NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme HSE University and University of London Double Degree Programme in "Data Science and Business Analytics"

UDC _____

Research Project Report

on the topic Data Preparation Tools

(interim, the first stage)

Fulfilled by the Student:
group #БПАД___ _____ _____
Signature Surname, First name, Patronymic, if any

_____
Date

Checked by the Project Supervisor:
Nikitin Sergey Aleksandrovich
Surname, First name, Patronymic (if any), Academic title (if any)

_____
Job

_____
Place of Work (Company or HSE Department)

Date_____2023 _____
Signature

Moscow 2023

Contents

1 Definitions and Abbreviations

BI - business intelligence

DataPrep - data preparation

DWH - data warehouse

ELT - extract, load, transform

ETL - extract, transform, load

OLAP - online analytical processing

2 Introduction

High quality and accurate data is crucial to every business. Useful data gives a greater chance of success to specialists who will come up with more credible decisions that make a difference. Data preparation is the process of cleaning and aggregating data for use in business analysis. It is the act of consolidating raw data and transforming it into a format that is simple to work with, which ensures that collected data is complete, accurate, and reliable. The raw data might be in any format and come from several sources. By cleaning up various types of data for different kinds of analysis, data preparation aims to provide a clean set of data for accurate reporting and forecasts of business analysts and data scientists.

Companies must make it possible for analysts to derive insights from data as business processes increasingly become digital. That is why, a lot of companies these days view data preparation as the key component to increasing their ability to efficiently utilize data to optimize business processes and the decisions that business leaders make are only as good as the data that they have to back it up. The accuracy and significance of analyses are increased by careful and thorough data preparation, which builds analysts' trust and enables them to ask better questions of their data. Better insights and, therefore, better outcomes that keep a company ongoing and relevant emerge from meaningful data analysis.

It is often true that the professionals will need to map data from several platforms together to acquire a view of the whole picture in order to extract relevant insights from the data. The aim of data preparation tools is to map data together. Data preparation tools refer to various tools used for discovering, processing, mixing, refining, cleansing and transforming data for future use. This enables the use of advanced busi-

ness intelligence and analytics tools to effectively integrate, consume, and analyze greater datasets.

Modern data preparation tools provide self-service capabilities for IT departments, data scientists, professional data analysts, and regular business owners. These tools make it possible to quickly and effectively integrate several data sources in one place.

The preparation, mixing, and refinement of data results in a much better and more efficient data analysis process. This helps businesses acquire useful insights to improve and enhance their functionality. Most data preparation software provide governance, control, data management, and machine learning features. These features enhance the functionality of the software as a whole and support efficient business management.

Relevance: In this constantly developing, dynamic industry it is extremely difficult to be kept up to date. However, there is a lack of scientific research and articles regarding the topic of data preparation and little information on best practice data preparation tools. This paper will cover aspects of data preparation tools that are not covered in current and previous works, and, therefore, help businesses and specialists gain info on the topic of research.

Based on the relevancy of DataPrep and DataPrep Tools, this paper will look into their features, functionality and areas of use.

Purpose: Research, analysis and determination of best-practice Data Preparation Tools on the market

Object of the research: Data Preparation Tools

Subject of the research: Features and functions of chosen tools

With consideration of project relevance, purpose and pre-research, there were established the following **steps:**

1. Determine the set of tools for further use and analysis:

Research the market of DataPrep Tools: look through online reviews of

the users, read articles and books of other services determining the best tools and their best analogues.

2. Determine the set of main features and criteria for comparison:

Look through users reviews on various platforms, describing favorable and preferable features of chosen tools, and search for articles highlighting necessary criteria. Compare and contrast their feedback, analyzing and defining the most useful of highly mentioned criteria by the users.

3. Develop a functional framework for analysis:

Come up with practical methodology for future analysis, including the steps that are going to be used and followed for consecutive and successful analysis of DataPrep Tools.

4. Analysis and classification of chosen tools based on selected features:

Analysis of chosen tools according to selected features and established framework. Structuring and classifying DataPrep Tools in terms of their characteristics and properties, while documenting the process. Concluding the process of analysis with important insights and possible recommendations in the end of the research.

5. Result presentation

Finishing off the documentation, preparing the presentation and defending the project on the due day.

3 Analytical review of sources

Research focused on data preparation, data preparation tools, and its rising concerns is scattered across the fields of databases, human-computer interactions, machine learning, and others. Although little research has been done on the subject, it has grown to be a significant force in the industry. Annual Market Guides for Data Preparation were released by reputable analytical firms like Gartner [4] together with Products In Data Preparation Tools Market on Peer Insights [5]. According to estimates, the market for data preparation technologies is valued 2.9 billion dollars and expanding quickly. The amount of importance is remarkable given that the subject of data preparation is new and not overly discussed.

David Stodder in Improving Data Preparation for Business Analytics [3] claims that combining views of diverse types of data is an important task of data preparation operations for many business users. For analysts to produce useful insights, organizations need well-integrated data that has been effectively converted. To meet these needs, some have started using conventional, non-self-service data preparation tools. Nevertheless, companies are able to meet self-service business and governance expectations for quicker generation of integrated views of data due to emerging solutions on the market.

Proving the point, specialists from Trifacta in their article [1] argue that the primary problem in data preparation is self-service. There is a huge need to enable the people who know the data best to prepare it themselves.

They also highlight several tasks that are included in the process of data preparation:

Unboxing: Discovery and Assessment. What the dataset focuses

on, what it looks like, who or what is in that data. It is difficult to analyze data without having an answer to these questions.

Structuring. Tables, matrices, dictionaries or lists – structuring focuses on formatting data in the shape that is easy and comfortable to operate.

Cleaning. Not just neat but thoroughly clean data can produce the best results in retrospect. The process of cleansing relies on detecting and correcting data, or completely deleting data for accuracy.

Enriching and Blending. Data requires additional changes for better context since data as in itself is not sufficient most of the time. Moreover, the data need to be constantly compared to the one of the larger scale for consistency.

Optimizing. A clean and structured dataset is often in need to do additional preparation to produce appropriate output for the next phase of analysis. It might also include data normalization or reduction prior to further publishing.

Other research articles contemplate whether the expensive aspect of data preparation comes from significant measures of programming tasks [6] due to the fact that data scientists either directly build data wrangling algorithms, use visual programming interfaces to create transformation scripts, or use workflows to combine data preparation procedures. This results in a substantial quantity of effort.

Data preparation tools tend to provide features that help to manage similar tasks (e.g., joining data sets, reformatting columns), but differ in how these are expressed by data scientists [7]. A data scientist still has fine-grained control over the specification of a data preparation task, even with tool help. There are instances where this is appropriate, but because the expenses are considerable and occasionally prohibitive, more automated methods are being investigated.

Even though early work on automating end-to-end data preparation appears promising, the potential of automation is still not fully realized. There is certainly still much more to be done. Consequently, there is still a need for data preparation tools.

IT has traditionally been responsible for locating, gathering, combining, cleaning, transforming, managing, and defining the metadata of various sources of data. However, Stodder says, the execution of these activities is now being carried out by non-IT professionals due to the growing popularity of data science and analytics. Business and data analysts are searching for more intelligent self-service technologies to ease challenges and speed up data preparation procedures. IT, on the other hand, is interested in solutions that can make data preparation easier, boost productivity, and allow IT to better serve users.

The survey conducted by Transforming Data With Intelligence (TDWI) [3] demonstrated that the majority of professionals are proficient at using spreadsheet programs like Microsoft Excel (81 percent). A similar number of people claimed to be skilled in SQL coding and/or programming for reporting-tool scripting languages (79 percent). A lot of participants said they are also capable of using data visualization and presentation technologies (75 percent) and specific BI or visual analytics software (72 percent). 63 percent of respondents have experience using ETL, ELT, data virtualization, and data integration technologies, while 68 percent are proficient in using relational database administration and data extraction tools.

Participants in the study use a variety of BI and analytics systems and technologies. The survey also showed that spreadsheets are the most widely utilized tools. ETL, ELT, or data integration tools took second place. The following were BI/OLAP systems run by central IT. These findings show that standard BI and data warehousing infrastructures

are used to a reasonable but not universal extent.

Another source, the research results of Gregorio Convertino and Andy Echenique [8], provided data on how analysts differed in their analytic skill level and tools use. The descriptions below characterize each user class and their needs.

Data Analyst Scenario: Data analysts are generally distinguished by the amount of time they devote to data analysis and their business skills. Their duties begin with a business problem that must be resolved utilizing the analyst's data. Sales data for a particular area or logistical information from a recent transaction are examples of business queries. Data analysts generally clean and filter the data using spreadsheet programs like Microsoft Excel in order to provide an answer to this question. They then use algorithms, pivot tables, and sample charts to offer solutions to the business difficulties of their clients.

Business Intelligence or Sr. Data Analyst Scenario: BI uses the same processes as data analysts but has access to a wider range of data sources. Senior Data Analysts spend the majority of their time preparing data. However, rather than receiving data from a specified location, Sr. Data Analysts are given more access within organizations and the capacity to acquire fundamental sorts of big data. BI analysts have more versatility in how they display the findings since they can choose from a wide range of business intelligence platforms, such Tableau or Microstrategy.

Data Scientist Scenario: The Data Scientists, who are regarded as the study's most technical users, use the most elaborate data preparation techniques. Data scientists also combine data sets and carry out statistical studies in addition to cleaning and filtering the data. Although they can use similar Business Intelligence tools, data scientists place less focus on representing and reporting data. The cycle used by

data scientists involves multiple cycles of data preparation as opposed to the single clean-represent-report loop.

Therefore, according to the variety of data preparation tools users, tools' main features, advantages and usage [2], it can be said that the main goal of the current research is to help users to select the right ingestion and preparation tool according to their needs and applications' requirements.

When creating an overall strategy, businesses shouldn't disregard the usage of spreadsheets for gathering data and preparation. However, the should consider technologies that could lessen the need for spreadsheets to prepare data, and then learn to use the new tools.

Overall, employing current tools continues to be the most popular strategy, as reported by 76 percent of research participants' firms. This demonstrates that businesses are at least somewhat interested in newer, more advanced solutions that provide better integration of data preparation into front-end self-service platforms. Self-service elements are being implemented with comparable interest by participants. Self-service features are crucial for giving users the ability to combine data preparation with visual analytics and data discovery in a single process.

4 Framework

This chapter will include methodology for future research.

5 Research

This chapter will include the research itself.

6 Attachment (KT1 Schedule)

Work schedule:

1. Determine the set of tools for analysis: second half of February, 2023.

2. Establish the set of features and criteria for comparison: second half of February, 2023.

3. Develop a functional framework for analysis: beginning of March, 2023.

4. Analysis and classification of tools based on features: March-April, 2023.

5. Result presentation and thesis defense: May-June, 2023.

7 Conclusion

Results of the research.

References

1. Joseph M. Hellerstein, Jeffrey Heer, and Sean Kandel. "Self-Service Data Preparation: Research to Practice." IEEE Data Eng. Bull. 41.2 (2018): 23-34.

2. Jaber Alwidian, Sana Abdel Rahman, Maram Gnaim, and Fatima Al-Taharwah. "Big Data Ingestion and Preparation Tools." Modern Applied Science 14.9 (2020): 12-27.

3. David Stodder. "Improving data preparation for business analytics." Transforming Data With Intelligence 1.1 (2016): 41.

4. Ehtisham Zaidi and Sharat Menon. "Market Guide for Data Preparation." Gartner (2020).

5. Gartner. "Peer Insights: Data Preparation".

6. Norman W. Paton. "Automating data preparation: Can we? Should we? Must we?" Proceedings of the 21st International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (2019).

7. Nikolaos Konstantinou. "VADA: an architecture for end user informed data preparation." Journal of Big Data 6.1 (2019): 1-32.

8. Gregorio Convertino and Andy Echenique. "Self-service data preparation and analysis by business users: New needs, skills, and tools." Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (2017).