An Open Source ETL Tool - Medium and Small Scale Enterprise ETL (MaSSEETL)

Rupali Gill
Research Scholar
School of Computer Sciences, CU Punjab

Jaiteg Singh
Associate Professor
School of Computer Sciences, CU Punjab

ABSTRACT
In Data Warehouse (DW) environment, Extraction-Transformation-Loading (ETL) processes consumes up to 70% of resources. Data quality tools aim at detecting and correcting data problems that affect the accuracy and efficiency of data analysis applications. Source data imported into the data warehouse often has different quality, format, coding etc. In order to bring all the data together in a standard, homogeneous environment, Extraction–transformation–loading (ETL) tools are used. ETL solutions provided so far are either proprietary and have limited functionality. Small and Medium Scale Enterprises(SME) and Small Scale Enterprises (SSE) cannot afford the licensing cost of these paid tools. The developed tool is capable of providing an integrated and open source data quality solution - MaSSEETL is to deal with naming conflicts, structural conflicts, date conversions, missing values and changing dimensions. *MaSSEETL* solves the appropriate errors with appropriate level of warning. In this paper, we are presenting the working of *MaSSEETL*.

The tool provides an pragmatic evidence of strategic intensification of quality data in the academic and business enterprises.

General Terms
Data warehousing, data cleansing, quality data, dirty data

Keywords:
Data inconsistency, identification of errors, organization growth, ETL, data quality

1. INTRODUCTION
An approach by Bill Inmon describes the Data warehouse as Subject Oriented, Integrated, Time-Variant and non volatile collection of data. This data helps in supporting decision building methods by analyst in an organization. The challenge in data warehouse environments is to incorporate, rearrange and consolidate large volumes of data over many systems, to provide a unified information base for business intelligence.

The core route of building a data warehouse is Extraction-transformation- Loading (ETL). According to TDWI report, 66% of data warehouse performance rely on the success of data warehouse ETL process.

ETL Data Quality Management Tools allows discovery of data quality issues and monitoring the quality measures. This process consumes one third of effort and expenses in the budget of the data warehouse. ETL or data integration processes run between the source and the staging layer, run between the staging layer and the operational data store and potentially run between the operational layer and the performance layer. ETL is a process of finding data, integrating it, and placing it in a data warehouse. To work in an operational environment several quality issues have been seen in an ETL environment. The quality of the information depends on quality attributes[3] of completeness, consistency, conformity, accuracy, pattern and common format , relevance, security and understandibility. To build a DW we must run the ETL tool. ETL tools are a category of Extraction – Transformation – Loading Tools with the job of dealing with data warehouse homogeneity, cleansing, transforming, and loading problems. Poor data quality affects customer satisfaction , economic aspects, and even strategic decisions.

The major market players mainly deals with discovery of quality issues . Moreover ,these are very expensive and have the licensing issues of paid tools which are not affordable by small scale and medium scale enterprises. For our research we present the working of a GPL bases open- source tool-MaSSEETL, for the benefit to SME's and SSE's.

2. RELATED WORK
E. Rahm et al. [14] classify data quality problems that can be addressed by data cleaning routines and provides an overview of the main solution approaches. The article also presents contemporary tool support for data cleaning process.

Muller and Freytag [13] classified quality problems into syntactical anomalies which concern data formats and values for data representation (e.g. lexical errors, domain format errors and irregularities). The authors also discussed the Semantic anomaly and coverage anomaly in context with integrity constraints, contradictions, duplicates and invalid tuples.

Singh and Singh in [9], highlights major quality issues in the field of a data warehouse. The review has collected various issues in data ware house process. The author has classified various causes of data quality data ware house process.

Rahul K. Pandey [1] has tried to gather various sources of data quality problems at various stages of an ETL process. The researcher has classified the problems as "problems at data sources, data profiling problems, staging problems at ETL, problems at data modelling".

Panos Vassiliadis et al.[10] in his research identified generic properties that characterize ETL activities. The researcher provided a taxonomy that characterizes ETL activities in terms of the relationship of their input to their output and the proposed taxonomy that can be used in the construction of larger modules which can be used for the composition and optimization of ETL workflows.

Ahmed Kabiri [6] has highlighted the review of open source and commercial ETL tools, along with some ETL prototypes coming from academic world, the modelling and design works in ETL field, ETL maintenance, review works for optimizing ETL.

K.Srikanth et al. [7] discusses issues related to Slowly Changing Dimensions - SCD type 2 to store entire history in

the dimension table. The implementation has been done on Informatica using employee sample data base.

Jasna Rodi´c et al. [13] have proposed various rules that can be used in data warehouse process. The researchers have generated metadata tables for these tables that store information about the rules. The information about the rules violations is stored to provide analysis of such data. Entire data quality process will be integrated into ETL process in order to achieve load of data warehouse that is as automated, as correct and as quick as possible.

The published work by Singh and Singh [11] substantiates that very diminutive information available on the quality assurance of ETL routines. The researcher suggested the au automated testing in extraction, transformation and loading routines independently.

Chinta et al.[8] provided data cleaning framework to provide robust data quality. The authors have worked upon missing values and dummy values using the "Indiasoft" data set.

Sujatha R.[5] in her research explores designed a framework for non-parametric iterative imputation based mixed kernel estimation in both mixture and clustered data sets. The research has implemented a framework to fill in incomplete instances.

The work by P. Saravanan [2]provided an integrated unit for imputing missing values for the right attribute. The kernel based iterative non-parametric estimators work for both continuous and discrete values.

The research by J. Anitha[4] has covered all the major aspects of ETL usage which can be used to compare and evaluate various ETL tools. The implementation of SCD Type has been done to show comparison.

3. DISCUSSIONS AND OBJECTIVES

The comprehensions from the previous work has given us an idea is various data quality issues in data warehouse environment. The aforementioned issues have been implemented through separate tools . But no single tool has provided a solution to all the above problems at a single place. The data quality issues along with their stages are described below:

Table 1 ETL Quality Issues

Quality Metric	ETL Stage	Scope	Example
Heterogeneous Data Source	Extraction	Integration	Integration of Flat file ,web data, databases, XML databases.
Naming Conflicts	Transformation and Cleaning	Synonyms	Sex/Gender, SID/StudentId /Rollno./ StudId
Structural Conflicts	Transformation and Cleaning	Gender, First Name Middle name Last name / Name/ Fname Lname	(“0”/”1” vs. “F”/”M”) for the Gender field.
Date Formats	Transformation and Cleaning	Various Date Separators and Date Formats	DD-MM-YY/Month,DD YY/ DD/MON/YY/ DATE TIME etc.
Missing Values	Transformation and Cleaning	Value Missing from the Data Set	Fees of the student missing from the data set.
Changing Dimensions	Loading	Versioning of data after every load and update operation.	SCD type 1,2,3

The table describes finding and implementation from various authors through separate tools. Moreover, the frameworks implemented which covers all the issues are implemented through paid tools. So we propose a *MaSSEETL* – an integrated open-source tool to implement the above issues.

4. PROPOSED WORK

The three step ETL process works as follows:

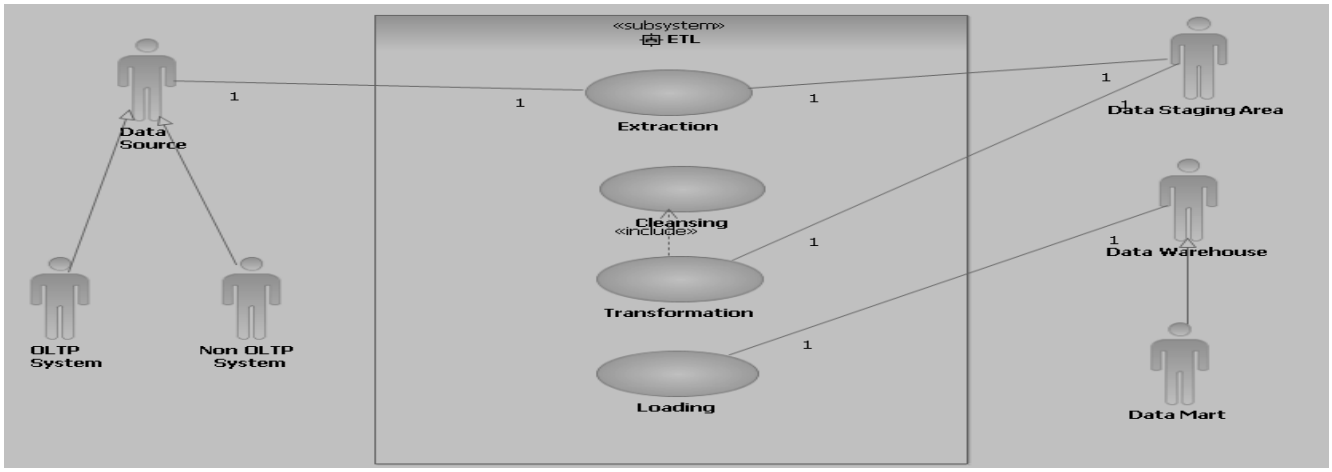

Fig 1 ETL Workflow Use-Case Diagram

EXTRACTION: The amalgamation of all of the disparate systems across the enterprise is the real challenge to getting the data warehouse to a state where it is usable. This step consolidates the data from different data sources. OLTP systems and Non OLTP systems like flat files or spreadsheets are the most common data sources. The main objective of extraction is to retrieve all the required data from the source system with as little resources as possible. It is also known as Data discovery phase. The validated data from extraction is backed up and archived at the staging area.

I. CLEANSING AND TRANSFORMATION: It applies a set of rules to transform the data from the source to the target. This includes converting the measured data to the same dimension using the same units so that they can be later joined.

II. LOADING: Loading data to the target data source structure is the final step in ETL. In this step extracted and transformed data is written into dimensional structures actually is accessed by the end user and application systems. Loading includes both dimensional tables and fact tables.

In a data warehouse, data quality is challenge in an idealized mono-cultural environment, but it is inflamed to epic proportions in a ETL environment. The data quality issues complicate the data warehouse process and hamper the implementation of Data warehouse ETL process in industry.

5. MaSSEETL WORKFLOW

The table below describes the problems that occur while building an MaSSEETL tool.

Table 2 MaSSEETL Quality Issues

	Data Quality Issue	**Problem**
1.	Data Integration Issues	Dealing with php data objects (pdo) in php.
2.	Date formats	Ms- Excel does not date as dd-mon-yy Ms-Access uses standard formats Date/Time My-SQL has format As DD-MON-YY
3.	Generation of source-id	Know the source –id for all the data sources, Ms-Excel does not have any source-id
4.	Generation of surrogate key	Surrogate key for Ms-Excel is difficult to be generated as it does not use any primary key
5.	Filling the missing values	Filling the missing values based on certain criteria.
6.	Domain Checks and conversion	Checking the domain of a particular column and changing the complete data set according to that value e.g. changing the numeric id field to varchar value.
7.	Structural conflicts	Identifying the values of those fields having same structural value , e. g. Gender (0/1) and marital status also having value (0/1) .

Taking into consideration the above issues we propose a MaSSEETL – an integrated ETL tool.

Following Rules can be applied to implement the above quality issues:

Table 3 Rules of MaSSEETL

Rule I	Integration Rule	{Source1(MySQL), Source2(FlatFile), Source3(MsAccess).....} → Sync(MySQL)
Rule II	Surrogate Key Generation	{SourceID1+Pk , SourceID2 +Pk, SourceID3 + Pk.....} → {SurrogateKey1, SurrogateKey2, SurrogateKey3.....}
Rule III	Date Format Mapping	{ DD-MON-YY, DD/MM/YY, Date/Time.....} → {DD/MM/YY}
Rule IV	Domain Conflicts Mapping	{varchar, char, text...} → varchar {date/time , date, varchar} → varchar {int, number, float....} → float {Boolean, varchar, numeric} →Boolean
Rule V	Structural Conflicts Mapping	{FirstName+MiddleName+LastName, Fname+Lname, Name} → {User-Specific Name} {Gender, Sex} → {User - Specific Name}
Rule VI	Missing Value Computation	Mean Value is used to compute the missing value Mode is used to fill the Non –numeric value.
Rule VII	Changing Dimensions	For every update on the data set Changing Dimension Version is added to the reporting data.

Sequence Diagram depicts the workflow of MaSSEETL as follows:

STEP 1: The user selects the data file. Once the file is selected, user can select the fields and the corresponding data types. Then the user can select the name of the column to be displayed in the reporting data.

STEP 2: The database generation of Step 1 is carried out in this step.
This step offers the user to create a merged data set or to update the prevailing data set.

STEP 3: For Create operation: All the cleansing operations are done and Cleansed and transformed data set is given to the end user.
For Update operation:
Version is added to every update operation on the record.

STEP 4:
Log table is maintained to depict the success and failure count of records.

STEP 5:
Report is generated in the form of a CSV File.
MaSSEETL follows the following Sequence Diagram

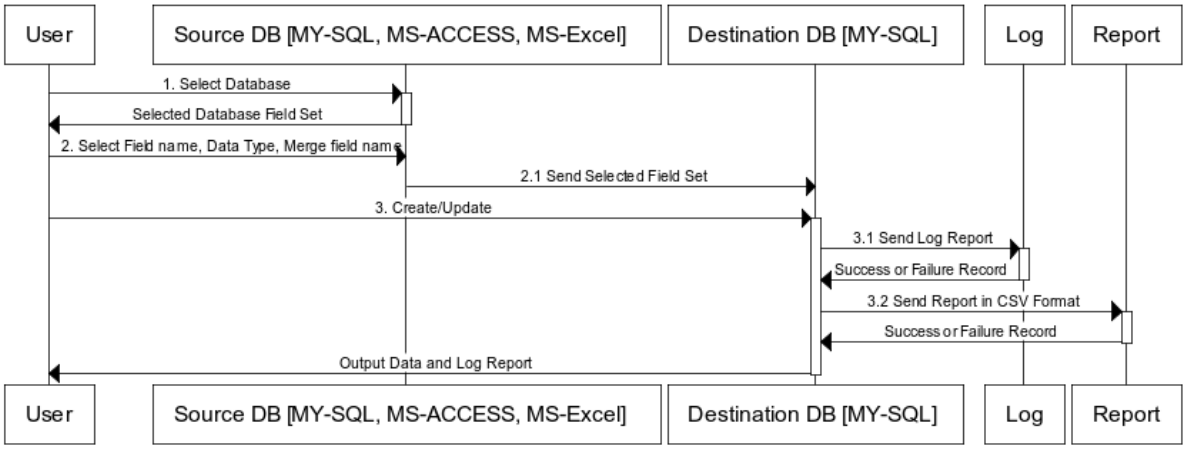

Fig 2 Sequence diagram of MaSSEETL

6. WORKING OF MaSSEETL

For the research, we have taken the data set from various schools of Patiala. The Data Set is represented in the following figures

Sample Data Set

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
ID	doa	regno	admno	name	gender	dob	class	section	rollno	mig	nigdeta	blood	categor	ationali	ranspor	usCardN	BusNo	busstop	fath
41	28-Feb-07	R004	4583	SIMR	Fema	29-Aug-03	Nur	A	1	No		O-		INDIA	Yes	-	4887	URBA	S.
42	06-Mar-07	R004	4584	CHIR	Male	21-Oct-02	Prep-	A	1	No		AB+	GENE	INDIA	Yes	-	4887	INCO	SHR
43	19-Apr-07	R004	4712	RAVJ	Male	15-Nov-90	Xl	C	7	No		O-		INDIA	No				S.
44	24-Mar-07	R004	4585	HARI	Male	23-Oct-02	Nur	A	0	No		AB+		INDIA	Yes	-	2204	BARS	S.G
45	24-Mar-07	R0045	4586	BHUVAN PASSEY	Male	27-Sep-90	XII (Non Med)	A	1	No				INDIA	No				SHRI. RAJES PASSE
46	24-Mar-07	R0046	4587	NEHA SHARMA	Female	03-Sep-90	XII (Med)	B	3	No		O+		INDIA	No				SHRI. J DEV SHARM

Fig 3 Data Set Generation

Stage Wise diagram to implement the MaSSEETL- an integrated ETL tool based on the following flow diagram.

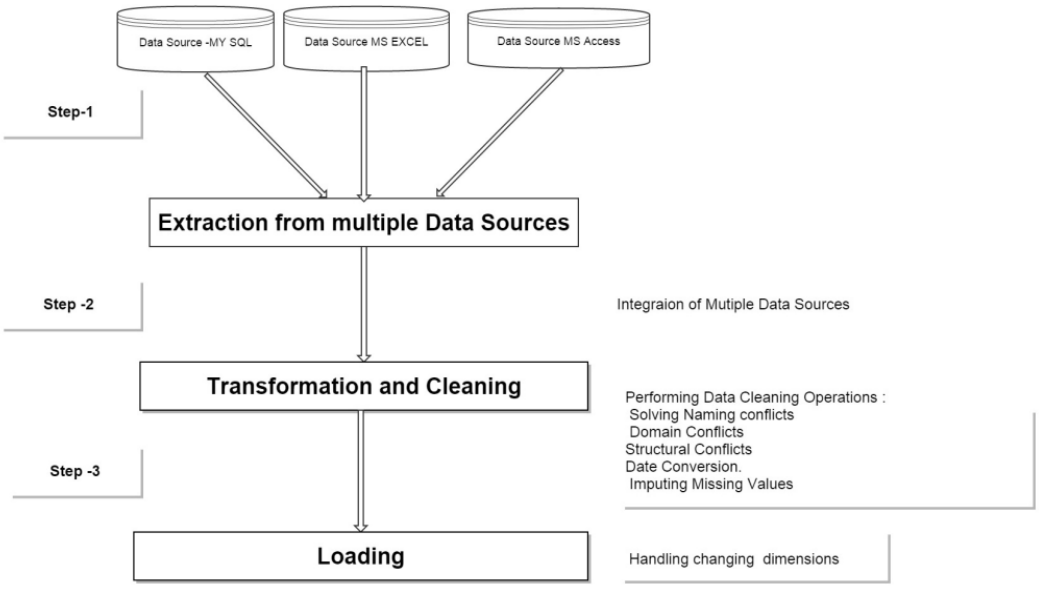

Fig 4 Stage- wise ETL Workflow of MaSSEETL

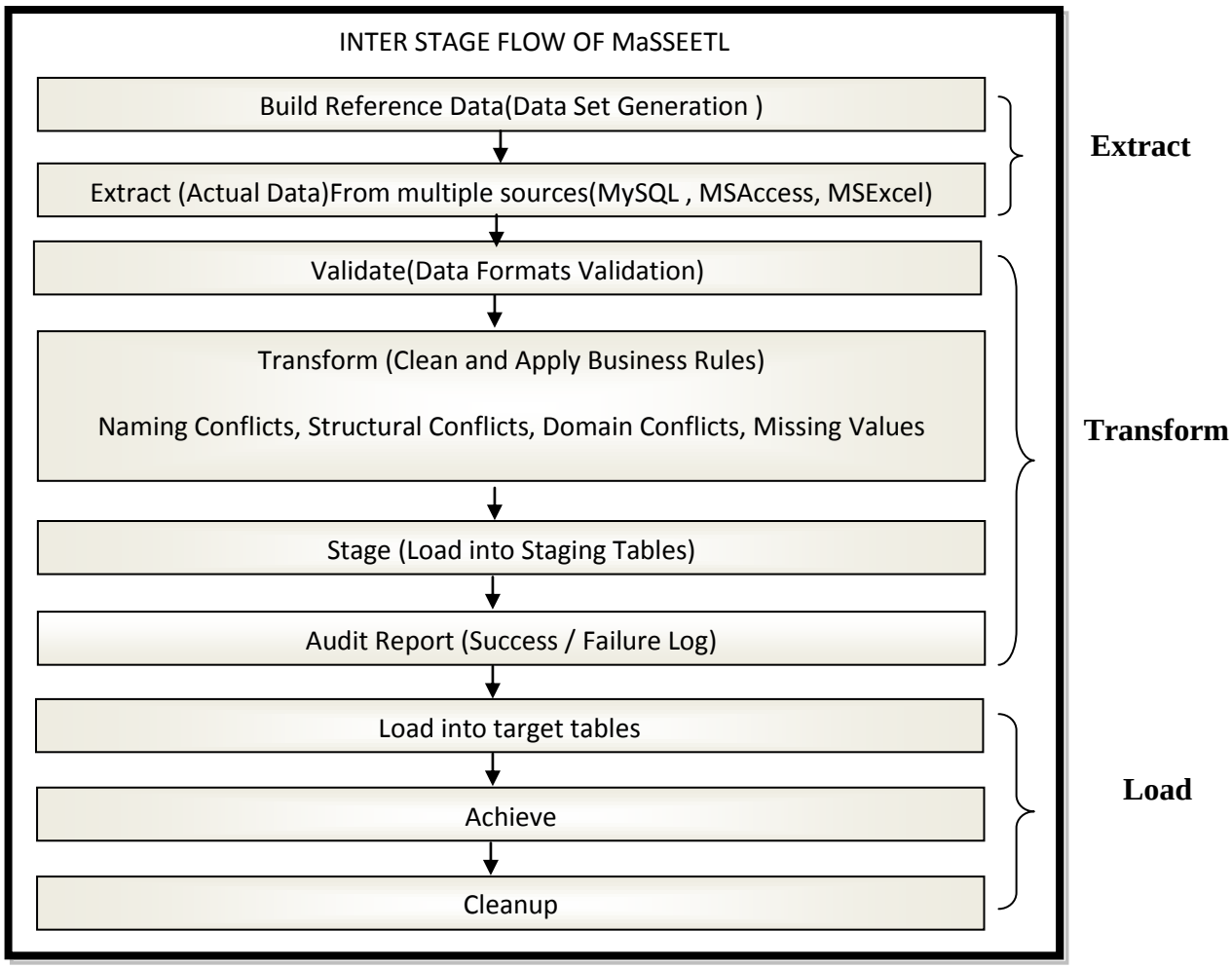

Fig 5 Inter Stage Flow of MaSSEETL

STEP 1:
Three data sets are taken in to consideration : MsAccess, MySQL, MSEXCEL.

STEP 2.
User can select the data source

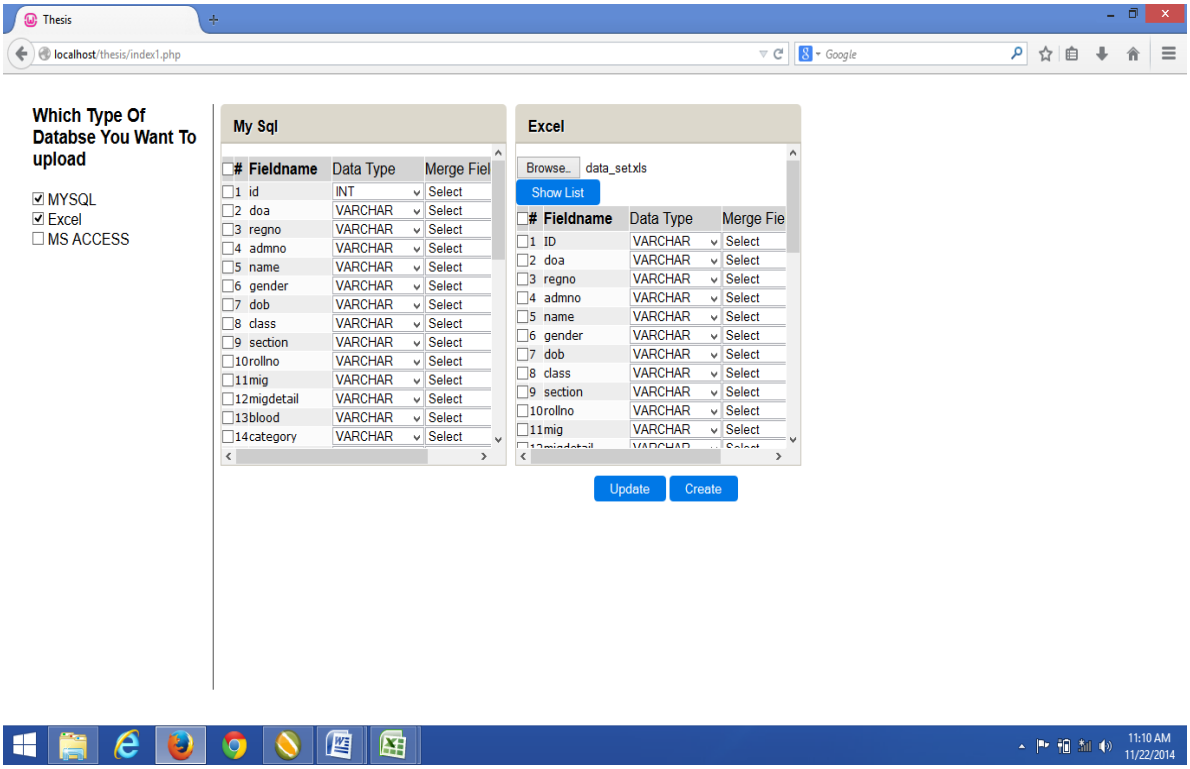

Fig 6 Selection of Data Sources

Step 3.
Select the data set by browsing the data source, perform transformation and log report.

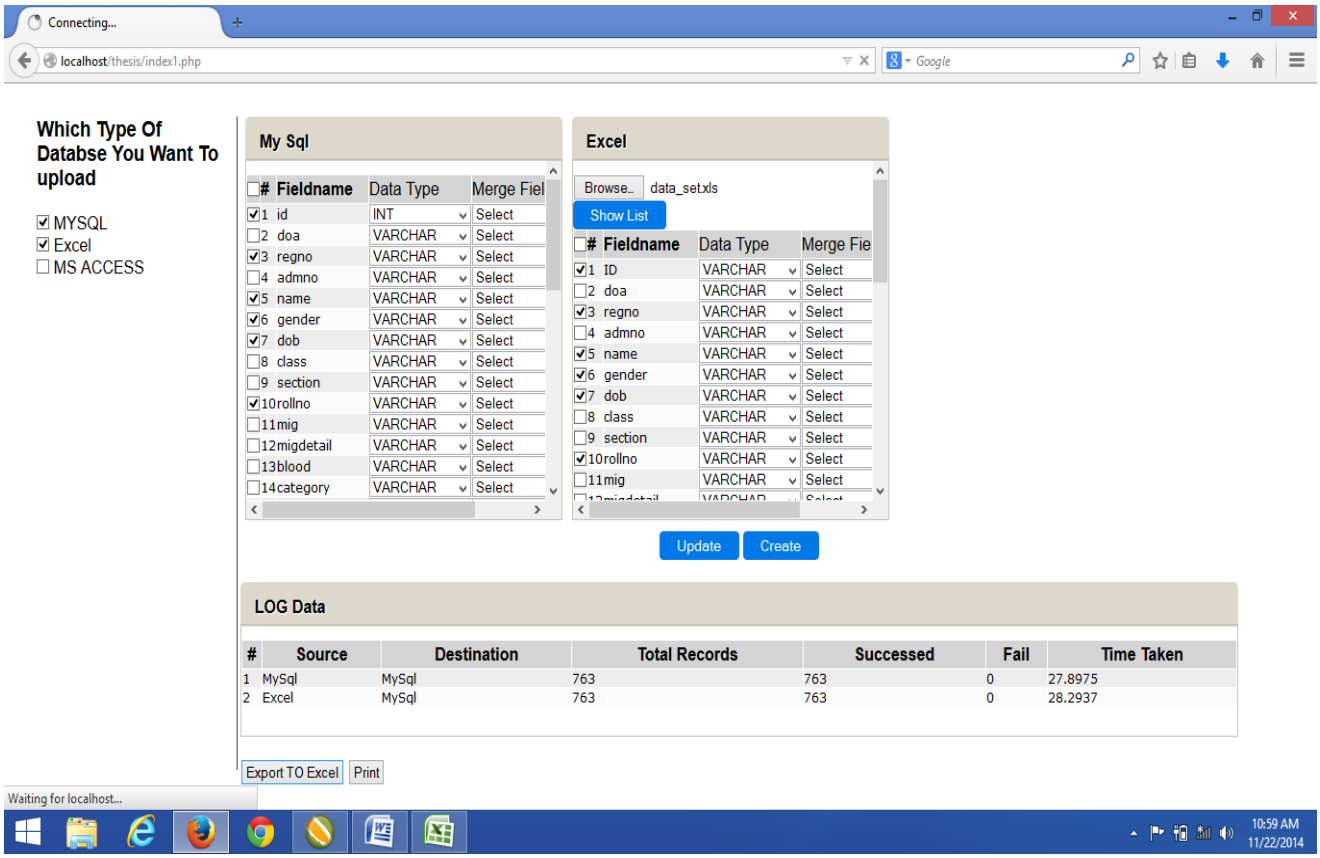

Fig 7 Extraction from Multiple Data Sources

Step 4.
Generate the report by exporting to EXCEL.

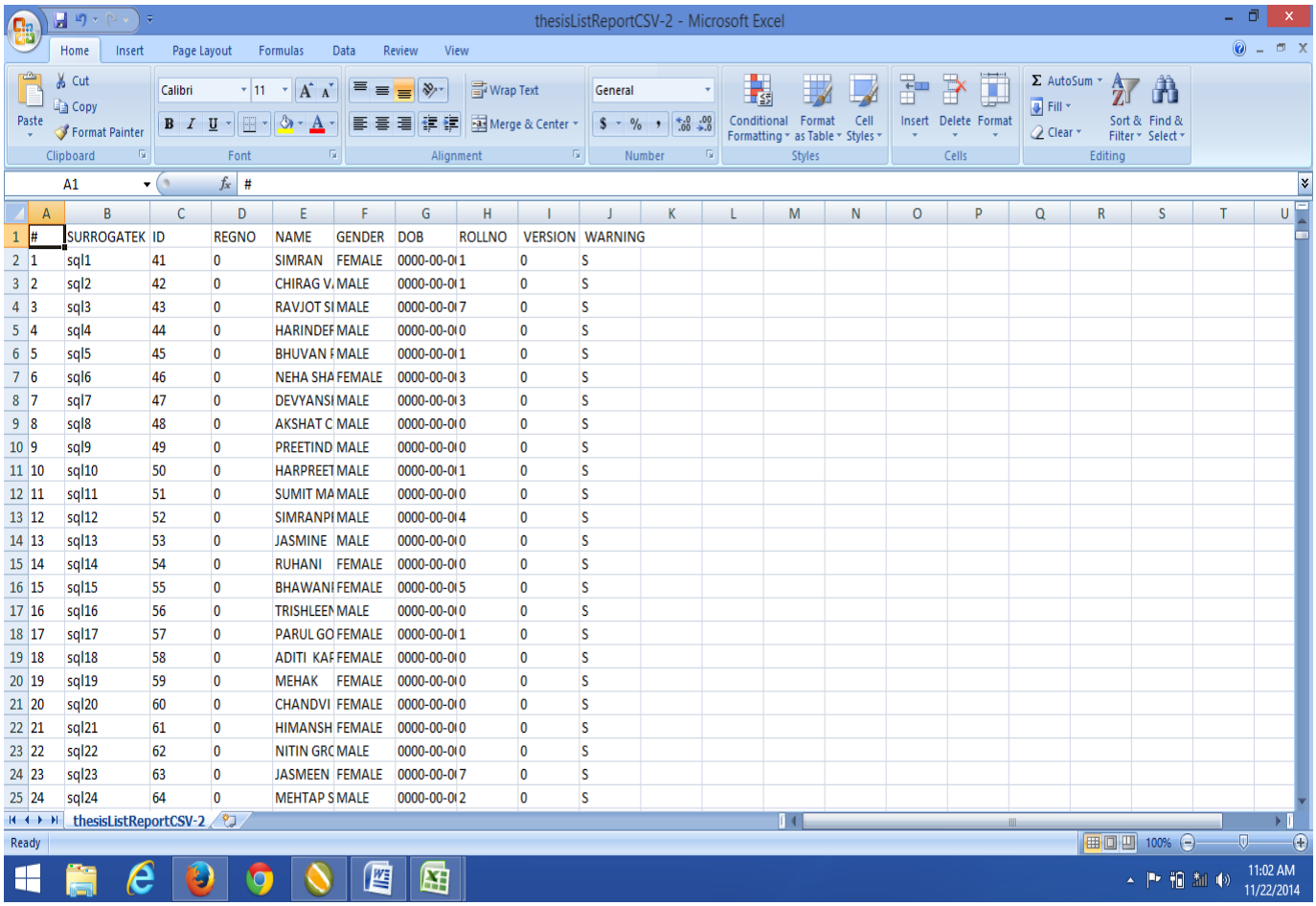

Fig 8 Report Generation

7. RESULTS

1526 records transformed and profiling is shown in figure below:

Detailed profile

Order	State	Time
1	Starting	38 µs
2	Checking Permissions	6 µs
3	Opening Tables	27 µs
4	Init	35 µs
5	System Lock	7 µs
6	Optimizing	4 µs
7	Statistics	13 µs
8	Preparing	10 µs
9	Executing	2 µs
10	Sending Data	329 µs
11	End	3 µs
12	Query End	5 µs
13	Closing Tables	8 µs
14	Freeing Items	143 µs
15	Cleaning Up	12 µs

Summary by state

State	Total Time	% Time	Calls	ø Time
Sending Data	329 µs	51.25%	1	329 µs
Freeing Items	143 µs	22.27%	1	143 µs
Starting	38 µs	5.92%	1	38 µs
Init	35 µs	5.45%	1	35 µs
Opening Tables	27 µs	4.21%	1	27 µs
Statistics	13 µs	2.02%	1	13 µs
Cleaning Up	12 µs	1.87%	1	12 µs
Preparing	10 µs	1.56%	1	10 µs
Closing Tables	8 µs	1.25%	1	8 µs
System Lock	7 µs	1.09%	1	7 µs
Checking Permissions	6 µs	0.93%	1	6 µs
Query End	5 µs	0.78%	1	5 µs
Optimizing	4 µs	0.62%	1	4 µs
End	3 µs	0.47%	1	3 µs
Executing	2 µs	0.31%	1	2 µs

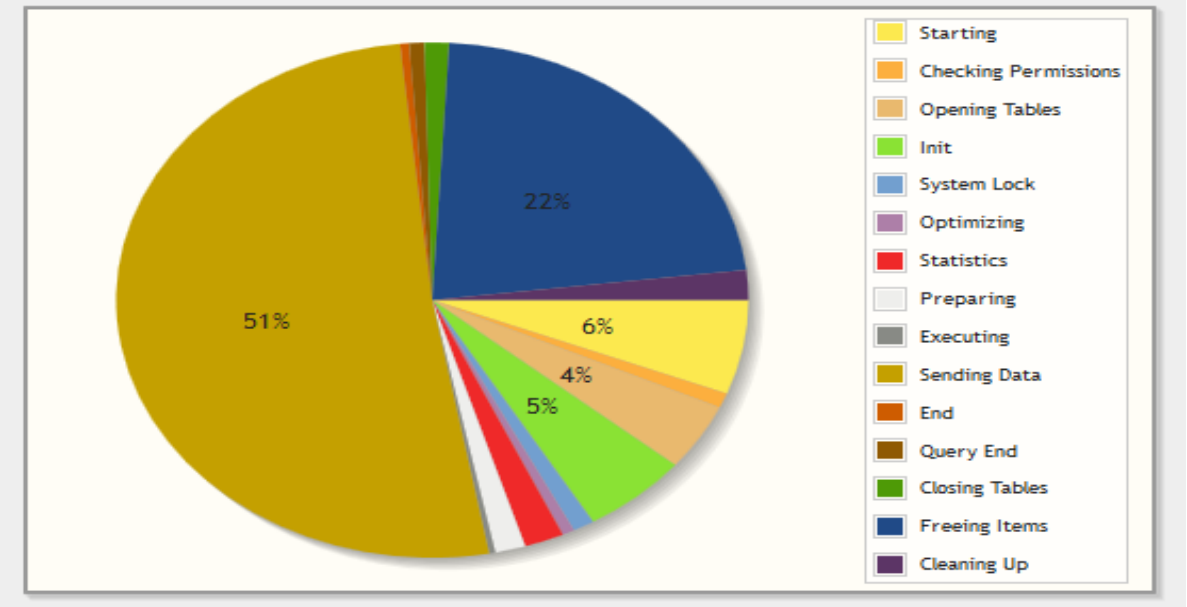

FIg 9 : Transformation Profiling of Output Records

8. CONCLUSION

Data quality has become a major concern activity performed by most organizations that have data warehouses. Every organization needs quality data to improve on its services it renders to its customers. In view of this, a thorough review of approaches and papers in that regard are discussed and their limitations also stated. This is to help future development and research directions in the area of ETL. The papers reviewed in this report looked at critical aspects of data quality, problems and shortcomings in the existing environment.

8.1 Future Work

In future work we propose to implement with data - de-duplication and handling semi structured data in the above tool. The tool performs loading in single destination. In future, multiple data sources can be selected to display the output.

9. REFERENCES

[1] Pandey K. Rahul (2014). Data Quality in Data warehouse: problems and solution.IOSR-Journal of Computer Engineering Volume 16 Issue 1 pp. 18-24.

[2] Saravanan P. (2014) "An Iterative Estimator for Predicting the Heterogeneous Data Sets", Weekly Science Research Journal ISSN: 2321-7871 Volume- 1 Issue -27 pp-1-15'

[3] Choudhary N. (2014) "A Study over Problems and Approaches of Data Cleansing/Cleaning", International Journal of Advanced Research in Computer Science and Software Engineering ISSN: 2277 128X Volume 4 Issue 2 pp- 774-779

[4] Srikanth K.; Murthy N.V.E.S; Anitha J. (2013) " Data Waehousing Concept Using ETL Process For SCD Type-3" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) ISSN: 2276-6856 Vol.2, Issue 5 pp-142-145

[5] Sujatha.R (2013) "Enhancing Iterative Non-Parametric Algorithm for Calculating Missing Values of Heterogeneous Datasets by Clustering" , International Journal of Scientific and Research Publication ISSN: 2250-3153 Volume 3 Issue 3 pp-1-4'

[6] Kabiri A.; Chiadmi D. (2013) "Survey on ETL Processes", Journal of Theoretical and Applied Information Technology. Vol. 54 No.2

[7] Srikanth K.; Murthy N.V.E.S.; Anitha J. (2013) "Data Warehousing Concept Using ETL Process for SCD Type-2", American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN: 2320-0936 Volume-2, Issue-4, pp-86-91' 2013

[8] Rao S. Chinta; Rajanikanth J.; Chandra Sekhar V.; MSVS Bhadri R. (2012) "Data Cleaning Framework for Robust Data Quality in Enterprise Data Warehouse" , IJCST e- ISSN : 0976-8491 p- ISSN : 2229-4333 Vol. 3, Issue 3, pp 36-41

[9] Singh R.; Singh K. (2009). "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing", International Journal of Computer and Electrical Engineering, Vol. 1, No. 4

[10] Vassiliadis P.; Simitsis A.; Baikousi E. (2009) "A Taxonomy of ETL Activities" DOLAP '09 Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP, pp 25-32

[11] Singh J.; Singh K. (2009) "Statistically Analyzing the Impact of Automated ETL Testing on the Data Quality of a Data Warehouse", International Journal of Computer and Electrical Engineering, Vol. 1, No. 4

[12] Rodi´c J.; Baranovi´c M. (2009) "Generating Data Quality Rules and Integration into ETL Process", DOLAP'09 ACM

[13] Muller H.; Freytag J. (2003). "Problems, Methods, and Challenges in Comprehensive Data Cleansing", pp. 21.

[14] Rahm, E.; Do; H.H. (2000). "Data Cleaning: Problems and Current Approaches" IEEE Data Engineering Bull. Vol 23 No. 4, pp. 3-13