Extraction, Transformation, Loading (ETL) and Data Cleaning Problems

Sunil Kumar[*] and Muhammad Nadeem
sksoftmind@gmail.com and nadeem@szabist.edu.pk
SZABIST
Karachi, Pakistan

Abstract: *Extraction, Transformation and Loading (ETL) is a process of enterprise data warehouse where process data is transferred from one or many data sources to the data warehouse. This research paper discusses the problems of the ETL process and focuses on the cleaning problem of ETL. Extraction, Transformation and Loading is a very time consuming process of data warehouse, taking 80% of time of the total data warehouse design time. During extraction and transformation of data, cleaning also is one of the parts where the process is especially required at the time of integration of metadata or heterogeneous data sources together with schema related data transformations. In data warehouses, data cleaning is a major part of ETL process. In this report, current tools are also discussed for data cleaning. Data warehouses require and provide extensive support for data cleaning. They load and continuously refresh huge amounts of data from a variety of sources, so the probability that some of the sources contain "dirty data" is high; data warehouses are used for decision making, so that the correctness of their data is vital to avoid incorrect conclusions. For instance, duplicated or missing information will produce incorrect or misleading statistics ("garbage in, garbage out"). So the correct cleaning of data is very important. Due to the wide range of possible data inconsistencies and the sheer data volume, data cleaning is considered to be one of the biggest problems in data warehousing.*

Keywords: *ETL extraction, transformation, loading, oracle*

1. INTRODUCTION

ETL (Extraction, Transformation and Loading) is a process to transfer data in data warehouse from one source to another. In this paper, author uses a case study, which is based on database facing some cleaning problems and future requirement of data cleaning in ETL [1].

Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to spelling mistakes during data entry, missing information or other invalid data. When multiple data sources need to be integrated, e.g., in data warehouses, federated database systems or global web-based information systems, the need for data cleaning increases significantly. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information becomes necessary.

2. CUSTOM BASE ETL SOLUTION

Custom base ETL development requires following steps:

- Execute stored procedures at the source to extract data, perhaps filtering some
- Store the extracted data into staging tables at the source
- Generate a file with the source staging data
- Transfer the file into the destination
- Load the data into staging tables at the destination
- Execute stored procedures at the destination to read the incoming staged data and compute what changed since the last time by comparing with the current value of the data warehouse
- Populate the data warehouse with new changes
- Supports incremental updates to sources
- Supports simple data cleansing
- Logs warning/errors

Oracle9i provides new server functionality in three areas: analytic capabilities, ETL (Extraction, Transformation, Loading), and data mining.

Utilizing Oracle9i as an ETL transformation engine, other tasks that are also essential for a successful ETL implementation; such as scheduling, monitoring and maintenance addressed by the Oracle solution with Oracle Warehouse Builder. ETL is facing some challenges and additional burden that they have not only to exchange but also need to rearrange, integrate and consolidate data over many applications and systems. During Extraction, following issues are to be considered:

- Different DBMS, OS, H/W, communication protocols
- Need a logical map, data movement view documents, data lineage report
  - – have a plan
  - – identity source candidates
  - – analyze source systems with a data profiling tool

[*] Student of MSCS Program at SZABIST

- receive walk-through of data lineage and business rules (from the DW architect and business analyst to the ETL developer)
- data alterations during data cleansing, calculations and formulas
- measure twice, cut once
- standard conformance to dimensions and numerical facts

Receive walk-through of the dimensional model and the desired data has to be identified and extracted from many different sources, including database systems and applications. Very often, it is not possible to identify the specific subset of interest; therefore, more data than is necessary has to be extracted; since the identification of the relevant data will be done at a later point in time. Depending on the source system's capabilities (e.g., OS resources), some transformations may take place during this extraction process. The size of the extracted data varies from hundreds of kilobytes to hundreds of gigabytes, depending on the source system and the business situation. Just as the size of the data extraction may vary widely, the frequency at which the data is extracted may also vary widely: the time span may vary between days/hours and minutes to near real-time. Web server log files, for example, can easily become hundreds of megabytes in a very short period of time, thus necessitating frequent extractions [2].

2.1 ETL Process Classified

Oracle classified the process in two main categories:

1. ETL processing outside the database
2. Loading into a database staging area for ETL processing

2.1.1 ETL Processing Outside the Database

Most of the transformation and cleansing is done outside the database, in separate standalone ETL engines/processes. These engines work with various data sources, trying to integrate them for the necessary ETL steps. If existing data in the target database is required, e.g., for data cleansing or ID lookup, the target database is treated like every other external data source involved in the ETL process. Many of the required exercises could be addressed with basic SQL capabilities: joins sorts, string manipulations, and validation of referential integrity.

2.1.2 Loading into a Database Staging Area for ETL Processing

Different sources in various formats reside outside the database. Rather than using an external engine as the single point of control, the database is used. All raw data is loaded mostly unchanged in neutral staging structures. If the source systems are relational databases, the staging tables will be typical relational tables. If the source systems are non-relational, the data may be staged in tables with columns like VARCHAR2 (4000), for further processing inside the database. After successfully loading the external data unmodified into the database, the transformation steps take place inside the database. This is the serial approach load-then-transform.

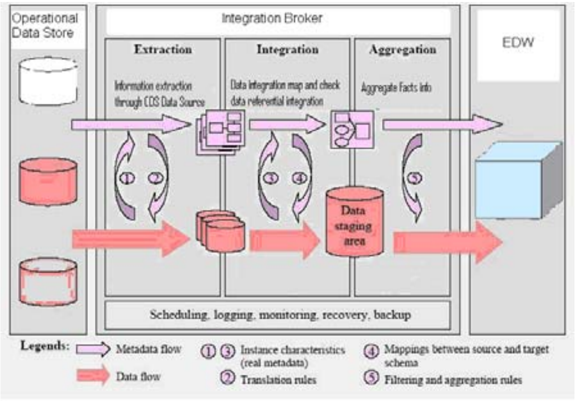
Operational Data Store
Integration Broker
EDW
Extraction
Information extraction through ODS Data Source
Integration
Data integration map and check data referential integration
Aggregation
Aggregate Facts into
①
②
③
④
⑤
Data staging area
Scheduling, logging, monitoring, recovery, backup
Legends:
Metadata flow
Data flow
① ③ Instance characteristics (real metadata)
② Translation rules
④ Mappings between source and target schema
⑤ Filtering and aggregation rules

Fig. 1. Steps of building DW: the ETL process

2.2 New Paradigm

Beginning with Oracle9i, Oracle's database capabilities are significantly enhanced to address specifically some of the tasks in ETL environments. The ETL process flow can be changed dramatically and the database becomes the integrated data transformation engine.

3. ETL IN ORACLE

Where Oracle9i is the first real business intelligence platform in history and in the market, it comes with most successful RDBMS data warehousing; by introducing new and best features and enhancing existing functionality. It increases the importance and the value of database in the business intelligence platform in every data warehouse environment. It also provides a broad tool kit of powerful functionality specifically targeted towards the ETL process, enabling you to define your own specific ETL process and data flow, and taking advantage of the reliability, scalability and performance of Oracle's proven database technology.

4. PROBLEMS IN ETL

- General ETL issues
- The ETL/DW refreshment process
- Building dimensions
- Building fact tables
- Extract
- Transformations/cleansing
- Load
- MS Integration Services
- A concrete ETL tool
- Demo
- Example ETL flow

- The most underestimated process in DW development
- The most time-consuming process in DW development
- Often, 80% of development time is spent on ETL
- Extract
- Extract relevant data

4.1 DW Phases

- Design phase
- Modeling, DB design, source selection
- Loading phase
- First load/population of the DW
- Based on all data in sources
- Refreshment phase
- Keep the DW up-to-date i.e. source data changes

4.2 Data Staging Area

- Transit storage for data underway in the ETL process, Transformations / cleansing done here
- No user queries (some do it)
- Sequential operations (few) on large data volumes
- Performed by central ETL logic
- Easily restarted
- No need for locking, logging, etc.
- RDBMS or flat files? (DBMS have become better at this)
- Finished dimensions copied from DSA to relevant marts
- Allows centralized backup/recovery
- Often too time consuming to initial load all data marts by failure
- Thus, backup/recovery facilities needed
- Better to do this centrally in DSA than in all data marts

5. ARKTOS

"ARKTOS" is a framework capable of modeling and executing the Extraction-Transformation-Load process (ETL process) for data warehouse creation. The authors consider data cleansing as an integral part of this ETL process which consists of single steps that extract relevant data from the sources, transform it to the target format and cleanse it, then load it into the data warehouse. A meta-model is specified allowing the modeling of the complete ETL process. The single steps (cleansing operations) within the process are called activities. Each activity is linked to input and output relations. The logic performed by an activity is declaratively described by a SQL-statement. Each statement is associated with a particular error type and a policy, which specifies the behavior (the action to be performed) in case of error occurrence.

Six types of errors can be considered within an ETL process specified and executed in the ARKTOS framework. PRIMARY KEY VIOLATION, UNIQUENESS VIOLATION and REFERENCE VIOLATION are special cases of integrity constraint violations. The error type NULL EXISTENCE is concerned with the elimination of missing values. The remaining error types are DOMAIN MISMATCH and FORMAT MISMATCH, referring to lexical and domain format errors.

The policies for error correction simply are IGNORE, but without explicitly marking the erroneous tuple, DELETE as well as WRITE TO FILE and INSERT TO TABLE with the expected semantics. The latter two provide the only possibility for interaction with the user.

The success of data cleansing can be measured for each activity by executing a similar SQL statement counting the matching/violating tuples. The authors define two languages for declaratively specifying of the ETL process. This is accompanied with a graphical scenario builder" [3].

6. DATA CLEANING PROBLEMS

This document is not a regular requirements specifications document. It comprises the intended future, direction for the ETL Integrator product. It will serve as foundation for the definition of individual future requirements (to be described in individual requirement specification documents). For that reason, it only covers a limited set of new basic requirements enabling planned additional, future requirements, which are outside the scope of this document. This document's foundation is ETL future versions providing support for most commonly used databases; transformation operators (functions) included in the SQL language standard (ANSI 92), and other user functionalities [4].

7. CASE STUDY

Traditionally, the refreshing of data in data warehouses has been performed in an off-line fashion. Active Data Warehousing refers to a new trend where data warehouses are updated as frequently as possible, to accommodate the high demands of users for refresh data.

Discuss the particular steps of their refreshment process and the differences to the view maintenance problem in a similar way. The authors define warehouse refreshment as a sequential process consisting of four steps:

1. Preparation: is performed for each source and includes the tasks of extraction, cleaning, and possibly data archiving.
2. Integration: consists of reconciliation of data originated from heterogeneous sources, and derivation of base relations (or base views) stored in an ODS.

3. Aggregation: during this step, aggregated views are computed from base views. According to the data warehouse architecture presented in the paper, aggregated views are stored in the corporate data warehouse, and are used for computing the contents of data arts.
4. Customization: refers to the derivation and customization (regarding various forms of presenting multidimensional data) of user views, which define the data marts.

Thus, similar to our viewpoint, view maintenance is considered as one step of the entire refreshment process. Furthermore, the authors discuss quality requirements related to design of warehouse refreshment system, considering source and warehouse parameters (e.g., availability, frequency of change, storage space limits, etc.), the focus of their future work.

In our framework, we have implemented ETL activities over queue networks and employ queue theory for the prediction of the performance and the tuning of the operation of the overall refreshment process. Due to the performance overheads incurred, we explore different architectural choices for this task and discuss the issues that arise for each of them [5].

7.1 Transformation Steps

Simple transformation and integration steps can be defined using a variety of predefined mappings. Extraction of operational data and loading data into the warehouse is supported by providing pre-defined interfaces to major data management systems, ERP systems etc. However, the extensibility of these systems is limited, e.g., with regard to:

- Integrating complex data types
- Defining complex or nested mappings
- Providing support for key management
- Extending the refreshment process by user-defined refreshment tasks
- Loading various target warehouses with different design and storage techniques (in relational or multidimensional database systems)
- Supporting history management techniques in the target warehouse

Furthermore, solutions for advanced history and key management issues are mostly missing. Thus, the majority of today's industry projects implement particular tasks of the refreshment process by several application programs and scripts, which are difficult to maintain, especially in complex data warehouse environment.

8. CONCLUSION

In this report, the ETL process is discussed in detail and their problems were focused on cleaning problem of ETL process, where data can be loss. It is defined as the sequence of operations intending to enhance to overall data quality of a data collection. There is only a rough description of the procedure in data cleansing, as it is highly domain dependent and explorative. Existing data cleansing approaches mostly focus on the transformation of data and the elimination of duplicates. Some approaches enable the declarative specification of a more comprehensive data cleansing process, still leaving most of the implementation details for the cleansing operation to the user. There still remain a lot of open problems and challenges in data cleansing.

They mainly concern the management of multiple, alternative values, the management and documentation of performed cleansing operations and the cleansing lineage, as well as the specification and development of an appropriate framework, supporting the data cleansing process.

REFERENCE

[1] Galhardas, H., Florescu, D., Shasha, D., and Simon, E. Ajax: An Extensible Data Cleaning Tool. Proceedings of the 2000 ACM SIGMOD international conference on Management of data, p.590, May 15-18, 2000, Dallas, Texas, United States.

[2] ETL Processing within Oracle9i, An Oracle White Paper, June 2001, Author: Hermann Baer (Copyright © 2000 Oracle Corporation. All rights reserved).

[3] Problems, Methods, and Challenges in Comprehensive Data Cleansing Heiko Müller, Johann-Christoph Freytag Humboldt-Universität zu Berlin zu Berlin, 10099 Berlin, Germany {hmueller, freytag}@dbis.informatik.hu-berlin.de

[4] ETL Integrator Future Ideas By Ahimanikya Satapathy.

[5] ETL Queues for Active Data Warehousing Alexandros Karakasidis, Univ. of Ioannina Ioannina, Hellas, alex@cs.uoi.gr, Panos Vassiliadis Univ. of Ioannina, Hellas pvassil@cs.uoi.gr Evaggelia Pitoura Univ. of Ioannina Ioannina, Hellas pitoura@cs.uoi.gr