Discovery of Knowledge by using Data Warehousing as well as ETL Processing

Arif Ali Wani, Bansi Lal Raina

Abstract*: Testing is very essential in Data warehouse systems for decision making because the accuracy, validation and correctness of data depends on it. By looking to the characteristics and complexity of Data warehouse, in this paper, we have tried to show the scope of automated testing in assuring best data warehouse solutions. Firstly, we developed a data set generator for creating synthetic but near to real data; then in synthesized data, with the help of hand coded Extraction, Transformation and Loading (ETL) routine, anomalies are classified. For the quality assurance of data for a Data warehouse and to give the idea of how important the Extraction, Transformation and Loading is, some very important test cases were identified. After that, to ensure the quality of data, the procedures of automated testing were embedded in hand coded ETL routine. Statistical analysis was done and it revealed a big enhancement in the quality of data with the procedures of automated testing. It enhances the fact that automated testing gives promising results in the data warehouse quality. For effective and easy maintenance of distributed data, a novel architecture was proposed. Although the desired result of this research is achieved successfully and the objectives are promising, but still there's a need to validate the results with the real life environment, as this research was done in simulated environment, which may not always give the desired results in real life environment. Hence, the overall potential of the proposed architecture can be seen until it is deployed to manage the real data which is distributed globally.

***Index Terms*: Data Quality, Data warehousing, ETL and Testing.**

I. INTRODUCTION

As we talk about the data framework, from 1970 onwards for decision making operational data is being used in the organizations to increase the utilization of this data. It includes the application to work upon for decision making purposes which can make an investigation regarding data recorded for recognizing the patterns, and make it ready for the execution. The data framework here imply to data distribution. Data warehousing is mainly picking up data from overall sources instead of one source. The organizations identify the distribution of data as an important process to merge the data from various diverse sources. This data sources may be data from inside or outside the company or organizations using OLAP. Latest and Best corporate houses present in India using the Technologies regarding information and communication are Asiana Paints Ltd.,TELCO., The National Stock Exchange of India (NSE)., Godrej Consumer Products Limited., ICICI Bank.

There are some patters used in the data utilization and distribution which are growing rapidly in databases of DW. It is not sure that these clients we have are always happy about the information or data utilized or distributed during the investigation of the data. We need to know about the market that is developing to match with as to build useful and best framework advancements. This data framework can increase the data quality, execution and efficiency to increase the utilization of data and improve its resources through which we get the information.

The literature survey of my research has been done into three parts. Initially the theoretical vision about data warehouse development has been presented, which includes issues associated with data warehouse design, deployment and data warehouse architecture proposed by various researchers. Secondly literature concerned with generating synthetic test data data warehousing has been reviewed[1]. Subsequent literature review is concerned about understanding ETL, Data Quality and Data Management.

After that, to ensure the quality of data, the procedures of automated testing were embedded in hand coded ETL routine. Finally the statistical analysis was done and it revealed a big enhancement in the quality of data with the procedures of automated testing. It enhances the fact that automated testing gives promising results in the data warehouse quality. For effective and easy maintenance of distributed data, a novel architecture was proposed.

II. RELATED WORK

A. Synthetic Test Data Generator

Huge investment is done by business houses on the generation of new databases in order to gain the competitive edge. A data warehouse require testing like any software project development. The complexity and size of data warehouse systems make comprehensive testing both “more difficult and more necessary”.The fact, queries that perform satisfactorily on small datasets may fail miserably in the real life environment. This necessitates establishing a system that runs queries on fully scaled data. Legal implications and business ethics do not allow performing testing with real business data. Hence this research has made efforts to generate synthetic test data ensuring correct balance and skew making sure that the

Revised Manuscript Received on July 22, 2019.

Arif Ali Wani, Computer Science and Engineering Department, Glocal University, Saharanpur, India.

Bansi Lal Raina, Computer Science and Engineering Department, Glocal University, Saharanpur, India.

ratio of fact to dimension is correct and so on [2].

B. The Data Set Generator

Taking under consideration the above stated and because of non-versatile behavior of ready to use synthetic data generators it was decided to develop a dataset generator (DSG) that can generate authentic personal records database while preserving integrity constraints. The architecture includes two sections. In the initial section desired fact table is generated by including relevant facts from multiple sources. In second section, dataset is synthetically puffed-up whereas protective integrity, bias, correct balance and skew, fact to dimension ratio and so on. Figure 1, illustrates projected design of the DSG.

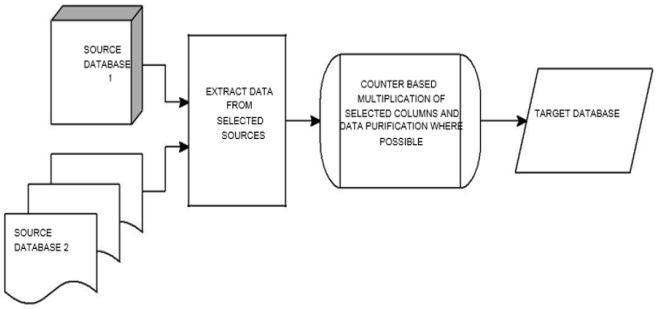
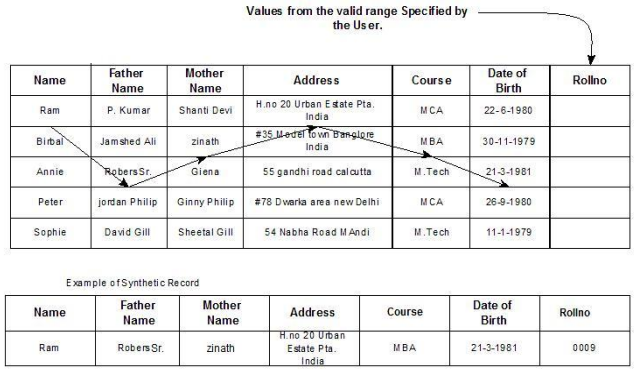


Figure 1. Synthetic Data Set Generator Architecture

In order to generate synthetic records a data generator uses data multiplication algorithm Figure 2, depicts the working of Data Multiplication Algorithm (DMA) used to puff up facts table by genetically mutating different available fields while preserving their integrity[3].

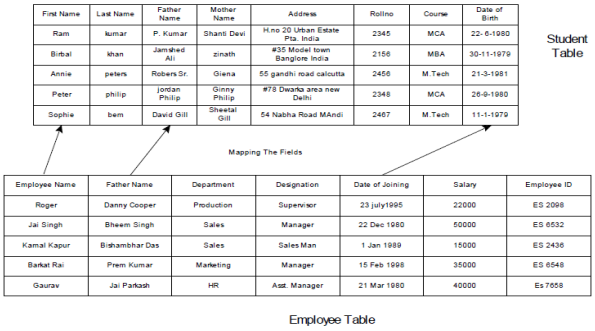


Name	Father Name	Mother Name	Address	Course	Date of Birth	RollNo
Ram	P. Kumar	Shanti Devi	H.no 20 Urban Estate Pta. India	MCA	22-6-1980	
Birbal	Jamshed Ali	zinath	#35 Model town Banglore India	MBA	30-11-1979	
Annie	RobersSr.	Giena	55 gandhi road calcutta	M.Tech	21-3-1981	
Peter	jordan Philip	Ginny Philip	#78 Dwarka area new Delhi	MCA	26-9-1980	
Sophie	David Gill	Sheetal Gill	54 Nabha Road MAndi	M.Tech	11-1-1979	

Name	Father Name	Mother Name	Address	Course	Date of Birth	RollNo
Ram	RobersSr.	zinath	H.no 20 Urban Estate Pta. India	MBA	21-3-1981	0009

Figure 2. The Multiplication logic of DMA Algorithm

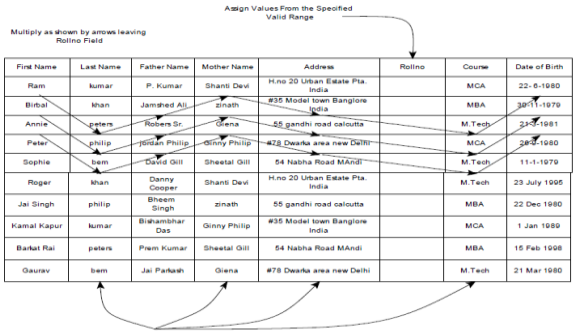
Example in figure 2 shows, the Name columns first instance is multiplied with third instance of Father Name column and second instance of Mother Name column than the first instance of address field is multiplied with the second and third instance of course and data of birth column respectively to generate the first record given in the new record example. The process of puffing up dimension table can be extended by crossover multiplication of even new synthetic data sets generated till a sizable dimension table is obtained. Knowingly multiplicative crossover was preferred over the randomized crossover so as to avoid auto correlation and to ensure more independence and uniformity of mimicking data sets through data set generator[4]. Unlike traditional test data generators the synthetic data set generator developed during this research is capable of exploiting resembling domain sets for different attributes at different locations. . As shown in figure 2, the employee ame and father name columns of employee table are mapped with first name and father name columns of student table respectively. Similarly, date of joining can be mapped with date of birth field. As depicted in figure 3, the data set generator can club the resembling domain sets from two sources in a table placed on the virtual staging area[5]. The fields for which resembling domain sets are not available like in last name, mother name and course fields of student table can be filled with random values from the same source. Such randomization cannot impose auto correlation as resultant records are generated with crossover multiplication. The taster can also specify the range for valid primary key values according to the need of the testing environments[4].



First Name	Last Name	Father Name	Mother Name	Address	RollNo	Course	Date of Birth
Ram	kumar	P. Kumar	Shanti Devi	H.no 20 Urban Estate Pta. India	2345	MCA	22-6-1980
Birbal	khan	Jamshed Ali	zinath	#35 Model town Banglore India	2156	MBA	30-11-1979
Annie	peters	Robers Sr.	Giena	55 gandhi road calcutta	2456	M.Tech	21-3-1981
Peter	philip	jordan Philip	Ginny Philip	#78 Dwarka area new Delhi	2348	MCA	26-9-1980
Sophie	bem	David Gill	Sheetal Gill	54 Nabha Road MAndi	2467	M.Tech	11-1-1979

Employee Name	Father Name	Department	Designation	Date of Joining	Salary	Employee ID
Roger	Danny Cooper	Production	Supervisor	23 july1995	22000	ES 2098
Jai Singh	Bheem Singh	Sales	Manager	22 Dec 1980	50000	ES 6532
Kamal Kapur	Bishambhar Das	Sales	Sales Man	1 Jan 1989	15000	ES 2436
Barkat Rai	Prem Kumar	Marketing	Manager	15 Feb 1998	35000	ES 6548
Gaurav	Jai Parkash	HR	Asst. Manager	21 Mar 1980	40000	Es7658

Figure3. The ability of DSG to exploit resembling domain sets at different locations



First Name	Last Name	Father Name	Mother Name	Address	RollNo	Course	Date of Birth
Ram	kumar	P. Kumar	Shanti Devi	H.no 20 Urban Estate Pta. India		MCA	22-6-1980
Birbal	khan	Jamshed Ali	zinath	#35 Model town Banglore India		MBA	30-11-1979
Annie	peters	Robers Sr.	Giena	55 gandhi road calcutta		M.Tech	21-3-1981
Peter	philip	jordan Philip	Ginny Philip	#78 Dwarka area new Delhi		MCA	26-9-1980
Sophie	bem	David Gill	Sheetal Gill	54 Nabha Road MAndi		M.Tech	11-1-1979
Roger	khan	Danny Cooper	Shanti Devi	H.no 20 Urban Estate Pta. India		M.Tech	23 July 1995
Jai Singh	philip	Bheem Singh	zinath	55 gandhi road calcutta		MBA	22 Dec 1980
Kamal Kapur	lumar	Bishambhar Das	Ginny Philip	#35 Model town Banglore India		MCA	1 Jan 1989
Barkat Rai	peters	Prem Kumar	Sheetal Gill	54 Nabha Road MAndi		MBA	15 Feb 1998
Gaurav	bem	Jai Parkash	Giena	#78 Dwarka area new Delhi		M.Tech	21 Mar 1980

Figure 4 The clubbing of resembling domain sets available at different locations

Once the process of declaration is over and data generation constraints are defined then the data set generator multiplies the records by exercising the logic discussed earlier. The multiplication process continues till the count of newly generated records matches with the count specified by the user. In the end the newly generated records are

appended in the target database thus jacking up the size of the target database by multiple times. This puffed up database can now be used to test a big database application like a data warehouse[4], [6].

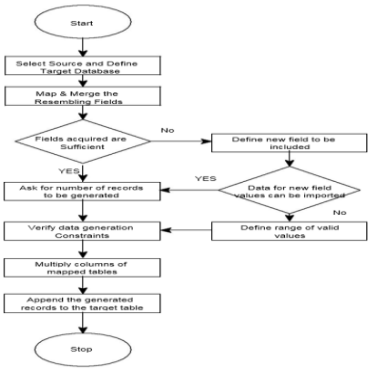


Figure 5 Flow Chart depicting the working of DSG

III. THE ETL DEVELOPMENT

A. Designing the Extract Module

Generally the data sources are operational they need to go offline as the time that data extraction takes place. Business organizations can't afford to suspend their operational systems for a long time hence the situation demands a balance in efficient ETL processing and to get the necessary source data within the allowed time slot. Prime requisite for a successful source extract is accurate field mapping. Mapping, selecting and merging data from the specified source databases is a challenging job because of different data definitions and high data Redundancy in operational data stores. On the fly data scrubbing and loading may need extra efforts for time consuming table lookups and cross referencing of specific keys, instances through extremely complex data extraction routines. This on the fly data-cleansing process is expected to slow down the extract process, which in turn would make the operational systems busy longer than is acceptable. These constraints of business environment forced the data warehouse industry to populate the BI databases in three different steps namely extract, transform and load. This concept introduced a staging area in between the source and target databases[7], [8].

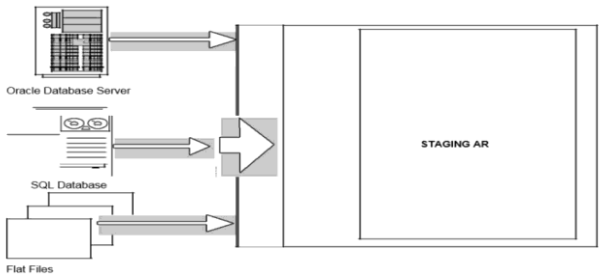


Figure 6 A Simple Data Extraction Module

Now the source extract is responsible to fetch the selected fields on to the staging area within the allocated time slot. The data cleansing and transformation takes place simultaneously at the staging area before loading the consolidated data onto the business intelligence database[9]. To develop an effective ETL routine one has to understand the prevailing business rules for applying transformations on the source data. The design of transformation logic initiates with activities like project planning, requirement definition and data analysis till application prototyping starts and is followed by Mata data analysis. A simple data extract module is shown in figure 6.

B. Designing the Transformation Module

There is a delusion in data warehouse business organizations that almost eighty percent of ETL processing effort is associated with data transformation routines. The extract and load routines are expected to contribute only twenty percent of the ETL processing effort [5] Involvement of heterogeneous sources with independent schema definitions makes it difficult to consolidate targeted business intelligence database. Decision needs to be taken in advance if the backward references from the target database to the source databases have to be kept or not? Managing inconsistent, inaccurate and duplicate data values is a crucial concern for transformation routines [10]. The varying data definitions and data formats in the source files demands specialized routines for consolidation and cleansing of data. Data cleansing is an ongoing process with every data load cycle. Further the transformation routines may transform the naming standards of source data according to the data warehouse business standards. Some data elements have to be split across different columns and on the other hand, some data elements from different operational systems may merge into a single column in the data warehouse target database. Working of transformation logic is shown in figure 7.

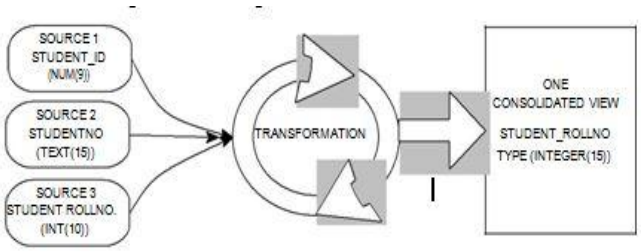


Figure 7 Data Transformation Logic

C. Designing the Data Load Module

The final phase in ETL processing is the loading phase. Its prime responsibility is to load the scrubbed and transformed records into the target data warehouse. Loading can be done in two ways; first way is the row by row insertion of data and the second option includes the

bulk load methodology. The organizations make a choice among the two according to the prevailing time and resource constraints.

D. ETL Sub Systems

A hand coded ETL routine to extract and unify data is the most popular option among small and mid-size enterprises[1], [11], [12]. The ETL is a big term having many small independent sub systems of its own like:

1. **Aggregate building System:** It is for creating and maintaining of physical database structures.
2. **Backup system:** It is responsible for backing up data and metadata.
3. **Cleansing system:** It is a dictionary driven system for parsing of names and addresses of individuals and organizations etc.
4. **Data Change identification system:** It keeps an eye over Source log file reader, source data and sequence number filters etc.
5. **Error tracker and handler:** It is a widespread system for identifying and reporting to all ETL error events.
6. **Fact table loader:** It is equipped with push/pull routines for appending/updating transaction fact tables.
7. **Job schedule Handler:** It is for scheduling and launching all ETL jobs.
8. **Late arriving fact and Dimension Handler:** It is for insertion of fact and dimension records that have been delayed in arriving at the data warehouse.
9. **Metadata manager:** It is for assembling, capturing and maintaining all ETL metadata and transformation logic.
10. **Pipelining system:** It is required for implementing streaming data flows.

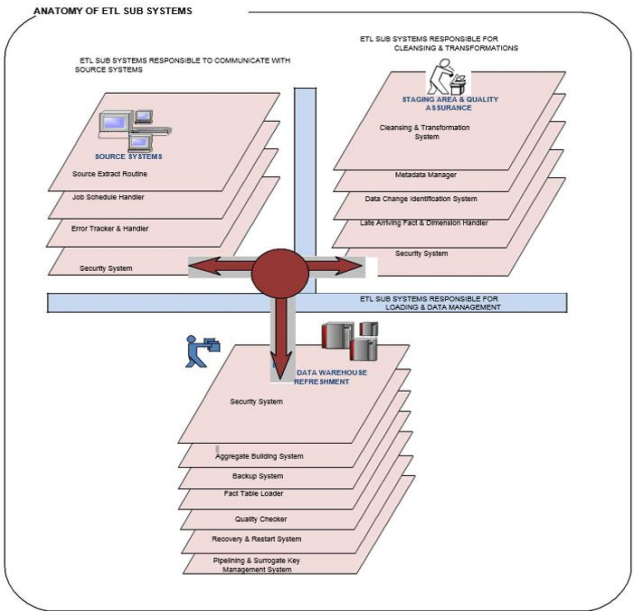


Figure 8 Anatomy of ETL Routine

11. **Quality Checker:** It is responsible to check the quality of incoming dataflow.
12. **Recovery and restart system:** It is responsible for restarting a job that has halted.
13. **Security system:** It is responsible for the security of data within an ETL.
14. **Source Extract system:** It includes Source data adapters along with push/pull routines for filtering and sorting at the source.
15. **Surrogate key Management System:** It is a Pipelined, multithreaded process for replacing natural keys of incoming data with data warehouse surrogate keys. All the aforesaid ETL subsystems are categorizes and bundled together in a layered fashion to provide a wall defined set of services. Keeping in view these categories the anatomy an ETL routine is shown with the help of figure 8.

E. Designing the Transformation Module

The Data Extractor is a tool developed for synthesizing data from multiple logical resources into the specified target database sink [13]. This primary ETL structure can extract records from numerous different resources; can scrub incoming data streams to the extent possible to ultimately load the cleansed records into a target database. Like any hand coded ETL routine The Data Extractor functionality is limited to managing personal records data.

Table 1 Workflow of ETL Prototype

Action	Description	
Start		
1	1.0	Select target database type
	1.1	Select target table
2	2.0	Select source database type
	2.1	Select source table
	2.2	Validate source availability and solve connectivity issues
3	3.0	Copy source data schema to staging area
	3.1	Copy target data schema to staging area
	3.2	Map source and target schema fields
4	4.0	On merge consult meta data repository
	4.1	Verify and validate data extraction rules
	4.2	Extract the specified source data onto staging area
5	5.0	Begin data validation, scrubbing and transformation procedure
	5.1	Maintain log file for rejected records
6	6.0	Load the purified records to the target database
	6.1	Commit the target data alterations
Stop		

F. Anomalies found in synthesized data

As a hand coded ETL routine is fanatical to a specific entity domain, this routine was customized to handle personal records from different sources. The synthetic data generated with the help of data set generator was placed on different logical locations and synthesized using this primary ETL structure. Following are the anomalies which were found in the data, synthesized from different locations.

a. Lexically Anomalies:

Syntactical errors are generally lexically errors which would name

discrepancies between the structure of the data items and the specified format. In a relational database the data is stored in the table form with each row representing a duple and each column representing an attribute. If there are five attributes of a relation or table than each tuple will also have five attributes.

But if some or all of the rows contain only four columns than the actual structure of the data will not confirm to the specified format. This wills result lexical errors. For example record no. 125 is lexically erroneous because the name column is representing a numeric value and the roll no. Column is representing a character string. Identified anomaly is shown in figure 9.

123	SAHIL	MAHAWAR	SH. RAKESH GUPTA	ASHA GUPTA	7084233546	#60, Sec-24A, St.No.7, Dilip Nagar, New
124	SANDEEP	KAUR	SH. SARABJIT SINGH	HARJIT KAUR	7084233547	Gobind Nagar, Near I.T.I., Sangrur Road, N
125	SANJEEV	KUMAR	SH. VED PARKASH	7084233548	Gulmohar Nagar,	9888153518
126	Santosh	Pun	Dam Bahadur Pun	MAN MAYA PUN	7084233549	#& Bihar Ncc BN Sub.Maj.D.B.Pun Sadh

Figure9. Lexical Anomalies

b. Format Errors:

These errors correspond to the value of a given attribute does not conform to the anticipated domain. For example record no. 56 is representing a correct name but in a wrong way. As the last name is specified first and the first name is specified as the last name hence it is violating the domain format. Format error is shown figure 10.

55	GURPREET	SINGH	SH. KRISHAN SINGH		7084233477	#42, Sec. - 10, Block - A, Guru Ki Nagri, Ne
56	KAUR	GURVEER	SH. SHINGARA SINGH	CHARANJEET KAUR	7084233478	Vill.Sahapur, Distt.Sangrur.
57	HARINDER	KAUR	SH. HARBANS SINGH	JASWANT KAUR	7084233479	#335, W.No.10, Sardar Colony, Bassi Road

Figure 10 Format Errors

c. Irregularities:

Irregularities are apprehensive with non-consistent use of values, units and abbreviations etc. For example in the target database where data has been merged from different sources the sex attribute has been represented as M/F and Male/ Female. Such anomalies will create a bottleneck for developing a generalized formula or query for data processing. Irregularities are shown in figure 11.

0165-5241607			5/29/1984	23 m
	GEN		5/28/1982	25 m
0165-5241908			8/27/1985	22 m
#3269, W.No.9, Nai Abadi, Sirhind Mondi, Distt. Fathegarh Sahib.	988841869	Gen	8/29/1984	23 female
Old Grain Market, W.No.10, Amloh	9417197114	Gen	12/11/1985	22 female
WH 4/8, Nai Abadi, Khanna.	01628-229284	Gen.	11/4/1986	21 male

Figure 11 Irregularities in fact representation

d. Constraint Violations:

Integrity constraint violation illustrates that set of tuples, which do not satisfy the imposed integrity constraints. Integrity constraints are essential to empathize the mini world by confining the set of valid instances. A constraint cans be defined as a rule for the representation of the knowledge about the domain along with the set of valid values for representing the concerned facts. For example the Dates 30th and 31st February are not possible and hence are violating the integrity constraints. Such values become apparent specially when data is imported from text/flat files where there is no provision to declare any integrity constraint. In the same way age cannot be zero if there is an entry for the Date of Birth column. Such anomaly is shown in figure 12.

Vill & P.O. Mangoo Teh.Arki, Distt. Sihan, H.P.	01796-247059	GEN	22-Dec-85	22 m
			28-Feb-85	22 m
			15-Nov-85	22 m
			15-Dec-87	0 m
			01-Mar-85	22 m
			02-Apr-82	25 m

Figure 12 Integrity Constraint Violations

e. Contradictions:

Contradictions refer to the values within a tuple of a relational database that contravene some kind of dependencies between values. For example the formula for calculating age is (Currant date (Date on which transaction was recorded) - Date of Birth). If this constraint is not followed correctly it may result in contradictions. For example in the figure given below the second record has date of birth value equal to 05 Apr 86 and the age is 24. A contradictory record is shown in figure 13.

			05-Apr-85	22 m
			05-Apr-86	24 m
#1625A, Old Gadda Khana, Near Arya Samaj Park, Patiala	0175-5004278	GEN	07-Jun-87	20 m
			14-Sep-84	23 m

Figure 13 Contradictions in records

f. Duplicates:

Duplicates can be stated as tuples representing the same world entity from the concerned mini-world. The values of these tuples do not need to be completely similar. These records may also result contradictions as they represents the same entity but with different values for all or some of its properties. Like in figure 4.20 the record no. 8 and 9 are representing the same entity from the mini world because the name and registration no. fields are same in both cases but information provided here is bewildering. Duplicate records may also lead to contradictions. An example of such duplicate record is shown in figure 14.

7	GUPTA	ANKUR	SH. NARESH KUMAR GUPTA	KUSUM LATA GUPTA	23250-ptu-98	#1625A, Old Gadda Khana, Near Arya Sa
8	ANKUSH	KHURANA	SH. SHIV K. KHURANA		23251-ptu-98	
9	ANKUSH	KUMAR	SH. SHIV K. KHURANA	USHA DEVI	23251-ptu-98	Sandeep Cloth House Kamal Road, Rama
10	ANTARPREET	KAUR	SH. CHANCHAL MEHBOOB S		23253-ptu-98	S.C.H Kamal Road, Raman Mandi, Distt.

Figure 14 Duplicate records with different serial number

g. Data Type Mismatch:

The data type mismatch can impose a serious problem while merging data from

various sources. For example it is possible to store an integer value in text data type but it is not possible to store a text value in an integer data type without conversion. Such conversions may or may not provide desired results.

h. Missing Values:

These values are the result of omissions during data collection.

The major reason for missing values is constraint violations as if we have null values for attributes where there exists a non null constraint for them. If missing values are present this means there is no such constraint imposed during data extraction. These values impose a major problem when one wants to substitute values for such records for example one cannot insert null values into an integer field because by default its value is set to zero. This insinuates that if we try to consolidate data by filling dummy values in null fields than numeric data types may require special attention because their default value is set to zero. Figure 15 shows missing values.

28	GAGANDEEP	KAUR	SH. GAMDUR SINGH	SHAMSHER KAUR	23271-ptu-98	#2567, Sector-70, Mohali
29	KAUR	GAGANDEEP	SH. MOHAN SINGH	MANJEET KAUR	23272-ptu-98	Plot No. 126, Ranjit Nagar, Sirhind F
30	GAURAV	SHARMA	SH. BRIJMOHAN SHARMA		23273-ptu-98	
31	GURMEET	SINGH	SH. AVTAR SINGH		23274-ptu-98	
32	GURPREET	KAUR	SH. DEVINDER SINGH		23275-ptu-98	
33	SINGH	GURPREET	SH. SWARN SINGH		23276-ptu-98	
34	GYPSY	ANEJA	SH. BAKSHI LAL ANEJA		23277-ptu-98	

Figure 15 Missing Values

i. Typographical Errors

Typographical errors are those errors, which grounds due to typing mistakes. These errors are almost impossible to identify with automatic checks. For example due to typing mistake male can be written as amle. The possible solution is to rearrange the data in ascending or descending order and analyze it manually afterwards.

j. Transformation & the preservation of meaning:

The transformation challenge is expressed in terms of successfully matching data fields. But this is not the best way to think about it. The real challenge is preserving the meaning held in records throughout the transformation process. To eliminate data inconsistency and integrity problems (as stated above) if one tries to purify data through data transformations it is again not an easy job.

k. Normalization Break-up:

Illogically or inconsistently stored data can cause a number of problems. A poorly designed database may provide erroneous information, may be difficult to use, or may even fail to work properly. Most of these problems are the result of two bad design features called: redundant data and anomalies. Redundant data is unnecessary reoccurring. Anomalies can be defined as any occurrence that weakens the integrity of your data due to irregular or inconsistent storage for example delete, insert and update irregularity that generates the inconsistent data. Basically, normalization is the process of efficiently organizing data in a database to reduce redundancy. Hence the two main objectives of the normalization process can be stated as to eliminate redundant data, which means storing the same data in more than one table, and to ensure that the data dependencies should make sense. Both of these are valuable goals as they reduce the amount of space a database consumes and ensure that data is logically stored. Generally it is considered as impossible to normalize a database but it can be done with the help of a smart ETL routine, which should be competent enough to normalize the target database before populating data into it.

IV. TESTING GOALS AND TEST CASES:

As ETL routines are accountable for synthesizing quality data for decision making hence assuring the quality of custom built ETL is vital [90]. Keeping in view the importance of data quality in decision making table 2 presents some general goals for testing an ETL application:

The aforesaid ETL testing goals provide vital information for developing prime test cases for a hand coded ETL routine. One can consider the following aspects of these testing goals while developing the test cases:

Table 2 Testing Goals for an ETL Application.

S NO.	TESTING GOAL	DESCRIPTION
1	**Completeness**	This test ensures that all expected data is loaded completely.
2	**Transformation**	It ensures that data transformations have followed business rules correctly.
3	**Data quality**	It is to ensure that the ETL application is aware of data quality definition and can reject, Corrects or ignores and reports invalid data.
4	**Performance**	It is required to verify that data loads and queries perform within expected time frames and that the technical architecture is also scalable.
5	**Integration testing**	It ensures that the ETL process is cooperating well with other data warehouse sub systems.
6	**User-acceptance testing**	It is to verify that the data warehouse ETL solution meets current user requirements and anticipates their future expectations.
7	**Regression testing**	It guarantees that the system performance remains intact each time a new release of code is completed.

Data Completeness

The basic test of data completeness is to ensure that all expected data is loaded into the data warehouse. This test validate that all records, all fields and the full contents of each field are loaded. The following test cases are possible in this category:

a) Boundary value analysis is required to find out database limitations if any.

b) Each data field should be tested for truncation of data values.

c) The count of source data records should be equal to the count

of data records loaded to the warehouse plus rejected records.

d) The rang and value distributions of the fields in a data set can be verified using a data profiling tool.

e) The unique values of key fields in source data should be preserved in the data warehouse.

A. Data Transformation

Testing the transformation logic of an ETL application is considered as the most difficult task of testing. One of the easiest technique is to select some sample records and "stare and compare" them to validate data transformations[4].

This technique can be useful but it involves step wise manual testing and experienced tasters who can understand the ETL logic. An amalgamation of automated data profiling tools and automated data movement validations can be a better methodology for ensuring ETL quality[3]. The following data movement is important from ETL prospective:

a) Correctness of ETL generated fields like surrogate keys have to be ensured.

b) If possible range and distribution of values in each field between source and target data should be compared.

c) One has to make sure that the data types in the data warehouse are same as declared in the design document.

d) Referential integrity should be maintained.

e) Synthetic test data should mutate the expected inputs of the ETL routine to ensure flexibility as the data definitions and validations may change according to changing business rules.

f) The type of input data required and expected results according to business rules should be recorded into a testing document. This technique can also be used to design the ETL routine.

g) There should be criteria to process incomplete and obsolete records.

B. Data Quality

The data quality management mechanism of an ETL routine is responsible for defining the data quality. This system manages data by selective substitution, rejection and correction of incoming data while preserving the actual information[3]. The data quality defined during design process can be ensured in following manner:

a) A wall defined mechanism is required to identify and eliminate duplicate records.

b) Data quality rules are usually not visible to the users hence it is important to decide how to process invalid data. Rejected invalid records can be stored in a log file which cans further be used to publish data quality reports. These data quality reports are vital in identifying logical and systematic issues with data sources.

c) Rectify and validate correct value if possible like city name can be rectified based upon the postal pin code.

d) Rejection of records is necessary if incoming data is not compliant with the data definitions of the data warehouse.

e) Validation of correct values can be made like age cans be calculated from date of birth.

f) Where to substitute null in case there is no value available.

C. Performance

With the expansion of data warehouse the ETL load times may increase and the response time of queries is also expected to grow because of rising complexity and huge volumes of data[5]. However these issues may deteriorate the performance of ETL routine but strong technical architecture and good ETL design can pose a remedy for such issues[2]. ETL performance can be evaluated on the following grounds:

a) Comparison of ETL processing times for extracting smaller data sets and larger data sets is necessary to anticipate scalability issues.

b) ETL should be tested for maximum expected data loads.

c) Simple and complex queries should be run on large database volumes to validate query performance.

d) Timing of reject process is crucial for analyzing huge volume of incoming data.

D. Integration Testing

Integration testing ensures that the ETL routine is cooperating wall with other data warehouse sub systems. Here instead of testing the ETL application alone one has to identify and test the possible interactions between various subsystems. Process failure handling mechanism and data recovery techniques at staging area also need to be tested during integration testing.

E. User-Acceptance and Regression Testing

To ensure user acceptance ETL routine should be tested with almost real looking data. The ETL routine should behave as per the expectations of the user. ETL interface should be user friendly as such changes at later stage may prove expensive[5]. The end users should be involved in interface design right from the beginning to develop an easy, acceptable and familiar user interface. Regression testing is required to validate the existing functionality with very software code developed to rectify a defect or to enhance the performance of the ETL routine [9], [11], [14]. The simplest possible technique for regression testing is to compare the results of previous successful results with the results of the newly developed code using the same source data sets. It is much quicker to analyze results only than to run an entire data validation procedure.

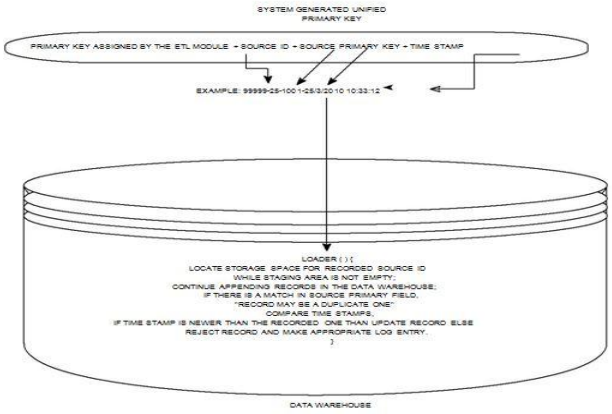


Figure 16 The Loading Logic of ETL Prototype

A successful test case is one which makes the system halt at the occurrence of any faulty event. The primary ETL structure was later analyzed for possible vulnerabilities. Accordingly, test cases of prime importance were developed and deployed to plug in the loop holes found in primary ETL routine. However it was also realized that to substantiate the need of automated testing, ETL performance needs to be analyzed statistically.

The aforesaid test cases were considered to be of prime importance for assuring the data quality hence all of these test were automated and included in the ETL routine developed. It has been observed that if quality checks are imposed during the data extraction stage than the effort required to refine and transform data can be reduced considerably resulting in the saving of time and money.

The primary reason for the convolution of the data extraction and transformation functions are the diversity of the source systems. This diversity includes be wildering combination of computing platforms, operating systems, database management systems, network protocols, and legacy source systems etc. Hence there is a need to pay special attention to the various sources and to begin with one should generate a complete record of the source systems. With this record as a starting point one should work out all the details of data extraction. The difficulties encountered in the data transformation function should also be related to the heterogeneity of the source systems. The loading procedure might seem to be the simplest one but it is solely responsible for consolidation and integration of targeted database hence its performance evaluation is also crucial[12]. Refined data from the staging area is to be loaded into the data warehouse according to the logic shown in figure 16 through the ETL prototype. The ETL data loading mechanism is responsible to generate surrogate keys for the data warehouse usage before loading the purified data into data warehouse tables. The surrogate key is a combination of four field values. The first field corresponds to a system generated primary key value. The second field is the identification number of the source to whom incoming data actually belongs to. Third field stores the primary key given to a specific record at its source database while the fourth field manages the time stamps for incoming data streams. As can be seen in the example shown in figure 16 amalgamation of all these four fields may result in a fool proof data management Strategy. While comparing source primary key field only one can identify the duplicates within the target database. Comparison of timestamp fields may help in the identification of records with slowly changing dimensions. The ETL performance was recorded before and after the imposition of automated testing. To verify the effectiveness of automated testing the recorded results needed to be analyzed statistically. Chapter 5 concludes the statistical analysis of the ETL routine performance.

V. STATISTICAL ANALYSIS

To verify the effectiveness of automated testing one has to check whether with the introduction of automated technique the count of errors have decreased considerably or not. Thus, to test the shift of location parameter paired t-test was applied following a generalized statistical hypothesis described below:

a) H_0: there is no shift of location parameter, i.e. mean value of samples taken before and after the introduction of automated testing is same, i.e. $\mu 1 = \mu 2$

b) H_A:location parameter is shifted on the lower side i.e. mean value of samples taken after the introduction of automated testing is less than that of the samples taken before the introduction of automated testing. i.e. $\mu 1 > \mu 2$.

Table 3 Anomalies Observed Before and After the Introduction of Automated Testing in Corresponding Samples.

Sp No	*Lexical Anomalies*		*Format Errors*		*Irregularities*		*Integrity Constraints*		*Duplicates*		*Semantic Errors*		*Contradictions*	
	Before x	After y	Before x	After y	Before x	After y	Before x	After y	Before x	After y	Before x	After y	Before x	After y
(x=sample values before automated testing, y= sample values after the introduction of automated testing)														
1.	65	4	75	6	50	2	40	3	120	3	65	6	45	2
2.	40	4	25	0	35	3	45	3	10	0	90	0	55	4
3.	50	2	35	1	45	0	60	4	65	6	70	6	40	0
4.	80	4	60	2	45	1	40	0	60	3	45	3	45	2
5.	40	2	45	0	75	3	65	2	62	4	50	1	55	2
6.	35	0	40	2	50	1	50	0	45	3	60	3	75	2
7.	75	6	80	3	60	0	65	3	60	2	55	4	50	1
SpNO:= Sample Number														

Testing in Corresponding Samples.

Here both $\mu 1$ and $\mu 2$ are the mean values of each data set of values before and after the introduction of automated testing. Initially H0 (null hypothesis) was followed with a perception that mean values of samples taken before and after the induction of automated testing are same. But a decrease in the mean value of samples taken after the induction of automated testing clearly represented a substantial decrease in errors. Table 4

Table 4 Comparisons of Calculated Values of t and p with Tabulated Values at 5% and 1% Level of Significance.

Anomalies	Actual calculated value of t test	Tabulated value of t using 5% level of significance	Tabulated value of t using 1% level of significance	p-values	Decision
Lexical	8.124	1.943	1.440	0.00013	Null Hypothesis Rejected
Format	6.824	1.943	1.440	0.00035	Null Hypothesis Rejected
Irregularities	10.38	1.943	1.440	0.000	Null Hypothesis Rejected
Integrity Constraints	11.593	1.943	1.440	0.000	Null Hypothesis Rejected
Duplicates	4.882	1.943	1.440	0.0013	Null Hypothesis Rejected
Semantic Errors	10.039	1.943	1.440	0.000	Null Hypothesis Rejected
Contradictions	11.997	1.943	1.440	0.000	Null Hypothesis Rejected

Above provides the details of applied paired t-test and its tabulated values at 5% and 1% level of significance along with the p-values to confirm the results. According to p-value concept lesser the value of p for each test there are more chances of selecting the alternative hypothesis, i.e. H1. Moreover, if the calculated value of t-statistic is more than the tabulated t-value at 5% and 1% level of significance than one has to reject the null hypothesis.

Thus the analysis of above table showed that in each category of error components we had rejected the null hypothesis and its alternative was accepted i.e. there is shift on the lower side of mean. This indicates a considerable decrease in data anomalies after the introduction of automated testing. The results of the statistical analysis performed earlier have been summarized with the help of Figure 17.

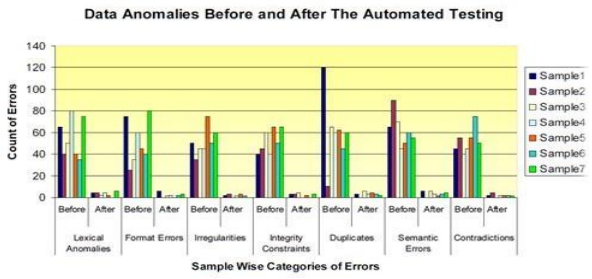


Figure 17 Bar Chart Representation of Statistical Analysis

The aforesaid analysis has revealed the fact that there is a significant gap between the perception and reality of prevailing data quality standards. The Y2K problem which led to modify software applications and databases using a two digit field to represent years is an archetype data quality issue turned up because of poor data quality standards. The concept of data is rapidly evolving, from structured data of relational databases, to semi structured data, unstructured data, documents, images, sounds, and maps resulting in a continuous change of the concept of data quality. Due to relative immaturity of data quality research area and absence of enforced de facto standards enacted by international organizations, it is extremely hard to formulate any standard data quality model for data warehouses. Hence there is an utmost need to classify and define the data quality dimensions and metrics from data warehouse point of view.

VI. FUTURE SCOPE OF THE RESEARCH

Following are the possible milestones that can be achieved with minor alterations in existing data warehouse design.

A. Intelligent Distributed Data Warehouse Systems

Multi database distributed systems are expected to gain importance over the next few years as most large corporations begin to take distributed technology for granted. Future systems will require not only the a techniques of distributed databases as available today to support interoperation of systems, but also the incorporation of techniques from artificial intelligence, knowledge-based systems and natural language processing, to provide the intelligence. In addition, richer information retrieval techniques capable of handling multimedia data will be required by future applications.

B. Contribution of the Web

Comparing the web marketing to traditional marketing, it is easy to see that the Web offers massive economic advantages for those needing to distribute information broadly. The break-even point for web deployment is much lower, and the variable costs are also less, meaning that any application successfully deployed in this environment will be much more profitable. Hence the scope of data sharehouse architecture proposed during the course of this research cannot be ignored as it seems to have immense potential to exploit the services offered by the refined computing technologies and the World Wide Web.

VII. CONCLUSION

During this research, customized ETL routine was developed, knowing the objectives to visualize the quality of performance. On the small data sets, the facts and the queries performance was satisfactory but it may not work well in real life environment. Therefore it is necessary to build a system which runs the queries on fully scaled data. Business ethics do not allow us to use the real business data for testing. So we made efforts to generate the test data, making sure the ratio of the facts and dimensions has a correct balance. To generate synthetic test data, we developed test data generator. Sub systems test scripts were written, keeping in mind the various ETL sub systems, to ensure quality data. Test cases are identified first, and then for Automated ETL testing, they are inculcated. By statistical analysis, it was revealed that the ETL testing can give us desired results.

For effective and easy maintenance of data, a novel architecture was proposed. Although the desired result of this research is achieved successfully and the objectives

are promising, but still there's a need to validate the results with the real life environment, as this research was done in simulated environment.

REFERENCES

[1] S. Chaudhuri, "<Sigrecord.Pdf>," no. March 1997, 1998.
[2] N. Rahman, "Refreshing Data Warehouses with Near Real-Time Updates," J. Comput. Inf. Syst., vol. 4417, no. Spring, p. 70, 2007.
[3] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Foundations of D ata W arehouse Q uality," no. 22469, pp. 2–13, 1998.
[4] S. Bruno, Nicolas; Chaudhuri and N. Bruno, "Flexible Database Generators," Vldb, pp. 1097–1107, 2005.
[5] A. A. Wani, B. L. Raina, "Issues and handy Solutions addressed at everystage in real time data warehousing , i . e . ETL (extraction , transformation & loading). - Literature Review.
[6] A. A. Wani, A. Khan, A. Jamal, and P. K. Gupta, "Cost Efficient Media Cloud Storage and Systematic Risks Involved in the Cloud Computing," no. 9, pp. 2466–2469, 2019.
[7] B. List, R. M. Bruckner, K. Machaczek, and J. Schiefer, "A Comparison of Data Warehouse Development Methodologies Case Study of the Process Warehouse," pp. 203–215, 2010.
[8] G. Swetha, D. Karunanithi, and K. A. Lakshmi, "Data Integration Models for Operational Data Warehousing," vol. 3, no. 2, pp. 508–516, 2014.
[9] R. J. Santos and J. Bernardino, "Real-time data warehouse loading methodology," p. 49, 2008.
[10] A. A. Wani and B. L. Raina, "Data in Data Warehouse and its Qualities Issues," no. 9, pp. 1753–1756, 2019.
[11] R. J. Davenport, "ETL vs ELT," no. June, 2008.
[12] K. Kakish and T. A. Kraft, "ETL Evolution for Real-Time Data Warehousing," Proc. Conf. Inf. Syst. Appl. Res., p. 12, 2012.
[13] A. A. Wani, U. Chandra, P. Bansi, and L. Raina, "Security Challenge in Big Data for Behaviour Analytics," vol. 5, no. 7, pp. 578–581, 2018.
[14] D. Agrawal, "The reality of real-time business intelligence," Lect. Notes Bus. Inf. Process., vol. 27 LNBIP, pp. 75–88, 2009.

AUTHORS PROFILE

Arif Ali Wani received his Bachelor's degree in Information and Technology from Model Institute of Engineering and Technology (MIET) affiliated to Jammu University, Jammu India. During the 2008 and M.Tech in Computer Science and Engineering from Gurgaon College of Engineering affiliated to Maharshi Dayanand University Rohtak, in the year 2013. He is having 9 years of teaching experience, his area of business is Data Warehouse and Data mining, Computer Network. He has published and his Research papers in peer reviewed International Journals, Book Chapters, international and national level conferences.

Bansi Lal Raina Backed by an exceptionally brilliant academic record, Prof. Raina has been engaged in administration, teaching & research for nearly 35 years now. He was awarded prestigious national fellowship of "TATA INSTITUTE OF FUNDAMENTAL RESEARCH" (T.I.F.R), Bombay, INDIA wherein he spent four years of research work and then proceeded to USA on an International fellowship to obtain his M.Tech (Computer Science Engineering & PhD. From 'USC', USA. Did he not only write an exemplary research paper at an early age of his career of 10+2 standard published by reputed 'American Mathematical Society' (January 1969 page 48-51), but his paper (part of which is noted below just for reference) was also widely acclaimed and often cited (e.g., See A. Del Cintel, 2008-SPRINGER) which in a dramatic development helped various eminent Scientists like Prof. ANDRE WILE then at PRINCETON UNIVERSITY, to draw a vital connection between the ELLIPTIC CURVES and MODULAR FORMS (See Ribet: Tanahama-Shimura Conjecture, 1986) leading him eventually to the famous solution in 1995 of even more famous CONJECTURE (See Annals of Mathematics, 142 (1995), which was unsolved for the last 350 years, earned Prof. Wiles a well-deserved 'KNIGHT HOOD' & the most prestigious award of 'FIELDS MEDAL'. Dr. Raina's above cited results have also immensely helped in the development of many subjects and more recently in 'CALABI-YAU' spaces & 'STRING Theory' in ASTRONOMY, thereby unifying the theories of 'NEWTON'S Gravitation, QUANTUM Physics & EINSTEIN'S Relativity.