International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Functional Code Development for Existing Data Transformations of Retail Industry Data Warehouse

Annlip Gour, Mayank Junankar, S Akshansh, Ankita Ghule

Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur
Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur
Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur
Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur

---

Abstract - Several ETL tools with graphical user interfaces and additional built-in functions have been developed to simplify the creation and maintenance of ETL processes. However, a drawback of using GUI-based ETL solutions is that they limit the flexibility of the software for customization by customers, as it remains closed-source. To overcome this issue, our project adopts a different approach that utilizes functional code written in the Python scripting language. ETL activities are defined by Python functional code, enabling users to easily and efficiently create complex ETL tasks with multiple resources and parallel tasks. Furthermore, our approach simplifies ETL development and surpasses GUI techniques in terms of testing, cost and flexibility.

***Key Words*:** ETL (Extract Transform Load), python, functional codes, data warehouse

1.INTRODUCTION

Data transformations are essential for performing operations on collected data. The ETL process, which stands for Extract, Transform, and Load, involves extracting data, transforming it, and loading it into a Data Warehouse. Various tools have been developed to facilitate the development and maintenance of transformations, offering graphical user interfaces and built-in functionalities such as transformation libraries and documentation generation. However, a disadvantage of using such GUI tools is their lack of open-source availability, which limits the flexibility of the software for customization by customers. In our project, Python code is used to define transformation tasks, enabling the implementation of various ETL transformations and allowing users to easily and efficiently define complex ETL tasks or modify existing ones using the full flexibility of Python.

2. Literature Review

1. "Data Warehousing for E-Commerce" by Jiawei Han and Jian Pei - This paper provides an overview of the challenges and opportunities in building a data warehouse for e-commerce. It discusses the data extraction, transformation, and loading (ETL) process, as well as the architecture and design of the retail data warehouse. The authors also present a number of techniques for optimizing the performance and scalability of the warehouse.

2. "Data Transformation Techniques for Retail Data Warehouses" by Wei Chen and Phillip B. Gibbons - This paper surveys various data transformation techniques for retail data warehouses, including data cleaning, data normalization, data enrichment, and data integration. The authors also discuss the trade-offs between different techniques and provide insights into how to choose the best approach for a given situation.

3. "Enhancing Data Warehouse Performance with Parallel Processing" by Jianxin Li and Wei Chen - This paper presents a parallel processing framework for enhancing the performance of retail data warehouses. The framework is designed to improve the scalability and reliability of the warehouse, as well as to reduce the time and resources required for data transformation. The authors also provide a detailed evaluation of the framework using a real-world retail data warehouse.

4. "Data Transformation for Retail Data Warehouses: A Case Study" by Xiaojun Wan and Wei Chen - This case study describes a data transformation project for a retail data warehouse. The authors present the challenges faced during the project, including data quality issues, data integration challenges, and performance optimization. They also discuss the solutions used to overcome these challenges, including data cleaning and normalization, data enrichment, and parallel processing.

5. "Data Transformation for Retail Data Warehouses: An Overview" by Wei Chen and Xiaojun Wan - This paper provides an overview of the data transformation process for retail data warehouses. The authors discuss the challenges and opportunities in transforming data for retail analytics, including data quality issues, data integration challenges, and performance optimization. They also provide a comprehensive survey of the existing data transformation techniques, including data cleaning, data normalization, data enrichment, and data integration.

These papers provide a comprehensive overview of the challenges and opportunities in functional code development of existing data transformations for retail data warehouses. They cover various data transformation techniques and provide insights into how to choose the best approach for a given situation.

3. Proposed System

We propose to develop a retail data management system based on Python functional codes for performing the transformations. This will save the cost of the software for the retailer and provide more customizability and options for transformations to the managers. The project will be based on SQL for storing data, and Python libraries like Pandas and NumPy for data transformations and analytics.

Table -1: Data Transformations offered by the system

Transformation	Description
Read from source	Asks for the location of file to be read by the system.
Write to target	Asks for a target file location where you wish to save the loaded/modified data.
Average	A transformation that can perform sum, average calculations on large groups of data.
Join	A transformation that joins data from two sources into a single one, by a primary key.
Convert	It is a transformation that performs conversion of table format data to csv or json format.
Clean	A cleanse function is used to standardize the content of the data.
Data Masking	The process of masking sensitive data to generate realistic test data for non-production environments is a transformation technique that is commonly used.
De-duplicate	This property can be used to find instances of duplicate elements in a data set and optionally remove duplicate records/ make a separate list of duplicated items.
Expression	A transformation that performs user defined calculations on separate rows of data.
Filtering	It is a transformation that filters data from the data stream.
Labeler	A labeler can be used to mark types of information and assign labels to datatypes.
Look-up	It searches for data and defines lookup conditions and return values.

After the creations of all the above stated functional codes, the data can be analyzed and visualized using libraries like Matplotlib which will give a clear presentation to the managers to take the required actions as per the data.

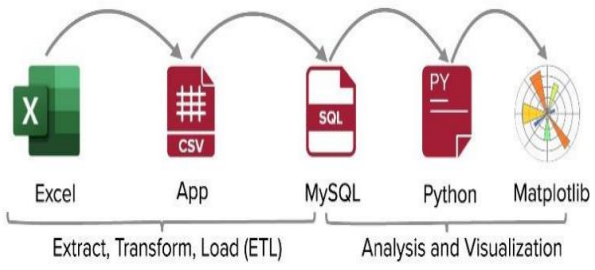

Fig -1: Figure

4. CONCLUSIONS

Data management is an important aspect in retail stores and keeping it secure is necessary as well. Using the sales data effectively can help in good inventory management which can act as a major factor in profitability. Excel sheets are not a secure repository for databases of important sales records. After loading the data, we combine different databases and excel sheets such as product details and transaction details into a normalized database and store them in SQL entirely using Python. This completes our ETL process. We have created functional, platform-independent Python code with support for various data sources such as products and transaction records for retail data. This includes writing short Python scripts, using different Python libraries, resulting in overall ease, fast project development and maintenance.

REFERENCES

1. ETLator – a scripting ETL framework Miran Radonić*1, Igor Mekterović*2 * Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

2. "PEP 249 - Python Database API Specification v2.0", https://www.python.org/dev/peps/pep0249/.

3. Jensen, S.K., Thomsen, C., Pedersen, T.B., Andersen, O. (2021). pygrametl: A Powerful Programming Framework for Easy Creation and Testing of ETL Flows. In: Hameurlain, A., Tjoa, A.M. (eds) Transactions on Large-Scale Data- and Knowledge-Centered Systems XLVIII. Lecture Notes in Computer Science(), vol 12670. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978- 3-662-63519-3_3

4. Biswas, N., Sarkar, A., Mondal, K.C. (2019). Empirical Analysis of Programmable ETL Tools. In: Mandal, J., Mukhopadhyay, S., Dutta, P., Dasgupta, K. (eds) Computational Intelligence, Communications, and Business Analytics. CICBA 2018. Communications in Computer and Information Science, vol 1031. Springer, Singapore. https://doi.org/10.1007/978-981-13-8581-0_22

5. Beyer, M.A., Thoo, E., Selvage, M.Y., Zaidi, E.: Gartner magic quadrant for data integration tools (2020)

6. www.datasciencecentral.com/profiles/blogs/10-open-source-ETL-tools.

7. J. Visser. Visitor combination and traversal control. In OOPSLA 2001 Conference Proceedings: Object-Oriented Programming Systems, Languages, and Applications, pages 270–282, 2001.

8. K. Fisher and R. Gruber. Pads: a domain-specific language for processing ad hoc data. In ACM PLDI, pages 295–304, 2005.

9. L. V. S. Lakshmanan, F. Sadri, and S. N. Subramanian. SchemaSQL: An extension to SQL for multi database interoperability. ACM Trans. Database Syst., 26(4):476–519, 200

10. Transformation types [17/11/2022- 02:20pm]

11. https://www.lucidchart.com/pages/examples/flowchart_software [20/10/2022- 08:56pm] [12]PostgreSQL[21/10/2022- 03:15pm].

12. Ali, S.M.F., Wrembel, R.: From conceptual design to performance optimization of ETL workflows: current state of

research and pen problems. VLDB J. (VLDBJ) 26(6), 777–801 (2017). https://doi.org/10.1007/s00778-017-0477-2

13. Andersen, O., Thomsen, C., Torp, K.: SimpleETL: ETL processing by simple specifications. In: 20th International Workshop on Design, Optimization, Languages, and Analytical Processing of Big Data (DOLAP). CEUR-WS.org (2018)

14. Beck, K.: Test Driven Development: By Example, pp. 194–195. Addison-Wesley Professional, Boston (2002)

15. Chandra, P., Gupta, M.K.: Comprehensive survey on data warehousing research. Int. J. Inf. Technol. (IJIT) 10(2), 217–224 (2018). https://doi.org/10.1007/s41870- 017-0067-y

16. https://www.informatica.com/resources/articles/what-is-etl.html

17. https://www.analyticsvidhya.com/blog/2020/03/understanding-transform-functionpytho