# Database Agnostic ETL Tool: An Economical Solution to Store and Process Data

Sumedha Sirsikar, Apoorva Jain, Dixha Raina, Ritu Munot, Warun Nimbhorkar
Maharashtra Institute of Technology, Pune

***Abstract*— People in society are producing data in an explosive manner. Various organizations use this data for decision support and to construct data-intensive products and services. A purpose of network management tool is to manage network devices within the data centers of an enterprise that is spread across the globe. Any organization's Decision Support Network Automation is the Business Intelligence (BI) system built for advance and custom reporting on the stored data by performing Extract, Transform and Load (ETL) operations using On-line Transaction Processing (OLTP) Database. Similarly ETL tool is a process in database usage and especially in data warehousing. This extracts data from homogeneous or heterogeneous data sources, transforms the data for storing it in proper format or structure for querying and analysis purpose and finally loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse). In this tool usually all the three phases execute in parallel as the data extraction process requires time. This paper provides an advance solution to existing database tools.**

***Keywords*— *Extract Transform Load (ETL), Organization Network Automation , Database Agnostic, Data Warehouse***

## I. INTRODUCTION

Organization Network Automation (ONA) is a network management tool for managing network devices within the data centers of a globally distributed enterprise. ONA stores all the information in OLTP database. Organization Decision Support Network Automation (ODS-NA) is the Business Intelligence (BI) system that is built to do advance and custom reporting on the data stored in ONA by performing ETL on ONAs OLTP Database. ONA extracts, transforms and loads (ETL) this data to a data warehouse. Currently ODS-NA uses Oracle Data Integrator (ODI) to perform the ETL operation and then there are built-in reports that run of SAP Business Objects BI Platform and IBM Cognos. User has the flexibility to create his own custom reports or modify the existing reports to meet his own requirements. Having a separate BI system consisting of Data Warehouse reduces the load on ONA OLTP database to avoid performance issues. It provides a more systematic way to organize the information by following the concepts of Dimensional Modeling and custom reporting. It also provides a way to store information for longer durations. Information from more than one ONA can be pushed to one data warehouse to provide a one holistic view to user by providing consolidated data of all the ONA system to one ODS-NA BI System.

Extract, Transform, Load; three database functions that are combined into one tool that automates the process to pull data out of one database and place it into another database. The database functions are described following [4].

Extract - the process of reading data from a specified source database and extracting a desired subset of data.

Transform - the process of converting the extracted/ acquired data from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining with other data.

Load - the process of writing the data into the target database.

Currently the product supports Oracle, SQL Server and PostgreSQL and ETL is done using Oracle Data Integrator. There is separate ETL developed for each of these database type and hence for every change to ETL we need to modify the ETL process for each database type. The objective of this project is to develop ETL tool that would replace Oracle Data Integrator and would be database agnostic so that making change at single place would make those changes applicable for any of database type supported.

Rest of the paper is organized as follows. Section II describes related work. Section III describes the Proposed System, its high level design and architecture.

## II. RELATED WORK

With the gradual evolution of Data Warehousing, organizations required a process to load and maintain data in a Warehouse. ETL is one of the most important sets of processes for the sustenance and maintenance of Business Intelligence architecture and strategy [5].ETL process evolved and gradually took control over the Data Warehousing market to fulfill this requirement. Initially, organizations developed their own custom codes to perform the ETL activity which was referred as hand-coded ETL process. Since the process was lengthy and quite difficult to maintain, vendors started developing off the shelf tools which could perform the same task as that of Hand-coded ETL but in efficient manner. In this era of ETL tools, the market saw different generations, different types and tools with their own pros and cons.

These are many ETL tools available in the market. However there are few problems with them. Such tools are

very expensive and does not support small size businesses. Configuration of such tools takes lot of time. Tools customization is not possible and thus, sometimes does not support the scenario provided. These tools have often been seen having common problems of:

a) Data dependencies.
b) Complexity of source code.

c) Poor Quality of data[1].

ODI that was introduced in January 2000 has been used by organization for performing ETL. It has following disadvantages:

- Increasing cost to organization
- It does not support heterogeneous databases together i.e. it is not database agnostic
- Focus on ETL solutions, rather than in an open context of data management
- Tools are used mostly for batch-oriented work and transformation rather than real-time processes or federation data delivery
- Long-awaited bond between Oracle Warehouse Builder(OWB) and ODI brought only promises - customers confused in the functionality area and the future is uncertain

Considering all this issues, there is a need to develop new system that is more efficient than the existing system:

The features of the new system are as follows:

1. Design simplicity
2. Cost effectiveness
3. Database Agnostic ETL tool
4. Facilitates Custom reporting using SAP Business Objects or IBM Cognos[1].

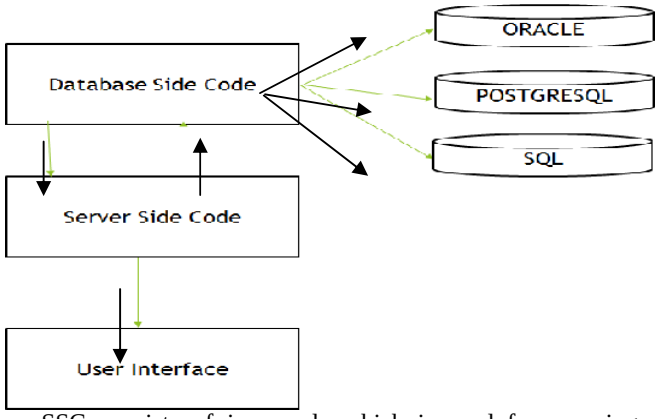
## III. PROPOSED SYSTEM

The proposed system includes building an ETL tool for the organizations that would replace existing Oracle Data Integrator (ODI) tool thus reducing immense cost to the company and also making it database agnostic.

Following diagram depicts 3 modules of our proposed system. These are:

a. DataBase Side Code (DBSC)
b. Server Side Code (SSC)
c. User Interface (UI)

DBSC consists of code which will be executed on the OLTP database. It comprises of query files and XML files such that each database has different query files.




SSC consists of java code which is used for mapping purpose. It maps heterogeneous source databases into target database. Thus, this allows provision of making the system database agnostic.

User Interface facilitates audit records, real time status check. It acts as a wrapper between DBSC and SSC.

The functionalities of the proposed system are as follows:

1. To provide more systematic way to organize information which is supported by target database format.
2. To store information for longer duration to be used for forecasting and analysis for future purposes.
3. To replace Oracle Data Integrator which reduces the cost as ODI is licensed and not an open source software
4. To make database agnostic so that queries from heterogeneous source databases could be run on target database irrespective of the database type the target database supports.
5. To generate holistic view of consolidated data from all the ONA systems
6. To reduce load on ONA OLTP as ONA can store data collected in 60 days and beyond this limit data cannot be stored in to ONA-OLTP

The following assumptions are made for the proposed system.

1. It is assumed that source database could be SQL Server, PostGre, Oracle and the queries from these heterogonous databases would be properly mapped to the database supported by the data ware house (BDS-NA).
2. While compressing the data, no data loss occurs.

Extract, Transform and Load (ETL) refers to a process in database usage and especially in data warehousing that: Extracts data from homogeneous or heterogeneous data sources, transforms the data for storing it in proper format or structure for querying and analysis purpose and loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse) [1] [3]. Usually all the three phases execute in parallel since the data extraction takes time, so while the data is being pulled another transformation process executes, processing the already received data and prepares the data for loading and as soon as there is some data

ready to be loaded into the target, the data loading kicks off without waiting for the completion of the previous phases. Organization Network Automation (ONA) has following functions:

1. Installing operating system on network devices
2. Configuration of network devices
3. Taking backup and restoration of the configuration
4. Enforcing government and enterprise wide policies
5. Pod and Network Container Management for private/public/hybrid clouds
6. Configuration of Software defined networking
7. Assuring network compliance and many others

*A. High Level Design*

High level design of proposed system is shown in Fig. 1. The components of the high level design of proposed system are described as follows:

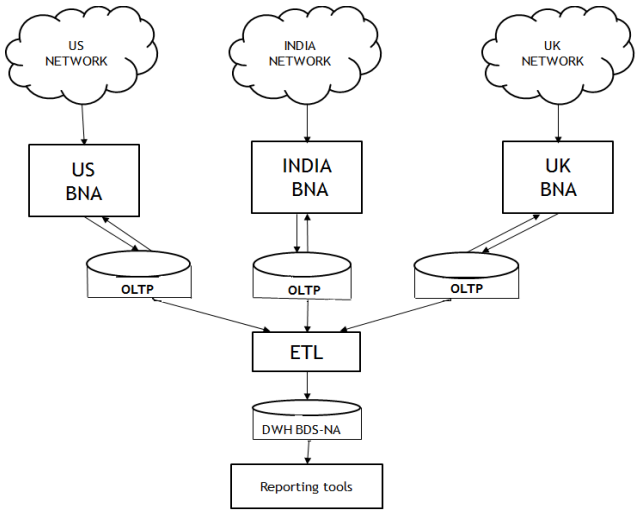



Figure 1: High level design of Proposed System

**ONA Site:** It is source of data which has different network devices spread across globe. These data may have information about configuration, settings, operating system used by different network devices like routers, switch, etc.

**OLTP:** It is a class of information systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing. It involves gathering of input information.

**ETL:** It is a tool which extract data from source, transform it into different form and load it into Data ware house.

**DWH ODS-NA:** It is Data ware house where all the data stored after being subjected to ETL tool. Data stored here will be transformed set of OLTP data.

**BusinessObject / Cognos:** It is set of reporting tool when applied to Data ware house generate different types of data.

*B. System Architecture*

The process involved in ETL Tool is described as follows:

Data from different OLTP are subjected to ETL. ETL is agnostic of databases used, such that it can accept data from various external databases like PostGre, SQL Server, MongoDB.

Fig. 2 shows the architecture of proposed system. The various databases in the ETL Tool are as follows:

**Staging:** A staging area is an intermediate storage area used for data processing during the ETL process. The data staging area sits between data sources and data target, which are often data-warehouse.

**MetaData Repository:** Metadata is data about data. For example, the index of a book serves as a metadata for the content in the book. So here, metadata is summarized data that leads us to detailed data. Meta data acts as directory. Metadata helps in decision support system for mapping of data when data is transformed from operational environment to data warehouse environment.

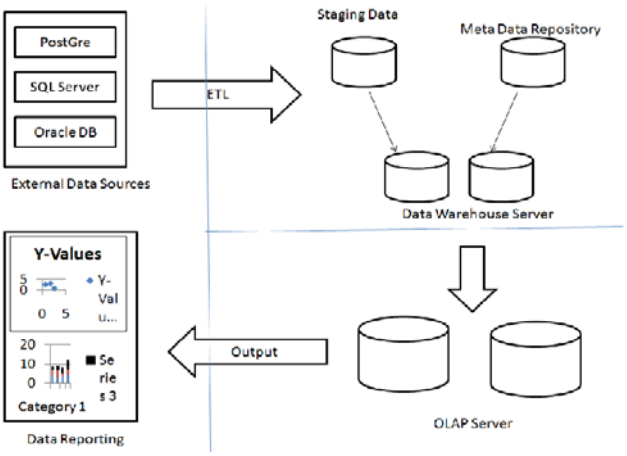



Figure 2: Proposed System Architecture of ETL Tool

**Data Ware house Server:** It has data which is to be loaded into OLAP.

**OLAP:** In OLAP, data stored is used for analytical purpose by which reporting can be done on data. Different reporting tools are used to generate report from OLAP for example BusinessObject , Cognos.

## IV. Conclusion

In this paper, we have proposed an innovative method to improve the performance of ETL by making it database agnostic. The primary goal of the proposed system is to make it economical by making it easily available to the organization. It facilitates data storage for a longer duration, thus overcoming the limitations of OLTP databases. Metadata layer allows provision of custom reporting with the help of System Application Products (SAP) Business Objects and International Business Machine (IBM) Cognos. Thus, this paper focuses on development of an ETL tool which is database agnostic. It will try to reduce the complexity of the tool and also reduces an immense cost to the company.

## REFERENCES

[1] Satkaur, Anuj Mehta Research scholar, S.K.I.E.T Proposed work on ETL. Volume 3, Issue 6, June 2013 ISSN: 2277 128X.

[2] M. GOLFARELLI, D. MAIO AND S. RIZZI, The Dimensional Fact Model: a Conceptual Model for Data Warehouses, International Journal of Cooperative Information Systems, Vol. 7(2&3), pp. 215–247.

[3] C. WHITE, Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise, The Data Warehousing Institute, October, 2005. Building an ETL Tool by Ahimanikya Satapathy SOA/Business Integration ,Sun Microsystems

[4] Fundamentals of database systems. 4th Edition. Pearson International and Addison Wesley. Ramez Elmasri and Shamkant B. Navathe.