www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 10 October 2017, Page No. 22545-22550
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i10.03

ETL Data conversion: Extraction, transformation and loading Data conversion

Gajare Harikishan Prakash[1], Prof. S. P. Rangdale[2]

[1] M.E. Information Technology,
Siddhant college of Engg.Sudumbre, Pune, India
[2] M.E. Information Technology,
Siddhant college of Engg.Sudumbre, Pune, India

Abstract:
The essential elements of decision support system are data warehousing and on-line analytical processing (OLAP), which are focus of the database industry. Recently, many commercial products and services are database management system vendors. Decision support places some rather different requirements on database technology compared to traditional on-line transaction processing applications. The researchers and developers had proposed various trials on new requirements of data warehousing and OLAP technologies to put a standard conceptual design of ETL processes in data warehouse. To avoid limitations of the previous trials, the author proposed: 1. The model for conceptual design of ETL processes and entity mapping diagram (EMD), which is built upon the enhancement of the models to support some missing mapping features. 2. The conceptual model in a prototype called EMD Builder and uses it in an illustration scenario. This is an ETL project,, which focuses on data transformation and loading in Oracle Data warehouse. Extraction–transformation–loading (ETL) tools are pieces of software responsible for the extraction of data from several sources, its cleansing, customization, reformatting, integration, and insertion into a data warehouse. Structure the ETL development be potentially one of the major tasks of construction a warehouse; it is difficult, time consuming, and consume most of data warehouse project's implementation efforts, costs, and resources. In spite of the importance of ETL processes, little research has been done in this area due to its complexity. There is a apparent lack of a ordinary sculpt that can be used to be a symbol of the ETL case scenarios. In this paper we will try to navigate through the efforts done to conceptualize the ETL processes. These projects try to represent the main mapping activities at the conceptual level. Due to the differences and variation stuck between the planned solutions for the intangible design of ETL processes and due to their limitations, this paper also will propose a model for conceptual design of ETL processes. The proposed model is built upon the enhancement of the models in the previous models to support some missing mapping features.
Keywords: Extract, transform, load, scripting, query cache, OLTP, OLAP, Data warehousing.

1. Introduction

Throughout the ETL process, data is removed from an OLTP database, altered to equivalent the data warehouse schema, and loaded into the data warehouse database. Many data warehouses also incorporate data from non-OLTP systems, such as text files, legacy systems, and spreadsheets; such data also requires extraction, transformation, and loading.

In its easy form, ETL is the course of action of repetition data from one database to a further. This simplicity is rarely, if ever, found in data warehouse implementations; in reality, ETL is often a complex combination of process and technology that consumes a significant portion of the large data warehouse enlargement efforts and need the skills of trade analysts, database(DB) designers, and application developers.

When defining ETL for a data warehouse, it is important to think of ETL as a process, not a physical implementation. ETL systems differ from data warehouse to another data warehouse and even sandwiched between department information marts within a data warehouse. A monolithic application, regardless of whether it is implemented in Transact-SQL or a traditional programming language, does

not provide the flexibility for change necessary in ETL systems. A combination of tools and technologies be supposed to be used to grow applications with the intention of each present a specific ETL task.

The ETL process is not a one-time event; new data is added to a data warehouse periodically. Classic periodicity possibly will be monthly, weekly, daily, or even hourly, on basis of the reason of the data warehouse and the type of business it serves. Because ETL is an integral, ongoing, and recurring part of a data warehouse, ETL processes must be automated and operational procedures documented. ETL too modify and mix as the data warehouse mix, so ETL processes have to be designed for ease of modification. A solid, well-designed, and documented ETL system is necessary for the success of a data warehouse project.

Data warehouses mix to get better their service to the business in addition to to adapt to modifications in business processes and requirements. Business rules change as the business reacts to market influences—the data warehouse must respond in order to maintain its value as a tool for decision makers. The ETL execution must become accustomed as the data warehouse mix. Microsoft® SQL Server™ 2000 provides significant enhancements to existing performance and capabilities, and introduces new features that make the development, deployment, and maintenance of ETL processes easier and simpler, and its performance faster.

Related Work

Despite of how they be implemented, every ETL systems encompass a general purpose: they move information from one database (DB) to another. Usually, ETL systems move information from OLTP systems in the direction of a data warehouse, but they can also exist used to travel data from one data warehouse to a new. An ETL system contains four discrete functional fundamentals:

Extraction: The ETL removal constituent is in charge for extracting data commencing the supply system. During removal, data may be removed from the resource system or a copy made duplicate and the original data retain in the source system. It is ordinary to move chronological data that accumulate in an prepared OLTP system to a data warehouse to maintain OLTP performance and competence. Heritage systems possibly will need too much exertion to employ such offload processes, so bequest data is often copied into the data warehouse, leaving the original data in place. Extracted records is encumbered interested in the data warehouse staging part (a relational database typically break up from the data warehouse database), for manipulation by the remaining ETL processes. Data extraction is usually performed surrounded by the foundation system itself, principally if it is a relational database to which extraction procedures can easily be added. It is moreover possible intended for the taking out logic to subsist in the data warehouse staging area and doubt the source system for data using ODBC, OLE DB, or other APIs. For heritage systems, the largest part ordinary method of data removal is for the heritage system to bring into being text files, although many newer systems offer direct query APIs or accommodate access through ODBC or OLE DB. Data taking out processes is be able to be implemented with the help of Transact-SQL stored procedures. Data alteration Services (DTS) tasks, or tradition applications urbanized in encoding or scripting languages.

Transformation: The ETL transformation element is responsible for data validation, data accuracy, data type conversion, and business rule application. It is the most complicated of the ETL elements. It may appear to be more efficient to perform some transformations as the data is being extracted (inline transformation); however, an ETL system that uses inline transformations during extraction is less robust and flexible than one that confines transformations to the transformation element. Transformations performed in the OLTP system impose a performance burden on the OLTP database. They also come apart the alteration logic involving two ETL elements and put in maintenance density when the ETL logic changes.

Loading: The ETL loading element is responsible for loading transformed data into the data warehouse database. Data warehouses are more often than not updated every so often somewhat than endlessly, and large figures of records are often loaded to multiple tables in a single data load. The data warehouse is often taken offline during update operations so that data can be loaded faster and SQL Server 2000 Analysis Services can update OLAP cubes to incorporate the new data. BULK INSERT, **bcp**, and the Bulk Copy API are the best tools for data loading operations. The design of the loading element should focus on efficiency and performance

to minimize the data warehouse offline time. For more information and details about performance tuning, see Chapter 20, "RDBMS Performance Tuning Guide for Data Warehousing."

Meta Data: The ETL meta data functional element is responsible for maintaining information (meta data) about the movement and transformation of data, and the operation of the data warehouse. It also documents the data mappings used during the transformations. Meta data logging provides possibilities for automated administration, trend prediction, and code reuse.

2. Title, Authors, Body Paragraphs, Sections Headings and References

3.1 Piyaporn Samsuwan, "Generation of Data Warehouse Design Test Cases", 2015 IEEE [1]

There are several data warehouse lifecycle for Relational database. However, most design modeling agrees on Conceptual schema, Logical schema, and Physical schema. This paper presents an approach to generating test cases for data warehouse design tests, including the level of conceptual, logical, and physical. The test cases contain SQL statements for verifying some certain predefined aspects. The minimum edit distance component is also available for correcting the garbled terms found in the SQL. The implemented system would enhance the quality of target data as well as lessen the rework.

In particular, the presented approach would reduce the number of defects injected in detailed designed prior to entering the ETL phase. This leads to less rework for preprocessing target data at the ETL phase. several data warehouse tests are performed, such as Conceptual design test, Workload refinement test, Logical design test, and Front end test (UAT). If defects are found during any testing, the data construction process will be back to the ETL phase for correction, resulting in the waste of resources.

3.2 Erhard Rahm, Hong Hai Do, "Data Cleaning: Problem And Current Approaches", Microsoft Research, Redmond, WA. [2]

In this paper the author proposed data cleansing or scrubbing, approaches to improve the quality of data by detecting and removing errors and inconsistencies from data. Some data quality problems are misspellings in data, missing information or invalid unstructured data. The author suggested data cleaning to avoid redundant data received from multiple data sources The elimination of duplicate information become necessary provide access to accurate and consistent data. The author suggested process i.e data analysis and data definition workflow and mapping rules to clean data. During data analysis phase various kinds of errors, inconsistencies are to be removed. Also suggested manual inspection of the data or data samples, analysis programs, data properties and detect data quality problems. By using definition of transformation workflow and mapping rules with respective to number of data sources and heterogeneity of the data, a large number of data transformation and cleaning steps can be executed. The schema translation is used sometime to map sources to a common data model. The data cleaning steps are carried to correct single-source instance problems and to prepare the data for integration. In next step duplication with schema/data integration is avoided and cleaning multi-source instance problems is done.

3.3 Er. Sonal Sharma, "Modeling ETL Process in Data warehouse: An Exploratory Study", Master Thesis Software Engineering Thesis no: MSE-2011-65 09 2011 [3]

The data warehouse facilitates knowledge workers in decision making process. A good DW design can actually reduce the report processing time but, it requires substantial efforts in ETL design and implementation. In this paper, the authors have focused on the working of Extraction, Transformation and Loading. The focus has also been laid on the data quality problem which in result leads to falsification of analysis based on that data. The authors have also analysed and compared various ETL modeling processes.

3.4 Prayag Tiwari, "Improvement of ETL through Integration of Query Cache and Scripting method", International Conference on Data Science and Engineering, 2016 IEEE [4]

ETL tools fetch data from data source and transform data into new format according to need and then load into Data warehouse for further processing of Data. Slowing ETL makes big trouble in business so ETL must be fast that could give us better performance in business. We toss light on the requirement for utilization of scripting advances to robotize ETL instruments preparing end to end process which diminish manual cerebral pain of

running ETL process taking care of furthermore prompt improvement of ETL devices in future. Constructing the ETL procedure is conceivably one of the greatest errands of constructing a warehouse. Query cache will store all record of executed query. Query cache will keep record of recently executed queries. The major objective of the query cache is to diminish the reaction time of query.

3. Proposed System Model

A. The system will execute using below procedure:

A data warehouse is a centralized repository that stores data from multiple information sources and transforms them into a common, multidimensional data model for efficient querying and analysis.

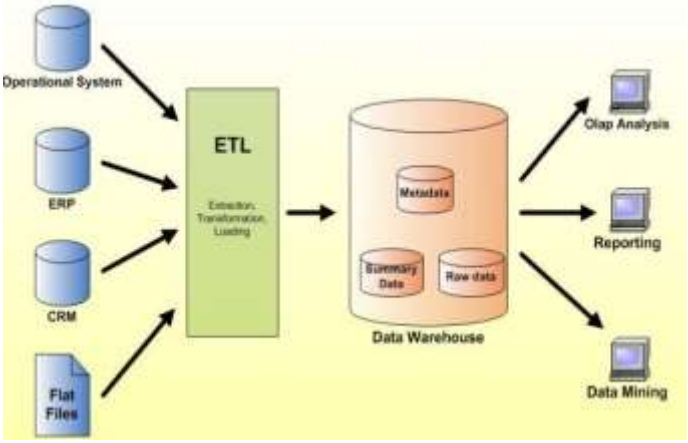

Fig. 1: proposed system architecture

In Figure 1-2, the metadata and raw data of a traditional OLTP system is present, as is an additional type of data, summary data. Histories are extremely valuable in data warehouses since they pre-compute lengthy operations in move ahead. For example, a typical data warehouse query is to retrieve something such as August sales. A summary in an Oracle database is called a materialized view.
This figure illustrates three things:

1. Data Sources (operational systems and flat files)
2. Warehouse (metadata, summary data, and raw data)
3. Users (analysis, reporting, and mining)

B. Keyword Points

Scripting method

Great working and a very much displayed recorded information distribution center requires organizing different operations crosswise over different applications, databases, and frameworks. In a huge organization, this may be upwards of 1000 discrete operations that should be performed in the right grouping, at the right time, and under the right conditions. Data warehouse operations manage greatly huge volumes of data. Missing a solitary stride in the process or executing a certain progression at the wrong time, can bring about a lot of squandered preparing time, or in the most dire outcome imaginable, terrible information. Proposed approach comprises of data source layer that can have diverse homogenous or heterogeneous frameworks on various hubs might be operational or level documents.

Query Cache method

For the quick get inside database we utilize the query cache. Query cache will store all record of executed query. Query cache will keep record of recently executed queries. The major objective of the query cache is to diminish the reaction time of query.

OLTP

OLTP (online transaction processing) is a class of software programs capable of supporting transaction-oriented applications on the Internet. Characteristically, OLTP systems be used for arrange entry, monetary transactions, retail sales and customer relationship management (CRM). Such systems have a large number of users who conduct short transactions. Database queries are usually simple, require sub-second response times and return relatively few records. An important attribute of an OLTP system is its ability to maintain concurrency. To avoid single points of failure, OLTP systems are often decentralized.

OLAP

OLAP (On-line Analytical Processing) is characterized by relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems a response time is an effectiveness calculate. OLAP applications are extensively used through Data Mining techniques. In OLAP DB (database) there is historical, aggregated data, stored in multi-dimensional schemas (usually star schema).[9]

Data warehousing

In one place you can find descriptions of ETL and BI tools, the most popular Data Warehouse architectures, solutions, engines and a lot of others. On our pages you be able to discover general idea

and overview of very useful information concerning entire Business intellect market. In the News section you can gather a piece of information about Data Warehouse and BI events and seminars. [2] If you wish for to expand optimum charge from information data resources in your association Datawarehouse4u.info is the key step that you should take.

C. Mathematical Module:

Transformation algorithm-

Step 1: Selecting only certain column to Load
Step 2:Translating coded values
Step 3:Encoding Free from value
Step 4: Deriving new calculating value
Step 5: Sorting
Step 6 : Joining data from multiple sources
Step 7:Aggregation,Generting surragoate key
Step 8:transposing and Spit column into multiple column
Step 9:Look up and validate the data

scripting algorithm

a) We need to indicate the different variable and global parameters, source, target path and other environment data, as required in design records as underneath.

Source_path=/app/source/…….
Target_path=/app/target/……..
Script path=/app/source/scripts
Log_path=/app/logs/err.log
Db_name=smg
Db_password=******
Db_servername=xyz1883c

b) Proceed source file format, source path and target file format, target path and other fundamental data to script by bringing config records characterized in step1 furthermore check information exists in sources file. We have separate script code for the normal function to handle certain job like logingerrors(), runtime(), runstatus(), dbcalling().

c) Execute the fundamental script code that summons ETL tool and pass all the essential data, output, config framework to tools as required relying upon tool. For our situation we have passed maps transformation record and configured document through command.

d) On the off chance that different jobs must be handled then loop all jobs them putting looping controls

4. Conclusion

ETL processes are very important research problem of data warehousing. This paper presents an approach to generating test cases for data warehouse design tests, including Conceptual test, Logical test, and Physical test. The structure and details of design document are predefined to adequately create the test cases for verifying the target data. The conceptual model EMD is a simplified model for representing ETL processes in data warehousing projects. The position of enterprises becomes growing real-time such as E-commerce sites, real-time BI will be increasing important to such companies. The ETL system efficiently extracts data from its sources, transforms and sometimes aggregates data to match the target data warehouse schema, and loads the transformed data into the data warehouse database. A well-designed ETL system supports automated operation that informs operators of errors with the appropriate level of warning. The results of our experiments show that the proposed ETL system has a high performance, especially for the massive data. We classified the data quality problems in data sources differentiating between single- and multi-source and between schema- and instance-level problems. We outlined steps for data transformation and data cleaning and provided an integrated way to cover schema- and instance-related data transformations. The ETL system is a primary source of meta data that can be used to track information about the operation and performance of the data warehouse as well as the ETL processes.

5. Future Work

The future scope of this system is to implement fully real time data streaming with ETL with big data applications. Data warehousing lets people look at corporate data as never before. Patterns, trends, seeing the forest and the trees -- all became a possibility with data warehousing. Commerce populace was capable to seem at their enterprise as no one had yet been capable to before and entire industries and multi-billion dollar companies grew out of this love of knowing more.

References

1. Piyaporn Samsuwan, “Generation of Data Warehouse Design Test Cases”, 2015 IEEE

2. Erhard Rahm, Hong Hai Do, "Data Cleaning: Problem And Current Approaches", Microsoft Research, Redmond, WA.
3. Er. Sonal Sharma, "Modeling ETL Process in Data warehouse: An Exploratory Study", Master Thesis Software Engineering Thesis no: MSE-2011-65 09 2011
4. Prayag Tiwari, "Improvement of ETL through Integration of Query Cache and Scripting method", International Conference on Data Science and Engineering, 2016 IEEE
5. V. Rainardi, "Data warehouse architecture," Building a Data Warehouse: With Examples in SQL Server, 2008, pp. 29-47.
6. M. Golfarelli and S. Rizzi, "A comprehensive approach to data warehouse testing," Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP. ACM, 2009.
7. M. Serrano, C. Calero, J. Trujillo, S. Luján-Mora, and M. Piattini,"Empirical validation of metrics for conceptual models of data warehouses," Advanced Information Systems Engineering. Springer Berlin Heidelberg, 2004.
8. C. Dell'Aquila, F. Di Tria, E. Lefons, and F. Tangorra, "Logic programming for data warehouse conceptual schema validation,"Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg, 2010, pp. 1-12.
9. S. Rizzi, A. Abelló, J. Lechtenbörger, and J. Trujillo, "Research in data warehouse modeling and design: dead or alive?," Proceedings of the 9th ACM international workshop on Data warehousing and OLAP. ACM, 2006.
10. "Data modeling - conceptual, logical and physical data models" ,http://www.1keydata.com."Minimun Edit Distance", https://web.stanford.edu.
11. K. U. Schulz and S. Mihov, "Fast string correction with Levenshtein automata," International Journal on Document Analysis and Recognition 5.1, 2002, pp. 67-85.