MIT-WPU
॥ विश्वशान्तिर्ध्रुवं ध्रुवा ॥

Standardization of ETL Process

End-Term Presentation: Group ID B56

Domain of Project - Software Engineering and Database Management

Industry Supported - Symphony Tech

Presented & Developed by :

S.NO	NAME	PRN	ROLL NO.
1.	Utkrist Agrawal	1032190030	PB 03
2.	Mehul Pansari	1032190041	PB 04
3.	Prabhat Panwar	1032190048	PB 05
4.	Nikunj Padia	1032190109	PB 11

External Guide:
Mr. Ajay Ghatpande &
Mr. Ravi Khare

Inhouse Guide:
Dr Rashmi Phalnikar

Agenda

TEAM

- ➔ Requirements Gathering
- ➔ Literature Survey
- ➔ System Design
- ➔ Implementation & Testing
- ➔ Programming & Ethics
- ➔ Security aspects
- ➔ Project maintenance
- ➔ Future aspects

INDIVIDUAL

- ➔ Individual Aim
- ➔ Individual Objective
- ➔ Individual contribution with module design

TEAM

Problem Statement & justification

The ETL Standardization 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.
- The back-end will be responsible for executing the ETL jobs specified by the user. 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.

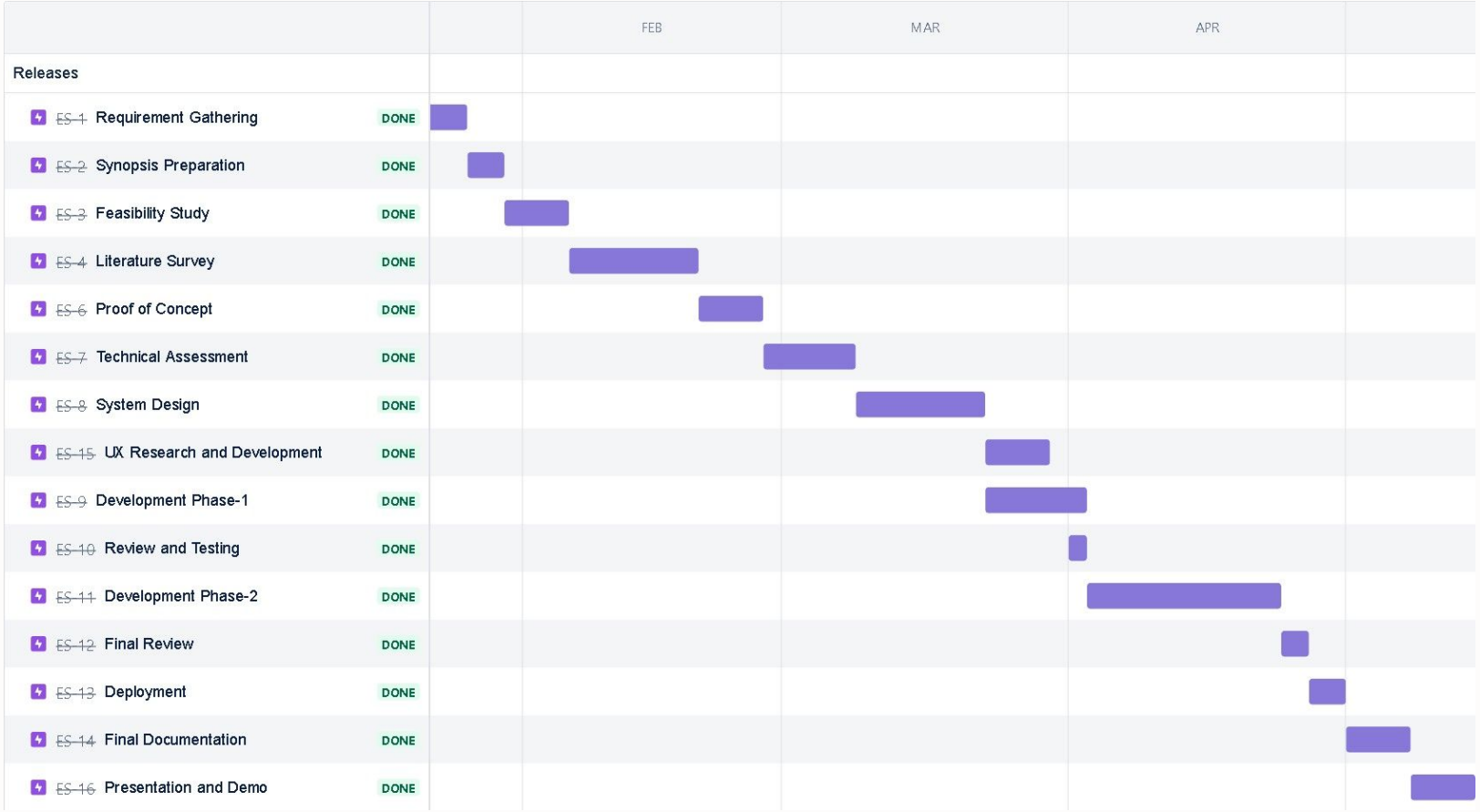
Requirements & timeline

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

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


FEB
MAR
APR
Releases
ES-1 Requirement Gathering DONE
ES-2 Synopsis Preparation DONE
ES-3 Feasibility Study DONE
ES-4 Literature Survey DONE
ES-6 Proof of Concept DONE
ES-7 Technical Assessment DONE
ES-8 System Design DONE
ES-15 UX Research and Development DONE
ES-9 Development Phase-1 DONE
ES-10 Review and Testing DONE
ES-11 Development Phase-2 DONE
ES-12 Final Review DONE
ES-13 Deployment DONE
ES-14 Final Documentation DONE
ES-16 Presentation and Demo DONE

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

Literature Survey

Paper name	Authors	Objective and Methodology	Research Gap /Future Scope	Conclusion
A Study of Extract–Transform– Load (ETL) Processes(2015)	S.Sajida, Dr.S.Ramakrishna	In Data 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 modeling in data 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. This paper also 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. ETL and data governance 4. ETL and machine learning	As the 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 business Intelligence an ETL tool helps to extract the data from one or more sources, cleanse it and loads the date into data warehouse. In data integration techniques, the ETL method is important..	1. Performance comparison between Talend-Data Integration and other ETL tools 2. Integration of Talend-Data Integration with cloud-based ETL platforms 3. 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.

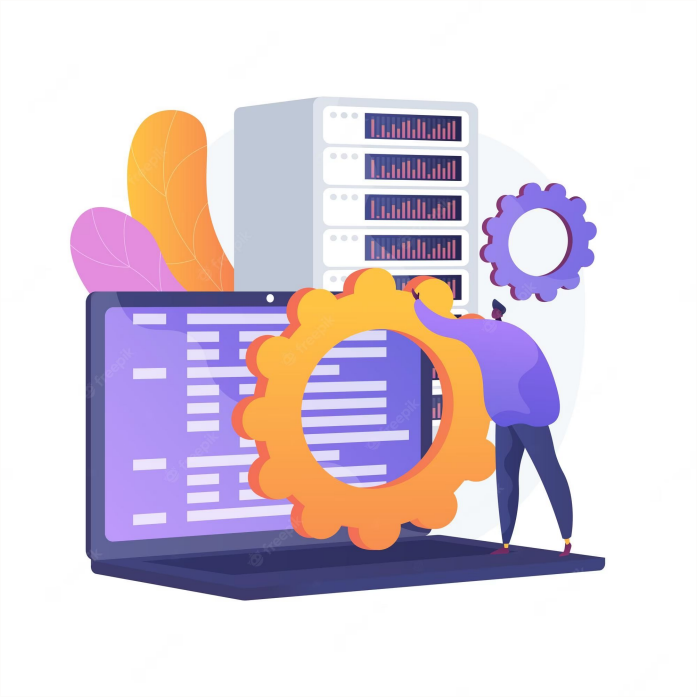
Paper name	Authors	Objective and Methodology	Research Gap /Future Scope	Conclusion
Data Integration in ETL Using TALEND(2020)	Sreemathy J,Infant Joseph V, Nisha S, Chaaru Prabha I, Gokula Priya RM	his paper describes the various steps involved in integrating data from various sources using the ETL process ,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 analyzing disparate data into a much more easy 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 involved in data warehouse projects.	In conclusion, 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 of the paper is to discuss the role and importance of data warehousing in today's business landscape.It concludes by defining a data warehouse as a subject-oriented, integrated, time-variant.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 prohibitively expensive costs associated with it.	The paper concludes by emphasizing the importance, complexity, and criticality of real-time BI and DW as a significant topic of research and practice that needs to be addressed in the future

Paper name	Authors	Objective and Methodology	Research Gap /Future Scope	Conclusion
Real-Time Data Warehouse Loading Methodology (2008)	Ricardo Jorge, Jorge Bernardino	The methodology is focused on four major areas: (1) data warehouse schema adaptation; (2) ETL loading procedures; (3) OLAP query adaptation; and (4) 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, which may be a concern for large-scale data warehouses.	In conclusion, the paper presents a methodology for supporting the implementation of Real-Time Data Warehousing (RTDW) by enabling continuous data integration while minimizing impact on query execution..
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 behavior 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 mechanisms that can be used to address this issue, such as 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 paper has presented the core technology of real-time analysis based on filtering, extracting, and capturing changed data in data log. 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, or PostgreSQL.	There are some missing features in the current implementation of JSON in relational database systems. The authors propose that a native support for JSON data type, similar to that of XML data type, should be implemented in the RDBMSs.	The paper highlights the fact that different RDBMSs have implemented JSON in different ways, with Oracle being the one that has implemented the most concepts specified in the ANSI SQL/JSON standard.

Paper name	Authors	Objective and Methodology	Research Gap /Future Scope	Conclusion
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.	The conclusion of the text is that a new technology for data management of complex and irregular structures, specifically for the representation of thermophysical properties of substances, has been proposed.
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 analyzed 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 analyzer. These components work together to collect data from a variety of sources, transform the data to a format that is suitable for analysis	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 paper concludes that 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. They demonstrate how the UML Activity Diagrams can be used to represent the flow of data through the ETL process	The research gap addressed by the paper is the need for more effective and accessible techniques for modeling ETL processes in data warehouse.	The paper concludes that UML Activity Diagrams can provide an effective way to model ETL processes in data warehouses, offering advantages such as flexibility, intuitiveness.

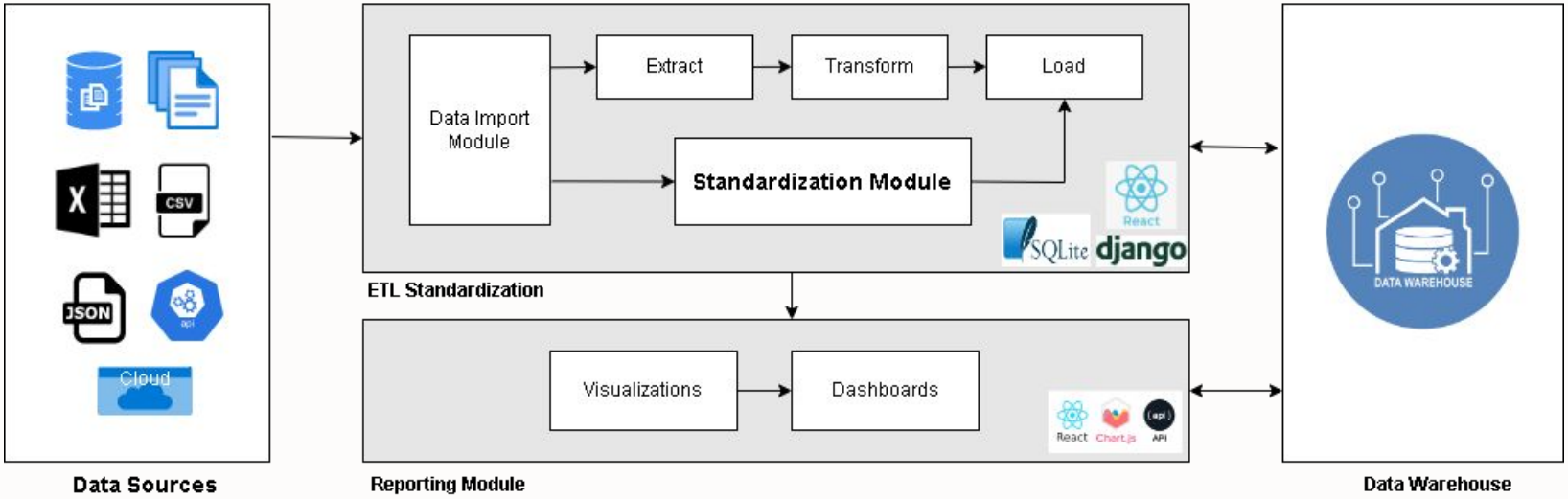
Paper name	Authors	Objective and Methodology	Research Gap /Future Scope	Conclusion
EMD: entity mapping diagram for automated extraction, transformation, and loading processes in data warehousing (2012)	Abdeltawab M.A. Hendawi and Shaker H. Ali El-Sappagh	The paper presents a case study of the EMD methodology 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 methodology can be used to automate the ETL process, including data extraction, transformation, and loading	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 paper concludes that the EMD methodology can help address the challenges associated with automating ETL processes in data warehousing, providing a more flexible and intuitive approach that can be easily understood by non-technical users.
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 concludes by highlighting the benefits of the proposed conceptual model and methodology, such as improved data quality, increased flexibility, and reduced development time and cost.
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 then describe the ETL process in detail, including data extraction, data cleaning, data transformation, and data loading.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 and implementation of data warehouses.	In conclusion, the paper provides a detailed description of the ETL process used in a Land and Resources Star Schema data warehouse.

Paper name	Authors	Objective and Methodology	Research Gap /Future Scope	Conclusion
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 (DF) model, and a semi-automated methodology to build it from pre-existing Entity/Relationship (E/R) schemes or relational database schemes, 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 or relational database schemes.	The paper proposes a conceptual model and a semi-automated methodology for designing data warehouses. The proposed Dimensional Fact (DF) model is independent of the target logical model .
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.
Simulating Secure Data Extraction in Extraction Transformation Loading (ETL) Processes (2009)	M Mrunalini, T V Suresh Kumar, K Rajani Kanth	The objective of the paper is to propose a simulation model for secure data extraction in ETL processes that addresses the security aspects in the conceptual modeling phase. The paper aims to develop a tool that can be used for implementing security policies of the system in ETL processes and validate its features with a case study.	Developing and testing the proposed UML model for secure data extraction in a real-world ETL system with large volumes of data.	This paper presents a simulation model of secure data extraction in ETL processes using UML 2.0. The authors test the model in a software and Business Process Outsourcing company.



System Design

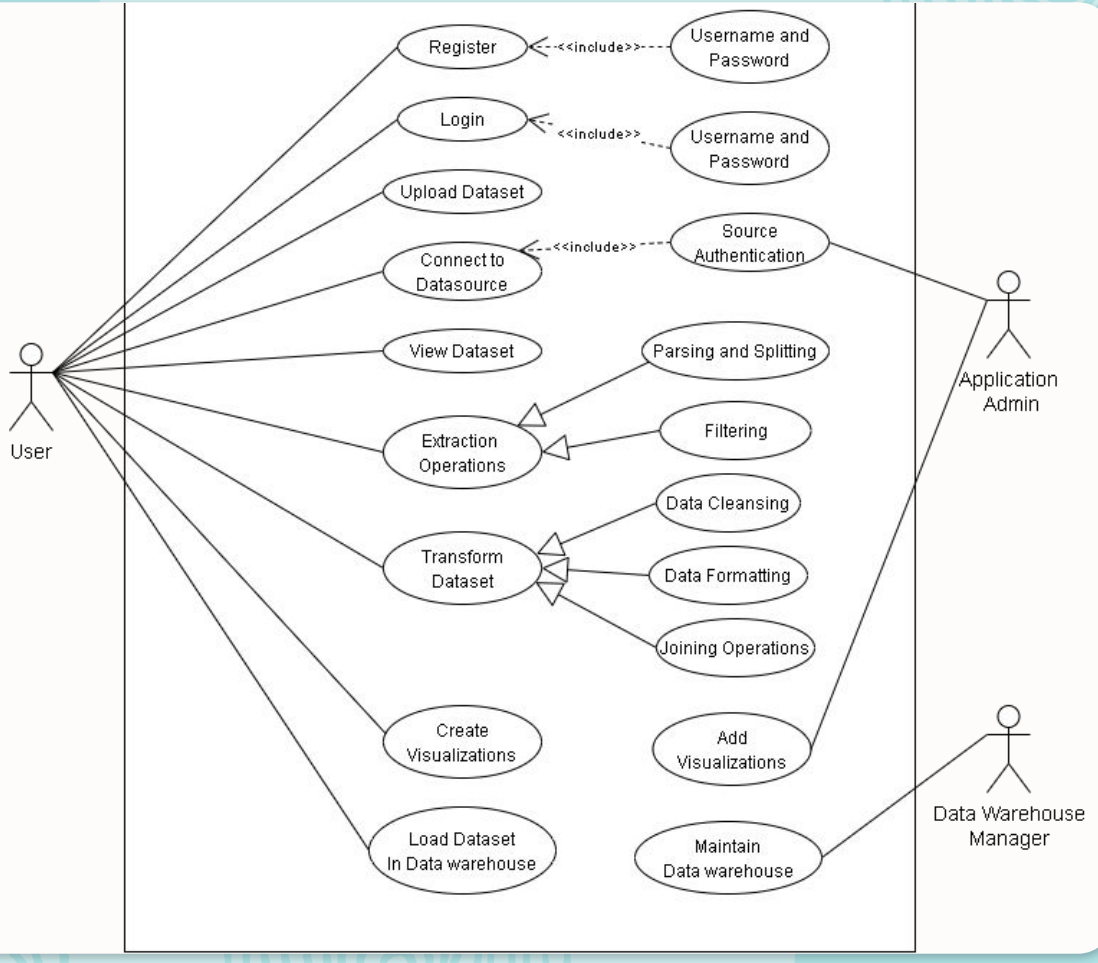
SYSTEM ARCHITECTURE



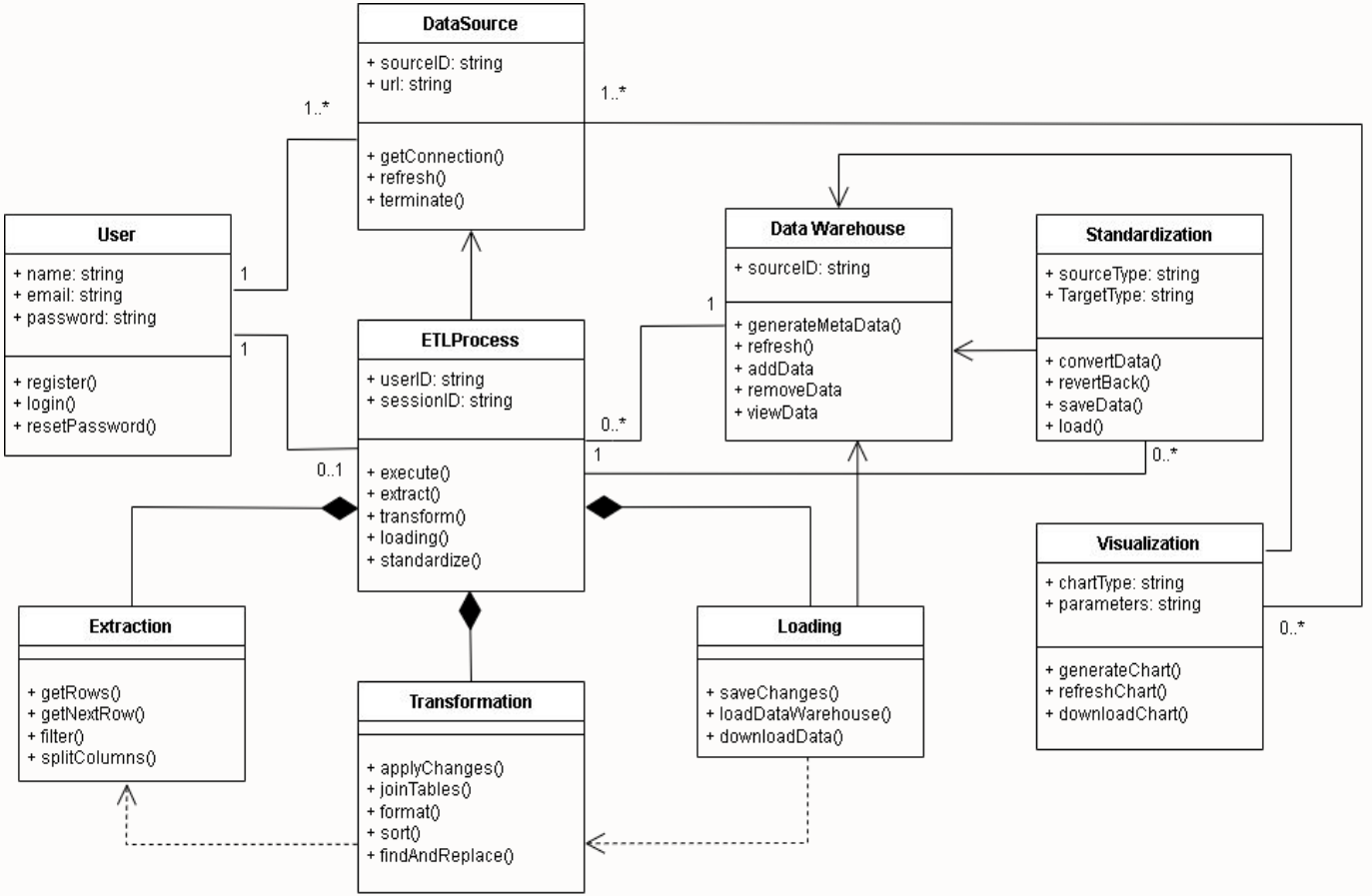
The block diagram for the ETL standardization application includes various components working in sequence corresponding to Data sources (include databases, files, APIs, web services, and other sources of data. Data sources may be located on-premises or in the cloud), Extract component(extracts data from the various data sources and prepares it for transformation), Transform component(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

USE CASE DIAGRAM

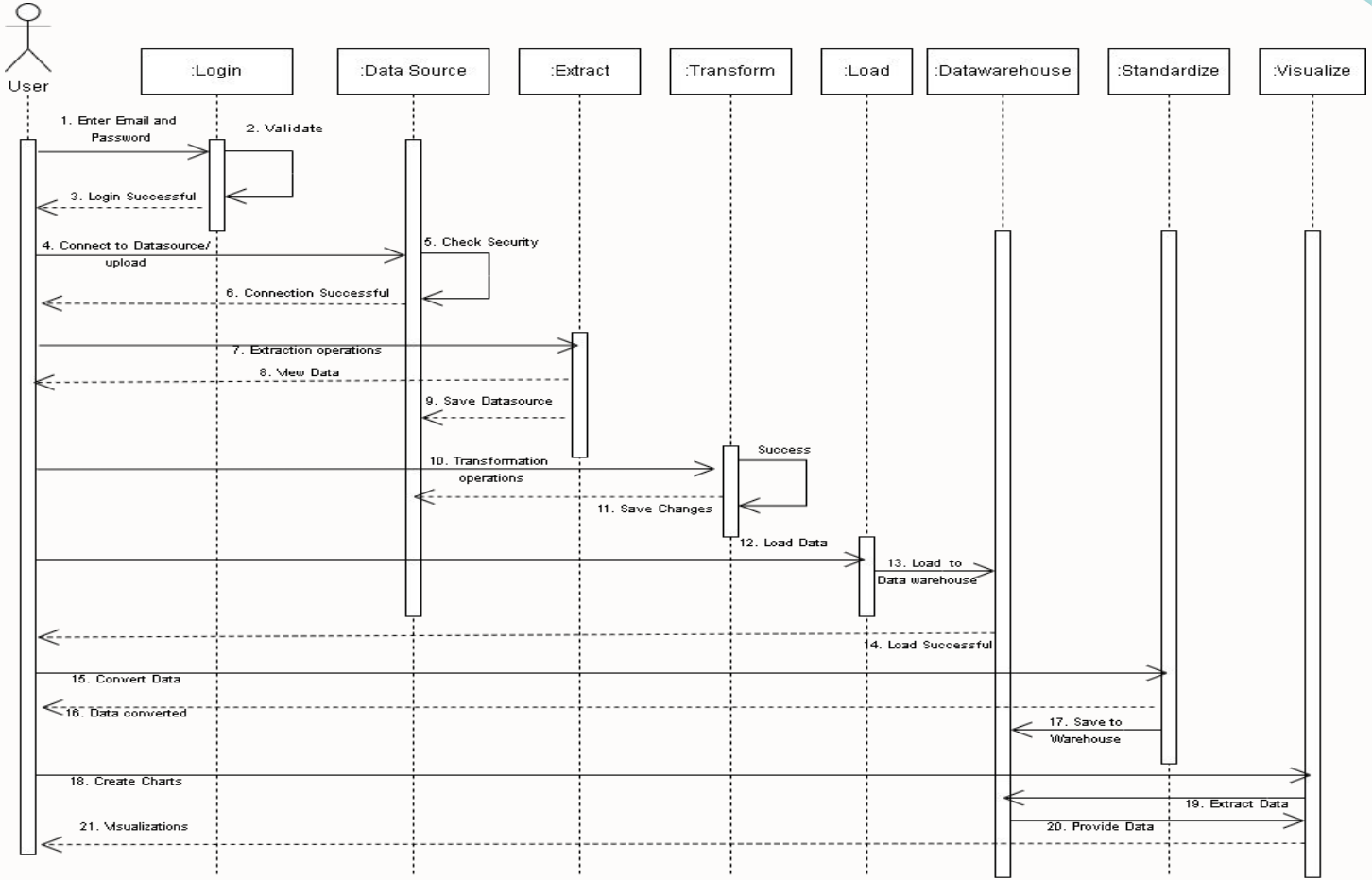
The Use case diagram depicts user, admin and data warehouse manager as actors, and various use cases such as extracting data from various sources, transforming data to meet business rules, loading data into target systems, creating reports, visualizations, managing validations on data, and data standardization



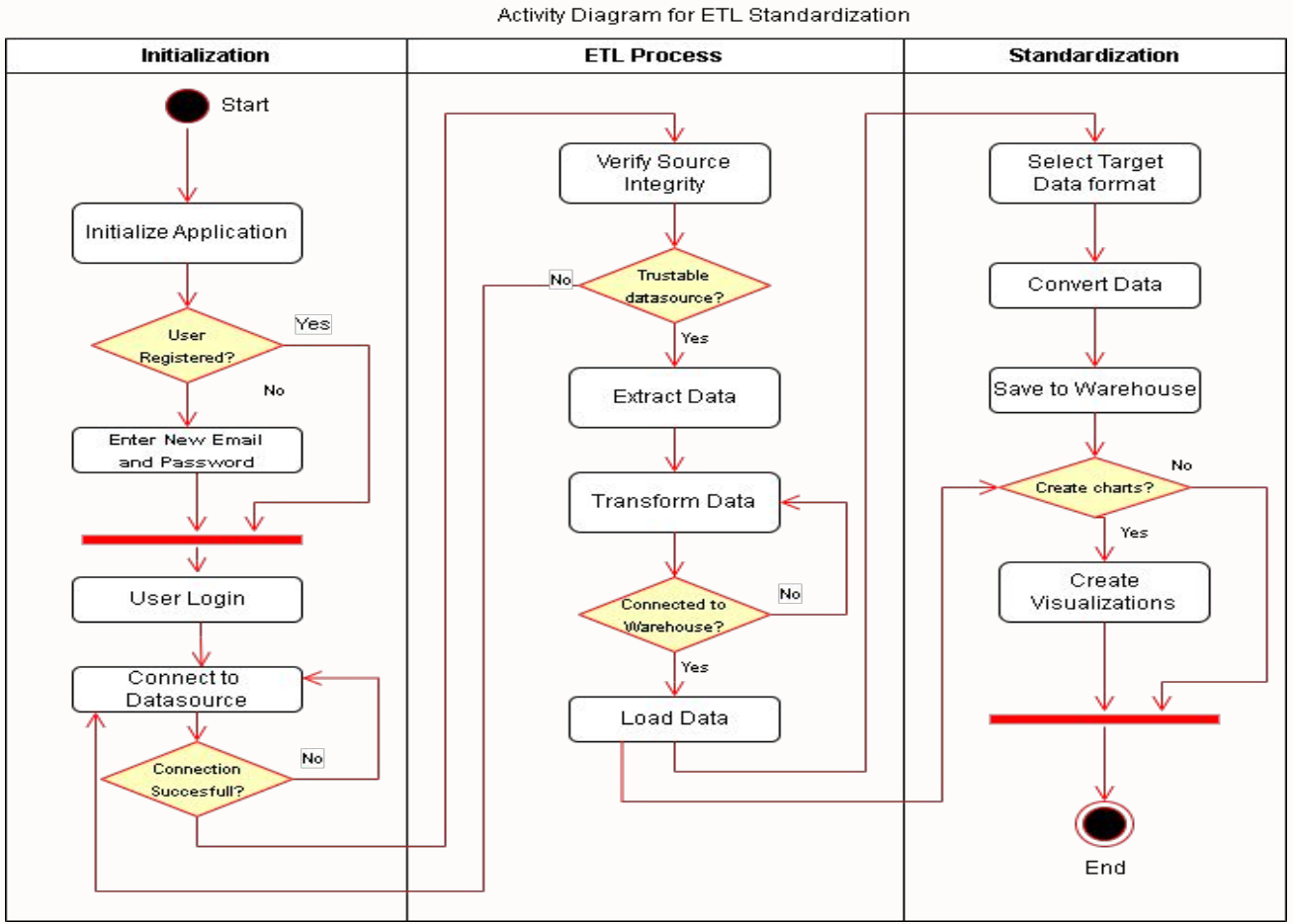
CLASS DIAGRAM



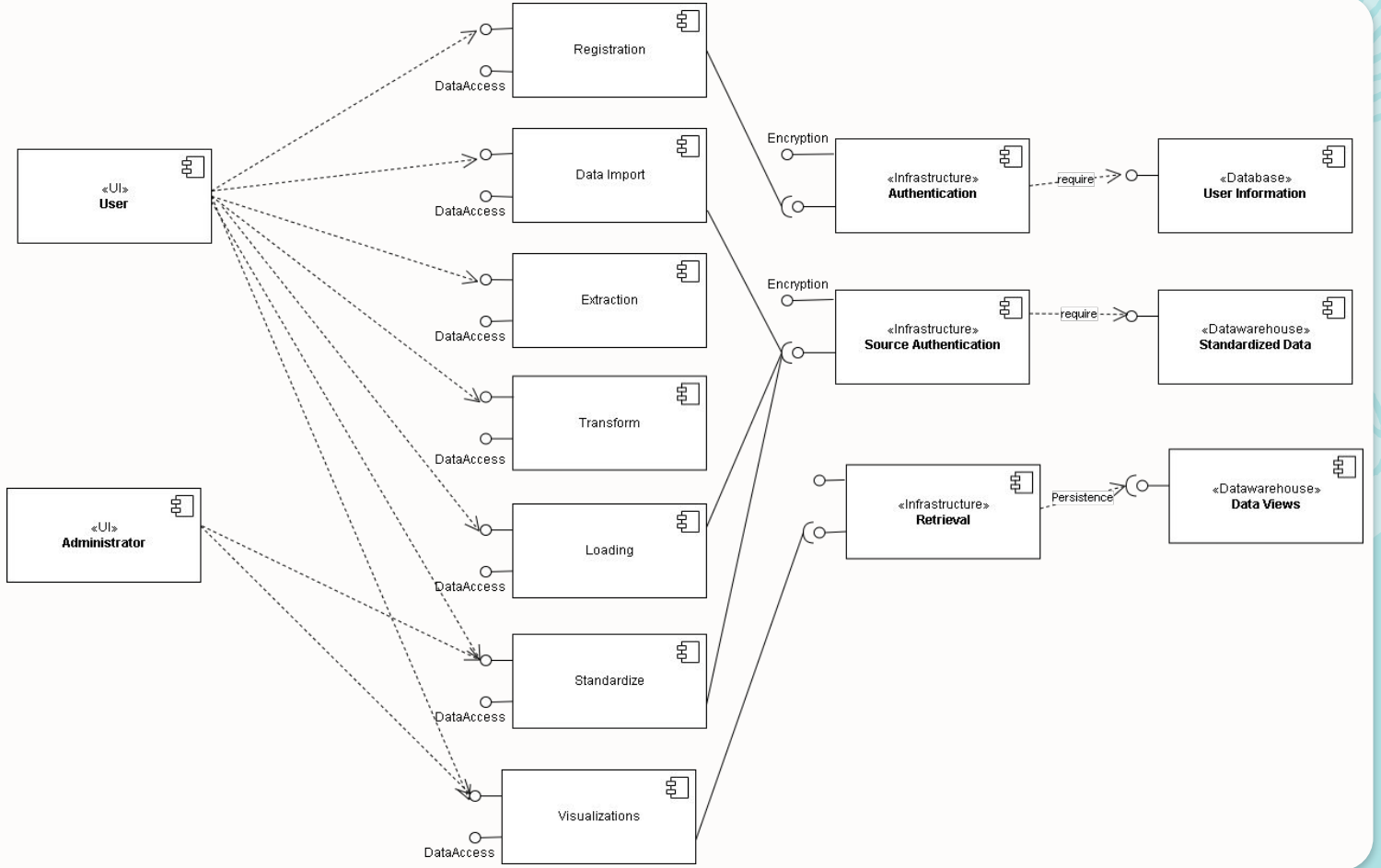
SEQUENCE DIAGRAM



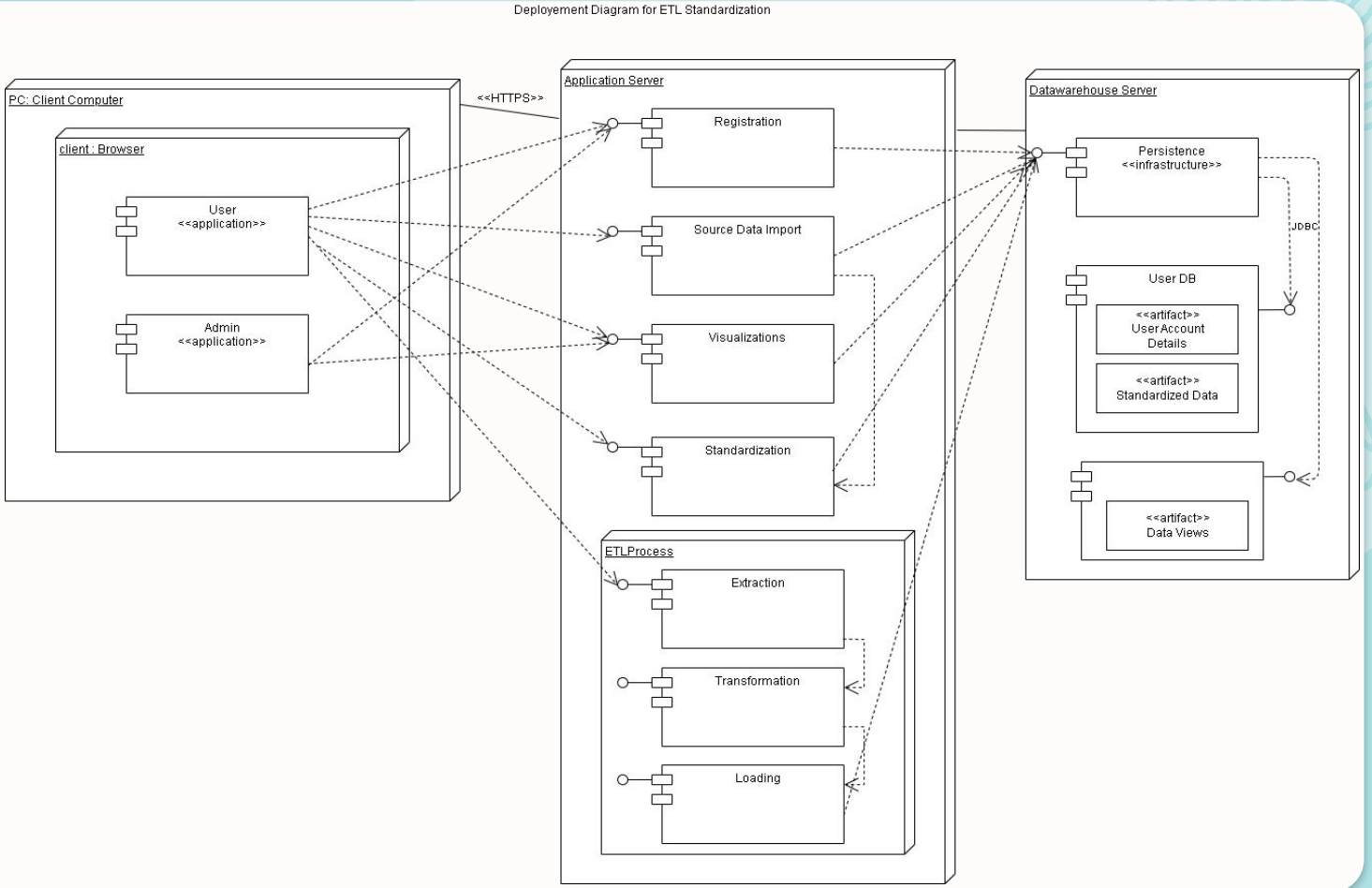
ACTIVITY DIAGRAM



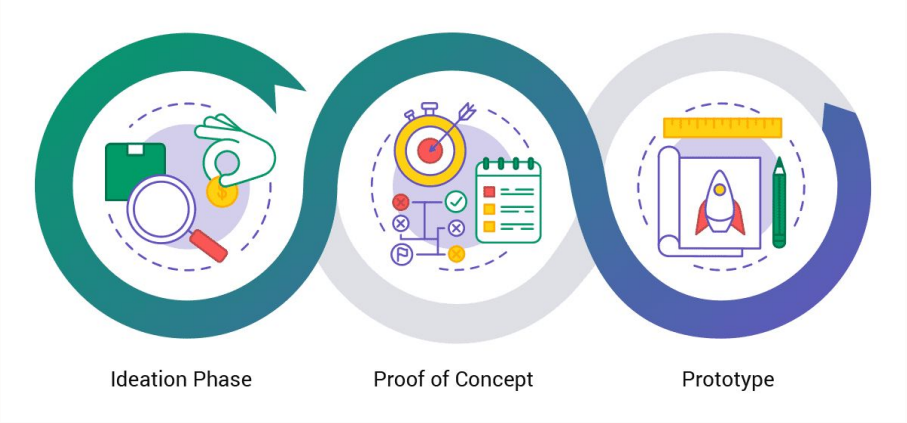
COMPONENT DIAGRAM



DEPLOYMENT DIAGRAM



Proof of Concept(PoC)



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 system's requirements and transformed data is loaded into the target system.
- The visualization module generates meaningful and useful insights from the transformed data, and standardization module successfully downloads data of new format.

Methodology:

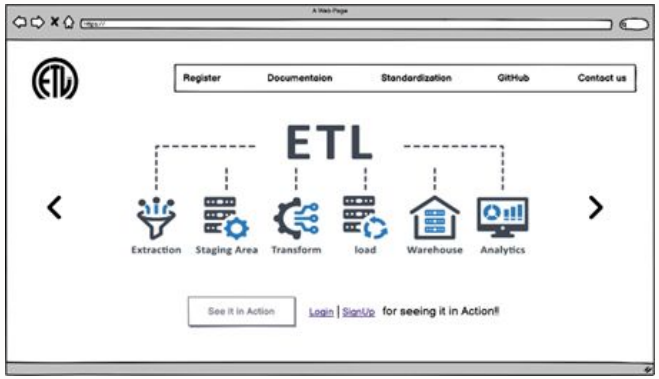
- Identify the source system from which data is to be extracted.
- Develop an ETL pipeline to transform the source data into a format suitable for loading and analysis.
- Identify the target system (data warehouse/database) and load the transformed data into it, followed by data format conversion as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

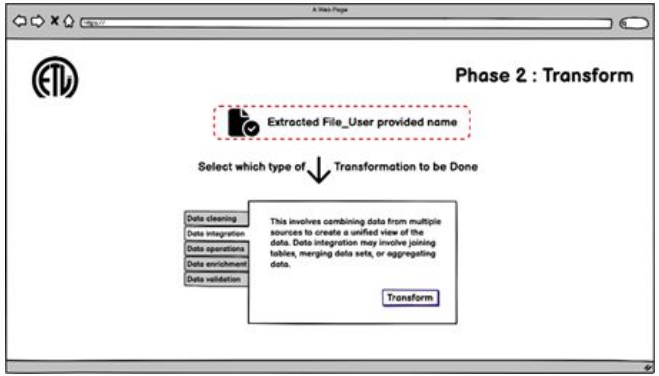
Results:

- Data was extracted from the source system using the ETL tool, which transformed the data into the target system's 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.

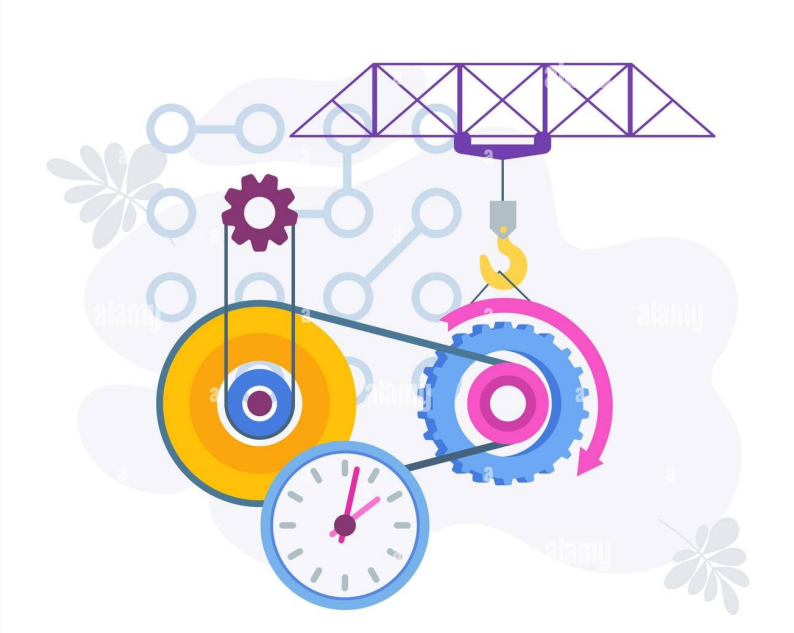
WIREFRAMES

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.





Implementation of Project



Technological Details

- **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.

- **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.

- **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.

Technological Details

- **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

Implementation Details

- 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.

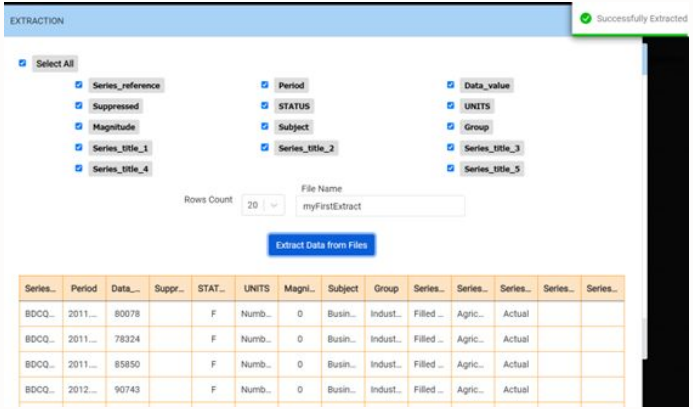
1. ETL Implementation

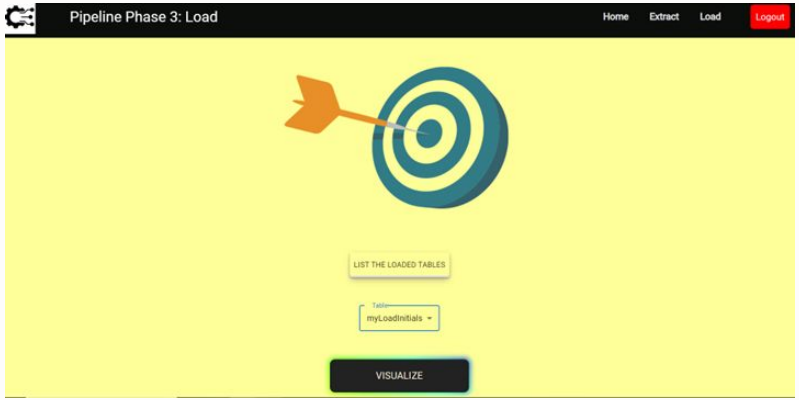
Extract phase: The Extract phase of the ETL (Extract, Transform, Load) process involves retrieving data from various sources and bringing it into a central location, such as a database or data warehouse, for further processing. The first step in the Extract phase is to identify the data sources that need to be extracted. These sources could be databases, files (CSV, Excel, JSON), APIs, or web services.

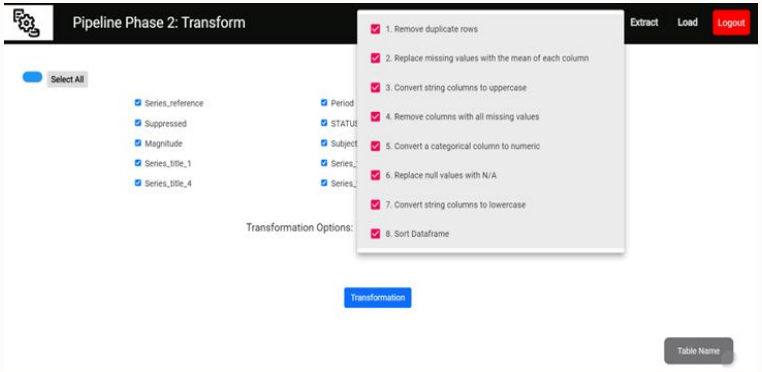
Transform phase: The Transform phase of the ETL (Extract, Transform, Load) process involves cleaning, enriching, and reshaping the extracted data into a format that can be easily loaded into the target system. Some basic transform operation performed by our application includes selecting specific columns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

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.

1. ETL Implementation

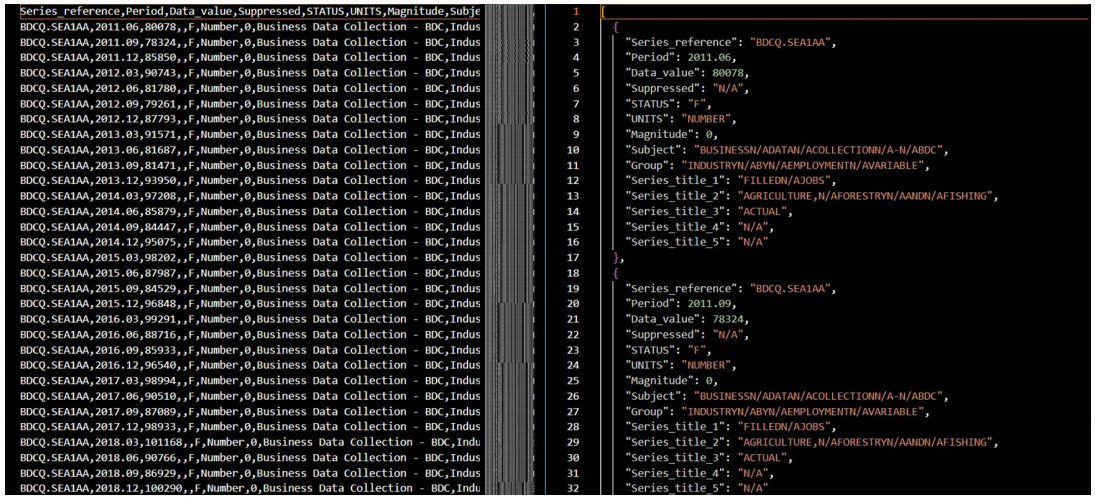






2. Data Standardization Implementation

- Define the data standardization requirements based on the company's needs and data sources. Designed 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.



```
Series_reference,Period,Data_value,Suppressed,STATUS,UNITS,Magnitude,Subje
BDCQ.SEA1AA,2011.06,80078,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2011.09,78324,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2011.12,85850,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2012.03,90743,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2012.06,81780,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2012.09,79261,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2012.12,87793,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2013.03,91571,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2013.06,81687,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2013.09,81471,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2013.12,93950,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2014.03,97208,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2014.06,85879,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2014.09,84447,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2014.12,95075,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2015.03,98202,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2015.06,87987,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2015.09,84529,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2015.12,96848,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2016.03,99291,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2016.06,88716,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2016.09,85933,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2016.12,96540,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2017.03,98994,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2017.06,90510,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2017.09,87089,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2017.12,98933,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2018.03,101168,,F,Number,0,Business Data Collection - BDC,Indu
BDCQ.SEA1AA,2018.06,90766,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2018.09,86929,,F,Number,0,Business Data Collection - BDC,Indus
BDCQ.SEA1AA,2018.12,100290,,F,Number,0,Business Data Collection - BDC,Indu
```

```
[
  {
    "Series_reference": "BDCQ.SEA1AA",
    "Period": 2011.06,
    "Data_value": 80078,
    "Suppressed": "N/A",
    "STATUS": "F",
    "UNITS": "NUMBER",
    "Magnitude": 0,
    "Subject": "BUSINESSN/ADATAN/ACOLLECTIONN/A-N/ABDC",
    "Group": "INDUSTRYN/ABYN/AEMPLOYMENTN/AVARIABLE",
    "Series_title_1": "FILLEDN/AJOBS",
    "Series_title_2": "AGRICULTURE,N/AFORESTRYN/AANDN/AFISHING",
    "Series_title_3": "ACTUAL",
    "Series_title_4": "N/A",
    "Series_title_5": "N/A"
  },
  {
    "Series_reference": "BDCQ.SEA1AA",
    "Period": 2011.09,
    "Data_value": 78324,
    "Suppressed": "N/A",
    "STATUS": "F",
    "UNITS": "NUMBER",
    "Magnitude": 0,
    "Subject": "BUSINESSN/ADATAN/ACOLLECTIONN/A-N/ABDC",
    "Group": "INDUSTRYN/ABYN/AEMPLOYMENTN/AVARIABLE",
    "Series_title_1": "FILLEDN/AJOBS",
    "Series_title_2": "AGRICULTURE,N/AFORESTRYN/AANDN/AFISHING",
    "Series_title_3": "ACTUAL",
    "Series_title_4": "N/A",
    "Series_title_5": "N/A"
```

3. Data Visualization Implementation

- Developed the visualization module using ReactJS as frontend programming language and leveraging the benefits of Chart.js API in creating the charts.
- Connected the data visualization tool to the database: Configure the data visualization tool to connect to the database and retrieve the required data. Created interactive and engaging data visualizations based on the organization's needs and requirements.

Performance Evaluation and Testing

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 (~60mb)
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.

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.

2. **Visualization Testing:**

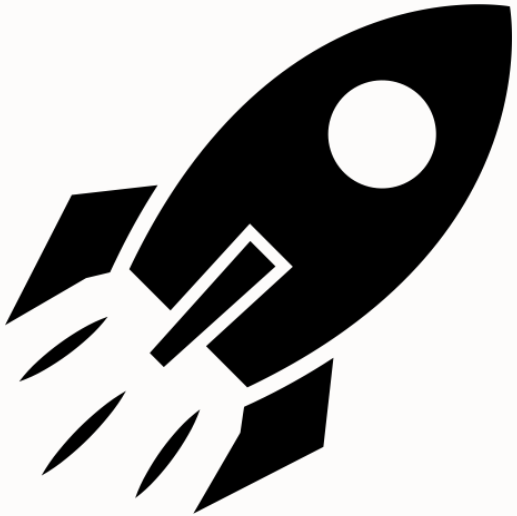
- Functional Testing: Verified that the visualization tool works as expected and displays the data correctly. Interactive charts are being rendered based on user selection of fields of dataset
- Performance Testing: Verified that the visualization tool can handle large amounts of data and respond quickly after any change

3. **Login/Registration Testing**

- Here, our aim is to test the Register and Login Functionalities with various combinations of test cases

4. **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.



Results and Deployment

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.
3. **Enhanced Efficiency:** The ETL application automates the data processing, reducing manual effort and errors.
4. **Better Decision Making**: With accurate, standardized data and powerful visualization tools, stakeholders can make more informed decisions.

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

Security Aspects

Here are some key security aspects to consider for ETL (Extract, Transform, Load), data standardization, and visualization application:

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.
5. **Secure Data Storage:** Safeguard standardized data by applying appropriate access controls, encryption, and monitoring techniques to the storage systems where the data resides.
6. **Secure Session Management:** Implement secure session management techniques, including session expiration, session token management, and protection against session hijacking

Future Aspects

In conclusion, an ETL 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. 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.

INDIVIDUAL

Individual Aim and Objectives

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

Following are the module information correlated with team members along with individual roles and responsibilities :-

- **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.

Publication details

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>

Thank you!!