ISSN:0975-9646

Enactment of Medium and Small Scale Enterprise ETL(MaSSEETL)-an Open Source Tool

Rupali Gill[1]
Assistant Professor
School of Computer Sciences, CU Punjab

Jaiteg Singh [2]
Associate Professor
School of Computer Sciences, CU Punjab

Abstract -Data quality is major concern area in an Data Warehouse environment. ETL tools focus on detection and correction of data quality problems that affect the success of a data warehouse. Data imported from source 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. Proprietary tools used for data cleaning have a very limited functionality. Small and Medium Scale Enterprises(SME) and Small Scale Enterprises (SSE) cannot afford the licensing cost of these paid tools. The solution to data quality problems is provided by open source data quality tool - MaSSEETL is to deal with naming conflicts, structural conflicts, date conversions, missing values and changing dimensions. This tool solves the integrity issues faced by various available GPL tools. *MaSSEETL* solves the appropriate errors with appropriate level of warning. In this paper, we are presenting the implementation of *MaSSEETL*. The tool provides an increased ease of use in a data warehouse environment.

General Terms -*Data warehousing, data cleansing, quality data, dirty data, surrogate keys*
Keywords: *Data inconsistency, identification of errors, organization growth, ETL, data quality*

1. INTRODUCTION

Extraction-Transformation and Loading (ETL) tools are accountable for the extraction of data from a number of sources, scrubbing , transformation and loading into a data warehouse. According to TDWI report 66% of respondents rely on correctness of data warehouse data for the efficient working of a business organisation. Bill Inmon defines Data warehouse as subject Oriented, Integrated, Time-Variant and non volatile group of data. 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.

This whole process depends on correctness of data warehouse ETL process. ETL and Data Cleaning tools consumes most of the data warehouse resources. ETL is a process of finding data, integrating it, and placing it in a data warehouse. For a successful business organisation, several quality issues quality issues have to be dealt with in an ETL environment. ETL tools are a category of Extraction – Transformation – Loading Tools with the job of dealing with data warehouse homogeneity, cleansing, transforming, and loading problems. The data preparation before their actual loading in the warehouse for further querying is necessary due to quality problems, generation of surrogate keys for uniqueness of data, merging the columns for representation in an standard environment, changing the domains, filling in the missing values maintaining the log report and generating warning is the major concern of all the ETL tools .

None of the open-source and proprietary tools covers the data quality issues of various stages of ETL collectively. The proprietary tools are very expensive. Moreover, the licensing issues of paid tools 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. [13] 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 [12] 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 constrains, contradictions, duplicates and invalid tuples.

Singh and Singh in [8], 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.[9] 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 [5] 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. [6] discusses issues related to Slowly Changing Dimensions - SCD type 2 to store entire history in the dimension table. The implementation has been done in Informatica using employee sample data base.

Jasna Rodi´c et al. [11] 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 [10] 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.[7] 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.[4] 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[3] 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:

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

Table 1 ETL Quality Issues

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. Implementation Of MaSSEETL

The three stage flow chart of MaSSEETL :

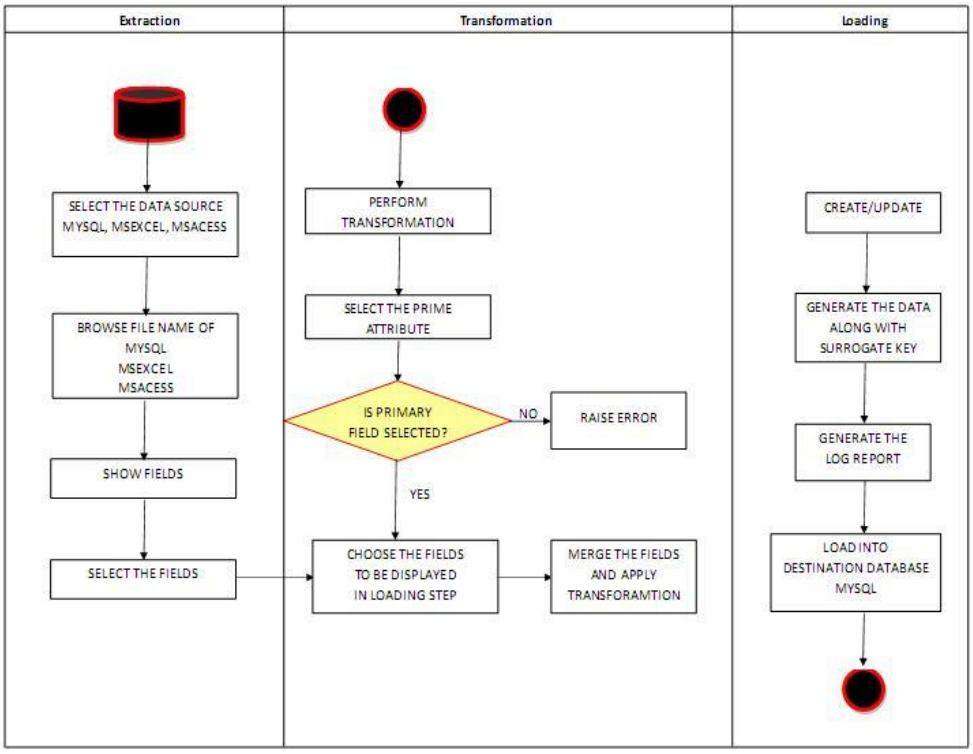

Fig 1 Three Stage Flow Graph of MaSSEETL

I. **EXTRACTION:** In Extraction stage of an MaSSEETL, user selects the fields data sources by clicking on the data sources checkboxes. For the implementation of MaSSEETL , three data sources - **MySQL, MSEXCEL, MSACCESS** are taken as a reference. Then the user browses for the files in the MaSSEETL.
Once the file is chosen user clicks in SHOW FIELDS button to retrieve the list of fields. When the field set if retrieved , user can select the fields to be displayed in the destination database.

II. **CLEANSING AND TRANSFORMATION:** In the transformation phase of MaSSEETL, user first of selects the prime attribute. If it is not selected , an alert message appears on the screen. On selecting , User can change the data type and name of the field to be displayed in the destination database.

III. **LOADING:** Loading data to the target data source structure is the final step in ETL. In this step extracted and transformed data is created by clicking on the CREATE or UPDATE buttons. Log report and loading into **MySQL** database id done in this phase of MaSSEETL.

User has an option to export the report to Excel or print the report. MaSSEETL has been hand-coded to execute all the above mentioned issues taking reference of data sets of **schools in Punjab region**.

5. CHALLENGES WHILE IMPLEMNTING MASSEETL.

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

S No.	Challenge	Problem
1.	Data Integration Issues	Dealing with php data objects (pdo) in php.
2.	Generation of source-id	Know the source –id for all the data sources, Ms-Excel does not have any source-id
3.	Exporting data from MS-Excel	MS Excel does not use any connectivity drivers so connection with MS Excel string was a major problem
4.	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
5.	Generation of surrogate key	Surrogate key for Ms-Excel is difficult to be generated as it does not use any primary key
6.	Filling the missing values	Filling the missing values based on certain criteria.
7.	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.
8.	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) .
9.	MS Excel Date format	Date formats used in MS-Excel are not supported in Databases . Data has to be converted to text format to retrieve the data.
10.	Blank Spaces	Spaces in MS Excel are considered as blank spaces while importing the data into MySQL.
11.	Integrity Constraints	MS Excel does not have integrity constraints so artificial keys have to inserted in order to apply integrity checks.
12.	File Formats	File formats supported for MS Excel is .xls . If the file is in .xlsx data is not extracted.
13.	WAMP Server	WAMP Server supported for the proposed tool is 32-bit. If the 64 bit is used, MS Access file could not be extracted

Table 2 Challenges for Implementing MaSSEETL

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

6. MASSEETL RULES

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

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.....} → {YYYY-MM-DD}
Rule IV	Domain Conflicts Mapping	{varchar| char|text...} → varchar {date/time| date| varchar} → varchar {int| number| float....} → float {Boolean| varchar| numeric} →Boolean (0/1)
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.

Table 3 Rules of MaSSEETL

In this paper , we are giving the details of first the rules .

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

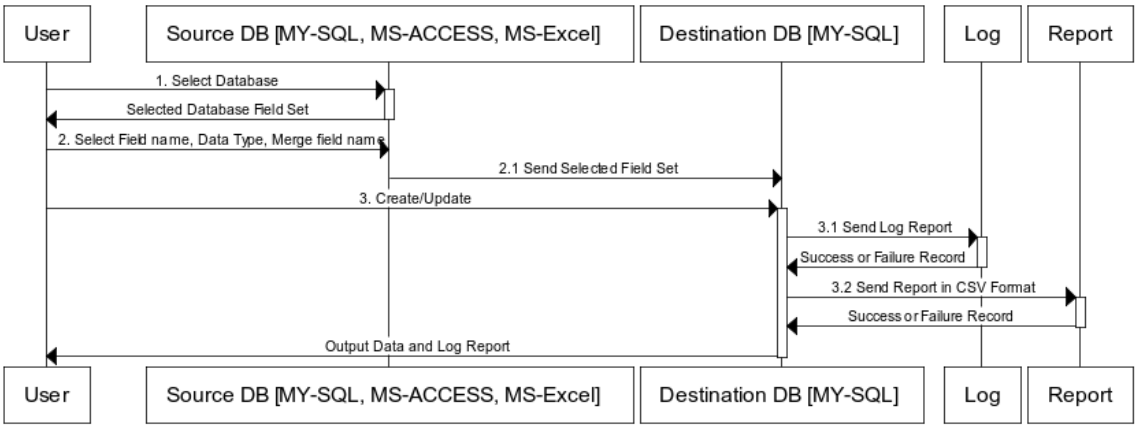

Fig 2 Sequence diagram of MaSSEETL

7. WORKING AND IMPLEMENTATION OF MASSEETL RULES

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

data set

ID	doa	regno	admno	name	gender	dob	class	section	rollno	mig
95	4/2/2007	0	4637	PARAS SINGH AJJEE	Male	7/29/1995	Vith	A	0	No
96	4/2/2007	0	4638	LOVEPREET SINGH	Male	6/13/2004	Nur	A	0	No
97	4/2/2007	0	4639	SANJAY KEWAT	Male	8/17/1996	IIIrd	A	1	No
98	4/3/2007	0	4640	KARANVEER SINGH DHILLON	Male	8/4/1991	XI (Non Med)	A	21	No
99	4/3/2007	0	4641	VARUN BHOLA	Male	2/9/1995	Vith	B	0	No
100	4/3/2007	0	4642	DISHANT SINGH	Male	7/15/1998	Ivth	A	0	No
101	4/3/2007	0	4643	GURCHAIN KAUR	Female	2/26/1997	IIIrd	A	2	No
102	4/3/2007	0	4644	GAGANJOT SINGH	Male	8/6/1999	IIIrd	A	0	No
103	4/3/2007	0	4645	RAMANDEEP SINGH	Male	1/5/1999	IIIrd	A	0	No
104	4/3/2007	0	4646	JASMEEN KAUR	Female	1/23/1998	Vth	A	0	No
105	4/3/2007	0	4647	HARISDEEP SINGH	Male	2/17/2001	Prep-II	A	3	No
106	4/3/2007	0	4648	NAVDEEP KAUR	Female	1/1/1990	XII (Non Med)	A	2	No
107	4/3/2007	0	4651	PIYUSH WALIA	Male	11/9/2004	Nur	A	21	No
108	4/3/2007	0	4649	SUMAN VEER KAUR	Female	6/12/1999	IIND	A	2	No
109	4/3/2007	0	4650	SAHILPREET KAUR	Female	11/3/1997	Vth	A	0	No
110	4/25/2007	0	4652	HARDEV SINGH	Male	10/19/1998	Ivth	A	3	No
111	4/25/2007	0	4653	SUKHPREET SINGH	Male	12/19/2002	Prep-I	A	3	No
112	4/25/2007	0	4654	AMREET NIRWAN	Female	4/2/2004	Nur	A	13	No
113	4/25/2007	0	4655	MEHAKJOT SINGH SANDHU	Male	7/16/2003	Nur	A	0	No
114	4/4/2007	0	4656	SANAM PREET	Female	2/20/2004	Nur	A	0	No
115	4/4/2007	0	4657	RAMANDEEP SINGH	Male	8/11/2001	Ist	A	0	No
116	4/4/2007	0	4658	GURJOT SINGH SIDHU	Male	3/28/2001	Ist	A	6	No
117	4/4/2007	0	4659	GURKIRAT SINGH	Male	9/28/2002	Prep-I	A	4	No
118	4/4/2007	0	4660	SIMRAN KAUR	Female	5/9/2003	Prep-I	A	0	No
119	4/4/2007	0	4662	HARMANDEEP SINGH	Male	7/28/2004	P-Nur	A	3	No
120	4/5/2007	0	4663	SUDEEP SINGH	Male	12/30/1990	XI (Non Med)	A	0	No
121	4/5/2007	0	4664	HARSHPREET SINGH	Male	8/10/1998	Ivth	A	4	No

Record: 1 of 763 No Filter Search

Fig 3 Sample MS Access Data Set

ID	doa	regno	admno	name	gender	dob	class	section	rollno	mig	migdetail	blood	category	nationality	transport	busCardNo	BusNo
41	28-Feb-07	R01	4583	SIMRAN	Female	29-Aug-03	Nur	A	1	No				INDIA	Yes	-	4887
42	06-Mar-07	R02	4584	CHIRAG VAID	Male	21-Oct-02	Prep-I	A	1	No			GENERAL	INDIA	Yes	-	4887
43	19-Apr-07	R03	4712	RAVJOT SINGH	Male	15-Nov-90	XI (Arts)	C	7	No				INDIA	No		
44	24-Mar-07	R04	4585	HARINDER PAL SINGH	Male	23-Oct-02	Nur	A	0	No				INDIA	Yes	-	2204
45	24-Mar-07	R05	4586	BHUVAN PASSEY	Male	27-Sep-90	XII (Non Med)	A	1	No				INDIA	No		
46	24-Mar-07	R06	4587	NEHA SHARMA	Female	03-Sep-90	XII (Med)	B	3	No				INDIA	No		
47	24-Mar-07	R07	4588	DEVYANSHU	Male	07-Nov-03	Nur	A	3	No				INDIA	Yes	-	3944
48	24-Mar-07	R08	4589	CHOUDHARY	Male	01-Jan-04	Nur	A	0	No				INDIA	No		
49	24-Mar-07	R09	4590	PREETINDER SINGH	Male	09-Sep-03	Nur	A	0	No				INDIA	Yes	-	2204
50	24-Mar-07	R10	4591	HARPREET SINGH	Male	09-Oct-04	P-Nur	A	1	No				INDIA	No		

Fig 4 Sample MS Excel Data Set

7.1 Implementation of Rule I : Integration Rule

During the extraction process, data are to be collected from multiple data sources. When this has to be done, different data sources have to be connected to each other. In the present work, we have hand-coded the connection string of multiple data sources at single location where we can view and select multiple columns at the single place. The chosen data sources are: MySQL, MS Excel and MS Access.

Integration Rule states that :

{Source1(MySQL), Source2(FlatFile), Source3(MsAccess).....} → Sync(MySQL)

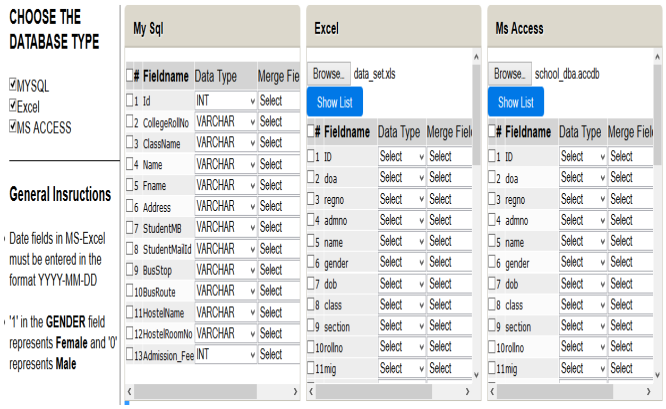

Fig 5 Integration of Multiple Data Sets : Rule I

On selecting the data sources , user have to browse the file of MS Excel and MS Access by clicking the browse button. On selecting the file from upload folder, press Show List button to retrieve the data set.

7.2 Implementation of Rule II: Surrogate Key Generation

In data warehousing, surrogate keys are used to join tables not business or natural keys. Its ETL's job to handle the updates of natural keys which comes into the data warehouse. This is done with lookups which are based on surrogate keys. The whole data warehouse is based on surrogate keys instead of natural keys. When talking about surrogate keys, there are some important points that should be mentioned, Surrogate keys should be,

1. simple integers starting with one and going up to the maximum number needed Surrogate keys should not be,
1. Smart which describes the record
2. Derived from a natural key or combined with natural key
3. Should contain only one column so that there are only one join between two tables.

A unique and common surrogate key is a one-field numeric key which is shorter, easier to maintain and understand, and independent from changes in source system than using a business key. Also, if a surrogate key generation process is implemented correctly, adding a new source system to the data warehouse processing will not require major efforts.

According to above discussion , Surrogate Key generation rule says:

{SourceID1+Pk , SourceID2 +Pk, SourceID3 + Pk.....} → {SurrogateKey1, SurrogateKey2, SurrogateKey3.....}

In this case, we have implemented the surrogate key by a combination of source Id of the data source and the business key of the data source. In case of Ms Excel, first artificial key is inserted as a primary key and then it is combined with the source id of MS Excel.Surrogate keys are generated in continuation.

Following snapshot depicts the following output screen:

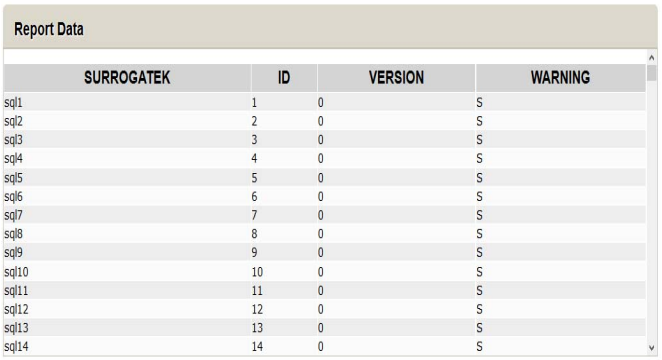

Fig 6 Surrogate Key MySQL :Rule II

Screenshot depicts the surrogate keys generated while various data sources are being selected.

7.3 Implementation of Rule 3 : Date Format Mapping Rule

Managing the date formats in a data ware house is a major issue. There various date formats supported in various databases

For e.g. User may enter the date in any of the following formats

FORMAT	SEPARATOR	DESCRIPTION
DD-MM-YY	-	Date in numeric and hyphen separated
DD-MON-YY	-	Three initial letters of Month in words
DD-MONTH-YYYY	-	Complete Month name and year with hyphen separated
MONTH DD,YY	,	Named month and year are comma separated
DD/MM/YY	/	Date in numeric and slash separated
DD/MON/YYYY	/	Complete Month name and year with slash separated

Table 4 Date Formats

The different sources have different formats. In order to improve the quality of data in a data warehouse a standard format has to be devised. Standard format for the date when the data comes from various sources are defined based on the following rule.

Date Format Mapping Rule states that :

{ DD-MON-YY, DD/MM/YY, Date/Time.....} → {YYYY-MM-DD}

User data source may have entered the date is any of the above mentioned formats but the standard conversion date format will be YYYY-MM-DD that of MySQL.

Also in an MS Excel file user has to specify the date format as YYYY-MM-DD. If the user does not convert the date value into this format a zero value in that column will be

displayed as a part of output from the MS Excel file. An instruction for the same has been written in the *General instruction division.*

Following screen shot depicts the following results :

In the current work we are representing the implementation of first three rule for working on MaSSEETL.

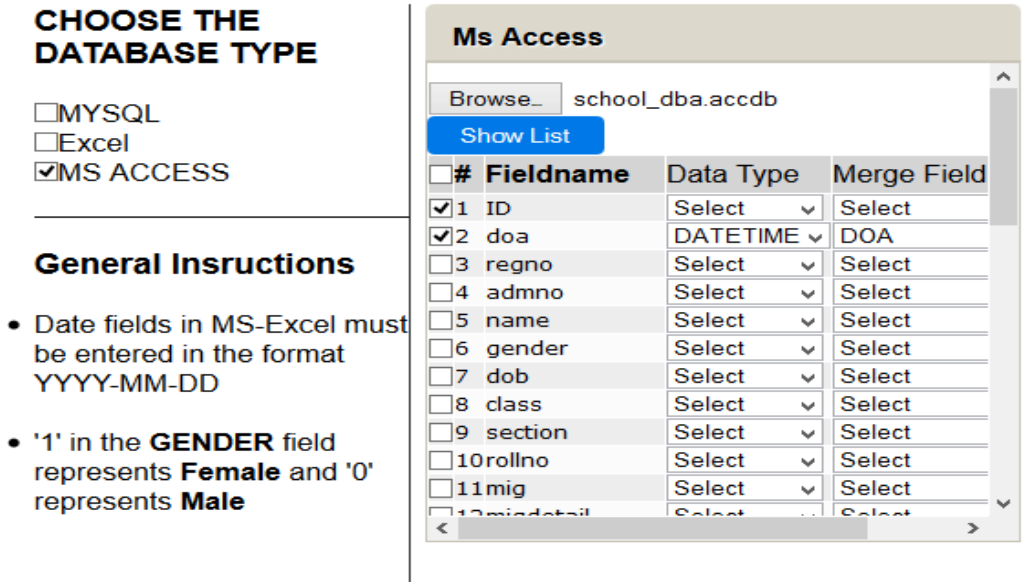

Fig 7 Selection of Date field - Rule III

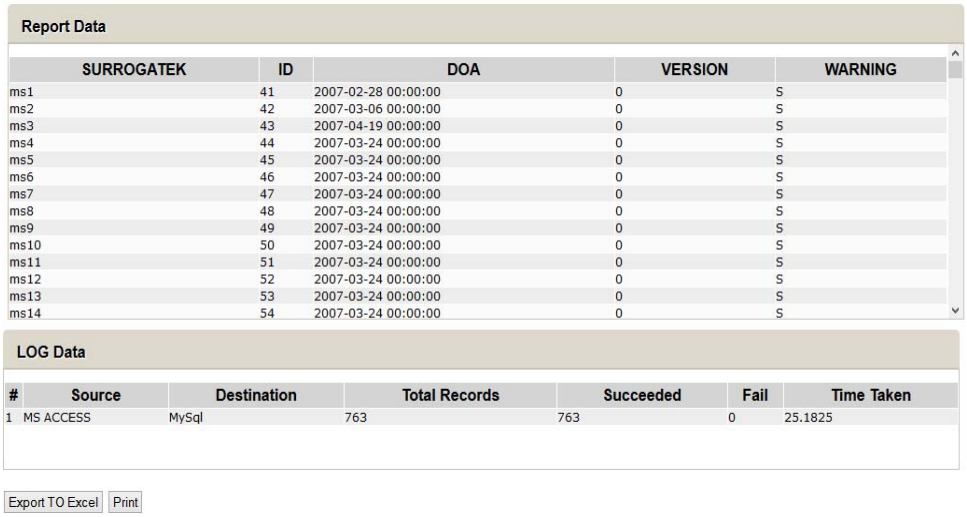

Fig 8 Output of Date Field Selection -Rule III

8. Conclusion And Future Scope

Enterprise needs quality data to improve on its services it renders to its customers. MaSSEETL provides data quality solution to medium and small scale enterprises. In the current work we have implemented only three rules and in future, will provide the implementation of all the above mentioned rules. In future work we propose to implement with data - deduplication and handling semi structured data in the above tool. In current scenario, three data sources are taken as a reference.

In future, other data sources, like Oracle, web data and XML sheets can be taken as input to provide better handling of data.

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] 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

[4] 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’

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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

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

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

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