ASKLEPION
INFORMAÇÃO EM SAÚDE

ARTIGO

DOI: https://doi.org/10.21728/ asklepion.2026v5n1e-132

Data de submissão: 11/03/2026 Data de aprovação: 20/04/2026Data de publicação: 22/04/202

SCALABLE ETL PIPELINE FOR HEALTH DATA INGESTION

application to the Brazilian Unified Health System (SUS)

Andre Massahiro Shimaoka[1]
Universidade Federal de São Paulo
andre.shimaoka@unifesp.br

Maria Elisabete Salvador[2]
Universidade Federal de São Paulo
elisabete.salvador@unifesp.br

José Marcio Duarte[3]
Universidade Federal de São Paulo
jm.duarte@unifesp.br

Antonio Carlos da Silva Junior[4]
Universidade Federal de São Paulo
acsjunior@unifesp.br

Luciano Rodrigo Lopes[5]
Universidade Federal de São Paulo
luciano.lopes@unifesp.br

Paulo Bandiera-Paiva[6]
Universidade Federal de São Paulo
paiva@unifesp.br

Abstract

The growing availability of health data within the Brazilian Unified Health System (SUS) increases the potential for data-driven analysis. However, large datasets also introduce challenges related to volume, structure, and data integration. This study develops and evaluates an automated Extract, Transform, Load (ETL) pipeline for ingestion and preparation of data from the Ambulatory Information System (SIA-SUS). The architecture uses cloud computing infrastructure to support scalable processing. The study follows the Design Science Research approach, which focuses on the development and evaluation of technological artifacts. A pilot experiment processes data from January 2024 for three Brazilian states. The experiment includes approximately 3.2 million ambulatory records. Each execution runs five times to estimate operational variability. Results show stable pipeline performance across scenarios. The extraction stage accounts for the largest share of total execution time. Throughput remains relatively consistent despite differences in data volume. Linear regression between processed records and execution time produces a coefficient of determination of $R^2 = 0.996$. The result indicates an approximately linear relationship between data volume and processing time. The pipeline demonstrates operational feasibility and scalability potential. The architecture reduces the complexity of preparing large datasets from the SUS. The solution supports the development of analytical environments for public health data.

Keywords: ETL; health information systems; data ingestion; health data integration; public health.

[1] Pesquisador no Departamento de Informática em Saúde, Escola Paulista de Medicina, UNIFESP
[2] Docente do Departamento de Informática em Saúde, Escola Paulista de Medicina, UNIFESP
[3] Pesquisador no Departamento de Informática em Saúde, Escola Paulista de Medicina, UNIFESP
[4] Pesquisador no Departamento de Informática em Saúde, Escola Paulista de Medicina, UNIFESP
[5] Docente do Departamento de Informática em Saúde, Escola Paulista de Medicina, UNIFESP
[6] Docente do Departamento de Informática em Saúde, Escola Paulista de Medicina, UNIFESP

PIPELINE ETL ESCALÁVEL PARA INGESTÃO DE DADOS EM SAÚDE

aplicação ao Sistema Único de Saúde (SUS)

Resumo

A crescente disponibilidade de bases em saúde no Sistema Único de Saúde (SUS) amplia o potencial para análises baseadas em dados, mas também impõe desafios relacionados ao volume, à estrutura e à integração das informações. Nesse contexto, este estudo teve como objetivo desenvolver e avaliar um pipeline automatizado de Extração, Transformação e Carga (ETL) para ingestão e preparação da base de produção ambulatorial do Sistema de Informações Ambulatoriais do SUS (SIA-SUS), utilizando arquitetura de computação em nuvem. A pesquisa adotou a abordagem de Design Science Research, voltada à construção e avaliação de artefatos tecnológicos. Um experimento piloto foi conduzido com dados de janeiro de 2024 para três Unidades Federativas (Santa Catarina, Espírito Santo e Rio Grande do Norte), totalizando aproximadamente 3,2 milhões de registros ambulatoriais processados. Cada execução foi repetida cinco vezes para estimar a variabilidade operacional. Os resultados indicaram estabilidade do pipeline e predominância da etapa de extração no tempo total de processamento. O throughput manteve-se relativamente constante entre os cenários analisados, e a regressão linear entre volume de registros e tempo de execução apresentou coeficiente de determinação $R^2 = 0.996$, indicando comportamento aproximadamente linear do sistema. Conclui-se que o pipeline proposto apresenta viabilidade operacional e potencial de escalabilidade, contribuindo para automatizar a preparação de grandes bases do SUS e apoiar o desenvolvimento de ambientes analíticos em saúde pública.

Palavras-chave: ETL; sistemas de informação em saúde; ingestão de dados; integração de dados em saúde; saúde pública.

PIPELINE ETL ESCALABLE PARA LA INGESTIÓN DE DATOS DE SALUD

aplicación al Sistema Único de Salud (SUS)

Resumen

La creciente disponibilidad de bases de datos en salud en el Sistema Único de Salud (SUS) amplía el potencial para análisis basados en datos, pero también introduce desafíos relacionados con el volumen, la estructura y la integración de la información. En este contexto, este estudio tuvo como objetivo desarrollar y evaluar un pipeline automatizado de Extracción, Transformación y Carga (ETL) para la ingestión y preparación de la base de producción ambulatoria del Sistema de Información Ambulatoria del SUS (SIA-SUS), utilizando una arquitectura de computación en la nube. La investigación adoptó el enfoque de Design Science Research, orientado a la construcción y evaluación de artefactos tecnológicos. Se realizó un experimento piloto con datos de enero de 2024 para tres Unidades Federativas (Santa Catarina, Espírito Santo y Rio Grande do Norte), con un total aproximado de 3,2 millones de registros ambulatorios procesados. Cada ejecución se repitió cinco veces para estimar la variabilidad operativa. Los resultados indicaron estabilidad del pipeline y predominio de la etapa de extracción en el tiempo total de procesamiento. El throughput se mantuvo relativamente constante entre los escenarios analizados, y la regresión lineal entre el volumen de registros y el tiempo de ejecución presentó un coeficiente de determinación $R^2 = 0.996$, lo que indica un comportamiento aproximadamente lineal del sistema. Se concluye que el pipeline propuesto presenta viabilidad operativa y potencial de escalabilidad, contribuyendo a automatizar la preparación de grandes bases del SUS y a apoyar el desarrollo de entornos analíticos en salud pública.

Palabras clave: ETL; sistemas de información en salud; ingestión de datos; integración de datos en salud; salud pública.

1 INTRODUCTION

The integration of health information plays a central role in public health management, especially in large universal health systems such as the Brazilian Unified Health System (SUS). Business Intelligence strategies enable the transformation of large data volumes into structured information for decision making (Torres et al., 2021). In this context, management reports based on reliable clinical and financial data support planning, performance monitoring, policy evaluation, and resource allocation across different levels of the health system (Antunes *et al*., 2021).

Brazilian public health operates in a context characterized by strong territorial diversity, demographic heterogeneity, and socioeconomic inequality. These factors increase the complexity of health system management. The SUS provides universal and comprehensive access to health care and receives international recognition for this achievement. However, important challenges remain in the integration and availability of information produced across different levels of care. Data fragmentation, the coexistence of multiple information systems, and the absence of structured integration processes limit the effective use of large administrative datasets for health system management (Paim *et al*., 2011).

Among national health information systems, the Ambulatory Information System of the Brazilian Unified Health System (SIA-SUS) plays a central role in recording outpatient care across the country (Shimaoka *et al*., 2025). The system uses the Individualized Ambulatory Production Bulletin (BPA-I) to store detailed information on consultations, procedures, diagnostic tests, associated diagnoses, patient demographic characteristics, and reported and approved financial values (Brasil, 2022). These data support the construction of clinical and financial indicators and provide an important data source for Business Intelligence applications in public health.

Despite its strategic relevance, analytical use of BPA-I data faces important operational challenges. The system produces large volumes of records every month. The data appear in multiple files and require extensive variable standardization and enrichment. Local computational infrastructure often lacks the capacity required for large scale processing. These factors limit the systematic preparation of the data for management and analytical purposes (Shimaoka *et al*., 2025).

Some initiatives already support automated processing of other national health databases, such as the Mortality Information System (SIM) and the Hospital Information System (SIH-SUS), through the Health Data Science Platform (PCDaS). However, few

documented and reproducible pipelines exist for SIA-SUS data. This limitation is particularly evident for BPA-I individual records, which require more complex processing (Fiocruz, 2019).

In Business Intelligence and data engineering, Extract, Transform, Load (ETL) processes play a central role in integrating heterogeneous data sources into analytical environments (Zarate *et al*., 2024). The ETL process extracts data from primary sources, transforms the data through standardization, cleaning, and variable enrichment, and loads the results into repositories for query and analysis (Souibgui *et al*., 2019).

Cloud computing provides an effective infrastructure for scalable data processing, structured storage, and integration with relational databases in distributed environments (Berisha; Mëziu; Shabani, 2022). Cloud architectures offer computational elasticity, horizontal scalability, and process reproducibility. These characteristics support the processing of large health datasets (Silva; Bonacelli; Pacheco, 2020). Understanding how computational effort distributes across ETL pipeline stages is essential for designing scalable and sustainable data architectures (Jyoti Aggarwal, 2025).

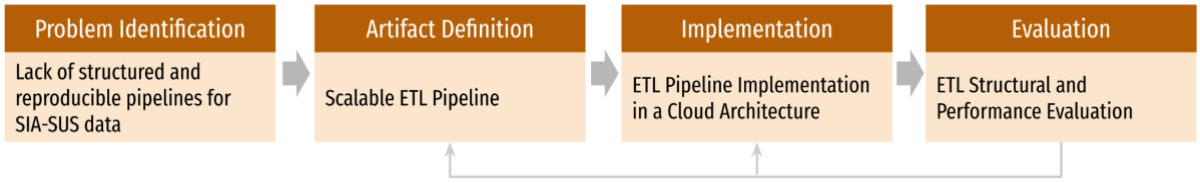
Automated and reproducible data pipelines play a central role in enabling data driven analytical environments in health systems. These pipelines reduce the operational complexity of data access and preparation. They also allow empirical evaluation of processing performance and operational bottlenecks when large volumes of records are processed (Krishnapur *et al*., 2026).

This study develops and evaluates an automated and scalable ETL pipeline for ingestion and processing of ambulatory data from the SIA-SUS system in a cloud computing architecture. It also examines the operational behavior of the pipeline and analyzes how computational resources distribute across the Extract, Transform, and Load stages.

2 METHOD

This research follows the Design Science Research (DSR) approach, which focuses on the development and evaluation of technological artifacts that address relevant information systems problems[13]. The study design follows the core stages of DSR: (i) problem identification, (ii) artifact definition, (iii) artifact implementation, and (iv) artifact evaluation, as illustrated in Figure 1.

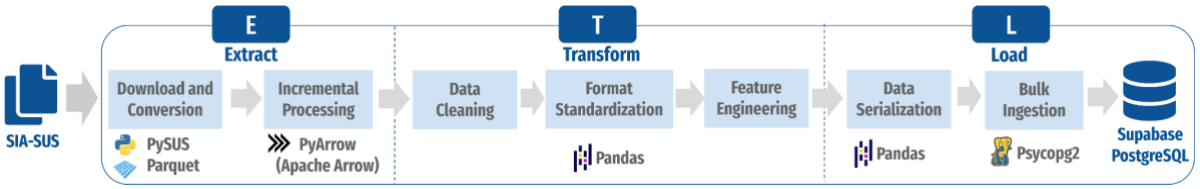
Figure 1 - Methodological model based on Design Science Research (DSR)



Source: Adapted from Peffers *et al*. (2007).

The SIA-SUS database presents operational complexity due to the large volume of records distributed across multiple files and the need for prior standardization for analytical use. The proposed artifact is an automated ETL pipeline designed to operate in a cloud computing environment. The architecture prioritizes scalability, reproducibility, and detailed performance monitoring for each pipeline stage (Figure 2).

Figure 2 - ETL pipeline architecture in a cloud environment.



Source: Prepared by the authors.

A pilot experiment evaluated the initial performance of the proposed artifact. The experiment processed data from three Brazilian states located in different regions: Santa Catarina (SC), Espírito Santo (ES), and Rio Grande do Norte (RN), which represent the South, Southeast, and Northeast regions, respectively. This selection provides regional, demographic, and healthcare diversity. The experiment used data from January 2024 to control dataset size and evaluate the operational behavior of the proposed architecture.

The pipeline extracts data using the PySUS library, which provides programmatic access to public datasets available from DATASUS. PySUS automatically identifies and downloads the files in Parquet format, which enables direct integration with Python analytical environments (Coelho, 2024). The pipeline reads the data using the PyArrow library and processes the records incrementally in batches (batch_size = 100,000). This approach avoids full dataset materialization in system memory (Apache, 2026).

The transformation stage processes each batch using the Pandas library. The pipeline standardizes column names, cleans textual fields, converts numeric variables, and normalizes

date formats. It also standardizes the sex variable, derives the Brazilian state from the municipality IBGE code, and converts age variables into numerical values.

The load stage writes the processed data to a PostgreSQL database hosted in a cloud environment (Supabase) through a secure SSL connection (Supabase Inc, 2026). The pipeline uses the native PostgreSQL bulk loading mechanism (COPY FROM STDIN) through the psycopg2 library (Di Gregorio; Varrazzo, 2025). For each processed batch of up to 100,000 records, the pipeline converts the transformed data into an in memory CSV stream and sends it directly to the database through a bulk loading operation. This strategy reduces transactional fragmentation, minimizes client server communication overhead, and improves the efficiency of large scale data ingestion in remote environments (Henke *et al*., 2023).

The artifact evaluation considers three complementary dimensions: (i) stage level execution time, (ii) behavior under different processing volumes to assess empirical scalability, and (iii) architectural analysis of the pipeline (Reddy Gujjala, 2024). For the temporal evaluation, the analysis records the download time, file reading time, transformation time, database loading time, and total pipeline execution time. As a complementary metric, the analysis also calculates pipeline throughput. Throughput is defined as the ratio between the number of processed records and the execution time.

To evaluate the empirical scalability behavior of the pipeline, the analysis fits a simple linear regression model between the number of processed records and the total execution time (Barnes *et al*., 2008). The analysis estimates the model using the Ordinary Least Squares (OLS) method through the implementation available in the scikit learn Python library (Pedregosa *et al.*, 2012). The analysis represents data volume as the number of processed records expressed in millions, while the dependent variable corresponds to the total pipeline execution time in seconds. The regression serves a descriptive purpose and examines how processing time increases with data volume in order to characterize the scalability behavior of the proposed architecture. The coefficient of determination (R2) indicates the goodness of fit of the linear model to the observed data (Nakagawa; Schielzeth, 2013).

Cloud computing environments introduce inherent variability in execution time. To account for this effect, the study runs each pipeline execution five times on different days and at different times for each analyzed Brazilian state. For each temporal metric (Extraction, Transformation, Loading, and Total Time), the analysis calculates the mean, standard deviation, and the 95% confidence interval. These measures evaluate the stability of the artifact behavior and reduce the influence of occasional variations in network latency or server workload.

The architectural analysis decomposes the pipeline into functional layers (acquisition, processing, and persistence) and examines how computational effort distributes across these layers (Khattach; Moussaoui; Hassine, 2025). This approach identifies the stage that dominates operational cost, characterizes system behavior under different data volumes, and evaluates the suitability of the architecture for cloud environments. The measurements do not aim to provide absolute performance benchmarks. Instead, they analyze the operational structure of the artifact and identify potential bottlenecks.

The architecture supports horizontal scalability, expansion to multiple historical periods, and replication in environments compatible with Python and PostgreSQL. The modular structure by Brazilian state and time period enables incremental reprocessing and future integration with Business Intelligence tools.

3 RESULTS AND DISCUSSION

The pilot experiment analyzes three Brazilian states with different volumes of ambulatory production for January 2024, representing large, medium, and small scale scenarios. Santa Catarina presents the largest processed volume (2, 195, 219 records), followed by Espírito Santo (662,622 records) and Rio Grande do Norte (361,330 records). Table 1 reports the average execution time for each ETL pipeline stage based on five independent runs conducted at different times. The table also includes the mean, standard deviation, and the 95% confidence interval.

Table 1 - Execution time of ETL pipeline stages

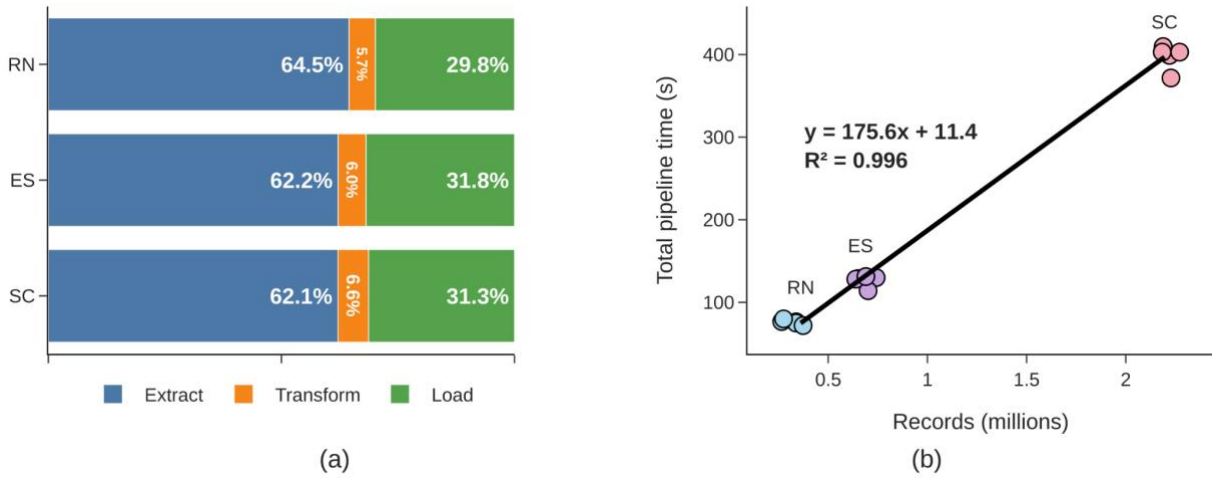
State	E (s)	T (s)	L (s)	Total (s)	95% CI
SC	245.65 ± 11.40	26.01 ± 1.26	123.70 ± 2.22	397.20 ± 14.82	378.8 - 415.6
ES	77.86 ± 5.13	7.48 ± 0.23	39.86 ± 3.35	126.42 ± 6.99	117.7 - 135.1
RN	48.06 ± 2.88	4.23 ± 0.21	22.22 ± 0.60	76.01 ± 2.91	72.4 - 79.6

Source: Prepared by the authors.

Repeated executions allow the estimation of pipeline operational variability and the calculation of confidence intervals for execution times. These measures strengthen the methodological robustness of the experimental evaluation. The observed standard deviations correspond to less than 5.6% of the mean execution time across all scenarios, indicating low variability between runs even under potential fluctuations in network latency and computational load in the cloud environment.

The extraction stage, which transfers files from the remote repository, accounts for the largest share of total execution time (approximately 62-64%). The database loading stage follows with 29-32%, while the transformation stage represents a small portion of total runtime (5-7%). These results indicate that data transfer, rather than data transformation, represents the main performance bottleneck of the pipeline (Figure 3a).

Figure 3 - ETL pipeline performance analysis: (a) stage time distribution; (b) Linear regression between number of records and total pipeline time.



Source: Prepared by the authors.

In data engineering, network latency and data transfer are major determinants of performance in large-scale ingestion pipelines, particularly when datasets are retrieved from remote repositories such as health information systems (Namli *et al.*, 2024). The pattern observed in this experiment reinforces this behavior, indicating that the main operational cost of the pipeline is associated with data transfer from the remote repository.

The mean throughput observed was approximately 5,527 rows per second for Santa Catarina (95% CI: 5.33–5.72 thousand), 5,242 for Espírito Santo (95% CI: 4.98–5.48 thousand), and 4,754 for Rio Grande do Norte (95% CI: 4.55–4.94 thousand). Despite differences in data volume across the analyzed states, throughput values remained within the same order of magnitude, suggesting stable pipeline performance and linear scalability. A Kruskal–Wallis test indicated a statistically significant difference in throughput across states ($H = 12.5$, $p = 0.0019$).

The relationship between the number of processed records and the total pipeline execution time was evaluated using simple linear regression. The fitted model yielded a coefficient of determination of $R2 = 0.996$, indicating a strong linear relationship between dataset size and execution time. As illustrated in Figure 3b, processing time increased proportionally with the number of records processed.

Although the experiment included only three Brazilian states and a single time period, the strong linear relationship observed between the number of processed records and execution time suggests that this pattern may extend to larger datasets. In this context, the regression model may serve as an approximation for estimating processing time in scenarios involving additional states or longer historical series.

Another relevant finding was the impact of the database persistence strategy. A complementary experiment using multiple INSERT statements resulted in an execution time approximately ten times longer than that achieved with bulk loading. This result highlights the efficiency of batch loading mechanisms for large-scale data ingestion, as they reduce transactional overhead, minimize client–server interactions, and improve database write performance (Martins *et al*., 2019; Shaik, 2020). In cloud environments, this effect is further amplified, as network latency and communication overhead increase the benefits of batch operations, making bulk loading particularly advantageous for large-scale data ingestion (Noll *et al*., 2020).

The implemented architecture separates data acquisition, processing, and persistence into distinct functional layers. The extraction and transformation stages run in a cloud computing environment (Google Colab), while the storage layer uses a cloud-hosted database (PostgreSQL/Supabase). This modular design allows each pipeline component to evolve independently and improves scalability and system maintainability (Wojciechowski, 2011; Yu, 2025). It also supports replication across different Brazilian states and time periods without structural changes to the ingestion process. The approach can be applied to other SUS databases with similar volume and structure, such as the Mortality Information System (SIM) and the Hospital Information System (SIH-SUS).

Pipeline instrumentation with stage-level timing metrics identified the main operational bottlenecks. The extraction stage showed the highest optimization potential. File transfer depends on the download mechanism implemented in the PySUS library, which currently performs sequential downloads. In scenarios with many files, this behavior increases extraction time. Parallel downloads, temporary caching, or geographic data replication may reduce total processing time (Liu, 2014).

Automated ETL pipelines reduce the operational complexity of accessing and preparing large-scale health datasets. In the SIA-SUS system, data are distributed across multiple files and often require prior standardization and cleaning before analytical use. Automating these steps improves the reproducibility of data extraction and preparation. This characteristic is particularly relevant for epidemiological studies, public policy evaluation, and health

monitoring systems, where periodic data updates and consistent processing workflows are essential for reliable analyses (Krishnapur *et al*., 2026).

This study has several limitations. The evaluation covered only one time period and a single cloud execution environment. Infrastructure metrics such as CPU usage, memory consumption, and cloud computing costs were not analyzed.

3 CONCLUSION

This study presented the development and evaluation of an automated ETL pipeline for the ingestion and preparation of outpatient production data from the Brazilian Outpatient Information System (SIA-SUS), implemented in a cloud computing architecture. The Design Science Research approach supported the development of an artifact aimed at automating and standardizing the extraction, transformation, and loading of administrative health data.

The experimental evaluation demonstrated the operational feasibility of the proposed architecture. Results indicated stable pipeline performance and linear scalability with increasing data volume. These findings suggest that the solution has potential scalability for larger scenarios, such as processing multiple states or longer historical series.

From an applied perspective, the proposed artifact reduces the operational complexity associated with preparing large administrative datasets from the Brazilian Unified Health System (SUS). The automated workflow improves reproducibility and facilitates data integration in analytical environments. The modular pipeline design also supports replication across different states, time periods, and other national health information systems, such as the Mortality Information System (SIM) and the Hospital Information System (SIH-SUS).

Future work may explore parallelization and distributed processing strategies, as well as evaluate infrastructure metrics such as computational resource consumption and cloud processing costs. Expanding the analysis to longer time periods and a larger number of states may also help assess pipeline behavior in higher-scale data scenarios.

REFERÊNCIAS

ANTUNES, F. M. *et al.* Informação como apoio para tomada de decisão de gestores públicos de saúde. **Revista de Administração em Saúde**, [*s. l.*], v. 21, n. 82, 2021. Disponível em: https://cqh.org.br/ojs-2.4.8/index.php/ras/article/view/283. Acesso em: 18 fev. 2026.

APACHE. **PyArrow - Python library for Apache Arrow**. versão 23.0.0. [*S. l.*]: ©2016-2026 The Apache Software Foundation. Disponível em: https://pypi.org/project/pyarrow/. Acesso em: 23 fev. 2026.

BARNES, B. J. *et al.* A regression-based approach to scalability prediction. *In*: ICS08: INTERNATIONAL CONFERENCE ON SUPERCOMPUTING, 2008, Island of Kos Greece. **Proceedings of the 22nd annual international conference on Supercomputing**. Island of Kos Greece: ACM, 2008. p. 368–377. Disponível em: https://dl.acm.org/doi/10.1145/1375527.1375580. Acesso em: 9 mar. 2026.

BERISHA, B.; MËZIU, E.; SHABANI, I. Big data analytics in Cloud computing: an overview. **Journal of Cloud Computing**, [*s. l.*], v. 11, n. 1, p. 24, 2022.

BRASIL. Ministério da Saúde. Sistema de Informações Ambulatoriais. **Manual Operacional do Boletim de Produção Ambulatorial**. Brasília, DF: Ministério da Saúde, 2022. Disponível em: https://wiki.saude.gov.br/sia/index.php/BPA. Acesso em: 26 fev. 2026.

COELHO, F. C. **PySUS**. versão 1.0.1. 2024. [*S. l.*]: ©2026 Python Software Foundation. Disponível em: https://pypi.org/project/pysus/1.0.1/. Acesso em: 23 fev. 2026.

DI GREGORIO, F.; VARRAZZO, D. **Psycopg2 - Python-PostgreSQL Database Adapter**. versão 2.9.11. 2025. [*S. l.*]: ©2026 Python Software Foundation. Disponível em: https://pypi.org/project/psycopg2. Acesso em: 23 jun. 2026.

FIOCRUZ. **Plataforma de Ciência de Dados aplicadas à Saúde**. Rio de Janeiro: FIOCRUZ, 2019. Disponível em: https://pcdas.icict.fiocruz.br/. Acesso em: 18 fev. 2026.

HENKE, E. *et al.* An Extract-Transform-Load Process Design for the Incremental Loading of German Real-World Data Based on FHIR and OMOP CDM: Algorithm Development and Validation. **JMIR Medical Informatics**, [*s. l.*], v. 11, p. 1–10, 2023.

JYOTI AGGARWAL. ETL pipelines for cloud-native data platforms: Architecting real-time analytics on integrated cloud services. **World Journal of Advanced Engineering Technology and Sciences**, [*s. l.*], v. 15, n. 2, p. 107–114, 2025.

KHATTACH, O.; MOUSSAOUI, O.; HASSINE, M. End-to-End Architecture for Real-Time IoT Analytics and Predictive Maintenance Using Stream Processing and ML Pipelines. **Sensors**, [*s. l.*], v. 25, n. 9, p. 2945, 2025.

KRISHNAPUR, P. K. *et al.* A Reproducible Python-Based Computational Pipeline for Real-Time Ingestion, Advanced Analysis, and Dynamic Reporting of Public Health Data: A Systems Validation Study. **Cureus**, [*s. l.*], 2026. Disponível em: https://www.cureus.com/articles/449538-a-reproducible-python-based-computational-pipeline-for-real-time-ingestion-advanced-analysis-and-dynamic-reporting-of-public-health-data-a-systems-validation-study. Acesso em: 10 mar. 2026.

LIU, X. **Optimizing ETL Dataflow Using Shared Caching and Parallelization Methods**. [*S. l.*]: arXiv, 2014. Disponível em: https://arxiv.org/abs/1409.1639. Acesso em: 10 mar. 2026.

MARTINS, P. *et al.* A performance study on different data load methods in relational databases. *In*: 2019 14TH IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES (CISTI), 2019, Coimbra, Portugal. **2019 14th Iberian Conference on Information Systems and Technologies (CISTI)**. Coimbra, Portugal: IEEE, 2019. p. 1–7. Disponível em: https://ieeexplore.ieee.org/document/8760615/. Acesso em: 9 mar. 2026.

NAKAGAWA, S.; SCHIELZETH, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. **Methods in Ecology and Evolution**, [*s. l.*], v. 4, n. 2, p. 133–142, 2013.

NAMLI, T. *et al.* A scalable and transparent data pipeline for AI-enabled health data ecosystems. **Frontiers in Medicine**, [*s. l.*], v. 11, p. 1393123, 2024.

NOLL, S. *et al.* Shared Load (ing): Efficient Bulk Loading into Optimized Storage. 2020. **CIDR**. [*S. l.*]: [*s. d.*], 2020.

PAIM, J. *et al.* The Brazilian health system: history, advances, and challenges. **The Lancet**, [*s. l.*], v. 377, n. 9779, p. 1778–1797, 2011.

PEDREGOSA, F. *et al.* **Scikit-learn:** Machine Learning in Python. [*s. l.*], 2012. Disponível em: https://arxiv.org/abs/1201.0490. Acesso em: 9 mar. 2026.

PEFFERS, K. *et al.* A Design Science Research Methodology for Information Systems Research. **Journal of Management Information Systems**, [*s. l.*], v. 24, n. 3, p. 45–77, 2007.

REDDY GUJJALA, P. K. Optimizing ETL Pipelines with Delta Lake and Medallion Architecture: A Scalable Approach for Large-Scale Data. **International Journal For Multidisciplinary Research**, [*s. l.*], v. 6, n. 6, p. 55445, 2024.

SHAIK, B. **PostgreSQL Configuration:** Best Practices for Performance and Security. Berkeley, CA: Apress L. P, 2020.

SHIMAOKA, A. M. *et al.* Big Data na Saúde Pública: Análise do Ecossistema das Bases Epidemiológicas no Brasil: Big Data in Public Health: Analysis of the Epidemiological Database Ecosystem in Brazil. **Revista de Epidemiologia e Saúde Pública - RESP**, [*s. l.*], v. 3, n. 1, p. 167–177, 2025.

SILVA, V. J.; BONACELLI, M. B. M.; PACHECO, C. A. O sistema tecnológico digital: inteligência artificial, computação em nuvem e Big Data. **Revista Brasileira de Inovação**, [*s. l.*], v. 19, p. 1–31, 2020.

SOUIBGUI, M. *et al.* Data quality in ETL process: A preliminary study. **Procedia Computer Science**, [*s. l.*], v. 159, p. 676–687, 2019.

SUPABASE INC. **Supabase**. [*S. l.*], 2026. Disponível em: https://supabase.com/docs. Acesso em: 20 fev. 2026.

TORRES, D. R. *et al.* Aplicabilidade e potencialidades no uso de ferramentas de Business Intelligence na Atenção Primária em Saúde. **Ciência & Saúde Coletiva**, [*s. l.*], v. 26, n. 6, p. 2065–2074, 2021.

WOJCIECHOWSKI, A. E-ETL: framework for managing evolving etl processes. *In*: CIKM '11: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 2011, Glasgow Scotland, UK. **Proceedings of the 4th workshop on Workshop for Ph.D. students in information & knowledge management**. Glasgow Scotland, UK: ACM, 2011. p. 59–66. Disponível em: https://dl.acm.org/doi/10.1145/2065003.2065016. Acesso em: 10 mar. 2026.

YU, X. Disaggregation: A New Architecture for Cloud Databases. **Proceedings of the VLDB Endowment**, [*s. l.*], v. 18, n. 12, p. 5527–5530, 2025.

ZARATE, G. *et al.* Evolution of Extract-Transform-Load (ETL) processes towards data product pipelines. *In*: ESAAM 2024: 4TH ECLIPSE SECURITY, AI, ARCHITECTURE AND MODELLING CONFERENCE ON DATA SPACE, 2024, Mainz Germany. **Proceedings of the 4th Eclipse Security, AI, Architecture and Modelling Conference on Data Space**. Mainz Germany: ACM, 2024. p. 25–32. Disponível em: https://dl.acm.org/doi/10.1145/3685651.3686662. Acesso em: 19 fev. 2026.