Researchers

Sweety Patel

Department of Computer Science, Fairleigh Dickinson University, USA

Mrudang D. Pandya

Ganpat University, Ganpat Vidyanagar, Mehsana, Gujarat, INDIA

Email-
sweetu83patel@yahoo.com

mrudangpandya007@gmail.com

How is Extraction important in ETL process?

Abstract:

In ETL there are three main phase: Extraction, Transformation, Loading. Where all process stands on Extraction as to extract a data from different databases to one database required most of the work done to be in correct direction as it is hard to work on the incorrect data with all wastage of time, cost and also human efforts. So extraction of data is also main important thing to be delivered to the database as it is a foundation for all other phases of ETL cycle. Extraction process is described in detail in the paper.

KEYWORDS: ETL process, Extraction, Transformation, Loading.

Introduction:

In ETL there are three main phase: Extraction, Transformation, Loading. Where all process stands on Extraction as to extract a data from different databases to one database required most of the work done to be in correct direction as it is hard to work on the incorrect data with all wastage of time, cost and also human efforts. So extraction of data is also main important thing to be delivered to the database as it is a foundation for all other phases of ETL cycle. Extraction process is described in detail in the paper.

Extraction

Out of three major elements of ETL, extraction is also the most important part of the ETL process. ETL stands for Extract transform and loading, where ETL is out of the most important area of business ct intelligence as it is allow creating warehouse in fast reliable way. Appropriate extraction process is the key process to the creation of data warehouse and data mart. Whole ETL process outline is shown in fig.1.

Creation of the logical data maps:

Important and required to do as a first step is, create a logical data map for the extract system design. The logical data map document generally includes the spreadsheet or table format which contain mapping of data from primary source to final point of data source. The logical data map document sometimes also referred as lineage report.

Linear report may contain following data as in it, as a part of the document.

- Components of the target table which content target table name, column name, table type, fact, dimensions or sub dimension, data type.
- Component of the source table like the source database, table name, column name, data type.

VOLUME 1 NUMBER 5 (Sep/Oct 2012)
ISSN: 2277–1891

© IJAITI 2012

- Transformation type with all perfect detail let's go for detail one for above monitored area.

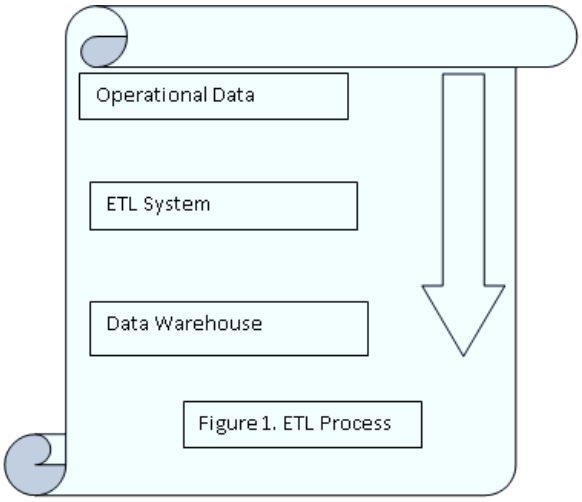

Fig.1 Target Component

The four component of target are compulsory while fifth is optional and that is SCD is defined as dimension which change with time. It is one type of indicator. It is translated as slowly changing direction. There are three types of SCD:

- SCD 1:- which overwrite old data with new data.
- SCD 2:- tracks historical data by creating multiple recodes in a dimensional table and its unlimited.
- SCD 3:- tracks changes using different separate columns and it is limited by number of columns as we designed.

The SCD type may be different for each Colum in the dimensions.

- *Source Component*: The source database is name of instance from where actual data is extracted which required as a initial point of ETL cycle.
- *Transformation*: The transformation is a kind of conversion on data as per the required format by the application it can be expressed in SQL or pseudo code. Sometimes transformation in a logical data map table is blank meaning that it is not required to do any transformation on that particular mapped data.

Advantages of Logical Data Map

- It returns all information to the one place from different desperate system so it improves extraction process.
- It make advantageous to improve information and function analysis.
- It is most important point of developers to do a work on extraction procedure as it is work as source of data and once source of data is not correct then target data result will be 100% incorrect. That's way it lead to do a hard work on getting data perfect in source component.
- Sometimes data are extracted and on that length may be decrease and so if data is 100 character and need to be directed up to 60 characters. At that time transformation procedure play important role and here we required particular tool that compromise with data length from source to target as per the specific requirement.

VOLUME 1 NUMBER 5 (Sep/Oct 2012)
ISSN: 2277–1891
© IJAITI 2012

How to Build the Logical Data Maps?

Before starting building of logical map we required to know how exactly the targeted data look like.

- *Identify the data source*: The data modeling session may be helpful in many times but it only indicates the major source systems. It depends on the team of developer that each and every smaller set of data source is extracted as an input to ETL process. However identifying a data source may be sometimes lead quit complex work. As it is not required that all actual data as a source with correct format and correct name is very complicated task. Solution for this problem is make all database as a central repository for getting it with same name in entire organization.
- Collecting and documenting source system which can work as final source of data for extraction process.
- Creating that source system which tracks the report it shows information about who is responsible for each of the source. It content many characteristics of the source like sub area, interface name, business complexity, transaction per day, priority, daily use count, department use, business use, platforms and some comments.
- Creating and analyzing the E-R (Entity relationship) diagram. It shows how different entities are related to each other.

E-R Diagram

It contains unique identifier: status_id, status_code, status_description.

- *Data types*: relationship between papers it is very important relationship characteristic and need for appropriate data extraction.
- *Discrete relationship*: It is necessary for mapping many of the dimensions. It is a single look up table storing all data from all the table make it central to work on that so data access becomes easy to work on that.
- *Cardinality of relationship and columns*: It is no. of element in diamond that is related. There are three cardinality types:
    - One to One- Relation of primary key between tables.
    - One to many-foreign key columns in tables.
    - Many to many- It is usually involves three tables with two one to many cardinality types.

After preparation of logical data map:

- Analyze the data content which allows making us before choice of data for the extraction procedure. But there are some traps which are solving as: example: if data field value is not as a date in date field on null then we must loose data .We are doing joining. At that time it required to do an outer join time it require to do an outer join in tables.
- Business rules must be complete for ETL process makes 100% workable.
- Integrate data source to single data source may lead to efficient working procedure. But if dimensions are entirely different then we are failed for integration. So this should be taken considered in to the system working procedure.
- The logical data map is nothing but specification for the physical plan of work. It shows the main point for the attention in source and target database. It sometimes gives efficient result. It may be also used as information source and be presented to end users.

Conclusions

As considering logical data map for the extraction process many efficient efforts are required to extract a data from the different databases to make an ETL cycle complete. As extraction process is done on many data bases with different format of databases may require much more concentration on the extraction process so logical mapping of data is done effectively on the database from source to destination. Exact extraction of the data from the database is the main advantageous to all other phase in ETL cycle as work load and work process is

VOLUME 1 NUMBER 5 (Sep/Oct 2012)
ISSN: 2277–1891
© IJAITI 2012

done on the right data with the right direction of the work procedure. Extraction is an initial point of the database ETL cycle where foundation is not right than building structure is also not as per the requirement and needed to be effective work to make out it as a main critical part of the ETL process.

References:

[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).

Author Details:

Sweety Patel: Department of Computer Science, Fairleigh Dickinson University, NJ- 07666, USA

Mrudang D. Pandya: GANPAT UNIVERSITY, Ganpat Vidyanagar,Mehsana-Gozaria HighwayMehsana - 384012, GUJARAT, INDIA

© IJAITI 2012