



Dr. Vishwanath Karad

**MIT WORLD PEACE
UNIVERSITY** | PUNE

TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

**Capstone Project Report
on
“Standardization of ETL Process”**

Submitted by

Project Members

1032190030 Utkrist Agrawal

1032190041 Mehul Pansari

1032190048 Prabhat Panwar

1032190109 Nikunj Padia

Under the Internal Guidance of

Dr. Rashmi Phalnikar

Under the External Guidance of

Mr. Ajay Ghatpande

Mr. Ravi Khare

(Symphony Technologies)

**School of Computer Engineering and Technology
MIT World Peace University, Kothrud,
Pune 411 038, Maharashtra - India
2022-2023**



Dr. Vishwanath Karad
MIT WORLD PEACE
UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY

C E R T I F I C A T E

This is to certify that,
Utkrist Agrawal (1032190030)
Mehul Pansari (1032190041)
Prabhat Panwar (1032190048)
Nikunj Padia (1032190109)

of BTech. (Computer Science & Engineering) have completed their project titled **“Standardization of ETL Process”** and have submitted this Capstone Project Report towards fulfillment of the requirement for the Degree-Bachelor of Computer Science & Engineering (BTech-CSE) for the academic year 2022-2023.

[Dr Rashmi Phalnikar]

Project Guide

School of CET

MIT World Peace University, Pune

[Dr. Vrushali Kulkarni]

Program Head

School of CET

MIT World Peace University, Pune

Internal Examiner: _____

External Examiner: _____

Date:

Acknowledgement

We would like to express our heartfelt gratitude and appreciation to all the people who have supported and contributed to the successful completion of our B. Tech Capstone Project.

Firstly, we would like to thank our project guide **Dr Rashmi Phalnikar** for her continuous guidance, valuable insights, and unwavering support throughout the project. We are grateful to our external project guides **Mr. Ajay Ghatpande** and **Mr. Ravi Khare** from **Symphony Technologies** for their support throughout the course of the project. Without their guidance, it would not have been possible to complete this project successfully.

We would like to thank the faculty members of School of Computer Engineering and Technology for their guidance and support during the project period. Their constant encouragement and feedback helped us to improve our skills and knowledge. We would also like to thank our classmates and friends, who have been a great source of motivation and encouragement during this project. Their inputs and feedback helped us to improve the quality of our work.

We are deeply indebted to **Dr. Vrushali Kulkarni** (HOS, SCET), who provided us with valuable resources and materials that helped us to complete the project on time and all the facilities provided by MIT-WPU to help students excel in their career.

Student Name

Utkrist Agrawal (1032190030)

Mehul Pansari (1032190041)

Prabhat Panwar (1032190048)

Nikunj Padia (1032190109)

Abstract

The ETL (Extract, Transform, Load) standardization project aims to develop an integrated solution that enables efficient management, processing, and analysis of large volumes of data specific to company's data requirements. The project includes researching and developing an ETL application along with standardization module for interconversions of various data formats, leverages ETL tools and technologies to extract data from various sources, transform it into a format suitable for analysis, and load it into a data warehouse. The project also includes the development of a user-friendly data visualization tool that provides intuitive and interactive access to the data warehouse's contents. The project's objective is to enable the company to make informed decisions based on accurate and up-to-date data, leading to improved business performance and maintain data coming from various sources. The Proof of Concept and research is aimed at arriving at a all-in-one application similar to other Business Intelligence tools available in the market, but is lightweight and specific to organization's technological and data requirements. The project's scope covers the entire ETL, data warehousing, and visualization process, including requirements gathering, design, development, testing, and deployment. The project utilizes an Agile development methodology to ensure flexibility and adaptability to changing requirements and stakeholders' feedback. Overall, the ETL, visualization, and data warehousing project addresses the critical need for effective data management, processing, and visualization in today's data-driven business landscape.

Keywords- *ETL, Data Warehouse, standardization, Business Intelligence, visualization*

List of Figures

S. No.	Name	Page No.
1	The ETL Process	1
2	Architecture Block Diagram for ETL Standardization Application	12
3	Use Case Diagram	12
4	Activity Diagram	13
5	Class Diagram with well-defined relationships	13
6	Sequence Diagram	14
7	Component Diagram for ETL Standardization Application	14
8	Deployment Diagram	15
9	Wireframes for ETL standardization application	15
10	Project Timeline	16
11	Extraction Page	20
12	Transformation Page	21
13	Data Loading Page	21
14	Data Visualization Module	22

List of Tables

S. No.	Name	Page No.
1	Literature Review Summary	4
2	Project Planning and Summary	16
3	Application Performance Metric	23
4	Sample Test Case for Extraction module	24
5	Individual team member contribution	33

Table of Content

	Topic	Page No.
	<i>Abstract</i>	<i>II</i>
	<i>List of Diagrams</i>	<i>III</i>
	<i>List of Tables</i>	<i>III</i>
1	Introduction	1
	1.1 Project Statement	2
	1.2 Objectives	3
	1.3 Project Domain	3
2	Literature Survey	4
3	Problem Statement	7
	3.1 Project Scope	7
	3.2 Project Assumptions	7
	3.3 Project Limitations	8
4	Project Requirements	9
	4.1 Software and Hardware requirements	9
	4.2 Functional and Non-functional requirements	11
5	System Architecture and UML Diagrams	12
6	Project Plan and Timeline	16
7	Methodology	17
	7.1 Proof of Concept	17
	7.2 Technological Details	18
	7.3 Implementation Details	19
	7.4 Security Aspects	22
8	Performance Evaluation	24
	8.1 Performance Metric	24
	8.2 Testing performed	24

9	Results and Deployment		26
10	Applications of Project		27
11	Conclusion and Future Prospects		28
12	References		29
13	Publication Details		30
14	Appendix		31
	14.1	Base Paper	31
	14.2	Plagiarism Report	32
	14.3	Individual Contribution	33

Chapter I: Introduction

ETL stands for Extract, Transform, Load. It is a data integration process that involves extracting data from various sources, transforming it into a consistent format, and loading it into a target database or data warehouse for analysis.

The Extract phase involves pulling data from different sources such as databases, spreadsheets, or web APIs. The data can be stored in various formats, and it may be structured or unstructured. The Transform phase involves cleaning and processing the extracted data into a format that is suitable for analysis. This phase can include several steps such as data cleansing, normalization, aggregation, and filtering. The goal of this phase is to ensure that the data is in a consistent format and is suitable for analysis. The Load phase involves loading the transformed data into a target database or data warehouse. This phase may involve defining data schema, mapping source data to target data, and optimizing data load performance.

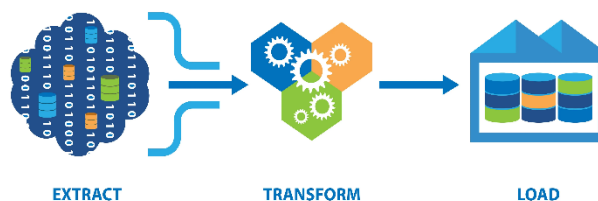


Fig 1. The ETL Process

Overall, the ETL process is critical for organizations that need to integrate data from multiple sources to gain valuable insights. By following a systematic approach to data extraction, transformation, and loading, organizations can ensure that their data is consistent and accurate, making it easier to analyze and draw insights from.

Data standardization is the process of transforming data into a consistent format that conforms to a set of predefined rules or specifications. This process involves identifying and resolving differences in data formats, values, and semantics to ensure that the data is accurate, reliable, and comparable. Standardization is crucial when dealing with data from multiple sources, as data may be stored in different formats, using different units of measurement, or with different codes or abbreviations. Without standardization, it can be challenging to compare data across different sources or to analyze data consistently. Data standardization typically involves several steps, including data cleansing, normalization, and formatting. Data cleansing involves removing or correcting invalid, incomplete, or inconsistent data. Normalization involves

converting data into a common format or structure, such as converting dates to a standardized format. Formatting involves ensuring that data is presented consistently, such as using the same units of measurement or abbreviations.

Overall, data standardization is critical for ensuring data consistency, accuracy, and comparability, which are essential for effective data analysis and decision-making. By standardizing data, organizations can make better-informed decisions.

Data visualization and reporting are crucial components of ETL (Extract, Transform, Load) processes. Once data has been extracted, transformed, and loaded into a database, it needs to be visualized and reported in a way that is easy to understand and analyze. Data visualization involves presenting data in a graphical or pictorial format, such as charts, graphs, or maps. The goal of data visualization is to make it easier for users to understand complex data and identify patterns, trends, or outliers. By using the right visualization and reporting tools, organizations can create a data-driven culture, improve data quality, and optimize their decision-making processes.

1.1 Project Statement

This project aims to build an Extract, Transform, and Load (ETL) application using ReactJS and Django, and perform standardization of given data file into user-specified data format. The application will be used to extract data from different sources, transform it, and load it into a target data store. The application enables the user to import data from any data source such as online or offline data files, process the data and convert it into standard format as per user requirement/format such as JSON, Excel, CSV, etc., all in one application for company's internal data requirements.

The project will have two parts: a front-end built using ReactJS and a back-end built using Django. The front-end will provide a user-friendly interface for users to configure and run ETL jobs. Users will be able to specify the source data, target data store, and any necessary transformations using a checkbox interface. The back-end will be responsible for executing the ETL jobs specified by the user. It will handle the extraction of data from various sources, perform the necessary transformations on the data, and load the transformed data into the target data store. The back-end will also provide an API for the front-end to communicate with it, and download the data files after all the operations are successfully completed, followed by visualization module to create interactive dashboards for uploaded data.

1.2 Objectives

1. To research and arrive at Proof of Concept for ETL standardization application in accordance with company's data requirements.
2. To perform technical assessment of various web technologies along with feasibility study of data warehouse for the given application.
3. To develop an application which will act as intermediary between different data formats and store standardized data in data warehouse.
4. To experience complete Software development Life cycle for the given project domain.
5. To enable the user to import data from any data source such as online or offline data files, process the data and convert it into standard format as per user requirement

1.3 Project Domain

The domain for creating an ETL application along with data standardization module followed by visualization application would be software engineering, web application development, data engineering and data analytics. Data analytics involves the process of examining and interpreting large data sets to derive insights and make informed decisions. In this project, the ETL application would be used to extract data from various sources, transform it, and load it into a target data store. The visualization application would then be used to analyze and present the data in a user-friendly format. The project would involve dealing with data in different formats and from various sources, performing data cleaning and transformation operations, and storing the transformed data in a database. It would also involve designing and implementing an interface to allow users to visualize the data and extract insights from it. Overall, the project would be focused on solving data engineering challenges and improving the efficiency and effectiveness of existing ETL applications by integrating all requirements of the user for integrated ETL standardization and visualization application.

Chapter II: Literature Survey

Paper name	Authors	Objective and Methodology	Research Gap	Conclusion
A Study of Extract–Transform– Load (ETL) Processes (2015)	S.Sajida, Dr.S. Ramakrishna	In Warehouse environment, ETL processes constitute the integration layer which aims to pull data from data sources to targets, via a set of transformations. By this work we intend to enrich the field of ETL processes, the backstage of data warehouse.	Standardizing models: no proposal becomes a standard, neither widely accepted by research community like multi-dimensional modelling in warehouse area.	This paper focused on ETL, the backstage of DW, and presents the research efforts and opportunities in connection with these processes.
Extraction Transformation and Loading (ETL) of Data Using ETL Tools (2022)	Manish Manoj Singh	This Paper discusses various ETL Tools Available in the Market. A huge piece of BI frameworks is a well-performing implementation of the ETL process, and focuses on the best ETL Tools and which tool can be the best for the ETL process.	1. Comparison of different ETL tools 2. ETL in the context of big data 3. Data governance and Machine Learning	ETL process plays the main role in Big data processing. Informatica PowerCenter is mostly the preferred tool used in data processing
Overview of ETL Tools and Talend-Data Integration (2021)	Sreemathy J, Brindha R, Selva Nagalakshmi M, Suvekha N, Karthick Ragul N.	BI leverages software and services to transform data into useful insights. In BI an ETL tool helps to extract the data from one or more sources, cleanse it and loads the data into data warehouse. In data integration techniques, the ETL method is important.	Performance comparison between Talend-Data Integration and other ETL tools. Automation of ETL workflows using Talend-Data Integration	We may assume that both Talend and Informatica are capable of executing the same shift and data integration tasks after evaluating all of their features.
Data Integration in ETL Using TALEND (2020)	Sreemathy J, Infant Joseph V, Nisha S, Chaaru Prabha I, Gokula Priya RM	The paper describes the various steps involved in integrating data from various sources using the ETL process, how the Talend Open Studio acting as a Data Integration and ETL tool helps in transforming heterogeneous data into homogeneous data for easy analysis and how all the integrated data is stored in a Data Warehouse	1. Performance comparison between Talend and other data integration tools 2. Data profiling and data quality assessment in Talend 3. Best practices for using Talend for data integration in ETL	The process of data integration is the main and the most important step in the process of integrating data from different sources. It makes the difficult process of analysing disparate data into a much easier process.
A UML Based Approach for Modeling ETL Processes in Data Warehouses (2003)	Sergio Luján-Mora, Juan Trujillo	The proposed approach involves using UML diagrams, such as activity diagrams, use case diagrams, and class diagrams, to model different aspects of ETL processes. The authors provide examples of how each type of diagram can be used to model different ETL process components, such as data extraction, transformation, and loading.	The paper does not discuss the use of other modeling languages or techniques. The paper assumes a certain level of knowledge and experience with UML modeling, which may not be the case for all stakeholders.	The paper provides a UML-based approach for modeling ETL processes in data warehouses. The approach is intended to provide a standardized and systematic approach for ETL process modeling.

A Survey of Real-Time Data Warehouse and ETL (2014)	Esmail Ali, F.S.	The objective is to discuss the role and importance of data warehousing in today's business landscape. The most popular data model for a DW is a multi-dimensional model, which can exist in the form of a star schema, a snowflake schema, or a fact constellation schema.	The trade-off between the overhead of providing real-time BI and the need for such analysis calls for serious research and consideration to avoid the resulting system having high costs.	The paper concludes by emphasizing the importance, complexity, and criticality of real-time BI and DW as a significant topic of research and practice.
Real-Time Data Warehouse Loading Methodology (2008)	Ricardo Jorge, Jorge Bernardino	The methodology is focused on four major areas: Data warehouse schema adaptation; ETL loading procedures; OLAP query adaptation; and DW database packing and reoptimization. The paper proposes a methodology of creating a replica of each table in the data warehouse, which is initially empty and has no constraints.	The method may not work well for data warehouse contexts where additive attributes are difficult or impossible to define for their fact tables. The text does not provide any information about the scalability of this method	In conclusion, the paper presents a methodology for supporting the implementation of Real-Time Data Warehousing by enabling continuous data integration.
An ETL Strategy for Real-Time Data Warehouse (2011)	Zhou, H., Yang, D., Xu, Y.	The paper explains the components of RTDW, including real-time behaviour and data warehousing, and highlights the importance of ETL in establishing and maintaining the data warehouse. The paper also discusses the challenges of capturing changed data in real-time and provides examples of message queues, database triggers, or streaming technologies.	The paper does not discuss the potential drawbacks or limitations of the real-time ETL process. The paper also does not compare the real-time approach with traditional batch processing.	The study has shown that the real-time ETL process provides accurate changing data loading and real-time data
JSON Integration in Relational Database Systems (2017)	Dušan Petković	The objective of the research paper is to explore the integration of JSON data format into relational database systems. The paper aims to investigate the challenges and benefits of incorporating JSON data into a RDBMS, such as MySQL, Oracle, etc.	There are some missing features in the current implementation of JSON in relational database systems.	The paper highlights the fact that different RDBMSs have implemented JSON in different ways.
Standardization of Storage and Retrieval of Semi-structured Thermophysical Data in JSON-documents Associated with the Ontology (2017)	A.O. Erkimbaev, V.Yu. Zitserman, G.A. Kobzev, A.V. Kosinov	The objective of this text is to highlight the challenges posed by the increasing volume and complexity of data on substances and materials properties, and to propose a set of solutions based on Big Data technology that can help to integrate diverse resources belonging to different organizations and states.	Overall, while the paper does provide an overview of the proposed technology and its potential benefits, there are several areas where it could be improved by providing more context.	A new technology for data management of complex and irregular structures has been proposed., specifically for the representation of thermophysical properties of substances
Batch to Real-Time: Incremental Data Collection & Analytics Platform (2017)	Ahmet Arif Aydin, Kenneth M. Anderson	The paper is designed to allow continuous data processing, allowing data to be analysed in real-time as it arrives, rather than being processed in batches at predetermined intervals. The platform consists of three main components: a data collector, a data transformer, and a data analyser. These components work together to collect data from a variety of sources, transform the data to a format.	The paper could be the need for more effective data processing and analysis systems that are capable of handling real-time data in dynamic and constantly changing environments.	The proposed platform represents a significant improvement over traditional batch processing systems, particularly in environments where data needs to be processed quickly and continuously.

Modelling ETL Processes of Data Warehouses with UML Activity Diagrams (2008)	Lilia Muñoz, Jose-Norberto Mazón, Jesús Pardillo & Juan Trujillo	The paper presents a case study of the proposed methodology applied to a real-world data warehouse. The authors use UML Activity Diagrams to model the ETL process of the data warehouse, including data extraction, transformation, and loading.	The research gap addressed by the paper is the need for more effective and accessible techniques for modeling ETL processes in data warehouse.	UML Activity Diagrams can provide an effective way to model ETL processes in data warehouses, offering advantages such as flexibility, intuitiveness.
Entity mapping diagram for automated extraction, transformation, and loading processes in data warehousing (2012)	Abdeltawab M.A. Hendawi and Shaker H. Ali El-Sappagh	The paper presents a case study of the EMD methodology applied to a real-world data warehousing scenario. It introduces the EMD methodology, which is based on a graphical notation for representing ETL processes using entities, attributes, and relationships. Demonstrates how EMD can be used to automate the ETL process.	The lack of a flexible and intuitive approach to automate the ETL processes in data warehousing. The authors propose the EMD methodology as a solution to address this gap.	The EMD methodology can help address the challenges associated with automating ETL processes in DW, providing a more flexible and intuitive approach.
Conceptual data warehouse modeling (2023)	Panos Vassiliadis, Alkis Simitsis and Sipros Skiadopoulos	Conceptual schema design: The authors describe the process of designing a conceptual schema that includes facts, dimensions, and hierarchies. They also provide guidelines for selecting appropriate levels of granularity and for identifying and resolving conflicts between dimensions.	Lack of discussion on new or emerging data warehouse technologies. The paper does not discuss how the proposed conceptual model can be adapted or applied to these new or emerging technologies.	The paper highlights the benefits of the proposed conceptual model, such as improved data quality, increased flexibility, and reduced development time.
Research on Extract, Transform and Load (ETL) in Land and Resources Star Schema Data Warehouse (2013)	Qin, Hanlin, Jin, Xianzhen; Zhang, Xianrong	The paper provides an overview of the Land and Resources Star Schema data model and the requirements for the ETL process. The paper discusses the challenges associated with the ETL process, such as data consistency, data accuracy, and data security.	The paper does not discuss the selection of ETL tools or frameworks for the implementation of the ETL process, assumes a certain level of domain knowledge and expertise in the design of DW.	In conclusion, the paper provides a detailed description of the ETL process used in a Land and Resources Star Schema data warehouse.
Conceptual Design of Data Warehouses from E/R Schemes (2002)	Matteo Golfarelli, Dario Maio, Stefano Rizzi	The objective of the paper is to propose a graphical conceptual model called the Dimensional Fact model, and a semi-automated methodology to build it from pre-existing ER schemes or RDBMS, for designing data warehouse (DW) systems.	The gap is the lack of a well-defined and understandable conceptual model for data warehouse design, particularly one that can be derived from E/R documentation.	The paper proposes a conceptual model and a semi-automated methodology for designing data warehouses.
A proposed model for data warehouse ETL processes (2011)	Shaker H. Ali, Abdeltawab M. Ahmed Hendawi, Ali Hamed El Bastawissy	The objective of the paper is to address the lack of a standard model for representing ETL scenarios and to explore the efforts that have been made to conceptualize ETL processes. The paper also highlights the importance of ETL processes in building a data warehouse.	The paper proposes a framework for using the EMD model and suggests future work to develop a prototype tool called EMD Builder.	The paper addresses the need for a standard conceptual model for representing ETL processes in data warehousing projects.

Table 1. Literature Review Summary

Chapter III: Problem Statement

3.1 Project Scope

The scope of creating an ETL application and a visualization application can be defined as follows:

1. **ETL Application:** The ETL application will be responsible for extracting data from various sources, transforming it, and loading it into a target data store. The scope of the ETL application includes identification and integration of data sources, i.e. The application will need to be able to connect to various data sources such as databases, APIs, and file systems, and integrate the data into a unified format, data transformation to perform various operations such as cleaning, filtering, and aggregating the data to ensure that it is consistent and accurate and finally data loading into a data store such as a database or data warehouse.
2. **Data Standardization:** The ETL standardization application would allow the user to convert the data file into specified data format such as JSON, CSV, Excel through automated ETL pipeline, and download the newly formatted file into local machine.
3. **Visualization Module:** The visualization application will be responsible for presenting the transformed data in a user-friendly format. The scope of the visualization application includes data exploration using various interactive visualizations such as charts, graphs, and tables, and customization of the visualizations to select the data they want to display.

The scope of the project includes the integration of various functionalities such as live data connection using APIs and importing large amount of data using different techniques. The project will involve selecting and implementing appropriate technologies as well as libraries for the ETL and visualization applications, designing and implementing user interfaces, and testing and debugging the applications.

3.2 Project Assumptions

Assumptions for the ETL (Extract, Transform, Load) standardization application project may include:

1. **Data source availability:** Assumption that the necessary data sources will be available for extraction and that the data will be of sufficient quality.

2. Data consistency: Assumption that the data being extracted will be consistent and in a format that can be transformed and loaded into the target visualization system.
3. Data volume: Assumption that the volume of data to be processed will be within the capabilities of the ETL standardization application and the target visualization system.
4. Data mapping: Assumption that the data mapping process will be properly executed, ensuring that data is correctly transformed and loaded into the target database/ data warehouse.
5. Visualization tool functionality: Assumption that the chosen visualization tool will be capable of handling the specific data sources and target systems required for the project.
6. Transformation rules: Assumption that the transformation rules have been properly defined and will be implemented correctly in the ETL application to generate meaningful insights.
7. Standardization module performance: We have to make the assumption that the application will perform within the required time frames and will be able to handle any potential bottlenecks or issues that may arise during the process.
8. Data security: Assumption that the data being processed will be properly secured during the ETL process and visualization and session is maintained during user operations to maintain data integrity and security.
9. System integration: Assumption that the ETL application will integrate seamlessly with the target visualization system and any other necessary applications.
10. Project timeline: Assumption that the project timeline will be realistic and achievable, allowing for any potential delays or unforeseen issues that may arise.

3.3 Project Limitations

The ETL standardization application is bounded by several limitations for the current scope of the project which includes software and hardware limitations, data integrity and security issues, dynamic requirements of the organization and time constraints. The limitations of the project are as follows:

1. **Data quality:** The accuracy and completeness of the data are critical for any data-driven project. Poor data quality can lead to incorrect analysis, which can undermine the value of the ETL and visualization project.
2. **Technical and integration issues:** The ETL and visualization application must be capable of handling the data volume and complexity of the target system. Inadequate hardware, software or bandwidth limitations may hinder performance. Integrating the ETL and visualization application with the target system may require significant coordination and management. Integration issues can arise from the different data models, application programming interfaces (APIs), or databases used by the target system.
3. **Cost constraints:** Implementing ETL and visualization application project can be expensive, requiring investments in hardware, software, including well-known data warehouses. Cost constraints may impact the scope, quality or timeline of the project.
4. **Data privacy:** The application may need to conform to specific privacy requirements to ensure that sensitive data is handled securely, and with the dynamic cyber security domain, it's difficult to ensure utmost security and privacy concepts in the initial phases of the application development.
5. **Changing requirements:** Changing project requirements can lead to delays and rework. Changes in the business requirements, data sources or visualization requirements can cause significant changes to the ETL standardization application. ETL standardization application projects need to be executed within specified time frames. Delays in data extraction or transformation, unexpected issues or changing requirements can cause significant schedule delays.

Chapter IV: Project Requirements

4.1 Software and Hardware requirements

4.1.1 Software Requirements

ETL (Extract, Transform, Load) visualization applications are designed to help users visualize and analyze data during the ETL process. These applications typically require the following software requirements:

1. **Programming Languages:** The given ETL standardization application requires knowledge of programming languages like SQL, Python, and Frontend programming language like ReactJS to manipulate data, write scripts, or develop custom solutions. The project is built using Django as backend language and ReactJS and its various libraries as Frontend programming language.
2. **Database Management Systems:** ETL standardization application require a database management system (DBMS) to store and manage data. Popular DBMSs include Oracle, SQL Server, MySQL, and PostgreSQL.
3. **Business Intelligence Tools:** ETL visualization module may require business intelligence tools to generate reports and visualizations based on the extracted data. Popular business intelligence tools include Tableau, QlikView, and Power BI. For our application, we have built a separate visualization tool using ReactJS, thus covering the need of using any third-party BI tool.
4. **Data Warehousing Tools:** ETL visualization applications may require data warehousing tools to create and manage a data warehouse. Popular data warehousing tools include Snowflake, Amazon Redshift, and Google BigQuery. For the current scope of the project, we have used MySQL database to load the data and use it for visualization purpose. The application may require data quality tools to ensure the data being extracted, transformed, and loaded is accurate and consistent. Popular data quality tools include Talend Data Quality, Informatica Data Quality, and IBM InfoSphere QualityStage.

4.1.2 Hardware Requirements

The hardware requirements for an ETL visualization application project will depend on several factors, such as the size and complexity of the data being managed, the number of users accessing the application, and the performance requirements of the application. Here are some general hardware requirements to consider:

1. ETL processes can be CPU-intensive, so it is recommended to have a multi-core CPU with a clock speed of at least 2 GHz or higher. The amount of RAM required will depend on the size of the data being managed and the number of users accessing the application. A minimum of 8GB of RAM is recommended, but larger datasets may require 16GB or more.

2. The storage requirements will depend on the size of the data being managed. Consider using fast SSD drives for improved read/write performance. Implementing a backup and recovery strategy to ensure data is protected against hardware failures or other disasters is also an important requirement.
3. A fast and reliable network connection is essential for efficient data transfer between systems, performing effective ETL process and load data quickly into the warehouse. Also, If the ETL visualization module may include complex visualizations or 3D graphics, a dedicated graphics card with at least 2GB of VRAM would be required.

4.2 Functional and Non-functional requirements

4.2.1 Functional Requirements

- The ETL standardization application uses Django as the backend programming language, ReactJS as the front-end language, Microsoft SQL for data loading and SQLite for application data storage.
- The application will enable users to select the data source of their choice, extract it, and then transform it into a format that can be easily loaded and analysed. The definition of the source connection will be stored and saved for later use. The transformed data will also need to be stored in the database.
- The React front-end will be used for interacting with the data and performing various analysis and visualization tasks using ReactJS chart APIs, while the Django back-end handles the data processing and storage.

4.2.2 Non-functional Requirements

- The application allows users to easily manage and organize their data, making it a valuable tool for company that need to extract insights from any amount of data, and perform basic validation on input data sources.
- The application needs to be scalable, secure and able to handle large volumes of data of any available format. The data can be imported in two ways: Batch-import or Real-time stream import from online/offline sources.
- The application must be interactive, secure and suitable for company's data requirements and flexible enough to adjust to changing requirements in future.

Chapter V: System Architecture

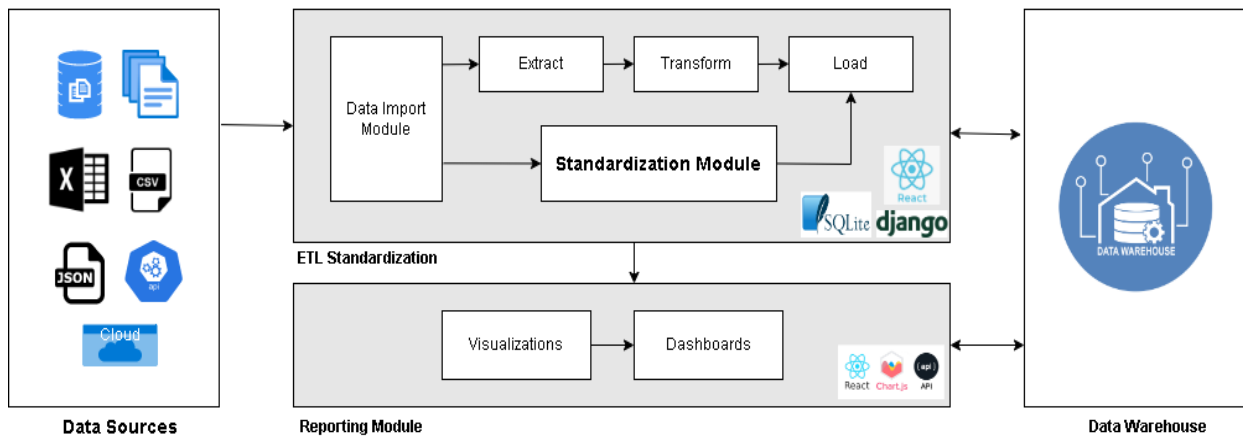


Fig 2. Architecture Block Diagram for ETL Standardization Application

The block diagram for the ETL standardization application Fig. 1 includes various components working in sequence corresponding to Data sources (include databases, files, APIs, web services, and other sources of data. Data sources may be located on-premises or in the cloud), Extract component(extracts data from the various data sources and prepares it for transformation), Transform component(it applies business rules, data cleaning, data enrichment, and other data processing operations to the extracted data), Load component(loads the transformed data into the target system, such as a data warehouse, data lake, or other analytical system)and the Reporting module(provides a user interface for users to interact with the data and generate reports, dashboards, and other visualizations) .



Fig 3. Use Case Diagram

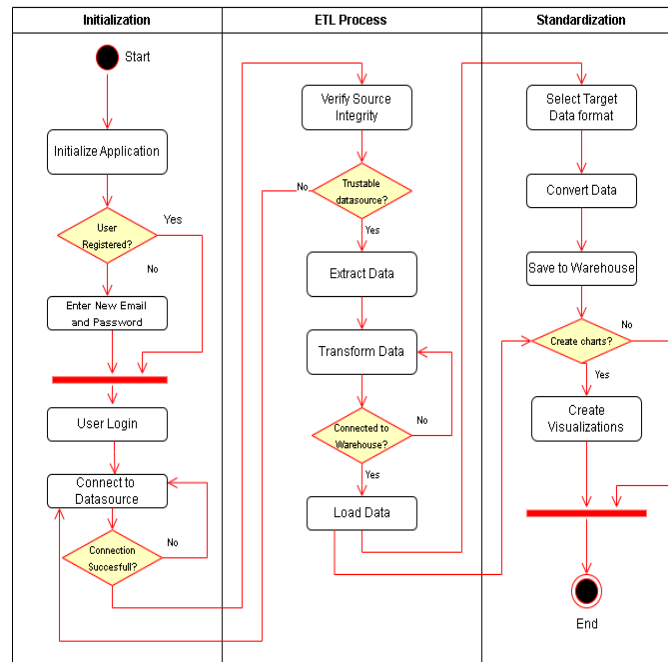


Fig 4. Activity Diagram

The Use case diagram in Fig. 2 depicts user, admin and data warehouse manager as actors, and various use cases such as extracting data from various sources, transforming data to meet business rules, loading data into target systems, creating reports, visualizations, managing validations on data, and data standardization. In Fig. 3, the activity diagram depicts the flow of activities and actions involved in the ETL process.

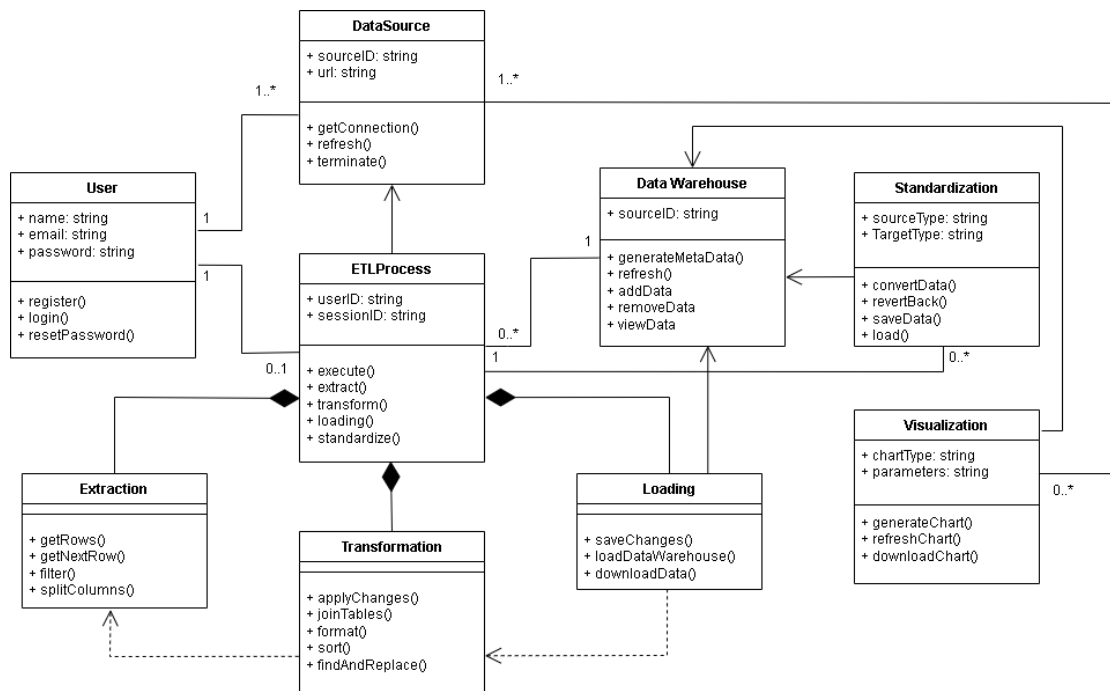


Fig 5. Class Diagram with well-defined relationships

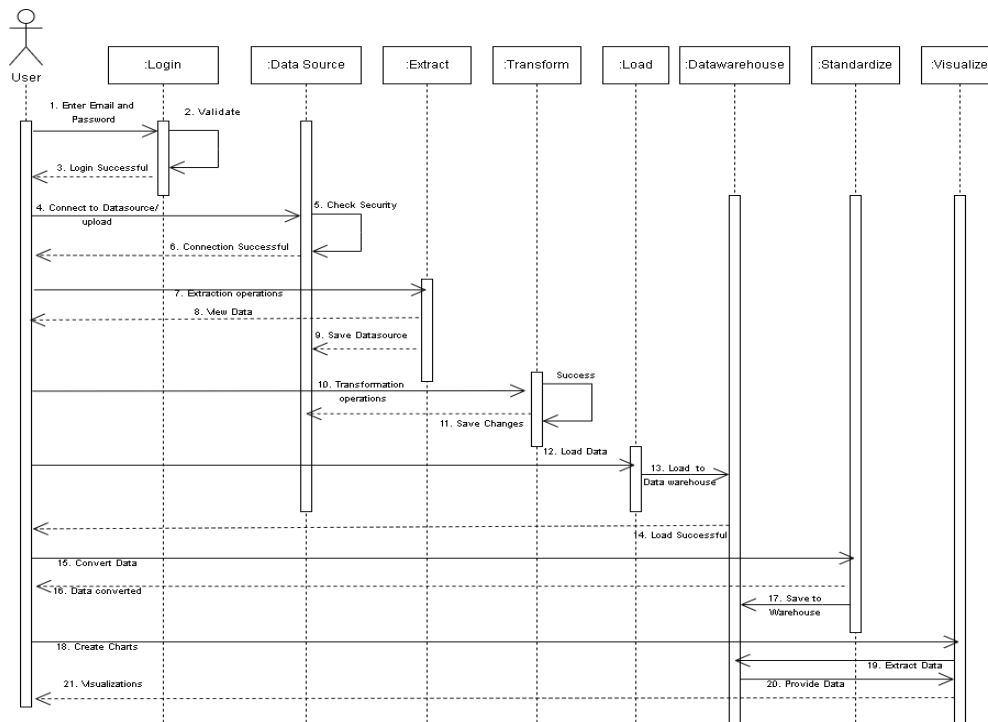


Fig 6. Sequence Diagram

Class diagram for an ETL standardization application (Fig.4) typically shows the various classes (User, ETL process, Extract, Transform, Load, Visualization, etc.) and their relationships involved in the application, while Fig. 5 shows the sequence of interactions between objects involved and flow of data in the ETL and standardization process.

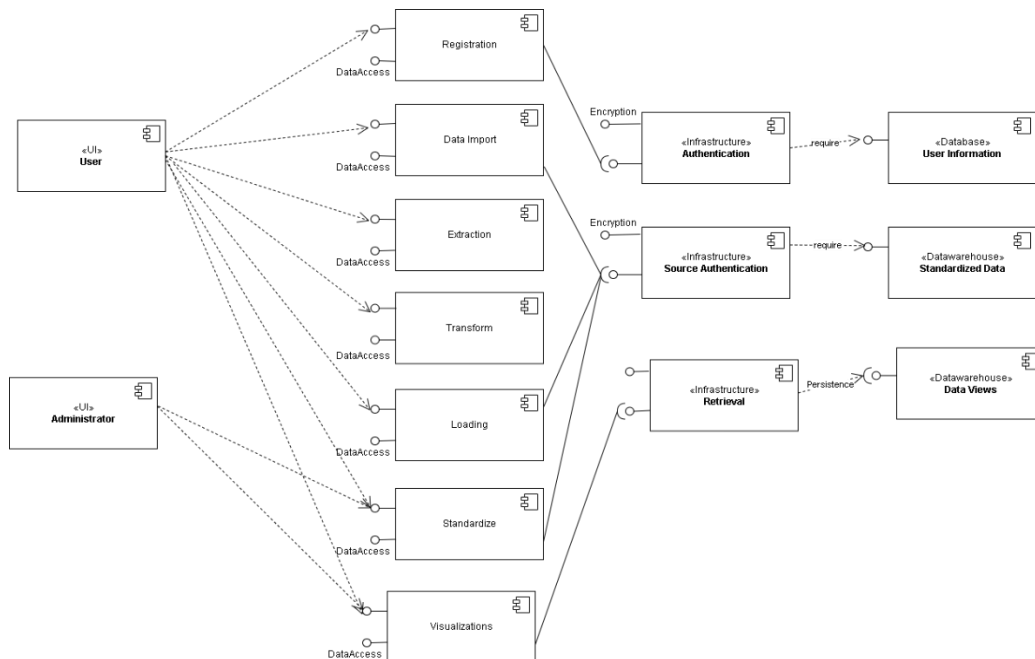


Fig 7. Component Diagram for ETL Standardization Application

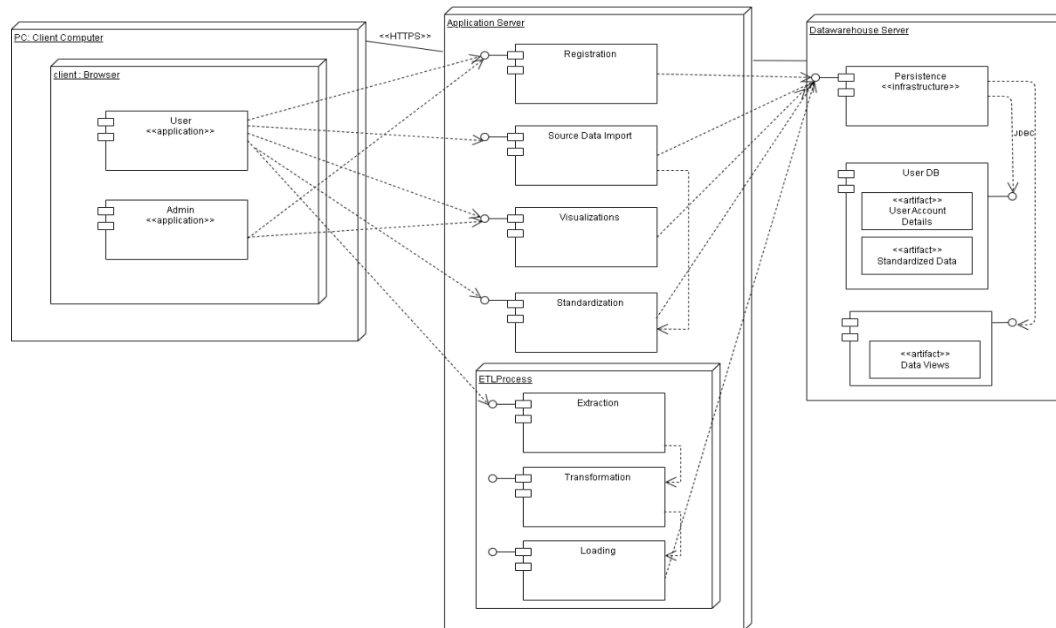


Fig 8. Deployment Diagram

The components of the system and deployment strategy are being shown the Fig. 6 and Fig.7 respectively for the ETL application.

Well before beginning the application development phase, we have worked upon the UI/UX development part for our ETL application using Balsamiq wireframing tool. Following are some example wireframes developed as part of UI framework development.

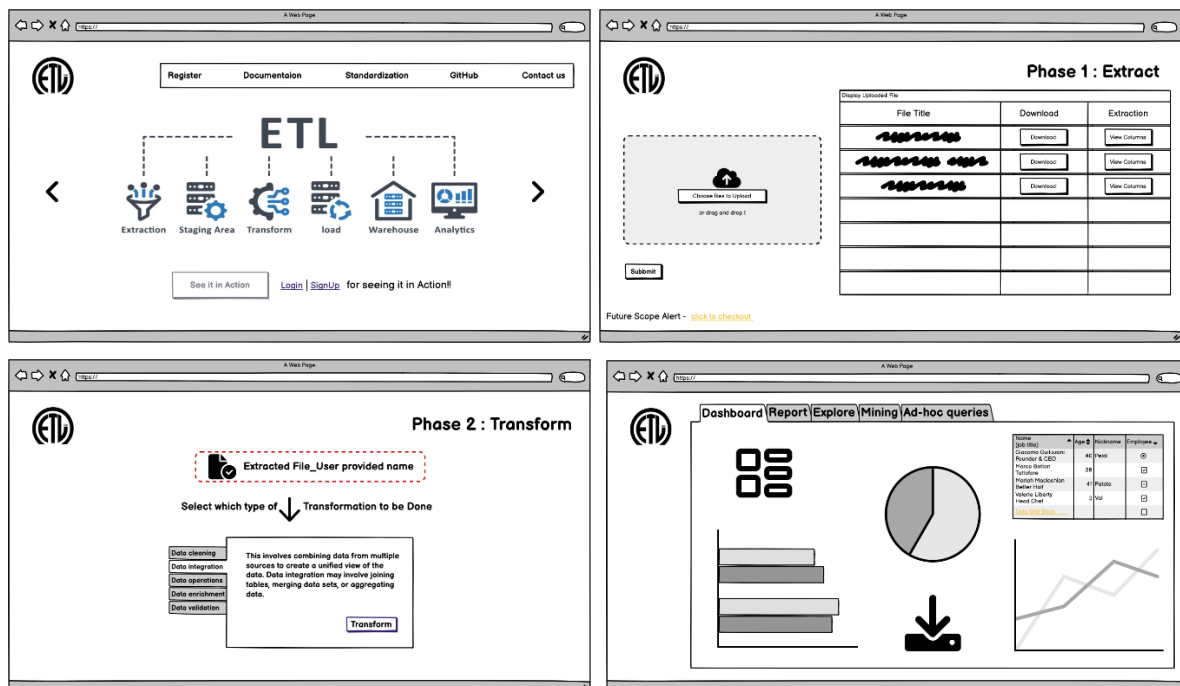


Fig 9. Wireframes for ETL standardization application

Chapter VI: Project Plan and Timeline

The project has been divided into several modules and assigned to different team members based on skillset and areas of interest, and these modules includes Visualization Module, Application Frontend and Backend development, UI/UX research, Standardization module, ETL pipeline module and data loading setup. Each member did literature surveys based on individual objectives and collaboratively worked upon the application Proof of Concept and development using tools like GitHub, JIRA, etc. The following is the timeline of the project planning from JIRA board roadmap application.

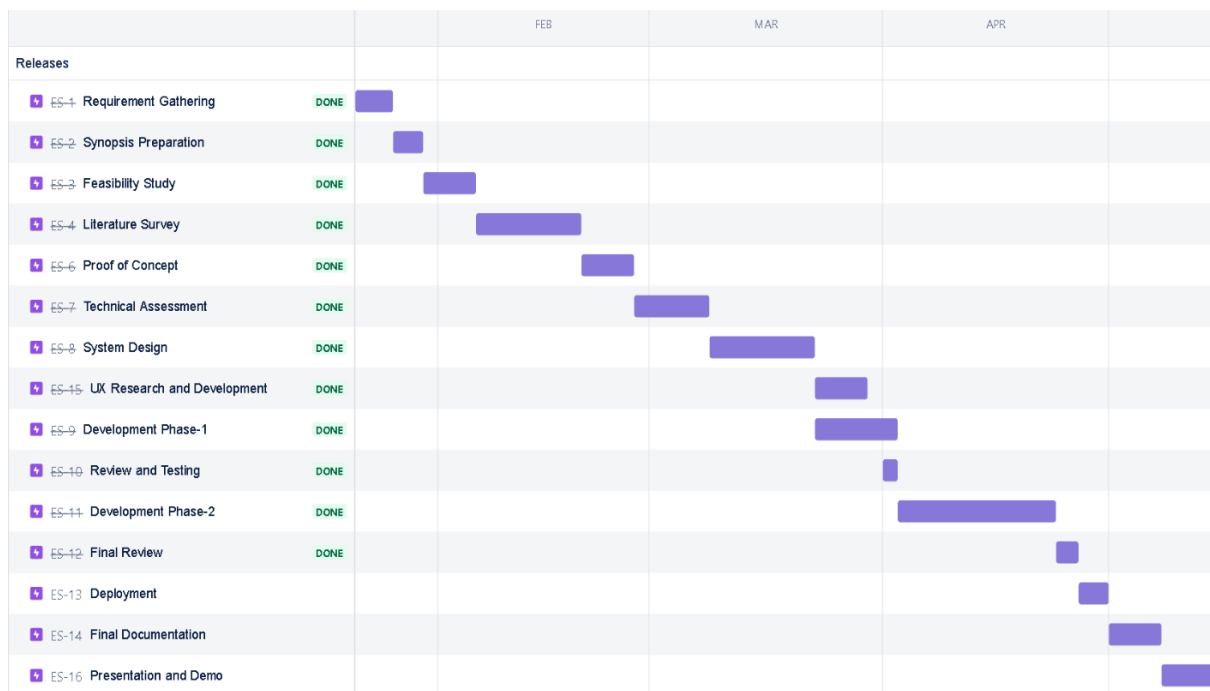


Fig 10. Project Timeline

Phase	Date	Tasks Performed
Planning and Requirements Gathering	23-01-2023 to 05-02-2023	Define project scope and objectives, collect requirements from stakeholders and document them, develop project schedule and planning, Define project deliverables and acceptance criteria
Research and Literature Survey	06-02-2023 to 08-03-2023	Performed literature research and Proof of Concept finalization for the ETL standardization application according the company's data requirements.
Design and Development	09-03-2023 to 23-04-2023	Designing the ETL standardization layout and user interface, coding the frontend and backend along with integrating in GitHub, developing visualization and loading module in data warehouse.
Testing and Deployment	24-04-2023 to 07-05-2023	Conduct system integration testing and acceptance testing, fix any defects found during testing and deploy the application on cloud/on-premise systems

Table 2. Project Planning and Summary

Chapter VII: Methodology

7.1 Proof of Concept

After finalizing the user requirements and carving a detailed project plan around the module design and development, the following Proof of Concept (POC) was formulated to support the development of the application before beginning to code for the same, so as to provide feasibility study and refine the concepts to be used in development life cycle of the ETL project.

Project Name: Standardization of ETL Process

Objective: To demonstrate the feasibility of the ETL Standardization and visualization project by building a proof of concept.

Assumptions: The source data is available in a structured format and the required hardware and software infrastructure are available.

Scope:

- Extract sample data from a source system.
- Transform the data using the ETL tool to meet the target system's data model requirements.
- Load the transformed data into a target system (preferably data warehouse) and standardize the given data to different formats.
- Visualize the transformed data using the reporting module.

Deliverables:

- A report on the proof of concept, including its objectives, methodology, and results.
- A demonstration of the proof of concept, including the ETL process and visualization output, followed by application development and deployment.

Success Criteria:

- The source data is successfully extracted from local/online environment.
- The data is transformed to meet the target system's requirements and transformed data is loaded into the target system without errors.
- The visualization module generates meaningful and useful insights from the transformed data, and standardization module successfully downloads data of new format.
- The POC report demonstrates the feasibility of the project.

Methodology:

- Identify the source system from which data is to be extracted.
- Develop an ETL pipeline to transform the source data into a format suitable for loading and analysis.
- Identify the target system (data warehouse/database) and load the transformed data into it, followed by data format conversion as per user selection.
- Select a data visualization method like specific chart to present meaningful insights from the transformed data.

Results:

- Data was extracted from the source system using the ETL tool.
- The ETL tool transformed the data into the target system's data model requirements.
- The transformed data was loaded into the target system without errors, and standardized data file downloaded successfully into local system.
- A data visualization tool was selected to generate meaningful insights from the transformed data.

7.2 Technological Details

The discussed ETL application is being developed using specific programming languages and their libraries based on the requirements setup by the stakeholders of the application and technological feasibility study conducted to ascertain the best performance and useability of the application in the business domain, and efforts are being made to make the application most cost effective and scalable. The following tools and technologies are being used in development of the ETL application: -

- **Django:** Django is a popular web framework written in Python that can be used for building ETL application. Django comes with built-in support for interacting with various data sources and the Django framework can be used in conjunction with libraries such as Pandas, NumPy, and SciPy to perform various data transformation operations. Django comes with built-in support for various databases, such as PostgreSQL, MySQL, SQLite, and Oracle. Developers can also use Django ORM (Object-Relational Mapping) to interact with databases and load transformed data.

Overall, Django can be a good choice for building ETL applications due to its ease of use, flexibility, and scalability.

- **ReactJS:** ReactJS is a popular JavaScript library for building user interfaces. It can be used in conjunction with other technologies to build ETL and data visualization applications. ReactJS provides a rich set of UI components that can be easily customized and combined to build rich and responsive user interfaces, and it can be used in conjunction with data visualization libraries such as D3.js and Chart.js to create interactive data visualizations.
- **MS SQL Server:** It is a popular relational database management system that can be used for data loading in ETL processes. MS SQL Server is designed to be scalable and can handle large amounts of data. It can be used to build ETL pipelines that can handle high volumes of data, and it provides built-in support for encryption, user authentication, and authorization.
- **SQLite:** SQLite is a lightweight and popular database management system that is widely used in web development projects. It's also one of the default database engines supported by Django, which is a popular Python-based web framework.
- **APIs: Chart.js** is a popular open-source library for creating charts and graphs using JavaScript. It's easy to use and highly customizable, making it a great choice for data visualization in ReactJS applications.
- **Balsamiq:** Balsamiq is a rapid wireframing tool that allows users to create simple and effective wireframes for software applications, webpages, and mobile apps. It is a popular tool used by designers, developers, product managers, and project stakeholders to quickly sketch out ideas and concepts for their applications.

7.3 Implementation Details

The implementation details of an ETL (Extract, Transform, Load), visualization, and data standardization application may have several pre-requisites to be considered. Firstly, to create front-end of the application, we have created UI/UX prototype of the application using the Balsamiq wireframing tool available online. After creating wireframes, the work on frontend of the web application started using ReactJS along with developing the backend of the application for ETL pipeline and data standardization using Django, keeping in mind the UML diagrams and workflow of the user's interaction while using the final product. Followed by this, the data loading module was developed and tested on the MS SQL server for various data

sources of varying formats, and then the visualization module was created using Chart.js API for ReactJS, allowing the user to create interactive visualizations for the loaded data and saving them for future use. Here are some common implementation details that can be considered for the project:

1. ETL Implementation:

- Design the ETL workflow: Plan and design the ETL workflow based on the data sources, data types, and data volumes.
- Extract phase: The Extract phase of the ETL (Extract, Transform, Load) process involves retrieving data from various sources and bringing it into a central location, such as a database or data warehouse, for further processing. The Extract phase is crucial to the ETL process because it sets the foundation for the rest of the data processing. The first step in the Extract phase is to identify the data sources that need to be extracted. These sources could be databases, files (CSV, Excel, JSON), APIs, or web services. Once the data has been extracted, it is important to validate and cleanse it to ensure its accuracy and consistency. This involves checking for errors, duplicates, and missing data, and correcting any issues that are identified and finally setting the stage for the Transform and Load phases of the ETL process

The screenshot shows the 'EXTRACTION' page with a 'Successfully Extracted' status. The form includes a 'Select All' checkbox and three columns of checkboxes for selecting fields: Series_reference, Suppressed, Magnitude, Series_title_1, Series_title_4, Period, STATUS, Subject, Series_title_2, Data_value, UNITS, Group, Series_title_3, and Series_title_5. Below the checkboxes are 'Rows Count' (set to 20) and 'File Name' (myFirstExtract) fields, and an 'Extract Data from Files' button. A table at the bottom displays the extracted data with columns: Series..., Period, Data..., Suppr..., STAT..., UNITS, Magni..., Subject, Group, Series..., Series..., Series..., Series..., and Series....

Series...	Period	Data...	Suppr...	STAT...	UNITS	Magni...	Subject	Group	Series...	Series...	Series...	Series...	Series...
BDCQ...	2011...	80078		F	Numb...	0	Busin...	Indust...	Filled ..	Agric...	Actual		
BDCQ...	2011...	78324		F	Numb...	0	Busin...	Indust...	Filled ..	Agric...	Actual		
BDCQ...	2011...	85850		F	Numb...	0	Busin...	Indust...	Filled ..	Agric...	Actual		
BDCQ...	2012...	90743		F	Numb...	0	Busin...	Indust...	Filled ..	Agric...	Actual		

Fig 11. Extraction Page

- Transform phase: The Transform phase of the ETL (Extract, Transform, Load) process involves cleaning, enriching, and reshaping the extracted data into a format that can be easily loaded into the target system. Some basic transform operation performed by our application includes selecting specific columns for consideration, removing duplicate values, sorting the data, handling the missing values by statistical method, and case conversion of specific columns.

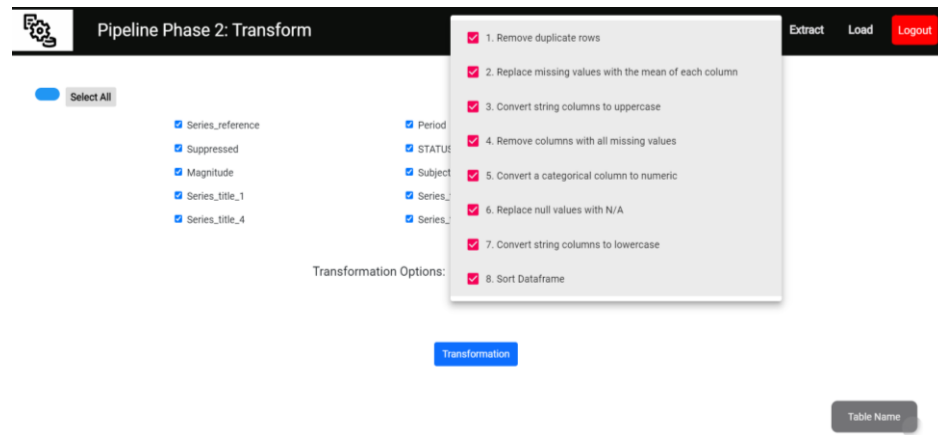


Fig 12. Transformation Page

- Loading phase: The Load phase of the ETL process involves loading the transformed data into the target system, here it is MS SQL database, but data warehouse can be considered for future scope. The first step in the Load phase is to define the target schema, which is the structure of the target system where the data will be loaded. This involves creating tables, defining fields, and specifying data types. After the target system has been set up, the next step is to load the transformed data into the target system. Validating the data involves checking for errors, duplicates, and missing data, and correcting any issues.

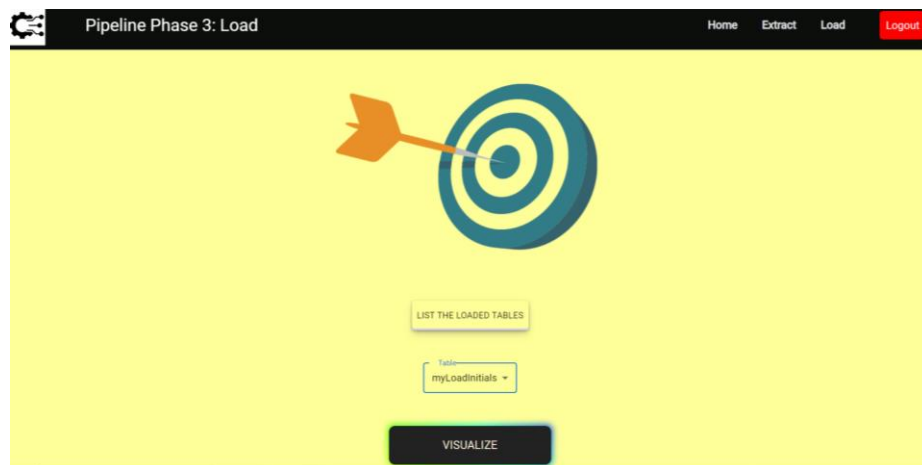


Fig 13. Data Loading Page

2. Data Standardization Implementation:

- Define the data standardization requirements based on the company's needs and data sources. Plan and design the data standardization workflow based on the data sources and data types. Data format can be inter converted into JSON, Excel and CSV formats at the Extraction and Transformation stage of the pipeline.

- Test and refine the data standardization workflow: Test the data standardization workflow to ensure that it is accurate, efficient, and meets the organization's needs.

3. Data Visualization Implementation:

- Develop the visualization module using ReactJS as frontend programming language and leveraging the benefits of Chart.js API in creating the charts.
- Connect the data visualization tool to the database: Configure the data visualization tool to connect to the database and retrieve the required data.
- Design and develop the data visualizations: Create interactive and engaging data visualizations based on the organization's needs and requirements. Test the data visualizations to ensure that they are accurate, efficient, and meet the organization's requirements.



Fig 14. Data Visualization Module

7.4 Security Aspects

When it comes to security aspects for ETL (Extract, Transform, Load), data standardization, and visualization application, there are several considerations to keep in mind. Here are some key security aspects to consider for each of these areas:

ETL Security:

1. **Data Encryption:** Ensure that data is encrypted both at rest and in transit during the ETL process. This includes securing connections between data sources, ETL servers, and target systems.

2. **Access Control:** Implement strong access controls to restrict access to ETL systems and databases. Only authorized individuals should have the necessary privileges to perform ETL operations.
3. **Data Validation and Sanitization:** Validate and sanitize input data to prevent injection attacks and protect against malicious code execution during the transformation process. Used only parameterized queries and input validation techniques.
4. **Error Handling and Logging:** Implement proper error handling mechanisms to handle exceptions and log errors securely. Avoid exposing sensitive information in error messages that could be exploited by attackers.

Data Standardization Security:

1. **Quality Assurance:** Implement rigorous data validation checks to ensure that standardized data meets predefined quality criteria. Validate and sanitize data inputs to prevent security vulnerabilities.
2. **Secure Data Storage:** Safeguard standardized data by applying appropriate access controls, encryption, and monitoring techniques to the storage systems where the data resides.

Visualization Application Security:

1. **User Authentication and Authorization:** Implement robust authentication mechanisms to verify the identity of users accessing the visualization application.
2. **Secure Communication:** Ensure that communication between the visualization application and backend systems is encrypted using secure protocols such as HTTPS.
3. **Secure Session Management:** Implement secure session management techniques, including session expiration, session token management, and protection against session hijacking.

Chapter VIII: Performance Evaluation

8.1 Performance Metric

S No.	Criteria	Summary	Results
1	Data completeness	Measures the percentage of data that was successfully extracted, transformed, and loaded into the target system.	The application worked perfectly fine for considerable amount of data (~100mb)
2	Data accuracy	This metric measures the degree to which the data in the target system reflects the original source data	The transformed data is accurate up to the changes made manually by the user.
3	Data consistency	Measures the degree to which the data in the target system is consistent across different sources and time periods	The standardization module successfully performs conversions to different data formats of varying size.
4	Processing time	The time it takes to complete each phase of the ETL process, from data extraction to data loading.	Loading large amount of data takes some time as connection has to be maintained with database server, followed by transformation process.
5	Visualization response time	The time it takes for the system to generate and display visualizations based on the data.	Charts are generated quickly from the loaded data in the database.

Table 3. Application Performance Metric

8.2 Testing performed

Manual testing for ETL (Extract, Transform, Load), visualization, and standardization applications involves several steps:

1. ETL Testing:
 - Data Extraction Testing: Verifying that the data extracted from the source system matches the expected data. Tested upload option for incompatible files types and error handling working according to expectations.
 - Data Transformation Testing: Verified that the data is transformed correctly according to the business rules. All the eight transformation methods tested successfully on different fields with different data types.
 - Data Loading Testing: Verify that the data is loaded into the target system without any data loss or data truncation. The data is stored without error into MS SQL database with user specified table name and fetched successfully for different purposes.

2. Visualization Testing:

- **Functional Testing:** Verified that the visualization tool works as expected and displays the data correctly. Interactive charts are being rendered based on user selection of fields of dataset, and these charts can be downloaded on local system.
- **Performance Testing:** Verified that the visualization tool can handle large amounts of data and respond quickly after changes were made into data present in database.

3. Standardization Testing:

- **Data Standardization Testing:** Verified that the data is standardized according to the specified data standards. The extracted and transformed data can be downloaded in JSON, Excel and CSV format interchangeably on local system of the user.
- **Integration Testing:** Verified that the application can integrate with other systems and processes, such as working seamlessly in conjunction with the ETL pipeline.

4. Login/Registration Testing

Here, our aim is to test the Register and Login Functionalities with various combinations of test cases, and the modules are working according to set user requirements and performance measures.

STEP ID	STEP DESCRIPTION	TEST DATE	EXPECTED RESULTS	ACTUAL RESULTS	PASS / FAIL
1	Navigate to the Data source page and upload the file	05-05-2023	Data file should be uploaded successfully	As Expected	Pass
2	Click the "Do Extraction" button	05-05-2023	Extraction page should be loaded with display of data	As Expected	Pass
3	Select the columns and rows to be extracted and click Extract	05-05-2023	Selected data extracted successfully	As Expected	Pass
4	Download standardized files in different data formats available	05-05-2023	Files downloaded successfully	As Expected	Pass

Table 4. Sample Test Case for Extraction module

Chapter IX: Results and Deployment

The final results of the ETL standardization project can provide numerous benefits to stakeholders and company after incorporating almost all the future prospects and use cases as already mentioned in the requirements, following are some of the benefits the ETL application would provide to the end-users once deployed:

1. **Accurate Data:** The ETL application extracts data from various sources, transforms it into a standard format, and loads it into a database, ensuring that the data is accurate, consistent, and reliable.
2. **Improved Data Analysis:** The data standardization component of the application ensures that the data is comparable and can be easily analysed. The data visualization tools enable stakeholders to view the data in a way that is easy to understand, identify trends and patterns, and make informed decisions.
3. **Enhanced Efficiency:** The ETL application automates the data processing, reducing manual effort and errors. The data standardization component eliminates the need for manual data cleaning, saving time and resources. The data visualization tools enable stakeholders to quickly access the data they need and make informed decisions faster.
4. **Better Decision Making:** With accurate, standardized data and powerful visualization tools, stakeholders can make more informed decisions. The data can be used to identify trends, predict outcomes, and optimize operations, resulting in better business outcomes.

Deploying the given application on-premise systems includes the consideration of various factors such as choosing a deployment environment, which can be on-premise, cloud-based, or a hybrid model, ensuring that the data sources are properly configured and that the application can connect to them, deployment involves setting up the ETL processes, data standardization, and loading the data into the data warehouse environment, and creating visualizations after connecting the module to stored data. It is essential to test the deployment thoroughly to ensure that the application is functioning correctly and meeting the organization's needs. Once the deployment is complete, it is essential to maintain and monitor the application regularly. This includes monitoring performance, ensuring data accuracy, and making updates and changes as needed.

Chapter X: Applications of Project

The ETL application along with the data standardization and visualization module has numerous use cases within the company's domain, but the similar project can be leveraged by other businesses and organizations to take best out of their data requirements and perform analytics on the raw data from ant available source. Some of the applications of the project are as follows: -

1. Finance: In the finance industry, the ETL application can be used to extract, transform, and load financial data from various sources, such as stock market data feeds, transaction data, and accounting data. Standardizing the data enables more accurate financial analysis, while data visualization tools enable stakeholders to quickly view trends, compare data, and make informed decisions.
2. Healthcare: In the healthcare industry, the ETL application can be used to extract, transform, and load patient data from various sources, such as electronic health records (EHRs) and medical devices. Data standardization enables easier data comparison and sharing, while data visualization tools enable physicians and other healthcare providers to visualize patient data, track patient progress, and make more informed treatment decisions.
3. Manufacturing: In the manufacturing industry, the application can be used to extract, transform, and load data from various sources, such as sensors, machines, and production lines. Standardizing the data enables more accurate production planning and optimization, while data visualization tools enable stakeholders to quickly view production metrics, identify bottlenecks, and make informed decisions.
4. Retail: In the retail industry, the application can be used to extract, transform, and load data from various sources, such as sales data, customer data, and inventory data. Standardizing the data enables more accurate sales forecasting and inventory management, while data visualization tools enable stakeholders to quickly view sales and customer data, identify trends, and make informed decisions.

Overall, an ETL, visualization, and data standardization application project has applications across different industries and domains. By leveraging the power of data, organizations can gain valuable insights, optimize their operations, and make better-informed decisions.

Chapter XI: Conclusion and Future Prospects

In conclusion, an ETL (Extract, Transform, Load) application project that includes data standardization and visualization components can be a valuable tool for organizations to gain insights from their data. The project involves extracting data from various sources, transforming it into a useful format, standardizing the data, and loading it into a database. With this foundation, powerful data visualizations and reports can be created to help users better understand and analyze the data. Successfully implementing this project requires a deep understanding of data processing, database management, data standardization techniques, and visualization tools. It also involves careful planning, task prioritization, and efficient execution to ensure that the project achieves its goals. Following are some of the future scopes the project:

- **Real-time Data Processing:** As the demand for real-time data processing continues to grow, the ETL application can be enhanced to support streaming data, enabling company to analyze large volume of online data in real-time.
- **Cloud Computing:** The ETL application can be deployed in the cloud, allowing company to scale their data processing and storage needs on-demand and reducing the costs associated with maintaining on-premise infrastructure.
- **Automated Data Quality Checks:** The ETL application can be enhanced to include automated data quality checks, ensuring that data is consistent, accurate, and complete, reducing errors and inconsistencies in analysis.
- **Incorporating more transformation operations and providing more dynamic visualization options to the user for creating interactive dashboards.**
- **The data source security validation and access control are under consideration for the application's security needs.**

Overall, an ETL, visualization, and data standardization project has the potential to provide significant value to the company by helping them make better-informed decisions based on their data. With the right approach, tools, and skills, organizations can successfully build a robust ETL application that meets their specific needs, empowers their users to gain valuable insights from their data, and ensures that their data is accurate, reliable, and comparable.

Chapter XII: References

- 1] S.Sajida, Dr.S.Ramakrishna, 2015, A Study of ExtractâTransformâLoad (ETL) Processes, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NCACI – 2015 (Volume 3 – Issue 18),
- 2] Manish Manoj Singh,2022,Extraction Transformation and Loading (ETL) of Data Using ETL Tools, IJRASET, ISSN : 2321-9653
- 3] Kraetz, D., Morawski, M. (2021). Architecture Patterns—Batch and Real-Time Capabilities. In: Liermann, V., Stegmann, C. (eds) The Digital Journey of Banking and Insurance, Volume III. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-78821-6_6
- 4] J. Sreemathy, R. Brindha, M. Selva Nagalakshmi, N. Suvekha, N. Karthick Ragul and M. Praveennandha, "Overview of ETL Tools and Talend-Data Integration," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 1650-1654, doi: 10.1109/ICACCS51430.2021.9441984.
- 5] J. Sreemathy, I. Joseph V., S. Nisha, C. Prabha I. and G. Priya R.M., "Data Integration in ETL Using TALEND," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 1444-1448, doi: 10.1109/ICACCS48705.2020.9074186.
- 6] Trujillo, J., Luján-Mora, S. (2003). A UML Based Approach for Modeling ETL Processes in Data Warehouses. In: Song, IY., Liddle, S.W., Ling, TW., Scheuermann, P. (eds) Conceptual Modeling - ER 2003. ER 2003. Lecture Notes in Computer Science, vol 2813. Springer, Berlin, Heidelberg.
- 7] Vyas, Dr Sonali & Vaishnav, Pragya. (2017). A comparative study of various ETL process and their testing techniques in data warehouse. Journal of Statistics and Management Systems. 20. 753-763. 10.1080/09720510.2017.1395194.
- 8] M. Golfarelli, D. Maio and S. Rizzi, "Conceptual design of data warehouses from E/R schemes," Proceedings of the Thirty-First Hawaii International Conference on System Sciences, Kohala Coast, HI, USA, 1998, pp. 334-343 vol.7, doi: 10.1109/HICSS.1998.649228.
- 9] Shaker H. Ali El-Sappagh, Abdeltawab M. Ahmed Hendawi, Ali Hamed El Bastawissy, A proposed model for data warehouse ETL processes,Journal of King Saud University - Computer and Information Sciences,Volume 23, Issue 2,2011
- 10] M. Mrunalini, T. V. S. Kumar and K. R. Kanth, "Simulating Secure Data Extraction in Extraction Transformation Loading (ETL) Processes," 2009 Third UKSim European Symposium on Computer Modeling and Simulation, Athens, Greece, 2009, pp. 142-147, doi: 10.1109/EMS.2009.111.
- 11] A. Simitsis, P. Vassiliadis and T. Sellis, "Optimizing ETL processes in data warehouses," 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 2005, pp. 564-575, doi: 10.1109/ICDE.2005.103.
- 12] Muñoz, L., Mazón, JN., Pardillo, J., Trujillo, J. (2008). Modelling ETL Processes of Data Warehouses with UML Activity Diagrams. In: Meersman, R., Tari, Z., Herrero, P. (eds) On the Move to Meaningful Internet Systems: OTM 2008 Workshops. OTM 2008. Lecture Notes in Computer Science, vol 5333. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88875-8_21
- 13] aisel.aisnet.org/hicss-50/st/big_data_engineering/2/
- 14] Abdeltawab M.A. Hendawi and Shaker H. Ali El-Sappagh, EMD: entity mapping diagram for automated extraction, transformation, and loading processes in data warehousing, Published Online:May 25, 2012pp 255-272, <https://doi.org/10.1504/IJIDS.2012.047003>
- 15] Ankorian, Itamar,Change Data Capture Efficient ETL for Real-Time BI, www.proquest.com/openview/ebb6b945abb5673c78aa513ff8c9489c/1?pq-origsite=gscholar&cbl=51938
- 16] Ricardo Jorge Santos, Jorge Bernardino Real-time data warehouse loading methodology, IDEAS '08: Proceedings of the 2008 international symposium on Database engineering & applications
- 17] Esmail Ali, F.S. (2014). A Survey of RealTime Data Warehouse and ETL. International Scientific Journal of Management Information Systems, 9 (3), 03-09.

Chapter XIII: Publication Details

^{ci} Submission Summary

Conference Name

2023 IEEE World Conference on Applied Intelligence and Computing

Paper ID

1484

Paper Title

Standardization of ETL Process

Abstract

The ETL (Extract, Transform, Load) standardization project aims to develop an integrated solution that enables efficient management, processing, and analysis of large volumes of data specific to company's data requirements. The project includes researching and developing an ETL application along with standardization module for interconversions of various data formats, leverages ETL tools and technologies to extract data from various sources, transform it into a format suitable for analysis, and load it into a data warehouse. The project also includes the development of a user-friendly data visualization tool that provides intuitive and interactive access to the data warehouse's contents. The project's objective is to enable the company to make informed decisions based on accurate and up-to-date data, leading to improved business performance and maintain data coming from various sources. The Proof of Concept and research is aimed at arriving at an all-in-one application similar to other Business Intelligence tools available in the market, but is lightweight and specific to organization's technological and data requirements. The project's scope covers the entire ETL, data warehousing, and visualization process, including requirements gathering, design, development, testing, and deployment. Overall, the ETL, visualization, and data warehousing project addresses the critical need for effective data management, processing, and visualization in today's data-driven business landscape

Created on

5/18/2023, 10:52:31 AM

Last Modified

5/18/2023, 10:52:31 AM

Authors

Prabhat Panwar (MIT WPU) < 1032190048@mitwpu.edu.in> ✓

Mehul Pansari (MIT WPU) < 1032190041@mitwpu.edu.in> ✓

Utkrist Agrawal (MIT WPU) < 1032190030@mitwpu.edu.in> ✓

Nikunj Padia (MIT WPU) < 1032190109@mitwpu.edu.in> ✓

Rashmi Phalnikar (MIT WPU) < rashmi.phalnikar@mitwpu.edu.in> ✓

Submission Files

ETL_ReserachPaper.docx (93.4 Kb, 5/18/2023, 10:50:00 AM)

Submission Questions Response

1. Status of using third-party material in your article.

Third party content is defined as any material within the manuscript which is not your original work. Third party content may consist of text passages, figures, photos, screenshots, etc. and be found in many places such as but not limited to - the Internet, print and online books and articles, theses, annual reports, conference material, photocopies, course packages, translations, visually impaired readers. In particular, pay close attention to sensitive images such as identifiable persons or human research participants, recognisable architecture, logos, brands/trademarks, and images from online photo libraries. Please refer to Springer Nature Guide to Copyright and Permissions for further guidance. You are responsible for clearing the rights for third party content under your publishing agreement. Please confirm if your manuscript contains any third-party content?

I am not using third-party material for which formal permission is required

2. Conflict of interest

The authors declare no conflict of interest and all authors are aware of this submission. In case of any conflict submitting author will be responsible.

Appendix

Base Paper

- **Reference:** Manish Manoj Singh, 2022, Extraction Transformation and Loading (ETL) of Data Using ETL Tools, IJRASET, ISSN: 2321-9653
- **Year of Publication:** 2022
- **Journal:** International Journal for Research in Applied Science & Engineering Technology



International Journal for Research in Applied Science & Engineering Technology (IJRASET)

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538

Volume 10 Issue VI June 2022- Available at www.ijraset.com

Extraction Transformation and Loading (ETL) of Data Using ETL Tools

Manish Manoj Singh

MCA, Thakur Institute of Management Studies, Career, Development & Research (TIMSCDR), Mumbai, India

Abstract: This Research Paper presents the Extract, Transform, Load (ETL) Process and discusses various ETL Tools Available in the Market. A huge piece of BI frameworks is a well-performing Implementation of the Extract, Transform, and Load (ETL) process. In BI projects, implementing the ETL process can be the big task ETL is the core process of Data integration which is associated with Data Warehouse. This paper also focuses on the best ETL Tools and which tool can be the best for the ETL process.

I. INTRODUCTION

Business intelligence has reached wide recognition in the last few years.

A data warehouse is only a social data set that is intended for inquiry and investigation rather than for exchange handling.

The Data warehouse information is only a mix of authentic information just as conditional information. We want to load the data warehouse consistently with the goal that it can fill its need of working with the business examination. To play out this interaction information from at least one functional framework should be separated and duplicated into the information distribution centre. ETL is a course of extracting data from source frameworks and bringing it into the data warehouse, which stands for extraction Transformation and loading. The procedure and undertaking of ETL have been notable for a long time, and are not remarkable to information stockroom conditions Extract, Transform and Load (ETL) process is One of the important components of Business Intelligence.

ETL processes take up to 80% of the effort in BI projects it is a data integration function that involves extracting data from outside sources (operational systems), transforming it to fit business needs, and eventually stacking it into an information distribution centre To tackle the issue, organizations use extract, transform and load (ETL) innovation, which incorporates perusing information from its source, tidying it up and arranging it consistently, and afterward composing to the objective vault to be taken advantage of.

The information which is utilized in ETL cycles can emerge out of any source like a centralized server application, an ERP application, a CRM device, a level document, or an Excel spreadsheet. ETL tool can gather, read and move information from various information structures and across various stages, similar to a centralized computer, server In this paper, we have analysed some of the ETL Tools.

II. ETL PROCESS

ETL (Extract, Transform, and Load) is a cycle that processes information from different sources and places it into a data warehouse.


The purpose of ETL is to provide the users, not only a process of extracting data from source systems and bringing it into the data warehouse but also provide the users with a typical stage to incorporate their information from different stages and applications

ETL is a cycle that extricates the information from various RDBMS source frameworks, then, at that point, changes the information (like applying estimations, connections, and so on) lastly stacks the information into the Data Warehouse system.

Extract, Transform, Load three database capacities that are consolidated into one instrument that computerizes the interaction to haul information out of one database and spot it into another database. The database functions are described following:

ETL involves the following tasks.

Plagiarism Report

		Similarity Report ID: oid:28480:35399540	
PAPER NAME			
B56 Capstone Report.docx			
WORD COUNT		CHARACTER COUNT	
9425 Words		54373 Characters	
PAGE COUNT		FILE SIZE	
38 Pages		2.0MB	
SUBMISSION DATE		REPORT DATE	
May 14, 2023 9:47 PM GMT+5:30		May 14, 2023 9:48 PM GMT+5:30	
<hr/>			
<p>● 30% Overall Similarity</p> <p>The combined total of all matches, including overlapping sources, for each database.</p> <ul style="list-style-type: none">• 19% Internet database• 15% Publications database• Crossref database• Crossref Posted Content database• 24% Submitted Works database			

Individual Contribution and Module Development

Following are the details of contribution given by each team member along with their roles, responsibilities, objectives and module development as part of Capstone project team: -

- **Utkrist Agrawal:** Worked on application UI and frontend development and focused on the data standardization module research and development.
- **Mehul Pansari:** Worked as Full Stack Software developer focusing on development of backend logic using Django and completed the ETL pipeline module after thorough research and demo products.
- **Prabhat Panwar:** Worked as Team Lead, QA tester, Frontend developer and managed the project workflow using JIRA and High/Low level diagrams development.
- **Nikunj Padia:** Responsible for researching the data warehousing concepts and developing the Data Loading module, developed user login and registration feature.

Name	Role	Research Objectives	Module Development
Prabhat Panwar	Team lead, Software architect, Frontend developer	ETL survey and Data warehouse/data lake exploration and comparison, Batch processing vs real time data handling	Visualization Module, System Architecture Development and QA testing
Utkrist Agrawal	UI/UX and Frontend developer	Different data formats and representations, Data transformation techniques and best python libraries for project	Data Standardization module and UI Wireframes, Application Frontend
Mehul Pansari	Full stack developer, ETL developer	Survey of suitable Backend and Frontend technology for data transformation and storage applications, Data visualization techniques for enterprise data with significance of each chart.	ETL pipeline module and Django backend integration with ReactJS
Nikunj Padia	Backend developer, Warehouse manager	API concepts for fetching the data from source and transforming the data, Find best data storage technology for unprocessed data.	Data Warehouse survey and Data Loading setup, Authentication module

Table 5. Individual Team Member Contribution

Project to Outcome Mapping

Objectives:

- 1.
- 2.
- 3.
- 4.

Sr. No.	PRN No.	Student Name	Individual Project Student Specific Objective	Learning Outcomes mapped (To be filled by Guide)	Marks