Data Integration and Modelling using Talend

[1]Avanthika S, [2]Aniruthram K P, [3]Anubama G and [4]Sreemathy J
[1,2,3,4] Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, India
[1]avanthikasivakumar24@gmail.com, [2]aniruthram08@gmail.com, [3]anubamagunasekaran@gmail.com,
[4]sreemathy.j@sece.ac.in

Article Info
A. Haldorai et al. (eds.), *2nd International Conference on Materials Science and Sustainable Manufacturing Technology*, Advances in Computational Intelligence in Materials Science.
Doi: https://doi.org/10.53759/acims/978-9914-9946-9-8_1

Abstract - The world revolves entirely around data. We are unable to find even the most fundamental facts without data. Data will grow as enterprises do. You can learn a lot of things with these data. Data integration must be considered in order to access this information. Data integration has a wide range of applications and is used to combine multiple heterogeneous data sets into homogenous data. Analysis, extraction, transformation, and loading are all components of integration. ETL is a key component of data integration and takes up the majority of the work. The concept of warehousing is crucial because it describes how data that has undergone the ETL step will be loaded and used for various purposes.

Keywords - Data Integration, Information, ETL, Heterogenous Data, Homogenous Data.

I. INTRODUCTION

Everyone uses data, which is present everywhere. There is a lot of data being handled by every organisation. Any new invention in this era is impossible without data. Even if it were conceivable, the wrong customers and the wrong places would not receive it. We'll have a limited amount of data when we first launch our company. We must deal with a large amount of data as our firm grows, and it is impossible to guarantee that all of this data will be loaded into files in the same manner. Combining data from numerous sources allows executives and data managers to analyse it and improve company choices. This procedure involves employing a human or system to find, retrieve, clean up, and present the data. Data managers and/or analysts can perform queries on this connected data to obtain business intelligence insights.

Because there are so many potential benefits, businesses should take the time to match their goals with the best plan. To further understand data integration, let's look at its five different forms (sometimes referred to as approaches or techniques). Combining data from numerous sources allows executives and data managers to analyse it and improve company choices. Locating, obtaining, cleaning, and presenting the data are all steps in this process.We'll go over each type's benefits and drawbacks as well as appropriate times to employ them.In other words, we will receive files in various formats within an organisation. The data needs to be transformed into a single format so that we can manipulate it easily.

II. DATA INTEGRATION

The process of collecting data from multiple sources or applications and creating it into a single unified view is known as data integration. Data Integration nowadays has become the most important thing in all the organizations to store information in different databases. [3]

It helps the business users to integrate the data from multiple sources.Businesses use data integration tools with various applications, technologies and techniques to collect data from different sources and to create it as a single unified view. The process of data integration involves system locating, retrieving, cleaning and presenting the data.[1]

Methods or strategies of data integration according to talend:

Manual Data Integration
- Reduced cost
- Greater freedom

Middle-ware data integration
- Better data streaming
- Easier access between systems

Application based integration
- Simplified process
- Easier information exchange

- Fewer resources are used

Uniform access integration

- Lower storage requirements
- Easier data access
- Simplified view of data

Common storage integration

- Reduced burden
- Increased data version management control
- cleaner data appearance
- Enhanced data analytics

III. ETL

Extraction, Transformation, and Loading is referred to as ETL. It is an essential part in data integration. This three-stage process plays a major role in business organizations. The first step in the processing procedure is extraction, during which we gather the data needed from various sources [10]. Then comes Transformation and Loading Data will be transformed during the transformation process in accordance with the specifications. The transformed data is put into a data warehouse or database during the loading stage, which is the last step in the ETL process [7]. Data analysis is just one of the many uses for the data that can be done with it after it has been put into the warehouse. ETL is majorly used to convert heterogeneous data into homogeneous data which may be then loaded into a database or a data warehouse. Hence with the help of ETL, organizations can reduce the time used on computing and can automate processes which are complex and difficult.

IV. TALEND

Fabrice Bonan and Bertrand Diard created the open-source software Talend in 2005[8]. It is currently among the most widely used ETL applications. Talend consists of variety of components which makes it easier for the ETL developer to use and process data. Talend offers a wide variety of products, such as Talend Big Data and Talend Data Integration. Since Talend was created using Java, it is much better suitable for those who have a foundational understanding of Java programming. The fact that Talend is affordable and open-source software is one of the main reasons why businesses favour it. Talend's design palette enables us to position components and later adjust them as necessary.

Overview Of Major Talend Components

File Components

Input Components

Talend offers a number of input components that enable us to retrieve input based on the sort of data. The entry components tFile Input Delimited and tFile Input Excel are two of the more popular ones. The component tFile Input Delimited can be used to obtain input data in the form of delimited files. To obtain Excel files as input, we can use tFile Input Excel.

Output Components

tFile Output Delimited and tFile Output Excel are two commonly used output components. Similar to input components, they will show the data in the designated formats rather than fetching input.

Management Components

tFile List is one of the management components. It helps us to list the files which are present in the specified directory.

Processing Components

tMap

It is one of the most crucial processing-related components. tMap is a multitasker because it can perform tasks that are typically handled by a variety of other components. The variables can be mapped, and a connection can be found between them. The following stages of processing could then make use of those variables.

tAggregate Row

Aggregation actions like sum, min, max, and other similar operations are performed using this component.

tFilter Row

The records are filtered using tFilter Row based on a defined criteria that can be changed in the configuration tab.

tSort Row

It is used to sort data based on a condition, as its name suggests.

tJoin

On the available records, join procedures like inner join and outer join are carried out using the tJoin component.

V. PERFORMING EXTRACTION

Extraction is the procedure used to obtain data from a variety of sources. We have produced data based on the provided use case.

The expected format is:

The following fields must be in String format: Company Name, First Name, Last Name, Designation, Salary, Email id, and Phone Number. Age and Employee Id must both be of the integer type. Date format is required for Date Of Birth and Date Of Joining. The commission must of double data type.

In the beginning, we processed three distinct files, each divided by area. Employee info_ India, Employee info_ UK, and Employee info_US are their respective names. We have used Talend's tRow Generator component, which can be customised, to generate data for a "n" amount of records. **Fig 1** shows tRow Generator customization.

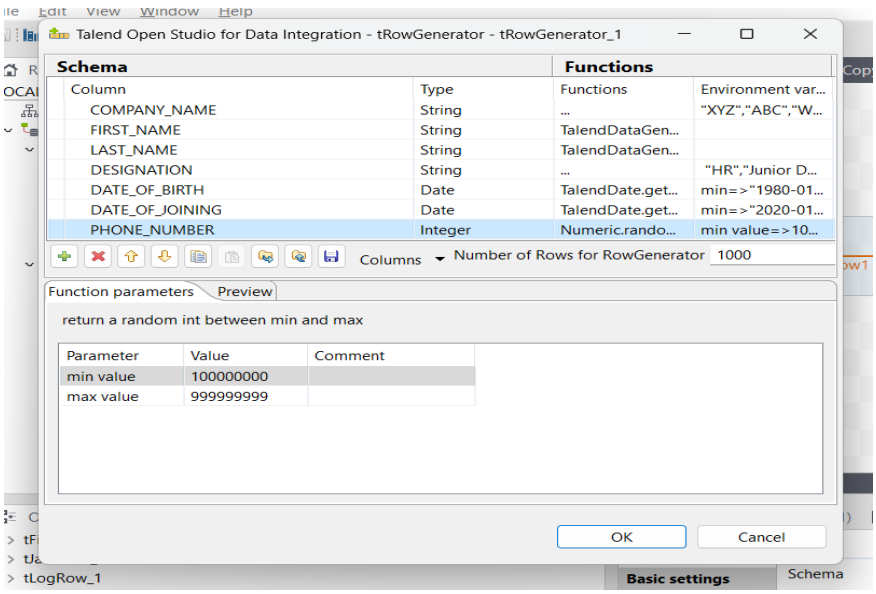


Fig 1. tRow Generator Customization.

Here, we've tailored the parameters using tRow Generator. The number of rows that must be generated for a specific task can be changed in the "Number of Rows for Row Generator" field. Using tMap's expression builder, we have customised EMPLOYEE_ID ,according to the COMPANY_NAME. Numeric.sequence() function generates numeric sequence according to the parameters specified inside. Similarly, SALARY has been generated based upon the DESIGNATION of the employee. Numeric.random() is used to generate random numeric records. **Fig 2** shows JOB 1.

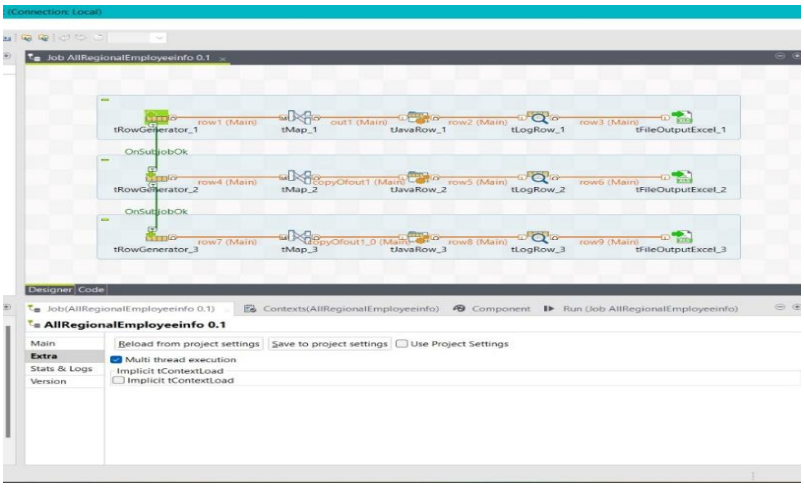


Fig 2. JOB 1

VI. PERFORMING TRANSFORMATION

Transformation is a process of converting the extracted data based on our requirements, Here we have transformed the extracted data based on the use case given.The expected outcomes of transformation are:

Company Name,Employee ID,First Name,Last Name,Designation,Date Of Birth,Salary,Date Of Joining should be fetched from the Source System "X". Age must be calculated by using Date Of Birth. Then, Email Id must generated by appending First Name,first character of Last Name,Company Name, and Date Of Birth .

Eg: nickj08101999@xyz.com . Experience must be calculated from Date Of Joining.Commission must be 10% from the Salary and Phone Number must be a 10 digit integer prefixed by country code.

Initially we have combined all regional employeeinfo data into one single file using tunite component and then tmap and tjavarow components are connected each other, where tmap is connected with tunite and tjavarow is connected with tmap component. Bascially tmap and tjavarow are most popular and widely used components for performing transformations. For more complex transformation opertions talend tool provides us the java component families like tjava,tjavarow,and tjavaflex. **Fig 3** shows JOB 2.

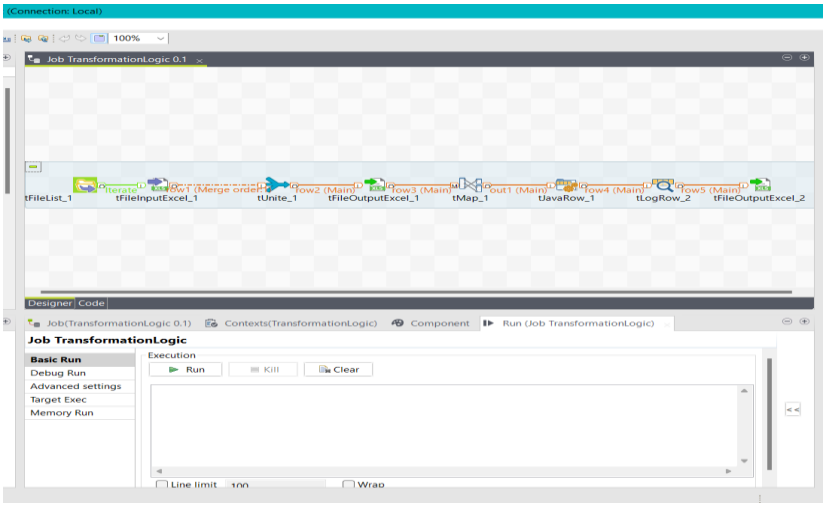
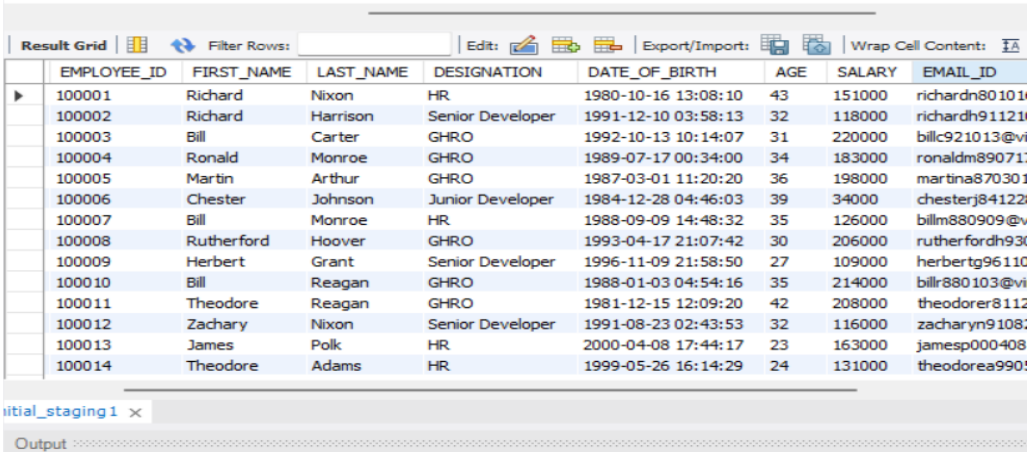


Fig 3. JOB 2

For calculating the Age we have used Talend diff date method which operates on current date and date of birth of the employee to generate age. For generating email id we performed java code to concat firstname and first character of the last name along with Date of birth in yymmdd format followed by "@" company name ".com". Next transformation is to calculate ten percent of commission, for that we have divided salary by 10.0. And then to calculate Experience the current year is subtracted with Year of joining, same is replicated for month and then both results are added and divided by 12, finally multiped by 100 to get desired output. Phone number is appended with the country code respectively.All these transformations are finally loaded into database table using tdboutput component for further SCD loading. **Fig 4** shows Transformation Results



Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

EMPLOYEE_ID	FIRST_NAME	LAST_NAME	DESIGNATION	DATE_OF_BIRTH	AGE	SALARY	EMAIL_ID
100001	Richard	Nixon	HR	1980-10-16 13:08:10	43	151000	richardn80101
100002	Richard	Harrison	Senior Developer	1991-12-10 03:58:13	32	118000	richardh91121
100003	Bill	Carter	GHRO	1992-10-13 10:14:07	31	220000	billc921013@vi
100004	Ronald	Monroe	GHRO	1989-07-17 00:34:00	34	183000	ronaldm89071
100005	Martin	Arthur	GHRO	1987-03-01 11:20:20	36	198000	martina870301
100006	Chester	Johnson	Junior Developer	1984-12-28 04:46:03	39	34000	chesterj84122
100007	Bill	Monroe	HR	1988-09-09 14:48:32	35	126000	billm880909@v
100008	Rutherford	Hoover	GHRO	1993-04-17 21:07:42	30	206000	rutherfordh93(
100009	Herbert	Grant	Senior Developer	1996-11-09 21:58:50	27	109000	herbertg96110
100010	Bill	Reagan	GHRO	1988-01-03 04:54:16	35	214000	billr880103@vi
100011	Theodore	Reagan	GHRO	1981-12-15 12:09:20	42	208000	theodorer8112
100012	Zachary	Nixon	Senior Developer	1991-08-23 02:43:53	32	116000	zacharyn9108
100013	James	Polk	HR	2000-04-08 17:44:17	23	163000	jamesp000408
100014	Theodore	Adams	HR	1999-05-26 16:14:29	24	131000	theodorea990

itial_staging 1

Output

Fig 4. Transformation Results

VII. PERFORMING LOADING

Loading is a process of storing the transformed data in one particular place where it can be used for further operations like analytical purposes [9]. Once the data is transformed it is then loaded into a table called staging table which then used for SCD loading. SCD stands for slowly changing dimensions which helps in managing and storing both current as well as historical data over a period of time in data warehouse. **Fig 5** shows JOB 3.

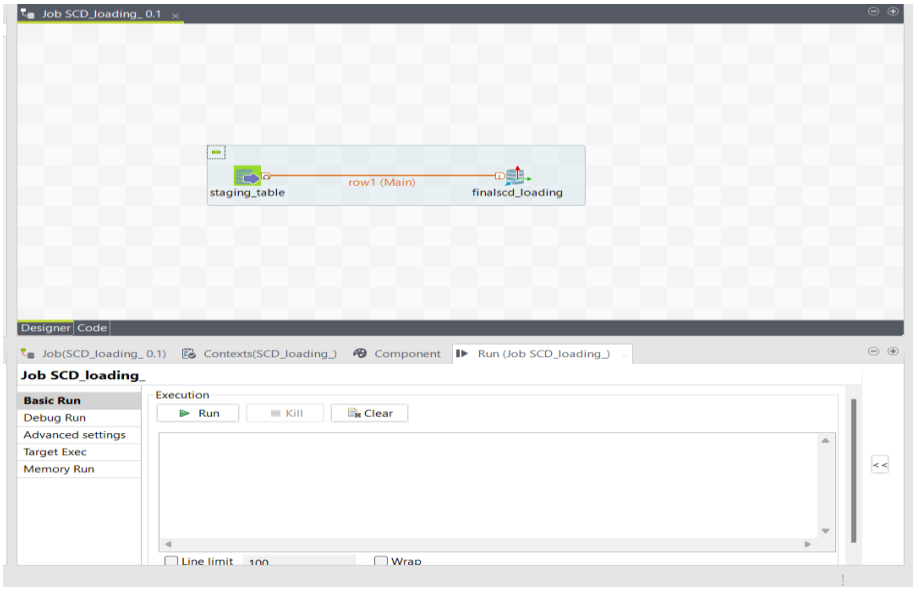


Fig 5. JOB 3

There are three SCD types, which is typically called as type 1 ,2 and 3. Type 1 is overwritting, which overwrites the existing data and type 2 stores both current and history of the record and it is the most effective and preferable type for data analytics, which makes their job more easier. In **Fig 6** we can see tdbinput component is linked with the tdbscd component, where tdbinput component fetches the data from the staging database table and flows the data to the scd component using row(main). For SCD type we have used type 2 and configured Surrogate key as EMPLOYEE_KEY, Start_date as DATE_OF_JOINING and end_date as NEW_DATE with ACTIVE flags.

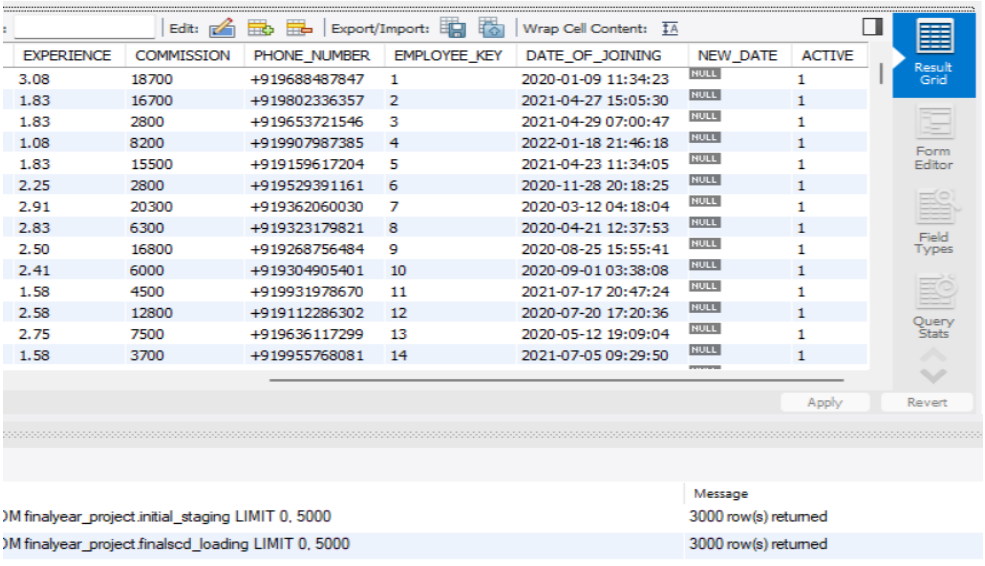


Fig 6. SCD Loading

VIII. RESULTS

The final output willl be:

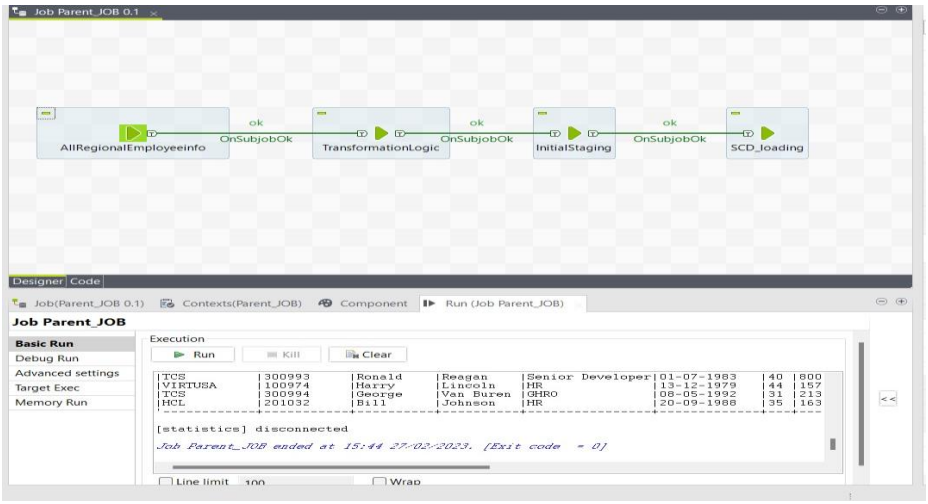


Fig 7. Final Output

Here Talend has helped to easily perform the operations through its different components. Especially tJavaRow and tMap has helped us to get the exact formats according to the requirements. **Fig 7** shows final output.

References

[1]. H. M. Gihan Chanuka Karunarathne, H. M. N. Dilum Bandara, and S. Herath, "WDIAS: A Microservices-Based Weather Data Integration and Assimilation System," 2020 Moratuwa Engineering Research Conference (MERCon), Jul. 2020, doi: 10.1109/mercon50084.2020.9185270.

[2]. M.-E. Vidal, S. Jozashoori, and A. Sakor, "Semantic Data Integration Techniques for Transforming Big Biomedical Data into Actionable Knowledge," 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Jun. 2019, doi: 10.1109/cbms.2019.00116.

[3]. M. Matyoqubov, A. Saidov, O. Kazakov, and O. Rustamova, "Enterprise Systems Data Integration," 2020 International Conference on Information Science and Communications Technologies (ICISCT), Nov. 2020, doi: 10.1109/icisct50599.2020.9351423.

[4]. Fivien Nur Savitri and H. Laksmiwati, "Study of localized data cleansing process for ETL performance improvement in independent datamart," Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, Jul. 2011, doi: 10.1109/iceei.2011.6021806.

[5]. P. S. Diouf, A. Boly, and S. Ndiaye, "Variety of data in the ETL processes in the cloud: State of the art," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), May 2018, doi: 10.1109/icird.2018.8376308.

[6]. D. Alvarez-Coello and J. M. Gomez, "Ontology-Based Integration of Vehicle-Related Data," 2021 IEEE 15th International Conference on Semantic Computing (ICSC), Jan. 2021, doi: 10.1109/icsc50631.2021.00078.

[7]. S. Oni, K. Pansare, S. S. Arneja, Z. Chen, A. Crainiceanu, and D. Needham, "RDFINT: A Benchmark for Comparing Data Warehouse with Virtual Integration Approaches for Integration of RDF Data," 2020 IEEE International Conference on Big Data (Big Data), Dec. 2020, doi: 10.1109/bigdata50022.2020.9378131.

[8]. M. HAJJI, M. Qbadou, and K. Mansouri, "Towards the Development of Talend Open Studio Components for the Support of Semantic Sources," 2019 1st International Conference on Smart Systems and Data Science (ICSSD), Oct. 2019, doi: 10.1109/icssd47982.2019.9002820.

[9]. Md. Badiuzzaman Biplob, G. A. Sheraji, and S. I. Khan, "Comparison of Different Extraction Transformation and Loading Tools for Data Warehousing," 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), Oct. 2018, doi: 10.1109/iciset.2018.8745574.

[10]. B. Pan, G. Zhang, and X. Qin, "Design and realization of an ETL method in business intelligence project," 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Apr. 2018, doi: 10.1109/icccbda.2018.8386526.

[11]. P. S. Diouf, A. Boly, and S. Ndiaye, "Variety of data in the ETL processes in the cloud: State of the art," 2018 IEEE International Conference on Innovative Research and Development (ICIRD), May 2018, doi: 10.1109/icird.2018.8376308.

[12]. H. Dhayne, R. Haque, R. Kilany, and Y. Taher, "In Search of Big Medical Data Integration Solutions - A Comprehensive Survey," IEEE Access, vol. 7, pp. 91265–91290, 2019, doi: 10.1109/access.2019.2927491.

[13]. Munawar, "Extract Transform Loading (ETL) Based Data Quality for Data Warehouse Development," 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), Oct. 2021, doi: 10.1109/iccsai53272.2021.9609770.

[14]. A. Ta'a, N. Ishak, E. M. Elias, and N. Mahidin, "An Impact Analysis Of Extract Transform Load Process For Maintaining The System Of Data Warehouse," Journal of Information System and Technology Management, vol. 7, no. 27, pp. 168–186, Sep. 2022, doi: 10.35631/jistm.727014.

[15]. 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), Mar. 2020, doi: 10.1109/icaccs48705.2020.9074186.