Automated Data Validation for Data Migration Security

Manjunath T. N
Research Scholar
Bharathiar University
Coimbatore, Tamil Nadu, India

Ravindra S. Hegadi
Asst.Prof, Dept of CS
Karnatak University
Dharwad, Karnataka, India

Mohan H S
Research Scholar
Dr. MGR University
Chennai, Tamil Nadu, India

ABSTRACT
In the present information age, with the invent of new packages and programming languages, there is need of migrating the data from one platform to another as versions are changing, to perform quality assurance of any migrated data is tedious task which require more work force, time and quality checks. To perform efficient data migration process, legacy data should be mapped to new system by considering extraction and loading entities. The new system should handle all the data formats, the good design of data migration process should take minimal time for extraction and more time for loading. Post loading into new system results in verification and validation of data accurately, by comparing with the benchmark values and derive accuracy of data migration process. The manual data validation and verification process is time consuming and not accurate so automated data validation improve data quality in less time, cost and attaining good data quality, Author's emphasis on Automation of Data migration process for quality and security across industries.

Keywords
Data migration, Verification, ETL, Mapping, Parsing, DV (data Validation)

1. INTRODUCTION
Data migration may sound simple. It isn't. In fact, industry experience has shown data migration to be one of the most risky, complex and time-consuming IT projects. Cost overruns and delays are common. Business performance can suffer if data migration doesn't support strategic objectives behind a merger or acquisition, legacy systems modernization, or an upgrade to a new ERP or CRM application. Those problems are also avoidable. Informatica's market-leading Data Migration Suite, offers your organization proven data migration technology and expertise to (i) Reduce data migration cost and risk,(ii) Accelerate time to value,(iii) Improve data quality and (iv) Maximize business value of data[17].

The objective of this paper is to present the method of automating the Data validation for data migration from mainframe machine to DB, author's emphasis to process the huge volume of data within specified time, to increase the processing speed. One of the well known customers is opting to move data from mainframes to Oracle on UNIX platform, the main objective is to add value to the capital raising and asset-management process by providing the highest quality and most cost-effective self-regulated marketplace for the trading of financial instruments [12][9]. This paper will help Quality Analysts who does such kind of migration from one platform to another across domains.

2. LITERATURE REVIEW
Authors has undergone literature review phase and evolved with the problem statement with the help of work, has published till today in the area of data quality and data validations.

Robert M. Bruckner, Josef Schiefer Institute of Software Technology (1999) - describes the portfolio theory for automatically processing information about data quality in data warehouse environments.

Kyung-Seok Ryu, Joo-Seok Park, and Jae-Hong Park (2006) - Discuss the empirical methods of issues as data management matured.

Bloor Research (2007) - Data Migration Projects Are Risky: 84% of data migration projects fail to meet expectations, 37% experience budget overruns, 67% are not delivered on time.

Virginie Goasdoue EDF R&D (2008) - Proposed A evaluation framework for data quality tools-A practice oriented.

American health information management Association (2009) - Proposed a graphical method for DQM domains as they relate to the characteristics of data integrity and examples of each characteristic within each domain. The model is generic and adaptable to any application.

Jaiteg Singh and Kawaljeet Singh (2009) - The data quality was observed before and after the induction of automated ETL testing. Statistical analysis indicated a substantial escalation in data quality after the induction of automated ETL testing.

Manjunath T.N, Ravindra S Hegadi Ravikumar G.K (2011) - Discussed and analyzed possible set of causes of data quality issues from exhaustive survey and discussions with SME's.

Authors are proposing the method of automating the data validation for Data migrations for quality assurance and security resulting in effort and cost reduction with improved data quality parameters.

3. METHODOLOGY

3.1 Steps to Automate Data Validation
Below we have discussed all the steps which are required to automate the data validation of data migration from mainframe machine to DB. Data Migration is a process of migrating the data from one platform to another here author's emphasis with the case of mainframe machine to Oracle database [9][6]. The steps involved in the process are as below:

1. Extracting data from Mainframe server requires a lot of time.

2. The data in mainframe server is in EBCDIC format.

3. In order to access the data quickly, DBA team transfers the data in to UNIX server i.e., converts data from EBCDIC format to ASCII format using SAS tool.

4. The data in UNIX machine is stored in files (ASCII Format)

5. The data from UNIX machine is transferred to Oracle 10g server in the form of External Tables (Sources for our Data Validation).

6. The data in Oracle server is stored in the form of tables.

7. We will generate the test data using the reference documents like Mapping document, Parsing rules document and ETL document, and thus create a QA sub table called test data or reference data.

8. We compare the target table (Dev team Responsible) with Source QA table (QA team Responsible).

9. The compare2tabs script compares all the columns between source-QA and target-DEV and generates the DIFF table.

10. Analyze the DIFF table using SQL Scripts and in case of discrepancy, we will recheck with the requirements and raise an issue.

11. We track the issue till it is resolved.

12. The process is repeated for 'n' number of files to be migrated.

Below are the documents we are using for processing the data:

Parsing Rules -Document containing Parsing rules to extract data from Data Ware house.

Mapping document- To map data between the schemas

ETL Document- ETL Rules while extracting the data.

3.2 Block Diagram for Automated DV

Figure-1 describes the various components involved in automating the Data validation process. Experiment has two environments QA and DEV,In DEV environment developers will prepare the migration script to load the data into target tables, author wanted to validate the data which Developer has loaded into target DB(tables),instead of doing conventional manual process for validation, author has automated the Data validation process, which will saves time and cost,befefits of this process are completeness of values, validity, accuracy to source,precision,Non duplication of occurrences, derivation Integrity, accessibility, timeliness and definition conformance by all these parameters one can obtain high data quality[6]. The components described in figure-1 are explained in section 3.1.

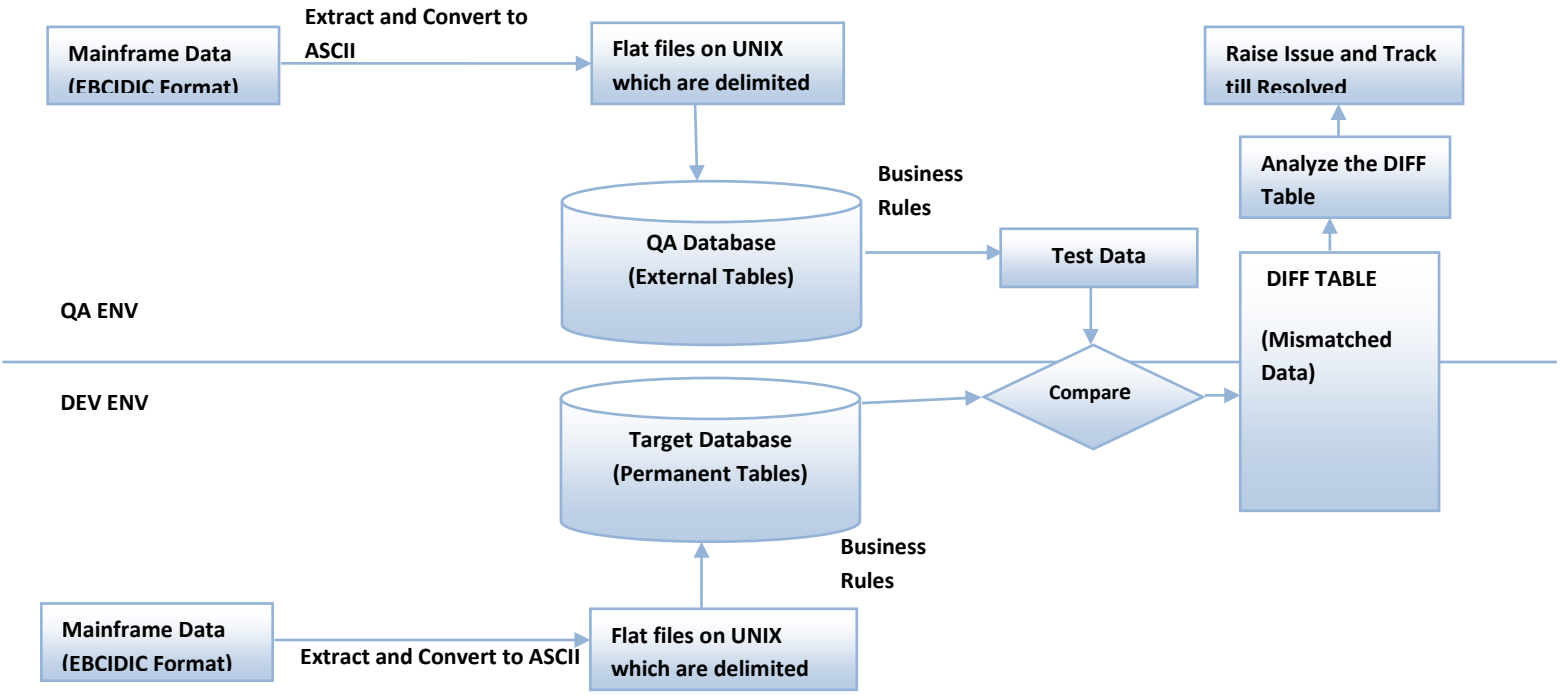

Figure-1: Block Diagram for Automating Data Validation

4. STAGES OF DATA MIGRATION

Ultimate goal of a data migration is to move data, a lot of upfront planning needs to happen prior to the move in order to ensure a successful migration. In fact, planning is the number-one success factor for any migration project, independent of the complexity. Not only does upfront planning help shorten the duration of the migration process, but also it reduces business impact and risk for example, application downtime, performance degradation, technical incompatibilities, and data corruption/loss. The migration plan the end result of the planning defines what data is moved, where it is moved, how it is moved, when it is moved, and approximately how long the move will take [9][7].

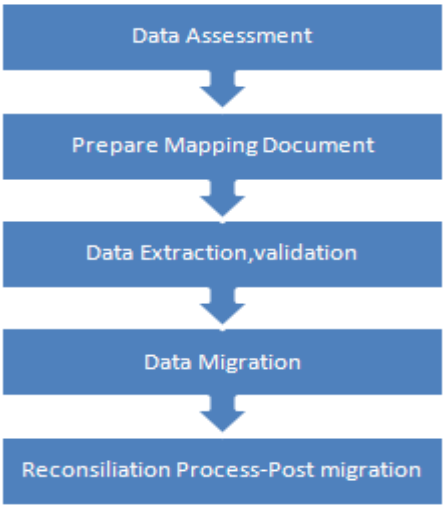

Figure-2: Stages of Data migration

Data Assessment key activities are (i) Identify data sources (ii) Run system extracts and queries (iii) Conduct user interviews on data migration process (iv) Review migration scope and validation strategy and (v) Create work plan and milestone dates.

The Outputs will be (i) Migration scope document (ii) Migration validation strategy document (iii) Work plan with milestone dates.

Mapping Document Key Activities are Prepare the excel sheet with these Columns (i) Source Table (or Filename) (ii) Source Column (or Fieldname) (iii) Source Data Type (iv) Validation Rules (source data is derived or direct values). (v) Typical Values. (vi) Transformation Rules. (vii) Target Table (viii) Target Column (ix) Target Data Type.

The outputs are (i) Modified source data that increases the success of automated data conversion (ii) Control metrics and dashboards.

Data Extraction, Validation and Loading key activities are (i) Create/verify data element mappings (ii) Run data extracts from current system(s) (iii) Create tables, scripts, jobs to automate the extraction (vi) Address additional data clean-up issues (v) Execute application specific customizations (vi)Run mock migrations (vii) Load extracts into the new system using ETL tools or SQL loader with bulk loading functions (viii) Conduct internal data validation checks including business rules and referential integrity checks (ix) Report exceptions to client team (x) Perform data validation.

The outputs are (i) Extracts from source system. (ii) Data jobs, scripts (iii) Application loaded with converted data (iv) Exceptions, alerts and error handling control points migration modules (packages) [7].

Migration Validation key activities are (i) Run final extracts from the current system(s) (ii) Execute specific customizations on target database (iii) Execute application specific customizations (iv) Run pilot migrations (v) Load extracts into the new system using ETL tools or SQL loader with bulk loading functions (vi) Conduct internal data validation checks including business rules and referential integrity checks (vii) Report exceptions to client team (viii) Perform data validation (ix) Prepare migration validation reports and data movement metrics (x) Review migration validation reports and metrics (xi) Record count verifications on the new system (xii) Reconcile or resolve any exceptions or unexpected variations (xiii) Sign off based on migration validation.

The output of this will be Signed-off migration validation document

Reconciliation Process-Post Migration key Activities are (i) Complete data migration reports and cross-reference files/manuals (ii) Data sanity reports (iii) Target system reports.

The output will be (i) Exception reports, cross-reference files/manuals (ii) Infrastructure dashboards (iii) Signed-off data migration activity. The output of all these stages of data migration is data quality, if the migration effort does not formally specify the level of end-state data quality and the set of quality control tests that will be used to verify that data quality, the target domain may wind up with poor data quality. This may result in the following issues.

- Costs associated with error detection
- Costs associated with error rework
- Costs associated with error prevention
- Time delays in operations
- Costs associated with delays in processing
- Difficulty and/or faulty decision making
- Enterprise-wide data inconsistency

We have successfully migrated data for one of the biggest customer with this proposed method with the following challenges (i) 35 million records/day/table (Aprox) (ii) Testing Historical Data (One month data) (iii) More no of colums (more than 120 column in source files) (iv) Datatype conversion from mainframe to Relational DB (v) More accesstime to pull data for validation from VM files (vi) Excel sheet limitation is 65000+ records to put for comparision.Figure-3 shows the historical data processed per day by authors [9].

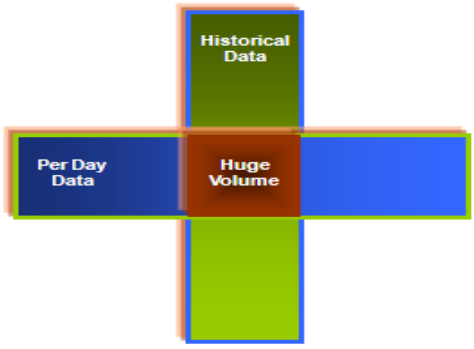

Figure-3: Different Parameters used for data migration

Derived fields are validated using the concept of divide and conquer strategy.Computing Derived field values from 35millionrecords/column/table/day(Aprox),One Derived field value may depend on 5 column from 5 dependent table. Joining time and conversion time is more.Complex Business Rules and More time for filtering and conversion.In figure-4 shows various derived fields are considered for business validations using joins and business rules.

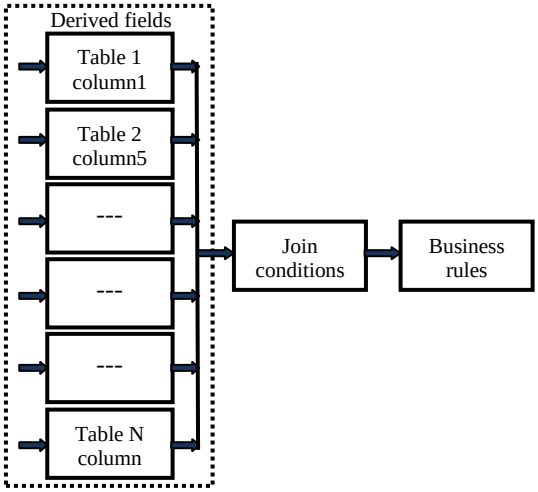

Figure-4: Evaluation of Derived Columns

4.1 Case Study of Data Migration using SQL

4.1.1. E-R Diagram for two tables in source

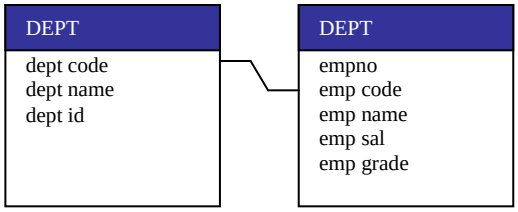

Figure-5: E-R diagram for two source tables

DEPT and EMP are two source tables with column names listed in the ER diagram we have 1: N relationship between DEPT and EMP. We have **EMP, DEPT** tables in source and **Customer** table in target.

4.1.2 Source Tables Structures

EMP Table

Emp_no	Emp_code	EMP_name	Emp_salary
123	1	Mahesh	20000
438	3	Rajesh	45000
156	2	Santosh	30000
888	4	Ramesh	25000

DEPT Table

Dept_code	Dept_name	Dept_id
1	Clerk	10
3	Assistant	20
2	Manager	30
4	Clerk	40

Figure-6: Source Table Structure

4.1.3 Target Table Structure

CUSTOMER:

Cust_no	Cust_code	Cust_name	Cust_salary	Cust_Dept_name
123	1	Mahesh	20000	Clerk
438	3	Rajesh	45000	Assistant
156	2	Santosh	30000	Manager
888	4	Ramesh	25000	Clerk

Figure-7: Target Table Structure

4.2 Validation Methods

4.2.1 Mapping Validation

This mapping shown below is from source to target.

Target			Source			
schema	Table	Column	schema	Table	Column	Validation Rule
Target	Customer	Cust_no	Migrate	Emp	Emp_no	Cust_no is mandatory

Figure-8: Mapping showing from Source to target

Select the source side data by using below query.

```
SELECT emp_no
FROM emp;
```

And select the target side data by using the below query.

```
SELECT cust_no
FROM customer;
```

Then perform "MINUS" operation on above queries.

```
SELECT emp_no
FROM emp;
MINUS
SELECT cust_no
FROM customer;
```

Then the output contains records of source which are not present in target.

4.2.2 Transformation rule Validation

```
SELECT cust_no
FROM customer
WHERE cust_no is null;
```

As per the transformation rule cust_no is not null. So use NEGATIVE test case to check transformation rule.

4.2.3 To Find Duplicates

We can find duplicates present in the cust_no by using following query.

```
SELECT cust_no, count (*)
FROM emp
GROUP BY cust_no
HAVING count (*) > 1
```

5. RESULTS

We have captured the experimental results on the real world data of well known customers we are working on. To automate the data validation for data migration process we used SQL queries. This method is very simpler and fast. First we need to write select query on source and target both side. Then perform MINUS operation on both select queries. This one will give the records which are not present in target, we have implemented a GUI based application which will handle this as shown in figure-9. The benefits from this process are speed, accuracy and timeliness, precision and other quality factors can be achieved. We have successfully validated the data after data migration for one of well known customer data.

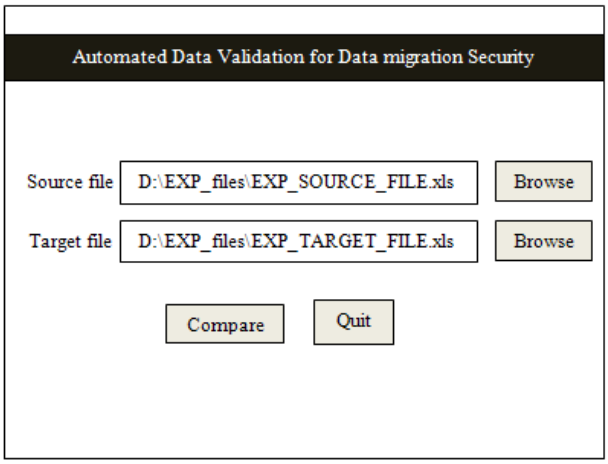

Figure-9: GUI Screen for Automated Data Validation for Data Migration Security

Figure-10 gives the details for manual and automated process based on data volume, it uses and time it takes to process in Secs.

Data Volume	Time (Secs)	Manual Process (Secs)	Automated Process (Secs)
10000	5	4	2.79
20000	10	6.5	5.58
30000	15	9	8.37
40000	20	13	11.16
50000	25	15	13.95
60000	30	18	16.74

Figure-10: Manual Process and automated Process across Data volume and time (Secs)

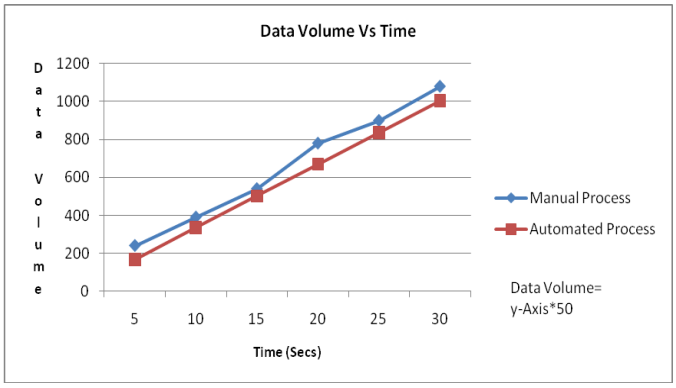

Figure-11: Data Volume Vs Time (Secs).

Figure-11 captured the experimental results of automated and manual process, the graph depicts automation is more efficient then manual process with respect to time and data volumes, approximately 11% efficiency is increased after automating this process.

Figure-12 depicts the manual and automated process with respect to data volume and defects raised.

Data Volume	Defects	Manual Process(cumulative defects)	Automated Process(cumulative defects)
5000	2	0	0
10000	4	2	3
15000	6	4	5
20000	8	5	8
25000	12	8	12
30000	15	11	15

Figure-12: Manual Process and Automated Process Across data Volume and Defects.

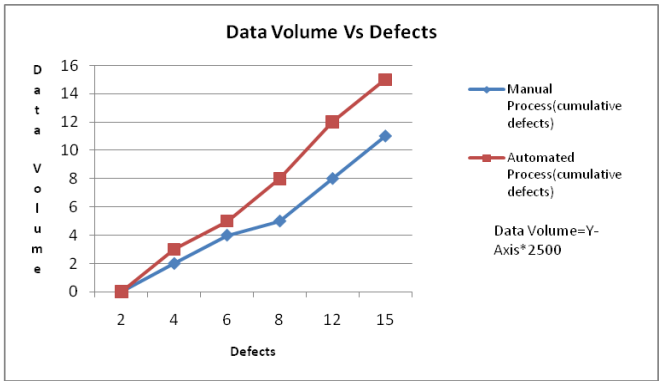

Figure-13: Data Volume Vs Defects

Figure-13 shows automated is more efficient then manual with respect to defect raised and time taken and quality parameters.

6. CONCLUSIONS

We have successfully automated the data validation used for data migration, the proposed framework will fit for any kind of source data, bit of customization is required based on complexity, This model is working fine with main frame source data migration into Oracle DB, this process and framework and methodology for Data migration where huge volume of data is involved between the different databases in less time more accurate and reducing effort by 25 to 30 %.This best practice and methodology is ideally suits for any data migration QA works. The benefits from this model and framework, one can save time, money and data quality will be high around 30% when compared to manual process and it is 11% efficient with respect to time and defects raised.

7. ACKNOWLEDGMENTS

Information presented in this paper was derived from Experimental results conducted on real data and conversations with Subject Matter Experts (SME) on Data migrations of various IT companies of India and abroad The authors gratefully acknowledge the time spend in this discussions provided by Mr.Shahzad, SME, CSC USA, Mr. Vasanth Kumar ITC Infotech INDIA. Mr. Govardhan (Architect) IBM India Pvt Ltd.

8. REFERENCES

[1] English, L. P. (1999). Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits, John Wiley and Sons, Inc.Data Quality Issues, Page 10.

[2] Fields, K. T., Sami, H. and Sumners, G. E. (1986). Quantification of the auditor's evaluation of internal control in data base systems. *The Journal of Information Systems*, 1(1), pp. 24-77.

[3] Firth, C. (1996). Data quality in practice: experience from the frontline, Paper presented to the Conference of Information Quality, 25-26 Oct.

[4] White paper by Vivek R Gupta, Senior consultant, System Services Corporation, "An Introduction to Data Warehousing".

[5] Potts, William J. E., (1997), Data Mining Using SAS Enterprise Miner Software. Cary, North Carolina: SAS Institute Inc.

[6] SAS Institute Inc., (1998), SAS Institute White Paper, From Data to Business Advantage: Data Mining, the SEMMA Methodology and the SAS® System, Cary, NC: SAS Institute Inc.

[7] Microsoft® CRM Data Migration Framework White Paper by Parul Manek, Program Manager Published: April 2003.

[8] Jaiwei Han, Michelinne Kamber, "Data Mining: Concepts and Techniques."

[9] DATA MIGRATION BEST PRACTICES NetApp Global Services January 2006.

[10] DATA Quality in health Care Data warehouse Environments Robert L.Leitheiser, University of Wisconsin –Whitepaper

[11] Badri, M. A., Davis, Donald and Davis, Donna (1995). A study of measuring the critical factors of quality management. *International Journal of Quality and Reliability Management*, 12(2), pp. 36-53.

[12] Larry P. English. Improving Data Warehouse and business Information Quality. Wiley & Sons, New York, 1999.

[13] Miles, M. B. & Huberman, A. M. (1994). *Qualitative Data Analysis - A Source Book of New Methods*, Sage Publications, Thousand Oaks.

[14] Ravikumar G K,Dr.Justus rabi, Manjunath T.N, Ravindra S Hegadi,"Design of Data Masking Architecture and Analysis of Data Masking Techniques for Testing - IJEST11-03-06-217- Vol. 3 No. 6 June 2011 p.5150-5159

[15] Ravikumar G K, Manjunath T N, Ravindra S Hegadi, Umesh I M "A Survey on Recent Trends, Process and Development in Data Masking for Testing"-IJCSI- Vol. 8, Issue 2, March 2011-p-535-544.

[16] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K."A Survey on Multimedia Data Mining and Its Relevance Today" IJCSNS. Vol. 10 No. 11-Nov 2010, pp. 165-170.

[17] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K."Analysis of Data Quality Aspects in Datawarehouse Systems", (IJCSIT)-Jan-2011.