Automated Workflow Management In Data Engineering Using ELT

Prof. Sheetal Shimpikar[1], Rajashree Deokar[2], Barsharani Padhy[3], Sakshi Sobale[4], Shravani Patil[5]
[1, 2, 3, 4, 5] Dept of Information Technology Engineering
[1, 2, 3, 4, 5] Pillai College of Engineering Navi Mumbai, India

Abstract- *Organizations face challenges in managing data from various sources, including databases, APIs, and external platforms. Data inconsistency, redundancy, and inefficiencies hinder effective analysis. To address these issues, an ELT (Extract, Load, Transform) approach, combined with MongoDB and data visualization solutions, can be implemented. This involves extracting data from diverse sources, loading it into a storage system, performing transformations within MongoDB to clean, standardize, and structure the data for analysis, and using data visualization tools to present the transformed data in a meaningful way. This approach offers a streamlined, scalable framework for managing multi-source data, optimizing data quality, and facilitating decision-making insights.*

Keywords- ELT, Extraction, Transformation, Loading, Data Visualization, NoSQL Database MongoDB .

I. INTRODUCTION

Without a centralized system to handle the inconsistencies in data formats, organizations face challenges such as:

Data inconsistency: Different formats (e.g., JSON, CSV, etc.) lead to difficulties in comparing and analyzing data.

Data accessibility: Lack of a unified view of the data makes it hard for stakeholders to access accurate and real-time data for decision-making.

Time and resource inefficiency: Manual processes to clean, transform, and integrate data across systems are labor-intensive and error- prone.

Implementing an ELT (Extract, Load, Transform) approach combined with NoSQL Database MongoDB and data visualization solution can provide a comprehensive answer to these challenges.

In this approach, data is extracted from diverse sources and loaded in its raw form into a storage system, like MongoDB, which accommodates flexible and unstructured data. Transformations are then performed to clean the data for analysis. The final stage involves using data visualization tools to present the transformed data in a meaningful way, making insights accessible and actionable.By leveraging MongoDB for centralized storage and data visualization for output, this approach provides a streamlined, scalable framework for managing multi-source data, optimizing data quality, and facilitating insights for decision-making.

a) Objectives

1. The ELT method aims to Centralize and clean data in MongoDB for easy access, accuracy, and consistency.
2. Build a scalable system that handles increasing data sources and volumes, leveraging tools like Tableau or Power BI for user-friendly visualizations. Enable real-time insights through efficient data processing, ensuring quick access to actionable information.
3. Establish data protection rules and ensure regulatory compliance while training the team on effectively using MongoDB and data visualization tools. Regularly optimize system performance to maintain fast, efficient data access and processing.

b) Scope

The project scope involves implementing an ELT approach with MongoDB and data visualization tools by identifying data sources, defining extraction and loading processes, and transforming data for analysis. It includes setting up MongoDB for scalability, ensuring data governance for security and compliance, and selecting visualization tools to create insightful dashboards. Additionally, it encompasses developing training programs, monitoring performance metrics, and establishing continuous improvement plans to support data-driven decision-making.

II. LITERATURE REVIEW

Data warehouses (DWs) primarily receive their information from operational systems through a series of extraction, transformation, and loading operations known as ETL. ETL plays a crucial role in DW. It begins with the

extraction of data from many heterogeneous data sources, converts it to the appropriate format, and then loads it into the DW. ETL is hence crucial to the creation of DW.

Numerous studies on effective and centralized data engineering techniques, such as ELT, ETL, and Reverse ETL, are represented in this study.

In a recent paper named "Security Thinking in Online Freelance Software Development" [2] published in 2023 by Bashar Nuseibeh et al. reveals that security issues in online freelance software development are multivalent and varied, with differing scopes and expectations among developers. Financial investment also influences their attention to secure coding, with compensation for secure software development seen as requiring specialist knowledge. These findings suggest that tailored security interventions may be needed to support secure practices and build sustainable software for society.

Another paper "ETL, ELT and Reverse ETL: A business case Study" [3] published in the 2022 by Bharat Singhal et al. et al. compares ETL, ELT, and Reverse ETL are different approaches for processing data in a data warehouse. ETL extracts, transforms, and loads data, while ELT loads data first and then transforms it. Reverse ETL extracts data from the data warehouse and pushes it back to operational systems. The best method depends on factors like data size, complexity, and business needs

In a paper published in 2021 "Data Migration using ETL Workflow" [4] by Saranya N et al. states that data validation ensures data accuracy and consistency before migration. ETL is a common data migration process for data warehouses. Cloud storage provides reliable and scalable solutions for data migration.

In the paper "Design and Realization of an ETL Method in Business Intelligence Project" [6] published in 2018 by Guomin Zhang et al. presents a promising approach for improving ETL processes in BI projects, offering benefits in terms of efficiency, flexibility, data quality, and data integration

a) Literature Review Summary

A literature review is an objective, critical summary of published research literature relevant to a topic under consideration for research. The summary is presented here.
Table 2.1 Literature review summary

Sr No.	Paper Name	Advantages and Disadvantages
1.	Ingest and Visualize CSV Files using AWSPlatform For Transition from Unstructured toStructured Data (2023). [1]	Advantage:- simple data handling and analysis. Disadvantage :- complexity and cost of AWS infrastructure
2.	Security Thinking in Online Freelance Software Development (2023). [2]	Advantages: Highlights crucial security considerations often overlooked Disadvantages: Focuses only on one contexts, limiting generalizability to other software
3.	ETL, ELT and Reverse ETL: A business case Study (2023). [3]	Provides study regarding different data integration methods that can be used for storing and visualizing the data
4.	Data Migration using ETL Workflow(2021). [4]	Advantages: Streamlined data transfer, improved data quality, reduced downtime during migration. Disadvantages: Complexity in setup, potential for data loss or corruption, resource-intensive process.

5.	Extract Transform Loading (ETL) Based Data Quality for Data Warehouse Development (2021). [5]	Advantages: Improves data quality in data warehouses. Disadvantages: Adds complexity and resource requirements to the ETL process, potentially increasing development and maintenance overhead.

III. METHODOLOGY

3.1 Existing System Architecture

The information in DW (data warehouse) is mainly obtained from operational systems through a set of processes of extracting, transforming and loading data called ETL into the DW. ETL is an important component of DW, starting with extraction of data from various heterogeneous data sources, transforming those data to the required format and then loading those data into the DW. Therefore ETL has a very important role in DW development.

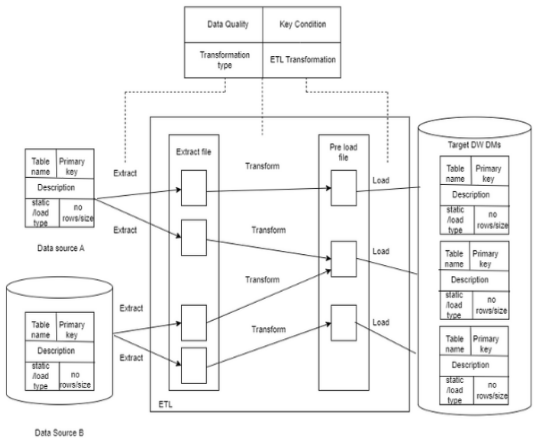

Fig. 3.1 Existing system architecture [5]

Data Source A and B are two different data sources, such as databases or files, each with tables with unique identifiers and descriptions. The ETL process involves extracting data from both sources, transforming it to match the target format, pre-loading a file for loading into the data warehouse, and loading the transformed data into the target data warehouse. The final destination contains similar metadata as the source tables, and the target data warehouse (DW) contains similar information. Key elements include data quality, which ensures accuracy and consistency during transformation, and key condition, which checks conditions based on the primary key. ETL transformation ensures the data is correctly transformed to fit the target data structure.

a) Advantages of the Existing Framework:

1. Improved Data Quality: By incorporating data quality checks throughout the ETL process, the framework aims to improve the overall quality of data in the data warehouse.
2. Reduced Errors: Early detection of data quality issues can help reduce errors and inconsistencies in the data warehouse.

b) Disadvantages of the Existing Framework:

1. ETL processes can become bottlenecks when dealing with large volumes of data or complex transformations, impacting performance and scalability.
2. Transforming data before loading can risk data loss or security breaches, while compliance regulations may require storing raw data, which ETL may not inherently support.

3.2 Proposed System Architecture

Data visualization in an ELT (Extract, Load, Transform) context leverages the flexibility of the ELT approach to load raw or semi-processed data into a data warehouse before transforming it into formats suitable for visualization and analysis. The main advantage of this approach is the ability to work with large volumes of data efficiently, transforming and preparing data for visualization directly in the data warehouse.

The ELT data pipeline simplifies data access and retrieval by extracting and loading data from various sources into a central data warehouse without extensive preprocessing. Transformations are deferred until after loading, allowing data teams to apply scaled transformations. This allows for aggregation, cleaning, and formatting of data for visualization. The immediate loading of data allows for real-time analysis of transformations, allowing for more dynamic and current insights.

ELT enhances visualization performance by enabling efficient transformations within the data warehouse, allowing for faster processing of complex calculations like aggregating sales data or joining tables across large datasets. It also offers scalability, allowing for the use of cloud or on-premise warehouses that can accommodate increasing data volumes, ensuring visualizations remain performant as data grows,

allowing organizations to visualize and explore historical data trends.Here's an outline of the proposed system:

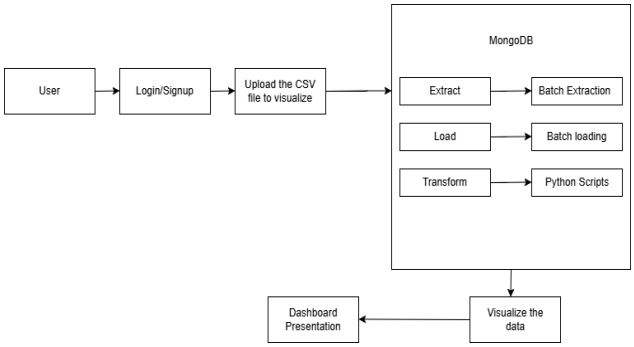

Fig.3.2 Data Flow Diagram

The system follows an Extract, Load, and Transform (ELT) methodology using Python, MongoDB NoSQL Database, and MongoDB Charts for efficient data processing and visualization. The user interface (UI) serves as the primary interaction point, enabling users to upload CSV files while ensuring format validation and data integrity. Real-time logs and progress indicators provide updates on extraction, loading, and transformation stages. Data extraction, implemented using Python scripts, retrieves data from uploaded files through batch extraction, which processes only new or modified records for efficiency, and full extraction, which loads the entire dataset for initial imports or full refreshes. The extracted data is stored in a staging collection in MongoDB, applying incremental extraction logic to optimize batch processing.

The loading phase involves inserting extracted data into MongoDB while ensuring efficient storage. Batch loading optimizes performance by loading only new or updated records, while full loading is used for database initialization or refreshes. Python scripts compare extracted records with existing ones and apply indexing strategies to improve query performance. Transformation, also handled via Python, structures and formats data before visualization, including field mapping for schema alignment, data field removal for storage efficiency, and rule-based classification for categorization. Finally, processed data is visualized using MongoDB Charts, where interactive dashboards provide analytical insights through bar charts, pie charts, and heatmaps. Users can filter data, drill down for details, and export reports in CSV or PDF formats, ensuring comprehensive data exploration and decision-making.

The technical requirements for MongoDB setup include a database server, sufficient storage, secure network configuration, ELT tools for data extraction, API access for external data pulling, and software like Tableau and Power BI for data visualization. Data requirements include source identification, catalog of data sources, and data quality assessment to ensure consistency and accuracy before extraction.tasks while maintaining high standards of performance and security.

IV. RESULTS

To address the challenges of data inconsistency, redundancy, accessibility, and inefficiency, I suggest implementing a centralized ELT (Extract,Load,Transform) pipeline using MongoDB and Python. This solution will automate data collection, transformation, and storage processes, providing a unified data repository accessible for real-time analysis and decision-making.

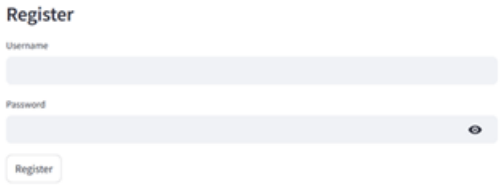

Fig 4.1 Sign up Page

Fig 4.2 Input Page

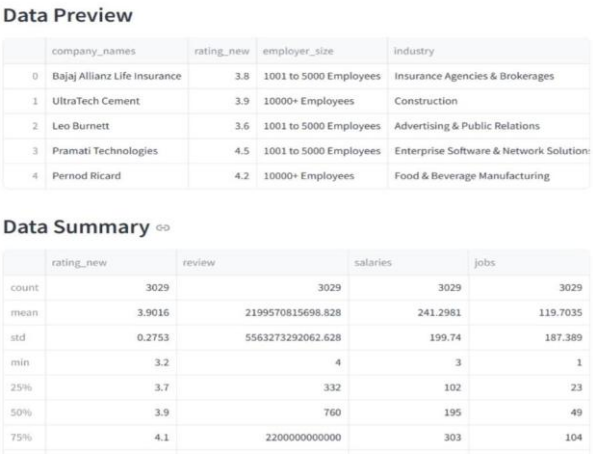

Fig 4.3 Dataset input preview

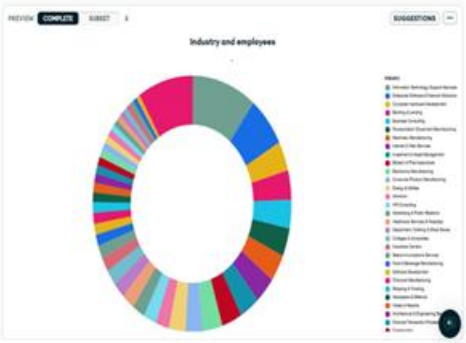

Fig. 4.4 Industry and employee Visualization

V. FUTURE DIRECTION

In consideration of the advancements achieved in this research endeavour, several promising avenues emerge for future exploration and development in the realm of managing a flexible workflow system.

The ELT-MongoDB-visualization approach can be improved through real-time data processing, advanced analytics, enhanced data visualization, cloud integration, and robust data governance and security measures. These advancements enable organizations to make timely decisions, uncover deeper insights, provide immersive data experiences, and ensure compliance with regulations, ultimately unlocking their data's full potential and driving innovation.

VI. CONCLUSION

The integration of ELT, MongoDB, and data visualization offers a robust solution to the complex data landscape faced by modern organizations. By centralizing data in a flexible and scalable database like MongoDB, organizations can efficiently manage diverse data sources, ensuring data consistency, accessibility, and quality. The transformation process within MongoDB empowers analysts to shape data into a format suitable for analysis, while data visualization tools bring these insights to life through interactive and engaging presentations. This holistic approach ultimately enables data-driven decision-making, fostering innovation and strategic growth.

REFERENCES

[1] Ayush Gupta , Dr. Namrata Dhanda,Kapil Kumar Gupta,“Ingest and Visualize CSV Files using AWSPlatform For Transition from Unstructured toStructured Data” 2023 11th International Conference on Emerging Trends in Engineering & Technology[2023]

[2] Irum Rauf, Marian Petre,Thein Tun, Tamara Lopez,”Security Thinking in Online Freelance Software Development”, IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society [2023]

[3] Lily Puspa Dewind, Michael Wongrd, Henry N. Palit, “ETL, ELT and Reverse ETL: A business case Study”,International Conference on Intelligent Cybernetics Technology & Applications [2023]

[4] Saranya N, Brindha R, Aishwariya N, Kokila R, Matheswaran P, Poongavi P, “Data Migration using ETL Workflow”International Conference on Advanced Computing & Communication Systems [2021]

[5] Munawar,"Extract Transform Loading (ETL) Based Data Quality for Data Warehouse Development” 1st International Conference on Computer Science and Artificial Intelligence [2021]

[6] Bin Pan, Guomin Zhang, Xuepei Qin,”Design and Realization of an ETL Method in Business Intelligence Project”IEEE International Conference on Cloud Computing and Big Data Analysis[2018]

[7] Ying Pei, Jungang Xu, Qiang Wang,”One CWM-based Data Transformation Method in ETL Process”,IEEE International Conference [2010]