 ISSN:2302-8149
e-ISSN:2540-9719

Design and Implementation of an ETL Pipeline for Prospective Student Data Analysis in Higher Education Admissions

[1]Nina Setiyawati*, [2]Dwi Hosanna Bangkalang, [3]Gilang Windu Asmara
[1]Department of Informatics Engineering, Satya Wacana Christian University
[2]Department of Information System, Satya Wacana Christian University
[3]Marikh Prigel Technology
[1,2]Jl. Dr. O. Notohamidjojo No.1 - 10, Blotongan, Kec. Sidorejo, Kota Salatiga, Jawa Tengah 50715, Indonesia
*e-mail: nina.setiyawati@uksw.edu

(***received***: 14 March 2025, ***revised***: 8 June 2025, ***accepted***: 8 June 2025)

Abstract

The number of universities in Indonesia continues to grow. This condition certainly makes the flow of new student admissions increasingly competitive between universities, thus encouraging universities to do branding, show quality, and do the right positioning. Therefore, it is important for universities to adopt a data-driven approach that can provide in-depth insights into prospective students and the effectiveness of marketing strategies. The purpose of this study is to design and build an ETL (Extract, Transform, Load) pipeline to collect, process, and analyze prospective student data as part of the business intelligence (BI) system to be built. The proposed ETL architecture design supports automated microservices-based data transformation in data cleaning, normalization, and integration. In addition, it can also be used as a solution to increase the scalability and flexibility of data mobilization in the BI system. This study introduces a novel approach by designing an ETL pipeline within a business intelligence framework aimed at enhancing university marketing efforts. Unlike prior research, which has primarily applied business intelligence tools to evaluate academic activities within learning management systems, this work shifts the focus to marketing analytics. Additionally, while existing studies on higher education marketing often center around digital marketing techniques and the marketing mix, this research fills a gap by proposing a technical infrastructure that supports data-driven marketing through automated ETL processes. The resulting ETL was tested using several methods, namely Source to Target Count Testing, Source to Target Data Testing, Duplicate Data Check Testing, and Data Transformation Testing. The results of each test are valid.

Keywords: data analysis, prospective students, business intelligence, ETL pipeline, data driven marketing

1 Introduction

The number of universities in Indonesia continues to grow. In 2022 there will be more than three thousand state and private universities [1], [2] in Indonesia. This condition certainly increases the flow of new student admissions competitively between universities which also encourages universities to foster their branding, accentuate their quality, and establish the right positioning. The use of digital for university marketing has also increased [3], [4]. This situation leads into the importance of marketing and market understanding roles for higher education institutions, not only to maintain the number of existing students, but also to attract new prospective students.

Admission process is one of the key aspects in marketing for higher education institutions [5]. An effective admission process can help institutions attract the right prospective students who fit the targeted profile. On the other hand, traditional approaches are often inadequate to capture the complexity and dynamics of today's education market. Therefore, it is important for universities to adopt a data-driven approach [6], [7] that can provide in-depth insights [8] into prospective students and the effectiveness of marketing strategies. Prospective student data collected during the admission process contains potentially useful information for decision-making [9] [10]. By analyzing prospective student data, institutions can profile and segment the prospective students, identify

 ISSN:2302-8149
e-ISSN:2540-9719

patterns in study program selection, and even predict the likelihood of graduation or academic success [11].

For those reasons, there are two formulated research problems in this study: 1) how universities can conduct data-based marketing; and 2) how to design and build a marketing analysis system architecture that can maintain and even improve the quality of data obtained from primary sources and ensure that the data is consistent, relevant, and accurate, which meets all university marketing needs and also helps make better decisions.

This study aims to design and implement an Extract, Transform, Load (ETL) pipeline to facilitate the collection, processing, and analysis of prospective student data, as a foundational component of a business intelligence (BI) system to be developed. This research distinguishes itself by focusing on the design and implementation of an ETL pipeline within a business intelligence framework specifically oriented toward university marketing. While a number of prior studies have addressed the utilization of business intelligence, their emphasis has predominantly been on analyzing academic activities within learning management systems. For example, the study *A Business Intelligence Framework for Analyzing Educational Data* [12] employed Excel-based datasets processed using the WEKA tool to identify unemployed students and evaluate their academic performance. Conversely, existing literature on marketing in higher education has mainly explored themes such as digital marketing practices [13], [14], and the marketing mix [15], with limited attention to the underlying data infrastructure. In response to this gap, the proposed ETL architecture adopts a microservices-based communication model and incorporates automated mechanisms for data cleansing, normalization, and integration, thereby enhancing the efficiency and scalability of business intelligence applications in the context of university marketing.

2 Literature Review

ETL Pipeline is a structured approach to managing and processing data, consisting of three main stages [16], [17], [18]:

a) *Extract* (data collection). This stage involves collecting data from various sources, such as registration forms, prospective student databases, surveys, and external data such as demographic information. The data are stored in several distinct databases and are structured in the form of relational databases. The volume of extracted data consists of approximately 3,000 records. Integrated and comprehensive data collection is key to getting a complete picture of prospective students. The extraction method employed is incremental extraction.
b) *Transform* (data processing). Once the data is collected, the transformation stage is carried out to clean, integrate, and normalize the data. This process is important to ensure that the data used in the analysis is consistent and reliable. Transformation also involves combining data from various sources into a uniform and easily analyzed format.
c) *Load* (data transfer). At this stage, the processed data is loaded into the BI system to be used for analysis. BI systems provide a variety of data analysis and visualization tools that allow universities to explore patterns, trends, and insights that can support strategic decision making. This architecture utilizes a full data loading approach.

ETL serves as a foundational component of data warehousing systems [19]. It plays a crucial role in filtering out irrelevant data, correcting inaccuracies, and ensuring data quality. Additionally, ETL processes support the documentation and quantification of data, facilitating effective measurement. Furthermore, ETL captures operational data flows for custodial and auditing purposes, and it harmonizes data from disparate sources to ensure it is accessible and usable by end users [20].

There are several studies related to the use of ETL in the academic world which include: 1) analysis of teaching and learning approaches in higher education [19]; 2) analysis of education cost levels [20], [21]; 3) analysis of student graduation [22]. However, there has been no research that applies ETL to university marketing context. In this study, ETL is employed to analyze data related to prospective student admissions based on data extracted from the university's existing transactional databases.

 ISSN:2302-8149
e-ISSN:2540-9719

3 Research Method

The research workflow begins with identifying and analyzing the requirements, followed by designing the ETL pipeline architecture, which is subsequently implemented. The final stage of this study involves testing the ETL pipeline. The detailed research workflow is illustrated in Figure 1.

Figure 1. Research flow

1. Identification and Needs Analysis

This stage aims to identify the needs of the university to define and identify the scope of high-level requirements, business processes, confirmation of the scope of the system being developed. The outputs of this stage are: a) analysis of resources and system requirements (software, hardware, infrastructure, and support); b) system requirements specifications; c) inception report.

2. ETL Pipeline Architecture Design

Following the requirements analysis, the solution was defined and the existing technical constraints were identified. Subsequently, and ETL pipeline architecture is designed to align with the characteristics of the source database and other relevant data sources. As a result, the output of this stage is a comprehensive architectural blueprint.

3. Implementation of ETL Pipeline and Data Warehouse

In the subsequent phase, the previously designed ETL pipeline was implemented, and the data warehouse architecture defined in the earlier stage was constructed. This phase resulted in the deployment of both microservices and the data warehouse as core components of the data integration framework. The microservices are designed based on the Representational State Transfer (REST) architectural style, which is widely adopted due to its statelessness, scalability, and compatibility with distributed systems. REST operates over the HTTP protocol using a request-response communication model [19]. The integration endpoints expose standard RESTful methods—namely GET, POST, PUT, and DELETE—returning data in JSON format for interoperability across heterogeneous systems [20].

4. ETL Pipeline Testing

Upon completion of the ETL implementation, the system underwent a testing phase to evaluate its functionality and performance. Three types of test were conducted in this study, namely: 1) Duplicate Data Check Testing (DDCT) to check for the existence of duplicate data in the target system. When a large amount of data is present in the target system, the risk of data duplication in the production system increases, which can negatively impact the accuracy of analytical test results [23]; 2) Source-to-Target Data Testing (STDT) to verify the accuracy of data transfer from the source to the target system. This involves checking the validity of the data after the data has undergone transformation, both in the source system and the target system [24]; 3) Incremental Testing (InT) to verify old and new data gradually [25].

4 Results and Analysis

The result of ETL design is an architectural blueprint as shown in Figure 2 and Figure 3. Figure 2 presents the constructed ETL pipeline architecture. The raw data is taken from the initial database/resource owned by the university which consist of several databases and applications. Therefore, data ingestion and data integration processes are essential. Following this, a data warehouse is developed, and a provisioning process is conducted to prepare and configure the necessary resources for operational use.

 ISSN:2302-8149
e-ISSN:2540-9719

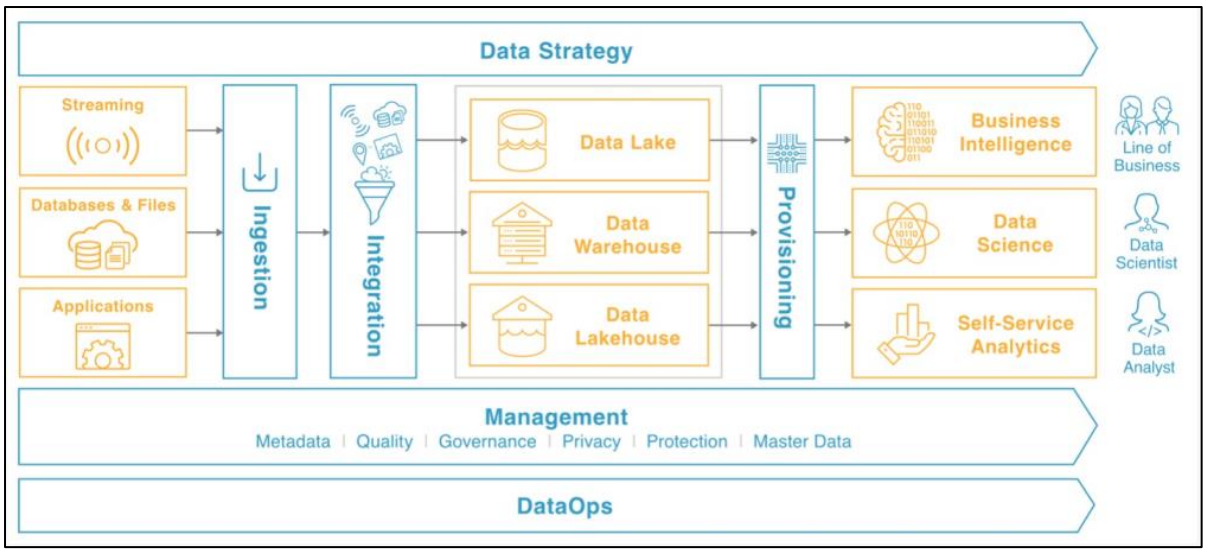


Figure 2. Blueprint of ETL pipeline architecture

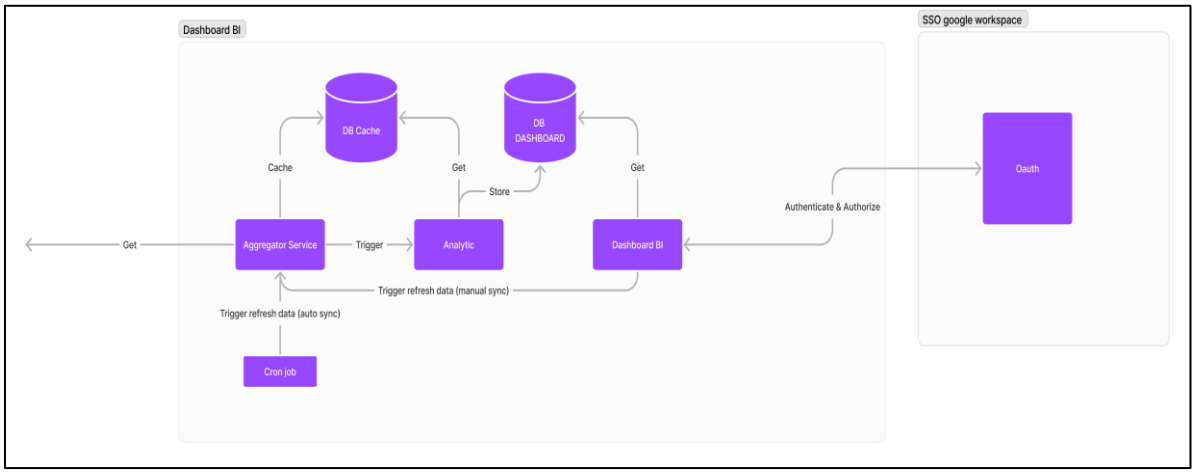


Figure 3. Data communication architecture in the system

Figure 3 shows the developed ETL data communication process. In the system, an aggregator service is implemented to retrieve data from various resources. To facilitate the automation of this retrieval process, a cron job is employed to execute predefined commands as scheduled intervals. Data retrieved from resources is stored in temporary storage, from which it is subsequently accessed for analysis. Then, the data is stored in the dashboard database where it becomes available for visualization and interpretation within the BI system.

In the implementation stage, ETL Pipeline and Data Warehouse produce microservices. One of the microservices produced is shown in Figure 4, namely a service to calculate the number of students per study program based on the school of origin.

 ISSN:2302-8149
e-ISSN:2540-9719

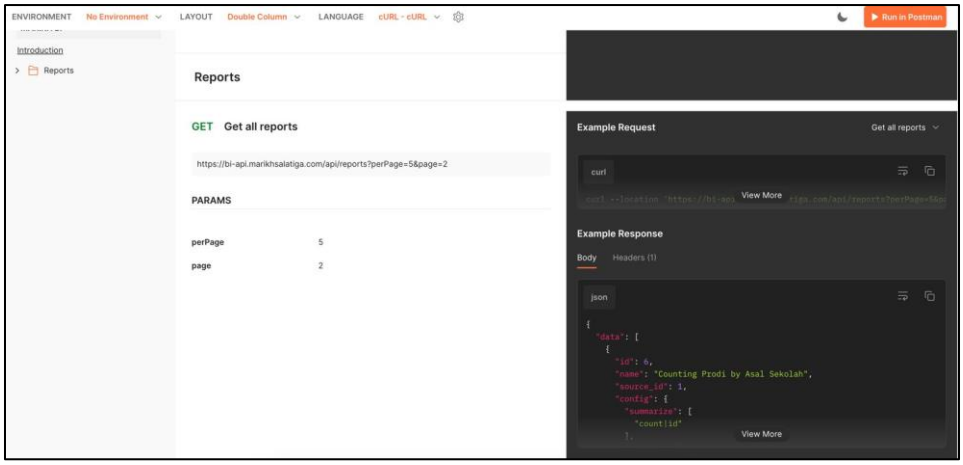


Figure 4. Service calculate the number of students per study program based on school of origin

After designing and implementing ETL, ETL testing is carried out to ensure the accuracy, completeness, and consistency of data when going through the ETL process [23]. In this study, several ETL testing techniques were carried out, where the test results are shown in Table 1. The testing phase was carried out using a dataset consisting of approximately 3,000 records to evaluate the performance and reliability of the implemented ETL pipeline. The tests carried out were:

1) Source to Target Count Testing
   This technique involves comparing the number of records in the source system with the number of records loaded into the target system. If the numbers do not match, it indicates that some data may be missing or incomplete. This test is part of data completeness testing.
2) Source to Target Data Testing
   This test is to verify the accuracy of data transfer from the source to the target system. In this test, the validity of the data is checked after undergoing transformation, both in the source system and the target system connected to it.
3) Duplicate Data Check Testing
   This test is to check for the presence of data duplication in the target system.
4) Data Transformation Testing
   This testing approach is time consuming because it involves executing multiple SQL queries for each individual record to verify the accuracy of the transformation rules. After executing the SQL queries, the resulting data in the source and target systems is compared.

Table 1. ETL testing results

No	Testing	Query dan Results	Results

1	Source to target count testing	`mysql> select count(*) from data_sources;`
		`+----------+`
		`| count(*) |`
		`+----------+`
		`| 3380 |`
		`+----------+`
		`1 row in set (0.00 sec)`
		`mysql>`
		count data report
		total
		3380
		Menampilkan 1 data
			Valid

 ISSN:2302-8149
e-ISSN:2540-9719

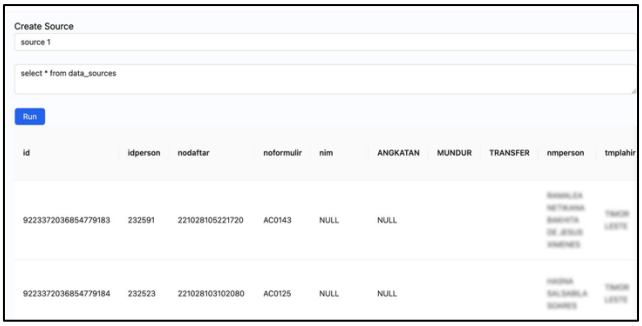
2	Source to target data testing		Valid
3	Duplicate data check testing		Valid
4	Data transformation testing		Valid

Table 1 (the ETL Test Results) indicates that the source system has the same number of data records loaded in the target system. All the data in the both systems are valid and no data duplication found. Furthermore, the data results in the source system and the verified system are in accordance with the transformation rules. Hence, the aforementioned results above are in line with Lokaadinugroho, et al. who also conducted research related to ETL and data warehouses in universities. The study showed that ETL and data warehouses reduce data errors [28]. As supported by Edhya and Susilowati, ETL provides a more efficient data transformation process, more structured data, and is safer from human error [29].

5 Conclusion

This study has developed and implemented an ETL architecture that serves as the foundation for a data warehouse aimed at analyzing prospective students. The ETL processes integrates data from multiple university-owned databases. The testing results indicate that the data volume is consistent, the accuracy of transferred data is validated, and no duplication records are found. The established ETL framework would serve as the basis for the subsequent development of a comprehensive business intelligence system.

Acknowledgement

This research was funded by the Vice Chancellor for Research, Innovation, and Entrepreneurship of Satya Wacana Christian University through the Directorate of Research and Community Service for the applied research scheme.

 ISSN:2302-8149
e-ISSN:2540-9719

Reference

[1] Badan Pusat Statistik, "Jumlah Perguruan Tinggi1, Dosen, dan Mahasiswa2 (Negeri dan Swasta) di Bawah Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Menurut Provinsi, 2022," 2023.

[2] C. M. Annur, "Jumlah Perguruan Tinggi di Indonesia Capai 3.107 Unit pada 2022, Mayoritas dari Swasta," 2023. Accessed: Jul. 03, 2024. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/03/01/jumlah-perguruan-tinggi-di-indonesia-capai-3107-unit-pada-2022-mayoritas-dari-swasta

[3] B. Williamson, "Making Markets Through Digital Platforms: Pearson, Edu-Business, and the (e)Valuation of Higher Education," *Critical Studies in Education*, Vol. 62, no. 1, pp. 50–66, 2021.

[4] S. Sellar and A. Hogan, *Pearson 2025: Transforming Teaching and Privatising Education Data*. Education International, 2019.

[5] I. Fauzi, M. Rachmawati, and A. Aziz, "Pelatihan Internet Marketing dalam Upaya meningkatkan Softskill Kewirausahaan pada Siswa SMK Bhakti Nusantara Salatiga," *Abdi Makarti*, Vol. 1, No. 1, pp. 34–48, 2022, Accessed: Sep. 02, 2024. [Online]. Available: https://jurnal.stieama.ac.id/index.php/abdimakarti/article/view/265/0

[6] M. A. Shareef, K. K. Kapoor, B. Mukerji, R. Dwivedi, and Y. K. Dwivedi, "Group Behavior in Social Media: Antecedents of Initial Trust Formation," *Comput Human Behav*, Vol. 105, No. 106225, 2020, Accessed: May 02, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0747563219304443

[7] J. R. Saura, "Using Data Sciences in Digital Marketing: Framework, Methods, and Performance Metrics," *Journal of Innovation and Knowledge*, Vol. 6, No. 2, pp. 92–102, Apr. 2021, Accessed: Feb. 12, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2444569X20300329

[8] A. T. Rosário and J. C. Dias, "How has Data-Driven Marketing Evolved: CHALLENGES and Opportunities with Emerging Technologies," *International Journal of Information Management Data Insights*, Vol. 3, No. 2, Nov. 2023.

[9] J. R. Saura, *Advanced Digital Marketing Strategies in a Data-Driven era*. IGI Global, 2021. doi: 10.4018/978-1-7998-8003-5.

[10] N. Akbar, "Perancangan SPK Tentang Keterampilan Mahasiswa dengan Metode SAW," *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, Vol. 8, No. 1, pp. 105–112, Jan. 2023, doi: 10.36341/rabit.v8i1.3033.

[11] H. Nurrifqi, F. Fikrillah, and D. Kurniadi, "Rekomendasi Pemilihan Program Studi menggunakan Algoritma Naïve Bayes," *Jurnal Algoritma*, Vol. 20, No. 1, pp. 42–49, 2023, [Online]. Available: https://jurnal.itg.ac.id/

[12] W. Villegas-Ch, X. Palacios-Pacheco, and S. Luján-Mora, "A Business Intelligence Framework for Analyzing Educational Data," *Sustainability (Switzerland)*, Vol. 12, No. 14, pp. 1–21, Jul. 2020, doi: 10.3390/su12145745.

[13] W. Vicente-Ramos and L. M. Cano-Torres, "The Effect of Digital Marketing on the Management of Relationships with University Students in Times of Covid-19," *International Journal of Data and Network Science*, Vol. 6, No. 1, pp. 59–66, Dec. 2022, doi: 10.5267/J.IJDNS.2021.10.004.

[14] N. S. Makrydakis, "The Role of Digital Marketing in Public Higher Education Organizations in Attracting Younger Generations," *Expert Journal of Marketing*, Vol. 9, No. 1, pp. 28–38, 2021.

[15] W. M. Lim, T. W. Jee, and E. C. De Run, "Strategic Brand Management for Higher educatIon Institutions with Graduate Degree Programs: Empirical Insights from the Higher Education Marketing Mix," *Journal of Strategic Marketing*, Vol. 28, No. 3, pp. 225–245, Apr. 2020, doi: 10.1080/0965254X.2018.1496131.

[16] C. Van der Putten, "Transforming Data Flow: Generative AI in ETL Pipeline Automatization," Politecnico di Torino, 2024.

 ISSN:2302-8149
e-ISSN:2540-9719

[17] H. Ashok, S. Ayyasamy, A. Ashok, and V. Arunachalam, "E-business Analytics through ETL and Self- Service Business Intelligence Tool," in *Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020)*, IEEE, 2020, pp. 747–751.

[18] S. Shankar Bhosale, Y. Kumar Sharma, S. S. Bhosale, Y. K. Sharma, F. Kurupkar, and S. Jagdishprasad Jhabarmal, "Role of Business Intelligence in Digital Marketing," *Int J Adv Innov Res*, Vol. 7, No. 1, pp. 113–116, Mar. 2020, [Online]. Available: https://www.researchgate.net/publication/339676330

[19] Q. Li, P. Duffy, and Z. Zhang, "A Novel Multi-Dimensional Analysis Approach to Teaching and Learning Analytics in Higher Education," *Systems*, Vol. 10, No. 4, Aug. 2022, Accessed: Dec. 19, 2024. [Online]. Available: https://www.mdpi.com/2079-8954/10/4/96

[20] A. A. Yulianto, "Study on Data Warehouse System for Supporting Decision Making in the Higher Education Institution (HEI) in Indonesia," Kanazawa University, 2020.

[21] A. A. Yulianto and Y. Kasahara, "Data Warehouse System for Multidimensional Analysis of Tuition Fee Level in Higher Education Institutions in Indonesia," *IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 6, pp. 541–550, 2020, [Online]. Available: www.ijacsa.thesai.org

[22] A. A. Yulianto, "Extract Transform Load Process in Distributed Database Academic Data Warehouse," *APTIKOM Journal on Computer Science and Information Technology*, Vol. 4, No. 2, pp. 61–68, 2019, Accessed: Dec. 19, 2024. [Online]. Available: http://aptikomjournal.com/index.php/CSIT/article/view/36

[23] D. Seenivasan, "Exploring Popular ETL Testing Techniques," *International Journal of Computer Trends and Technology*, Vol. 71, No. 02, pp. 32–39, Feb. 2023, doi: 10.14445/22312803/ijctt-v71i2p106.

[24] N.F. Oliveira, "ETL for data science? A case study," M.S. dissertation, ISCTE-Instituto Universitario de Lisboa, Portugal, 2021.

[25] N. Biswas, A. Sarkar, and K. C. Mondal, "Efficient Incremental Loading in ETL Processing for Real-Time Data Integration," *Innov Syst Softw Eng*, Vol. 16, No. 3, pp. 53–61, Mar. 20201.