Information Dynamics and Applications
https://www.acadlore.com/journals/IDA

A Scalable Framework to Analyze Data from Heterogeneous Sources at Different Levels of Granularity

Iqbal Hasan[1], S.A.M. Rizvi[1], Majid Zaman[2*], Waseem Jeelani Bakshi[3], Sheikh Amir Fayaz[3*]

[1] Department of Computer Sciences, Jamia Millia Islamia, 110025 Delhi, India
[2] Directorate of IT & SS, University of Kashmir, 190006 J&K, India
[3] Department of Computer Sciences, University of Kashmir, 190006 J&K, India

[*] Correspondence: Majid Zaman (zamanmajid@gmail.com); Sheikh Amir Fayaz (skh.amir88@gmail.com)

Received: 09-05-2022 **Revised:** 10-11-2022 **Accepted:** 10-31-2022

Citation: I. Hasan, S. A. M. Rizvi, M. Zaman, W. J. Bakshi, and S. A. Fayaz, "A scalable framework to analyze data from heterogeneous sources at different levels of granularity," *Inf. Dyn. Appl.*, vol. 1, no. 1, pp. 26-34, 2022. https://doi.org/10.56578/ida010104.

Abstract: There is an enormous amount of data present in many different formats, including databases (MsSql, MySQL, etc.), data repositories (.txt, html, pdf, etc.), and MongoDB (NoSQL, etc.). The processing, storing, and management of the data are complicated by the varied locations in which the data is stored. If combined, this data from several sites can yield a lot of important information. Since many researchers have suggested different methods to extract, examine, and integrate the data. To manage heterogeneous data, researchers propose data warehouse and big data as solutions. However, when it comes to handling a variety of data, each of these methods have limitations. It is necessary to comprehend and use this information, as well as to evaluate the massive quantities that are increasing day by day. We propose a solution that facilitates data extraction from a variety of sources. It involves two steps: first, it extracts the pertinent data, and second, then to identify the machine learning algorithm to analyze the data. This paper proposes a system for retrieving data from many sources, such as databases, data sources, and NoSQL. Later, the framework was put to the test on a variety of datasets to extract and integrate data from diverse sources, and it was found that the integrated dataset performed better than the individual datasets in terms of accuracy, management, storage, and other factors. Thus, our prototype scales and functions effectively as the number of heterogeneous data sources increases.

Keywords: Heterogeneous sources; MongoDB; Staging server; Big data; Geographical data

1. Introduction

Data is growing at an exponential rate, with the last two years alone generating 80-90% of the world's data [1]. There are currently 555 million websites, up 300 million from the previous year [2], making for a diverse range of data sources, each with its own framework and structure. The issue of user-desired quantitative methodology must be recognized and resolved as soon as possible. Big data extraction now faces additional difficulties as heterogeneous data sets proliferate quickly. Since there are numerous ways of sources, including social media, news, government departments, the health sector, the agriculture sector, and other more departments, and this is increasing on a daily basis, scholars and researchers are aware of the value of the data in order to extract patterns from the raw data into information. These sources of information can contain different types of data, which include data stored at database levels, and various file systems like pdf/txt and HTML file formats. So in order to deliver real insights, we need a unique system to function and execute queries from one or more data sources [3].

The handling of data from numerous sources presents challenges for many departments and organizations. These data sources include duplicate data, time series data, and internet transactions. Hospitals are one of the best examples of heterogeneous data environments, where information about patients is kept in one database while information about X-rays, medications, and other information is kept in other databases due to the need to manage both tabular and image data for every single patient. To meet accreditation requirements, this data needs to be normally maintained, organised, accessed, and assessed in accordance with a defined format. Other organizations, such as meteorological departments, also struggle with handling the data because different rainfall data

characteristics are stored on a regular basis. This poses a major threat to the efficiency of database systems. Satellite data, GIS data, and data gathered from RADAR systems are all included in the data that the meteorological agencies retain. These forms of data are extremely difficult to manage and require a lot of processing power for storage, retrieval, and management. Furthermore, many of this data is stored in an unstructured manner using a number of languages and formats. Traditional methods of data management are useless because there is so much data and it is so complex [4, 5].

Consequently, presenting the diverse unstructured material in a systematic fashion will always be difficult. Data warehousing, big data, and other approaches have all been proposed as ways to address the heterogeneity challenges from various levels of granularity, but each has its own set of drawbacks when it comes to effectively managing this varied mix of data [6].

2. Review of Literature

One common method for addressing data heterogeneity problems is to transform and combine them into a single data source [7-10]. Lexical, syntactic, and geometrical inconsistencies that arise during data integration may cause some information to be lost when all of the data sources are integrated into one source, increasing the storage capacity requirement to that of a data warehouse.

Rehman et al. [11] offer a method for getting better data collection outcomes. The model is said to reduce data in the beginning stages. The suggested approach in this study did not take into account the usage of ML approaches or data processing; instead, it concentrated simply on data minimization.

Machine learning is now required in order to manage the data due to the increase in data from sources such as literature, databases, and repositories. The author's research [12] aims to review works involving knowledge discovery in various data sources using machine learning techniques. The fundamental ideas, typical tools, and implementation of ML in this are also summarized. This data generated from heterogeneous sources have been analyzed using machine learning techniques.

In the study [13], the authors covered the issues that can occur when integrating diverse data at a metallurgical facility. They introduced a model of information for defining the specifications of information on metallurgical output. The authors examined integration approaches and the potential for applying them to integrate metallurgical businesses with diverse data.

Hashim et al. [14] provide an integrated, mediated, and data warehouse-based architecture for cognitive integration. They employ two different kinds of ontologies—local and global ontologies—to direct the integration and manage syntactic, structural, and semantic heterogeneity.

The ideas, relations, and characteristics are treated as primordial and, as such, irreducible entities in the [15] study's suggested investigation of ontology-driven semantic data integration in an open environment. In light of the intentional conceptualization model, the formal intentional account of both ontology and ontological commitment is also proposed. Additionally, a proposed intentional model for ontology-driven mediated data integration in an open context. The suggested model takes into account the open environment's dynamic nature and intentionally describes the data from data sources. The link between global and local ontologies is then defined, together with the formal intentional semantics of query response.

All of the above studies had limitations, which leads us to offer a possible solution to the heterogeneity problem. Data processing, data management, and, most importantly, data storage is among the current problems. Many researchers are motivated by data processing to create complex heterogeneous systems; nevertheless, evaluating vast volumes of data demands greater computer capacity. Furthermore, data management and storage are perpetual issues, with no comprehensive answer in sight. As a result, the primary objective for undertaking this research is to address these current issues. All of the systems under discussion support just a limited number of data sources. Wrappers must be generated manually or hard-coded, and only a few data sources [16-22] appear to be supported.

3. Towards the Solution

Many academics have proposed a variety of approaches to handle the heterogeneous data at different granularity levels, however all systems have some limitations. Data warehouse and big data are a couple of these. These already-used options each have their own set of drawbacks. The following is a discussion of the drawbacks of each currently in use solution:

3.1 Data Warehouse: Limitations

In a variety of industries, data warehouses have been developed [23-26]. On the other hand, modern DWs face brand-new scientific challenges. Today's data sources are, in fact, numerous, independent, adaptable, and distributed. Due to these difficulties, traditional data warehouses are constrained in some ways, including in terms of data essence, availability, storage methods, and so forth. According to reference [27], this is because of a lack

of scalability caused by processing difficulties combined with inherent data problems and restrictions on the underlying hardware, application software, and other infrastructure. When employing data warehouses to address data heterogeneity, many researchers currently encounter a number of problems, including the following:

1. A major organization's data warehouse construction is a difficult operation that can take years to complete [28].
2. Data warehouse administration is also challenging and needs a team with advanced technological knowledge [29].
3. Monitoring data ware issues for quality when heterogeneity of data is taken into account, both quality and consistency of data are not up to par [30].

3.2 Big Data: Limitations

In order to analyze massive amounts of data, big data systems need high-speed computer infrastructure, which can be costly in terms of data collection, storage, processing, and visualization [31-39].

Big data systems have a number of challenges, one of which is the need to protect corporate and individual privacy through verification and security. These problems include data management, which may be expensive and time-consuming when done at a heterogeneous level. Additionally, managing the storage of data that is impressively large in size can be tough since it is always a difficult process when storage is taken into account. Processing such a vast volume of data continues to be a major difficulty since it always seems to be problematic. The management of data from several heterogeneous sources is the final but not the least problem. When you have data from one or more sources with diverse structures and from different platforms, a number of challenges are associated.

Due to the fact that each of the two workable options so far has its own set of limitations. Therefore, in addition to this, we need to find a better solution that will take care of all the problems that may arise in the future.

3.3 Proposed Solution

The framework that enables researchers to examine data from several heterogeneous sources, such as various databases, data repositories, and other sources, is the subject of this section. The difficulty of data analysis from various sources will rise because there are n databases with n*m tables and that number is growing exponentially.

By separately extracting the data from each of these data sources, our framework enables the utilization of the data from a variety of data sources for analysis. In our method, the user queries the system, and based on the query, the analytics engine invokes some analytic models, fetching the data to the staging server regardless of structure. The model sends data requests to the data source interface while it is running. Different data interface procedures have been created in the data source interface (Staging server) to retrieve data in the format required by the analytical model. The element that engages with the various data sources is called wrappers [40]. DI routines are in charge of query creation, query execution, formatting, and returning the result data to the analytical system. It should be independent of the type, size, and structure of the data in order to evaluate the heterogeneous data and test the scalability.

We make the assumption that there is no centralization of the data from the indicated sources prior to conducting the experiment. When sources are identified, it also indicates that meta data, which contains details about the data, is present. The method is that all sources—including different databases (MySQL etc), data repositories (txt, html, pdf, etc.), and NoSQL (MongoDB) —are shown in below (Figure 1).

Figure 1. Heterogeneous data sources

A staging server [41, 42] or temporary server is also available. There is no data source stored on the staging server. The data is inserted into this staging server in line with the structure after being taken from various sources (databases, data repositories, and NoSQL). Thus, the data in staging server is flushed without storing the data, every time the process finishes. The potential issue with this is that it may result in data redundancy or that the data we get is not normalized, but the important focus here is data retrieval alone. Following data retrieval, we can undertake any normalization strategy.

3.3.1 Architectural framework

The experiment to analyze the data from the heterogeneous sources follows a two fold process:

1. Extract the relevant information.
2. Identification of machine learning algorithm to analyze the data.

Algorithm

For the extractions of relevant information we have:

(1). Identification of sources:

The server administrator is in responsible of this, thus there may be some manual work involved. Every repository that is present must have some basic information added by the administrator.

(2). Data extraction:

Keywords, sets of numbers, characters, and other data extraction items provide the type of data for which information is required from various sources.

Accessing the meta data is the initial stage in the data extraction process. The data may be present from the source in one of three ways. Databases, data repositories, and NoSQL are all included in this. Each of the n kinds of the database must be visited. We also have several repositories and NoSQL. Because every database has a unique structure, there is a lot of heterogeneity present in databases. However, because NoSQL databases all have the same structure, there is very little heterogeneity (Figure 2).

The staging server acts as a data source interface where requests are sent to the module from different analytics models of the analytics engine. The data interface function is aware of the schema of the data source it needs to connect to. In order to retrieve the required data, the data interface procedure formulates the query in accordance with the data source's format. The wrappers that are created to access a certain sort of source are really the information that is extracted. Consequently, the process of identification and extraction is finished.

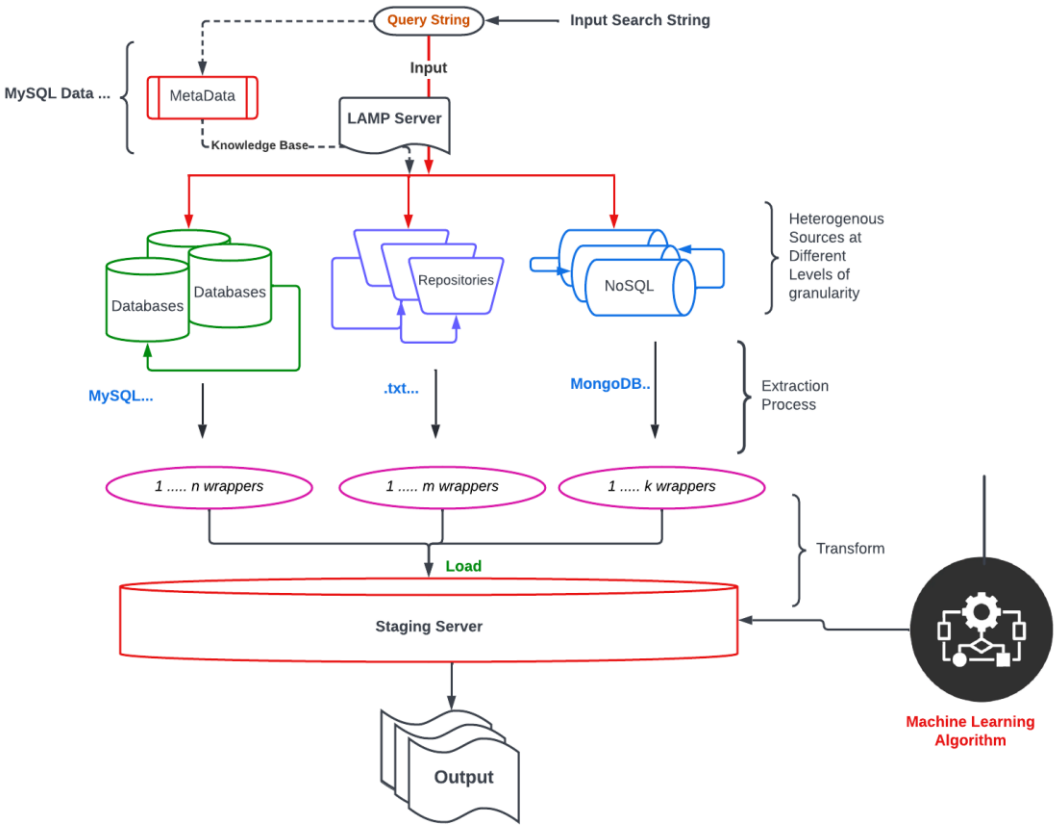

Figure 2. Extraction and transform process from different sources

3.3.2 Implementation & results

The framework was implemented on following structure shown below (Figure 3). The first phase information was distributed into multiple sources. Subsequently decision tree (Iterative Dichotomiser 3) algorithm was implemented on it. A decision tree is constructed by recursively dividing data divisions into smaller partitions based on splitting rules or criteria. A heuristic for selecting criteria that optimally separates a class labelled training dataset into different classes is an attribute selection measure or splitting rules. The attribute selection measure should be such that splitting produces pure partitions, meaning that all entries in a given partition belong to the same class. The ID3 Algorithm is a decision tree algorithm that works on the information gain and entropy. In the next phase framework was implemented, data was transformed and loaded.

Once the data is accumulated and stored in the staging server appropriate algorithm is applied on the data based on structure, volume etc of the data. After the data from the different sources have been extracted, it is time to select a machine learning algorithm to look at the data. Since the primary purpose of this work is to examine a variety of data sources without identifying the data, consequently, Iterative Dichotomizer 3 (ID3) algorithm have been applied on the data which was extracted from different sources and the results are generated accordingly.

The experiment was carried out on a geographical dataset collected from different data sources as shown in the Figure 3. In this paper we have implemented a basic classification algorithm in order to check the accuracy level at each different sources and latter on the same data is integrated into one file and the same algorithm is implemented to check the final accuracy after data integration.

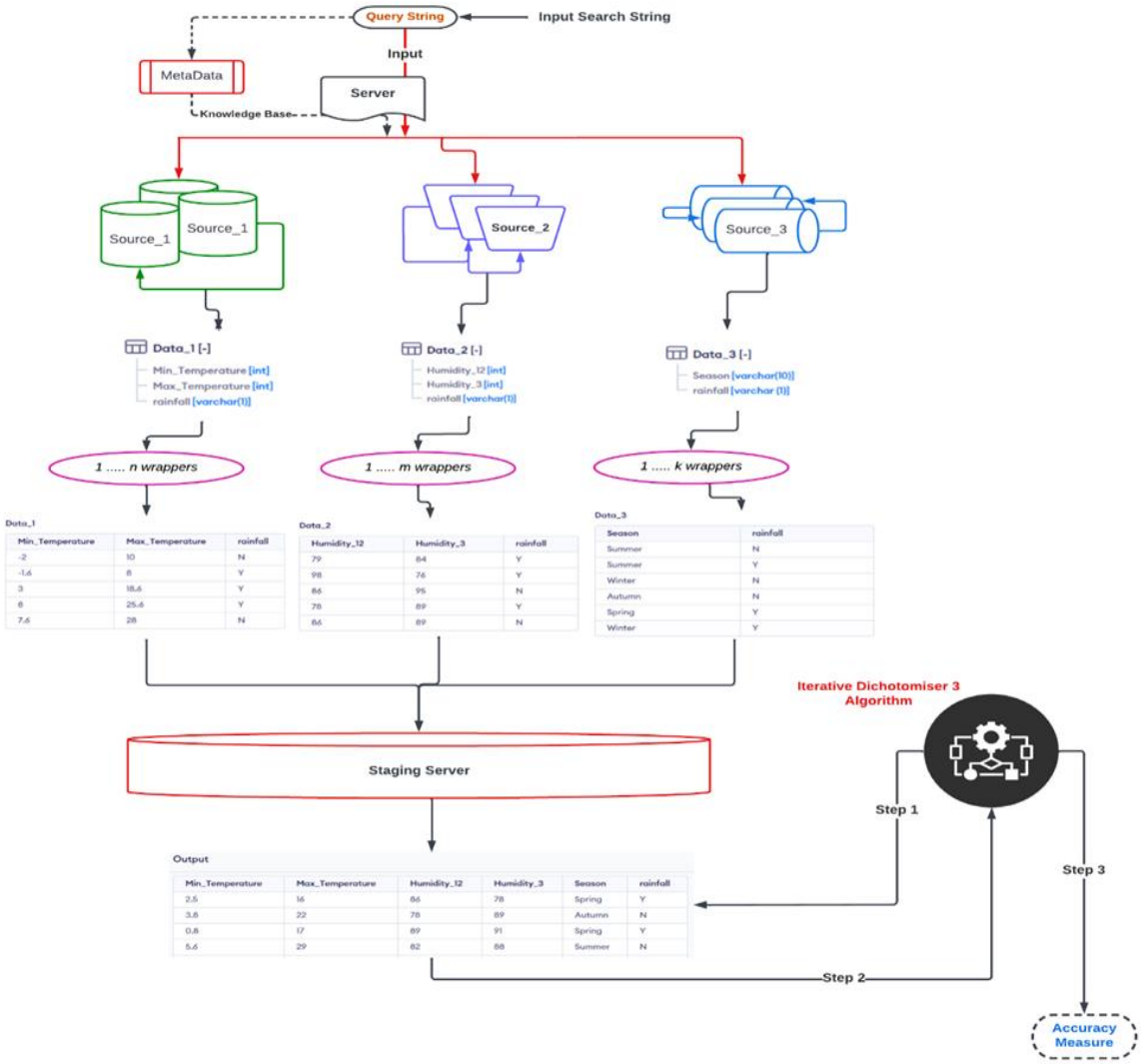

Figure 3. Implemented framework

The data collected from different heterogeneous sources consists of different meteorological parameters including Humidity at 12 am and 3 pm, minimum and maximum temperature, season and rainfall as the target parameter. The integrated dataset contains around 6,000 records. In order to check the accuracies at individual

levels and at the final level, we have divided the dataset into 70% training and 30% testing ratio. The experiment results which we obtained in processing the data is shown below Table 1. It was discovered that the source 1, source 2, and source 3 datasets had few parameters, resulting in lower accuracy when compared to integrated data. The reason for the high accuracy measure is that the integrated data contains all of the parameters present in all of the source data, and it is obvious that the more parameters in a data set, the greater the chances of high and correct accuracy measures, and the results are as per the ID3 algorithm.

Table 1. Accuracy statistics

Model (ID3)	Source_1 Data	Source_2 Data	Source_3 Data	Integrated Data
Records	*5491*	*5491*	*5491*	*5491*
Test set	*1650*	*1650*	*1650*	*1650*
Training	*3841*	*3841*	*3841*	*3841*
Correctly classified	*1281*	*1352*	*1179*	*1502*
Accuracy	*77.6%*	*81.9%*	*71.4%*	*91.03%*

The graphical representation of the experimental results is shown below (Figure 4). As we can see the accuracy measures at different data sources are comparatively less and the same integrated dataset has an accuracy measure of 91%, which shows the proposed solution works well as compared to data stored at different granularity levels.

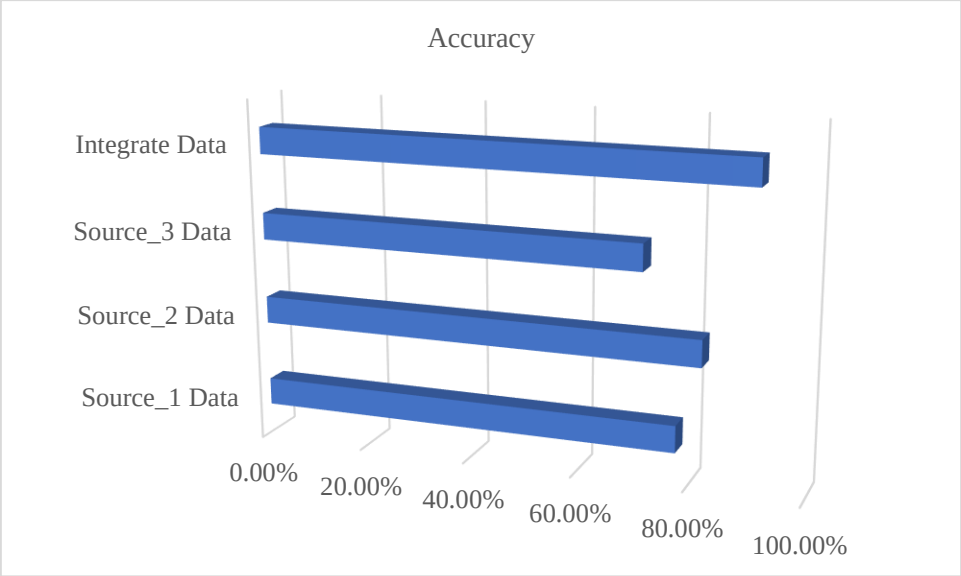

Figure 4. Accuracy statistics at heterogeneous levels of data

4. Discussion

Heterogeneity with volume makes integration difficult and if the same is complemented by multiple sources a problem is manifold [43-45]. Machine learning over the years is progressing and has attained standardization based on type of algorithms. However scalability and granularity is still random not standardized.

In this paper, a framework is proposed which will retrieve data from numerous sources including databases, data sources and NoSQL. Stored on varying infrastructure it is not humanly possible to have machine learning algorithm implemented on multiple sources in isolation. This will generate contradictory results and may lack correlation between varying parameters. Accordingly, a framework is proposed which shall work on varying sources, structures and shall also integrate data into staging server where upon machine learning algorithms are used.

Staging server is temporary storage data storage (buffer) which is used to hold integrated data temporarily while as meta data (knowledge base) is permanent data storage having information about whole framework including source, also deletion, modification of repositories.

Furthermore, a typical (ID3) technique was then used on data obtained from various sources and then on an integrated source leads to accuracy measurement. The accuracy was computed using the decision trees IF-THEN else rules. Since we may improve the data's correctness by applying other decision tree extensors such as random

forests, extra trees, and several ensembled techniques, this will serve as a benchmark for all other methods implemented.

5. Applications of Proposed Methodology

Due to the fact that numerous methods have been offered by numerous researchers throughout the years in order to examine the data from diverse sources. The currently suggested approach has a lot of problems and restrictions, such management and data storage. The difficulties that now exist and the suggested solutions that are offered in this paper are tabulated below (Table 2) as part of our attempt to address them.

Table 2. Proposed solutions results provided in this paper

Issues	Proposed solution results
Management	*It is clear that since we are not transferring the data to new infrastructure, there are no additional administration charges or handling requirements.*
Storage	*We used the idea of a staging server to address this problem. The staging server is flushed each time the request has been handled. We can avoid using additional storage by using this.*
Processing	*Non-managed files are not covered by the staging server for the purposes of delta data capture and data propagation. The non-managed files must be manually copied from the staging server to the production server. This doesn't affect the overall processing of the system servers.*
Heterogeneity	*A framework is proposed which shall work on varying sources, structures and shall also integrate data into staging server where upon machine learning algorithms are used. This will help us to tackle with the heterogeneous sources.*

6. Conclusion

For extracting and analyzing the data from the heterogeneous source, we described a framework in this study. This framework performs distributed cross-source join operations and allows users to specify changes that enable joinability instantly at query time. This framework uses a two-step procedure in which it first extracts data from various sources and then uses any conventional machine learning algorithm to analyze the information. By depending on the connectors of the latter, this architecture eliminates users from manually creating wrappers, which is a major bottleneck in enabling data variety throughout the literature. It was discovered that the integrated dataset outperformed the component datasets in terms of accuracy, management, storage, and other characteristics when the framework was tested on a variety of datasets to extract and integrate data from various sources. The performance of the suggested framework on various datasets with increased levels of variability will be examined in the future. This will enable us to test the framework's scalability in diverse settings.

Author Contributions

Iqbal Hassan, S.A.M. Rizvi & Majid Zaman: Framed the main idea of the work, implementation, interprets the results, Data curation, Data collection and Visualization. Waseem Jeelani Bakshi: Study plans with all authors. Provides the basic idea of the work, Design and draft of the manuscript. Sheikh Amir Fayaz: Study conception and Investigation, testing of the results, editing of the manuscript and paper writing.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. Zaman and M. A. Butt, "Information translation: A practitioners approach," In World Congress on Engineering and Computer Science, (WCECS 2012), San Francisco, USA, October 24-26, 2012, pp. 45-47.

[2] S. A. Fayaz, A. I. Altaf, A. N. Khan, and Z. H. Wani, "A possible solution to grid security issue using authentication: An overview," *J. Web Eng. Technol*, vol. 5, no. 3, pp. 10-14, 2018.

[3] R. Zaki, A. Barabadi, J. Barabady, and A. N. Qarahasanlou, "Observed and unobserved heterogeneity in failure data analysis," *J. Ris.*, vol. 236, no. 1, pp. 194-207, 2022. https://doi.org/10.1177/1748006X211022538.
[4] J. B. Christopher, T. Elizabeth, and S. Y. David, "Behavioural science is unlikely to change the world without a heterogeneity revolution," *Nat. Hum. Behav.*, vol. 5, no. 8, pp. 980-989, 2021. https://doi.org/10.1038%2Fs41562-021-01143-3.
[5] S. H. Matthewes, "Better together? Heterogeneous effects of tracking on student achievement," *Econ. J.*, vol. 131, no. 635, pp. 1269-1307, 2021.
[6] N. Fathy, G. Walaa, B. Nagwa, and H. Mohamed, "Ontology-based data access to heterogeneous data sources: State of the art approaches and applications," *Int. J. Intell. Comput. Inform. Sci.*, vol. 2022, pp. 1-10, 2022. https://doi.org/10.21608/ijicis.2022.110450.1144.
[7] S. A. Fayaz, M. Zaman, and M. A. Butt, "Numerical and experimental investigation of meteorological data using adaptive linear M5 model tree for the prediction of rainfall," *Rev. Comput. Eng. Res.*, vol. 9, no. 1, pp. 1-12, 2022. https://doi.org/10.18488/76.v9i1.2961.
[8] S. Kaul, S. A. Fayaz, M. Zaman, and M. A. Butt, "Is decision tree obsolete in its original form? A burning debate," *Revue d'Intelligence Artificielle*, vol. 36, no. 1, pp. 105-113, 2022.
[9] S. A. Fayaz, M. Zaman, and M. A. Butt, "Knowledge discovery in geographical sciences—A systematic survey of various machine learning algorithms for rainfall prediction," In International Conference on Innovative Computing and Communications, Singapore, 2021, Springer, pp. 593-608.
[10] S. A. Fayaz, M. Zaman, and M. A. Butt, "An application of logistic model tree (LMT) algorithm to ameliorate Prediction accuracy of meteorological data," *Int J. Adv. Technol. Eng. Explor.*, vol. 8, no. 84, pp. 1424-1440, 2021. http://dx.doi.org/10.19101/IJATEE.2021.874586.
[11] M. H. U. Rehman, V. Chang, A. Batool, and T. Y. Wah, "Big data reduction framework for value creation in sustainable enterprises," *Int. J. Inform. Manag.*, vol. 36, no. 6, pp. 917-928, 2016. https://doi.org/10.1016/j.ijinfomgt.2016.05.013.
[12] M. E. Günay and R. Yıldırım, "Recent advances in knowledge discovery for heterogeneous catalysis using machine learning," *Catal. Rev.*, vol. 63, no. 22, pp. 1-45, 2021. http://dx.doi.org/10.1080/01614940.2020.1770402.
[13] S. Porshnev, O. Ponomareva, A. Borodin, and S. Mirvoda, "Problems and methods for integrating heterogeneous data (on Example, Metallurgical Production)," In 2019 International Multi-Conference on Industrial Engineering and Modern Technologies, (FarEastCon), Vladivostok, Russia, October 1-4, 2019, IEEE, pp. 1-5. http://dx.doi.org/10.1109/FarEastCon.2019.8934336.
[14] H. Hashim, A. Ahmed, N. Salim, A. Osman, O. Y. S. Heng, A. Sim, A. Bakri, N. H. Zakaria, R. Ibrahim, and S. S. Omar, "A new database integration model using an ontology-driven mediated warehousing approach," *J. Theor. Appl. Inform. Technol.*, vol. 58, no. 2, pp. 392-409, 2005.
[15] M. A. Islam, "Ontology-driven semantic data integration in open environment," In Electronic Thesis and Dissertation Repository, USA: The University of Western Ontario, 2020.
[16] S. A. Fayaz, M. Zaman, and M. A. Butt, "To ameliorate classification accuracy using ensemble distributed decision tree (DDT) vote approach: An empirical discourse of geographical data mining," *Procedia Comput. Sci.*, vol. 184, pp. 935-940, 2021.
[17] A. Fayaz, M. Zaman, S. Kaul, and M. A. Butt, "Is deep learning on tabular data enough? An assessment," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 466-473, 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130454.
[18] S. A. Fayaz, M. Zaman, and M. A. Butt, "Performance evaluation of GINI index and information gain criteria on geographical data: An empirical study based on JAVA and Python," In International Conference on Innovative Computing and Communications, Singapore, 2021, Springer, pp. 249-265.
[19] I. Altaf, M. A. Butt, and M. Zaman, "A pragmatic comparison of supervised machine learning classifiers for disease diagnosis," In 2021 Third International Conference on Inventive Research in Computing Applications, (ICIRCA 2021), Coimbatore, India, September 2-4, 2021, IEEE, pp. 1515-1520.
[20] L. Mutanu and Z. Khan, "Enhancing the QoS of a data warehouse through an improved ETL approach," *J. Lang. Technol. Entrep. Afr.*, vol. 12, no. 2, pp. 89-102, 2022.
[21] B. Kormeier and K. Hippe, "Data warehousing of life science data. integrative bioinformatics," In *Integrative Bioinformatics*, Singapore: Springer, pp. 85-96, 2022.
[22] M. Golfarelli and R. Stefano, "Modern principles and methodologies," In Data Warehouse Design, Philadelphia: McGraw-Hill, Inc., pp. 480-480, 2009.
[23] A. Sebaa, F. Chikh, A. Nouicer, and A. Tari, "Research in big data warehousing using Hadoop," *J. Inform. Syst. Eng. Manag.*, vol. 2, no. 2, pp. 1-1, 2017. http://dx.doi.org/10.20897/jisem.201710.
[24] L. Wu, R. Sumbaly, C. Riccomini, G. Koo, H. J. Kim, J. Kreps, and S. Shah, "Avatara: OLAP for web-scale analytics products," *Proc. Vldb. Endow.*, vol. 5, no. 12, pp. 1874-1877, 2012. https://doi.org/10.14778/2367502.2367525.

[25] K. Krishnan, Data Warehousing in the Age of Big Data, USA: Newnes, 2013. https://doi.org/10.1016/c2012-0-02737-8.
[26] R. Mohd, M. A. Butt, and M. Z. Baba, "GWLM–NARX: Grey Wolf Levenberg–Marquardt-based neural network for rainfall prediction," *Data Technol. Appl.*, vol. 2020, 2020.
[27] S. A. Fayaz, S. Kaul, M. Zaman, and M. A. Butt, "An adaptive gradient boosting model for the prediction of rainfall using ID3 as a base estimator," *Revue d'Intelligence Artificielle*, vol. 36, no. 2, pp. 241-250, 2022. https://doi.org/10.18280/ria.360208.
[28] S. A. Fayaz, M. Zaman, and M. A. Butt, "A hybrid adaptive grey wolf Levenberg-Marquardt (GWLM) and nonlinear autoregressive with exogenous input (NARX) neural network model for the prediction of rainfall," *Int. J. Adv Technol. Eng. Explor.*, vol. 9, no. 89, pp. 509-522, 2022. http://dx.doi.org/10.19101/IJATEE.2021.874647.
[29] V. Christos, H. Vasilis, and Z. Nabil, "Introduction to big data in education and its contribution to the quality improvement processes," In Big Data on Real-World Applications, Croatia: InTech, pp. 41-64, 2016. http://dx.doi.org/10.5772/63896.
[30] C. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inform. Sci.*, vol. 275, pp. 314-347, 2014.
[31] M. M. Rathore, A. Ahmad, A. Paul, and A. Daniel, "Hadoop based real-time big data architecture for remote sensing earth observatory system," In 2015 6th International Conference on Computing, Communication and Networking Technologies, (ICCCNT), USA, 13-15 July, 2015, IEEE, pp. 1-7. https://doi.org/10.1109/ICCCNT.2015.7395242.
[32] "Institutional analytics is hard work: A five-year journey," EDUCAUSE Review, 2016, http://er.educause.edu/~/media/files/articles/2016/8/erm1656.pdf.
[33] S. Aljawarneh and J. A. Lara, "Special issue on quality assessment and management in big data part I," *ACM T. Embed. Comput. S.*, vol. 13, no. 2, pp. 1-3, 2021. https://doi.org/10.1145/3449052.
[34] P. K. Premkamal, S. K. Pasupuleti, and P. J. A. Alphonse, "Efficient escrow-free CP-ABE with constant size ciphertext and secret key for big data storage in cloud," *Int J. Cloud Appl. Com.*, vol. 10, no. 1, pp. 28-45, 2020. https://doi.org/10.4018/ijcac.2020010103.
[35] R. A. Habeeb, F. H. Nasaruddin, A. Gani, I. A. Hashem, E. Ahmed, and M. A. Imran, "Real-time big data processing for anomaly detection: A Survey," *Int. J. Inf. Manag.*, vol. 45, pp. 289-307, 2019.
[36] W. Ahmad and B. Alam, "An efficient list scheduling algorithm with task duplication for scientific big data workflow in heterogeneous computing environments," *Concurr. Comp.-Pract. E.*, vol. 33, no. 5, Article ID: e5987, 2021.
[37] S. A. Fayaz, M. Zaman, and M. A. Butt, "A super ensembled and traditional models for the prediction of rainfall: An experimental evaluation of DT versus DDT versus RF," In Communication and Intelligent Systems, Singapore, 2022, Springer, pp. 619-635. http://dx.doi.org/10.1007/978-981-19-2130-8_48.
[38] A. F. Sheikh, S. Jahangeer, Z. Majid, and A. B. Muheet, "Machine learning: An introduction to reinforcement learning," In Machine Learning and Data Science, Fundamentals and Applications, Hoboken, USA: Wiley-Scrivener, pp. 1-22, 2022.
[39] N. N. Dalvi, R. Kumar, and M. A. Soliman, "Automatic wrappers for large scale web extraction," ArXiv, vol. 2011, 2011.
[40] A. Sabtu, N. F. Azmi, N. N. Sjarif, S. A. Ismail, O. M. Yusop, H. M. Sarkan, and S. Chuprat, "The challenges of Extract, Transform and Loading (ETL) system implementation for near real-time environment," In 2017 International Conference on Research and Innovation in Information Systems, (ICRIIS), Langkawi, Malaysia, 16-17 July 2017, IEEE, pp. 1-5. https://doi.org/10.1109/ICRIIS.2017.8002467.
[41] S. A. Fayaz, M. Zaman, S. K. Kaul, and M. A. Butt, "How M5 Model Trees (M5-MT) on continuous data are used in rainfall prediction: An experimental evaluation," *RIA*, vol. 36, no. 3, pp. 409-415, 2022. http://dx.doi.org/10.18280/ria.360308.
[42] H. Mehmood, E. Gilman, M. Cortes, P. Kostakos, A. Byrne, K. Valta, S. Tekes, and J. Riekki, "Implementing big data lake for heterogeneous data sources," In 2019 IEEE 35Th international conference on data engineering workshops, (ICDEW), Macao, China, 08-12 April 2019, IEEE, pp. 37-44. https://doi.org/10.1109/ICDEW.2019.00-37.
[43] A. Cuzzocrea, "Big data lakes: models, frameworks, and techniques," In 2021 IEEE International Conference on Big Data and Smart Computing, (BigComp), Jeju Island, Korea 17-20 January 2021, IEEE, pp. 1-4. https://doi.org/10.1109/BigComp51126.2021.00010.
[44] A. M. Olawoyin, C. K. Leung, and A. Cuzzocrea, "Open data lake to support machine learning on Arctic big data," In 2021 IEEE International Conference on Big Data, (Big Data), Orlando, FL, USA, 15-18 December 2021, IEEE, pp. 5215-5224. https://doi.org/10.1109/BigData52589.2021.9671453.
[45] N. M. Mir, S. F. Khan, M. A. Butt, and M. Zaman, "An experimental evaluation of Bayesian classifiers applied to intrusion detection," *Indian J. Sci. Technol.*, vol. 9, no. 12, pp. 1-7, 2016.