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.2022.44939

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

Extraction Transformation and Loading (ETL) of Data Using ETL Tools

Manish Manoj Singh
MCA, Thakur Institute of Management Studies, Career, Development & Research (TIMSCDR), Mumbai, India

***Abstract:** This Research Paper presents the Extract, Transform, Load (ETL) Process and discusses various ETL Tools Available in the Market. A huge piece of BI frameworks is a well-performing Implementation of the Extract, Transform, and Load (ETL) process. In BI projects, implementing the ETL process can be the big task ETL is the core process of Data integration which is associated with Data Warehouse. This paper also focuses on the best ETL Tools and which tool can be the best for the ETL process.*

I. INTRODUCTION

Business intelligence has reached wide recognition in the last few years.
A data warehouse is only a social data set that is intended for inquiry and investigation rather than for exchange handling.
The Data warehouse information is only a mix of authentic information just as conditional information. We want to load the data warehouse consistently with the goal that it can fill its need of working with the business examination. To play out this interaction information from at least one functional framework should be separated and duplicated into the information distribution centre. ETL is a course of extracting data from source frameworks and bringing it into the data warehouse. which stands for extraction Transformation and loading. The procedure and undertaking of ETL have been notable for a long time, and are not remarkable to information stockroom conditions Extract, Transform and Load (ETL) process is One of the important components of Business Intelligence.
ETL processes take up to 80% of the effort in BI projects it is a data integration function that involves extracting data from outside sources (operational systems), transforming it to fit business needs, and eventually stacking it into an information distribution centre To tackle the issue, organizations use extract, transform and load (ETL) innovation, which incorporates perusing information from its source, tidying it up and arranging it consistently, and afterward composing to the objective vault to be taken advantage of.
The information which is utilized in ETL cycles can emerge out of any source like a centralized server application, an ERP application, a CRM device, a level document, or an Excel spreadsheet. ETL tool can gather, read and move information from various information structures and across various stages, similar to a centralized computer, server In this paper, we have analysed some of the ETL Tools.

II. ETL PROCESS

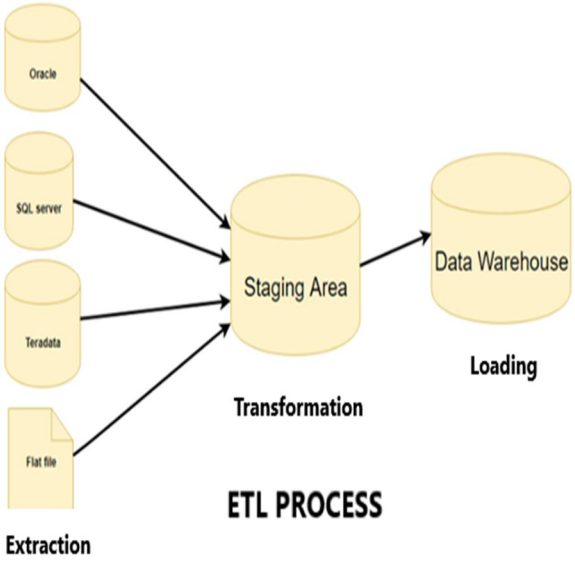
ETL (Extract, Transform, and Load) is a cycle that processes information from different sources and places it into a data warehouse. The purpose of ETL is to provide the users, not only a process of extracting data from source systems and bringing it into the data warehouse but also provide the users with a typical stage to incorporate their information from different stages and applications
ETL is a cycle that extricates the information from various RDBMS source frameworks, then, at that point, changes the information (like applying estimations, connections, and so on) lastly stacks the information into the Data Warehouse system.
Extract, Transform, Load three database capacities that are consolidated into one instrument that computerizes the interaction to haul information out of one database and spot it into another database. The database functions are described following:

ETL involves the following tasks.

1) *Extract:* The most common way of perusing information from a predefined source database and extracting an ideal subset of information.
2) *Transform:* The most common way of changing over the removed obtained information from its past structure into the structure it should be in so it very well may be set into another database. Change happens by utilizing rules or query tables or by joining with different information.
3) *Load:* The method involved with composing the information into the objective database.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Let us discuss briefly all the three processes.



III. EXTRACTION

Extraction is the first part of an ETL process. Every time it is not easy to collect data from various sources and store it in a data warehouse but it can be done using the ETL Process.

In many cases, this addresses the main part of ETL Most information warehousing projects consolidate information from various source systems. the different framework may likewise utilize an alternate information association and organization Common information source designs incorporate social data sets, XML, JSON, and level records.

In simple words, we can say that Extract is the process of reading data from a database. In this process, the data is collected, from multiple and different types of sources Data Extraction can be done from the various source system

The Extract step covers the information extraction from the source framework and makes it available for additional handling.

- The principal objective of the concentrating step is to recover every one of the necessary information from the source framework with as few assets as could be expected.
- The concentrate step ought to be planned such that it doesn't adversely influence the source framework as far as execution, reaction time, or any sort of locking.

Later the extraction, this information can be changed and stacked into the information distribution centre.

None of the extraction processes, today address the security during the extraction process, thus there are possibilities for the data to be hacked during the process. If the data that is extracted contains any confidential data, then just providing security after building a data warehouse cannot make data secure as it would have been hacked during the building process itself.

There are multiple ways to perform the extract:

1) *Update Notification:* In this process if the source system is not providing a notification that a record has been changed and describes the change, this is the easiest way to get the data.
2) *Incremental Extract:* In this interaction, a few frameworks can't give a warning that an update has happened, however, they can recognize which records have been changed and provide an extract of such records. During the following ETL steps, the framework needs to distinguish changes and execute them, By utilizing every day separately, we will be unable to deal with erased records appropriately.
3) *Full Extract:* In this cycle, a few frameworks can't distinguish which information has been changed by any means, so a full concentrate is the main way one can get the data out of the system. full extraction requires keeping a duplicate of the last concentrate in a similar arrangement to have the option to distinguish changes. Full extract handles delete operation as well.

If we are using Incremental or Full extracts, the extracted frequency is extremely important. Particularly for full focuses, the data volumes can be in a few gigabytes.

Some validations are done during Extraction:

- Reconcile records with the source data
- Make sure that no spam or unwanted data is loaded
- Data type check
- Remove all types of duplicate/fragmented data
- Check whether or not all the keys are set up

IV. TRANSFORMATION

Transformation is only an interaction that changes over the separated information from its past structure into the structure it should be in with the goal that it tends to be set into another data set.

Transformation is happened by utilizing a few guidelines or query tables or by consolidating the information with different information.

Information extracted from the source server is crude and not usable in its unique structure. Accordingly, it should be purified, planned, and changed. Indeed, here the ETL cycle adds worth and changes information with the end goal that it tends to be justifiable and precise and by which the BI reports can be created.

In this process, you apply a bunch of capacities to extricate information. Information that doesn't need any change is known as an immediate move or pass-through information, rich information.

- Transformation process includes cleaning, filtering, validating, and applying rules to extracted data
- The main objective of this step is to load the extracted data into the target database with a clean and general format
- This is because we extract data from various sources and each has its format
- The transformation process has a series of rules to transform the data from the source to the target.
- The change likewise requires joining the information from a few sources, creating totals, arranging, inferring new determined qualities, and applying progressed approval rules.

The ETL change component is answerable for information approval, information exactness, information type transformation, and business rule application. It is the most muddled of the ETL components. It might seem, by all accounts, to be more proficient to play out certain changes as the information is being separated.

A. For Example

There are two sources A and B
A date format is dd/mm/yyyy
B date format is yyyy/mm/dd
In transformation, these dates bring it in a standard format

B. Validations are Done During this Stage

1) Filtering – Select just specific sections to stack
2) Utilizing rules and query tables for Data normalization
3) Transformation of Units of Measurements like Date Time Conversion, money changes, mathematical transformations, and so on
4) Data threshold validation check. For example, age can't be multiple digits.
5) Required fields ought not to be left clear.
6) Cleaning (for instance, planning NULL to 0 or Gender Male to "M" and Female to "F" and so on)
7) Split a segment into products and blend different sections into a solitary segment.
8) Transposing rows and columns,
9) Use lookups to merge data
10) Utilizing any complicated information approval (e.g., assuming the initial two sections straight are unfilled then it naturally reject the line from handling)

V. LOADING

Data extracted and transformed is of no use until it is loaded in the target database In this step the extracted data and transform data is loaded to the target database To make information load proficiently it is fundamental

- During the heap step, it is important to guarantee that the heap is performed accurately and with as few assets as could be expected.
- The referential uprightness should be kept up with by the ETL apparatus to guarantee consistency.

Stacking information into the objective data warehouse data set is the last advance of the ETL cycle. In an ordinary Data warehouse, a tremendous volume of information should be stacked in a moderately brief period (nights). Subsequently, the load cycle ought to be upgraded for execution.

In the event of burden disappointment, recuperate systems ought to be arranged to restart from the weak spot without information trustworthiness misfortune. Data Warehouse administrators need to screen, continue, drop loads according to winning server execution.

Every one of the three stages in the ETL cycle can be run equally. Information extraction sets aside time thus the second step of the change process is executed all the while. This gets ready information for the third step of loading.

When a little information is prepared it is stacked without hanging tight for the culmination of the past advances.

A. Types of Loading

1) *Initial Load:* It populates all the Data Warehouse tables

2) *Incremental Load:* In this process applies ongoing changes when needed periodically.

3) *Full Refresh:* It erases the substance of at least one table and reloads with new information.

B. Load Verification

1) Guarantee that the key field information is neither missing nor invalid.

2) Test demonstrating sees dependent on the objective tables.

3) Check that combined values and calculate measures.

4) Data checks in dimension table as well as in history table.

5) It Checks the BI reports on the stacked truth and aspect table.

VI. DATA STAGING

As the data is extracted from the source, the next step is transformation if unfortunately, the transformation step fails, it is not necessary to restart the Extract step. We can do this by carrying out appropriate arranging An organizing region (DSA) is a brief stockpiling region between the information sources and an information stockroom. Where data from source systems is copied It is a process where we perform several operations The staging area is also used in the ETL process to store the results of processing.

The staging area has quickly extracted the data from its data sources, minimizing the impact of the sources.

As the data is loaded into the staging area, a staging area is combined data from multiple data sources, transformations, validations, data cleansing.

A staging area is usually in a Data Warehousing Architecture for timing reasons. It means all required data must be available before data can be integrated into the Data Warehouse.

VII. ETL TOOLS

An ETL apparatus is a product, principally utilized for Extracting, Transforming, and Loading information. ETL tools empower associations to make their information open, significant, and usable across information frameworks. When it comes to tools, you have a lot of options for choosing the right ETL (Extract, transform, load) tools that were used to simplify the data management by reducing the absorbed effort. These are designed to save time and money when a new data warehouse is developed Depending on the needs of customers there are many types of tools and you have to select the appropriate one for you. Most of the ETL tools are quite expensive, some tools are complex to handle. The most important aspect to start with defining business requirements is the selection of the right ETL tool. The working of the ETL tools depends on ETL (Extract, transform, load) process.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

There are the Following ETL Tools which Is used in Data Processing

- Informatica PowerCenter
- Skyvia
- IBM Infosphere DataStage
- Oracle Data Integrator
- Microsoft SQL Server Integration Services

There are so many best ETL tools available in the market but Informatica PowerCenter is one of the best tools which is used in the ETL process

Informatica PowerCenter

Informatica is the best ETL apparatus in the commercial centre It can remove information from various heterogeneous sources, changing them according to business needs and stacking to target tables. It's utilized in Data movement and stacking projects

Informatica is one of the Software development companies, which offers data integration products. It offers items for ETL, data covering, data Quality, data replication, data virtualization, master data management, and so forth.

Informatica nearly talks with all significant information sources (centralized computer/RDBMS/Flat Files/XML/VSM/SAP and so forth), can move/change information between them. It can move huge volumes of data in a very operational way, many times better than even bespoke programs written for specific data movement only.

Informatica PowerCenter is used for Data integration. It offers the ability to interface and brings information from various heterogeneous source and handling of information.

For instance, you can associate with a SQL Server Database and Oracle Database both and can coordinate the information into a third framework. The well-known customers involving Informatica PowerCenter as an information coordination device are U.S Air Force, Allianz, Samsung, and so forth. The popular tools available in the market in competition to Informatica are IBM Data stage, Oracle OWB, Microsoft SSIS, Skyvia.

Let us consider one example which works with a tool Informatica PowerCenter

Let us consider We have a flat file that contains data about different products.it stores details like the name of the product, its description, category, date of expiry, price, etc.

The user requires to fetch each product record from the file, generate a unique product id corresponding to each record and load it into the target database table. There are several conditions products which either belong to the category 'C' or whose expiry date is less than the current date.

Product_name	Prod_description	Prod_category	Prod_expiry_date	Prod_price(in Rs.)
ABC	This is product ABC.	M	8/14/2017	150
DEF	This is product DEF.	S	6/10/2018	700
XYZ	This is product XYZ.	M	8/14/2016	1000
PQRS	This is product PQRS.	M	5/23/2019	1500
GHI	This is product GHI.	C	8/14/2017	600

VIII. FLAT FILE

Based on the condition stated above, the database table (Target) should look like this:

Table name: Tbl_Product

Prod_ID (Primary Key)	Product_name	Prod_description	Prod_category	Prod_expiry_date	Prod_price
1001	ABC	This is product ABC.	M	8/14/2017	150
1002	DEF	This is product DEF.	S	6/10/2018	700
1003	PQRS	This is product PQRS.	M	5/23/2019	1500

Presently, say, we have fostered an Informatica work process to get the answer for my ETL prerequisites. The hidden Informatica planning will peruse information from the level record, go the information through a switch change that will dispose of columns which either have item class as 'c' or expiry date, then I will be using a sequence generate to create the unique primary key values for Prod ID column in Product Table.

Finally, the records will be loaded to the Product table which is the target for Informatica mapping.

Informatica Mapping addresses the information stream between the Source and target tables or we can basically say that it characterizes the principles for information Transformation.

A. Why Informatica is the best ETL tool compared to others?

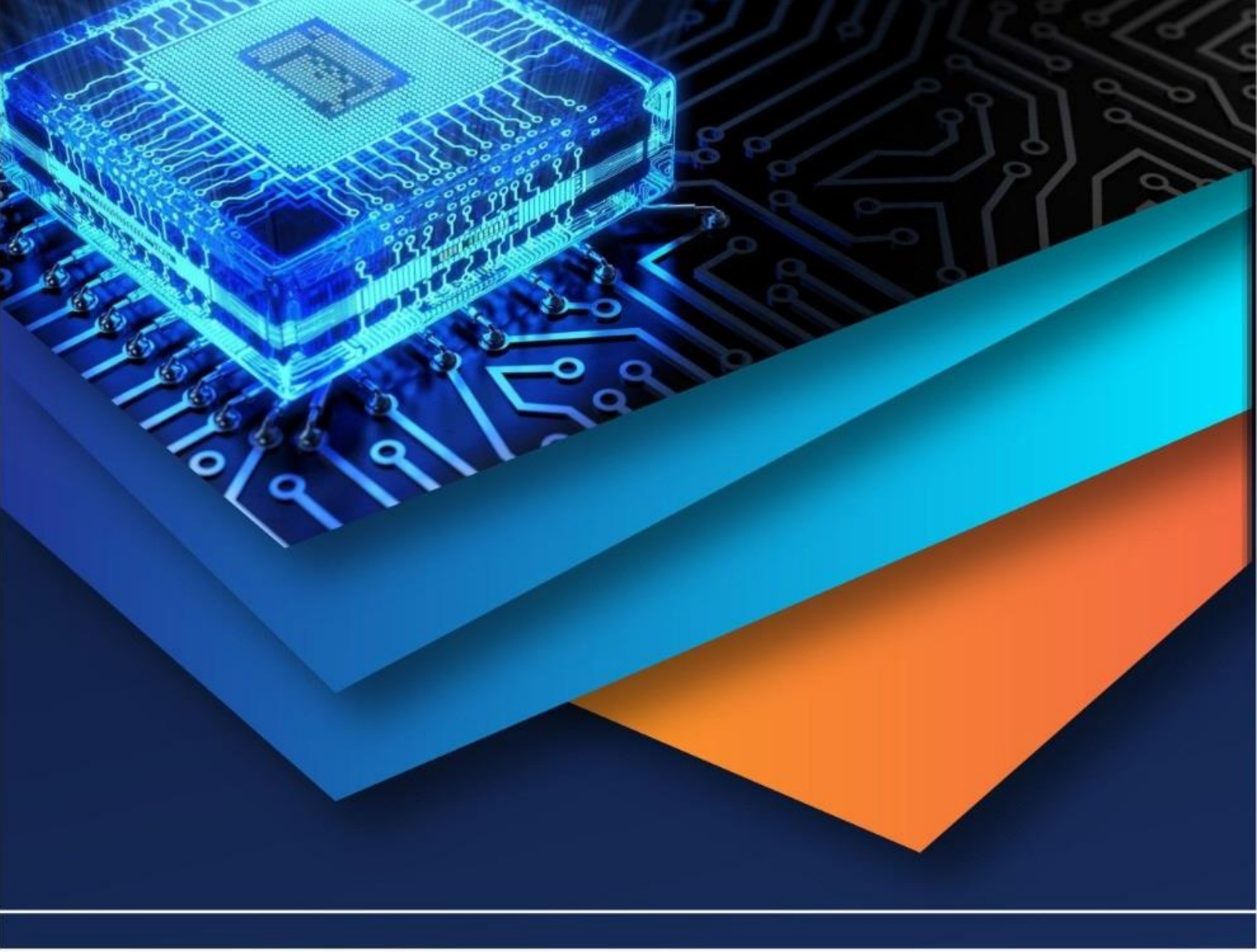
ETL tools are the better way to handle the database and Data Warehouse. There are several good ETL tools available in the market which we had seen But still, Informatica is one of the best ETL tools, it is the most used ETL tool. We will analyse why Informatica is the best ETL tool. There are several features by which we can say that Informatica is a best ETL tool

1) *Integration:* Informatica without a doubt is the market's leading data integration platform. It is a highly efficient data integration solution that can integrate more data in less time compared to any other solution. One key reason for Informatica's success is its ability to enable lean Integration. One significant justification for Informatica's achievement is its ability to empower Lean Integration, Lean assembling is the normal idea in the assembling industry to stay away from squandering.
2) *High Performance:* Informatica uses advanced technology to optimize performance in terms of quality, speed, and cost that enable companies to keep up with SLAs while transforming the business processes utilizing automation, reusability, and debugging. Informatica creates an environment with is quicker for analysts to perform different analyses, and is much easier to maintain. Informatica assists with adjusting the heap between the data set box and ETL server, with coding capacity. It has a High-speed loading of target data warehouses.
3) *Support For Different Databases and Data Types:* Different databases and data types get support from Informatica. Regular ODBC drivers, Teradata, Parallel Transporter as well as Fast Load are some examples. Informatica supports different data types thus providing the flexibility for the ETL process it can handle enterprise data type
4) *Maintenance:* Using Informatica Workflow Monitor, monitoring jobs is mighty easy. Identification and recovery in case of slow-running jobs or failed jobs are easier. The great feature of the ability to restart from failure row/step is handy. Features like runtime monitoring and automatic job logging make Informatica ideal for BI-managed services projects.
5) *Error Handling:* Informatica gives a brought-together blunder logging framework that works with logging mistakes and dismissing information into social tables or level records adequately, further empowering your specialized group to survey and approve the blunders. Informatica makes available a centralized error logging system that makes logging errors and rejecting data into relational tables or flat files effortless, enabling the technical team to review and authenticate the errors.
6) *Training and User Friendly:* Easy to learn with no programming. Easy training available and tool availability has made easy resource availability for the software industry, This helps companies in reducing training costs.
7) *Cost-Effective:* Informatica ETL isn't that costly device, what other place apparatuses like stomach muscle initio is pricey which enjoys many added benefits in a specialized viewpoint. Same time others ETL apparatuses are having difficulties like convenience, re-ease of use, troubleshooting, availability which makes Informatica an ideal ETL device.

Informatica has many advantages over other tools. But still, there are lots of options available in the market, we can choose the ETL tool which is best according to requirement. Which can also help to improve the business ability.

IX. CONCLUSION

As the ETL process plays the main role in Big data processing. ETL processes are a very important research problem As we have discussed the process of ETL in detail and also we focused on various ETL Tools There are several commercial and open-source ETL tools available in the market. By analysing all tools, we found that Informatica PowerCenter is mostly the preferred tool used in data processing Which is one of the best tools available today. The reason behind that it makes the data processing easier and faster it is cost-effective and this tool is the best solution in large enterprises because it is information base unbiased and consequently, it can speak with any data set the most impressive information changes device. It can be integrated with other tools if required.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY