Sweety Patel, Piyush Patel, Saumil Patel / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp.909-911

WHY TRANSFORMATION IS IN DATA WAREHOUSE OR IN ETL PROCESS?

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 – ETL stands for extraction transformation and loading where transformation make a bridge between extraction and loading with process of transformation from one extreme point of data value to end extreme point of data value. Without transformation, extraction and loading can't be stand in whole ETL cycle and that a non integral part of ETL cycle. Transformation may need data to be consistent with all required respect to the databases values and make it efficient proof to work on ETL process. Extraction only extract data from the database and loading loads data to other end point of data destination but transformation make it workable of data from extraction to loading with required process of transforming data to one end to other end.

KEYWORDS – ETL process, extraction, transformation, loading

I. INTRODUCTION

1.1 Transformation is nothing but operation of data movement from one system to another system. In data warehouse system the most common requirements for transformation are in moving data,
1.1.1. From source to staging area
1.1.2. From source to data warehousing system
1.1.3. From staging area to data warehouse and
1.1.4. From data warehouse to data mart

1.2 As it is a simple but most important part of ETL and also in data warehouse. If from 100 percent of ETL process, 40 percent work is allocated for the transformation as it is digging process in data for make it to proper format as considering data layout with respect to transforming one data to required destination layout form of data.

All transformation types are below.

II. CREATING COMMON KEYS

Different keys on different system may represent same entity at that moment. Master data management required as even it is tedious to get it done with different mapping of keys in the heterogeneous or in different system.

2.1 SURROGATE KEYS CREATION

Creating common keys is important to link in production system. Surrogate key is used in star schema or in data from loaded one. To create a surrogate key is not simple method but it is very complex method. If we don't go with lower step of the floor than creation of surrogate key involves many problems and choose it with below criteria.

2.11 Choose dimension as you wish to create a key.
2.12 Find out maximum instances of chosen set of dimension.
2.13 Choose serial number, (combination of numbers and character) for dimension.
2.14 Dimension product key should be mapped with key range of serial number with same specific order like descending or ascending or any other specific series wise.

2.2 TEXTUAL ATTRIBUTES AND KEY DESCRIPTION MUST BE IN STANDARD FORMAT

Making of standard version for attributes description works in efficient manner to get data effectively without much more confusion, but, in whole system, from production, development, in channel and hosting contents different name to be given them, make confusing and make lots of extra efforts to make them standard also if not standardized then lot of work to be put to mapping in logical way to get data out from one extreme starting point to extreme end point with cost and time consideration.

2.3 FOLLOW STANDARDIZATION IN ORGANIZATION THOROUGHLY FOR STANDARD AND STRUCTURE

Make unique form for different accounting terms. Operational entity may contain different norms of ending period of year or different ending cut offs for reports. Currency conversation may not be ignored as it important not only to money wise but also make it suitable for human being to deal it with as per its connivance.

In the analysis phase, where 'As is' and 'To be' value must be defined for above all working areas. To make a standard form unique which identity for each textual as well as defined criterion, it is really hard but once it mapped with standard format thoroughly in organization then it is make all other remaining process technically easy and efficient to work.

III. DATA QUALITY CHECKS

3.1 DATA QUALITY

It is a first step of transformation process, data quality include data cleansing which include manipulating , aggregate, data management of

deployed data, duplication data removed, mapping of one formatted data to another specific value of data.

Work out for short forms of data like a' bad to Ahmadabad. All these small issue must be considered while approaching data to transformation state.

Smaller these issues make transformation process ugly in the source code, & make it harder to understand created problem because of mentioned above problem areas.

3.2 RELEVANCE OF DATA FOR QUALITY CHECKING OF DATA

Data which required actual transformation only included as it is handled to work on unnecessary data if is not recommended or required to be transform. As mobile number validation may not include zip code data in transformation.

If analysis is done through selected set of entities then it must be ignored as once it is not required.

3.3 MERGING OF DATA

In some of the time in transformation, on many to one process required for mapping where actually it is used as, like, girl and women both should follow the category female. So which data should merge and on which basic from it merged is must be commenced while in transforming procedure.

3.4 DATA TYPE CONVERSION

Data type conversion is most important thing where it is a critical part of the ETL process.

India's zip code is in numeric may lead confusion with other countries zip code as it is in alpha numeric may lead confusion of data in transaction with final end of production system and so on. Attention on data conversion process lead towards direction or goal in transformation process.

IV. DATA SETS WITH DATA TRANSFORMATION FOR PRESENTATION PURPOSE

4.1 DE-NORMALIZATION

Not in all but in some cases de-normalization is also required to get data to be easily from one data domain than all disparate domains of data it leads time and efforts to be much more little at one end up to considering smaller database length to bigger one.

4.2 NORMALIZAION

Production system never go for 100% normalization, as it takes higher efforts with time length to get data back from data domain and make it harder for developer to create a easy query(in result wise) and so normalization up to one end is must but not for 100% and for each time.

4.3 DERIVED ATTRIBUTES CREATED

Systems never contain the entire dimension with the source as a required data. Some data are also derived to work as week of day from given date.

4.4 CALCULATION OF DATA VALUE

Sometimes data values are calculated as it is not easy to store all result and all values in database. Pre-stored values are easy to get & print it on paper but not easy to store in all times may leads to do calculation at run time with mathematical or other logical operation in report from data.

V. DESIGN OF TRANFORMATION PROCESS

For given types of transformations, is not required to follow one time and with one directional method but contain thousands of way to reach it to destination type of required formatted transformation. Still some rules are there may make easier this produce as before.

5.1 CREATION OF PRE-DATA SET

List of all transformation code, list of input values to transformations, set of activity done on transformations, which are the output fields, table and their specific specification data?

Create all above in advanced to make up the transformation procedure easy and workable.

VI. FIND DATA SET CREATION FOR MAKE TO LOAD

Data is created to make it loaded into production as a final destination of ETL process as in loading process.

6.1 POST COMPLETION DATA QUALITY & COMPLETENESS CHECKS

Quality completeness makes it 100% over on its post data completion process.

6.2 MATCH

For match, search is pre required for appropriate data for matching of data. Match and search is done to be effectively to get data easy and simply to work on it.

VII. CONCLUSION

So in ETL, mainly, transaction introduces magic and makes it surely workable for thoroughly to the entire organization for all high level operations as they are most fundamental part of the system for making ETL as a complete one cycle. Transaction is most critical and also simple once it is defined structurally unique across whole organization as otherwise it is harder to even find out simple error which may be occurred as of data mapping. Data transactions join extraction and loading and make it as a bridge to pass data from source to destination.

REFERENCES

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).