Extract Transform & Load Tool using Oyster

Rahul Sharma[1] Kartik Vijan[2]

[1,2]Department of Information Technology

[1,2]MAIT, Delhi, India

Abstract— The data in the world is ever increasing and with that comes the problem of storing the data in an efficient manner using limited storage. This containment of data will be very useful in controlling 'Big Data' by the methods of Data Warehousing. ETL tool is one such tool that does the job for us because it – Extracts the data from the source, Transforms it according to the user requirements and then Loads it back for the efficient usage of the quality data. This ETL process can be achieved using trademarked and open source tools. Oyster is an open source tool which is employed to resolve entities existing in the real world. Even though there exist other trademarked tools in the market, in this paper we demonstrate the applicability of Oyster, an open source tool which can perform equally.

Key words: ETL, Oyster, Entity Resolution, Identity Capture

I. Introduction

ETL tool was born because of Data Warehousing. The term" Data Warehouse" was first coined by Bill Inmon in 1990. According to him, Data warehouse is subject Oriented, Integrated, Time-Variant and non-volatile collection of data that supports decision making process in an organization. The operational database undergoes several day to day transactions which makes the process of data analysis more and more complex and time consuming. [1] The basic need for warehousing has originated from the increasing transactional data with time. The transactional data of an organization must be kept in easy access for various scenarios of analysis of data.

Challenge:

a) The tool must be designed keeping in mind the requirements of the organizations.
b) The data can be in any form, i.e., it can be either a heterogenous database or a non heterogenous database.
c) There are a lot of inconsistancies present in a transactional database.

Above presented reasons have led to designing an ETL tool which verges these confines and must be processed within service level agreements. Oyster is one open source tool that has the capability of overcoming all these mentioned challenges. Any kind of a dataset can be handled using Oyster.

II. The Process of ETL

A. Extract

The first step of an ETL process that is 'Extract' refers to extracting data files from the source. This is the beginning and the most important aspect of the whole process. An intrinsic part of the extraction involves data validation to confirm whether the data pulled from the sources has the correct/expected values in a given domain (such as a pattern/default or list of values). If the data fails the validation rules it is rejected entirely or in part. The rejected data is ideally reported back to the source system for further analysis to identify and to rectify the incorrect records. In some cases, the extraction process itself may have to do a data-validation rule in order to accept the data and flow to the next phase. [2]

B. Transform

The second step of the process is when certain rules and functions are applied to the retrieved data extracted from the source. In this method, Oyster comes into play. The rules are defined in accordance to the columns present in the file. The transformation of the data entities targeted is deduplication in this case. The method of Entity Resolution is employed to do so.

C. Load

The last phase of the process is loading the transformed data in to end target depending on the organizations requirements. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis. Other data warehouses (or even other parts of the same data warehouse) may add new data in a historical form at regular intervals—for example, hourly. [3]

Extract
Transform
Load

Fig. 1: ETL Tool

III. Oyster

The OYSTER Open Source Project is sponsored by the Center for Advanced Research in Entity Resolution and Information Quality (ERIQ) at the University of Arkansas at Little Rock. It is intended to provide an entity resolution system that includes functionality for entity identity information management (EIIM). Originally developed as a teaching tool, it now has enough capability to support record linking, EIIM, and master data management (MDM) processes in small to medium-sized organizations.

OYSTER is designed to be easily configurable through the use of several, run-time XML scripts that define such things as the format and locations of reference sources to be processed, access to previously defined identity structures, identity rules and associated matching algorithms, as well as many parameters that adjust system performance to particular ER applications. These scripts allow OYSTER to be configured to run in different ER modes or architectures including record linking/merge-purge, identity resolution, identity capture, and identity update. [4]

Oyster runs three files:

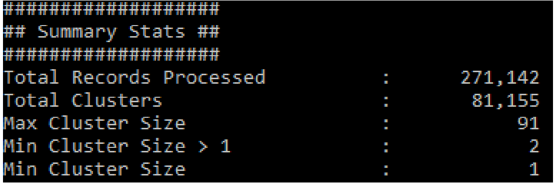
- RunScript
- SourceDescriptor
- Attributes

The matches are made on the basis of:

- Direct Matching
- Transitive Equivalence
- Relationship Resolution
- Asserted Equivalence

Identity capture is a form of identity resolution in which the system builds (learns) a set of identities from the references it processes rather than starting with a known set of identities. [5]

After the transformation is completed, every unique reference in the database is assigned an OysterID.



```
###################
## Summary Stats ##
###################
Total Records Processed          :      271,142
Total Clusters                   :       81,155
Max Cluster Size                 :           91
Min Cluster Size > 1             :            2
Min Cluster Size                 :            1
```

Fig. 2 The output of the ETL Tool

IV. Result

In order to calculate the obtained accuracy, we compare the truth set with the obtained match set.

$$TWi(T,P) = \frac{\sqrt{|T|\cdot|P|}}{|V|}$$

Fig. 3: Calculation of the Talburt-Wang Index

Here T and P are two partition sets of the same data S. T corresponds to the truth set, while P is the set whose accuracy we want to measure.

- V is the total number of overlaps.
- The value of Talburt-Wang Index lies between 0 and 1. The Talburt-Wang Index of 1 signifies 100% accuracy. [6]
- An accuracy of 49.85% was obtained for the dataset of 271,142 entities.

V. Conclusion

A well designed ETL tool is created using OYSTER Open Source Project. It was inferred that data warehousing is done by the application of the ETL tool.

This tool can facilitate the data warehousing projects of proprietary and government organizations efficiently.

References

[1] Ranjith Katragadda, Sreenivas Sremath Tirumala, David Nandigam. “ETL tools for Data Warehousing: An empirical study of Open Source Talend Studio versus Microsoft SSIS” Retrieved from the HYPERLINK “ https://www.researchgate.net/publication/271207443_ETL_tools_for_Data_Warehousing_An_empirical_study_of_Open_Source_Talend_Studio_versus_Microsoft_SSIS “

[2] Retrieved from the HYPERLINK “ https://en.wikipedia.org/wiki/Extract,_transform,_load ”

[3] Retrieved from the HYPERLINK “ https://sourceforge.net/p/oysterer/home/Home/ ”

[4] Fumiko Kobayashi, John R. Talburt, “ Introduction to Entity Resolution with OYSTER v3.3”, Document Version: 1.10, Date: 02 December 2012, Copyright © 2012 ERIQ University of Arkansas at Little Rock

[5] John R. Talburt , Emily Kuo , Richard Wang, Kimberly Hess. “AN ALGEBRAIC APPROACH TO QUALITY METRICS FOR CUSTOMER RECOGNITION SYSTEMS” Retrieved from the HYPERLINK “http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202004/Papers/AnAlgebraicApproach2QualityMetrics.pdf”