HDSS Data Cleaning and Integration Using SAS Business Intelligence Tool

Venkatanaveen Dasari [1] and S S Suresh[2]

[1]*Department of Advanced Software and Computing Technology*

[2]*IGNOU-I2IT Centre of Excellence for Advanced Education and Research*

ABSTRACT

HDSS's (Health and Demographic and Surveillance Systems) are typically structured around subjects within the DSA (Demographic Surveillance Area). These subjects have both a conceptual and a logistical rationale. Major subject areas are: birth, death, in-migration, out-migration and delivery. Field workers collect HDSS data in various points of time and feed the data into computer systems manually. Further data analysis takes place. To have a concise view about major elements data cleaning and integration are required. Due to many dimensions and large volume of data, integration is becoming a challenging issue. Researcher performs data cleaning, transformation and integration manually. Which is a time consuming task? The current paper highlights the use of Business intelligence tools (SAS) for HDSS data integration. The experiment starts with Vadu-HDSS site (which is a member of INDEPTH NETWORK, Ghana), Pune, and Maharashtra. The current paper addresses the complexities in the HDSS data Integration, implementation, advantages and results.

Key words: HDSS, DSA, iSHARE, Business Intelligence, Data Integration Studio, Data Repository.

Introduction

Health and Demographic Surveillance System (HDSS) is a set of field and computing operations to handle the longitudinal follow-up of well-defined entities or primary subjects (individual, households and residential units) and all related demographic and health outcomes within a clearly circumscribed geographic area [1]. Demographics are currently statistical characteristics of a population describes the historical changes in population over time. This includes gender, race, age, mobility, employment status and even location [2].

INDEPTH, An international platform of sentinel demographic sites that provides health and demographic data and research to enable developing countries to set health priorities and policies based on longitudinal evidence [3]. The Indepth Network has deployed data sharing system (iSHARE) to increase research capacity.

In the current work, the required data has been collected from iSHARE repository. The main goal of this paper is to describe the need of BI tools for HDSS data integration. Data integration is the process of consolidating data from a variety of sources in order to produce a unified view of the data. Data integration includes Extraction, Transformation and Loading (ETL). ETL is a process periodically extracts data from various source, transforms the data and loads the data into unique format [5].

The project is implemented using SAS/BI platform. Statistical Analysis System (SAS) is an integrated system of software solutions that enables to perform data entry, data retrieval and management, report writing and graphics design, statistical and mathematical analysis, decision support and application development [6]. The SAS system provides the functionality and flexibility that is required to a HDSS data. SAS system provides a number of products where SAS/Data Integration Studio is one among those products. SAS/Data Integration Studio is an integration tool. The SAS/integration studio has effective metadata repository management system (MRMS). The MRMS enables different applications to access the data in a meaningfully way. The MRMS conveys the unique meaning about datasets. This is important for data integration projects.

The current paper describes the following the sections. The section II describes complexities in HDSS data integration. Section III describes Meta data of the dataset under consideration, section IV describes proposed solution and its technical architecture, section V describes implementation and results followed by acknowledgements and conclusions.

Complexities in HDSS data integration

The HDSS data integration gives many advantages to the HDSS users. For example, a HDSS data user can use the integrated data for studying or analyzing cross-data analysis. The integration may involve multiple site data. Each site (a site represents an HDSS data centre) has its own data standards and policies. Hence, integration is a complex task. The complexity increases as the number of sites increases. So, traditional tools and technologies do not work effectively. Hence, SAS is chosen to implement this project. As per our observations, the complexities in the HDSS data integration are as follows:

1. Different regions follow different data standards.
2. Unique format is missing. Hence, data sharing is difficult.
3. Too many errors hamper the result.
4. Since data is Voluminous, integration is time consuming task.
5. Too many rules for data cleaning.
6. Mismatch in Variables.

7. Missing variables
8. Lack of Meta data repository.
9. Different intervals in data collection
10. Difference in geographic cultures
11. Difference in Govt. policies
12. Difference in data standards

Various authors have proposed solutions for data integration. Those are: Balen (2000), Calvanese et al. (2001), Devlin (2003), Erickson (2003), Fox (2003), Holland (2000), McCright (2001), Meehan (2002), Nash (2002), Orovic (2003), Vaughan (2003), Pelletier, Pierre and Hoang (2003) and Whiting (2002) and many others discussed about data integration issues, models and solutions [5,p.no 224].

The next section explains the Meta data of the dataset used in this work.

Meta Data

In this current study the required data is collected through iSHARE site with proper access and permission. The site is available online at [http://www.indepth-ishare.org] and is under maintenance at present. The minimal data sets are as follows [4].

1. Base Table
2. Pregnancy Table
3. Death Table
4. In-Migration Table
5. Out-Migration Table

The iSHARE team has prepared the Meta data for integrating HDSS data of different countries. It represents uniqueness across HDSS users.

The base table (or data set) records the details of basic information about each individual in the population. The meta data is given in the following table 2.1

Table 1: Base Table; Size: 133082

Ordinal Position	Field	Description
1.	PID	Individual identifier. It should be in GUID format.
2.	MID	Mother identifier.
3.	DOB	Date of Birth (YYYY-MM-DD)
4.	SEX	01-Male, 02-Female, 03-Unknown
5.	OCC	Occupation decoded as site descriptions
6.	MS	Marital Status

7.	EDU	Education Status
8.	HHSEXP	Household (geographic location) at start exposure
9.	SEXP	Start of exposure to DSA, Date of first data collection
10.	SOBS	Date of first observed in DSA
11.	EEXP	End of exposure to DSA, Date of last data collection

The pregnancy, death, In-Migration, Out-Migration tables are event tables which explains the details of every individual and brings data into the database. A few key variables are recorded on each event of which a very important one is date of the event. The Meta data of these event tables are as follows:

Table 2: Pregnancy Table; Size: 5746

Ordinal Position	Field	Description
1.	EID	Person identifier
2.	MID	Mother identifier
3.	HID	Household or Geographic identifier of the event
4.	ANC	ANC visits (no. of visits start from "0"/ NULL–not known)
5.	DVD	Date of delivery
6.	OBS	Observation date of delivery
7.	DP	Place of delivery
8.	AB	Attended by
9.	POC	Pregnancy Outcome
10.	BC	Birth Certified/Registered

The death table describes the place and cause of the death. The Meta data of this table are as follows:

Table 3: Death Table; Size: 1317

Ordinal Position	Fields	Description
1.	EID	Event identifier
2.	PID	Person identifier
3.	HID	Household or Geographic identifier of the event
4.	COD	Cause of Death
5.	POD	Place of Death
6.	DOD	Date of Death
7.	OBS	Date of observation of death
8.	DC	Death Certified/Registered

Both In-Migration and Out-Migration tables explain the reason for mobility and records his/her frequency of moving from one place to another place.

Table 4: In-Migration Table; Size: 46578

Ordinal Position	Variable	Description
1.	EID	Event identifier
2.	PID	Person identifier
3.	HID	Household or Geographic identifier of the event
4.	ISI	Is Internal migration(Internal or External migration)
5.	RSM	Reason for this Mobility (IN-Migration)
6.	IMD	Date of In-Migration
7.	OBS	Date of observation of migration

Table 5: Out-Migration Table; Size: 33194

Ordinal Position	Variable	Description
1.	EID	Event identifier
2.	PID	Person identifier
3.	HID	Household or Geographic identifier of the event
4.	ISI	Is Internal migration(Internal or External migration)
5.	RSM	Reason for this Mobility (out-Migration)
6.	OMD	Date of Out-Migration
7.	OBS	Date of observation of migration

System Block Diagram

The figure 1 shows the system block diagram and followed.

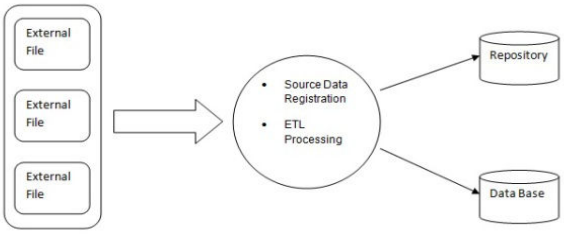


Figure 1: System Block Diagram

The system collects data from external sources and sends the data for processing. The processing involves the data cleaning, transformation and integration.

1. Proposed solution

The following Figure 2 shows the detailed technical architecture of the project.

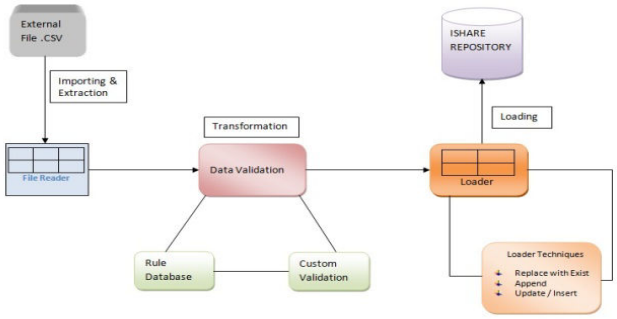


Figure 2 :Technical Architecture

Before performing integration tasks of ETL, registration of source data and target table into Meta data repository is compulsory. This would facilitate unique accessing. The

Implementation: The implementation involves the sequence of steps. The details of these steps are as follows.

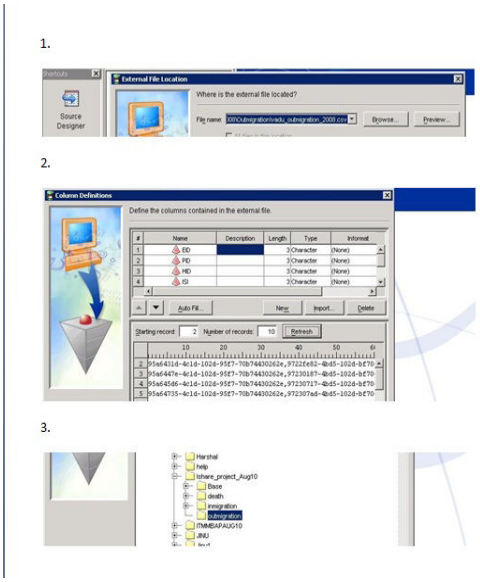


Figure 3: Source Registration

Using Source Designer tool from desktop window of SAS/DIS external file .CSV is imported into Meta data repository by defining all the variables and specifying the location path to save source data set.

Target Registration

The following Figure 4 shows the registration processes of target table. Target table is been created using the target designer tool from desktop window of SAS/DIS by selecting the library and variables of the source table for unique data processing.

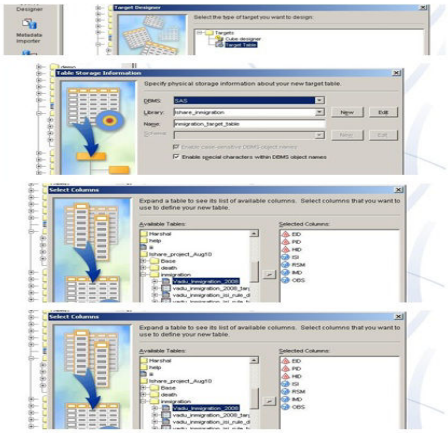


Figure 4: Target Registration

The implementation contains JOBS. Each job is process. Which takes data, processes data, validate data and store the data into the Staging area or repository. The following Figure 5 shows the execution flow of technical architecture in detailed.

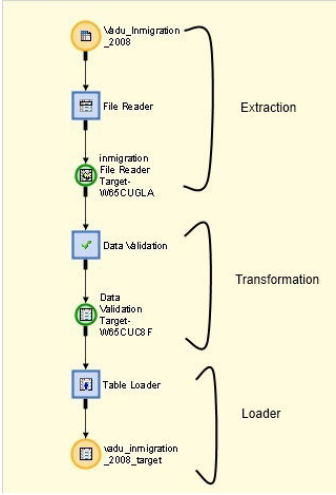


Figure 5: Implementation

Extraction and Cleaning of Data

File reader in Figure 5 is a tool which reads an external file and writes it to the temporary file upon which all the validation rules are implemented. This is connected automatically to a process flow when an external file is specified as a source.

Transformation and Validation of Data

The following Figure 6 shows the transformation and validation of data.

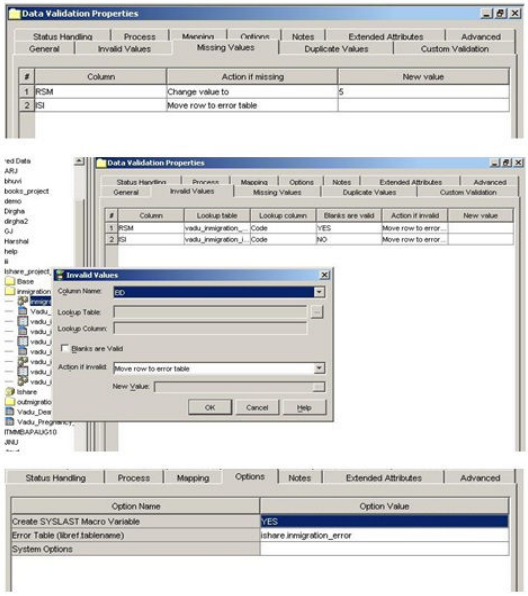


Figure 6: Transformation

All the rules and constraints are stored in a rule database. The database constraints are of 30 rules. While processing the validation against the data, these rules are applied from database when we are performing integrity checks.

1) Data Validation transformation node is applied on the temporary target table.

2) The following validations are performed on the RSM and ISI columns:

1. Invalid values: The NULL values in RSM and ISI column have been considered as invalid and moved to Error table.
2. Missing values: The Missing values in RSM and ISI column have been replaced by Metadata code=NULL as specified is HARE Metadata codes.
3. Error table: The Error table i.e. error_inmigration_table is created and saved at the repository location.

3) The ETL process separates invalid data from valid data and stores in an Error table. The Error data would be sending for sites for further review. Valid data stores in a target table.

5.1.5 Load Strategy:

The following Figure 7 shows the data load strategy. SAS gives various options for data loading. The options are: Replace, Append to existing and Update or Insert. At present, Replace strategy is chosen.

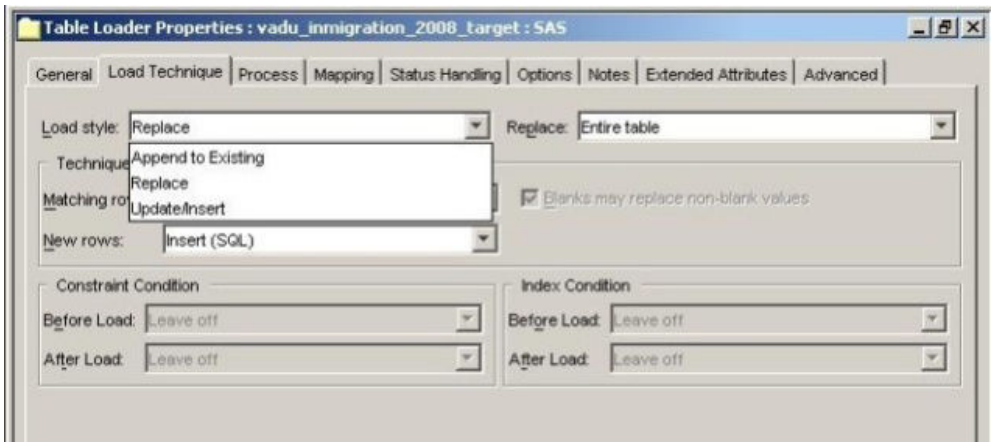
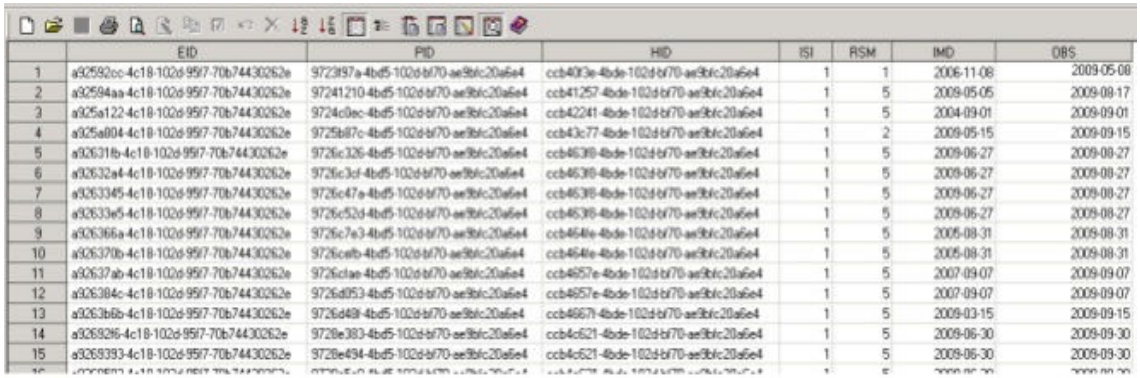


Figure 7: Loading Strategy

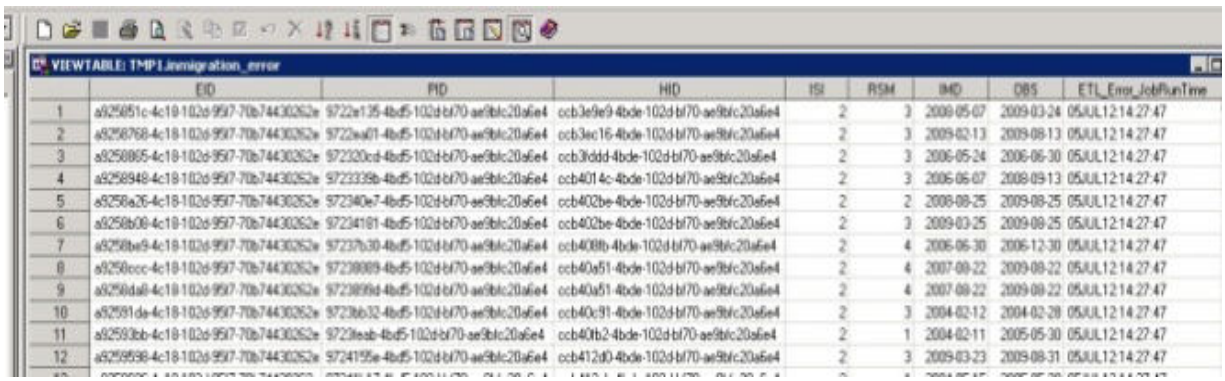
Results

After the load process, the target table contains the valid data. The following fig shows the table data after load and subsequently error table.



	EID	PID	HID	ISI	RSM	IMD	OBS
1	a92592cc-4c18-102d-9597-70b74430262e	9723f97a-4bd5-102d-bf70-ae9bfc20a6e4	ccb403e-4bde-102d-bf70-ae9bfc20a6e4	1	1	2006-11-08	2009-05-08
2	a92594aa-4c18-102d-9597-70b74430262e	97241210-4bd5-102d-bf70-ae9bfc20a6e4	ccb41257-4bde-102d-bf70-ae9bfc20a6e4	1	5	2009-05-05	2009-08-17
3	a925a122-4c18-102d-9597-70b74430262e	9724c0ec-4bd5-102d-bf70-ae9bfc20a6e4	ccb42241-4bde-102d-bf70-ae9bfc20a6e4	1	5	2004-09-01	2009-09-01
4	a925a804-4c18-102d-9597-70b74430262e	9725b87c-4bd5-102d-bf70-ae9bfc20a6e4	ccb43c77-4bde-102d-bf70-ae9bfc20a6e4	1	2	2009-05-15	2009-09-15
5	a92631fb-4c18-102d-9597-70b74430262e	9726c326-4bd5-102d-bf70-ae9bfc20a6e4	ccb4638b-4bde-102d-bf70-ae9bfc20a6e4	1	5	2009-06-27	2009-08-27
6	a92632a4-4c18-102d-9597-70b74430262e	9726c3cf-4bd5-102d-bf70-ae9bfc20a6e4	ccb4638f-4bde-102d-bf70-ae9bfc20a6e4	1	5	2009-06-27	2009-08-27
7	a9263345-4c18-102d-9597-70b74430262e	9726c47a-4bd5-102d-bf70-ae9bfc20a6e4	ccb46393-4bde-102d-bf70-ae9bfc20a6e4	1	5	2009-06-27	2009-08-27
8	a92633e5-4c18-102d-9597-70b74430262e	9726c52d-4bd5-102d-bf70-ae9bfc20a6e4	ccb46398-4bde-102d-bf70-ae9bfc20a6e4	1	5	2009-06-27	2009-08-27
9	a92636a-4c18-102d-9597-70b74430262e	9726c7e3-4bd5-102d-bf70-ae9bfc20a6e4	ccb4648e-4bde-102d-bf70-ae9bfc20a6e4	1	5	2005-08-31	2009-08-31
10	a926370b-4c18-102d-9597-70b74430262e	9726cefb-4bd5-102d-bf70-ae9bfc20a6e4	ccb4648e-4bde-102d-bf70-ae9bfc20a6e4	1	5	2005-08-31	2009-08-31
11	a92637ab-4c18-102d-9597-70b74430262e	9726cfae-4bd5-102d-bf70-ae9bfc20a6e4	ccb4657e-4bde-102d-bf70-ae9bfc20a6e4	1	5	2007-09-07	2009-09-07
12	a926384c-4c18-102d-9597-70b74430262e	9726d053-4bd5-102d-bf70-ae9bfc20a6e4	ccb4657e-4bde-102d-bf70-ae9bfc20a6e4	1	5	2007-09-07	2009-09-07
13	a9263b6b-4c18-102d-9597-70b74430262e	9726d49f-4bd5-102d-bf70-ae9bfc20a6e4	ccb4667f-4bde-102d-bf70-ae9bfc20a6e4	1	5	2009-03-15	2009-09-15
14	a926926-4c18-102d-9597-70b74430262e	9726e383-4bd5-102d-bf70-ae9bfc20a6e4	ccb4c621-4bde-102d-bf70-ae9bfc20a6e4	1	5	2009-06-30	2009-09-30
15	a9269393-4c18-102d-9597-70b74430262e	9726e494-4bd5-102d-bf70-ae9bfc20a6e4	ccb4c621-4bde-102d-bf70-ae9bfc20a6e4	1	5	2009-06-30	2009-09-30

Figure 8: Valid Data



VIEWTABLE: TMP1.inmigration_error

	EID	PID	HID	ISI	RSM	IMD	OBS	ETL_Error_JobRunTime
1	a925851c-4c19-102d-9597-70b74430262e	9722e135-4bd5-102d-bf70-ae9bfc20a6e4	ccb3e9e9-4bde-102d-bf70-ae9bfc20a6e4	2	3	2009-05-07	2009-03-24	05JUL12:14:27:47
2	a9258768-4c18-102d-9597-70b74430262e	9722ea01-4bd5-102d-bf70-ae9bfc20a6e4	ccb3ec16-4bde-102d-bf70-ae9bfc20a6e4	2	3	2009-02-13	2009-08-13	05JUL12:14:27:47
3	a9258865-4c18-102d-9597-70b74430262e	97232ccd-4bd5-102d-bf70-ae9bfc20a6e4	ccb3fddd-4bde-102d-bf70-ae9bfc20a6e4	2	3	2006-05-24	2006-06-30	05JUL12:14:27:47
4	a9258948-4c19-102d-9597-70b74430262e	9723339b-4bd5-102d-bf70-ae9bfc20a6e4	ccb4014c-4bde-102d-bf70-ae9bfc20a6e4	2	3	2006-06-07	2009-09-13	05JUL12:14:27:47
5	a9258a26-4c18-102d-9597-70b74430262e	972340e7-4bd5-102d-bf70-ae9bfc20a6e4	ccb402be-4bde-102d-bf70-ae9bfc20a6e4	2	2	2008-08-25	2009-08-25	05JUL12:14:27:47
6	a9258b08-4c19-102d-9597-70b74430262e	97234181-4bd5-102d-bf70-ae9bfc20a6e4	ccb402be-4bde-102d-bf70-ae9bfc20a6e4	2	3	2009-03-25	2009-09-25	05JUL12:14:27:47
7	a9258be9-4c18-102d-9597-70b74430262e	97237b30-4bd5-102d-bf70-ae9bfc20a6e4	ccb408fb-4bde-102d-bf70-ae9bfc20a6e4	2	4	2006-06-30	2006-12-30	05JUL12:14:27:47
8	a9258ccc-4c18-102d-9597-70b74430262e	97238089-4bd5-102d-bf70-ae9bfc20a6e4	ccb40a51-4bde-102d-bf70-ae9bfc20a6e4	2	4	2007-09-22	2009-09-22	05JUL12:14:27:47
9	a9258da8-4c19-102d-9597-70b74430262e	97238f9d-4bd5-102d-bf70-ae9bfc20a6e4	ccb40a51-4bde-102d-bf70-ae9bfc20a6e4	2	4	2007-09-22	2009-09-22	05JUL12:14:27:47
10	a92591de-4c18-102d-9597-70b74430262e	97239b32-4bd5-102d-bf70-ae9bfc20a6e4	ccb40c91-4bde-102d-bf70-ae9bfc20a6e4	2	3	2004-02-12	2004-02-28	05JUL12:14:27:47
11	a92593bb-4c18-102d-9597-70b74430262e	9723feab-4bd5-102d-bf70-ae9bfc20a6e4	ccb40fb2-4bde-102d-bf70-ae9bfc20a6e4	2	1	2004-02-11	2005-05-30	05JUL12:14:27:47
12	a9259598-4c18-102d-9597-70b74430262e	9724195e-4bd5-102d-bf70-ae9bfc20a6e4	ccb412d0-4bde-102d-bf70-ae9bfc20a6e4	2	3	2009-03-23	2009-08-31	05JUL12:14:27:47

Figure 9: Error Data

Advantages

The following list shows the uses of BI tools for data cleaning and transformation:

1. Developers use PL/SQL stored procedures for ETL operations. This is obsolete now. SAS/Integration Studio gives quick solution with minimal effort. It has effective inbuilt tools developed using Artificial Intelligence and Sql technologies.
2. Development time is minimal
3. Requirement changes can be done easily.
4. Efficient rule management
5. Less coding
6. Customization is easy
7. Integration with other ETL tools or applications becomes easy.
8. Due to efficient indexing techniques, SAS tools can process large volume of data.

Conclusion

The system which was generated now can process raw data from any individual HDSS sites, present the project contains five JOBS. All the jobs are successfully tested and recorded their outputs. The running time for all the jobs executed one after another is 4 seconds. At present job scheduling is not performed. If scheduling is performed, parallel execution takes place. This would reduce the execution time further. In the next phase, we have planned to implement scheduling to run the jobs eventually. This project work fulfils my master's degree credits.

References

Ponniah P, (2001), Data ware Housing Fundamentals: A comprehensive Guide for IT Professionals, John Wiley and Sons.

Efraim turban, jay E Aronson, and Ting-Peng Liang, Decision Support Systems and Intelligent Systems, Pearson.

iSHARE [Online] Available at http://www.indepth-ishare.org.

http://en.wikipedia.org/wiki/Demographics, Accessed on 10th January 2013.

Indepth Network. [Online] Available at,http://www.indepth-network.org, Accessed on 10th January 2013

(2010),iSHARE Site Data Set Specifications,6.2.

SAS| Business Analytics and Business Intelligence, [online] available at www.sas.com.