IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 12, December 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

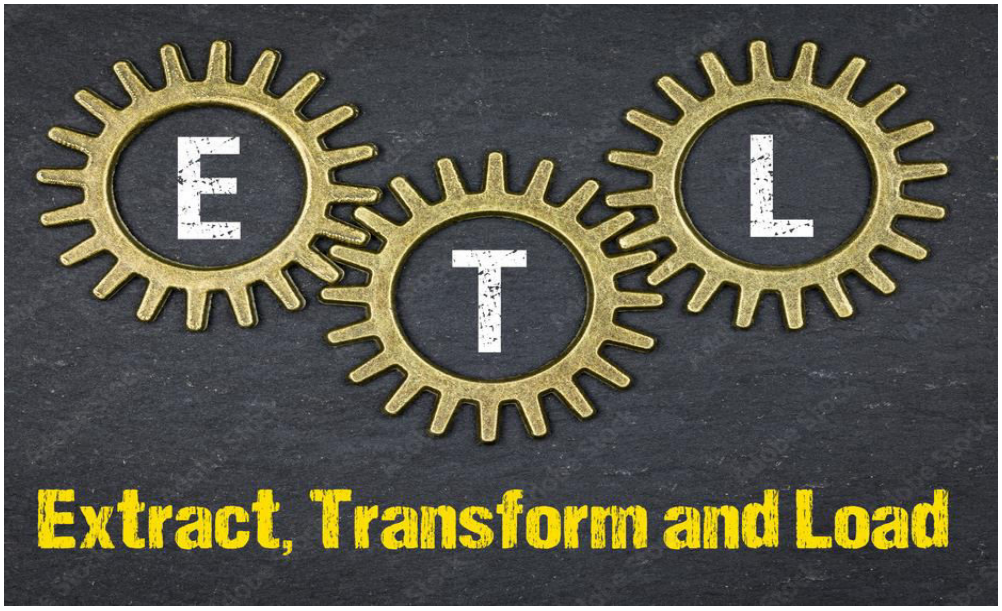
|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112022 |

Study of Various ETL Process and Their Testing Techniques

Tushar Narayan Patil, Soham Satish Ghali, Dr. Mrs.Sampada Gulavani

Department of Master of Computer Application, Bharti Vidyapeeth (Deemed to be University), Pune, Institute of Management, Kolhapur, India



ABSTRACT: Extraction–Transformation–Loading (ETL) tools are fraction of software which extract the data from various sources, and then it clean, customize, reformat, and integrate the data and insert into a data warehouse. ETL process in data warehousing is responsible for pick the data out from the operational systems and fixed it into the data warehouse. Construction of the ETL process is one of the bulky tasks of construct a warehouse. In this paper , an attempt is done to explain the ETL process, ETL testing challenges and ETL testing techniques.

I. INTRODUCTION

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. 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 1.. 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.

II. 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

i)To extract 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.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112022 |

ii)Transforms the data and information for performing data cleansing operations
iii)Loads the data to store into the Data Warehouse.
In simple terms , it is a process 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.

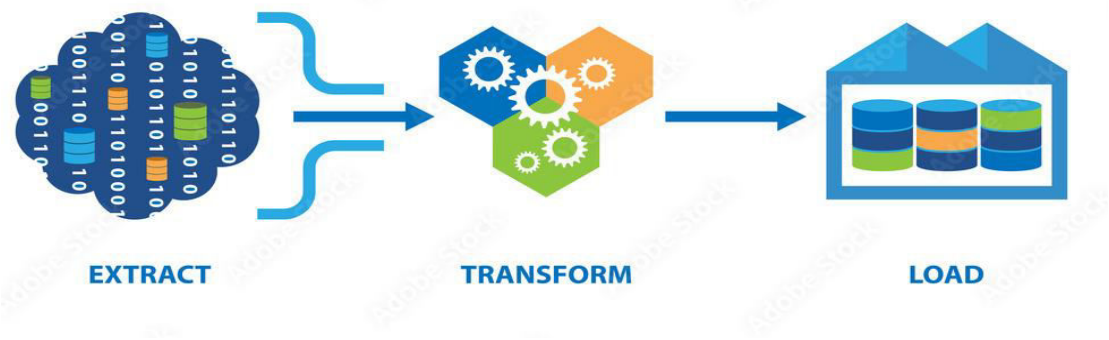


Fig1. : ETL Process

2.1 Extraction Process : In extraction process, the needed data is analysed and pulled up from one or more different sources, such as database systems and applications.The size of the extracted data can be from hundreds of kilobytes up to gigabytes. It totally depends on the source system and the business situations. The extract process should be built such as it does not negatively affect to the source system in terms of performance, response time or any type of locking. Extraction methods are of the types like Logical Extraction Method and Physical Extraction Method

Further logical extraction is divided into Update notification, Incremental extract and Full extract:

i)Update notification - If the source system provides us a notification that a record has been changed and explain the changes, this is the simplest way to get the data.

ii)Incremental extract - Some systems are not be able to provide a notification that a change has occurred, but they are able to find out which records have been updated and provide an extract of that records.

iii)Full extract - Some systems are not able to find out which data has been updated at all, so a full extract is the only way where we can get the data out of the system. The full extract depends upon keeping a copy of the last extract in the same format in order to be able to find out the modifications. Full extract also handles deletions.
Physical Extraction is further divided into :

i)Online Extraction: -The data is retrieved directly from the source system itself. The extraction process can be connecting directly to the source system to access the source tables or to a central systemthat stores the data in a pre-defined manner (for example, snapshotlogs or change tables). The central system is not physically differingfrom the source system.

ii) Off-line Extraction: - The data is not pulled up or extracted directlyfrom the source system but it is explicitly organized out of the originalsource system. The data already has an existing structure (such as,redo logs, archive logs) or was created by an extraction routine.When using Incremental or Full extracts, the extract recurrence isextremely important. Especially for full extracts; the data volumescan be in tens of gigabytes.

2.2 Transformation : In this process the data is transformed into the applicable format that canbe easily stored into a Data Warehouse system. Transformation processassociate with applying calculations, DML operation, joins, constraint, primary key and foreign keys on the data. For example, if you want average of total annuity, you will apply average formula in transformation and load the data.
2.3 Loading : In this process, data is loaded into the targetmultidimensional structure, extracted and transformeddata is stored into the dimensional structures accessed by the end usersand application systems.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112022 |

III.TESTING TECHNIQUES OF ETL

Commonly used ETL testing techniques are discussed below :

i)**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.

i)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.

ii).Source-to-target Count Testing : It includes checking amount of information in source and target systems. This testing does not include analysis of the data in target system.

iii)Data Integration / Threshold Value Validation Testing : It analyses data series where every origin part in target system is verified if values are as per the desired results. Data integration in target system from number of source systems included after transformation and loading.

iv)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 tasks like Foreign Key, Primary Key, NULL, NOT NULL & UNIQUE.

v)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.

vi)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.

vii)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.

viii)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.

ix)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 and it performs step by step order.

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

xi)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.

xii)Regression Testing: It is used to insert new functionality, which in turn supports the tester to search new errors, after that tester can do desired changes in data transformation and aggregation rules.

xiii)Navigation Testing : This testing is also called as front-end testing of the system and used to verify all the factors of the front-end according to the end user requirement.

IV. ETL TESTING CHALLENGES

ETL testing and database testing both are different and we haveto face many challenges during ETL testing process. Some common challenges which is listed below.

- It is very complicated to perform ETL testing in the objective systemwhen Dataware system contains real data, so the size of data can betoo vast.
- Wrong, insufficient or repeated data can be in database.
- Data may be disappearing throughout the ETL process.
- ETL testers do not know the consumer outline demands and tradeout flow of data.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112022 |

V. COMPARISON OF ALL ETL TESTING TECHNIQUES

TESTING NAME	COMPARE ON TYPE OF DATA	TESTING OBJECTIVE	TIME DURATION
ProductionValidation Testing	Source data	Testing of data aboutproduction system	It depends on Productionsystem
Source to Target count Testing	Source data and Target data	Itcan 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 valuevalidationTesting	On multiple sources of data	For checking span of datathen tester can use this testing	It'stime Consuming
Application MigrationTesting	Dataof application	It can be applied fromexisting application to the currentapplication 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, anddelete statements in database.	Step by Step check
Regression Testing	Any type of Data	To test and verify the newfunctionality Tester can use Regression testing	Depend on type of data
Retesting	Checks everything	For rechecking the code anddata tester can use this testing type	Depend on type of data
System IntegrationTesting	It mainly focuses on interfaces and flow of Data or Informationbetween the modules	It is used for testing thecomponent of systemindividually	Convenient for small system
NavigationTesting	Include data from various fields	It is used for checking front endoutput	Time Consuming

VI. CONCLUSION

ETL processes are considered as very critical problem. Value and convolution are the two tags in which ETL is identified. By considering the importance of ETL processes, the paper is focused on ETL with the backstage of Data Warehouse and it shows the research efforts and opportunities and 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, more focus will be given on to apply checks with constraints. 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://www.matillion.com/resources/blog/what-are-the-basics-of-etl-testing
2) https://www.softwaretestinghelp.com/etl-testing-data-warehouse-testing
3) Miroslav Dzakovic, "Industrial Application of Automated Regression Testing in Test-Driven ETL

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.118|

|| Volume 11, Issue 12, December 2022 ||

| DOI:10.15680/IJIRSET.2022.1112022 |

Development",2016
4) https://www.talend.com/resources/etl-testing
5) https://databricks.com/glossary/extract-transform-load
6) Philip Woodall, Torben Jess, Mark Harrison, "A Framework for Detecting Unnecessary Industrial Data in ETL Processes", 2014.
7) W. Inmon. Building the Data Warehouse. John Wiley & Sons, 2nd edition, 1996.
8) Philip Woodall, Torben Jess, Mark Harrison, "A Framework for Detecting Unnecessary Industrial Data in ETL Processes", 2014.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details