Helix (2021) 11 (5): 47-54

https://doi.org/10.29042/2021-11-5-47-54

Dimensions of Automated ETL Management: A Contemporary Literature Review

[1]G. Sunil Santhosh Kumar, *[2]M. Rudra Kumar
[1]Research Scholar, Department of CSE, JNTU Ananthapuramu
gsunilsanthosh105@gmail.com
[*2]Professor, Dept. of CSE, GPCET, Kurnool, A.P, India
drrudrakumarcse@gpcet.ac.in

Received: 3rd August 2021, Accepted: 18th October 2021, Published: 30th October 2021

Abstract

Emerging dynamics of data systems and reliance on quality data sources, data processing for informed and strategic decision-making enhance the scope of using the ETL solutions. In the current scenario, one of the critical aspects focused on software engineering is about focusing on using the data management tools that can help gain insights for functional and operational aspects. While many academic and industrial research studies have focused on data management dynamics and the application of ETL tools as a profound solution, there is an imperative need to upscaling the ETL efficiency over real-time applications. In this literature review, the scope of the current ETL frameworks, limitations, and scope are discussed. Categorically, the objective is to explore if the machine learning models are adapted in the ETL systems. However, from the literature review, it is evident that many academic studies have advocated using machine learning models to improve and optimize the use of ETL solutions. But very few tools in the market are using the comprehensive range of machine learning models in ETL processing. Focusing on the current constraints and the scope for improvement, this study advocates the need for designing and developing machine learning-based models for ETL-based data management optimization. If such processes could be developed, it can help the organizations have potential systems in place for decision-making.

Keywords: *Data quality (DQ), Entity Relationship Diagram (ERD), Transform-Extract-Load, Information systems, Extract Transform and Load (ETL)*

Introduction

Information systems have become an integral part of business operations. With every sector and business focusing on the e-business models, the scope and reliance on the organization's information and enterprise systems have become an imperative need. With a distinct set of technological solutions available, the organizations must choose the right kind of systems and practices paramount to managing their information systems for better operational excellence [1].

The quality of services from the applications used as information systems in the organization relies on the applications' build quality and the effective data management integrated into the application systems. The domain of software engineering, which plays a vital role in the development of software applications, is the combination of tools, practices, processes, techniques, technology, and human resource experience in handling the development of a niche range of software applications, wherein the data stands pivot at the development stage and implementation of the application [2]. In a simple instance, developing a website for a small business can be handled in distinct scenarios. One simple process is using a plug-play theme-based web system, selecting a theme, fixing the content about the organization, and launching the site within hours. The other level is working on the requirements, understanding the scope, choosing the right kind of technology, and getting the website developed in a customized fashion. At a profound level, the other opportunity for the process is about developing a project requirement document, develop the scope document, work a project charter for the project, in terms of the release for the

application in the first stage, inclusions to the second stage, etc. and accordingly develop a more comprehensive system that can be resourceful for the organizations. Based on the case mentioned above, it is pragmatic to presume the wide scope of aspects integral to developing an application system. Thus, with the role of data management being impeccable in developing a modern system, it is of paramount importance that the scope of executing the software engineering has to be more robust and significant [1] [3]. The domain of software engineering has evolved, and there are phenomenal ways in which software solutions have emerged for data handling. Some data management solutions like the ETL (Extract Transform and Load) approaches are gaining prominence and still evolving in terms of best practices or models that can enable the optimum utilization of resources and dynamics to improve the teams' operational excellence in software engineering. Right from developing a simple operational research-based implementation plan to machine learning models, a profound set of developments are used to improve the overall process of managing software engineering [4].

Despite such developments and numerous implementation case scenarios, one of the critical challenges that face the software engineering teams is about optimum utilization of data sources, developing an internal ecosystem practice wherein the multiple stakeholders involved in the process can address the requirements and gain insights. In general, effective software engineering practices are the combination of having competent resources, technologies that can support effective outcomes, and appropriate kind of project management practices to ensure an effective and successful outcome for the business process. Thus, taking such factors into account, it is of paramount importance that the businesses focus on combining factors that lead to a quality outcome. Some of the earlier proposed models for ETL and currently in real-time practice, like the machine learning models, can help the project management teams adapt the best possible solutions for data handling, which can play a significant role in the enterprise systems [5], [6]. Thus, this recent research literature review focuses on applying the machine learning models in the ETL process, certain key elements of the software engineering process like the techniques, tools, systematic approaches, and wider implementation scope. Also, assessing the current set of machine learning models that can be resourceful for the ETL management domain, the model's emphasis is about understanding the trends and assessing if there are any specific gaps in the model that needs to be addressed with more vehement solutions [5].

In the other sections of this report, the literature review related to key elements of ETL models and machine learning models' role in the ETL engineering systems is discussed. Based on the literature review, the gaps, if any, or the problems observed with the current process shall be discussed. Followed by the objectives for the future research scope are presented in the report.

Software Engineering

Software engineering can be defined as a process wherein analyzing the user requirements and designing, developing, testing, and roll-out of a software application shall address the requirements effectively [3]. Some of the significant definitions for the software engineering imperative from the literature are as per the IEEE standards 610.12-1990, software engineering is a systematic, disciplined, and computable approach towards developing, managing, and operating software. Boehm defines software engineering as "the practical application of scientific knowledge to the creative design and building of computer programs. It also includes associated documentation needed for developing, operating, and maintaining them." Fritz Bauer refers to software engineering as the establishment and usage of standard engineering principles. It supports in attaining the software which is reliable and effective on real machines. In line with the definitions mentioned above and pragmatic application of software engineering, software engineering can be advocated as a systematic process that enables the developments right from the conceptual stage to the implementation stage for an application system development. While the process from the conceptual stage to implementation can be executed in any manner, adapting the right steps and procedural developments can lead to optimal utilization of resources, developing more comprehensive solutions, which can yield better results [5], [6]. Handling software engineering successfully comprises scores of data to be handled as integral to the system development and management. It is important that the organizations have to work on meaningful solutions that can improve the overall data management, which can help in effective decision making. For instance, the simple approach of following the WBS (work breakdown schedule) for the project tasks helps in

having a right understanding of the task duration, predecessors and impact, resource requirements, etc. In the absence of such structured data, following the process flow and exercising better control of the process can be a complex challenge for organizations. Thus, there is a need for the organizations to focus on garnering an adequate set of data and effectively using the data for managing the software engineering projects [3], [5].

Data Handling in Software Engineering

The digital trends prevailing in the industries lead to significant volumes of data being generated, assessed, and updated. Effective use of such data management is critical for software engineers. Such data can work as a potential and actionable for improvisation in the software engineering process and its accurate delivery [6]. In the other dimension, effective ways in which the organizations garner the data indicate how the teams can monitor the work progress, the quality, and the quantum outcome. Towards effective project management and in software engineering management, such data processes help effectively. In some of the project management organizations, metrics-based key performance indicators termed KPIs are used as deliverable metrics to assess whether the systems and practices adapted in the model result in effective outcomes during the project in progress. In general, KPIs and KPIs can be the highly resourceful value-driven mechanism for software engineering of the ETL frameworks. Unless the ETL tools are designed more in customization to the functional flow essential for the organization, the outcome can be challenging [6].

In an illustrative scenario, for aviation software to be integrated into aircrafts controls and ground communication, there are thousands of vital data points for successful implementation. In terms of improving the software quality, accessibility to earlier software solutions, lags, pros, and cons can certainly help improve the quality of the current application development. However, one of the challenges remains in terms of how much data to be assessed or what set of data can be seen as more vital in data management in software engineering. As too much analysis could lead to more complexities and time-consuming processes, it is paramount that the data is garnered inaccurate and adequate volumes, which can provide maximum insights into the system. Thus, the ETL systems' execution for the aviation companies or airline operations or aircraft controls needs more customized data management requirements. In [6], the studies highlight the need for the team-work approach and metrics-based program as the critical factor that can garner adequate and appropriate information.

The study discusses the importance of resources being vigilant about the data and coordinating adequate data retrieval for improving the work. The critical point highlighted in the study is about how the human side of the software engineering process is of paramount importance for addressing the KPI or OKR initiative. The other study focusing on the metrics-based approach in software engineering has highlighted the conditions [7] as when the KPIs are not effectively executed, the challenges are imperative in the form of the design and development of the ETL application shall be more centric to the deliverables from the management perspective rather than focusing on the technicalities and technical feasibilities integral to the process. The other critical point highlighted in the study about the metrics-based approach is consensus management among the various stakeholders. In the absence of an effective mechanism, developing consensus among the stakeholders, approvals for the process can be time-consuming. It refers to the conditions wherein the channels of processes become hurdle for completing the project. In terms of improving the conditions of applying the metrics as the data for software management, careful selection of the metrics and assessments must be vital.

Authors of [8] have developed a pictorial representation of the key element’s integral to data-enabled software engineering work. The figure 1 refers to intrinsic ways in which the data-enabled software engineering work is integral to the system. The depiction reflects a broad scenario of the software engineering process based on the data-enabled conditions; however, the company vision can be seen as what the objective is in terms of micro-level. The positive behavioral metrics refer to the context of how the team is targeting the development and the selection of data sources accordingly, alongside the software metrics depicting the scope of compliance factors, standards, and the best practices that need to be incorporated for developing a sustainable application system.

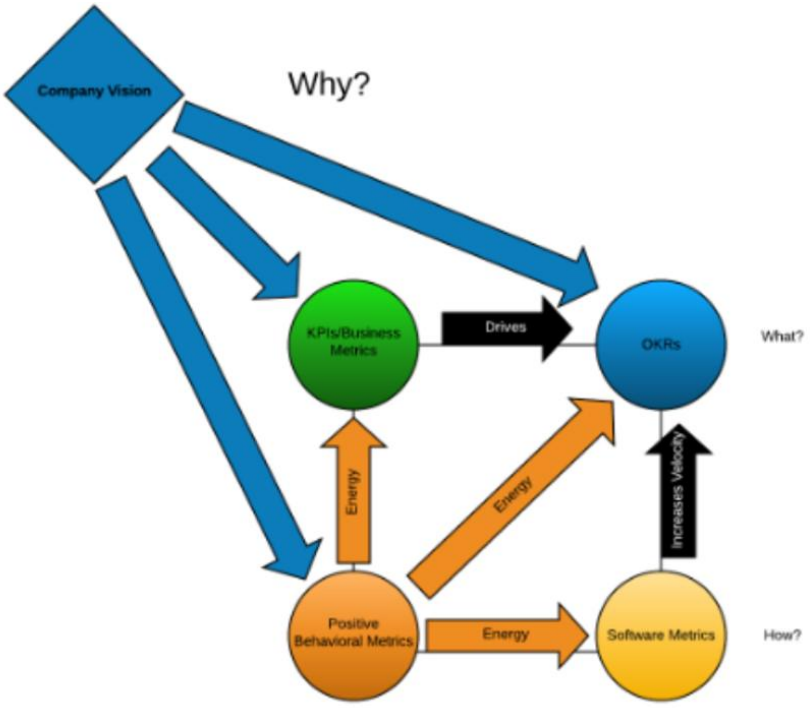

Figure 1: The data-enabled software engineering work is integral to the system

While many research studies support distinct set of solutions for improving data management in software applications, some of the studies highlighted the challenges, models, and solutions that are more relative to the usage of data management in the software engineering process.

ETL Systems

In [9], the authors have reviewed the importance of data quality (DQ) in the ETL process. By focusing on the four key solutions vividly used in the industry for extraction, transformation, and loading process, the data analysis carried out in the study refers to the conditions wherein the ETL processes are managed effectively. The experimental analysis of the three critical components is like the performance, reliability, and deduplication factors. However, the challenge remains in developing a pragmatic, systematic approach wherein the ETL role is effectively managed. The authors of the study claim that the open-source tools tested in the experimental analysis have many limitations in processing the NoSQL data scenarios. Iterating the importance of the non-relational models like the key-value, document-centric, graphs, or symbolic representations centric and column related scheme less representation, the models currently adapted has significant constraints. The critical issue is how the shortcomings could be affected ting the performance. There is an imperative need for the ETL framework that and be effective as on-demand and extensible conditions. In [10], the study's author shave discussed the scope of a new framework for data warehouse ETL processes. Focusing on the key aspects of the ETL process as the source area, destination area, and the mapping area, the model targeted the improvement of using the ETL processes. The study underlines the standardization of the models for source and destination as ERD (Entity Relationship Diagram) and the Start Schema, respectively, and the absence of any structured model for the mapping area. The model proposed in the study is profoundly targeting the feature selection for developing a conceptual model for the mapping procedures. The authors of the study proposed EMD (entity mapping diagram) as an effective model. Critical functionality proposed in the model is about handling the mapping part wherein the required transformation functions are garnered and are performed based on the incoming data from the base source or the provisional results handled over temporary tables in the staging area. The study fails to exhibit any experimental analysis of the model. Still, it has proposed the model with a traditional two-layer approach as an abstraction layer and the expansion to the abstraction

layer. Also, the study highlights the flexibility for the users to add more layers as deemed necessary. One of the key aspects imperatives in the model is how the ETL process layer of mapping needs importance and the mechanism that can be focused on such robust development. In their study, the authors of [11] iterate the importance of choosing the right kind of source for the data extraction. Emphasizing the importance of how the source data from distinct sources, if not cleaned effectively, could be misleading the complete process, propose the model of query processing for the data extraction. Conceptualizing the model based on the data in terms of managing the active and non-active query processing. The study claims that the proposed model, when implemented in the right execution with the other metrics, can reduce the response time and improve the data warehouse's overall performance. One of the important points highlighted in the study is how important the scope of using query processing is for getting the right kind of data. In the structured and unstructured conditions, the query processing needs to be much accurate in terms of retrieving the adequate and appropriate kind of data. This study stands significant as a limited set of studies have targeted the scope of effective query processing in ETL-related studies. Testing the models and their performance is a more effective part. While the testing of the models in the simulated environment can refer to the system's efficacy, the system's actual efficiency is imperative only in the real-time implementation. In [12], the authors of the study have discussed the importance of testing and the various set of testing techniques that are integral to testing ETL models and frameworks. The study elaborates and provides very insightful views on the distinct set of testing models, the type of data for tests, and the anticipated time duration required for the testing process. The study is more informative over the scope of testing in the ETL domain. However, focusing on this study's objectives to understand the emerging developments and effectiveness of the ETL frameworks, the models referred to in this study can help develop the experimental analysis for any new solutions developed in the future.

A review model [13] reviews the existing models and frameworks for the ETL processes by highlighting the various challenges, constraints, and positives for the process at each of the stages. By focusing on the critical issues, the study also has discussed the pros and cons of various models proposed earlier. Based on the literature's insights, the study highlights the strengths of various models and their effectiveness for the system's practical application over a period. Some of the common issues highlighted in the model are about pivoting, data mapping, and data lineage. When implemented in real-time conditions, it is of paramount importance that the models need to exhibit the accurate processes that can help meet the objectives of all the three stages like extraction, transformation, and loading.

A review of the study offers insights into wide gaps in the process management at all three critical stages and the need for the right kind of systems to improve the model's overall process outcome. The first insight from the model concerns the concerns related to the construction of commonly accepted conceptual and logical modeling tools for handling the ETL processes using a standardized approach. The second critical factor considered in the model is the efficiency of the individual ETL operators. In [14], a new solution in TEL (Transform-Extract-Load), a novel approach, is proposed for ETL. The model relies on developing a virtual table for realizing the transformation stage before the extraction stage. Some of the highlights mentioned for the model are about adapting the scope of reducing the transmission load factor and improving the performance query from the accessible layers. Also, in terms of experimental results imperative for the model, the benchmarks are indicated for the model's feasibility. While the models reviewed in the above section refer to the conventional solutions for ETL processing, the study's objective is also to understand the scope of reducing the human intervention in the data handling processes and how ETL processes can be automated for significant improvements in the system. One of the key models that can be adapted for the automation of the ETL process is applying the machine learning models that can help improve the overall solution.

In [15], a detailed review of how ETL processes' automation can be handled for the respective domain is discussed in detail. Based on the inputs assessed in the ETL requirements for various domains like financial and marketing, the study proposes the possible architecture that could be considered for the process's automation. Although the model refers to various key aspects to be considered in the systematic approach, the research paper does not discuss any specific model or an experimental outcome for the scope discussed. Thus, the study can be seen as a qualitative

insight into the scope, technicalities, and metrics that could be seen as a potential system for automation of the ETL process across the domains.

A quantitative study assessing the application of conventional models and the automated models for the ETL implementation refers to some of the conventional solutions used in the process and how the modern-day implementations envisage the gap of automated systems for managing the process outcome. It is of paramount importance that the conjunction of human efforts and the machine learning models' capability is explored effectively, ensuring optimization of the results in the process. Though the study has limited scope in detailing the processes, there are various key areas depicted for implementing the automated systems in the ETL models [16]. Another study focusing on the ETL models refers to the pragmatic application of how the semantics of new data sources can map the sources to the common data model. While the complexities of the semantic and heterogeneities are evident, it offers a well-established approach for managing the facets of clean datasets, which are viable for machine learning systems. Thus, the study [17] refers to the dynamic ETL framework model, which supports integrating multiple data sources over real-time to generate necessary results, which can support value and impact for an automated approach.

Authors of [18], in a detailed review of the conditions leading to the ETL limitations and challenges, point to the key areas wherein the software engineering solutions are facing constraints for dynamic solutions. Targeting the design and the performance optimization of the ETL solutions, various factors are considered pragmatic by the organization for improving the ETL executions for the organizations over a real-time environment. While managing the costs of operations is one of the key discussions, the key recommendation is about the adoption of machine learning models for the ETL solutions implementation. In [19], the study profoundly targets the marine systems environment and identifies the gaps in handling the heterogeneous data schema mapping model that depends on the multi-analyzer machine learning solutions. The analysis refers to the conditions wherein the mapping model has the underlying solution in a fuzzy comprehensive evaluation system for analyzing the multifactor quantitative judging conditions. The study claims improvement with the model for addressing the error-rate issues.

Gap Analysis

ETL systems have become an essential need for large-scale computation processes and solutions. There is a need for organizations to ensure the processes of ETL solutions are improvised. Accordingly, developing a more comprehensive range of ETL solutions can be launched to enhance software engineering quality [20]. Across the business verticals, there are scores of performance data for the business operations available. If the ETL solutions play a vital role in ensuring appropriate data is sourced for managing the operations. For instance, in the aviation systems, or the in-healthcare services, there are various sets of data and data sources that can help the organization make informed decisions in the process. However, when the organizations do not extract the right kind of information or have constraints in garnering the information from data sources, it shall affect the decision-making capabilities [20], [15].

There are a profound set of practices wherein the organizations have many dimensions. The analysis can strengthen the performance of the internal systems and improve the system's overall efficiency. Considering the current practices of ETL solutions, which are more of a manual or framework-based systematic approach, implementing the machine learning models in the ETL solutions can be a profound experience. Various aspects are essential in addressing data management with an ETL process [17], [19]. As discussed in the literature review, at every level of the ETL process, there are certain levels of complexities that are integral to the functions. For instance, at the extraction level, focusing on the semantic and syntactic models is important to garner adequate and accurate data. Based on the information collated in the model, at the storage stage, the processing of data into a structured format requires choosing the right set-up, formats, and the scope for using the data in multiple dimensions [20].

In an illustrative scenario, the medical health records of individuals collected from various sources using the extraction process can be resourceful for formulating into different structures. Right from using the same data for demographic analysis to the disease symptom analysis or the treatment-related analysis, there are dimensions in

which the data can be analyzed. The organizations must choose the right data transformation format for each objective analysis and load the data for visual representation and detailed analysis [18], [21]. However, the challenge remains in the current forms about how the ETL solutions have limited machine learning models. Thus, it is evident that more significant ways in which the machine learning models can be adapted to implement the ETL process effectively. In a summary of the gap analysis, based on the literature review, it is to admit about the limited action of machine learning in the ETL sphere. While many studies have recommended a holistic level about the scope of implementing the machine learning and AI models for ETL process automation, some of the market's ETL tools are offering such services. However, in terms of the system's scope and efficacy, there is a wide gap for analysis. If the companies can target the right kind of solutions, it can help achieve the process outcome successfully.

Conclusion and Future Research Scope

Focusing on the future trends, with most organizations relying on the big data kind of solutions, there is a need for more robust analytic solutions to support effective and efficient decision-making. With the proven mettle of AI and machine learning models across the verticals, it is evident that the right kind of implementation in the ETL systems can turn out to be an effective solution. Although most of the studies advocated using the machine learning models for ETL implementation, an extremely limited range of studies have explored the actual dynamics of implementing the AI integration to the ETL domain. Some of the conventional approaches to the ETL frameworks have highlighted the pros and cons. In [6], the study has highlighted how reliability, deduplication, and performance are the critical factors alongside many other features important for analyzing the ETL performance. Focusing on such features and models, if the machine learning processes can be used for detailed analysis, can help in improving the overall efficiency of using ETL systems in the engineering of software applications.

Thus, the following are some of the key objectives considered for exploring the scope of using the AI-based solutions in the ETL management for the software applications.

- Automating the processes of extraction, transforming, and loading of the progression data of the software engineering strategies using machine learning
- Handling changes impacts and concept drifts of the progressive or streaming data in ubiquitous and pervasive computing software engineering strategies.
- The research includes defining and developing the measures related to feature selection and optimizing supervised and unsupervised learning methods to automate data handling, managing, and storing in ubiquitous and pervasive computing.
- Change and Drift impact tolerance shall achieve in automate progressive data handling in software engineering strategies.

Exploring the scope of contemporary models that align with the objectives above can help improve overall sustainable developments.

References

[1] Vyas, S., & Vaishnav, P. (2017). A comparative study of various ETL process and their testing techniques in data warehouse. Journal of Statistics and Management Systems, 20(4), 753-763.

[2] Rahman, N., Kumar, N., & Rutz, D. (2016). Managing application compatibility during ETL tools and environment upgrades. Journal of Decision systems, 25(2), 136-150.

[3] El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. Journal of King Saud University-Computer and Information Sciences, 23(2), 91-104.

[4] Machado, G. V., Cunha, Í., Pereira, A. C., & Oliveira, L. B. (2019). DOD-ETL: distributed on-demand ETL for near real-time business intelligence. Journal of Internet Services and Applications, 10(1), 1-15.

[5] Satkur, A. M. (2013). A Review Paper on scope of ETL in retail domain. Internation al Journal of Advanced Research in Computer Science and Software Engineering, 3(5).
[6] Wang, T., Hu, J., & Zhou, H. (2011). Design and implementation of an ETL approach in business intelligence project. In Practical Applications of Intelligent Systems (pp. 281-286). Springer, Berlin, Heidelberg.
[7] Golfarelli, M., Rizzi, S., & Turricchia, E. (2011, August). Modern software engineering methodologies meet data warehouse design: 4WD. In International Conference on Data Warehousing and Knowledge Discovery (pp. 66-79). Springer, Berlin, Heidelberg.
[8] Gupta, P. (2016). Data warehousing and ETL Processes: An Explanatory Research.
[9] Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. B. (2019). Data quality in ETL process: A preliminary study. Procedia Computer Science, 159, 676-687.
[10] El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. Journal of King Saud University-Computer and Information Sciences, 23(2), 91-104.
[11] Gour, V., Sarangdevot, S. S., Tanwar, G. S., & Sharma, A. (2010). Improve performance of extract, transform and load (ETL) in data warehouse. International Journal on Computer Science and Engineering, 2(3), 786-789.
[12] Vyas, S., & Vaishnav, P. (2017). A comparative study of various ETL process and their testing techniques in data warehouse. Journal of Statistics and Management Systems, 20(4), 753-763.
[13] Vassiliadis, P. (2009). A survey of extract–transform–load technology. International Journal of Data Warehousing and Mining (IJDWM), 5(3), 1-27.
[14] Guo, S. S., Yuan, Z. M., Sun, A. B., & Yue, Q. (2015). A new ETL approach based on data virtualization. Journal of Computer Science and Technology, 30(2), 311-323.
[15] Mondal, K. C., Biswas, N., &Saha, S. (2020, January). Role of Machine Learning in ETL Automation. In Proceedings of the 21st International Conference on Distributed Computing and Networking (pp. 1-6).
[16] Nwokeji, J. C., Aqlan, F., Anugu, A., &Olagunju, A. (2018). Big Data ETL Implementation Approaches: A Systematic Literature Review (P). In SEKE (pp. 714-713).
[17] McCarthy, S., McCarren, A., &Roantree, M. (2019, October). A Method for Automated Transformation and Validation of Online Datasets. In 2019 IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC) (pp. 183-189). IEEE.
[18] Ali, S. M. F., &Wrembel, R. (2017). From conceptual design to performance optimization of ETL workflows: current state of research and open problems. The VLDB Journal, 26(6), 777-801.
[19] Yan, W., Jiajin, L., & Yun, Z. (2014). A multianalyzer machine learning model for marine heterogeneous data schema mapping. The Scientific World Journal, 2014.
[20] Mondal, K. C., Biswas, N., &Saha, S. (2020, January). Role of Machine Learning in ETL Automation. In Proceedings of the 21st International Conference on Distributed Computing and Networking (pp. 1-6).
[21] Mali, N., &Bojewar, S. (2015). A survey of ETL tools. International Journal of Computer Techniques, 2(5), 20-27.