

IMPLEMENTATION OF RULE BASED AND STRING MATCHING IN THE ETL PROCESS OF AGGREGATED POPULATION DATA MIGRATION

Nasrul Aditri Rahmandika ^{1*}, Ismail Abdurrozzaq Z¹, Ghulam Asrofi Buntoro¹

¹Faculty of Engineering, Muhammadiyah University of Ponorogo, Indonesia

*Corresponding author: nasrul@gmail.com

DOI:

ABSTRACT

Keywords:

ETL;
Rule-Based
System;
Population
Data;

Population data is one of the strategic information sources used in development planning, public services, and data-driven decision-making. The management of large volumes of population data often faces issues such as file structure heterogeneity, differences in naming formats, header variations, and inconsistencies in regional codes, which can potentially lead to errors during the data migration process. In this study, an analysis of the implementation of the Extract, Transform, Load (ETL) process on the Rule-Based ETL System for Aggregate Population Data (SERDADU) was conducted to manage an aggregate population dataset consisting of 123 Excel files with 33 groups of file name variants and 13 main data categories. The approach used combines rule-based methods and string matching to identify data categories, attribute mapping, data structure normalization, and matching variations in file names, sheets, headers, and regional identities. The implementation results show that the approach is capable of automating the process of transforming heterogeneous data into more structured and consistent data, making it ready to be loaded into the target database. Thus, SERDADU can enhance the efficiency of the data migration process, reduce mapping errors, and support the provision of more integrated demographic data for visualization and further analysis needs.

Article Info:

Submitted:

01/1/2026

Revised:

10/02/2026

Published:

27/02/2026



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International \(CC BY-SA 4.0\)](https://creativecommons.org/licenses/by-sa/4.0/)

How to cite: Author 1, A. (2020). Title of The Article. *Falah: Jurnal Ekonomi Syariah*, 5(1), 1-14.

INTRODUCTION

The research on the development of a population data visualization model in Medan Amplas District shows that interactive dashboards, thematic maps, and distribution graphs can present demographic information more efficiently and intuitively to support decision-making [1]. The application of web-based population data and information visualization also demonstrates that population data can be monitored thru more easily readable visual displays, making demographic development information easier for users to understand [2].

In this research object, the initial inventory of the AGREGAT_202501 dataset folder shows that there are 123 aggregate Excel files with 33 groups of file name variants, and it allows for processing data from the previous and next semesters with a file count reaching into the thousands. On the other hand, the active import configuration in SERDADU maps 13 main categories, namely education, occupation, religion, gender, marital status, head of family, mandatory ID card, family card ownership, birth certificate, marriage/divorce certificate, age group, and single age.

The number and variation of these categories pose real problems in the data migration process. Some files use the header structure IDEM, KODE, and WILAYAH, while other files only use KODE and WILAYAH. The format of the region codes is also inconsistent because some are written with dots and others without dots. These differences risk causing errors in data category identification, column mapping errors, regional inconsistencies, and data loading failures into the target database if not handled with a controlled transformation mechanism.

To address these conditions, SERDADU implements the Extract, Transform, Load process using a rule-based and string matching approach. The rule-based approach is used to select category mapping rules.

and target attributes, while string matching is used to recognize variations in file names, sheet names, column headers, aliases, and regional identities. Based on these conditions, this research focuses on the analysis of ETL implementation in SERDADU to see how heterogeneous aggregate data is directed into more structured data.

LITERATURE REVIEW

Underpinning theories (Sub-Section)

Aggregate population data is population data presented in summary form based on certain categories. This data differs from individual data because it does not display the identity of each resident, but rather presents the number or recapitulation based on specific groups. Aggregate data is usually grouped based on region, gender, age, education, occupation, religion, marital status, and possession of population documents [8].

In the context of SERDADU, aggregate population data becomes the main source transferred from Excel files to the target database. The data has the characteristics of a summary by category, by region, and by period. Therefore, the processing requires precise mapping so that each aggregate value is stored in the appropriate category, region, and year.

Errors in the processing of aggregate data can impact the accuracy of the resulting information. For example, a mismatch in regions can cause data from one village to be recorded under another village. Category errors can also lead to data being placed in the wrong target table. Therefore, aggregate population data needs to be processed with clear validation and transformation mechanisms.

METHODOLOGY

The waterfall method is a software development model that has sequential stages, starting from requirements analysis, system design, implementation, testing, to maintenance. Each stage in the waterfall produces an output that serves as the basis for the next stage. This model is suitable for use when system requirements are relatively clear from the beginning and the development process can be planned systematically [21]. The advantages of the waterfall method are that its flow is easy to understand, the documentation at each stage is clearer, and the development process can be controlled sequentially. Each stage has clear work boundaries, making it easier for researchers to explain the research process. However, the waterfall model also has limitations because it is less flexible if there are significant changes in requirements during the development process. In this research, the waterfall model is used as a framework to explain the stages of ETL implementation analysis on SERDADU. These stages include problem formulation, literature study, needs analysis, ETL process design, implementation design, and black box testing. With the waterfall framework, the research flow can be arranged more systematically so that each stage has a clear connection to the research objectives.

RESULT AND DISCUSSION

This research is a qualitative descriptive study with a technical case study approach on the SERDADU system. The research method is used to describe and analyze the process of importing available aggregate data, while the stages of system development and design are structured using the waterfall model so that the flow of needs analysis, design, implementation, and testing can be explained sequentially.

The initial stage includes problem formulation and literature review, followed by the system development section, which consists of needs analysis, system design, implementation or coding, and testing. After the testing phase, the research is directed toward the formulation of conclusions and recommendations based on the analysis of the ETL process. The use of the waterfall model in this research is not interpreted as the complete development of a new system, but rather as a framework for organizing the research stages of the SERDADU ETL process. With this framework, each stage has a clear output, starting from problem identification, data requirements, flow design, import process implementation, to the testing plan.

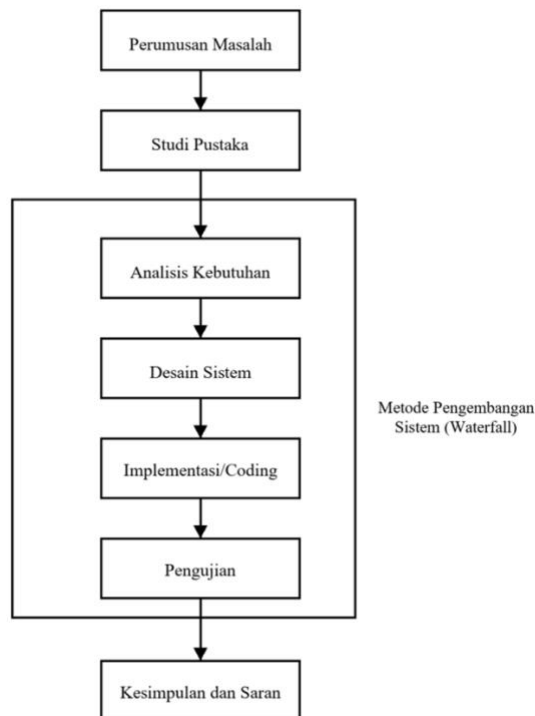


Figure 1. Research Stages with the Waterfall Model

The problem formulation stage was carried out by identifying the constraints in the process of migrating aggregate population data into the SERDADU system. The main issues lie in the variations of Excel file formats, differences in sheet names, variations in column headers, the use of term aliases, and the need for regional matching before the data can be loaded into the target database.

Based on these conditions, the research is directed to explain how the ETL process in SERDADU is carried out thru the admin import page, as well as how rule-based and string matching approaches are used to help the system recognize and transform source data into target data ready for presentation.

The output of this stage includes the formulation of the problem, the limitations of the problem, the research objectives, and the scope of the discussion. With this output, the research can be limited to the ETL process thru the admin import page, rather than the entire SERDADU features or system development from scratch.

Result/Data Analysis/Findings (Sub-Section)

The ETL implementation plan for SERDADU is divided into three main stages.

1. The first stage is extract, which is the process of receiving Excel files thru the admin import page, validating the parameters of year, semester, category, and file, then reading the workbook along with the sheets contained within it.
2. The second stage is transform. At this stage, the system applies a rule-based approach to determine the dataset category, target table, column mapping, unique keys, and calculation rules. The string matching approach is used to match variations of file names, sheet names,

- column headers, term aliases, and region names so that the source data can be directed to a standardized internal structure.
- The third stage is load. At this stage, the system forms the final payload and stores it in the target table using the upsert mechanism. If data with the same unique key is already available, the system updates that data. If it is not available, the system adds new data. All statuses of success, failure, and skipped rows are recorded in the import logs.

Testing is conducted using the black box testing method. This testing aims to ensure that the ETL import function operates as expected without directly discussing the program's code structure. The focus of the testing is directed at the system's behavior when receiving input, performing transformations, storing data, displaying process status, and handling invalid conditions.

Table 1. testing

No	Test Case	Hasil yang Diharapkan
1	Membuka halaman import admin.	Sistem menampilkan form import berisi pilihan tahun, semester, kategori, dan unggah file.
2	Mengirim form tanpa file atau parameter wajib.	Sistem menolak proses dan menampilkan pesan validasi.
3	Mengunggah file Excel dengan tahun, semester, dan kategori valid.	Sistem menyimpan file sementara dan meneruskan file ke proses extract.
4	File memiliki nama sheet atau nama file yang sesuai rule.	Sistem mengenali kategori dataset dan memilih rule import yang sesuai.
5	Header kolom memiliki variasi penulisan atau alias.	Sistem menormalisasi header dan memetakan kolom ke atribut target.
6	Baris data memuat nama atau kode wilayah.	Sistem mencocokkan wilayah dengan master kecamatan dan desa.
7	Data hasil transformasi valid.	Sistem melakukan upsert ke tabel target sesuai kunci unik.
8	Terdapat error, sheet dilewati, atau baris gagal.	Sistem mencatat status, jumlah baris sukses, jumlah baris gagal, dan pesan error pada import logs.

CONCLUSION

Based on the analysis of the implementation of the Extract, Transform, Load (ETL) process in the Rule-Based ETL System for Population Aggregate Data (SERDADU), it can be concluded that the management of population data faces challenges such as the heterogeneity of file structures, variations in file and sheet naming, differences in header formats, and inconsistencies in regional codes. These conditions have the potential to cause data identification errors, attribute mapping issues, and migration process failures if not handled systematically. The application of a rule-based and string matching approach in SERDADU can provide a more structured data transformation mechanism by identifying data categories, matching attributes, normalizing regional formats, and automatically adjusting various dataset structure variations. The implementation of ETL shows that the initially diverse and inconsistent aggregate population data can be processed into more consistent data ready to be loaded into the target database. Thus, SERDADU can improve the efficiency of the population data migration process, reduce the risk of data processing errors, and support the availability of more integrated data as a basis for information presentation, data visualization, and data-driven decision-making based on population data.

REFERENCES

1. D. A. Syafiq and Y. Sembiring, "Pengembangan Model Visualisasi Data Kependudukan Berbasis Teknologi Informasi untuk Mendukung Pengambilan Keputusan di Kecamatan Medan Amplas," *Jurnal Penelitian Inovatif*, vol. 5, no. 2, pp. 1329-1338, May 2025, doi: 10.54082/jupin.1500.
2. Y. Akbar, D. Mayangsari, and S. Bebriani, "Penerapan Visualisasi Data dan Informasi Kependudukan Berbasis Web," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 786-793, Jul. 2024, doi: 10.57152/malcom.v4i3.1290.
3. L. R. Putri, "Inovasi Pelayanan Administrasi Kependudukan dengan Online (PAKDE ON) melalui Aplikasi WhatsApp dan Inovasi Pelayanan Dokumen Administrasi Kependudukan dan Pengadilan Agama (DIDUPAG) di Dinas Kependudukan dan Pencatatan Sipil Kabupaten Madiun," *Jurnal Registratie*, vol. 5, no. 1, pp. 53-68, Sep. 2023, doi: 10.33701/jurnalregistratie.v5i1.3229.
4. S. Rohajawati, "Effectiveness of Digital Transformation in Data Warehouse Technology," *The Indonesian Journal of Computer Science*, vol. 13, no. 1, Feb. 2024, doi: 10.33022/ijcs.v13i1.3759.
5. R. Tarigan, E. Nuryani, S. K. Sianturi, N. Hidayanti, and Widyawati, "Migrasi Data Hasil Optimalisasi Pemuatan Kontainer Pada Legacy System ke SAP untuk Mempercepat Pembuatan Surat Jalan dengan Batch Data Communication (BDC)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 6, pp. 1213-1222, Dec. 2024, doi: 10.25126/jtiik.2024117365.
6. K. Saputra, K. A. Laksitowening, and A. Herdiani, "Performance Analysis of Extract, Transform, and Load Methods for Business Intelligence in E-Learning Systems Using Pentaho Data Integration," *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 1, pp. 75-86, Feb. 2025, doi: 10.52436/1.jutif.2025.6.1.4173.
7. Y. G. Berliana and A. D. Indriyanti, "Sistem Deteksi Kemiripan Validasi Surat Permohonan Keabsahan Akta Dengan Metode Fuzzy String Matching," *JSI (Jurnal Sistem Informasi) Universitas Suryadarma*, vol. 12, no. 2, pp. 100-108, Jul. 2025, doi: 10.35968/jsi.v12i2.1507.
8. Pemerintah Republik Indonesia, "Peraturan Pemerintah Republik Indonesia Nomor 40 Tahun 2019 tentang Pelaksanaan Undang-Undang Nomor 23 Tahun 2006 tentang Administrasi Kependudukan sebagaimana telah diubah dengan Undang-Undang Nomor 24 Tahun 2013," 2019.

9. L. Dinesh and K. G. Devi, "An Efficient Hybrid Optimization of ETL Process in Data Warehouse of Cloud Architecture," *Journal of Cloud Computing*, vol. 13, art. no. 12, 2024, doi: 10.1186/s13677-023-00571-y.
10. B. Khan, S. Jan, W. Khan, and M. I. Chughtai, "An Overview of ETL Techniques, Tools, Processes and Evaluations in Data Warehousing," *Journal on Big Data*, vol. 6, no. 1, pp. 1-20, 2024, doi: 10.32604/jbd.2023.046223.
11. T. A. Pratami, Tursina, and R. Septiriana, "Rekomendasi Pemilihan Model Sepeda Menggunakan Rule Based System," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 2, pp. 1350-1361, Apr. 2024, doi: 10.23960/jitet.v12i2.4239.
12. M. Alfian, W. Aprianti, H. Rhomadhona, and J. Permadi, "Sistem Pakar Penyakit Sapi Menggunakan Rule Based Reasoning dengan Forward Chaining," *Teknomatika: Jurnal Informatika dan Komputer*, vol. 17, no. 2, pp. 22-29, Dec. 2024, doi: 10.30989/teknomatika.v17i2.1541.
13. J. Song, C. Nanayakkara, and P. Christen, "Privately Evaluating Sensitive Population Record Linkage Without Ground Truth Data," *International Journal of Data Science and Analytics*, vol. 20, pp. 2971-2986, 2025, doi: 10.1007/s41060-024-00657-z.
14. F. A. Valensyah and O. Irnawati, "Sistem Informasi Berbasis Website Menggunakan Framework Laravel," *INSANtek*, vol. 5, no. 1, pp. 7-14, May 2024, doi: 10.31294/insantek.v5i1.3408.
15. B. D. Tiert, A. Dharmawan, and J. Gondohanindijo, "Perancangan Sistem Informasi Berita Otomotif Berbasis Website Dengan PHP Dan MySQL," *INTECOMS: Journal of Information Technology and Computer Science*, vol. 7, no. 4, pp. 1250-1256, Jul. 2024, doi: 10.31539/intecom.v7i4.10955.
16. R. A. Wilantara, N. Azizah, and R. S. Razaqi, "Rancang Bangun Sistem Informasi Administrasi Menggunakan Metode Scrum Berbasis Framework Laravel (Studi Kasus: SDN 3 Trebungan)," *Jurnal Komputer Teknologi Informasi Sistem Komputer (JUKTISI)*, vol. 4, no. 3, pp. 1718-1726, Feb. 2026, doi: 10.62712/juktisi.v5i1.506.
17. A. W. Pramudita and R. Somya, "Sistem Filtering Data Mahasiswa Menggunakan Framework Laravel dan Library Laravel Excel," *Jurnal Ilmiah Informatika*, vol. 9, no. 1, pp. 38-42, Jun. 2021, doi: 10.33884/jif.v9i01.3716.
18. I. Mardiana, Wahyudin, and E. Junaeti, "Pengembangan Learning Management System dengan Framework Laravel dan Tailwind CSS," *Multinetics*, vol. 10, no. 1, pp. 40-49, Jul. 2024, doi: 10.32722/multinetics.v10i1.6678.
19. F. Haykhal, "Rancang Bangun Sistem Informasi Taku untuk Pengajuan Judul Tugas Akhir Berbasis Web di Lingkup Jurusan Teknik Informatika," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3S1, pp. 1129-1137, Oct. 2025, doi: 10.23960/jitet.v13i3S1.8016.
20. M. S. A. Amrullah, Aminudin, and G. I. Marthasari, "Otomatisasi Proses Deployment dengan Metode CI/CD Menggunakan Jenkins dan Docker Pada Web Service i-Lab," *Jurnal Repositor*, vol. 6, no. 4, pp. 313-322, Nov. 2024.
21. N. A. Al Azfar and S. D. Anggita, "Penerapan Metode Waterfall Pada Sistem Informasi E-Rapor," *Information System Journal*, vol. 7, no. 01, pp. 45-55, Jun. 2024, doi: 10.24076/infosjournal.2024v7i01.1582.
22. S. Mayang, A. Amirudin, and S. Lestari, "Perancangan Sistem Informasi Retensi dan Penyusutan Berkas Rekam Medis di Rumah Sakit," *Jurnal Sains dan Kesehatan*, vol. 3, no. 1, pp. 35-43, May 2024, doi: 10.57151/jsika.v3i1.370.
23. M. Z. Abdillah and I. G. S. Pranata, "Analisis dan Perancangan Sistem Informasi Manajemen Gereja Menggunakan UML (Unified Modelling Language)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4831.
24. V. Tesaloniki, P. A. Jusia, and Mery, "Perancangan Sistem Informasi Pengolahan Data Transaksi Penjualan dan Persediaan Barang pada Sumber Abadi Motor," *Jurnal Manajemen Teknologi*

- Dan Sistem Informasi (JMS), vol. 4, no. 1, pp. 680-690, Apr. 2024, doi: 10.33998/jms.2024.4.1.1718.
25. C. E. E. S. Putri and A. Susanto, "Feasibility Analysis of Bengkel Koding Website Using Black Box Testing and Boundary Value Analysis," Sinkron: Jurnal dan Penelitian Teknik Informatika, vol. 8, no. 2, pp. 764-776, Mar. 2024, doi: 10.33395/sinkron.v8i2.13589.