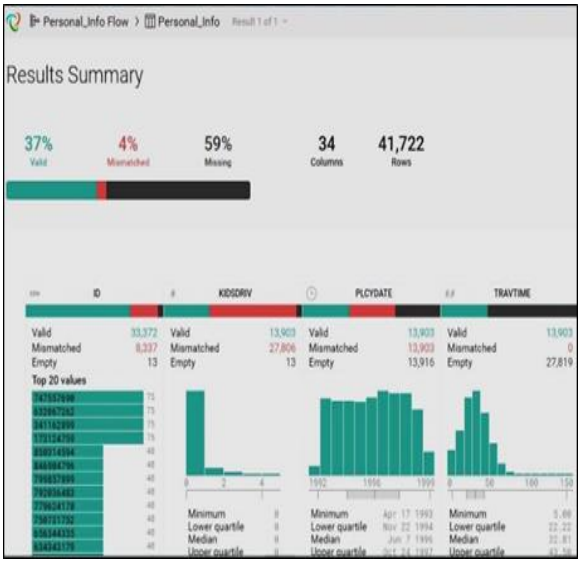

Fig.8. Results summary

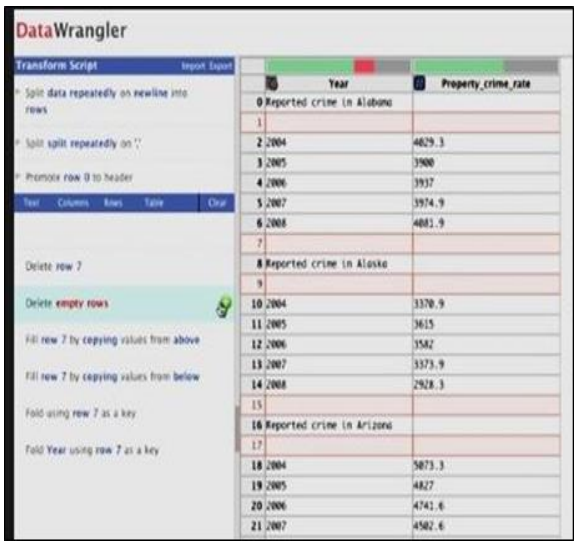

Fig.9. Preliminary version of wrangler tool

IV. Conclusion

It is concluded that data processing is a process of more sophisticated. The future 'wrangling' tools should be aggressive in high throughput and reduction in time to carry out wrangling tasks in achieving the goal of making data more accessible and informative and ready to explore and mine business insights. The data wrangling solutions are exploratory in nature in arriving at analytics initiative. The key element in encouraging technology is to develop strategy to skill enablement of considerable number of business users, analysts and to create large corpus of data on machine learning models. The

technology revolves around to design and support data integration, quality, governance, collaboration and enrichment [23]. The paper emphasizes on the usage of Trifacta tool for data warehouse process, which is one of its unique kind. It is understood that the data preparation, data visualization, validation, standardization, data enrichment, data integration encompasses a data wrangling process.

REFERENCES

[1] Cline Don, Yueh Simon and Chapman Bruce, Stankov Boba, Al Gasiewski, and Masters Dallas, Elder Kelly, Richard Kelly, Painter Thomas H., Miller Steve, Katzberg Steve, Mahrt Larry, (2009), NASA Cold Land Processes Experiment (CLPX 2002/03): Airborne Remote Sensing.

[2] S. K. S and M. S. S, "A New Dynamic Data Cleaning Technique for Improving Incomplete Dataset Consistency," Int. J. Inf. Technol. Comput. Sci., vol. 9, no. 9, pp. 60–68, 2017.

[3] A. Fatima, N. Nazir, and M. G. Khan, "Data Cleaning In Data Warehouse: A Survey of Data Pre-processing Techniques and Tools," Int. J. Inf. Technol. Comput. Sci., vol. 9, no. 3, pp. 50–61, 2017.

[4] Stodder. David (2016), WP 219 - EN, TDWI Best practices report: Improving data preparation for business analytics Q3 2016. © 2016 by TDWI, a division of 1105 Media, Inc, [Accessed on: May 3rd, 2017].

[5] Richard Wray, "Internet data heads for 500bn gigabytes | Business the Guardian," www.theguardian.com. [Online] Available:https://www.theguardian.com/business/2009/may/18/digital-content-expansion. [Accessed: 24-Oct-2017].

[6] Aslett Matt Research analyst Trifacta of 451 Research and Davis Will head of marketing, Trifacta, "Trifacta maintains data preparation" [July ,7, 2017], [Online] Available: https://451research.com [Accessed on: 01 August 2017].

[7] Kandel Sean, Paepcke Andreas, Hellersteiny Joseph and Heer Jeffrey (2011), Wrangler: Interactive Visual Specification of Data Transformation Scripts, ACM Human Factors in Computing Systems (CHI) ACM 978-1-4503-0267-8/11/05.

[8] Chaudhuri. S and Dayal. U (1997), An overview of data warehousing and OLAP technology. In SIGMOD Record.

[9] S. Kandel et al., "Research directions in data wrangling: Visualizations and transformations for usable and credible data," Inf. Vis., vol. 10, no. 4, pp. 271–288, 2011.

[10] Chen W, Kifer.M, and Warren D.S, (1993), "HiLog: A foundation for higher-order logic programming". In Journal of Logic Programming, volume 15, pages 187-230.

[11] Raman Vijayshankar and Hellerstein Joseph M, frshankar, (2001) "Potter's Wheel: An Interactive Data Cleaning System", Proceedings of the 27th VLDB Conference.

[12] Norman D.A, (2013), Text book on "The Design of Everyday Things, Basic Books", [Accessed on:12 April 2017].

[13] Carey Lucy, "Self-Service Data Governance & Preparation on Hadoop", www.jaxenter.com. [Online] (May 29, 2014) Available: https://jaxenter.com/trifacta-ceo-the-evolution-of-data-transformation-and-its-impact-on-the-bottom-line-107826.html [Accessed on01 April 2017].

[14] Google code online publication (n.d), www.code.google.com. [Online] Available: https://code.google.com/archive/p/google-refine https://github.com/OpenRefine/OpenRefine [Accessed on; 28 March 2017].

[15] Data wrangling platform (2017) publication, www.trifacta.com. [Online] Available: https://www.trifacta.com/products/architecture//, [Accessed on: 01 May 2017].

[16] Little Cinny, The Forrester Wave™: Data Preparation Tools, Q1 2017 "The Seven Providers That Matter Most and How They Stack Up", [Accessed On:March 13, 2017]

[17] Parsons Mark A, Brodzik Mary J, Rutter Nick J. (2004), "Data management for the Cold Land Processes Experiment: improving hydrological science HYDROLOGICAL PROCESSES" Hydrol. Process. 18, 3637-3653.

[18] T. Furche, G. Gottlob, L. Libkin, G. Orsi, and N. W. Paton, "Data Wrangling for Big Data: Challenges and Opportunities," EDBT, pp. 473–478, 2016.

[19] Kandel Sean, Paepcke Andreas, Hellersteiny Joseph and Heer Jeffrey (2011), published Image on Papers tab, www.vis.stanford.edu. [Online] Available: http://vis.stanford.edu/papers/wrangler%20paper [Accessed on: 25 May 2017].

[20] Ahuja.S, Roth.M, Gangadharaiah R, Schwarz.P and Bastidas.R, (2016), "Using Machine Learning to Accelerate Data Wrangling", IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, Barcelona, Spain, pp. 343-349.doi:10.1109/ICDMW.2016.0055.

[21] "Data catalog" Insurance dataset, [Online]www.data.gov. Available: https://catalog.data.gov/dataset. [Accessed: 24-Oct-2017].

[22] Piringer Florian Endel Harald, florian. (2015), Data Wrangling: Making data useful again, IFAC-PapersOnLine 48-1 (2015) 111-112.

[23] V. kumar, Tan Pang Ning, Steinbach micheal, Introduction to Data Mining. Dorling Kindersley (India) Pvt. Ltd: Pearson education publisher, 2012.

Authors' Profiles

Malini M. Patil is presently working as Associate Professor in the Department of Information Science and Engineering at J.S.S. Academy of Technical Education, Bangalore, Karnataka, INDIA. She received her Ph.D. degree from Bharathiar University in the year 2015.

Her research interests are big data analytics, bioinformatics, cloud computing, image processing. She has published more than 20 research papers in many reputed international journals. Published article, Malini M patil, Prof P K Srimani, "Performance analysis of Hoeffding trees in data streams by using massive online analysis framework", 2015, vol 12, International Journal of Data Mining, Modelling and Management (IJDMMM),7,4pg293-313,Inderscience Publishers. Malini M Patil, and Srimani P K, "Mining Data streams with concept drift in massive online analysis frame work", WSES Transactions on computers, 2016/3, Volume 15 Pages 133-142.

Dr. M Patil is a member of IEEE, Institution of Engineers (India), Indian Society for Technical Education, Computer Society of India. She is guiding four students. She has attended and presented papers in many international conferences in India and Abroad. She is a recipient of distinguished woman in Science Award for the year 2017 from Venus International Foundation. Contact email: drmalinimpatil@gmail.com

Basavaraj N Hiremath is a research scholar working in the field of artificial intelligence at JSSATE Research Centre, Dept. of CSE, JSSATE, affiliated to VTU Belagavi, INDIA. He is pursuing research under Dr. Malini M Patil, Associate Professor, Dept., of ISE, JSSATE. Completed M. S. in Computer Cognition Technology in 2003 from Department of studies in Computer science, University of Mysore, Karnataka INDIA.

He has also worked in information technology industry as SOLUTION ARCHITECT in data warehouse, business Intelligence and analytics space in various business domains of Airlines, Retail, Logistics and FMCG. Published article, Basavaraj N Hiremath, and Malini M Patil, "A Comprehensive Study of Text Analytics", CiiT International Journal of Artificial Intelligent Systems and Machine Learning, 2017/4, vol 9, no 4,70-77

Mr Hiremath is a Fellow of Institution of Engineers (India), member IEEE, member Computer Society of India and Life member Indian Society for Technical Education, Member Association for the Advancement of Artificial Intelligence.

How to cite this paper: Malini M. Patil, Basavaraj N. Hiremath, "A Systematic Study of Data Wrangling", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.1, pp.32-39, 2018. DOI: 10.5815/ijitcs.2018.01.04

