ETL PIPELINE

Mid-Term Presentation
Domain of Project - Software Engineering and Database Management
Industry Supported - Symphony Tech

PRESENTED & DEVELOPED BY :

UTKRIST AGRAWAL
1032190030
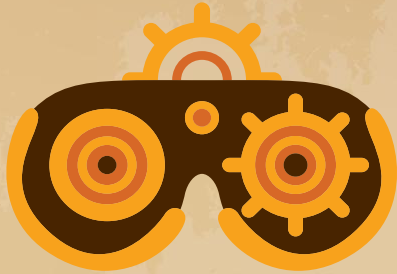PB - 03

MEHUL PANSARI
1032190041
PB - 04

PRABHAT PANWAR
1032190048
PB - 05

NIKUNJ PADIA
1032190109
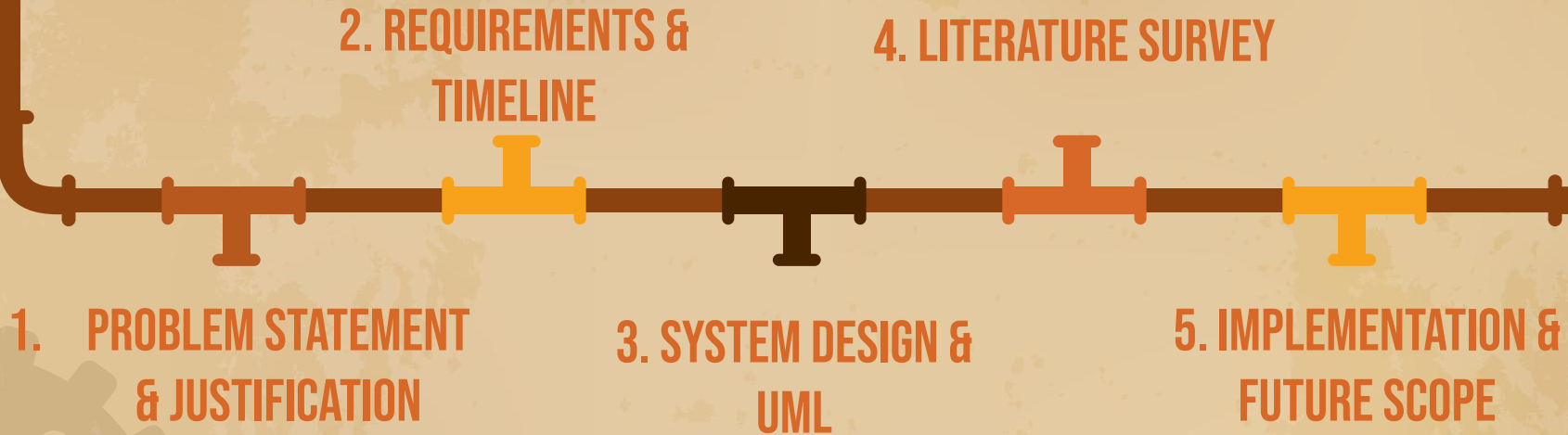PB - 11

GUIDED & MENTORED BY :

EXTERNAL GUIDE:
MR. AJAY GHATPANDE &
MR. RAVI KHARE

INHOUSE GUIDE:
DR RASHMI PHALNIKAR

CONTENT

2. REQUIREMENTS & TIMELINE

4. LITERATURE SURVEY

1. PROBLEM STATEMENT & JUSTIFICATION

3. SYSTEM DESIGN & UML

5. IMPLEMENTATION & FUTURE SCOPE

01

PROBLEM STATEMENT & JUSTIFICATION

STANDARDIZATION OF ETL

The application enables the user to import data from any datasource 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, SQL, etc,and create interactive dashboards, all in one application for Symphony Tech.

OBJECTIVE

1. To research and arrive at Proof of Concept for standard ETL application.
2. To perform technical assessment of various web technologies along with feasibility study of data warehouse for 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.

02

REQUIREMENTS & TIMELINE
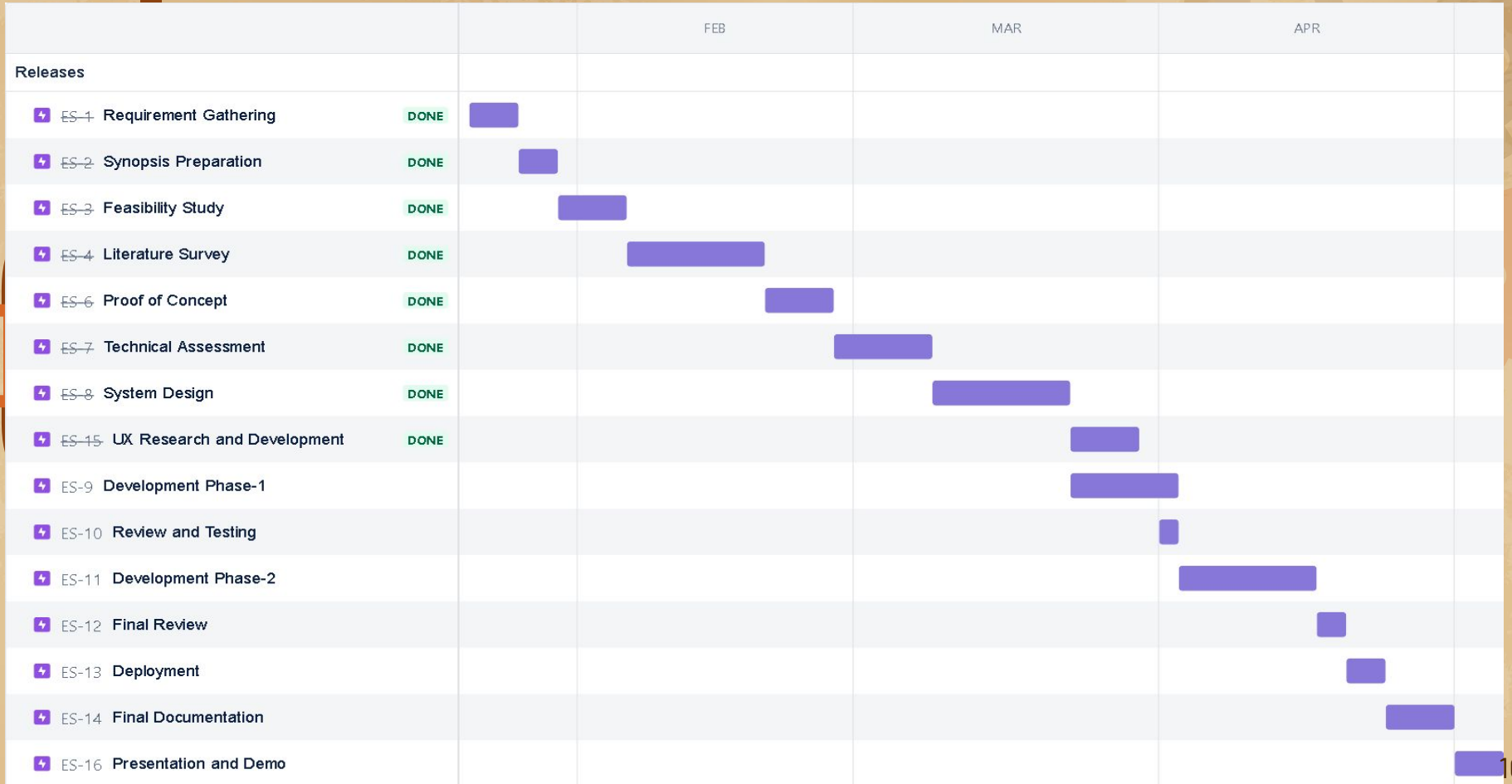
REQUIREMENTS

FUNCTIONAL

- The ETL (Extract, Transform, Load) standardization application uses Django as the back end and ReactJS as the front-end.

- 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 chat APIs, while the Django back-end handles the data processing and storage

REQUIREMENTS

NON-FUNCTIONAL

- The application allows users to easily manage and organize their data, making it a valuable tool for businesses and organizations that need to extract insights from any amount of data and from standard data sources.

- The application needs to be scalable 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 organization's data requirements and flexible enough to adjust to changing requirements in future.

TIMELINE GRAPH

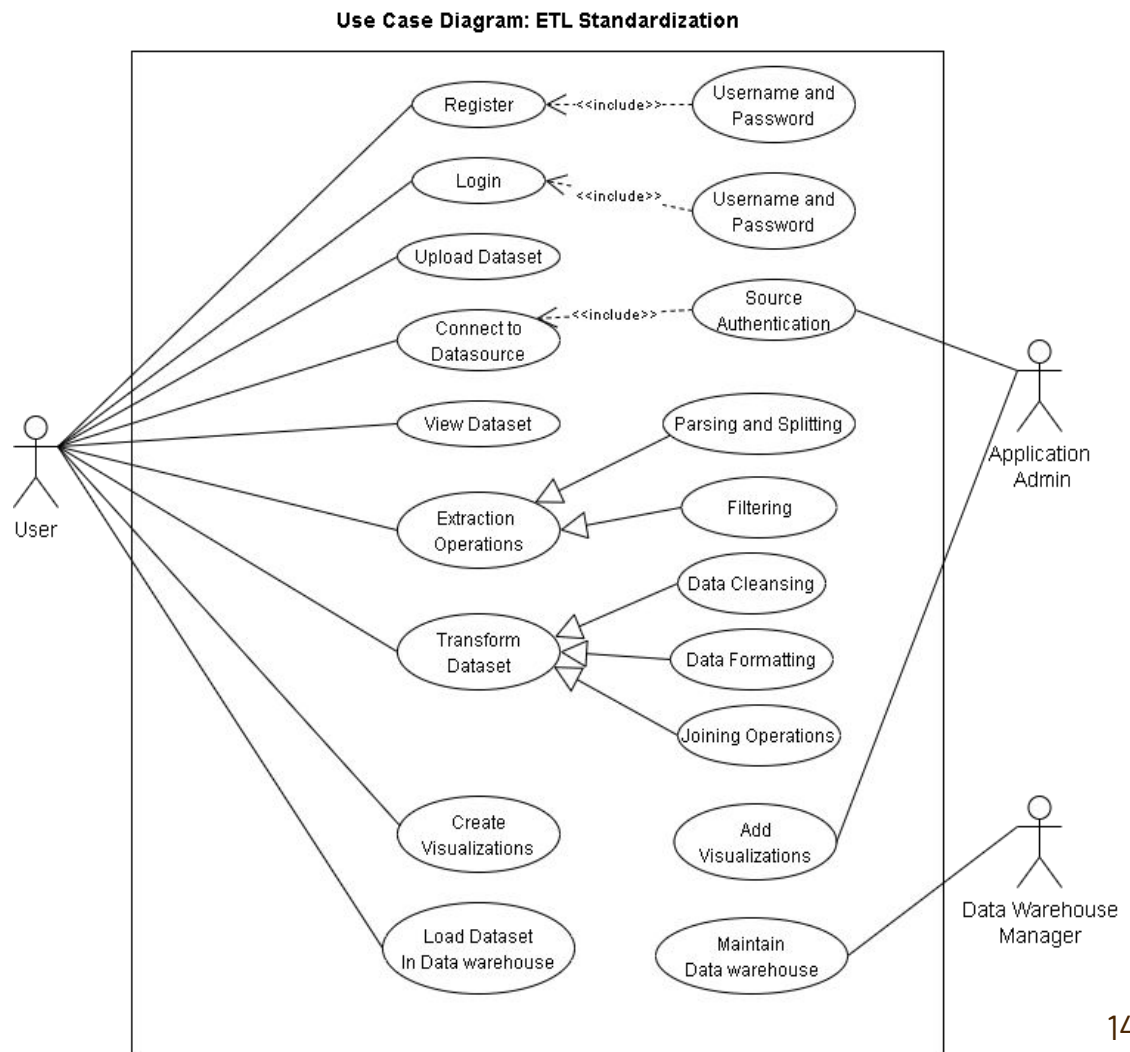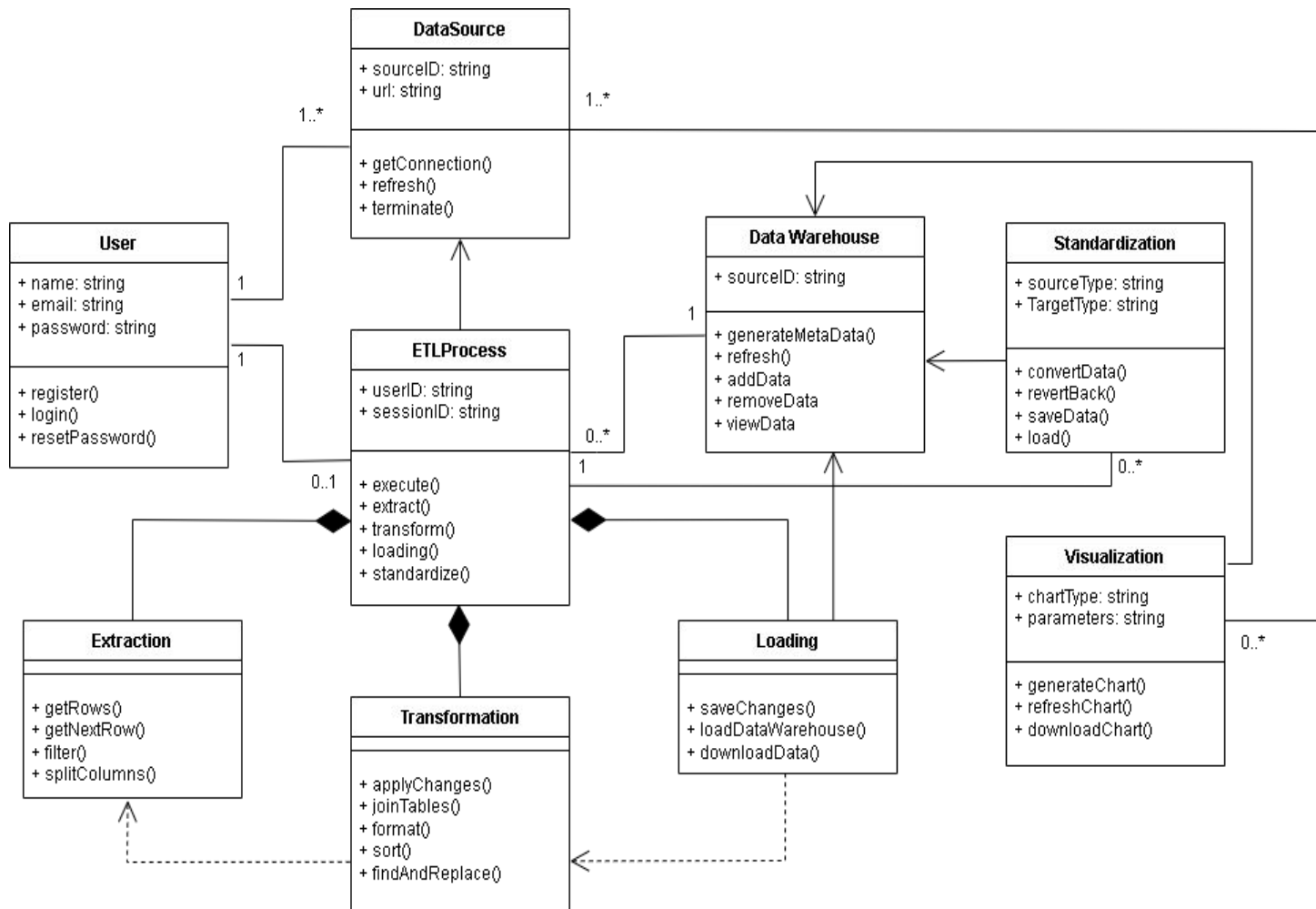|		FEB	MAR	APR	
Releases					
⚡ ES-1 Requirement Gathering	DONE	▮			
⚡ ES-2 Synopsis Preparation	DONE	▮			
⚡ ES-3 Feasibility Study	DONE	▮			
⚡ ES-4 Literature Survey	DONE	▮			
⚡ ES-6 Proof of Concept	DONE		▮		
⚡ ES-7 Technical Assessment	DONE		▮		
⚡ ES-8 System Design	DONE		▮		
⚡ ES-15 UX Research and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

03

SYSTEM DESIGN & UML

SYSTEM ARCHITECTURE FOR ETL STANDARDIZATION

Use Case Diagram: ETL Standardization

USE CASE DIAGRAM

14

CLASS DIAGRAM

DataSource

+ sourceID: string
+ url: string

+ getConnection()
+ refresh()
+ terminate()

User

+ name: string
+ email: string
+ password: string

+ register()
+ login()
+ resetPassword()

ETLProcess

+ userID: string
+ sessionID: string

+ execute()
+ extract()
+ transform()
+ loading()
+ standardize()

Data Warehouse

+ sourceID: string

+ generateMetaData()
+ refresh()
+ addData
+ removeData
+ viewData

Standardization

+ sourceType: string
+ TargetType: string

+ convertData()
+ revertBack()
+ saveData()
+ load()

Visualization

+ chartType: string
+ parameters: string

+ generateChart()
+ refreshChart()
+ downloadChart()

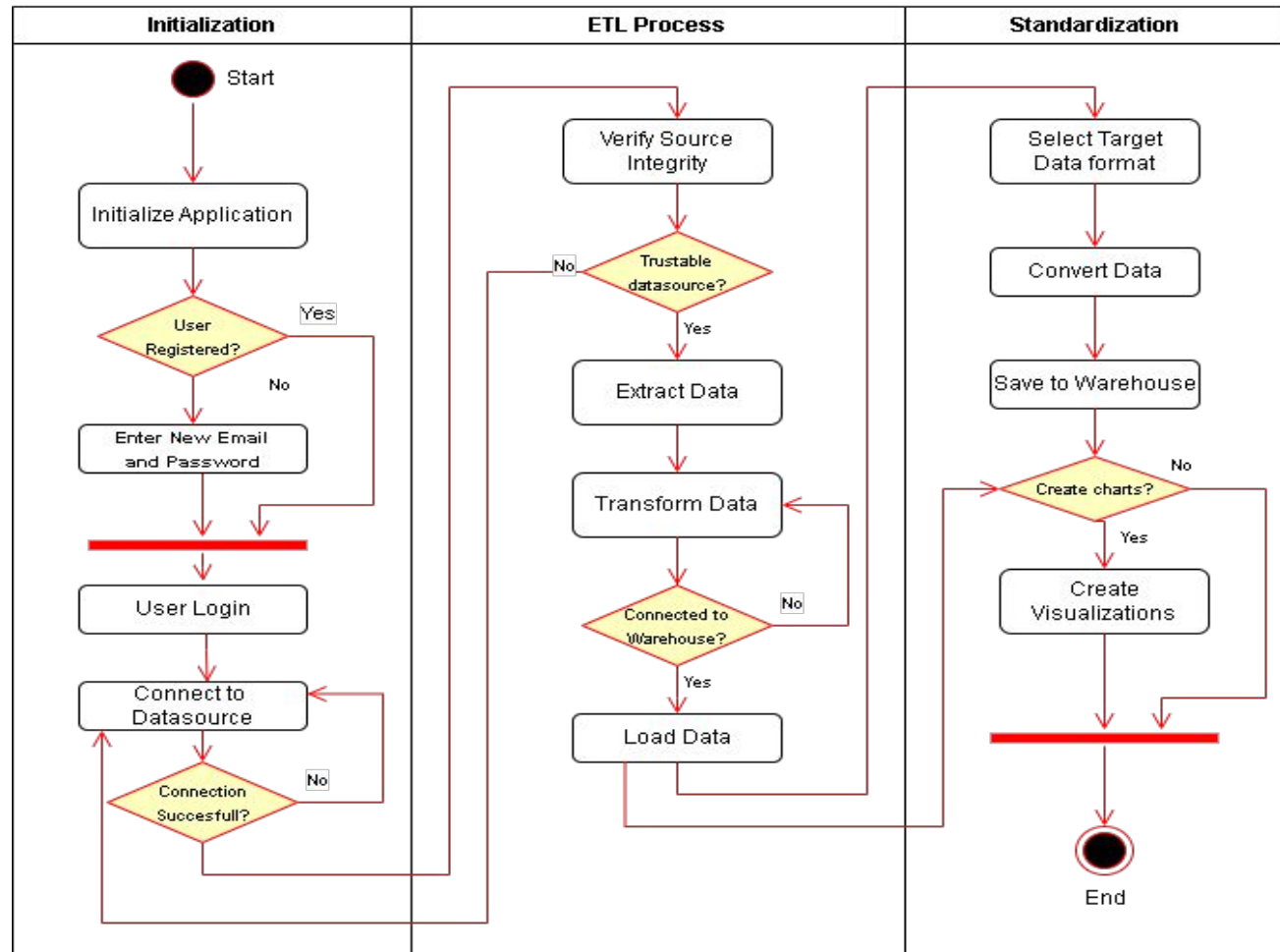**Extraction**

+ getRows()
+ getNextRow()
+ filter()
+ splitColumns()

Transformation

+ applyChanges()
+ joinTables()
+ format()
+ sort()
+ findAndReplace()

Loading

+ saveChanges()
+ loadDataWarehouse()
+ downloadData()

15

SEQUENCE DIAGRAM

16

ACTIVITY DIAGRAM

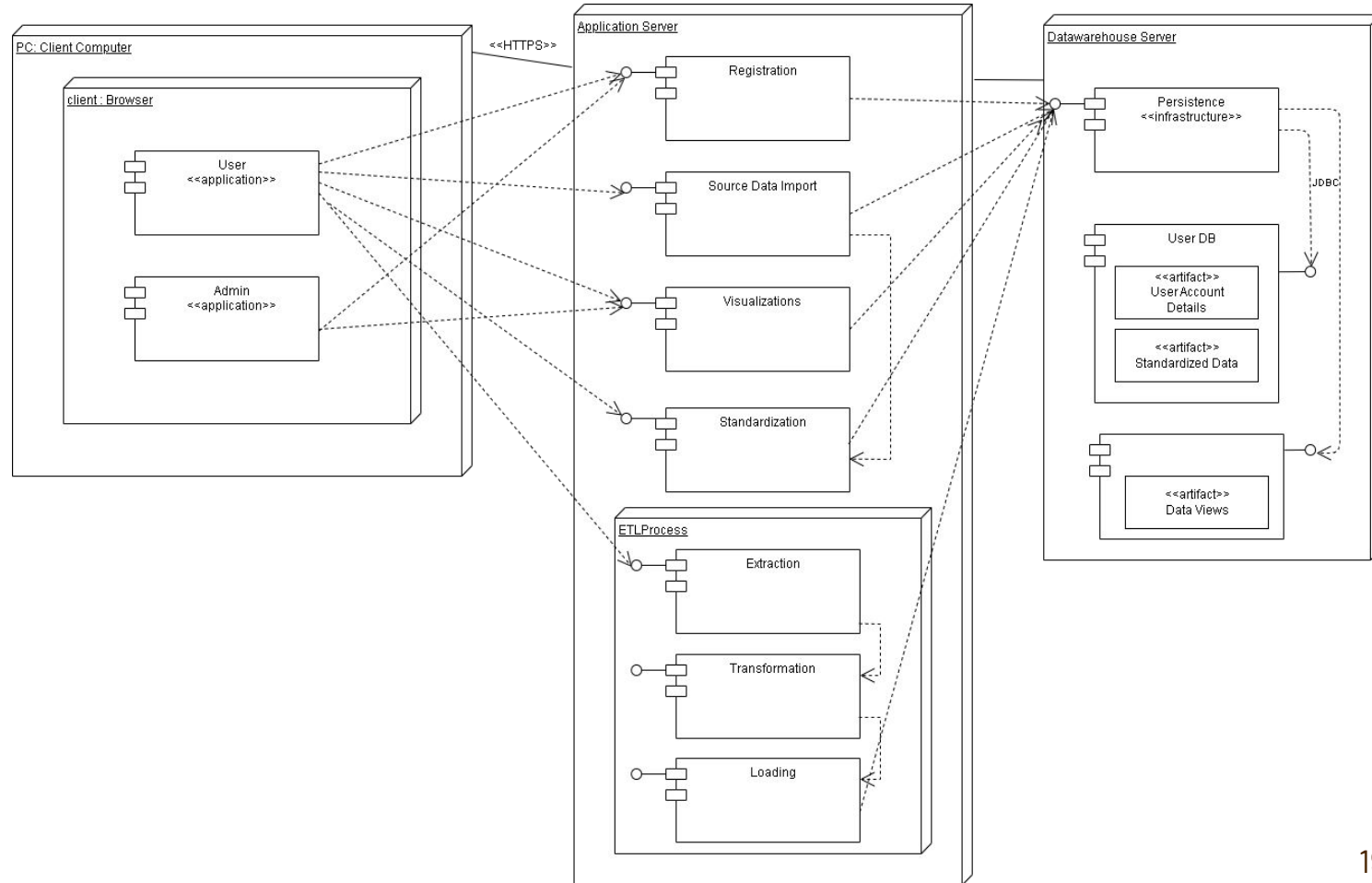Activity Diagram for ETL Standardization

17

COMPONENT DIAGRAM
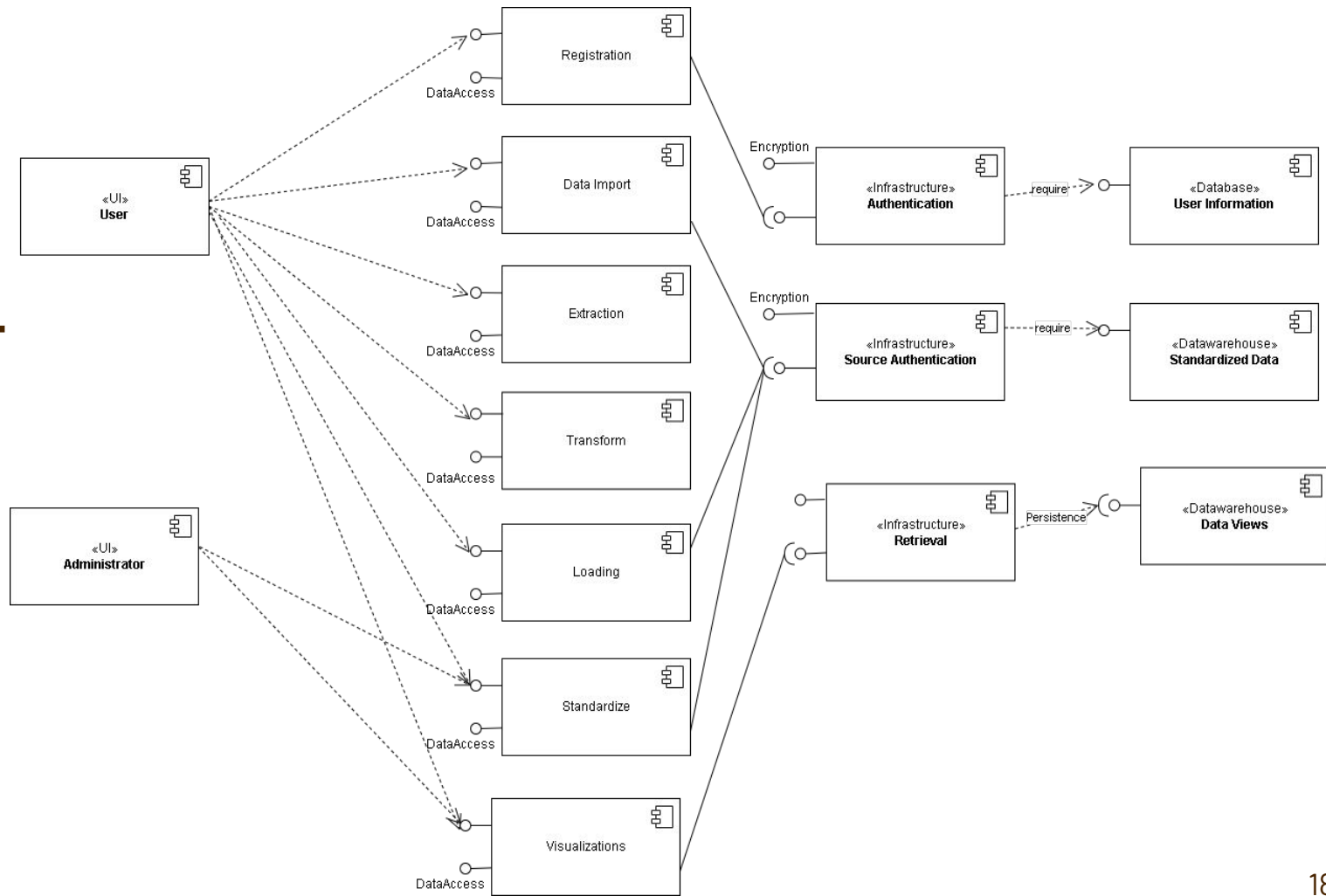
DEPLOYMENT DIAGRAM

04

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.

21

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.

05

IMPLEMENTATION & FUTURE SCOPE

DEMO FLOW

DEMO 1 — Wireframe & Prototype

Implemented website Walkthrough — **DEMO 2**
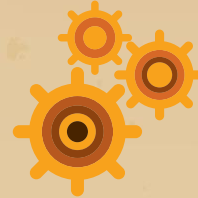
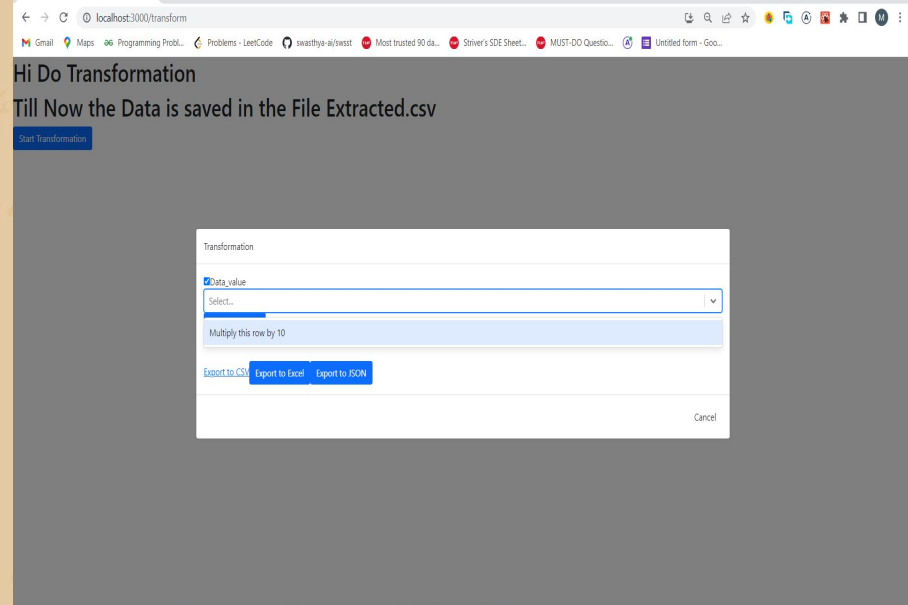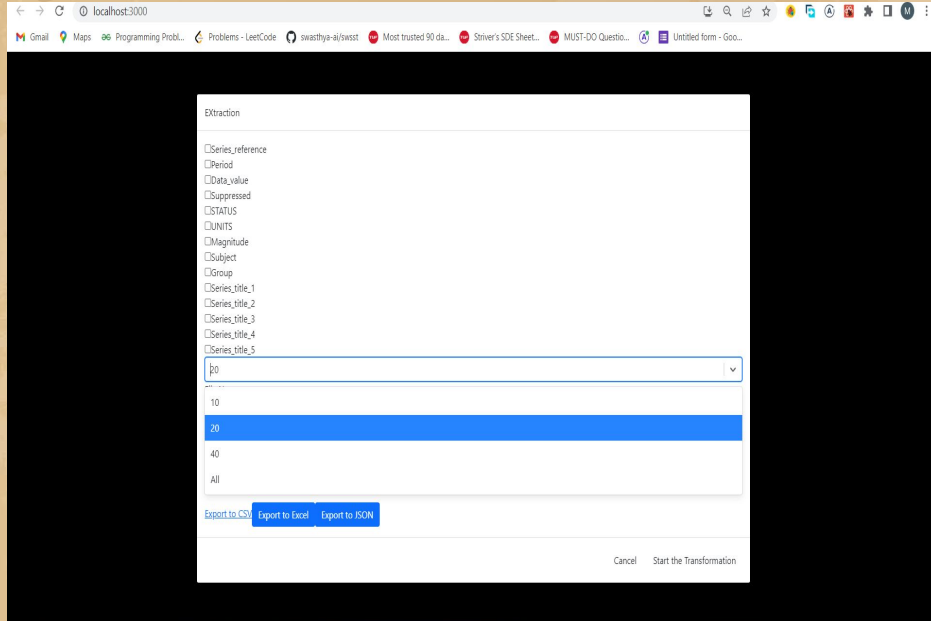**FUTURE SCOPE** — Future work demonstration

WIREFRAME SCREENSHOTS

CODE IMPLEMENTATION

Temp   College Stuff   Others

ETL
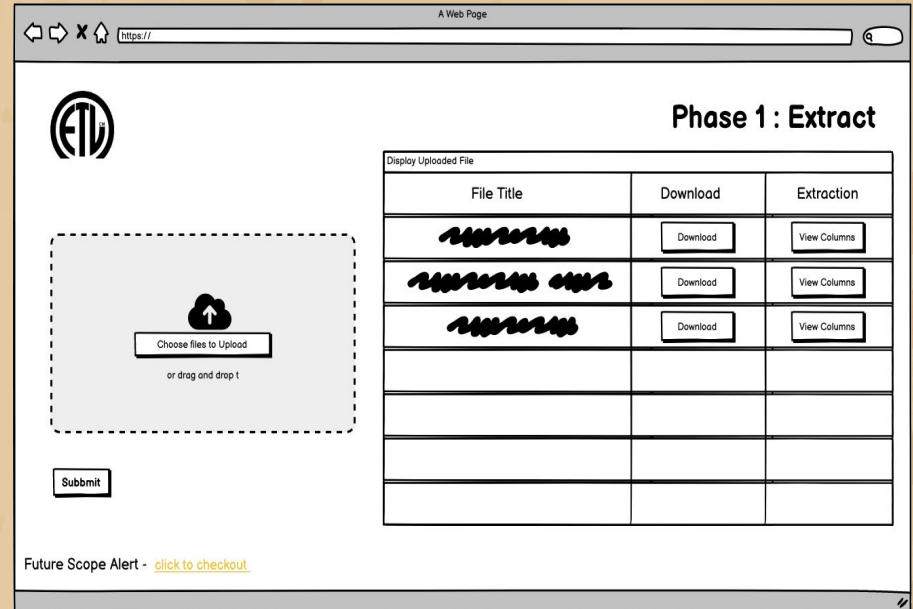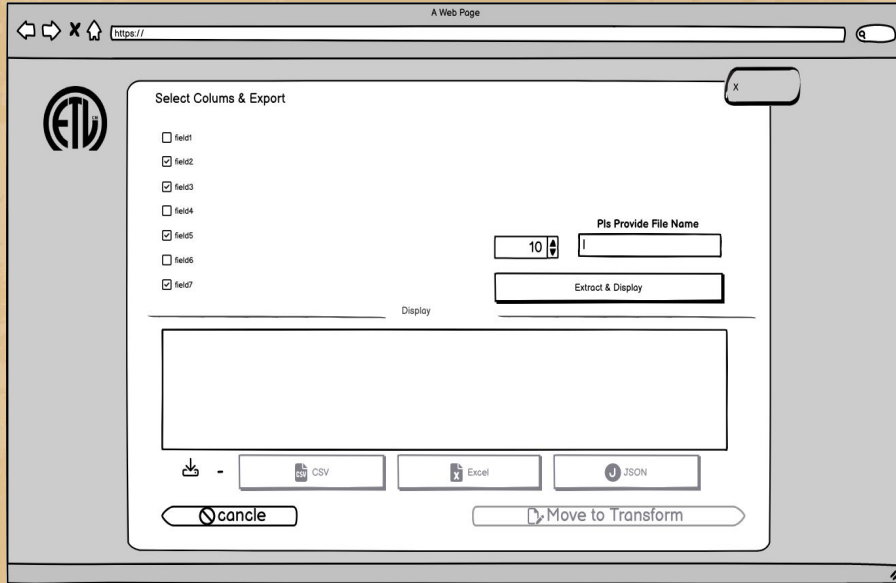
mehul272 Merge pull request #6 from mehul272/enhancement/MovingButtons ...   b2a4e8d last week   🕐 27 commits

📁	Lib/site-packages	Initial Commit	2 weeks ago
📁	Scripts	Initial Commit	2 weeks ago
📁	django_ractjs	The Beginning Of TRANSFORMATION	last week
📄	pyvenv.cfg	Initial Commit	2 weeks ago
📄	requirements.txt	Initial Commit	2 weeks ago

⭐ 0 stars

👁 1 watching

🍴 0 forks

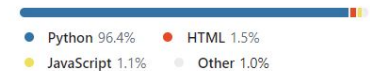Report repository

Releases

No releases published

Packages

No packages published

Contributors 3

mehul272

Utkrist1508 Utkrist Agrawal

Mehul2718

Languages

● Python 96.4%   ● HTML 1.5%
● JavaScript 1.1%   ● Other 1.0%

Temp   College Stuff   Others

```javascript
37
38    const handleDoTransformation = () => {
39      updateModal(false);
40      isTransformation(true);
41    };
42
43    let api = "http://127.0.0.1:8000/api";
44
45    const handleFilterColumnNames = (event, option) => {
46      const isChecked = event.target.checked;
47      const isIncluded = columnNamesArray.includes(option);
48
49      if (isChecked && !isIncluded) {
50        columnNamesArray.push(option);
51      } else if (!isChecked) {
52        columnNamesArray.splice(columnNamesArray.indexOf(option), 1);
53      }
54
55      setColumnNamesArray(columnNamesArray);
56    };
57
58    const handleExtraction = async () => {
59      await axios
60        .get(
61          api + `/filter-files-data/${title}`,
62          {
63            params: {
64              stringArray: JSON.stringify(columnNamesArray),
65              numRows: numRows,
66            },
67          },
68          url
69        )
70        .then((res) => {
71          setData(res.data.result);
72        });
```

32

INDIVIDUAL CONTRIBUTION AND ROLES

Name	Role	Survey Objective	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 and APIs, Application Frontend
Utkrist Agrawal	UI/UX and Frontend developer, QA tester	Different data formats and representations, Data transformation techniques and best python libraries	Standardization module and UI Wireframes
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

REFERENCES

1) S.Sajida, Dr.S.Ramakrishna, 2015, A Study of Extractâˆ'Transformâˆ' 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.

REFERENCES

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/IJIIDS.2012.047003

15) Ankorion, 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.

THANK YOU !!