iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.43931

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Study of ETL Process and Its Testing Techniques

Mr. Sujit Prajapati[1], Mrs. Sarala Mary[2]
[1]*Student,* [2]*Professor, Institute of Computer Science, Mumbai Educational Trust- MET ICS, Mumbai, India*

Abstract: ***ETL stands for Extract Transform and Load. ETL is a process in Data Warehouse. In this process ETL tools are used to extracts the data from various data source systems from where data can be retrieved, transforms it in the staging area, and at the end it loads the data into the Data Warehouse system. This process picks data from operational system and fixes it into the data warehouse. Construction of an ETL process is one of the bulky task of data warehouse. In this paper we will try to cover ETL process, Different ETL testing techniques and their challenges.***
Keywords: Extract, Transform, Load, ETL testing

I. OBJECTIVES

A. To study the importance of ETL in Data Warehouse.
B. To study the process of ETL in Data Warehouse.
C. To study the various type of ETL testing.
D. Understanding the challenges of ETL testing.

II. INTRODUCTION

A. What is ETL process?

Data from various operational systems has to be extracted and used into the information warehouse regularly so it can achieve its purpose of simplify business analysis. Data warehouse provides a brand-new collective information base for business intelligence by integrating, rearranging and increase great deal of knowledge on many systems. ETL is the process of extracting data from operational systems and settling it into the data warehouse. ETL stands for extract, load and transform. The processes and tasks of ETL have been well-known for past years, and are not unique to data warehouse environments: for any business all proprietary applications and database systems are the IT backbone [as shown in figure]. Data should be shared between those applications or required systems, trying to mix them. This data sharing was mostly addressed by mechanisms almost like what we now call ETL.

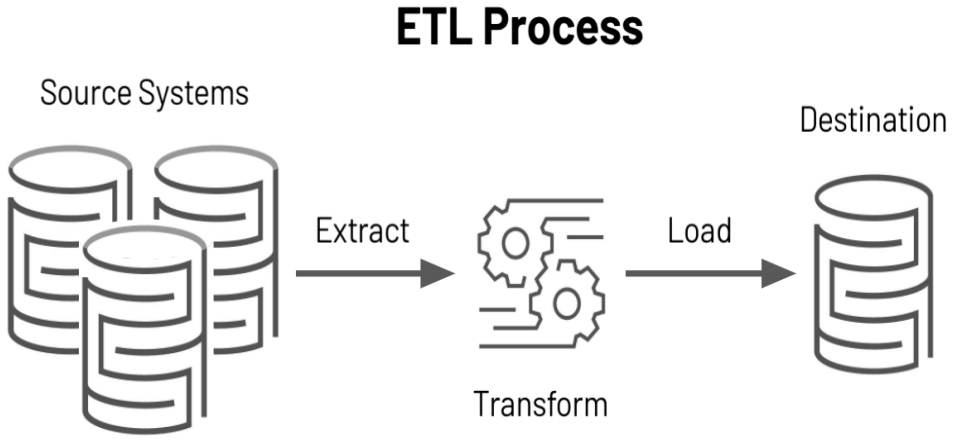


Fig. ETL process

1) *ETL Process:* After performing all the operations of ETL process, the data from the source system gets loaded into Data Warehouse system. ETL process is performed in the three operations as it Extracts the information and required data from operational system which might be any kind of relational database like Microsoft SQL, My SQL, IBM DB and oracle, Transforms the data and information for performing data cleansing operations and Loads after performing these two operations the information stores into the Data Warehouse. [1] In simple terms it can be said that it is a process data engineers use to extract data from different sources, transform the data into a usable and trusted resource, and load that data into the target systems so that end-users can access them and use them to solve business problems.

Different ETL tools are used to perform this process and by using ETL tools we are able to also extract data from flat files like spreadsheets, XML files and CSV files and store it into an information warehouse for further processes like data analysis and reporting.

B. Extraction Transformation and Loading

1) *Extraction:* In extraction process, the needed data is analyzed and fetched up from one or more different sources, like database systems and different applications. It may be possible the size of the extracted data from these sources always be different it can be from hundreds of kilobytes up to gigabytes, it is totally based on the source system and also the business requirements. The main objective of the extract process is gaining all the required business data from the source system with the less amount of resources as much as possible. The extract process should be built in such a way that it doesn't have negative impact on the source system in terms of performance, their reaction time or any variety of locking mechanism. Extraction can be either Logical Extraction or Physical Extraction. Logical Extraction is completely used to extract the data from source system. Logical extraction doesn't require addition logic to extract data from source. Sometimes drawing outdated data from outdated system is not possible through Logical Extraction that is why we require Physical Extraction.
2) *Transformation:* In this process the information is transformed into the applicable format of data that may be easily stored into a Data Warehouse system as it is required. Transformation process is mainly associated with applying calculations, DML operation, joins, constraint, primary key and foreign keys on the data. as an example, consider if you wish to calculate average of total annuity then you'll have to apply average formula in transformation logic of the data. For a few data there's no requirement to perform any transformation which is direct moveable into the data warehouse system; that data is also known as direct move or pass through data. Data transformation process also involves data correction, cleansing of data, removing incorrect or any duplicate data, incomplete data formation, and fixing the data errors, data integrity and formatting incompatible data that may cause problem before loading it into Data Warehouse system.
3) *Loading:* The final step in the ETL process is of Loading the data in to the target in multidimensional structure such as cube. During this process, extracted and transformed data is stored into the dimensional structures which is then accessed by the top users and application systems. [2] Loading Process includes both loading dimension tables and loading fact tables. It is an important point to make sure that the load operation is performed properly and it should be with as less resources as much as possible. The main target may be a database of the Load process. To form the load process efficiently, it's advisable to disable any constraints and indexes that are present before the load and enable them back only after the load process completes to make sure that consistency referential integrity should be maintained by ETL process.

III. LITERATURE REVIEW

Data warehousing is very important. It facilitates the analysis of business processes in an efficient way by putting all of your important data in one easily accessible location. to attain this, organizations must integrate, rearrange and consolidate large amounts of data across multiple systems. This process is observed as ETL, or extraction, transformation, and loading. ETL may be a complex system of processes that's achieved in three simple steps. it's a fancy combination of process and technology.

ETL also takes an excessive amount of time and expensive process, which may leave decision makers in industrial organizations anticipating the critical data they have to form decisions and perform their operations. ETL process doesn't transfer the specified orders to the order release system in time, and so the top result can be that there are delays while the production line waits for the parts to be available. Moreover, this might cost the business further, if the worth for the parts increases within the time it takes to release the orders.

IV. STAGES IN ETL TESTING

We perform ETL testing which is designed to validate and verify ETL process to get correct data after reducing data redundancy or any information loss. Effective ETL testing plan can be used to detect problem with the source data as early as possible before it's loaded to the data repository. It can be used to detect any inconsistency or ambiguity in given business rules that are intended to guide integration and data transformation. [3] ETL testing can be broken down into eight steps which can be used while performing testing:

A. Identify Business Requirements

The first and most important step in ETL testing is to capture the business requirements of given model. It can be achieved by designing the data models, schematic lane diagrams, reports and by defining the business flow based on the client expectations. It's important to start here and to understand the business requirements so that testers can be aware of what to be tested and to understand the scope of the project. The team needs to thoroughly document the project to fully understand its scope.

B. Validate Data Sources

The next step in the ETL testing is to identify the source data and perform the preliminary check. It can involve checking data counts, schema check, validation of table, column data type check, etc. in order to check whether given ETL process aligns with the specification of table. Tester needs to make sure that primary keys are checked and in place and remove duplicate data. If this is not done correctly, the aggregated report could be inaccurate or it can be misleading.

C. Design Test Cases

The next step is to create a mapping between source and target and to design the test cases. It includes transformation logic according to the business requirements, SQL script to perform and get source to target count comparison, and execution flow of the test cases, etc. To ensure that mapping document align with the business needs and it contains all of the information, we need to validate it as well.

D. Extract Data from Source Systems

This step involves executing ETL test cases as per business requirement. Identifying different types of bugs or defects that are encountered during testing and make a report of it. It is important to detect and reproduce any defects, report, fix the bug, resolve, and close bug report before continuing to further steps.

E. Apply Transformation Logic

In this step tester needs to ensure that data is transformed according to business logic to match schema of target data warehouse. Check data threshold, alignment, and validate data flow of source data. This ensures that the data type is matching the mapping document for each column and table.

F. Load Data into Target Warehouse

It involves performing a record count check before and after data is moved from staging to the data warehouse to ensure that every record is moved to the target system. It confirms that invalid data is rejected as per business logic and that the default values are accepted.

G. Summary Report

In this tester verifies layout, options, filters and export functionality of summary report. This summary report lets decision-makers/stakeholders know the details and results of the testing process and if any step was not completed that is “out of scope” and why.

H. Test Closure

It’s final step in ETL testing where tester closes the testing file and completes all kind of testing.

V. ETL TESTING CATEGORIES

ETL testing can be categorized into four general categories depending upon the data process and their journey throughout the testing phases [4].

A. New Data Warehouse System Testing

In this type of testing new data Warehouse system is built and gets verified. It takes data input from different end users and customers as well as from various data sources. Later, this using this data new Data Warehouse system gets created and these data are verified using different ETL tools. Following ETL testing can be performed in this category:

1) Data quality testing
2) Metadata testing

B. Migration Testing

Different ETL tools can be used for ETL testing and each of them ay vary in terms of performance, time cost, etc. In Migration testing instead of creating new Data Warehouse system, customers have already existing Data Warehouse and ETL, but to improve the efficiency and time cost they look for different ETL tools. It involves migration of existing Data Warehouse system using new ETL tools. This testing ensures that data is migrated from one system to other without loss of any data and by applying transformation rules and strict checking. It may involve following ETL testing techniques:

1) Data quality testing
2) Source to target count testing
3) Source to target data testing

4) Performance testing
5) Data transformation testing
6) Data integration testing

C. Change Testing

In this category, data from different data sources are added to existing Data Warehouse system which means changes takes place in existing system. In Change testing, customer can change the existing rules as well as can add new rules to the system. It may involve following ETL testing techniques:

1) Data quality testing
2) Source to target count testing
3) Source to target data testing
4) Production validation testing
5) Data integration testing

C. Report Testing

Reports are final output of any Data Warehouse system to check data validation and data quality. Report testing involves creating reports for Data Warehouse system and its data validation. Reports are tested based on data of the report, their layouts and their calculated values. It involves following ETL testing techniques:

1) Report testing

VI. ETL TESTING TECHNIQUES

Defining the appropriate ETL Testing technique is important prior to starting the testing process. We must get approval from every team member to make sure that appropriate testing method is chosen to apply ETL testing. Following are the various testing methods which can be used:

A. Production Validation Testing

To Perform Analytical Reporting and Analysis, Maintaining the accuracy of data is important aspect. Production Validating testing performs on data which is sent to the production system. It includes information analysis in the system and then compare it with the source data of the system. It is also called as Table balancing or Production reconciliation [2]. To support business decisions, data in the production system has to be in correct order. This validation is used to check the data against false logic, failed loads or the different operational processes that are not loaded correctly in the production system.

B. Source-to-target Count Testing

It includes checking number of information in source and target systems. Having less time, we can select this source to target count technique to move testing operation. This testing does not include analysis of the data in target system. Even if data is ascending or descending order after mapping of the data then this testing can be use.

C. Source-to-target Data Testing

If tester wants to validate data values between the source and target system then they can use Source-to-target testing. It checks for value of data after transformation in the system source & the interrelated target system values.

D. Data Integration / Threshold Value Validation Testing:

In Data Integration / Threshold Value Validation Testing Tester analyses data series. Every origin part in target system are verified if values are as per the desired results. Data integration in target system from number of source systems included after transformation and loading.

E. Application Migration Testing

Application migration testing is performed automatically whenever we verify data from previous application to current application. If data pulled up from previous application is same as current application system then we use this type of testing and it saves a lot of time.

F. Data Check and Constraint Testing

This testing is used to verify many checks constraints like data length check, data type check, and index check. In this type of testing Tester performs the following task – Foreign Key, Primary Key, NULL, NOT NULL & UNIQUE.

G. Duplicate Data Check Testing

When there is huge amount of data in the target system it can cause the incorrect data in Analytical testing and the possibilities of having duplicate information in the production system are found. This testing checks the repeated data in the target system.

H. Data Transformation Testing

This testing is time consuming and it uses multiple SQL queries for each row to verify the transformation rules. For each tuple in the table the tester needs to run SQL queries so target data can be compared with results.

I. Data Quality Testing

This testing is used to test the data quality and tests number check, date check, null check, precision check, etc. To address invalid characters, wrong upper- or lower-case order etc Tester apply SYNTAX TEST and when the data is according to the given model Reference Test applied.

J. Incremental Testing

To use this testing technique, Insert, delete and Update statements should run as same as the desired output. Incremental testing check from old and new data it performs step by step.

K. Regression Testing

To insert new functionality, we use Regression Testing which in turn supports the tester to search new errors, after that tester can do desired changes in data transformation and aggregation rules. The errors in data comes in regression testing can be called Regression.

L. Retesting

Retesting is applied at the time when we execute the test after completing the codes.

M. System Integration Testing

System integration testing plays important role in verifying the parts of a system one by one and after that adding the modules in the system. System integration is of three types: hybrid, top down, and bottom up.

N. Navigation Testing

This testing is also called as front-end testing of the system. To verify all the factors of the front-end according to the end user requirement this testing included, information of various fields, aggregates, and calculation etc included in reports.

VII. COMPARISON OF ETL TESTING TECHNIQUES

TESTING NAME	COMPARE ON TYPE OF DATA	TESTING OBJECTIVE	TIME DURATION
Production Validation Testing	Source data	Testing of data about production system	It depends on Production system
Source to Target count Testing	Source data and Target data	It can be performed when tester do not have plenty of time	Time saving
Source to Target Data Testing	Source data and Target data	It can be performed in medical and academic project	It is time Consuming
Data Integration value validation Testing	On multiple source of data	For checking span of data then tester can use this testing	It's time Consuming

Application Migration Testing	Data of application	It can be applied from existing application to the current application which is performed automatically	It is time Saving
Duplicate data check Testing	Data of Table	Tester can use this testing when there are duplicate data in table	Depend ondata
Data Transformation Testing	Compares the output of target data	It can be used when we want to compare the target data with result	Time Consuming
Data Quality Testing	For all the data	To check quality of data use data quality testing	Time Consuming
Incremental Testing	Data of database	To test insert, update, and delete statements in database.	Step by Step check
Regression Testing	Any type of Data	To test and verify the new functionality Tester can use Regression testing	Depend on type of data
Retesting	Checks everything	For rechecking the code and data tester can use this testing type	Depend on type of data
System Integration Testing	It mainly focuses on interfaces and flow of Data or Information between the modules	It is used for testing the component of system individually	Convenient for small system
Navigation Testing	Include data from various fields	It is used for checking front endoutput	Time Consuming

VIII. ETL TESTING CHALLENGES

ETL testing and database testing both are different and we have to face many challenges during ETL testing process. Some common challenges which is listed below. [5] ETL testing is also very different from typical software testing because it's primarily about data, not code. The data type can be different for the data coming from each source, so the testing must accommodate heterogeneous data types.

A. Data may be disappearing throughout the ETL process.
B. Wrong, insufficient or repeated data can be in database.
C. It is very complicated to perform ETL testing in the objective system when Data ware system contains real data, so the size of data can be too vast.
D. It is robust to develop and frame test cases, if data size is very vast and complicated.
E. ETL testers do not know the consumer outline demands and trade outflow of data.
F. For data validation in the Target system, ETL testing involves various complex SQL concepts.
G. To target mapping information. The testers most of the time have no idea about the source.
H. In the development and testing of a process the utmost delay is observing

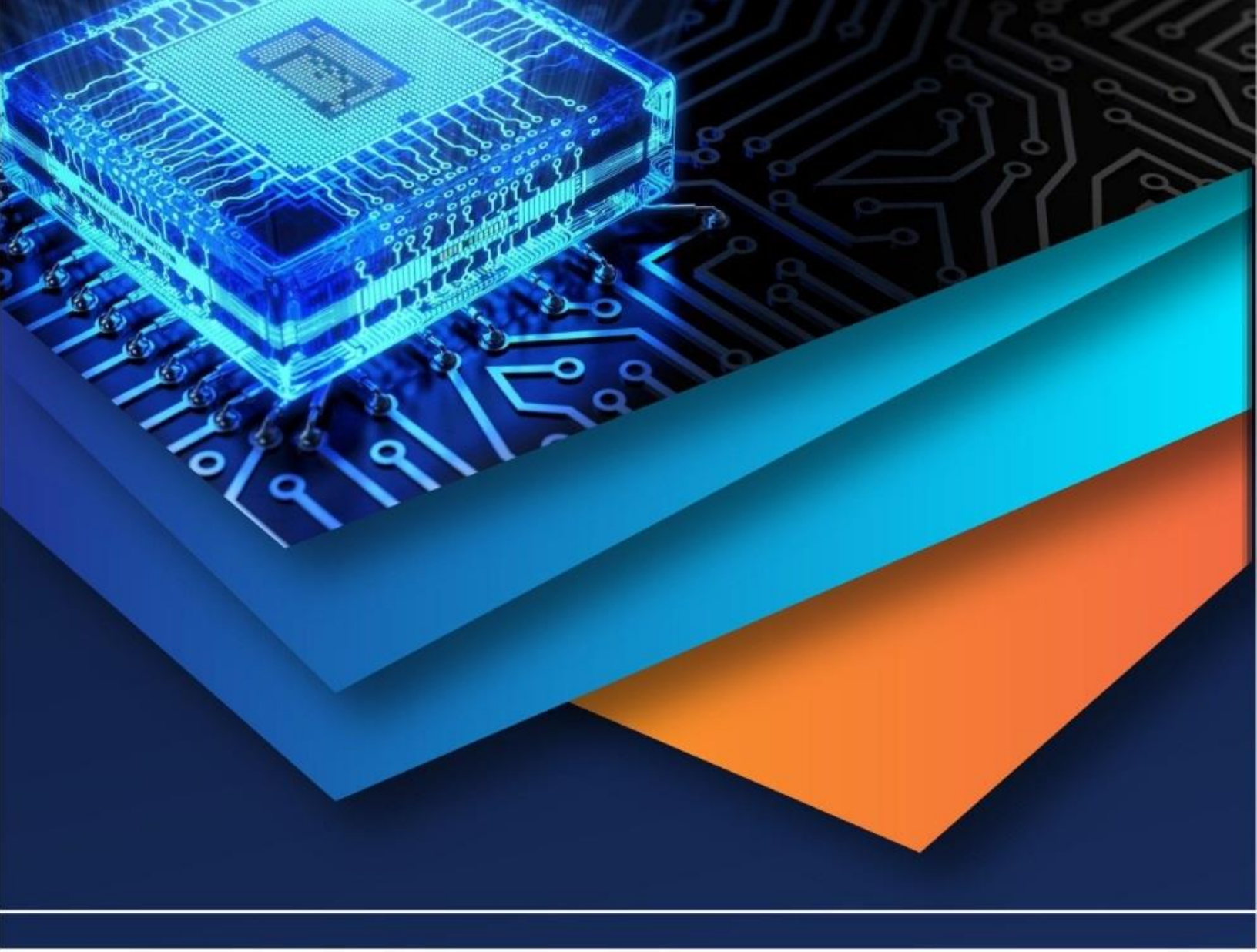
IX. CONCLUSION

In current research of data warehousing and ETL, ETL processes are considered as very critical problem. [2] Value and convolution are the two tags in which ETL is identified. Due to ETL processes importance, the paper is more pointed on ETL, the backstage of DW and it shows the research efforts and opportunities. [6] In the area of ETL Processes and testing techniques,

it is familiar that building ETL processes is costly and consumes time, money and efforts. Here, in this paper we define the challenges related to ETL and their testing techniques. It facilitates the tester for selection of best ETL techniques that can be used and provides various challenges related to ETL processes. In future work we will be defining a various testing technique for ETL processes that we can apply to check all constraint and also consider context of data. So, with the help of this technique all the dependency on data quality and required data checks will be reduced. This paper will show new challenges regarding ETL testing and help people in research opportunities related to ETL process of data warehouse.

REFERENCES

[1] https://databricks.com/glossary/extract-transform-load
[2] Philip Woodall, Torben Jess, Mark Harrison, “A Framework for De-tecting Unnecessary Industrial Data in ETL Processes”, 2014.
[3] https://www.talend.com/resources/etl-testing
[4] https://www.softwaretestinghelp.com/etl-testing-data-warehouse-testing
[5] https://www.matillion.com/resources/blog/what-are-the-basics-of-etl-testing
[6] Miroslav Dzakovic, “Industrial Application of Automated Regres-sion Testing in Test-Driven ETL Development”,2016



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY