Optimizing Execution Time in Big Data Warehouses

1st Lakshmi Deepthi Nukala
Assistant Professor
dept of Data Science
Institute of Aeronautical Engineering
Hyderabad, India
n.lakshmideepthi@iare.ac.in

2nd Gadipelly Varshitha
dept of Data Science
Institute of Aeronautical Engineering
Hyderabad, India
varshithagadipelly297@gmail.com

3rd Kalagani Vivek
dept of Data Science
Institute of Aeronautical Engineering
Hyderabad, India
vivekkalagani@gmail.com

4th Talari Srinidhi
Department of Data Science
Institute of Aeronautical Engineering
Hyderabad, India
srinidhitalari15@gmail.com

***Abstract*—Traditional data warehouses (DWs) have long been fundamental to business intelligence and decision support systems. However, with the exponential growth of data produced by modern applications, traditional data warehousing systems face challenges that require adaptation to new paradigms. In the context of big data, it's essential to modify current warehouse systems to address these emerging limitations. One major drawback of traditional Extract, Transform, Load (ETL) processes is their inability to efficiently handle massive volumes of data, especially when dealing with unstructured or semi-structured formats. The processing speed is also significantly hindered, leading to high execution times.To overcome these challenges, a new approach is proposed, which introduces a model based on four layers: Extract, Clean, Load, and Transform (ECLT). This model is specifically designed to optimize data processing and improve efficiency when managing large-scale and complex datasets. This aims to reduce execution time in ECLT and compare time between ETL and ECLT and adding analysis phase to ECLT named ECLTA (Extract Transform Load Transform Analysis) for queries and making analysis of the data.**

***Index Terms*—Big Data,ETL (Extract Transform Load) ,ECLT (Extract clean Load Transform),ECLTA (Extract Clean Load Transform Analysis).**

I. INTRODUCTION

In today's digital age, every online interaction whether it's through social media, financial transactions, sensor data, or digital communications contributes to a massive influx of data. Additionally, organizations accumulate extensive amounts of data, including customer records, sales information, and operational logs[1]. This has led to an unprecedented volume of data that needs to be processed and analyzed. A data warehouse functions as a central hub for organized, cleaned, processed, and stored data that has been consolidated from multiple sources, offering business intelligence users and decision-makers a comprehensive view of the information.[2].

Big data refers to large and complex datasets that traditional database management systems struggle to store and process efficiently[3]. Big data is often described by the 5Vs: volume (the size of the data), velocity (the speed at which data is generated and flows from source to destination), variety (the different formats and types of data), veracity (the accuracy and reliability of data), and value (the importance or usefulness of data prior to analysis). Over time, additional characteristics, such as volatility (the rate at which data changes) and visualization (how data can be represented), have been added to this list[2].Organizations today face the challenge of integrating big data into data warehouses to provide a comprehensive overview of their data. This integration is intricate and presents multiple challenges, particularly in constructing, storing, transforming, and analyzing data warehouses in a way that is scalable and efficient, all while ensuring minimal impact on end users. [5]. To address these difficulties, the concept of a data lake has emerged, offering a solution for managing and utilizing big data effectively. Researchers and developers are continuously working on ways to incorporate big data into traditional data warehousing systems. As the data is can't processed over the system warehouse is not needed here. We used database to store data. These data are processed by an Extract–Transform–Load (ETL) process[4]. We have other process like ELT, DETL etc for processing the data.They invlove more execution time.[6] The idea introduces a model based on four layers: Extract, Clean, Load, and Transform (ECLT). This model is specifically designed to optimize data processing and improve efficiency . This aims to reduce execution time in ECLT and compares time between ETL and ECLT and adding analysis phase to ECLT named ECLTA (Extract Transform Load Transform Analysis) for queries and making analysis of the data.

II. LITERATURE SURVEY

[1] The author explores the rapidly expanding landscape of big data and its implications for various industries. The paper highlights key statistics and trends, including market size, growth projections, and the increasing demand for data-driven solutions across sectors. It emphasizes the importance of big data in business strategies, decision-making, and technological advancements. Additionally, the study presents a detailed analysis of industry growth patterns, revealing how companies

leverage big data analytics for competitive advantage. The paper is an essential resource for understanding the current state and future potential of big data.

[2] The author explores the evolving functions of data lakes and data warehouses in managing enterprise data. They provide a comparison of the two systems, outlining their architectures, use cases, and benefits. The authors explain that while data warehouses are optimized for handling structured data and executing complex queries, data lakes are designed to accommodate various data types. Additionally, the paper looks into the integration of data lakes and warehouses to form a more comprehensive data management strategy.

[4] The authors provide a comprehensive review of data warehousing process models, tracing their evolution from traditional methodologies to modern trends. The paper analyzes key features such as data integration, storage, and retrieval, while comparing classical and emerging approaches. It highlights innovations like cloud-based data warehousing and the integration of AI technologies. The authors emphasize the growing need for flexible, scalable models to address the complexity of big data in contemporary enterprises. The paper also discusses the increasing complexity associated with big data and the corresponding need for adaptable and robust models. As enterprises face diverse data sources and growing volumes of information, the necessity for data warehousing solutions that can easily scale and evolve becomes paramount. The authors argue that the future of data warehousing will require a blend of traditional practices and innovative technologies .This enables organizations to effectively utilize their data for informed strategic decision-making.

[7] The authors provide an insightful overview of the evolution of ETL (Extract, Transform, Load) technology. The discussion covers the historical development of ETL, from its early use in data warehousing to its modern-day applications in big data and cloud environments. The talk also explores current trends such as automation, real-time ETL, and integration with AI and machine learning technologies. Looking toward the future, the authors speculate on advancements in ETL, emphasizing the need for adaptability in increasingly complex data ecosystems.

[10] The authors examine the challenges posed to ETL processes by the increasing variety of data formats and sources in modern computing. The study highlights how traditional ETL processes, designed primarily for structured data, are adapting to accommodate different types of data in big data environments. paper discusses new strategies and tools to handle diverse data types efficiently, emphasizing the importance of flexible and scalable ETL frameworks in heterogeneous data ecosystems. This work provides valuable insights into the evolution of ETL in response to the growing complexity of data.

III. EXISTING SYSTEM

An ETL (Extract, Transform, Load) model is crucial component in data warehousing, facilitating process of integrating data from various sources into a centralized data warehouse for unstructured data[7].

1. Extract: This is the first step of the ETL process, which involves retrieving data from multiple and possibly disparate sources. These sources can include databases, files, cloud storage, or APIs. The goal is to pull relevant data efficiently while minimizing the load on the source systems.

2. Transform: Once the data is extracted, it undergoes transformation to ensure consistency, quality with the data warehouse schema. In this phase, the extracted data is processed and transformed to fit the destination format and requirements.

This step ensures that data is compatible and useful for downstream systems like reporting tools or business analytics platforms.

3. Load: The final step is loading the transformed data into a target system, such as a data warehouse, database, or data lake. This process may involve full load (loading all the data at once) or incremental load (updating only the new or changed data). The goal is to store the data in an optimized format that enables fast querying and analysis.[9].

With ETL ,all models one after the other came into existence choosing the execution time and the memory allocation.

IV. METHODOLOGY

ECLT:

ECLT stands for Extract, Clean, Load, and Transform. This approach emphasizes cleaning as a distinct step that happens before loading or transforming the data, ensuring that only high-quality data gets stored and used in the final analysis.

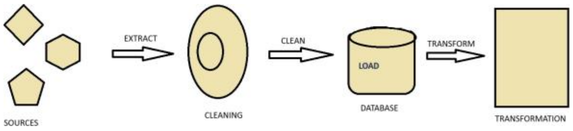

Fig. 1. ECLT

1. Extract:Collecting unprocessed data from multiple sources like databases, APIs, and spreadsheets., or cloud storage. The goal is to collect the data in its raw form without making any major changes.

2. Clean:After extraction, the data often contains errors, inconsistencies, or missing values. It focuses on improving the correctness of the data by removing duplicates, fixing errors, handling missing data, and ensuring it meets the required quality standards for analysis.

3. Load:Once the data is cleaned, it needs to be moved or loaded into the destination system, usually a database or data warehouse. This step ensures that the cleaned data is stored in a structured and accessible format for further use.

4. Transform:Finally, the transform step involves reshaping or modifying the data to suit specific business or analytical needs. This can include operations like data aggregation,

normalizing data, joining different datasets, or applying business rules. Here transformations are done within the database itself,which makes it more efficient.

ECLTA:

ECLTA stands for Extract, Clean, Load, Transform and Analysis.Here more Analysis phase is added which is used to querying and for the analysing the queries and output the graph according to the query asked and the remaining all the phases are same as in ECLT.

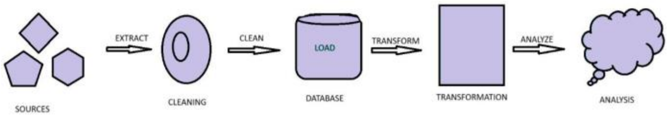

Fig. 2. ECLTA

Analysis Phase:After the completing of the four phases of ECLT, sql queries are written analysis phase and they generate graphs according to the query.This is just for the analysis, there is no comparsion of execution time with ETL and ECLT.

V. IMPLEMENTATION

ECLT model is specifically designed to optimize data processing and improve efficiency . It aims to reduce execution time in ECLT and compare time between ETL and ECLT and adding analysis phase to ECLT named ECLTA (Extract Transform Load Transform Analysis) for queries and making analysis of the data for given covid19 dataset consisting of confirmed cases, recovered covid cases,death covid cases files. To compare the execution time of ETL and ECLT , run the ETL first and then ECLT.Data should be selected first and extract it for both models. perform the ETL process ,next ECLT and ECLTA.After three completion of three models there will be comparision of ETL and ECLT Execution Time. In all this python libraries plays a major role like pandas,numpy,matplot.They are used for reading the csv files and for generating graphs.

ETL: The purpose of the extraction phase is to obtain data from three datasets using the read CSV and collect functions. (Covid19 confirmed, deaths, Recovered.csv) during full extraction.Transformation is done after the extraction like arranging the data in correct format and removing errors and duplicates.After transformation loading is done.All the transformed data is loaded into the database.

ECLT:

1. Extraction Phase :

The purpose of this is to extract data from three datasets or other sources . In this data is extracted from given dataset covid19 containing confirmed ,recovered and death cases. This is done through functions like read csv() and collect() to load the data from these files during the full extraction process. The extraction process aims to gather the relevant data while minimizing the impact on the source systems. In some cases, the extraction can be a full load, where all the data is pulled at once, typically during the initial setup.The sources can include databases, files, cloud storage, or APIs. Extraction may involve handling data that need to be aggregated and prepared for further processing.Upon completion of this phase the data is ready for the next phase of the process. The data present in the dataset like confirmed number of cases,death cases and transformed cases are extracted and kept ready for next phase.

2. Cleaning Phase :

In the Cleaning phase of the ECLT process, the goal is to ensure that the data is free from errors, inconsistencies, and irrelevant elements that could affect analysis. One of the main tasks is handling missing data, which may involve either filling in gaps with estimated values or removing incomplete records. Additionally, detecting and addressing outliers is critical to prevent skewed results; this can be done by transforming, capping, or removing extreme values that don’t align with the rest of the dataset. Another important aspect is the standardization of data formats, such as dates, units, or text casethat don’t align with the rest of the dataset. Another important aspect is the standardization of data formats, such as dates, units, or text case, to ensure consistency and compatibility across the dataset. Duplicate entries are also common when consolidating data from various sources, and these need to be identified and removed to avoid redundancy Similarly, ensuring data accuracy involves correcting Similarly, ensuring data accuracy

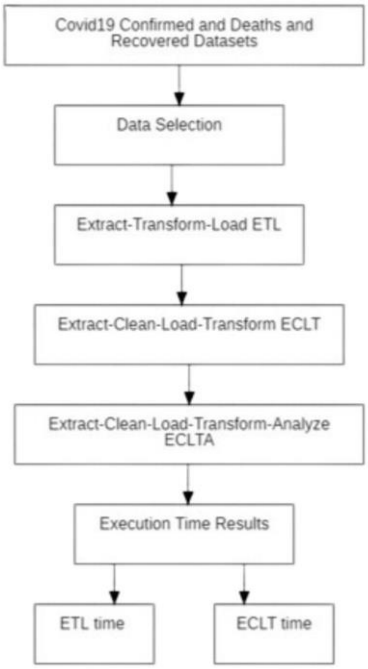
Covid19 Confirmed and Deaths and Recovered Datasets
Data Selection
Extract-Transform-Load ETL
Extract-Clean-Load-Transform ECLT
Extract-Clean-Load-Transform-Analyze ECLTA
Execution Time Results
ETL time
ECLT time

Fig. 3. Flow chart

involves correcting typos, spelling errors, and aligning the data to a uniform structure. In the used dataset some dates are missed. Cleaning phase helped to get the missed dates and missed information with NaN values. This cleaning phase is vital for maintaining data integrity, as it prepares the raw information for the subsequent steps of transformation and loading, ensuring that the final data is accurate, consistent, and reliable for decision-making.

3. Loading Phase :

Loading data into a database after the cleaning phase involves several important steps to ensure that the data is efficiently and accurately transferred from the cleaning layer to the storage system. Database Connection: Establish a secure connection to the target database. This connection should be configured with appropriate credentials and security measures such as encryption, to ensure that data is securely transmitted and stored.we used mysql as database. Schema Definition: Define the schema in the target database to match the structure of the cleaned data. This includes specifying the table names, column names, data types, and any constraints such as primary keys or unique constraints. Now the data is loaded into the database and we can see cleaned data there and the stored files in the database .This is crucial as ETL loading is the last step ,here transformations are done within the database.

4. Transformation Phase :

The primary goal of this is to restructure the data,so it can effectively support business intelligence (BI) or analytical tasks. Not every piece of extracted data needs to be transformed—only the data relevant to the specific analysis is processed. The choice of data for transformation is guided by the particular needs of the BI analysis. Transformations are often implemented as functions or operations that are evaluated lazily. This means they are deferred, creating an execution plan rather than running immediately,that is only triggered when an action, such as count(), is called. This lazy evaluation approach helps minimize the processing time needed for transformations. Once the transformation is complete, the result is immutable, ensuring data integrity. Applying actions to the transformed data helps reduce processing time, making the system more efficient and the data ready for further analysis. Here after the data is loaded in to the database then transformations are done which may help in improving the efficiency of the data and decreasing the execution time of the data.Here is corrected to correct format and counts the number of increased covid cases, recovered and death cases.

ECLTA:

Analysis Phase:After completing of four phases of ECLT, sql queries are written in the Analysis phase.They generate graphs according to the query.This is just for the analysis, there is no comparsion of execution time with ETL and ECLT.With given dataset one query is countries with most confirmed cases in march which gave as graph showing the countries.

Comparision of ETL and ECLT:

After running ETL and ECLT both execution time are displayed on the page.With help of execution times graph is generated for comparision.

VI. RESULTS

The idea focuses on comparing the execution times of ECLT and ETL models when applied to a given dataset. The ECLT model will show the less Execution time when compared to other given model,like shown in the picture .The execution time of ETL model is high and ECLT is less with the given dataset.The results demonstrate that ECLT consistently outperforms ETL in terms of speed, exhibiting significantly shorter execution times. This efficiency advantage of ECLT is particularly pronounced with larger and more complex datasets.In summary, the comparison underscores that ECLT not only enhances execution speed but also provides a more effective framework for managing complex data workflows.

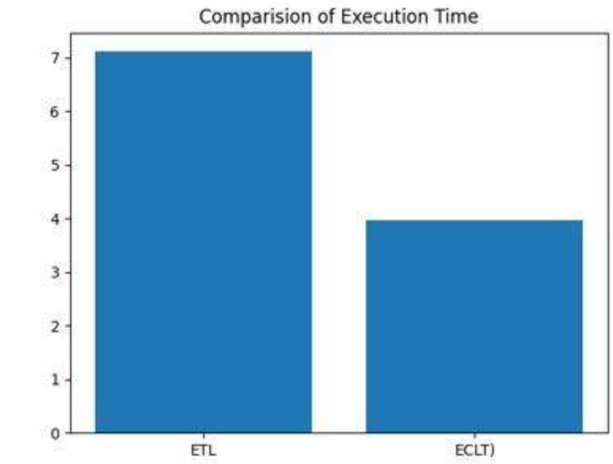

Fig. 4. Comparision of ETL and ECLT

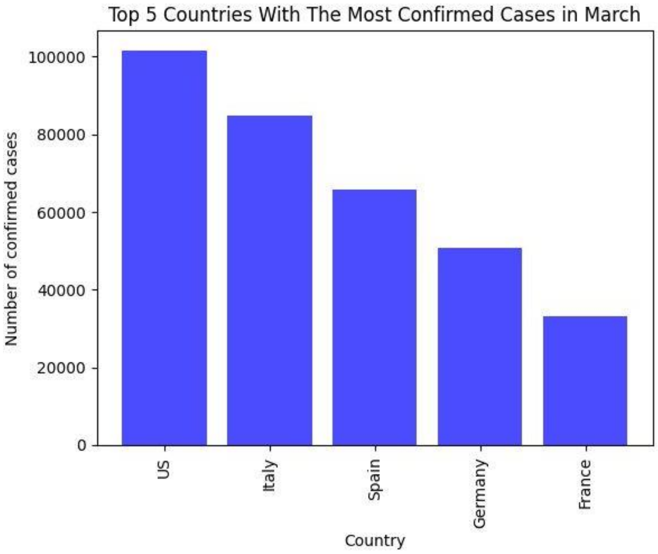

Fig. 5. Query graph

VII. Conclusion and Future enhancements

The proposed solution focuses on comparing the efficiency of the ECLT process against the traditional ETL (Extract, Transform, Load) model, with an emphasis on reducing execution time during data processing. The introduction of a variant named ECLTA further enhances this process, providing an queries and anlaysis to ECLT.The experimentation involved running queries on a chosen dataset, comparing execution times between ECLT and ETL processes. The results demonstrate that ECLT not only shortens the time required to load data but also performs better in data transformation, which is crucial for real-time or large-scale data applications. The graphs generated from this study visually depict the improved performance of ECLT, highlighting the time reduction achieved.In conclusion, the study provides substantial evidence that the ECLT model delivers superior performance in data processing tasks. The significant reduction in execution time positions it as a highly efficient tool, offering practitioners and researchers a faster and more scalable alternative to traditional ETL.

For future enhancements of the ECLT process, several improvements can be made to enhance its efficiency and scalability:

1. Parallel and Distributed Processing: As datasets grow larger, ECLT could be optimized to run across distributed systems, such as Hadoop to process data in parallel. This would allow the ECLT process to handle big data more effectively, distributing the workload across multiple nodes and reducing the time required for extraction, cleaning, loading, and transformations.

2.Adaptive Cleaning and Transformation Logic: Incorporating adaptive or AI-driven cleaning and transformation algorithms could further enhance the ECLT process. These enhancements would allow the system to automatically detect and adjust to different data patterns, anomalies, or quality issues, thereby improving the accuracy and reliability of the processed data without manual intervention.

References

[1] Wise, J. Big Data Statistics 2022: Facts, Market Size Industry Growth. Available online: https://earthweb.com/big-data statistics/ (accessed on 27 October 2022).

[2] Nambiar, A.; Mundra, D. AnOverview of Data Warehouse and Data Lake in Modern Enterprise Data Management. Big Data Cogn. Comput. 2022, 6, 132.

[3] Khine, P.P.; Wang, Z.S. Data lake: A new, ideology in big data era. ITM Web Conf. 2018, 17, 03025. [CrossRef].

[4] Dhaouadi, A.; Bousselmi, K.; Gammoudi, M.M.; Monnet, S.; Hammoudi, S. Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons. Data 2022, 7, 113.

[5] Martins, A., Martins, P., Caldeira, F., Sa´, F. (2020). An Evaluation of How Big-Data and Data Warehouses Improve Business Intelligence Decision Making. In: Rocha, A´ ., Adeli, H., Reis, L., Costanzo, S., Orovic, I., Moreira, F. (eds) Trends and Innovations in Information Systems and Technologies. WorldCIST 2020.

[6] Farhan, M.S.; Youssef, A.; Abdelhamid, L. A Model for Enhancing Unstructured Big Data Warehouse Execution Time. Big Data Cogn. Comput. 2024, 8, 17.

[7] Simitsis, A.; Skiadopoulos, S.; Vassiliadis, P. The History, Present, and Future of ETL Technology. Invited Talk. 2023. Available online: https://dblp.org/rec/conf/dolap/SimitsisSV23.html (accessed on 25 January 2024)

[8] Astriani, W.; Trisminingsih, R. Extraction, Transformation, and Loading (ETL) Module for Hotspot Spatial Data Warehouse Using Geokettle. Procedia Environ. Sci. 2016, 33, 626–634. [CrossRef].

[9] Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., Yahia, S. B. (2019). Data quality in ETL process: A preliminary study. Procedia Computer Science, 159, 676–687.

[10] Berkani, N., Bellatreche, L., Guittet, L. (2018). ETL Processes in the Era of Variety. Lecture Notes in Computer Science, 98–129.

[11] Santos, M. Y., Martinho, B., Costa, C. (2017). Modelling and implementing big data warehouses for decision support.

[12] Yang, Q.; Ge, M.; Helfert, M. Analysis of Data Warehouse Architectures: Modeling and Classification. In Proceedings of the 21st International Conference on Enterprise Information Systems, Heraklion, Greece, 3–5 May 2019; pp. 604–611.

[13] Vassiliadis, P., Vagena, Z., Skiadopoulos, S., Karayannidis, N., Sellis, T. (2001). Arktos: towards the modeling, design, control and execution of ETL processes. Information Systems, 26(8), 537–561.

[14] Astriani, W.; Trisminingsih, R. Extraction, Transformation, and Loading (ETL) Module for Hotspot Spatial Data Warehouse Using Geokettle. Procedia Environ. Sci. 2016, 33, 626–634.

[15] Zagan,E.; Danubianu, M. Data Lake Approaches: A Survey. In Proceedings of the 2020 International Conference on Development and Application Systems (DAS), Suceava, Romania, 21–23 May 2020; pp. 189–193.