Optimize ETL For Banking DDS : Data Refinement Using ETL Process For Banking Detail Data Store(DDS)

Akshay S. Lohiya¹, Dayanand M. Gojare², Sachin M. Karde³, Akshay A. Patil⁴
Department Of Information Technology, JSCOE, Pune, India

Abstract—Our project deals with unstructured and raw data and converts it to structured and consistent data by applying modern data warehousing, business intelligence concept and ETL tool Informatica. Throughout the project we are in continuous communication with clients and get updated files for correction these modern tool we help us for generating audit reports. We classify data quality problems that are addressed by data cleaning and provide an overview of the main solution approaches. Data cleaning is especially required when integrating heterogeneous data sources and should be addressed together with schema-related data transformations. In data warehouses, data cleaning is a major part of the so-called ETL process. We also discuss current tool support for data cleaning.

Keywords: datawarehouse, extraction transformation AND loading(ETL), database

I. INTRODUCTION
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 misspellings 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 become necessary.

II. LITERATURE REVIEW AND RELATED WORK
Why ETL Tool are Used:
In today’s businesses, decision-making processes and daily operations often depend on data that is stored in a variety of data storage systems, formats, and locations. In order to turn this data into useful business information, the data typically needs to be combined, sanitized, standardized, and summarized. For instance, information may need to be converted to a different data type or heterogeneous database servers may store the necessary data using different schemas. Dissimilarities like these must be resolved before the data can be successfully loaded to a target system. After the design and development of data warehouse in accordance with the business requirements, the process of consolidating the data into the data warehouse from various sources is to be addressed. Extract Transform Load (ETL) processes are critical in the success of the Data Warehousing projects. The process of extracting data from one system (extract), transforming it in accordance with the design of the data warehouse (transform) and loading it data warehouse system (load) constitute ETL. In other words, ETL is the process of extracting data from various data sources, transforms it as per the requirements of the destination data warehouse and successfully loading it into the destination data warehouse (database). In the transformation process data is actually standardized to make it compatible with the destination database along with data cleansing (cleaning) operations. The process of ETL is detailed in the following sub sections.

A. Extract:
The first step of ETL process is extraction of data from various data sources that contain the information that need to be transferred to the data warehouse. Some of these sources might be relational, some of them might be just single flat files without any data integrity rules. In the extract process, data is extracted from the source system and is made accessible for further processing. The main objective of the extract step is to extract the required data from the source systems utilizing least possible little resources. Further, the extract process should be designed in such a way that it does not affect the source system in terms of performance, response time or any kind of locking. Data extraction can be performed in several ways such as update
notifications, incremental extract, full extract etc. The frequency i.e. number of times an extract to be performed or the time interval between each extract is very crucial in the case of incremental of full extracts as the volumes of the data can be in tens of gigabytes.

**B. Transform:**
The most complex part of ETL process is the transformation phase. At this point, all the required data is exported from the possible sources but there is a great chance that data might still look different to the destination schema of the data warehouse. At times data itself need to be formatted to conform to the data types and other constraints of the data warehouse. In some cases data need to be represented differently in order to make it more fruitful. In the transform process, a set of rules are applied to transform the data from the source to meet the requirement of the target which is a data warehouse. This includes converting any measured data to the same dimension (i.e. conformed dimension as per the requirements of the data warehouse) using the same units so that they can later be joined. The transformation process also involves joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules.

**C. Load**
Once the data is extracted and transformed according to the requirements of the target data warehouse, the data is assumed to be ready for loading. However, several aspects like how the data to be loaded, the impact and implication of loading process as well as handling such implications are to considered before loading the data into the data warehouse. The process of loading may impact the processing speed of the server both for loading as well as analysis, can occur. It is also crucial to avoid database crippling while loading the data. ETL processes can be performed using almost any programming language. Building such program from scratch can be complex which makes it is necessary to use ETL tools. ETL process can be carried out manually or by using a tool of automation. When the ETL process is carried out with a automation tool, data mappings are fed into the tool of automation and the code that performs the mappings is created. Generally mappings are done manually when there are only few procedures to be written. However, it is more efficient to use an automation tool for ETL process. The widely used open source ETL tool is Talend Open Studio and proprietary tool is SQL Server Integration Services.

**III. Architectural Design**

**IV. PROPOSED SYSTEM**
To conceptualize the ETL processes used to map data warehouse schema, we studied the previous projects, made some integration, and added some extensions to the approaches mentioned above. We are using ETL tool for Transforming banking detail data store by applying Extraction, Transformation and Loading operation. We are also using Mapping between different Tables to get the desired Result.

To conceptualize the ETL processes used to map data from sources to the target data warehouse schema, we studied the previous research projects, made some integration, and added some extensions to the approaches mentioned above. We propose entity mapping diagram (EMD) as a new conceptual model for modeling ETL processes scenarios. Our proposed model mainly follows the approach of modeling based on conceptual constructs. The proposed model will fulfill six requirements
1. Supports the integration of multiple data sources.
2. Is robust in view of changing data sources.
4. Can be easily deployed in a suitable implementation environment.
5. Is complete enough to handle the various extraction, transformation, and loading operations.
6. Is simple in creating and maintaining.

**V. CONCLUSION AND FUTURE SCOPE**
Data cleaning will be considered one of the most important frontiers and one of the most promising
interdisciplinary developments in Information technology. In this paper we try to briefly review the data cleaning process, sources of error in data. We further outlined the importance of data quality its challenges and provide an overview of the main solution approaches. Furthermore we provide an overview of data cleansing in data warehouses using ETL Tool. This study would help the researchers to focus on the various issues of data cleansing. Data Cleansing is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains such as finance, marketing, banking, insurance, health care and retailing. Data cleansing is commonly used in these domains to increase the sales, to reduce the cost and enhance research to reduce costs, enhance research.

VI. REFERENCES


