The Management of Conformed ETL Architecture

Ashima Jain
I.T Department, Northern India Engineering College, GGSIPU, New Delhi,India

Sonam Garg
I.T Department, Northern India Engineering College, GGSIPU, New Delhi, India

Neha Sharma
Assistant Professor
I.T Department, Northern India Engineering College, GGSIPU, NewDelhi, India

ABSTRACT

This paper deals with the core research on the architecture of ETL process which is applied on BI environment along with the advent of metadata at each corresponding layer that can be applicable to all the scenarios of BI. The management of extraction process has been done using several operators which help in reducing its complexity. New operators have been developed to easily understand each and every layer of extraction process. ETL stands for extraction, transformation and loading, and it plays a vital role in the area of business intelligence. Extraction is the process of extracting heterogeneous data from disparate source system for further analysis in a data warehouse environment. Transformation is a process of storing data in a correct, unambiguous and consistent format which is compatible to the format of existing data warehouse. Loading is a process that loads data into the end target which may be simply a data warehouse or a data mart.

Keywords

Extraction, Transformation, Loading, Business Intelligence

1. INTRODUCTION

The process of plucking out data from disparate source systems and invoking it into the data warehouse or data marts is commonly called ETL, which implies extraction, transformation, and loading. ETL is responsible for procuring data from source systems and loading into the data warehouse in a very fast and reliable way. The initial part of an ETL process involves educing the data from the various source systems is converted into one polarized data warehouse format which is ready for transformation processing. An innate part of the extraction involves data validation to confirm if the data plucked from the sources have the consistent values in a given domain. The data transfiguration stage applies a set of protocols and some methods to the squeezed data from the source to derive the data for loading into the destination store room of data. In some cases, data do not require any transformation at all; which is nothing but direct move of data in our technical aspects. An important function of data transformation is purification of data that aims to pass only compatible data to the target. When different systems interact with each other; based on how these systems store data, there is a challenge in interacting with each other. Certain character set that may be available in one system may not be available in other. These cases must be handled properly leading to a number of data quality related issues. Now, the turn of loading comes which is the back end of the ETL process. This phase cram the data into the destination which is nothing but the data warehouse only. According to the requirements of various chamber, this process assorted widely. Some data warehouses may copy existing pool of information with conglomerativeinformation; updating excerpt data is frequently done on a daily, hourly or weekly basis.

Business intelligence (BI) has gained wide recognition in the last years. It also got high business impact and is seen as a key enabler for increasing value and performance" [1]. Business Intelligence (BI) is considered to have a high impact on businesses. Research activity has risen in the last years. An important part of BI systems is a well performing implementation of the Extract, Transform, and Load (ETL) process [2]. Business Intelligence is a technique which includes various services i.e. IT innovated functionality often offered by a provider, because of the characteristic of service the operation and performance management of workflow meet some new difficulties [3].By the time, in this world of business new techniques are growing day by day so there was an urgent need to manage our ETL process. The high level management of ETL processes as a generic approach to enable their flexible re-use, optimization, and rapid development. To this, there was an introduction of a set of basic operators on ETL processes, such as merge or invert, and motivate their used in some scenarios [4]. Usage of Extract-Transform-Load (ETL) tools to provide business object transformation and to solve some of the issues by using pluggable components and introduce customized operators for ETL tools. [5].

After studying these research papers we came to our conclusion that there are certain issues regarding these papers which are described as below:

- Management of metadata was difficult which leads in difficulty for the organizations to track and monitor data flows within Business Intelligence environment and also the evaluation was not done accurately and the performance was analyzed using multi-indicators which make it complex as we need different indicators at each step. This problem was resolved by adding metadata layer corresponding to each layer in our ETL architecture. Every metadata layer has an inbuilt concept of doubly linked list so that misinterpretation of data is avoided at each step during tracking of data flow.
- Another issue was of complexity, we were unable to alter the data corresponding to a particular layer. Reverting back to its previous data was unable. This problem was resolved by introducing a new operator at extraction process i.e. Undo operator using which we can go back to its previous state.

2. LITERATURE REVIEW

In the era of the research in ETL an attempt has been to make a next move towards developing a standardized architecture which was formerly proposed by prominent personalities. Some of their proficient work has now helped us as a backbone in our research paper. One important component of the architecture is the Extract, Transform and Load (ETL) process. It describes the gathering of data from various sources (extract), its modification to match a desired state (transformation) and its import into a database or data warehouse (load). A data warehouse is predominantly used to store detailed summary data and metadata. Metadata include information on data themselves.

Anand proposed facilitate a process of extracting, transforming and loading data through presenting sources of data in the layout of data warehouses. [2]

Watson and Wixom proposed that Business intelligence (BI) has gained wide recognition in the last years. It also got high business impact and is seen as a key enabler for increasing value and performance". [1].

The service has become the most important element of Business Intelligence. Morris {et al Liao, Padmanabhan, Srinivasan, Lau, Shan, and Wisnesky} proposed that Business objects represent the key concepts that a business needs to operate such as people, services, and products [5].

Liu and Fan focused on Service-oriented workflow shows many new characteristics in aspects of execution mechanism and performance evaluation. [3]

As to implement Business Intelligence in service oriented environment we need a systematic framework for management of ETL processes.

Albrecht and Naumann presented the main idea of an ETL management platform is to reduce the amount of programming needed to develop or maintain ETL processes. To establish ETL management we introduce a set of basic operators such as search, create, modify etc. [4].

To establish a statistical model we need a prominent graph to uphold its characteristics. Vassiliadis {et al Simitsis, Skiadopoulos} proposed that simple graph transformations are provided to reduce the complexity of the graph [6].After developing the graphical structure we need to implement the structure in the architecture.

Minton and Steffen presented the idea that creating timely, maintainable and extensible framework for converging like data sources into homogenized entities is the goal of any successful ETL architecture. [7]

3. EXISTING ARCHITECTURE

After taking a tour of the above literature review of the researches done before we came across various anomalies of these researches.

To analyze the performance at every level multi-indicators were used which makes it complex and toilsome to use. Metadata management was a serious issue in previous researches as organizations were unable to track and monitor the data flow within the Business Environment.

Also the architecture's use was restricted to limited scenarios of Business Intelligences.

For management of various ETL processes different operators were introduced such as Create, invert, merge, match, etc.

But once any operation is applied on any data we can't revert back to its previous state. In case of any mistake, no method is defined to make it correct which was the major demerit in the architecture.

So to overcome this case UNDO operator is introduced.

Rest of the proposed approach is mentioned in the upcoming sections of the paper.

4. PROPOSED ARCHITECTURE

In the highly autonomous, distributed environment, performance issue is of great importance. So the evaluation and analysis of workflow performance have attracted great attention.

There are six layers shown as follows:

1. Business operation layer builds process model using modeling tools, and saves models in model DB (database) for process execution or simulation. Workflow execution data and log are stored in instance DB and log DB, while workflow simulation data and log are stored in simulation DB and log DB.

2. Original data layer includes several DBs which are sources of data analysis and data mining so that data can be extracted for further analysis and also the corresponding data source metadata is attached to it.

3. Data extraction layer extracts interesting information from original data by ETL (extract, transform and load) tools, and stores the information in business process information data warehouse. Meanwhile, considering the different data formats of data sources, an interface is added for transformation of data formats. Data warehouse management tools are responsible for maintaining data alongwith the tracing of metadata form the corresponding data warehouse metadata layer.

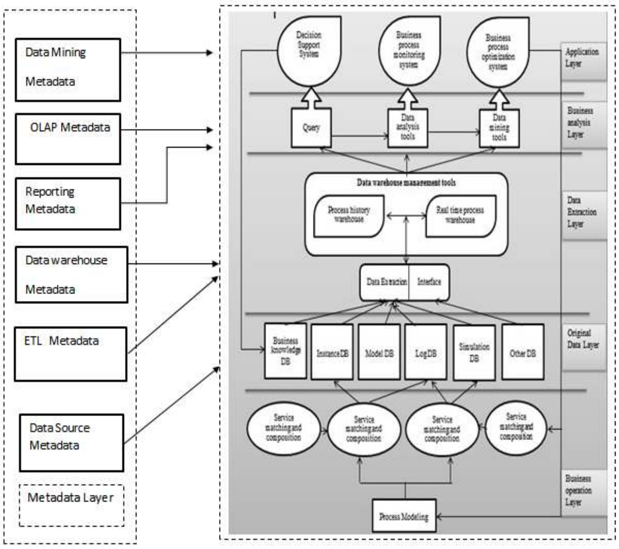

Fig 1. Meta ETL Architecture

4. Business analysis layer has three kinds of tools for different purposes. Query obtains related data from data warehouse and generates reports for users. Data analysis tools utilize the function of OLAP, and provide services on data analysis and decision support in Application layer. Data mining tools operate deep analysis, predict future development tendency, and discover relationships and rules among data. This layer stores all details regarding the data in OLAP metadata and Reporting metadata source.

5. Application layer includes three systems facing end users. Analysis of data done by user and also his/her interaction with that pool of data is carried out at this layer through data mining metadata. The result of business process optimization feeds back to process modeling and simulation. Business process monitoring system monitors the operation status and notifies exception to users. Decision support system assists decision-making activities and summarizes business rules and knowledge which are fed back to business knowledge DB.

6. Metadata Layer is included in each layer in the ETL architecture so that misinterpretation of data is avoided at each step during tracking of data flow. Adding Metadata layer having inbuilt concept of doubly linked list it makes the architecture applicable to all scenarios in Business Intelligence Environment.

A. *Extraction Process*

- Create – procreates an ETL process from a number of non-ETL data transfiguration steps by mapping between various data sources.
- Import – generates a tool-independent representation for a product-specific ETL process to support ETL management in a tool-independent manner.
- Match – finds for a given ETL process all corresponding ETL processes that extract, transforms, or load common data in an analogous way.
- Merge – takes independently developed ETL processes as input and return back a final merged ETL process.
- Invert – feeds back the output of one ETL process to the sources in order to share the benefit of any data cleansing and integration stage.
- Undo- Using undo operator we can change the operations done within a layer i.e. intralayer which makes the working simpler and accurate.

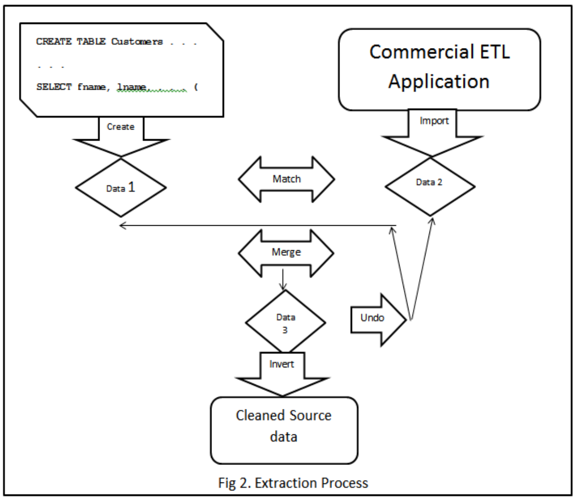

Fig 2. Extraction Process

5. CONCLUSION

This paper proposes the architecture of ETL process which is applied on BI environment along with the advent of metadata at each corresponding layer that can be applicable to all the scenarios of BI. The management of extraction process has been done using several operators like Create, Invert, Merge and Undo operator was also introduced which help in reducing its complexity. At every level Metadata layer is added so that the organization can outline and monitor the data flow within the Environment via the concept of doubly linked list at every layer of the architecture.

In this paper an attempt has been made to reduce the complexity of the system and to manage the various ETL processes up to a possible level.

6. REFERENCES

[1] H. J. Watson and B. H. Wixom. The current state of business intelligence. Computer, 40(9):96-99

[2] Nitin Anand, ETL and its impact on Business Intelligence, International Journal of Scientific and Research Publications, Volume 4, Issue 2, February 2014 1 ISSN 2250-3153

[3] Bo Liu and Yushun Fan, Research on Architecture and Key Technology for Service-Oriented Workflow Performance Analysis, Department of Automation, Tsinghua University, Beijing 100084, China

[4] Alexander Albrecht Felix Naumann, Managing ETL Processes, Hasso Plattner Institute at the University of Potsdam, German

[5] Huong Morris*, {Hui Liao, Sriram Padmanabhan, Sriram Srinivasan},

a. {Phay Lau, Jing Shan, Ryan Wisnesky}**, Bringing Business Objects into Extract-Transform-Load (ETL) Technology, IBM T. J. Watson Research, 19 Skyline Drive, Hawthorne, NY 10532.

[6] Panos Vassiliadis, Alkis Simitsis, Spiros Skiadopoulos , Modeling ETL Activities as Graphs, National Technical University of Athens, Dept. of Electrical and Computer Eng., Computer Science Division, IroonPolytechniou 9, 157 73, Athens, Greece.

[7] Pat Minton and Don Steffen,The Conformed ETL architecture.