SAMRIDDHI
A Journal of Physical Sciences, Engineering and Technology
ISSN(P): 2229-7111
ISSN(O): 2454-5767

AN ETL FRAMEWORK: FOR IMPLEMENTATION OF A MODERN DATA WAREHOUSE FOR REAL TIME ANALYTICS

Monika Patel
Department of Computer Science, Gujarat Vidyapith, Ahmedabad, Gujarat, India.
monapatel18r@gmail.com

Dr. Dhiren B Patel
Department of Computer Science, Gujarat Vidyapith, Ahmedabad, Gujarat, India.
dhiren_b_patel@gujaratvidyapith.org

Abstract: Data warehousing is the concept of storing historical data to get the insight of past and guidance for future. Previously data usually generated in relational formats, hence it was easy to manage them for analytics purpose. Nowadays data comes in verity of forms such as structured, semi-structured and unstructured forms, and that makes it difficult to store them into data warehouse and analyze them. Recently, NoSQL databases have become popular because of semi-structured and unstructured handling capacity. Many data warehouses are available to manage structured data but there are few or very less research done in the field of NoSQL based data warehousing.

The proposed solution provides an ETL framework with automated pipeline for modern data warehouse which can store and manage semi-structured data. To implement the proposed research authors have used MongoDB-document oriented database to store semi-structured documents. MongoDB also adds benefits of scalability and schema free structure for storing and accessing data. Moreover in the proposed work machine leaning techniques are used for accurate prediction. This analytics ready data at data warehouse reduces the execution time and provides real time analytics. The work is validated through the experiment and it shows the improvement in execution time with compared to manual process.

Keywords: Data Warehouse, NoSQL/SQL, ETL, Real Time Analytics

1. Introduction

Real time analysis of stored and streaming data is the key to success of any industry for accurate decision making. From many years storing of historical data is done implemented through data warehouses in the form of repository. Previously, these data warehouses were competent to control structured data only. In present era, various form of data produced from businesses, social media, blogs and e-commerce websites. Modern data warehouses are required to hold such semi-structured, unstructured data and structured [1], [2].

Extract-Transform-Load - ETL is the practice of preparing the records for the data warehouse. Cleaning, splitting, joining, transporting and aggregating the data are the subtasks

of ETL process. Data Warehouse is the storage for historical business data, which can be used to take better decisions based on past data. ETL is the process used to load data in the data warehouse. The proposed work addresses the problem of storing and analyzing semi-structured and structured data with non-relational data model.

2. Related Work

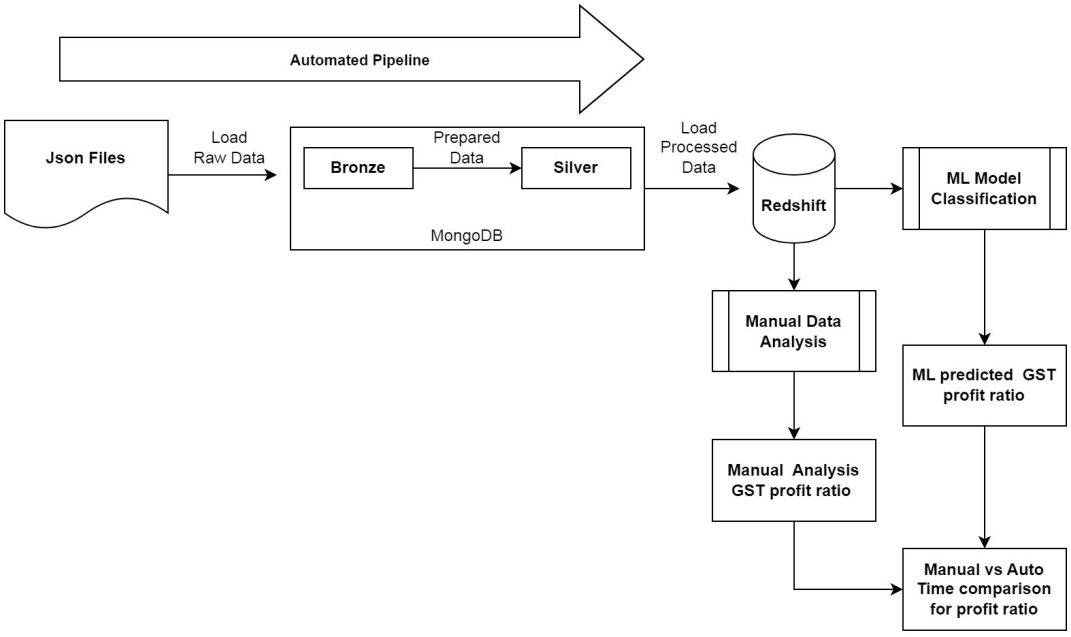
[2] Authors have designed the ETL framework for new age data warehouse, which is capable of handling and analyzing semi-structured and unstructured data through document oriented database (MongoDB). In the stated research, authors have done depth literature review for the types of ETL tools and its uses. Also this paper highlights and provides the guideline for area of research in the field of ETL process[1]. This paper focuses on design of big data warehouse schema for storage, transform and analyzing data. Paper also covers the criteria and features for data warehouse modeling[3]. This paper proposes the service oriented ETL model for big data for adaptability and extensibility for more consistent and reliable analytics[4]. In this research paper authors have focused on ERP systems and Business Intelligence for the big data. To reduce the load on the system, ETL technique with Capture Data Change is designed[5][6]. This research paper focuses on data warehouse technologies for data maintenance, collection and how to increase the usability of data from it[7]. This paper at hand provides the review of real-time ETL progression for data warehousing and found that there are many opportunities of research in the area of ETL and data warehousing[8]. This paper offers the extraction technique using Model Driven Architecture and designing of transformation rules to store document-oriented data in the data warehouse[9].The authors have highlights the increase in unstructured data due to many online platforms. The schema extraction approach is proposed using Google Cloud Platform[10]. The proposed research provides the solution for multi model DBMS using star schema, using classical star schema and the native form of schema[11]. This paper automatically creates big data pipeline for a data warehouse, which is able to manage semi-structured and unstructured data[12][13]. The authors have developed a automated pipeline – SIRAD, that joins the datasets from various government agencies and provides unique information[14]. This paper focuses on challenges of ETL process at the various stages of its execution. They have proposed the solution for better ETL process using artificial intelligence[15][16]. In this paper authors have discussed about ETL process for NoSQL based data warehouses. They have proposed the transformation rules using BPMN to convert multidimensional model into document oriented model[17], [18].

3. Proposed Work

SAMRIDDHI
A Journal of Physical Sciences, Engineering and Technology
ISSN(P): 2229-7111
ISSN(O): 2454-5767

ETL skeleton commonly used to renovate data into structured type for the storage in data warehouses for further analysis. Nowadays data comes in the various forms like structured, semi-structured or unstructured, and it is time consuming to convert these data into structure shape prior to loading them into the data warehouses. Also the demand for accurate and immediate results increased for survival in such a competitive market. There is a need for an improved ETL framework for a modern data warehouse, to grip such unstructured and semi-structured data without converting them into structured form for the data warehouse.

Projected ETL framework is intended to address the said problem for real time analytics with the modern technologies. In this research, the ETL framework moves step by step to provide the desired results. Step 1: Raw data (JSON documents) are loaded in the MongoDB - Extraction. Step 2: The raw data will be transform into bronze and silver collection – Aggregation / Transformation. Step 3: Data will be loaded into the data warehouse – AWS Redshift. Step 4: Multi class classification will be applied on data Step 5: Machine Learning algorithms will be applied for analysis ready data by the side of the data warehouse in support of real time analytics. Step 6: Manual facts analytics will perform. Step 7: Comparison of manual analysis results and Machine learning analysis results. The flow of research is revealed in the Fig.1.



[Flow of Proposed work]

Fig.1. Flow of planned work

SAMRIDDHI
A Journal of Physical Sciences, Engineering and Technology
ISSN(P): 2229-7111
ISSN(O): 2454-5767

4. Implementation Methodology

For the implementation of the research, GST (Goods and service tax) JSON documents are used. Each client has sales (GSTR 1) and purchase (GSTR 2) files monthly, with various tax rates.

Bronze Data Collection: total 42900 JSON documents

Silver Data Collection: total 25500 JSON documents (aggregated)

Silver data will be loaded in the Amazon Redshift to create data warehouse.

The proposed framework will predict the profit ratio for the client based on tax rate. To predict the profit category, annual profit is calculated based on monthly tax files. Based on the annual profit multi class classification is done on the data. We have classified 4 classes based on dataset.

Class 0: When profit ratio is negative or below zero.

Class 1: When profit ratio is between 0 and 5000000.

Class 2: When profit ratio is between 5000000 and 10000000.

Class 3: When profit ratio is greater than 10000000.

After classifying the data into 4 categories, four efficient and relevant ML algorithms applied for the further analysis. The process of modeling is as follows:

1. Select important features from the data

['gstin', 'year', 'rt', 'purchase_txval', 'sales_txval', ‘opening stock’, ’closing stock’, 'profit_cat’]

2. Select input feature

['gstin', 'year', 'rt', 'purchase_txval', 'sales_txval', ‘opening stock’, ’closing stock’]

3. Select target column: ['profit_cat']

4. Feature cleanup: label encoding

5. Scale the feature: standard scaling

6. Split the data into train data: 80 % and test data: 20 %

After modelling, it is seen that Random Forest Classifier performs best in the given situation and gives 99 % Accuracy.

5. Result and Discussion

SAMRIDDHI
A Journal of Physical Sciences, Engineering and Technology
ISSN(P): 2229-7111
ISSN(O): 2454-5767

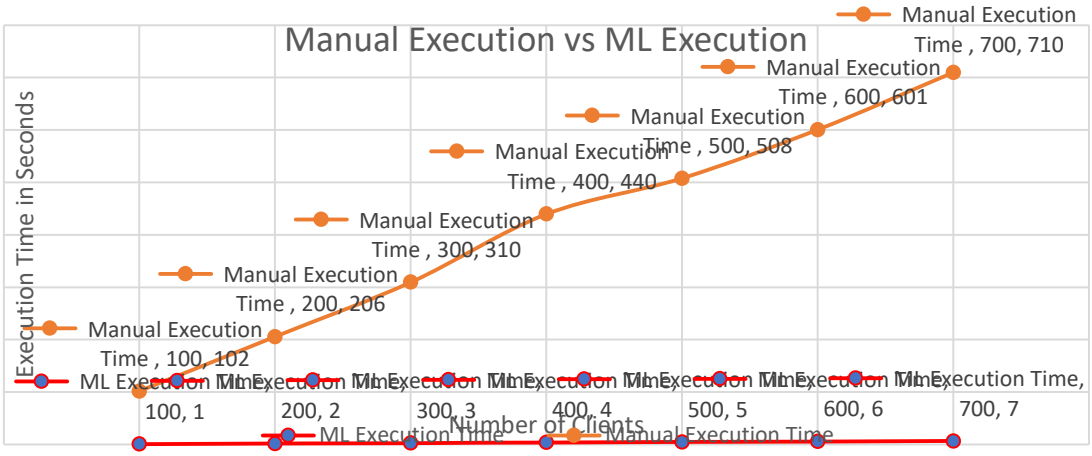


Fig.2. Execution time Manual vs ML

To evaluate the proposed research, MongoDB as Database, Redshift as Data Warehouse, and Python as Programming Language is used. Analysis is performed manually as well as automatically using machine learning algorithm. Execution time is recorded after each phase of execution and it has been clearly seen that using ML prediction time is much lesser than manual approach. As we can see in the Fig. 2, prediction for 700 clients only takes approx 7 seconds, whereas manually it takes approx 710 seconds.

6. Conclusion

This research work is carried out to implement and design an ETL framework for data warehouse creation for semi-structured data without converting them into structured form. To implement this task, an automated ETL pipeline is created which cleanse, aggregate and transforms data and also stores into the cloud based data store. Then, the statistics in the data warehouse is classified using multi class classification technique. Set of ML algorithms have applied to the classified data to know the best algorithm for profit ratio prediction. The random forest classifier offers the utmost accurateness with compared other Classifier. The proposed research have successfully reduces the execution time for prediction with 99% accuracy.

7. References

[1] M. Patel and D. B. Patel, *Progressive Growth of ETL Tools: A Literature Review of Past to Equip Future*, vol. 1187. Springer Singapore, 2021. doi: 10.1007/978-981-15-6014-9_45.

[2] M. Patel and D. B. Patel, “Data Warehouse Modernization Using Document-Oriented ETL Framework for Real Time Analytics,” *Lect. Notes Networks Syst.*, vol. 434, pp. 33–41, 2022,

doi: 10.1007/978-981-19-1122-4_5.

[3] A. Dhaouadi, K. Bousselmi, M. M. Gammoudi, S. Monnet, and S. Hammoudi, "Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons," *Data*, vol. 7, no. 8, Aug. 2022, doi: 10.3390/data7080113.

[4] B. Oliveira, M. Leite, Ó. Oliveira, and O. Belo, "A Service-Oriented Framework for ETL Implementation BT - Progress in Artificial Intelligence," 2022, pp. 636–647.

[5] S. Thulasiram and N. Ramaiah, "Real Time Data Warehouse Updates Through Extraction-Transformation-Loading Process Using Change Data Capture Method," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 44, Springer, 2020, pp. 552–560. doi: 10.1007/978-3-030-37051-0_62.

[6] F. Hamami, I. A. Dahlan, S. W. Prakosa, and K. F. Somantri, "Big Data Analytics for Processing Real-time Unstructured Data from CCTV in Traffic Management," *2020 Int. Conf. Data Sci. Its Appl. ICoDSA 2020*, pp. 6–10, 2020, doi: 10.1109/ICoDSA50139.2020.9212858.

[7] L. S. A. Nayak, K. Das, S. Hota, B. J. R. Sahu, and D. ahuti Mishra, "Implementation of Data Warehouse: An Improved Data-Driven Decision-Making Approach BT - Intelligent and Cloud Computing," 2022, pp. 419–427.

[8] F. de A. Vilela and R. R. Ciferri, "A Survey of Real-Time ETL Process Applied to Data Warehousing Environments BT - ITNG 2022 19th International Conference on Information Technology-New Generations," 2022, pp. 91–96.

[9] F. Abdelhedi, R. Jemmali, and G. Zurfluh, "Relational Databases Ingestion into a NoSQL Data Warehouse," *arXiv Prepr. arXiv2203.06949*, 2022.

[10] A. N. Istiqamah and K. R. S. Wiharja, "a Schema Extraction of Document-Oriented Database for Data Warehouse," *Int. J. Inf. Commun. Technol.*, vol. 7, no. 2, pp. 36–47, 2021, doi: 10.21108/ijoict.v7i2.584.

[11] S. R. Sandro Bimonte, Enrico Gallinucci, Patrick Marcel, "Data Variety, Come As You Are in Multi-model Data Warehouses Sandro," *Elsevier*, 2022.

[12] K. Ghane, "Big data pipeline with ML-based and crowd sourced dynamically created and maintained columnar data warehouse for structured and unstructured big data," *Proc. - 3rd Int. Conf. Inf. Comput. Technol. ICICT 2020*, pp. 60–67, 2020, doi: 10.1109/ICICT50521.2020.00018.

[13] M. Y. Santos, C. Costa, and C. Andrade, "Enhancing Big Data Warehousing for Efficient , Integrated and Advanced Analytics : Visionary Paper The final authenticated version is available online at : https://doi.org/ Enhancing Big Data Warehousing for Efficient , Integrated and Advanced Analytics V," no. June, 2019, doi: 10.1007/978-3-030-21297-1.

[14] M. Howison and M. Goggins, "SIRAD: Secure Infrastructure for Research with Administrative Data," *Softw. Impacts*, vol. 12, p. 100245, 2022, doi:

https://doi.org/10.1016/j.simpa.2022.100245.

[15] M. V.-T. J. of C. and Mathematics and undefined 2021, "Novel Solution for Real Time Challenges of ETL in Big Data," *Turcomat.Org*, vol. 12, no. 10, pp. 3661–3674, 2021, [Online]. Available: https://www.turcomat.org/index.php/turkbilmat/article/view/5054

[16] S. Gorhe, "ETL in Near-Real Time Environment: Challenges and Opportunities," no. November, 2020, [Online]. Available: https://www.researchgate.net/publication/340938742

[17] R. Yangui, A. Nabli, and F. Gargouri, "ETL based framework for NoSQL warehousing," *Lect. Notes Bus. Inf. Process.*, vol. 299, no. August, pp. 40–53, 2017, doi: 10.1007/978-3-319-65930-5_4.

[18] H. Mallek, F. Ghozzi, O. Teste, and F. Gargouri, "BigDimETL with NoSQL Database," *Procedia Comput. Sci.*, vol. 126, pp. 798–807, 2018, doi: 10.1016/j.procS.2018.08.014.