Survey on Data Ingestion for AutoML

1st Gabriel Mac'Hamilton Renaux Alves
Department of Computer Engineering
University of Pernambuco
Recife, Brazil
gmra@ecomp.poli.br

2nd Alexandre Magno Andrade Maciel
Department of Computer Engineering
University of Pernambuco
Recife, Brazil
amam@ecomp.poli.br

***Abstract*—Automated machine learning (AutoML) is an increasingly popular approach to building machine learning (ML) models without the need for extensive human intervention. One key component of AutoML is automated data ingestion, which involves automatically collecting, cleaning, and preparing data for use in ML models. This paper aims to analyze the literature in order to identify how automated data ingestion is being developed in the literature. To achieve this goal, a survey was conducted on the state-of-the-art of automated data ingestion using a method based on a systematic literature review, in order to identify the existing practices. A total of 12 articles were initially found, however, after applying filters, only six of them were ultimately utilized in the research, showing that visual data navigation and validation as well as metadata inference are important features for automated data ingestion focused in AutoML.**

***Index Terms*—automl, automated data ingestion, machine learning**

I. Introduction

Although data are important organizational assets, they are difficult to handle, requiring skilled professionals to process, and analyze them. However, most people do not have professional training in technology [1], thus generating the need for software that allow data mining and analysis in an automated way, called AutoML systems. Such systems abstract several operations inherent to ML, allowing users to perform data analysis activities in a faster and more accessible way [2].

The first step in machine learning is data ingestion, which has as its main goal to consistently bring data into the ML pipeline [3]. To achieve that, the ingested data needs to be handled in a way that it can be consumed in the next steps of the flow. Since the main task of the user in an AutoML system is, according to F. Hutter, L. Kotthoff and J. Vanschoren [2], to provide the data to the flow, Data Ingestion becomes a key system component.

Automated data ingestion is particularly important in machine learning because it can help to reduce the time and effort required to prepare data for modeling. By automating the process of data wrangling, ML practitioners can spend more time building and refining their models. This can lead to faster and more accurate models, as well as more efficient teams [4].

With the motivation to identify the ideal way to perform the data ingestion step for AutoML systems, this work analyzes the literature to obtain information about methods that best suit this purpose. With the information gathered, this paper seeks to list the practices and techniques used in the articles analyzed that fit the needs of AutoML.

DOI reference number: 10.18293/SEKE23-097

II. Background

A. AutoML

Machine learning can be defined as an area of knowledge that is designed to bring to computers the ability to learn without the need for explicit programming [5]. As described by H. Hapke and C. Nelson [3], the ML model pipeline can be summarized into nine well-defined steps, starting at data ingestion and generating a loop that restarts upon reaching model feedback.

AutoML is a broad term, generally used to define systems that seek to automate ML activities. Its fundamental motivation for existence is due to some of the main limitations of the development of activities related to ML, being the need for a professional specialized in the development of such algorithms and the need for domain knowledge linked to the subject on which the algorithm is based, besides the fact that some related activities eventually become tedious and repetitive. AutoML tools can help accelerate the development of ML models, and democratize access to ML for individuals who do not have extensive expertise in data science [6].

Although automation of ML activities can be developed for any of its steps, according to F. Hutter, L. Kotthoff and J. Vanschoren [2] the most basic tasks for automation are hyperparameter optimization (HPO), meta-learning and neural architecture search. With the automation of such activities the authors indicate that several advantages are generated, such as: reduction of human effort in the development of data mining activities, improved performance of algorithms, improved reproducibility of academic work, reuse of successful models, improved feature selection, etc.

B. Automated Data Ingestion

The data ingestion is characterized by obtaining and transporting data from an external source into the ML workflow. This process is typically used by organizations to streamline the data collection process and reduce the need for manual data entry, improve data quality and accuracy, and save time and resources, which can improve decision-making and reduce the risk of costly mistakes. The data ingestion is usually suitable for tabular data, commonly in CSV, XML or database tables [3].

Within the methods found in the literature for developing data ingestion, we can highlight batch and streaming. Batch data ingestion is usually done through ETL (extract, transform, and load) routines that collect data from an external source and bring it into the workflow. The batch technique is used for data that does not need to be consumed in real time, such as a database snapshot. Streaming data ingestion, on the other hand, is used in cases where there is a need for real-time data consumption and specific technologies are used to support this type of need [7].

The data ingestion process can be done through a variety of methods, including Application Programming Interfaces (APIs) and other data integration tools. This means that data can be transferred between systems in a structured format, reducing the risk of errors and ensuring data consistency. File transfer protocols, such as FTP and HTTP, are also commonly used for automated data ingestion, these protocols allow for the transfer of large files between systems, making it possible to collect data in bulk [8].

C. Related Work

This section shows secondary studies that sought to identify data ingestion methods for various purposes or related activities. The main difference of the present work to the others found in the literature is the focus on the analysis of methods relevant to AutoML systems during the search for data ingestion solutions, resulting on the search for automated data ingestion.

In paper [7] an analysis of the state-of-the-art of data ingestion and integration is performed, due to the wide range of existing tools and methods for these types of operations. Throughout the paper, the main features and technologies used are analyzed, dividing them into three main groups: data integration tools, stream data ingestion tools and cloud data processing tools. On the subject of data integration tools, it is indicated that usually the platforms have a graphical interface and are batch processing oriented, being represented mainly by Microsoft SQL Server Integration Services[1], Talend [2] and Sqoop[3]. Regarding the Stream data ingestion tools, the authors present as main characteristics the support to data pipelining, support to streaming processing and a wide variety of designs and support for users, such as Apache tools[4], Oracle Stream Analytics[5] and IBM Streams [6]. Finally, regarding cloud-based tools, the main functionalities listed are the support for batch processing and streaming with the ability to be applied both for local and cloud situations, highlighting in this context tools such as Azure Stream Analytics[7], IBM Streams[6] etc.

In [9] the authors list the data ingestion methods used in the literature and develop a tutorial, through a survey, for developing Hybrid Transactional/Analytical Processing (HTAP) solutions. The authors list two design options for HTAP: a single system for OLAP and OLTP that allows an organization for data ingestion in common databases and analytical processing in columnar databases, and the second option would be two separate systems for OLTP and OLAP, this way the data stored in the OLTP system would have the original data that would be ingested in the OLAP system through an ETL routine, keeping data stored in both formats, allowing both types of analysis, but there are alternatives where the data for both OLTP and OLAP are stored in the same database, even if they come from different systems.

[10] aim through a survey to review the state-of-the-art of recent big data solutions, to support the selection of the more appropriate technologies. Throughout the work, comparisons are made between the technologies present in the market for the most different layers of big data. In the context of data ingestion, the authors present that this technology is inserted in the data access layer and the main tools used for development are Apache Sqoop[3], Apache Flume[8] and Apache Chukwa[9], characterized by stream data processing.

III. Methodology

For this study, a state-of-the-art survey methodology, based on the systematic literature review proposed by Kitchenham and Charters [11], was used. To achieve the research objectives, five steps were taken to plan the research, conduct exploratory research, and refine the research results. Each stage will be detailed in the following subsections.

A. Research Planning

In the planning stage were defined the research method, the general and specific objectives of this work, as well as the research questions. Considering the main focus of this work as mapping the methods used in the literature for automated data ingestion, the research questions were defined as follows:

- **RQ1:** How is automated data ingestion implemented?
- **RQ2:** What are the metrics observed for data ingestion performance?
- **RQ3:** Which are the most suitable techniques for automated data ingestion in AutoML systems?

Research question 1 (RQ1) aims to identify broadly what are the main features of automated data ingestion systems, as well as the technologies used to develop them. In research question 2 (RQ2), the goal is to identify how to quantitatively evaluate data ingestion methods, so that comparisons between them are possible. Finally, research question 3 (RQ3) seeks to compare the characteristics of existing automated data ingestion systems with the needs of an AutoML system.

B. Search Strategy

The following procedure was made iteratively to build the search string used to perform the state-of-the-art analysis:

[1]https://www.microsoft.com/pt-br/sql-server/sql-server-2022
[2]https://www.talend.com/
[3]https://sqoop.apache.org/
[4]https://www.apache.org/
[5]https://www.oracle.com/middleware/technologies/stream-processing.html
[6]https://www.ibm.com/cloud/streaming-analytics
[7]https://azure.microsoft.com/pt-br/products/stream-analytics/
[8]https://flume.apache.org/
[9]https://chukwa.apache.org/

1) Definition of terms that would serve as keywords based on the research questions and objectives of this paper;
2) Identification of synonyms for the defined keywords, aiming to increase the search results;
3) Use of Boolean operators such as "AND" and "OR" to filter the results;
4) Evaluation of the results obtained, by analyzing the titles of the papers listed as research results.

After the completion of the procedure mentioned above, it was defined as search string: ("automated data ingestion"). The string was used to obtain papers from the research sources, searching for journals and conference papers: IEEExplore, ACM, Scopus, Science Direct. As a result of this phase of the search, a total of 12 papers were obtained, one from IEEExplore, two from ACM, three from Scopus and six from Science Direct.

C. Study Selection

By reading the title and abstract of the articles, inclusion, and exclusion criteria were applied according to Table I. The application of such criteria allowed a reduction in the initial number of articles, leaving a total of six articles, of these: one from IEEExplore, zero from ACM, one from Scopus and four from Science Direct.

TABLE I
INCLUSION / EXCLUSION CRITERIA TABLE

Inclusion criteria	**Exclusion criteria**
Conference, Proceedings or Journal papers	Papers that are not in English
Primary studies	Papers under 3 pages in length
Complete papers	Duplicate or similar papers
Papers published after 2017	Papers unavailable for download or viewing
Papers that respond at least two research questions	

D. Quality Criteria Application

Aiming to filter the previously selected papers, for the present study, one quality criterion was chosen, namely: *paper brings information about automated data ingestion*. After applying the quality criteria, 6 articles were selected.

E. Evidence Collection

In the evidence collection stage, the selected papers were read, aiming to identify whether they could answer the research questions listed in the planning stage. A scoring system was developed for the level of responsiveness of each research question by the article, requiring a sum greater than 1.5 for the work to be considered relevant to the research. The scoring was defined as follows: answer completely (score: 1), partially answers (score: 0.5), do not answer (score: 0).

IV. Data Analysis and Synthesis

A. RQ1 – How is automated data ingestion implemented?

The work developed by Otamendi et al. [12] presents an automated data ingestion module that provides the user with the ability to manipulate geospatial data, images obtained from satellites and sensors. The user must ingest the first data manually and specify the features and characteristics of each image. To facilitate this step, the authors have developed data-driven methods to extract metadata from other databases and attempt to infer the metadata of the selection being ingested. In addition, the automated data ingestion module is capable of processing the data periodically after the ingestion is configured, and can capture data from satellites on premise or on cloud.

The paper authored by Hacker et al. [13] features a prototype, called Experiment Dashboard (ED), with the goal of demonstrating automated data ingestion capabilities for research lab data with a minimal amount of effort and expertise. The data ingester automatically captures the data at the source and ingests it in real time into the database, where the information is queried for the ED. The user manipulates the data capture and analyzes the results, when satisfied with the collected data, presses an approval button and ingests the data into the project repository.

Deagen et al. introduce in their paper [14] an Autonomous Experimental System for polymer data analysis. In the system, there is a data ingestion module based on interfaces and data standards that allow end users with little knowledge in technology to use the platform without difficulty. The result of the project is a REST API developed in Python language for data access and a web interface to browse and organize the data visually.

Regarding data ingestion, the Pałys & Pałys [15] article presents the SAP system module called SAP Document and Reporting Compliance (DRC), used for electronic taxes. The module has support for data inputs such as XML, JSON and direct reporting. Throughout the process, there is transparency and traceability of the data by the user, allowing a data ingestion that is configurable and can be automatized by the customer.

B. RQ2 – What are the metrics observed for data ingestion performance analysis?

Among the analyzed papers, the only one that brought the performance metrics was the one developed by Hacker et al. [13]. The authors present that in cases where real-time data is needed, due to the requirement for data availability, synchrony with time becomes an essential factor. Based on that, ingest reliability, speed, and latency become relevant factors.

C. RQ3 – Which are the most suitable techniques for automated data ingestion in AutoML systems?

Analyzing the papers to answer the research question 3, it was considered as the main requirements for AutoML systems as to bring support to non-expert users, such as user interfaces and the possibility for easy data ingestion automation.

The work of Otamendi et al. [12] provides a data analysis service focused on non-expert users. However, it is indicated in the article that the project has certain limitations regarding the automated data ingestion, due to the Open Data Cube

used in the project, which makes the automated data ingestion module not user-friendly for non-expert users, due to the need that the user previously knows the database to configure it. This problem is mitigated through a model that performs the extraction of metadata from other databases so that inference can be made about the characteristics of the current metadata, alleviating this need for the user.

In the paper by Hacker et al. [13] several features are listed that made the prototype a success among non-expert users, such as:

- A reliable, functional, and performing data management framework for data collection and analysis;
- When a real-time data analysis is needed, time synchrony is a key factor;
- The tool should be easy to operate without the need for special expertise, allowing the user to review, check and approve data before the ingestion.

The programmatic interface presented in the Deagen et al. [14] article allows developers to configure automated data ingestion, while for end users, there is a GUI that allows visual validation of the data that has been uploaded and easy data navigation. It is important to emphasize that the project was developed with a focus on the end users (non-experts), who participated in the development stage through interviews.

Although not indicated to focus on non-expert users, the SAP system presented in the work of Pałys & Pałys [15] brings the possibility of automation of data ingestion, abstracting the difficulties of this step of the process, thus making this section of the system relevant to the current research.

V. Conclusion

As presented throughout this paper, the democratization of data analysis is a decisive factor for organizational success today, since it empowers decision makers. In this context, this paper look to contribute by defining the most suitable data ingestion techniques based in what is used in the literature to abstract this step for non-expert users in AutoML systems.

Exploring the research results, we can see that automated data ingestion can be developed in several ways, almost always focusing on the non-expert user. From the paper analysis we can conclude that important metrics for the evaluation of data ingestion are ingestion reliability, speed, and latency. Relevant features to be present in such systems can be observed as the navigation on the data to be ingested with a visual interface, providing the user the ability to approve or disapprove the data ingestion, based on the visualized data and have a wide range of data input options such as CSV, XML, JSON, etc.

In addition to the general features, important factors can be observed regarding the support of the non-expert user, the focus of AutoML systems, in the data ingestion stage. These factors can be listed as the existence of a user interface that is easy to use, the inference of metadata from the databases being used and the reliability of the data that is being manipulated.

Machine learning has an increasingly important role in society, so, the development of automated techniques for building ML models will be critical for advancing the field. As the field of AutoML continues to grow and evolve, we can expect to see many more improvements in the area of automated data ingestion.

VI. Future Work

This work acts as the initial phase of an ongoing research that aims to develop an automated data ingestion module to support business users in obtaining data for the development of activities in AutoML software.

VII. Acknoledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] Bouneffouf, D., Aggarwal, C., Hoang, T., Khurana, U., Samulowitz, H., Buesser, B., Liu, S., Pedapati, T., Ram, P., Rawat, A., Wistuba, M., & Gray, A. (2020, July). Survey on Automated End-to-End Data Science? 2020 International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/ijcnn48605.2020.9207453

[2] Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). Automated Machine Learning. The Springer Series on Challenges in Machine Learning. https://doi.org/10.1007/978-3-030-05318-5

[3] Hapke, H., & Nelson, C. (2020, July 28). Building Machine Learning Pipelines: Automating Model Life Cycles with TensorFlow.

[4] Patel, J. (2020, August). The Democratization of Machine Learning Features. 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). https://doi.org/10.1109/iri49571.2020.00027

[5] Simon, P. (2013, March 5). Too Big to Ignore: The Business Case for Big Data.

[6] Nagarajah, T., & Poravi, G. (2019, March). A Review on Automated Machine Learning (AutoML) Systems. 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). https://doi.org/10.1109/i2ct45611.2019.9033810

[7] Hlupic, T., & Punis, J. (2021, September 27). An Overview of Current Trends in Data Ingestion and Integration. 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO). https://doi.org/10.23919/mipro52101.2021.9597149

[8] Sangaiah, A. K. (Ed.). (2019, July 26). Deep Learning and Parallel Computing Environment for Bioengineering Systems. Academic Press.

[9] Özcan, F., Tian, Y., & Tözün, P. (2017, May 9). Hybrid Transactional/Analytical Processing. Proceedings of the 2017 ACM International Conference on Management of Data. https://doi.org/10.1145/3035918.3054784

[10] Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2018, October). Big Data technologies: A survey. Journal of King Saud University - Computer and Information Sciences, 30(4), 431–448. https://doi.org/10.1016/j.jksuci.2017.06.001

[11] Kitchenham, & Charters. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering.

[12] Otamendi, U., Azpiroz, I., Quartulli, M., Olaizola, I., Perez, F., Alda, D., & Garitano, X. (2021, July 11). Geo-Imagery Management and Statistical Processing in a Regional Context Using Open Data Cube. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. https://doi.org/10.1109/igarss47720.2021.9553940

[13] Hacker, T., Dyke, S., Ozdagli, A. I., Marshall, G., Thompson, C., Rohler, B., & Yeum, C. M. (2017, December). A Researcher-oriented Automated Data Ingestion Tool for rapid data Processing, Visualization and Preservation. Advances in Engineering Software, 114, 134–143. https://doi.org/10.1016/j.advengsoft.2017.06.019

[14] Deagen, M. E., Walsh, D. J., Audus, D. J., Kroenlein, K., de Pablo, J. J., Aou, K., Chard, K., Jensen, K. F., & Olsen, B. D. (2022, November). Networks and interfaces as catalysts for polymer materials innovation. Cell Reports Physical Science, 3(11), 101126. https://doi.org/10.1016/j.xcrp.2022.101126

[15] Pałys, A., & Pałys, M. (2022). The challenges for tax compliance in multinationals within the SAP environment. Procedia Computer Science, 207, 2384–2394. https://doi.org/10.1016/j.procs.2022.09.297