Proceedings of the 2013 International Conference on
Pattern Recognition, Informatics and Mobile Engineering, February 21-22

Novel Approach in ETL

A. Prema
Department of Research and Development
Bharathiar University
Coimbatore, TamilNadu, India
latharaman2012jr@gmail.com

Dr. A. Pethalakshmi
Department of Computer Science
MVM Government Arts College for Women
Dindigul, TamilNadu, India.

***Abstract*-The term ETL which stands for Extraction, Transformation, and Loading is a batch or scheduled data integration process that includes extracting data from their operational or external data sources, transforming the data into an appropriate format, and loading the data into a data warehouse repository. Through the study of Extract, Transform, and Load (ETL) hypothesis, a new ETL is designed, which is titled as Hyper-ETL for increasing an efficiency of ETL process. Hyper-ETL allows the integration of XML document file and Oracle data warehouse to reduce an execution time and to remove the mismanagement of metadata in an existing ETL process.**

Key Words- ETL; metadata; XML; Oracle; and Data Warehouse

I. INTRODUCTION

This ETL Tool is used to simplify the process of migrating data, standardize the method of data migration, store all data transformation logic as Meta data which enable the users, managers and architects to understand, review, and modify the various interfaces and reduce the cost and effort associated with building interfaces. Extraction is the process of reading data from a specified source database and extracting a desired subset of data. Transformation phase applies a chain of rules or functions to the extracted data to derive the data to be loaded. Three forms of transformations are utilized, that is, Subsets of tables, Formatting Data and Primary Keys and Indexes. Subsets are created to remove personally individual information. All tables except the reference table are transferred to the Data warehouses using an ETL process. Primary keys are created to make sure uniqueness within a table and to facilitate the fusion of tables. Indexes are created to expedite queries. Loading is the process of writing the data into the target database.

The ETL process includes designing a target, transforming data for the target, scheduling and monitoring processes. The purpose of using ETL tools is to save time and make the whole process more reliable. The ETL tools are customized to provide the functionality to meet the enterprise requirements. Hence, many of them choose to build their own data warehouse themselves.

Section 2 of this paper deals with related work done in the Extract, Transformation and Loading into the data warehouses. Section 3 explains an actual process of Extract, Transform and Load. Section 4 explains the steps of proposed work to design an ETL Engine. In section 5 Experimental analysis, results are given, and finally, section 6 presents a conclusion of this paper.

II. RELATED WORKS

Different varieties of approaches for the integration of ETL tool in data warehouses have been proposed. A data warehouse gives a set of numeric values (called facts) that are based on a set of input values in the form of dimensions [6]. Over the years, data warehouse technology has been used for analysis and decision making in enterprises [4]. A concrete ETL service framework was proposed and talked about in metadata management service, metadata definition services, ETL transformation rules service, process definition service etc [3]. Two heuristic algorithms with greedy characteristics were proposed to reduce the execution cost of an ETL workflow [11]. Li Jian defeated the weak points of traditional Extract, Transform and Load ETL tool's architecture and proposed a three-layer-architecture based on metadata. That built an ETL process more flexible, multipurpose and efficient and finally they designed and implemented a new ETL tool for drilling data warehouse [7].

Lunan Li recommended intensively managing ETL by metadata repository and made metadata easier to understand, therefore metadata management became more direct, simple, and centered [8]. A systematic review method was proposed to identify, extract and analyze the main proposals on modeling conceptual ETL processes for Data Warehouses. The main proposals were identified and compared based on the features, activities and notation of ETL processes and concluded the study by reflecting on the approaches being studied and providing an updated skeleton for future study [9].Numeric values of a classical data warehouse can be difficult to understand for business users, or may be interpreted incorrectly [10].

Therefore, for a more accurate interpretation of numeric values, business users require an interpretation in meaningful non-numeric terms. However, if the transition between terms is crisp, true values cannot be measured and smooth, transition between classes cannot take place [1]. At last, definition method and related algorithms of ETL rules are designed and analyzed. A data mart contains data from a particular business area and multiple data – marts can form a data warehouse [5].

978-1-4673-5845-3/13/$31.00©2013 IEEE

III. EXTRACT, TRANSFORM AND LOAD (ETL)

ETL is the process to allow business to combine their data while moving it from source system to data warehouse. Data can be taken from any source. A detailed explanation of the ETL process is Extract, Transform and Load (ETL). The three data base functions are combined into one tool that automates the process to pull data out of one database into another database. Extraction is referred as extracting the data from various heterogeneous systems. Transform means applying the business rules on data which are derived from different sources. The process of pumping the data into the data warehouse for end user access is referred as "Loading" [2]. The Testing of ETL mainly deals with how, from, when, what and where we carry in our data base.

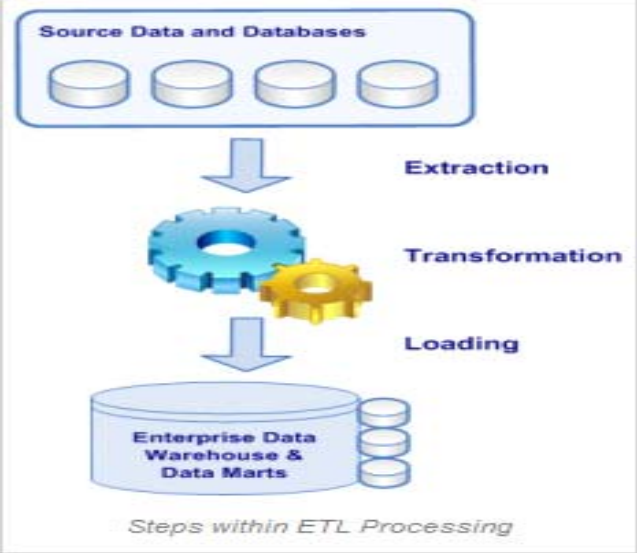


Figure 1. Process of ETL

The ETL tools [14] were created to improve and facilitate data warehousing. ETL eliminates the step of loading the text files into an intermediate storage, saving significant space and time. The purpose of using ETL Tools is to save the time and make the whole process more reliable.

The ETL process consists of the following steps:

1. Initiation
2. Build reference data
3. Extract from sources
4. Validate
5. Transform
6. Load into stages tables
7. Audit reports
8. Publish
9. Archive
10. Clean up.

Transformation has been applied to achieve migrating from one database to another the goals of the Data warehouse. End-users can access the data via several methods (i.e. ODBC, JDBC, OLE, etc.). The star schema is the simplest data ware house schema looks like entity – relationship model with points baking from a central table. The center of the star contains number of fact table. Snow-Flake Schema is a complicated data warehouse schema than a star schema. . This schema, normalizes data dimensions by grouping data into multiple tables rather than one giant table. All tables except the reference table are transferred to the Data warehouse using an ETL process [7]. Many of the tables are split into smaller tables in order to expedite queries. The ETL process [12] includes designing a target, mapping sources to target, extracting data from sources, transforming data for the target, scheduling and monitoring processes, and managing the overall BI environment. Benefits of an ETL tool [13] are given below:

- to simplify the process of migrating data
- to standardize the method of data migration
- to store all data transformation logic/rules as Meta data
- to enable Users, Managers and architects to understand, review, and modify the various interfaces.
- to reduce cost and effort associated with building interfaces.

IV. PROPOSED WORK

This paper, presents the design of Hyper- ETL. To design the Hyper ETL, the researcher used java, Oracle and XML. Here ETL rules are designed and analyzed to remove the mismanagement of metadata in ETL processes and also improve the ETL efficiency. SQL code generation optimizes the services of sales promotion. We created three tables namely Campaign, Product and Customer and this structure reduces the storage space by merging the three different tables into single table and split the column, based on the requirement of sales promotion. Hence, XML is used for Meta-Data structure.

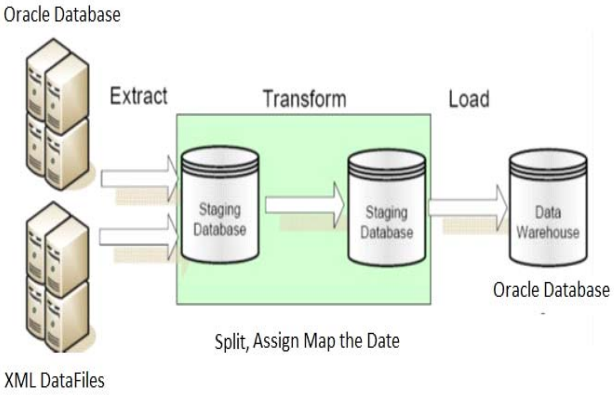


Figure 2.The process of Proposed Hyper ETL

The steps for designing the Hyper ETL are given below.

1. Extract the data from operational data source. Data extraction is one of the three main functionalities of the ETL tools. A main consideration to assess is the product's ability to extract from a variety of various data sources.

2. Creating Table with relevant attributes based on user requirement.

3. Transform it to fit operational needs. Generate the xml document file for the collected data.

4. Meta-Data for XML document file

This Research work implements three protocols namely Oracle Database, XML Data File and JDBC. The Protocol will be part of the url attribute of the target or source node. Every transformation will have a source and target.

<source url="xml://localhost/etl/test.xml“>

…

<target url="jdbc:oracle:thin:@localhost:1521:XE"

5. Eliminate the inconsistent data.

6. Split the table.

7. Assign the data.

8. Loading it into the end target. Pump the data into Oracle data warehouse. The loading phase is the last step of the ETL process. The information from data sources are loaded and stored in a form of tables. There are two types of tables in the database structure: Fact tables and Dimensions tables. Once the fact and dimension tables are loaded, it is time to improve the performance of the Business Intelligence data by creating Aggregates.

9. Audit reports

10. Publish

11. Archive

12. Clean up.

V. Experimental analysis and Result

We have proposed a new Hyper ETL for increasing the performance of the ETL which is different from the traditional ETL tool. This proposed Hyper ETL transforms the XML document file into the Oracle data base and we found that, 1 hour 57 minutes for nearly 2 millions records for which execution time is less than the previous one, were transformed and loaded into relevant tables. Experiments were conducted in Java, Oracle and XML language. The significance of this ETL was demonstrated through some sample Sales records and we used high configured system Intel® Xeon® E5-4600 Series Processor, 32GB DDR-III RAM for testing the Hyper ETL . In this paper, a systematic, an uncomplicated, and understandable Hyper ETL is proposed. Experimental and analysis results are given below.

Result :

A. Input Table : (Campaign, Product, Customer)

Table 1.

CAMPAIGN TABLE FIELDS

Column Name	Data Type	Nullable	Default	Primary Key
CAM_ID	NUMBER(4,0)	No	-	1
LOCATION	VARCHAR2(25)	Yes	-	-
START_DATE	DATE	Yes	-	-
END_DATE	DATE	Yes	-	-
CAM_NAME	VARCHAR2(20)	Yes	-	-
DESCRIPTION	VARCHAR2(50)	Yes	-	-

Table 2.

Product Table Fields

Column Name	Data Type	Nullable	Default	Primary Key
PROD_ID	NUMBER(4,0)	No	-	1
PROD_TYPE	VARCHAR2(50)	Yes	-	-
PROD_NAME	VARCHAR2(50)	Yes	-	-
PRICE	NUMBER(12,2)	Yes	-	-

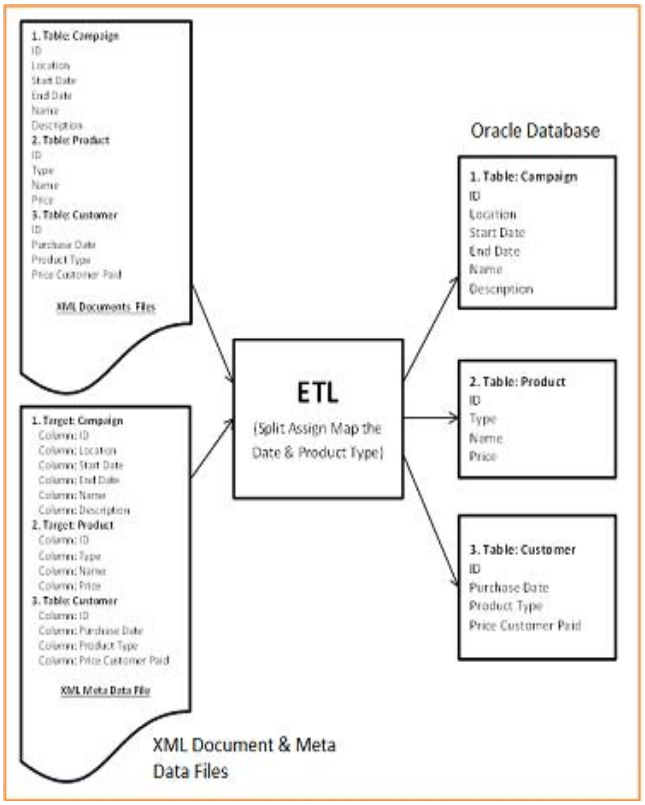
Table 3.

Customer Table Fields

Column Name	Data Type	Nullable	Default	Primary Key
PUR_ID	NUMBER(4,0)	No	-	1
PUR_DATE	DATE	Yes	-	-
PROD_TYPE	VARCHAR2(50)	Yes	-	-
PRICE_CUS_PAID	NUMBER(12,2)	Yes	-	-

B. Intermediate Process (Transformation – Split, assign, map the data)

Table 4.



C. Output Table (loading)

Table 5.

Product Table Records

PROD_ID	PROD_TYPE	PROD_NAME	PRICE
1	Stationary	Pencil	120
2	Stationary	Ink Pen	500
3	Stationary	A4 Paper	890
4	Stationary	Ball Pen	400
5	Stationary	Eraser	200
6	Fancy	Box	450
7	Fancy	Water Bottle	800
8	Fancy	School Bag	950

Table 6.

Customer Table Records

PUR_ID	PUR_DATE	PROD_TYPE	PRICE_CUS_PAID
1	03-NOV-12	Pencil	120
2	05-NOV-12	Ink Pen	500
3	08-NOV-12	A4 Paper	890
4	09-NOV-12	Ball Pen	400
5	04-NOV-12	Eraser	200
6	12-NOV-12	Box	450
7	15-NOV-12	Water Bottle	800
8	10-NOV-12	School Bag	950
9	20-NOV-12	Sharp	360
10	25-NOV-12	Note	712

Table 7.

Campaign Table Records

CAM_ID	LOCATION	START_DATE	END_DATE	CAM_NAME	DESCRIPTION
1	Madurai	01-NOV-12	30-NOV-12	Super	Super Scheme
2	Madurai	01-NOV-12	30-NOV-12	Super	Super SCheme
3	Madurai	01-NOV-12	30-NOV-12	Normal	Normal SCheme
4	Madurai	01-NOV-12	30-NOV-12	Normal	Normal SCheme
5	Chennai	01-NOV-12	30-NOV-12	Super	Super SCheme
6	Chennai	01-NOV-12	30-NOV-12	Super	Super SCheme
7	Chennai	01-NOV-12	30-NOV-12	Normal	Normal SCheme
8	Chennai	01-NOV-12	30-NOV-12	Normal	Normal SCheme
9	Trichy	01-NOV-12	30-NOV-12	Super	Super SCheme
10	Trichy	01-NOV-12	30-NOV-12	Normal	Normal SCheme
11	Salem	01-NOV-12	30-NOV-12	Super	Super SCheme
12	Salem	01-NOV-12	30-NOV-12	Normal	Normal SCheme

Result will be displayed based on the location.

XML data file :

<Record TableName=”Campaign”>
<Cam_ID>1</Cam_ID>
<Location>Madurai</Location>
<Start_Date>2012-11-01 0:00:00.0</Start_Date>
<End_Date>2012-11-30 00:00:00.0</End_Date>
<Cam_Name>Super</Cam_Name>
<Description>Super Scheme</Description>
</Record>

Sample XML Meta-Data

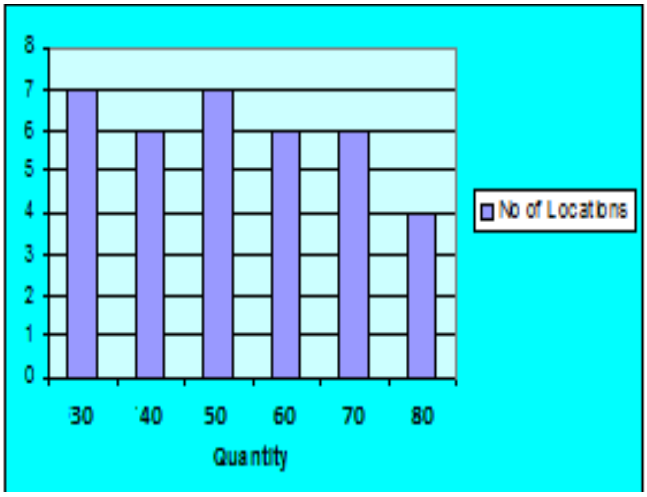
<target name="Campaigns" source="Sales"
driver="oracle.jdbc.driver.OracleDriver"
url="jdbc:oracle:thin:@localhost:1521:XE"
username="system" password="manager">
<column name="ID" type="Number (4)"
key="yes" source="ID"/>
<column name="Location" type="VarChar (25)"
source="Location"/>
<column name="Start_Date" type="Date"
source="Start_Date"/>
<column name="End_Date" type="Date"
source="End_Date"/>
<column name="Name" type="VarChar (20)"
source="Name"/>
<column name="Description" type="VarChar (50)"
source="Description"/>
</target>

This Hyper ETL is tested in some sales data and implemented by using Assignment problem. Here, the quantity of item and Locations are taken into account. The result of assignment problem shows that, product A is assigned to Location L3, B to L1, C to L4, D to L6, E to L2 and F to L5.

D. Histogram analysis

The result of assignment problem is given in the form of histogram analysis.

Quantity	No of Locations
10	7
20	6
30	7
40	6
50	6
60	4



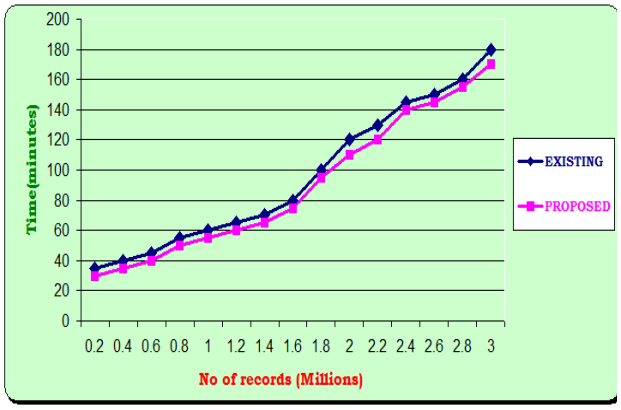


Figure: 3.Time Comparison for an Existing and Proposed ETL

VI. CONCLUSION

We have proposed the sophisticated structural design of Hyper ETL which accomplishes, the mapping multiple sources into multiple targets, merging the relevant field from three different tables and purging of all data and eliminating the duplicate fields from the table. This hyper ETL was demonstrated through sample Sales records and it is suggested that this ETL reduces the time, and improves the decision-making process and also automates the key activities of the process (extraction of metadata, transformation of information). Hence, we conclude with the expectation of the above mentioned delivers will be tackled in an enterprise and industry. The work presented in this paper is implemented by using the Assignment Problem. The researcher intends to extend our work by applying the other data mining methods.

REFERENCES

[1] D. Fasel and D. Zumstein, "Fuzzy data warehouse approach for web analytics,". In MD.

[2] Lytras, E. Damiani, J. M. Carroll, R. D. Tennyson, D. Avison, A. Naeve, A. Dale, P. Lefrere,F. Tan, J. Sipior, and G. Vossen, "Visioning and Engineering the Knowledge Society - A Web Science Perspective," volume 5736 of Lecture Notes in Computer Science, pages 276–285. Springer, 2009.

[3] Hariprasad T, "ETL testing Fundamentals," on March 29, 2012.

[4] Huamin Wang, "An ETL Services Framework Based on Metadata," 2nd International Workshop on Intelligent Systems and Applications, May 2010.

[5] W. H. Inmon, "Building the Data Warehouse," Wiley Publishing, Inc., 4 edition, 2005.

[6] Inmon, William, "Data Mart Does Not Equal Data Warehouse," DMReview.com, (2000-07-18).

[7] R. Kimball_and M. Ross, "The Data Warehouse Toolkit," Wiley Publishing, Inc., 2002.

[8] Li Jian, et al, "ETL tool research and implementation based on drilling data warehouse," Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Aug 2010.

[9] Lunan Li, "A framework study of ETL processes optimization based on metadata repository," International Conference on Computer Engineering and Technology, April 2010.

[10] Munoz L., et al, "Systematic review and comparison of modeling ETL processes in data warehouse," Iberian Conference on information Systems and Technologies, June 2010.

[11] A. Prema and A. Petha Lakshmi, "An approach to construct the Fuzzy Data Mart using Soft Computing," International journal on computer Application, Nov 2012.

[12] Simitsis, et al, "State-space optimization of ETL workflows," IEEE Transactions on Knowledge and Data Engineering, vol. 17, Issue 10, Oct 2005.

[13] "Business Objects Data Integrator overview," www.business objects.com.

[14] Sunmicrosystem.com, "Data Warehouse ETL Tools," .www.dwhetltool.blogspot.in