International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 2, Mar-Apr-2019, ISSN (Online): 2395-566X

An Advanced ETL Technique for Error Free Data in Data Warehousing Environment

Marzana Ifat Moly molyifat@gmail.com
Ovijit Roy ovijit786@gmail.com
Md. Alomgir Hossain alomgir.hossain@iubat.edu
Department of Computer Science & Engg.
International University of Business Agriculture and Technology
Dhaka, Bangladesh

Abstract- **A large store of data accumulated from a wide range of sources within a company and used to guide management decisions. Data Warehousing is one of the common words for last 10-20 years, whereas Big Data is a hot trend for last 5-10 years. Now we are trying to find out that how the ETL process work with the Data Warehouse and how to manage after transform the data into data warehouse. ETL is the stand for extract, transform and load. This is a very new concept for Data warehouse. Many of know about the ETL but it's just a theoretical concept. So we try to find out how ETL works with Data warehouse. And we try to build a prototype of ETL and that's ETL design include in this paper.**

Keywords **– Data Warehouse, ELT Process, ETL Process, ELT vs ETL, ETL Algorithm, ETL Queries**

I. INTRODUCTION

"Data warehousing" a hot topic in 1990′s and some part of 2000′s came into existence as the business requirements changed from transaction processing to analyzing the data. The data had to be cleaned, transformed and loaded (ETL) onto the data warehouse which was then used for generating reports. As reported on different websites, "Data warehousing" is always considered to be an architecture. Extract, Transform and Load, abbreviated as ETL is the process of integrating data from different source systems, applying transformations as per the business requirements and then loading it into a place which is a central repository for all the business data that is capable to do reporting.\

II. LITERATURE REVIEW

A data warehouse is populated using Extract, Transform and Load (ETL) process that (1) extracts data from various sources, (2) integrates, clean, and transform it into a common form and (3) load it into a data warehouse [1]. To provide the effective policy making for manager, and to give the reader a better service, the paper proposed a new decision support system based on the data warehouse technology that is ETL and used SQL Server for the solution of ETL [2]. Existing research has conducted reviews of warehousing research and has suggested ETL for the future work[3].

III. METHODOLOGY

According to the analysis, model design and algorithm and some other different criteria we think that our analysis goes to qualitative method both. For making this paper, we collect data from secondary source and we also see some research paper according to our topic which are already published. There are few steps that we follow to make this paper:

Step 1: Collect Data from Internet
Step 2: Collect Data from Research Paper
Step 3: Know about Data Warehouse and ETL
Step 4: Finding the existing model for Data Warehouse
Step 5: Find the design for ETL
Step 6: Find the simple and common algorithm for ETL in Data Warehouse.

IV. DATA WAREHOUSE

A data warehousing is a technique for collecting and managing data from varied source to provide meaningful business insight. It is a blend of technology and components which allows the strategic use of data [4]. Data warehouse system is also known by the following name:

- ETL Strategic Process
- ELT Strategic Process
- Data Warehousing Process
- Batch Process
- Stream Process
- Data Warehouse
- Decision Support System (DSS)
- Executive Information System
- Management Information System

Fig 1 Data Warehouse and Its System

International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 2, Mar-Apr-2019, ISSN (Online): 2395-566X

V. LIMITATION OF DATA WAREHOUSE

Data warehouses aren't regular databases as they are involved in the consolidation of data of several business systems which can be located at any physical location into one data mart. With OLAP data analysis tools, you can analyze data and use it for taking strategic decisions and for prediction of trends [5]. Even with the innumerable benefits, implementing a data warehouse model for your business might have some drawbacks.

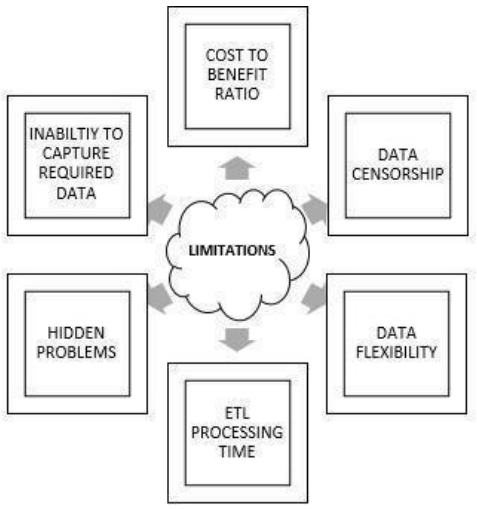

Fig. 2 Data Warehouse Limitations

VI. ETL IN DATA WAREHOUSE

The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading. Note that ETL refers to a broad process, and not three well-defined steps [6]. The acronym ETL is perhaps too simplistic, because it omits the transportation phase and implies that each of the other phases of the process is distinct. Nevertheless, the entire process is known asETL.

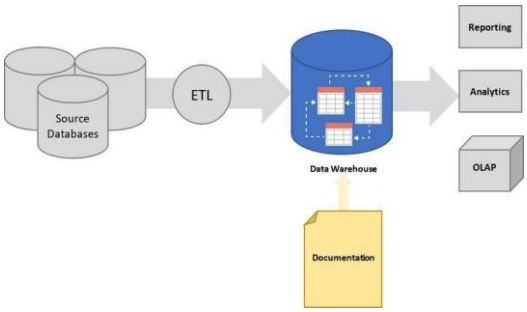

Fig. 3 ETL in Data Warehouse.

VII. COMPARISON BETWEEN ELT AND ETL PROCESS

ELT: ELT is appropriate for databases like Teradata. I heard that the Tera data engine is extremely fast and efficient at ELT (Extract, Load, Transform) processing to aggregate and resave data

Table 1: ELT vs ETL

Please Put Here table proper

VIII. EXTRACT-TRANSFORM-LOAD PROCESS

1. Building an ETL Pipeline with Batch Processing: Building ETL with batch processing, following ETL best practices, involves [7]:

1. Reference data - create a set of data that defines the set of permissible values your data may contain. For example, in a country data field, you can define the list of country codes allowed.

2. Extract from data sources - the basis for the success of subsequent ETL steps is to extract data correctly. Most ETL systems combine data from multiple source systems, each with its owndata.

3. Data validation - an automated process confirms whether data pulled from sources has the expected values - for example, in a database of financial transactions from the past year, a date field should contain valid dates within the past 12 months.

4. Transform data - removing extraneous or erroneous data (cleaning), applying business rules, checking data integrity (ensuring that the data was not corrupted in source, or corrupted by ETL, and that no data was dropped in previous stages), and creating aggregates asnecessary.

5. Stage - you will not typically load transformed data directly into the target data warehouse. Data should first enter a staging database, making it easier to roll back if something goe swrong.

6. Publish to data warehouse - loading the data to the target tables. Some data warehouses overwrite existing information every time the ETL pipeline loads a new batch - this might happen daily, weekly or monthly.

Fig. 4 ETL in Batch Processing.

7. Building an ETL Pipeline with Stream Processing- Modern data processes often include real time data - for example, web analytics data from a large ecommerce website. In these use cases, you cannot extract and transform data in large batches [8]: the need arises to perform ETL on data streams. This means that as client applications write data to the data source, data should be treated, transformed and saved immediately to the target data store.

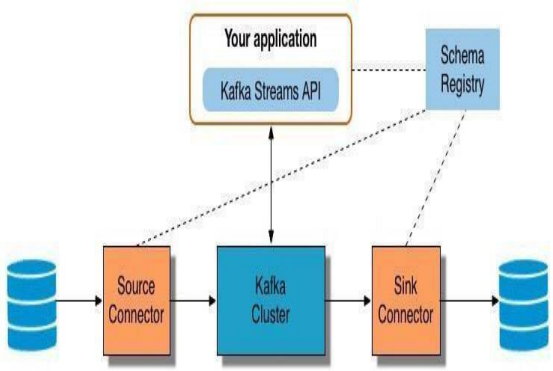

Fig. 5 ETL in Stream Processing

IX. ETL AND AUTOMATED CLOUD-BASE

1. Building a Pipeline without ETL using Automated Cloud-Based Data Warehouse- New cloud-based data warehouse technology makes it possible to achieve the original goal of ETL without building an ETL system at all [9]. Building a data pipeline without ETL in Panoply involves:

1. Select data sources and import data - select your data sources from a list, enter credentials and define destination tables, click Collect and Panoply automatically pulls the data for you. Panoply automatically takes care of schemas, data preparation, data cleaning, and more.
2. Run transformation queries - select a table and run a SQL query against the raw data. You can save the query as a transformation, or export the resulting table into your own system. Panoply supports both simple views and materialized transformation views [10]. You can run several transformations, until you achieve a data format that enables analysis. Panoply's Query Log allows you to easily roll back to previous processing steps. You shouldn't be concerned about "ruining" the data - Panoply lets you perform any transformation, but keeps your raw data intact.
3. Data analysis with BI tools - you can now connect any BI tool such as Tableau or Looker to Panoply and explore the transformed data.

X. ALGORITHM FOR ETL

T: = 0; Compute initial data D0;
While
Stopping condition not fulfilled **DO**
Begin
select individuals from source;
Extract off springs by crossing individuals;
transform mutate some common form;
compute new generation and load
End

XI. ETL SCRATCH SQL

Query Source Query

```
SELECT Demographics.ParticipantId,
Demographics.StartDate,
Demographics.Gender,
Demographics.PrimaryLanguage,
Demographics.Country,
Demographics.Cohort,
Demographics.TreatmentGroup
FROM Demographics
WHERE Demographics. Gender = 'm'
AND  Demographics. Treatment Group = 'Natural
Controller'
```

XII. ETL PROCESS QUERY

```
<etlxmlns="http://labkey.org/etl/x ml">
    <name>Demographics >>> Patients
(Males)</name>
    <description>Update data for
study on male
patients.</description>
    <transforms>
        <transform id="males"> <source
schemaName="study"
queryName="MaleNC"/>
            <destination
schemaName="study"
queryName="Patients"
targetOption="merge"/>
        </transform>
    </transforms>
    <schedule>
<poll interval="1h"/> </schedule></etl>
```

XIII. OUR MODEL OF ETL

As the data warehouse is a living IT system, sources and targets might change. Those changes must be maintained and tracked through the lifespan of the system without overwriting or deleting the old ETL process flow information. To build and keep a level of trust about the information in the warehouse, the process flow of each individual record in the warehouse can be reconstructed at any point in time in the future in an ideal case.

International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 2, Mar-Apr-2019, ISSN (Online): 2395-566X

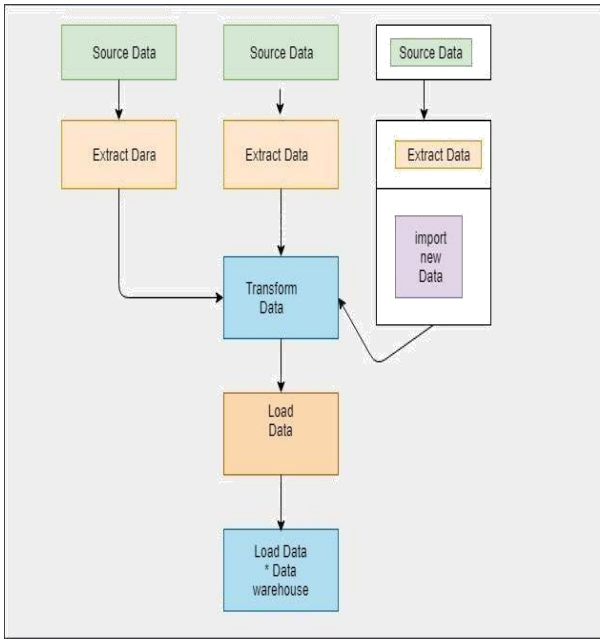

Fig. 6 ETL Process Model

1. Source System- A database, application, file or other storage facility from warehouse is derived.
2. Mapping- The definition of the relationship and data flow between source and target object.
3. Staging Area- A place where data is processed before entering the warehouse.
4. Cleansing- The process of resolving inconsistencies and fixing the anomalies.
5. Transformation- The process of manipulating data
6. Transportation- The process of moving copied or transformed data from source to destination.
7. Target System- A database, application, file or other storage facility to which the transformed data is loaded.

XIV. APPARATUS REQUIRED

1. Hardware- Core 2 Duo/ Core i3/ Core i5/ Core i7
2. OS-Windows 7/ 8/ 8.1/ 10
3. Front-End- Microsoft Visual Studio (C#)
4. Back-End- Microsoft SQL Server

XV. FINDINGS

We created a ETL process by using Microsoft SQL server and Microsoft visual studio(c#). We simulate this project to find out how ETL can load data from different source. And extract the data from it's original source, transform it to a common form for all types of data and finally how it load into final destination that is called main data warehouse. After simulating this project when we run and debug it at that time we see each and individual terms are working and after finish their work they give us green signal. If any term are not able to work then it will gives red signal.

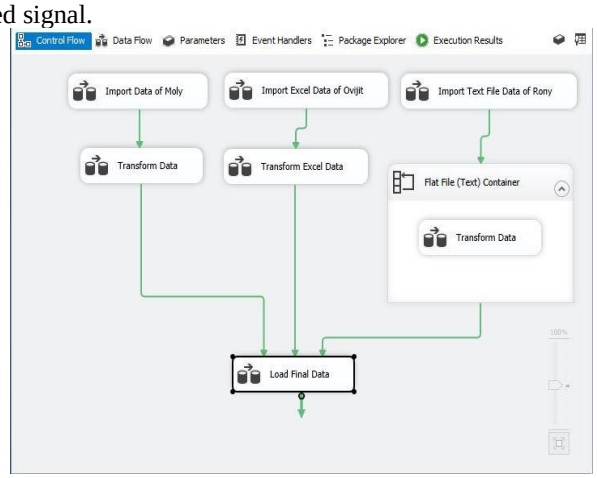

Fig. 7 Main Interface of System.

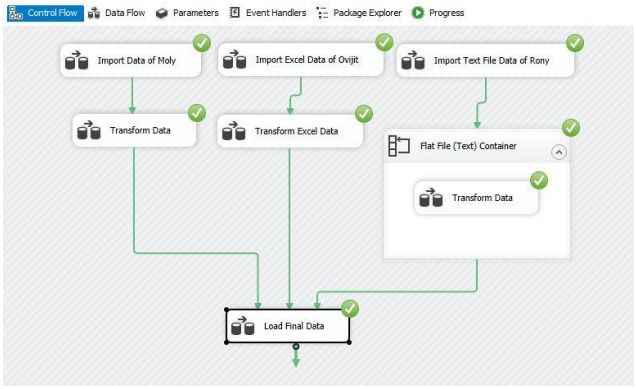

Fig. 8 Output of the System.

XV. CONCLUSION

We design our ETL process and we create this ETL process by using Microsoft SQL Server and Microsoft Visual Studio(C#). After creating this ETL we saw that this ETL can find out the source of the data and after detecting the data it extract from it's original source and transform it to a general form for any data and after transforming finally it loaded into main Data warehouse.

REFERENCES

1. HAJAR Homayouni, Testing Extract-Transform-Load Process in Data Warehouse Systems 2018 IEEE.
2. Fushan Wang, Application Research of Data Warehousing Technology in Library Decision Support System. 17-28-2000. Philip Woodall, Gokeen Yilmaz,

3. Vaggelis Giannikas, New Directions for Warehousing Data Management Research: Extensions to an Existing Review,2018.

4. A. V. Aho and J. D. Ullman, Foundations of Computer Science, Ced. W. H. Freeman, 1994.
5. A. Davoudian, L. Chen, and M. Liu, "A Survey on No SQL Stores," ACM Comput. Surv.,vol. 51, no. 2, pp. 1–43, 2018.
6. S. Ren, T. Wang, and X. Lu, "Dimensional Modeling of Medical Data Warehouse based on Ontology," in 3rd IEEE International Conference on Big Data Analysis, Shanghai, China, 2018, pp. 144– 149.
7. IJCSI International journal of Computer Science issues, Vol7, Issuse3, No 2, May2010
8. H. Davarzani and A. Norrman, "Toward a relevant agenda for warehousing research: literature review and practitiones' input", vol.8 no. 1,2015.
9. Kimball, R. The data Warehouse Toolkit. John Wiley,1996.
10. P. N. S.-B. Furtado, Evolving Application Domains of Data Warehousing and Mining: Trends and Solutions, 1st ed. Hershey, PA: Information Science Reference - Imprint of: IGI Publishing,2009.
11. Arup Kumar Bhatta charjee, AtanuMa k, Arnab Dey and Sananda Bandyopadhyay,
12. "Data Cleaning in Text File", Dept. of MCA, RCC Institute of Information Technology, India. F. Hinshaw, "Data Warehouse Appliances:
13. Driving the Business Intelli-gence Revolution," DM Review Magazine, pp. 30–34,2004.
14. Joyti Sheoran, Issues of Data Quality in Data Warehouses, (IJCA) (0975-8887). Dr. Mortadha M. Hamad and AlaaAbdulkhar Jihad, "An Enhanced Technique to Clean Data in the Data Warehouse". Computer Science Department. University of Anbar, Ramadi,Iraq.

Author Profile

Ovijit Roy

This is Ovijit Roy. I passed my SSC from Tajgaon Gov,T High School at 2010 and passed my Diploma in Engineering from Dhaka Polytechnic Institute from 2015. Now I completed my Bachelor of Science in Computer Science and Engineering from IUBAT. I live in Dhaka. Besides my study I like to write some article and work on thesis paper. I love to coding on Java and my hobby is photography. I am interested to listening song, watching movie and playing games in android.

Marzana Ifat Moly

This is Marzana Ifat Moly.I have passed my SSC from M.E.H Arif college at 2011 and HSC from Milestone College at 2013. Now I am doing my Bachelor of Science and Engineering degree from IUBAT. I am from Gazipur. I like to write and work on thesis paper. I am interesting on coding. My hobby is dancing, travelling.

Md. Alomgir Hossain

I am Alomgir Hossain. I passed my Secondary School Certificate at 2002 and passed my Diploma in Engineering at 2006. I passed my Bachelor of Science in Computer Science and Engineering from IUBAT at 2011. I passed my Master of Science from Jahangirnagar University at 2013. Now I am doing my PhD. The experience of my teaching profession is more than 7 years. I admitted myself as Lecturer in IUBAT at 2014. I am interested to do research in relevant area. My hobby is travelling.