Sweety Patel, Piyush Patel, Saumil Patel / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 2,Mar-Apr 2012, pp.906-908

# OVERVIEW OF ETL PROCESS WITH ITS IMPORTANT

**Sweety Patel[1], Piyush Patel[2], Saumil Patel[3]**
[1]Department of Computer Science, Fairleigh Dickinson University, NJ- 07666, USA
[2, 3] Department of Computer Science, Rajasthan Technical University, India

**ABSTRACT - While integrating different heterogeneous platform not only considering data file format and type but involving operating system, hardware require to map data for further usage for making out decision and report generation. ETL stands for extraction, transformation and loading where extraction done from operational data source ,transforming involve cleansing, filtering, validation and finally apply business governing rules and loading loads data to application database where it actually required to reside.**

**KEYWORDS – ETL process, extraction, transformation, loading**

## I. INTRODUCTION

1.1 Process contains, to get data out from one source and load it to data warehouse. It is a process of applying data from source to destination databases it means from one data base to another database. In many data warehouses, it deals on text files and spreadsheets where it leads to make a use of extraction, transformation and loading. ETL is not a physical implementation but it's a process. It is a complex combination of the process and many technologies that requires significant portion of data warehouse development efforts. Skilled development of the business analyst from database designer and application developers. Second important thing to be considered is it is not a onetime process. Adding data of one database to another by single time make it finished process but it requires lot of work on regular basis as monthly, yearly, daily, hourly quarterly and so on. ETL is an integral part of data warehouse. It is occurred frequently and continuous with repeated time frame in data warehouse.ETL is an automated process and if once defined in well format it runs automatically as per the schedule of the process.

1.2 It is well documented, if it is taken in brief then data source in text file or in spread sheet become easily readable, editable format after the ETL process. That makes it easy to use and put it to a highest corner for a working it on as effectively as well as efficiently, easily changeable. After ETL process data is in such a form that can be easily changeable to required steps.

1.3 ETL operation should be performed on a separate relational database server. ETL operation is never performed on source database and also not on a data warehouse data by ETL process. If ETL process is not conducted automatically on periodic basis then it make for developer or data workers hard to work on data for transferring from source database to data warehouse. It is a time consuming method, nobody can define fix time slot for this work as a this work has to be done all previous methodology for transferring data from data source to data warehouse and also sometimes it deals with upcoming new problems that makes it a hard to transfer every time from source database to destination database.

1.4 ETL make above process scheduled. So any transformation which done previously successfully that does not again any extra hard work procedure to transform it. But many new problems take extra time to split it to define answers as once it is defined, it again become part of periodic ETL scheduled process.

## II. EXTRACTION

In fig.1 data is extracted from the heterogeneous data source. Above define to different database platforms. Data need to be integrated into one and then resulted data should be captured. If all of the data are on different or disparate system for single or multiple enterprise need to make it usable then it is impossible to make it in workably simple usable format without ETL process. It is really challengeable thing to be put data on usable and or one platform as ETL make it simpler. Each data source has its own distinct set of characteristics that should be followed when it is extracted transform and integrated. For getting perfect result it has to be followed by all the rules otherwise it is harder to get true result back on the basis of perfection.

ETL process required to integrate systems which contain different DBMS, Operating System, hardware and communication protocol.

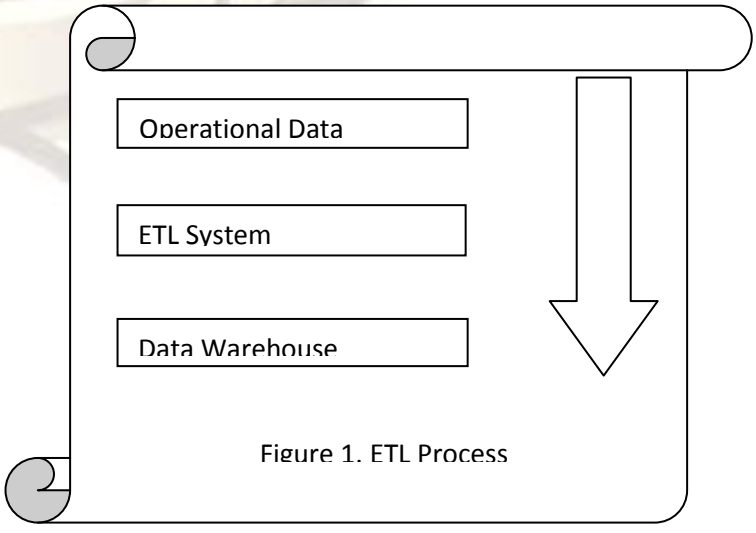


Figure 1. ETL Process

Logical mapping is most important while transferring a data and make it to be integrated and then physical mapping of data is possibly done.

Logical data map describe the relationship between the first starting points to extreme ending points of ETL system. Usually this ETL system presented in table or spreadsheet that makes it easy to map and checking procedure to mapping.

| Target | | | Source | | | Transfor--mation |
|---|---|---|---|---|---|---|
| Table Name | Column Name | Data Type | Table Name | Column Name | Data Type | |

Figure2. Transformation Mapping Process table in Database

Logical data mapping is not sufficient enough only common element of data but it is also required to map a critical element to be mapped efficiently in ETL plan. This Primary rule of fig.2 is providing clear-cut blueprint of exactly what is required for ETL process. This table must not have any questions and transformation must be involved what action has to be taken. Data Extraction (Analysis of the system) broken in to two phases, data discovery phase and second anomaly detection phase.

Once a target is finalized like what exactly it should required on, at the end point then it is easy to make a data discovery phase to be easier for the developer side.

## III. DATA TRANSFORMATION

The process performed in the staging area it is main important step of the ETL where values are added by ETL Process. Data quality check is done for transformation process like, correct, unambiguous, consistent, complex.
Cleaning data required the step which invoiced first anomaly detection as per example sampling with count (*) of the rows from the required table of source to destination.

Column Property rules should not be NULL at any time .Numeric value should be fall it into the define range not to high not too low. Column length must not be to short too long. Column which required some fixed patterned value set must not be outside of the set like weekdays not be a month of January.

Structure must be ruled of table, proper primary and foreign keys are defined on table.

Data values must be rule. Simple business rules should be followed by transaction process.

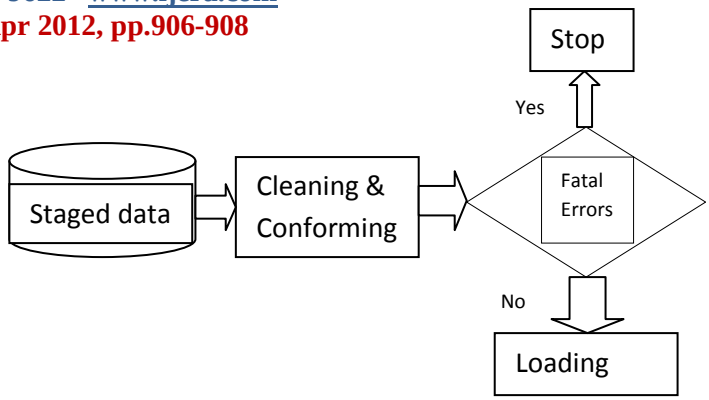


Figure 3 Data Transformation Process

Fig.3 shows a data transformation process clearly where fetal error data must be stopped before they go for loading.

## IV. LOADING

Starting point of staging area which containing a clean and conformed package of data ready for an upload to the data warehouse. This Package should not be logically connected. It means as per fig.4 it represents a snapshot as of valid time with social specific time limit with structural integrity constrains such as entity integrity and referential integrity must be hold by data.

Logically loading establishes the most resent picture of the history database even if dimension design is not followed by rigidly followed, we can distinguish between dimension data which may lead to insert and updating that data which should required leading insertion.

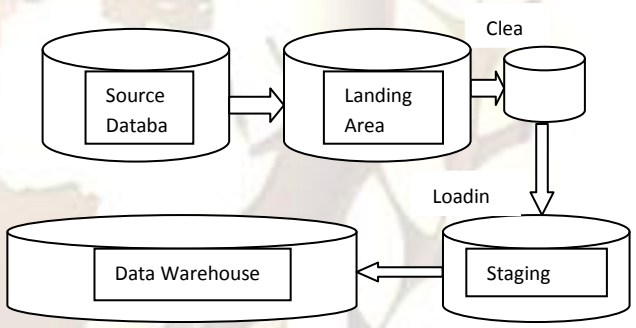


Figure 4 ETL Process with showing loading.

Loading requires all foreign keys mapping to the appropriate surrogate.

## V. CONCLUSION

As mentioned ETL is a very required process for making different data source to pull at one end. If it is conducted not effectively then too much cost wastage behind that. But may intelligence of ETL process may lead of burden of developer reduced & make it is a easy, simple and use scheduling again for repetitively do a same procedure on other data with periodically fixed time interval .ETL tools are there for make a ETL process easy, including some of them are,

Informatica - Power Center

IBM - Websphere DataStage(Formerly known as Ascential DataStage)

SAP - BusinessObjects Data Integrator

IBM – Cognos Datastream

Microsoft - SQL Server Integration Services

Oracle - Data

SAS - Data Integration Studio

Oracle - Warehouse Builder

AB Initio

Information Builders - Data Migrator

Pentaho - Pentaho Data IntegrationEmbarcadero Technologies - DT/Studio

IKAN - ETL4ALL

IBM - DB2 Warehouse Edition

Pervasive - Data Integrator

ETL Solutions Ltd. - Transformation Manager

Group 1 Software (Sagent) - DataFlow

Sybase - Data Integrated Suite ETL

## REFRENCES

### BOOKS

[1] Nong Ye, The Handbook of Data Mining (Lawrence Erlbaum Associates, Mahwah, NJ. Publication, 2003).

[2] Jiawei Han and Micheline Kamber, Data Mining:Concepts and Techniques (Morgan Kaufmann Publishers, University of Illinois at Urbana-Champaign).

[3] Bharat Bhushan Agarwal and Sumit Prakash Taval, Data Mining and Data Warehousing (Laxmi Publications, New Delhi - 110002, India).

[4] Ralph Kimball,Joe Caserta, *Data Warehouse. ETL Toolkit. Practical Techniques for. Extracting, Cleaning,. Conforming, and. Delivering Data* (Wisely Publication,Inc).