# Comprehensive Design Flow for Effective Data Processing and Visualization

Samudyata A
Department of Electronics and Telecommunication Engineering
R V College of Engineering
Bangalore, India
*Email: samudyata.a [AT] gmail.com*

P Nagaraju
Department of Electronics and Telecommunication Engineering
R V College of Engineering
Bangalore, India
*Email: nagarajup [AT] rvce.edu.in*

***Abstract*** **- Data is pervasive in the modern digital world. Data is continually being created, gathered, and analysed, from social media interactions to financial transactions. Businesses face both opportunities and challenges due to the data boom. On one hand, data may be a useful resource for companies looking to differentiate themselves from the competition, spur innovation, and enhance decision-making. However, the sheer amount and complexity of data can be daunting, making it challenging to glean insightful information and take timely action on it. Organisations need a strong and well-designed data processing design flow that can allow effective data exploration and creative visualization in order to meet this challenge. This work focuses on the complete data processing lineage and integrating that with various tools such as Talend and Power BI for data analytics to deliver diverse visualizations and insights which can cater businesses to stay ahead of the curve, providing a foundation for success in the ever-evolving digital landscape.**

***Index Terms – Talend, Power BI, Data Analytics, Exploration***

## I. INTRODUCTION

Data is the new oil for organization engines. Any organisation that is driven by data must possess solid data architecture. It entails creating, putting into place, and maintaining a data management framework that makes it possible to handle, store, and analyse data effectively. Businesses may gain useful insights and trends from their data to make smart decisions and remain ahead of the competition through setting an appropriate design pipeline for data in place. This necessitates the use of sophisticated analytical tools and methodologies to derive insights and make predictions, as well as reliable and well-designed system architecture that can support the efficient processing and storage of data.

The proposed system aims to help businesses leverage the use of various tools available for data processing and visualization.

Data as it is unstructured when obtained as input from various sources. The complete processing of data is performed stage by stage to get the structured well formatted data for analytics and visualization. For this purpose, data has to go through a series of steps that includes Extract, transform and Load (ETL). Later the required Key Performance Indicators (KPI) based on business needs must be catered to perform analytics and visualization. To achieve the contemplated outcomes the system uses the following:

1) Talend:

Leading data integration platform Talend offers a full range of tools for developing, evaluating, and implementing data integration workflows. It is a potent open-source solution that enables organizations of all sizes to swiftly and effectively combine data from diverse sources.

Talend's approach to data integration is mostly based on the Extract, Transform, and Load (ETL) methodology. With this strategy, data is extracted from numerous sources, formatted as required, and loaded into the target system. Data quality technologies from Talend work in tandem with its ETL capabilities to guarantee the accuracy and completeness of data throughout the integration process.

Data from databases, flat files, cloud-based programs, and big data platforms can all be processed using Talend's data integration solutions. A range of data integration use cases, including data warehousing, data migration, and data synchronization, are also supported by the platform. For this system, Talend serves as the main integration and transformation tool.

2) Power BI

Microsoft's Power BI is a service for business analytics that offers interactive visualisations and business intelligence features with a user interface that is easy enough for end users to utilise to build their own reports and dashboards. With a user interface that is simple enough for end users to utilise to create their own reports and dashboards, Microsoft's Power BI tool for business analytics includes immersive illustrations and business intelligence features. . It helps to obtain the business insights by figuring out the positive and negative growth rates of the business and monitoring the business performance in the long-run. It furthermore forecasts its future impacts, and analyses the performance of the business [1]. With Power BI, innovative visualizations can be designed for the system.

## II. LITERATURE SURVEY

Jijun et al. [2] propose a multi-layer power enterprise data management architecture based on big data, addressing challenges in data collection, storage, processing, and analysis in the power industry. Mao et al. [3] discuss multi-dimensional data distribution monitoring using OLAP techniques, highlighting the benefits of OLAP in analyzing large volumes of data and presenting a monitoring framework. Yi and Ye [4] present a practical workflow for cleaning master data, addressing the challenges and providing step-by-step guidance for data profiling, cleansing, and validation. Kaminskyi and Nehrey [5] propose an information technology model for CRM in nonbank lenders, integrating customer profitability analysis and risk assessment to support effective customer relationship management.

Sreemathy et al. [6] focus on data validation in ETL (Extract, Transform, Load) using Talend, providing insights into the importance of data validation and discussing the role of Talend as an ETL tool. Riasetiawan et al. [7] discuss data modeling and analysis for big data mapping and management in the energy data platform, offering insights into the challenges and techniques involved in managing big data in the energy sector. Sreemathy et al. [8] explore data integration in ETL using Talend, highlighting the significance of data integration and discussing Talend as a tool for ETL processes. Sreemathy et al. [9] provide an overview of ETL tools and specifically focus on Talend as a data integration tool, discussing its features and benefits in the context of ETL processes.

Diouf et al. [10] review the variety of data in ETL processes during cloud migration and validation, providing an overview of the state-of-the-art techniques and challenges in handling diverse data types. Balakrishnan et al. [11] propose a cloud computing-based approach for data validation and migration in ETL using Talend, emphasizing the advantages of cloud computing and Talend for efficient data management. Akbar et al. [12] discuss the implementation of business intelligence for visualizing transaction data analysis using a dashboard system in a case study of a convenience store, highlighting the benefits of business intelligence tools in data analysis. Zhang et al. [13] provide a review of data quality management for big data, discussing the challenges, techniques, and best practices in ensuring data quality in the context of big data analytics.

## III. PROPOSED SYSTEM ARCHITECTURE

In today's data-driven economy organisations are finding it increasingly difficult to successfully manage and use their data assets. Data is being collected from several sources, is being kept on various platforms and systems, and comes in a variety of formats and quality levels. As a result, organisations frequently have trouble making informed decisions, collaborating across teams, and staying in compliance with regulatory standards. This frequently results in inconsistent, incomplete, or erroneous data. In order to acquire insights, enhance operations, and stimulate business growth, organisations are realising the importance of a well-designed data architecture that can offer a foundation for efficient data administration and utilisation. Figure 1 represents the proposed system architecture.

### *A. Data Sources*

Loads of voluminous data is gathered from numerous sources such that these external data is made available to ingestion layer of architecture.

There can be two main types of data sources:

The internal source: These involve data controlled and owned within the organisation. These can include various relational and non-relational databases, business required schemas and information like customers details, product descriptions, sales, web analytics and information from various plants present all over the world.

The second source is external source which are not in control of organisation. It may include from other businesses posing requirements or those which are in collaboration with business firms. The open source data from government, twitter or facebook and so on might be required for market research, trends analysis and to perform various other analytics based on metrics.

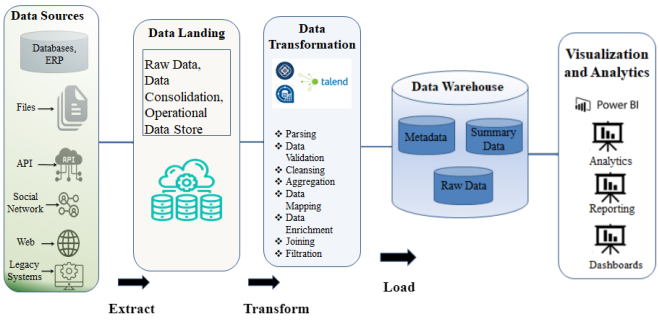


Fig 1. System Architecture

### *B. Data Landing*

The extracted data from various sources will be unprocessed and this data has to be ingested to data warehouse. Before that the vast amount of data is stored in the data lake and further process is done to refine the data to make it available for operational use. For this purpose an Operational Database Store (ODS) which is a centralised database is used. It serves as an intermediate between the data landing and operational systems in an organization. The ODS follows a structured and refined data model compared to raw data in landing.

The landing process can involve various methods like data validation, metatdata tagging and extraction of data.

Validation of data ensures the data that is fed into the system is meeting certain quality requirements like it needs to be concise, precise, accurate and reliable. To facilitate data discovery and organization, descriptive information related to the data, such as timestamps, data sources, or data formats are attached as part of metadata tagging.
In short the landing data contains all the data including historical information whereas the ODS consists only of the most recent data required.

*C. Data Transformation*
The data stored in ODS is subjected to various transformations and processing as per the business requirements. It can include numerous transformations like cleansing, filtering, formatting, enrichment, aggregation and mapping. These transformations are performed using the Talend tool which is an ETL software written in java.

*Extract, transform and Load*
The ETL process stands for Extract, Transform and Load. It is a method of unification and warehousing of data. Data is extracted from numerous sources, transformed to fulfil specific criteria, and loaded into a target system or database for analysis, reporting, or other reasons. ETL is critical in organisations for guaranteeing data quality, consistency, and availability for decision-making.

*Extraction:*
As shown in figure 2 data is extracted from various sources, including databases, flat files, cloud-based apps, web services, and APIs, during the extraction stage, the first step in the ETL process. Different formats and storage places for the data are both possible. The subsequent temporary storage space known as a staging area is where the data that was extracted is kept.

*Transformation:*
The extracted data is then transformed into a format that the target system may use in the following stage. The transformation process includes data integration, enrichment, normalisation, and validation. Data cleansing requires removing superfluous data, data integration entails merging data from many sources, data enrichment entails adding new data to the extracted data, data normalisation entails standardising data, and data validation entails verifying the data for mistakes and inconsistencies.

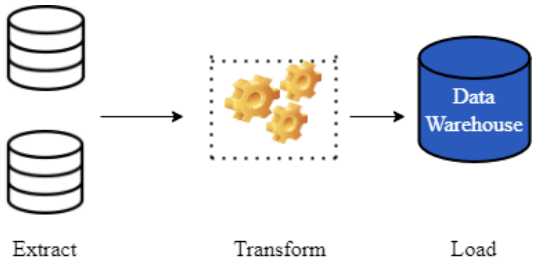

Fig 2. ETL pipeline

*Loading:*
The modified data is then imported into the target system, such as a data warehouse, a data mart, or a business intelligence tool, during the loading stage, which is the last stage. The mapping of the modified data to the destination system, data integrity checks, and error handling are all part of the loading process.
A graphical interface is offered by ETL systems like Talend for building ETL procedures. These processes can be created to automatically do ETL, including data extraction and loading. Furthermore, Talend offers a large variety of connectors that enable data to be extracted from a variety of sources, including databases, flat files, cloud-based programmes, and online services. Data mapping, data conversion, and data aggregation are a few examples of data integration techniques that can be used to alter the data. Finally, utilising connectors like JDBC, ODBC, or SOAP, the modified data can be loaded into the destination system.

Numerous transformations and modifications can be performed using Talend Open Studio. They are as follows:

1. Filtering: The data can be filtered based on specific criteria using components like "tFilterRow" or "tMap" as shown in figure 5 with expressions. This allows excluding or including certain rows in the dataset.
2. Sorting: Talend provides sorting components like "tSortRow" or "tMap" with sorting functions to sort data based on one or multiple columns. Ascending or descending order can be chosen for sorting.
3. Aggregation: Talend offers aggregation components such as "tAggregateRow" or "tAggregateSortedRow" that allows performing aggregations like sum, average, minimum, maximum, count, etc., on groups of rows in the dataset. Joining: Talend supports various types of joins, including inner join, left outer join, right outer join, and full outer join. Components like "tJoin" or "tMap" can be used to join data from multiple sources based on common keys.
4. Splitting and Merging: Talend allows splitting data into multiple streams using components like "tSplitRow" or "tFilterRow." Conversely, data can be merged from different sources using components like "tUnite" or "tJoin."

5. Data Conversion: Talend provides a range of data conversion functions to transform data types. For instance, a string can be converted to numeric value or vice versa, the date format can be altered, or there can be changeover between various date and time formats.
6. String Manipulation: There are multiple functions related to string manipulation in Talend for jobs like extracting substrings, concatenating strings, replacing characters, converting case (uppercase/lowercase), and more. Components like "tExtractRegexFields" or "tMap" can be used for string manipulation.
7. Data Cleansing: Talend provides components like "tDataMasking" or "tStandardizeRow" to clean and standardize data. Duplicates, trim leading or trailing spaces can be removed, data can be normalized, or custom cleansing rules can be applied.
8. Outliers Data handling: Handle values that are missing in customer and product data by either deleting entries with erroneous values or imputing missing values depending on certain criteria or algorithms.
9. Standardisation: To maintain consistency and increase data quality, standardise customer and product data elements such as addresses, phone numbers, and product names.
10. To ensure data integrity and avoid redundancy, identify and eliminate redundant data within the datasets.
11. Data Splitting and Denormalization: Talend allows splitting and denormalizing data using components like "tDenormalize" or "tNormalize." This is useful when it is needed to transform data from a normalized format (multiple related tables) to a denormalized format (single table).
12. Categorical Encoding: Use techniques such as one-hot encoding or label encoding for encoding categorical parameters in consumer and product information, including gender, categories of products, or customer segments.
13. Feature Engineering is the process of developing derived features or qualities from customer and merchandise data in order to collect more data or trends that may be used to improve research or models based on machine learning. Calculating client age from birth dates, for example, or producing product characteristics based on existing qualities.
14. Custom Transformations: Talend also supports custom transformations where own code can be written or user-defined functions (UDFs) in Java, SQL, or scripting languages like Perl or Python can be used. This gives flexibility to perform complex transformations based on specific requirements.

The tmap is a key component used for transformation as shown in figure 3.

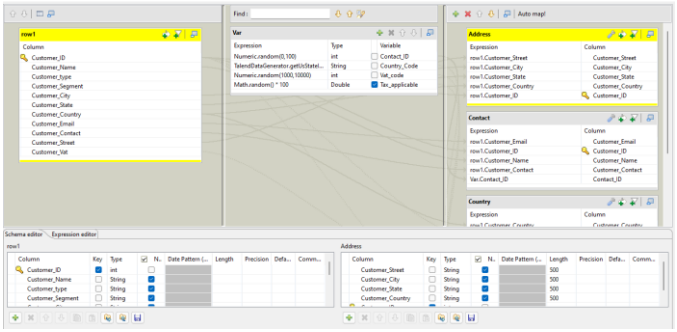

Fig. 3 tMap Component

The flow diagram in figure 4 illustrates how a talend job is proceesed to integrate data from multiple sources. The raw data or unstructured data is acquired and preprocessed. This data is then checked for reliability ,which is whether it is a valid data or not. In case it appears as erroneous data then it will be reverted back to the preprocessing stage. Otherwise, if data is not prone to any errors then it will be stored in operational data store. ODS can be any database like MySQL, Oracle , MongoDB and so on. This reposited data is pulled from the database to the Talend. There are many features supported by Talend for instance the database connection can be established by providing the credentials that include hostname, username and password or if it's a cloud database then URL or an endpoint address that identifies the database's position in the cloud is provided. The URL normally contains more details such as address of the server, port number, and any extra characteristics unique to the cloud database service. Components tPrejob and tPostjob are components used to carry out particular tasks or operations both prior to and following the main job execution. Within the prejob the database connection can be established for instance which ensures that the main job will not execute until the connection is established successfully. Next, the main job begine the exection with already being connected to the database and required transformations can be performed using tMap. Once the job finishes the connection is disconnected from the database.

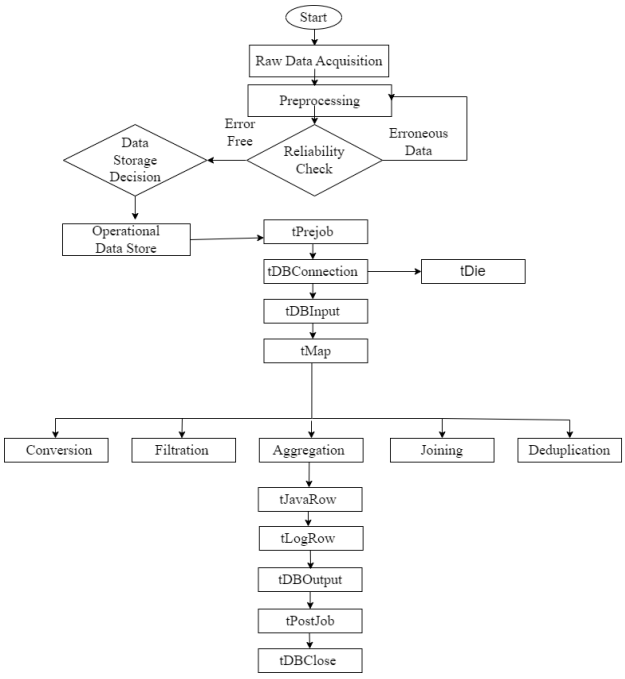


Fig. 4 Flowchart

*D) Data Warehouse*

A data warehouse is a central repository that consolidates and organizes data from various sources within an organization. It serves as a foundational component in the design flow for processing data from sources to visualization. A data warehouse provides numerous benefits, including improved data quality, enhanced decision-making, and simplified data analysis.

The purpose of a data warehouse is to create a unified and consistent view of an organization's data. It involves extracting data from operational systems, transforming it into a standardized format, and loading it into the data warehouse. This process ensures that data from disparate sources is integrated and harmonized, enabling efficient and accurate analysis.

The data warehouse acts as a single source of truth, providing a comprehensive and historical view of data. It allows users to perform complex queries and analysis across different dimensions, such as time, geography, or product lines. With its optimized structure and indexing, a data warehouse enables fast retrieval of data, facilitating timely decision-making.

Moreover, a data warehouse supports data quality initiatives. It provides mechanisms for data validation, cleansing, and enrichment, ensuring that the data stored is accurate, complete, and consistent. This high-quality data serves as a reliable foundation for analytics and visualization.

In the design flow for processing data from sources to visualization, the data warehouse plays a critical role. It serves as the central repository where data is stored and made available for further processing, analysis, and visualization. By leveraging a data warehouse, organizations can unlock the full potential of their data, gain actionable insights, and make informed decisions that drive business success.

*E) Analytics and Visualization*

In this stage the data is retrieved from the data warehouse and multiple views are created on top of it. To communicate the insights and findings effectively Power BI, the most sought tool is used for this purpose. Various analyses can be evaluated using several functions and queries. These are some of the things carried out in this stage:

*Data Analysis:* The transformed data is subjected to a variety of mathematical techniques and algorithms in order to find patterns, correlations, and trends. This could include:

*Descriptive statistics*: They are used to compute metrics such as mean, median, mode, variance, and so on. Conducting hypothesis testing, confidence intervals, and regression evaluation are all examples of inferential statistics.

*Time Series analysis*: Time series analysis is the process of analysing data and anticipating future values based on time-dependent trends.

*Data Visualisation:* The analysed data is visually portrayed to aid comprehension and interpretation. Creating charts, graphs, dashboards, and other visualisations is part of this. Among the most common visualisation approaches are:

a) Demonstrating patterns of distribution, trends, and proportions using visualisations such as bar charts, line charts, and pie charts.
b) Scatter plots and bubble charts are tools for displaying relationships between variables.
c) Heatmaps and treemaps are tools for visualising hierarchical and multidimensional data.
d) Geographic maps are used to display spatial data and geographic trends.
e) Dashboards that allow users to explore and engage with visualisations are known as interactive dashboards.
f) Infographics and storyboards: Narrative presentations of facts and insights.

Analytics and visualisations aid in the extraction of relevant insights from data, allowing for data-driven decision making. These insights can help with strategic planning, identifying opportunities, addressing difficulties, and making evidence-based decisions across several areas.

*Report Generation:* To relay the outcomes to stakeholders, summarised discoveries, analysis results, and visualisations are frequently put into reports or presentations. These documentations can be in PDF, PowerPoint, or dynamic web-based dashboard formats.

*Iterative Analysis and Refinement:* The analytics and visualisation process may involve an iterative approach, with the analysis and visualisations being refined and adjusted

based on feedback, extra information or changing requirements.
Finally, these visualizations can be shared by integrating with other systems or results can also be archived.

## IV. CASE STUDY

As an extension to all the processes explained above another use case was considered with respect to customer churn analysis for a telecom dataset.
Churn rate is a key performance indicator (KPI) that measures the rate at which consumers or subscribers terminate their relationship with a company or cease utilizing its products or services over a given time period. It's usually expressed as a percentage. Churn rate is an important number for businesses since it measures client retention and loyalty. A high churn rate shows that a large number of customers are departing the company, which can be detrimental to the company's growth and profitability. It emphasizes the need of focusing on maintaining existing customers and identifying the reasons behind their churn. Companies can detect patterns and trends that may contribute to client attrition by analyzing churn rate. They can then take proactive steps to improve customer happiness, improve their products or services, answer customer concerns, or develop targeted retention initiatives in order to reduce turnover and boost customer loyalty.

For this project, the dataset was obtained from kaggle consisting of 7043 customer's data belonging to a telecom company. The dataset had certain missing values which were handled by averaging and replacing it in there. Next the dashboards were created. The first page of dashboard consisted of all the overview related to demographics and these were created based on the KPI's that include Tenure, Month-to-month contracts, Number of tech tickets, Tech Support, Fibre optic internet service, Online security and Churn Rate.

## V. RESULTS AND DISCUSSION

The interactive dynamic dashboards for customer churn analysis were created taking into account the KPIs. The churn analysis of the dataset used in the telecom industry is shown in Figure 5. There were a total of 7043 customers, it was noted. Nearly 49.5% and 50% of the company's consumers, respectively, were males and women. The fact that only 25% of the company's former customers are seniors suggests that the younger demographic has a greater churn rate. Partner study revealed that 64.21% of people who left had no relationships. Again, the higher turnover rate among younger clients can be blamed for this. Also 87% of the customers had no dependents. All these attribute to demographic information. Furthermore, in the customer accounts overview, Customers who had been with the company for less than 10 months accounted for 49.38% of customer attrition. The majority of clients that left the company had month-to-month contracts. Customers who departed used electronic checks 45.29% of the time, postal checks 19.20% of the time, bank transfers 16.73% of the time, and credit cards 15.25% of the time.

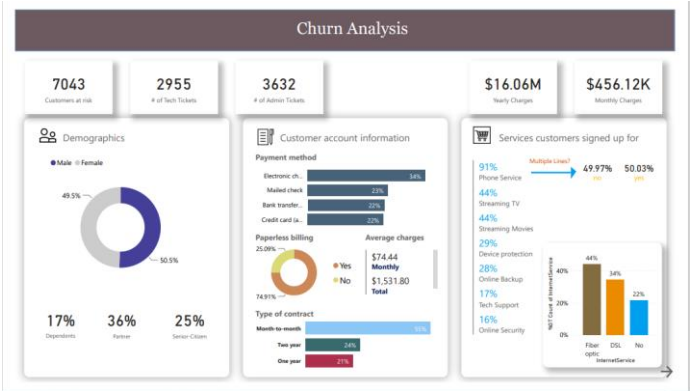


Fig.5. Churn Analysis Dashboard

Customers who have opted to paperless billing accounted for 74.91% of customer attrition. Whereas, customers on month-to-month contracts had the highest average monthly charges, but their overall yearly charges were much lower. Customers with two-year contracts, on the other hand, had the lowest average monthly prices, but their overall expenses were substantially higher.
Customers who departed had 45.48% phone service and numerous lines, while 45.43% had phone service but no multiple lines as shown in figure 6. Nonetheless, they departed. This implies that numerous lines and phone service were not significant determinants of customer attrition. Customers who had fibre optic and streamed movies left the company in 69.4%, those who had DSL and streamed movies in 24.56%, and those who had no internet access but streamed movies in 6.05%.


Fig 6. Customer Risk Analysis

This shows that Fibre optic may be a concern, and churn among clients who have subscribed to it is likely to increase. In terms of the number of tech tickets and tech assistance provided, 77.37% of customers with tech tickets but no tech support left.

For the case study few suggestions are as follows:

1. Pay special attention to new clients (those who have been with company for less than ten months). Offer them services that will entice them to stay with the organisation longer.
2. Make one or two year contracts more appealing to clients than month-to-month contracts so that customers do not feel compelled to leave the company before the contract time expires. This will also encourage them to stay longer, extending their term.
3. Find novel approaches to provide online security to clients while increasing their trust in the systems and the firm as a whole.
4. Employ extra technical personnel to provide adequate technical support to all clients.
5. Provide consumers with a variety of internet services in addition to fibre optic and DSL.

## VI. CONCLUSION

An effective design flow for processing data from many sources and transforming it into relevant visualizations was illustrated in this study. This design strategy helps user or businesses to extract useful insights and effectively communicate discoveries by understanding the peculiarities of data sources, assuring data quality through pretreatment and integration, and utilizing appropriate visualization techniques.

A case study of customer churn analysis was taken as a use case to illustrate the power of visualization tool to analyze data efficiently. Moreover, this approach which combines iterative exploration and refinement of visual designs with interactive tools and modern technology enables academics and practitioners to interactively analyze and explore data, resulting in improved decision-making and data-driven insights. This brief and practical guidance streamlines the workflow of organizations and researchers by using the potential of data visualization.

## REFERENCES

[1] H. Ashok, S. Ayyasamy, A. Ashok and V. Arunachalam, "E-business Analytics through ETL and Self-Service Business Intelligence Tool," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 747-751, doi: 10.1109/ICIRCA48905.2020.9183336.

[2] W. Jijun, C. Yongqiu and C. Li, "Research on Multi-layer Power Enterprise Data Management Architecture Based on Big Data," 2022 IEEE 8th International Conference on Computer and Communications (ICCC), Chengdu, China, 2022, pp. 1194-1198,

[3] Y. Mao, S. Huang, S. Cui, H. Wang, J. Zhang and W. Ding, "Multidimensional data distribution monitoring based on OLAP," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, 2020, pp. 298-302, doi: 10.1109/ITCA52113.2020.00070.

[4] S. Yi and H. Ye, "A Practical Workflow in Cleaning Master Data," 2020 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Exeter, United Kingdom, 2020, pp. 1311-1314

[5] A. Kaminskyi and M. Nehrey, "Information Technology Model for Customer Relationship Management of Nonbank Lenders: Coupling Profitability and Risk," 2021 11th International Conference on Advanced Computer Information Technologies (ACIT), Deggendorf, Germany, 2021,pp. 234-237

[6] J. Sreemathy, S. Priyadharshini, K. Radha, K. Sangeerna and G. Nivetha, "Data Validation in ETL Using TALEND," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Coimbatore, India, 2019, pp. 1183-1186, doi: 10.1109/ICACCS.2019.8728420.

[7] M. Riasetiawan, F. Anggara, A. Ashari, S. Winardi and B. N. Prastowo, "Data Model and Analysis for Big Data Mapping and Management in the Energy Data Platform," 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), Medan, Indonesia, 2021, pp. 1-6,

[8] 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.

[9] 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.

[10] P. S. Diouf, A. Boly and S. Ndiaye, "Variety of data in the ETL processes in the cloud migration and validation : State of the art," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, 2018, pp. 1-5.

[11] M. Balakrishnan, N. M, K. Ramya and S. K, "Cloud Computing based Data Validation and Migration in ETL using Talend," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 1349-1355, doi: 10.1109/ICECA55336.2022.10009262.

[12] R. Akbar, Wahyudi, J. Rahmadoni and H. A. Lubis, "The Implementation of Business Intelligence on Visualisation of Transaction Data Analysis using Dashboard System Case Study: XYZ Convenience Store," 2022 International Symposium on Information Technology and Digital Innovation (ISITDI), Padang, Indonesia, 2022, pp. 42-49, doi: 10.1109/ISITDI55734.2022.9944494.

[13] M. Zhang, J. Zhou, Q. Wang and X. Zhang, "Data Quality Management for Big Data: A Review," in IEEE Access, vol. 7, pp. 70102-70114, 2019