Analysis on Extract, Transform and Load (ETL)

Nidhi Arora¹, Nitika Batham², Swati Madan³ and Ritu Hans⁴
¹,²,³,⁴HMR Institute of Technology and Management, Guru Gobind Singh Indraprastha University.
¹nidhiarora.gimt@gmail.com, ²demure.neeti@gmail.com, ³swatimadan26@gmail.com and ⁴ritu_hans@ymail.com

ABSTRACT
Business enterprises invest lots of money to develop data warehouse that gives them real, constant and up to date data for decision making. For processing large volume of data and accessing to heterogeneous multi-data sources, how to efficiently shorten the execution time and possess the ability of general accessing data sources is a big challenge. Data warehouses are typically assembled from a variety of data sources with different formats and purposes. As such, Extraction Transformation Loading (ETL) is a key process to bring all the data together in a standard, homogenous environment. These are the tools used to extract the data from heterogeneous distributed databases, clean it, transform it and load into data warehouses. The time available to extract data from source systems may change, which may mean the same amount of data may have to be processed in less time. Some ETL systems have to scale to process terabytes of data to update data warehouses with tens of terabytes of data.

This research paper will analyze the ETL process. This paper proposes a solution of a general ETL process, realizing such functions as accessing several heterogeneous data sources and monitoring ETL process. The paper results prove this solution to be efficient and valid. The goal of extract, transform and load is to increase the productivity, quality and time to market in Data warehouse. It will allow flexibility for job execution and scheduling. It will provide better performance if coded properly, and can take advantage of parallel processing capabilities when need arises.

KEYWORDS
Data Warehouse, Extraction, transformation and Loading (ETL)

1.INTRODUCTION
As competition is becoming more violent, businesses are getting more complex and multifaceted. In such competitive environment executives and managers want their company to maintain its position in market and meet the industry benchmark. The executives of the companies, therefore, wanted some important facts information which can help them in making some useful and strategic decision in favors of their company’s upliftment. The data for making strategic decision is gathered from the databases maintained for every company. But before using these raw data it is converted into useful information using Extraction, Transformation and Loading (ETL) process.

ETL processes are responsible for reshaping the data of the source system into useful information that is stored in the data warehouse. These are the requirement for the data warehouse and are the back-end process. Therefore its function of the ETL to maintain updated information extracted from the database.

2.ETL ENVIRONMENT
ETL is a step wise step process starting from extraction, then transformation of data and finally loading the obtained information. Extraction is the operation of extracting data from the source systems for further use in data warehouse environment. Transformation is improving the quality of the data to make it more useful. It deals with reformatting, restructuring etc. At last is the loading process that includes the loading of the information in data warehouse.
The very first part of ETL process is extraction. The extraction process involves extraction of the data from all the source system. These source system can be different format like one source system is simply a flat file and other source system can be Relational Database Management System (RDBMS). The varieties of formats necessitate the parsing of extracted data to determine their cohesiveness and conformity with underlying structures. Extraction is very time-consuming job. This is because here it is very important to determine which data source must be extracted. Sometime it is needed to extract the data from the source more than once to keep up-to-date data and this extraction, therefore, must be done from time to time. Here it must be noted that there is no effect on source data and the performance of the source.

EXTRACTION

Logically Extraction Methods
In logical extraction there can again be two types of extraction:-
- Full Extraction
- Incremental Extraction
In full extraction the whole (full) data of the data source is extracted assuming that the available data is indeed complete. This method is of course very simple and easy to do. On the other hand, there is Incremental Extraction, in which only that portion of data is extracted that is updated or that has changed since a marked historical event. Therefore to perform incremental method of extraction it needs to incorporate some of the change capture technique which is an important consideration for extraction. This extraction is better than full extraction as it is efficient and it reduces the overhead of loading the whole data source instead it require to identify those data that is most recently changed. But change capture technique is a big challenge and quite typical. Below are the several techniques for implementing a self-developed change capture on Oracle source systems:
- Timestamps: - Some operational system tables have timestamp columns. The timestamp specifies the time and date that a given rows was last modified which helps in identifying the latest data.
- Partitioning: - In this source tables are partitioned along the date keys. Like one partition is on the basis of identifying current weeks’ data of the tables.
- Triggers: - Triggers can be created in operational systems to keep track of recently updated records. Triggers are the stored procedures. They can be used in conjunction with timestamp columns to identify the exact time and when the given row was last modified.

Physically Extraction Methods
On the basis of the chosen logical extraction method and the capabilities and restrictions on the source side, the data can be physically extracted. In physical extraction there are two types of extraction:
- Online Extraction
- Offline Extraction
The extraction process can connect directly to the source system to access the source tables themselves or to an intermediate system that stores the data in a preconfigured manner in online extraction. In offline extraction the data is staged openly outside the original source system. The data already has an existing structure or was created by extraction routine. Here following structure are considered:-
Analysis on Extract, Transform and Load (ETL)

- Flat files: data in defined and generic format.
- Dump files: oracle specific formats.
- Redo and archive logs: Information is in a special, additional dump file.
- Transportable table spaces: A powerful way to extract and move large volumes of data between Oracle databases.

2.2 CHALLENGES IN EXTRACTION PROCESS
Following are the problem faced in extraction process:-
1) Diversity of the source system is the greatest issue for the problem faced in extraction as in the large system where we have the network protocols, database management system, combination of computing platforms and the operating system.
2) Time available to extract the data from the source system may vary, which means that the small amount of data may have to be processed in the less time or tones of data are required to be processed within small time interval.
3) Data has to be extracted not once, but number of times.

2.3 TRANSFORMATION
Data transformation deals with preparing the extracted data for loading to the target database or databases such that they do not violate any business rules as data sources are dissimilar. By doing transformation the quality of the data is improved. Unequal transformation is applied to every data item i.e.; if some data is according to the rule then very less or no transformation is required otherwise transformation is used to restructured, cleaned, and organized before transportation. The ratio of particular effort of data transformation should be as high as 80 percent of enforcing business data domain rules and business data integrity rules and just 20 percent of technical rules of conversion of the data. Therefore primary estimated ETL transformation efforts should be multiplied by four and on may expect missing deadlines due to the amount of dirty data.

The transformation process includes:-

Selection: Selection of the data from the source system forms part of the extraction function itself. But in few cases the composition of the source structure may not amenable to selection of the necessary parts during data extraction.

Splitting/joining: This task includes the types of data manipulation, selecting relevant information only.

Conversion: It involves converting the extracted data into single format by applying the standardization process explained below.

Summarization: it is required to add the data at its lowest level. Lowest granularity data is very helpful for the users in query processing.

2.3.1 Transformation Methods
Typical methods of data transformation include:
1) Sorting: - Sorting of the data means of arranging the data in any order either ascending or descending order. This would increase the understandability of the data. Sorting and merging of data takes place in large amount in data staging area.
2) Filtering: - Filtering process involve the fragmentation of the needed data from the unwanted data. This is because it is not necessary that all the data of any table or file have relevant data.
3) Cleansing: - it is process in which data that violates the business rules is changed to obey these rules. It is usually done by ETL programs that determine or find the correct values and then write them to target databases.
4) Aggregation: - in this data values from various data tables are bring together to roll up and aggregate. Aggregation merely summarizes of the most granular data higher levels along the dimension hierarchies.
5) Pivoting: - It is the process of converting the relational tables into more concise and more typical relational form.
6) Standardization: - Standardization of data means standardizing data types and field lengths for same data elements. In this there is semantic standardization which is also important task. Resolving of synonyms means that two or more terms from different source system source system having same meaning. On the other hand when a single term means many different things in different source systems, it is resolving homonym.

2.4 CHALLENGES IN TRANFORMATION PROCESS
The challenges which are faced during transformation process are as follows:
1) Transformation process’s greatest issue is to deal with heterogeneous data, that is, data of various formats.
2) Poor quality data that have been identified in an extraction phase make transformation process more difficult.
3) Transformation deals with 40% of time of the whole ETL process.

4) If transformation is to be performed manually it requires lots of coding and to make them store metadata they are require to code another way. Maintaining metadata becomes more difficult.

5) Transformation process requires lot of knowledge to be performed.

2.5 LOADING

Loading is the process of moving transformed data into the data warehouse repository. It is the final job and the simplest job in ETL process. It is a straightforward process. The process changes from organization to organization. There are certain data warehouses which may choose to overwrite the data week to week. At the same time, there are also data warehouse which may choose to write data in hourly basis.

2.6 LOADING METHODS

The three types of loading processes are:-

Initial load: - populating all the data warehouse tables for the very first time.

Incremental Load: - applying ongoing changes as necessary in a periodic manner.

Full refresh: - completely erasing the contents of one or more tables and reloading with fresh data (initial table is a refresh of all the tables).

2.6.1 Problem Faced In Loading Process

Some of the problems which are faced in ETL process are as follows:-

1) The problem deals with the indexes, materialized views, defined over any data warehouse relation. Overhead of the indexes and materialized views are incurred if any update is made to data warehouse.

2) Discrimination of the records and inserted first time and the records and data that are acting as an update in the already loaded data is another issue. But this problem is solved by DBMS’s MERGE command).

3) Issue of efficient performance is also a problem which includes overheads of parsing insert statements, maintenance of log files and rollbacks on any failures.

4) Another problem in loading process is that it is time consuming as it may involve loading of millions of data in data warehouse. Loading and managing images is also not an easy task.

5) Dependencies of the data and tables also tend to reduce parallel loading operations.

6) Managing the load run is a challenging task.

CONCLUSION

In this paper, we discussed about the ETL process which is a kind of development process, the methods of Extraction, Transformation and Loading. We included some challenges that are faced in dealing with ETL process. The biggest challenges are that the process is time consuming and complex. For this reason we provide a solution of applying parallel processing with ETL process. The proposed mechanism saves the time and cost as compared to periodic update and other existing techniques for ETL. Thus the main contribution is to provide more flexible and effective ETL tool.

FUTURE SCOPE

The proposed analysis will be more effective and would be less time consuming. It will ensure that creating reports and business score card is much easier, enabling the business users to select the information relevant to them in the blink of an eye. A better approach of ETL process will help more because the reliability and stability of data warehouses built using a better ETL tool is much higher, since more aspects can be checked and monitored in relation to each other, metadata being a case in point. The new approach obtained through research will be more flexible and will implement by us on data.

REFERENCES

[2] Rachel Agheyisi “Business Intelligence and ETL.pdf”