Data Quality Assessment Model for Data Migration Business Enterprise

Manjunath T.N [#1], Ravindra S Hegadi [*2]
[# 1]Research Scholar, Dept of Computer science, Bharathiar University, Tamil Nadu, India
[*2]Assoc Professor, School of Computational Sciences, Solapur University, Maharastra,India
[#1] manjunath.tnarayanappa@rediffmail.com
[*2]rshegadi@gmail.com

Abstract-**In today's business world, decision making is challenging task due to emerging technologies and strategies applied for enterprise. To make accurate decisions, the data present in the decision databases should be of good quality, in this relation we emphasis on building data quality evaluation model for data migration or Extraction Transformation and Loading (ETL) business enterprises. However, there are no standard tools in the market for evaluating the ETL output in the decision databases. The proposed model evaluates the data quality of decision databases and evaluates the model at different dimensions like accuracy derivation integrity, consistency, timeliness, completeness, validity, precision and interpretability, on various data sets after migration.**

Keyword- DQ Assessment, Data Migration, Business Enterprise, ETL

I. INTRODUCTION

Data migration is an integral part of business knowledge process improvement and makes it accessible from the new system. Data migration is a set of activities that moves data from one or more legacy systems to a new system, it is actually the translation of data from one format to another format. Data migration is necessary when a company upgrades its database or software, either from one version to another or from one platform to another [6] [7]. After data migration or ETL process one should ensure the data quality of the target system, there is necessity to compute the data quality dimensions of the migrated data. Methods to compute the Data Quality Parameters have been explained as part of data quality Parameters Presentations and data quality indexes at the parameter level should be computed to generate a data quality assessment report to make proper business decisions. 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, which may result in (i) Costs associated with error detection,(ii) Costs associated with error rework (iii) Costs associated with error prevention (iv) Time delays in operations (v) Costs associated with delays in processing (vi) Difficulty and/or faulty decision making and (vii) Enterprise-wide data inconsistency [2][1] .Author developed the data quality assessment framework with different data quality dimensions to derive the quality of the target system using decision tree method. Author hopes this study will help in analyzing and ensuring the quality assurance of the warehouse system or any decision making databases. Author illustrated the data quality assessment model with the case study of loan Repayment data mart along with its derivates.

II. RELATED WORK

In present days decision making is so important for any business, Inaccurate, incomplete reporting leads to wrong decisions and redundant data handling leads to cost replications and synchronizations. Good quality helps to increase customer value. so with the implementation of data warehouse systems can be achieved, one of the challenge is when data is loaded into datwarehouse from legacy systems, need to verify the quality of the data present in data warehouse or target system, there is a need to develop uniform framework, author emphasis in constructing the framework to check the data quality for data quality testing. Data Quality is the measure of accuracy of data which meets the business requirements and supports to the decision makings. Data quality can be assessed with various dimensions such as completeness, accuracy, precision, consistency, derivation integrity.etc. In 2001 by Virginie Goasdoue, Sylvaine Nugier, Dominique Duquennoy and Brigitte Laboisse proposed an Evaluation Framework for Data Quality Semantically [5]. In 2006 by Kyung-Seok Ryu, Joo-Seok Park, and Jae-Hong Park proposed a Data Quality Management Maturity Model and showed empirically that data quality improves as data management matures [6]. A survey done by Bloor research group in 2007, highlighted 84 percent of projects were running over time or budget. The most commonly cited reason for the project was legacy migration i.e. 27 percent. Bull presented in London conference and emphasized on the data migration importance and its data quality for business decision, i.e. they wanted to migrate from a mainframe-based navigational database to a relational database on a non-mainframe platform. It used Reverse toolset to migrate the existing COBOL code so that it runs directly and without change, against the PostgreSQL database. The next phase will be to re-develop the application software. A Survey done on Poor data quality most common in business intelligence problem by Jeff Kelly, News Editor Published in 20 Sep 2010. The most

noticeable change is that we nearly always record that the biggest complaint is query speed, said by Barney Finucane, a BARC analyst and lead author of the BI Survey's. BARC has been conducting the survey of BI end users since 2001. This year's survey included responses from nearly 2,200 end users, or consultants on behalf of end users, most from Europe and North America. Impact of data quality on business domains, survey emphasis that bad data means big business intelligence problems. The International Association for Information and Data Quality (IAIDQ) was established in 2004 to provide a focal point for professionals and researchers in this field [5]. IAIDQ have recently completed survey in 2010 with full report shows the importance of data quality in the present real business world. Manjunath T N et.al [2011] highlighted different data quality dimensions in different stages of data migration or ETL cycle. No literature found on the developing uniform data quality assessment model using Key Performance Indicators (KPI) and decision tree method of accessing the data quality for data migration business enterprise.

III. DATA QUALITY ASSESSMENT FRAMEWORK AND ALGORITHM

After Data migration | ETL process from legacy system to data warehouse / target system, need to assess the data quality of the target system with respect to underlying data, author derives the data quality characteristics in terms of KPI's (Key performance indicators), compute the KPI values for target system, compare these computed KPI's with the threshold values, construct the decision tree to predict the data quality of the target system and give the feedback on the data quality for the end user[6] [7].

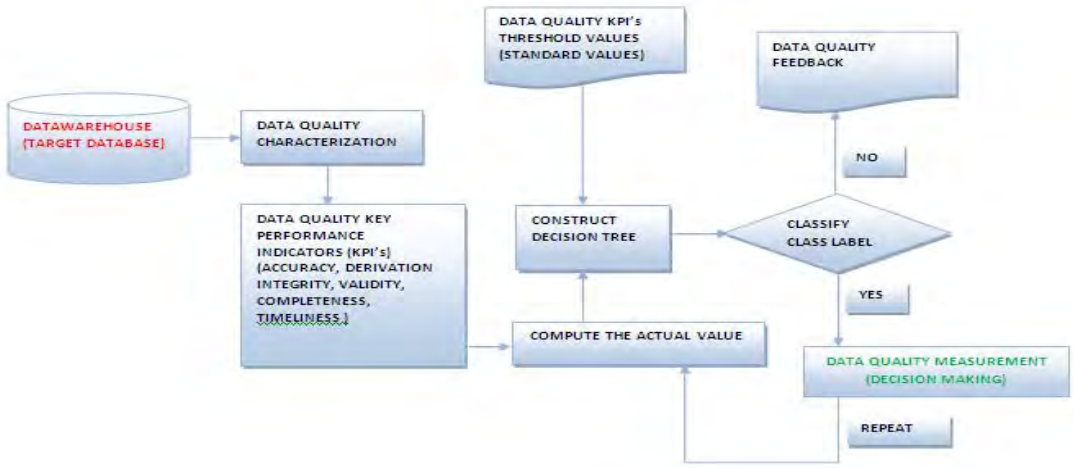

Figure-1: Data Quality Assessment Framework

Data Quality Measurement is necessary for any business decisions; figure-1 describes the components used for the framework. Accuracy: Is a measure of the degree to which data agrees with data contained in an original source. Validity: Is a measure of degree of conformance of data values to its domain and business rules. Derivation integrity: Is the correctness with which two or more pieces of data are combined to create new data. Completeness: Is the characteristic of having all required values for the data fields. Timeliness: Is the relative availability of data to support a given process within the timetable required to perform the process [4] [5]. The figure-2 confers the algorithm for data quality assessment.

Algorithm-Data Quality assessment Framework

```
Begin algorithm
for each table in the source database SD (1...N)
        for each attribute Aij in SD (i rows and j columns)
                for each record in the target database TD (1...M)
                        for each attribute Bij in TD (i rows and j columns)
                                Evaluate DQ KPI's
                                  for i=1 to n (where n=no of Entities in D)
                                   Begin
                                        Accuracy ()
                                        Derivation integrity ()
                                        Validity ()
                                        Completeness ()
                                        Timeliness ()
                                        Accessibility ()
                                        Consistency ()
```

End
Store all DQ KPI's in Temp table
end for.
end for.
end for
end for.
Construct Decision tree for different KPI's i.e. DT
If (DT is Acceptable) break
Else not acceptable for decision making
End algorithm.

Figure-2: Algorithm for data quality assessment framework

IV. DATA QUALITY COMPUTATION CASE STUDY: LOAN REPAYMENT DATA MART

Concise indication of Loan Repayment Data Mart. Leading Bank is having operations in all the states across the country. The Bank offers a wide range of banking products and financial services to corporate and retail customers through a variety of delivery channels and through its specialized subsidiaries and affiliates in the areas of investment banking, venture capital and asset management. The Bank has ERP systems and legacy systems for capturing the transaction data. The bank has decided to go ahead with the implementation of data warehousing solution to solve the repayment business need. Figure-3 describes the complete business requirements to perform ETL Process | data migration from legacy to data mart.Figure-4 gives the multi-dimensional model with start schema.

4.1 Typical Requirements Specification for legacy Data system

Bank wants to design a data mart that will meet the following requirements. The bank should be able to view the following information from the data mart.

1. List of defaulters
2. List of customers who have repaid the loan amount completely.
3. List of customers who have opted for partial prepayment of loan.
4. List of customers who have opted for full prepayment of loan.
5. List of customers who have completed 25 percent of their loan.

4.2 Design: Loan Repayment Data Mart

The data mart will have the following tables.

1. Loan product
2. Customer
3. Branch
4. Payment mode
5. Time

4.3 Mapping Document with typical business rules

Field Name	Data Type	Description	Business Rules
Actual_date_payment	Date type	Date when the loan installment (EMI) is actually paid by the customer	Entered in dd-mm-yyyy format
Address	Text	The address of the customer	Maximum characters: 40, can also include special characters
Balance_loan_amount	Currency	The outstanding loan amount for the customer	Can take only integers (no decimals). Values from Re 1 to Rs 10 million
Branch	Text	The name of the branch from where the loan is availed.	Can take values only mentioned in the drop down box
Branch_id	Text	The unique id given to the branch	Can take values only mentioned in the drop down box
City	Text	The city of the customer	Maximum characters: 20, cannot include integers or special characters
Credit_rating	Text	Credit rating given to the customer.	Can take values only mentioned in the drop down box. Four values: A, B, C & D
Customer_id	Text	The unique customer id allotted to the customer.	Unique customer id allotted to the customer. Same for all banking

			transactions for the customer in the bank.
Customer_Type	Text	The type of customer such as salaried or self-employed.	Should not accept any other value than those identified.
Date_key	Date type	This contains the format of the date	
Day	Text	This contains all the days of the week	
Delayed_payment_days	Number	This is the payment delay (in number of days). It is equal to the "Actual_date_payment" minus "Due_date_payment".	Would be an integer.
Description	Text	Loan description (Housing loan, Vehicle loan, Personal loan)	Can take values only mentioned in the drop down box
DOB	Date type	Date of Birth of the Customer	Entered in dd-mm-yyyy format
Due_date_payment	Date type	Date when the loan installment (EMI) need to be paid by the customer	Entered in dd-mm-yyyy format
Duration (months)	Number	This is the total number of months for which the loan has been sanctioned.	It should be between 12 months and 240 months. Must be an integer.
Eff_date	Date type	This is the data from which the revised interest rate is effective.	
Equated_Monthly_Installment	Currency	The monthly installment which the customer has to pay every month.	Calculated amount. The decimals values are rounded off.
Exception_id	Number	The id for all the exceptions which can happen during the tenure of the loan	
First_name	Text	The first name of the customer	Maximum characters: 20, cannot include integers and special characters
Fulldate	Date type	This is the date in dd-mm-yyyy format.	
Last_name	Text	The last name of the customer	Maximum characters: 20, cannot include integers and special characters
Loan_End_date	Date type	Date when the loan installment for the loan shall be paid by the customer	Entered in dd-mm-yyyy format
Loan_id	Text	This is the id allocated to the loan taken by the customer.	Should be unique. One customer id can have multiple loan ids.
Loan_product_id	Text	The id allocated to various types of loans like HL, PL and VL.	Should accept only allocated ids.
Loan_Start_date	Date type	This is the date when the disbursement of the loan takes place.	
Month	Text	This contains all the months of the year.	
No_of_installments_defaulted	Number	This is the number of months for which the customer has defaulted.	Can accept up to one decimal.
Pay_mode_description	Text	The description of various modes of payments such as Cash, ECS and Cheque.	Should not accept any other value.
Payment_made	Date type	This is the date when the payment is actually made by the customer.	Entered in dd-mm-yyyy format
Prepaid_full_penalty_charges	Number	This is the penalty (in %) which is charged on the outstanding principal amount when the customer wants to foreclose the loan.	
Repayment_no	Auto Number	This is the installment number being paid by the customer. Should be increased by one over the previous repayment number.	Auto generated
ROI	Number	This is the interest rate being charged to the customer.	For HL, it should be equal to bank rate plus spread. Can include up to 2 decimals.
ROI_type	Text	Includes two types: fixed rate and floating rate.	Should not accept any other value.
Salesperson_id	Text	The id allocated to the sales person	Should be unique for every sales person.

Salesperson_name	Text	Name of the sales person against whom this loan id is tagged.	Maximum characters: 40, cannot include integers and special characters
Spread	Number	This would be the percentage mark up for the customer over the bank rate. Depends on credit rating.	Should increase with a fall in the credit rating. Can include up to 2 decimals.
Status_flag	Text	It contains two flags: Yes and No.	Should not accept any other value.
Total_Loan_Amount	Currency	The total loan amount sanctioned by the bank	Can take values between Rs 5000 to Rs 10 million.
Week_Number	Auto number	The auto generated week number of the year.	

Figure-3: Typical business rules while performing ETL| Migration

4.4 Multidimensional Data Model: Loan Repayment data mart

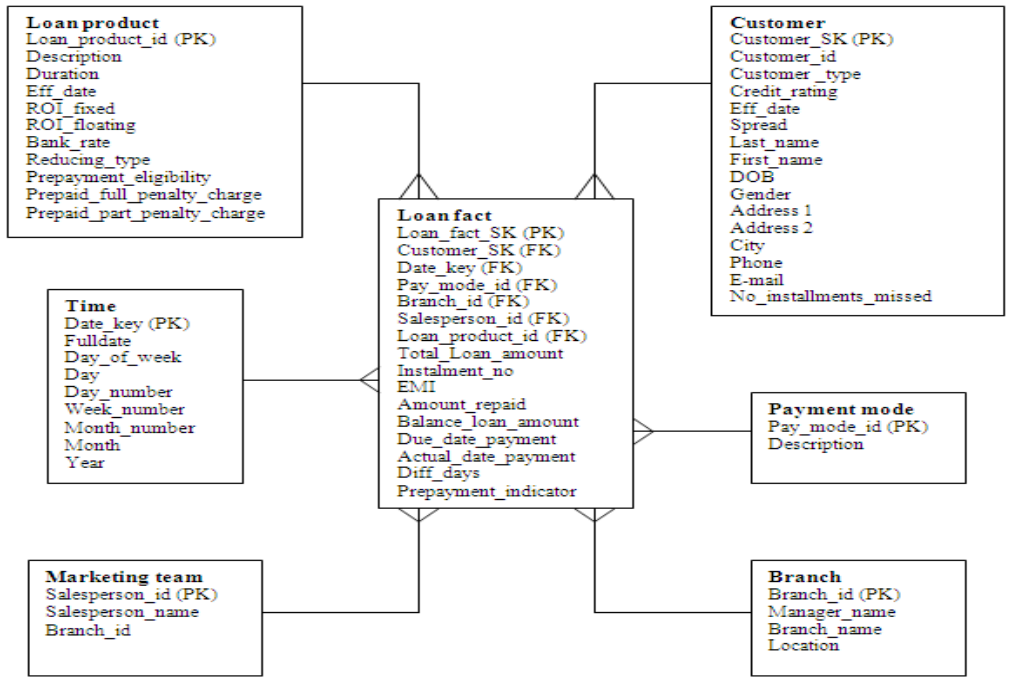

Figure-4: Multidimensional model to hold Bank data: Star Schema

The exercised data mart consist of six dimension tables namely loan product, customer, time, payment mode, time, marketing team and branch and one fact table loan fact and all these tables have be arranged in start style. The transactional data present in legacy system is loaded into data mart on the daily basis. To ensure the data quality of the data mart is challenge and tedious task, the proposed KPI (Key Performance Evaluation) method helps in ensuring the data quality. The exercised requirements are highlighted in mapping document in figure-3 and multidimensional model in figure-4

4.5 Analytical Model for Data Quality Assessment

Dimension	Table name	Fields
Accuracy	Loan	Spread
	Repayment	Closing Principal
	Repayment	Interest
Derivation integrity	loan	ROI
	Repayment	Delayed_Payment_Days
	Repayment	Preclosure_Penalty
Validity	Customer	DOB
	Customer	Customer Type
	Loan	Loan_product_ID
Completeness	Loan	Customer_ID
	Loan	Equated_Monthly_Installment
Non-Duplication	Customer	First_Name, Last_Name, DOB

Figure-5: Data Quality Analysis with respect to different KPI's

Accuracy: Is a measure of the degree to which data agrees with data contained in an original source.

$$Accuracy = \frac{Sum\ of\ All\ \ E(i)\ (\ Where\ i = 1,2,3\ \dots\dots n)}{Total\ number\ of\ entities\ in\ the\ source\ database}$$

$$Accuracy\ of\ an\ Entity(E_i) = \frac{Number\ of\ Accurate\ fields}{Total\ number\ of\ fields\ in\ the\ entity}$$

$$Accuracy\ of\ each\ Field(A_i) = \frac{Number\ of\ values\ of\ a\ field\ without\ error\ in\ the\ target\ database}{Total\ number\ of\ values\ in\ a\ field\ of\ the\ source\ database}$$

Derivation integrity: Is the correctness with which two or more pieces of data are combined to create new data.

$$Deviation\ from\ Derivation\ Integrity\ in\ the\ field = \frac{Number\ of\ error\ records}{Total\ number\ of\ records}$$

$$Derivation\ Integrity\ = 1 - Average(Deviation\ in\ integrity\ of\ each\ field)$$

One should be very cautious, when computing the Derivation Integrity, which would not be literally calculated by one single formula, but should be considered under different circumstances, conditions based on the database instances.

Validity: Is a measure of degree of conformance of data values to its domain and business rules.

$$Validity\ = \frac{Total\ Number\ of\ records\ in\ the\ target\ database\ satisfying\ the\ business\ rules}{Total\ number\ of\ records\ in\ the\ source\ database}$$

Automated data assessments can test validity and reasonability, but they will not be able to assess accuracy of the values

Completeness: Is the characteristic of having all required values for the data fields i.e. completeness can be measured by taking a ratio of the number of incomplete items to the total number of items and subtracting from 1

$$Validity\ = 1 - \frac{Number\ of\ incomplete\ items}{Total\ number\ of\ items}$$

Timeliness: Is the relative availability of data to support a given process within the timetable required to perform the process. Timeliness is measured as a maximum of one of the two terms: 0 and one minus the ratio of currency to volatility.

$$Timeliness = 1 - \frac{Currency}{Volatilty}$$

Where, Currency is the age plus delivery time minus the input time. Volatility refers to the length of time the data remains valid. Delivery time refers to when the data is delivered to the user.

Consistency: It can be measured by a ratio of violations of a specific consistency type to the total number of consistency checks subtracted from one.

$$Consistency = 1 - \frac{Violation\ of\ a\ specific\ consistency\ type}{Total\ number\ of\ consistency\ checks}$$

V. RESULTS AND DISCUSSIONS

The proposed methods have greater data quality assessment with respect to the different KPI's for different data sets. Target database is classified for decision making which meets customer expectation and confidence to rely on the underlying data. The data is validated with different characteristics such as, which are the exact data (The data user need), with the right completeness (All the data user need), in the right context (Whose meaning user know), with the right accuracy (User can trust and rely on it), in the right format (User can use it easily), at the right time (When user need it), at the right place (Where user need it), for the right purpose (User can accomplish business objectives).Figure-6 and 7 shows the relationship between existing and proposed system.

DQ Dimension	Proposed	Existing
Completeness	100	75.2
Business Conformance	100	78.7
Non Duplicates	100	94.825
Accuracy Field	100	100
Accuracy Entity	66.66	66.66
Derivation Integrity	100	67.7
Consistency	100	98

Figure-6: KPI's between Proposed and Existing Method

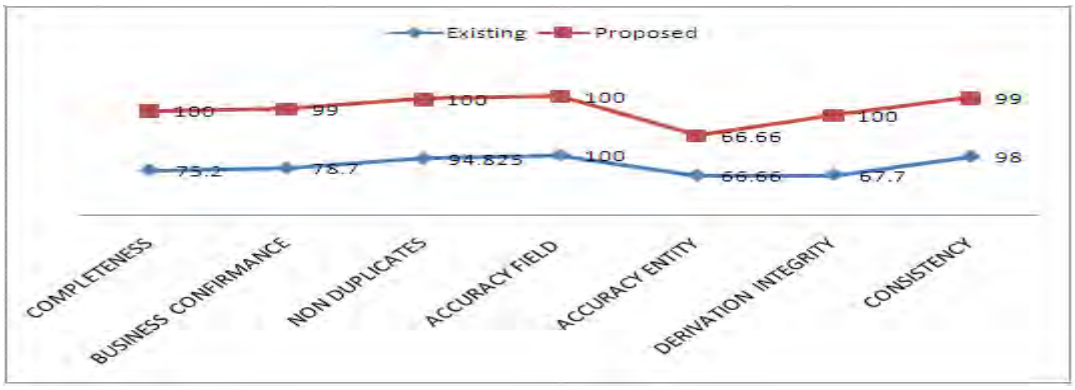

Figure-7: Relationship between existing and proposed method

To ascertain successful and qualitative loading of data from the source database | legacy system to the target database the following additional verification should be conducted.

Summary Tests: These include checking for total number records based on certain key columns or while considering rejections. Similarly, summary tests can be conducted by applying suitable filters.

Total Values Tests: These include checking for SUM (or related values) as applicable to key numeric columns.

Audit and Control Log Verification: This helps to get an idea about the load at the summary level.

Rejection Log Verification: This helps to get an idea on expected data quality issues

5.1 Classification with Decision tree Method

Validated the data quality of the decision / target database using decision tree method: For a given a collection of records (training set), each record contains a set of attributes, and one of the attributes is the class. Find a model for class attribute as a function of the values of other attributes.

Goal: Previously unseen records should be assigned a class as accurately as possible.

A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Decision tree is a flow-chart-like tree structure, internal node denotes a test on an attribute, Branch represents an outcome of the test and Leaf nodes represent class labels or class distribution. At start, all the training examples are at the root and Partition examples recursively based on selected attributes. The below are the training data set and test data set with different databases of loan repayment data mart. Figure-8 represents training set for different KPI's and figure-9 shows the test set for the different databases with respect to KPI's.Figure-10 represents the decision tree construction for data quality classification with respect to KPI's.

Training Data set		
Dimension	**Value**	**Class**
Completeness	80 to 100	**Yes**
Business Conformance	80 to 100	
Non Duplicates	100	
Accuracy Field	90 to 100	
Accuracy Entity	80 to 100	
Derivation Integrity	80 to 100	
Consistency	80 to 100	

Figure-8: Training set for different KPI's

Test Data Set			
Data Sets	**Dimension**	**Value**	**Class**
Oracle	Completeness	98	Yes
	Business Conformance	98	
	Non Duplicates	100	
	Accuracy Field	100	
	Accuracy Entity	66.66	
	Derivation Integrity	100	
	Consistency	100	
SQL Server	Completeness	77	No
	Business Conformance	78	
	Non Duplicates	84	
	Accuracy Field	100	
	Accuracy Entity	85	
	Derivation Integrity	76	
	Consistency	80	
Flat Files	Completeness	97	Yes
	Business Conformance	96	
	Non Duplicates	77	
	Accuracy Field	100	
	Accuracy Entity	100	
	Derivation Integrity	98	
	Consistency	100	
DB2	Completeness	100	Yes
	Business Conformance	100	
	Non Duplicates	100	
	Accuracy Field	98	
	Accuracy Entity	77	
	Derivation Integrity	96	
	Consistency	98	

Figure-9: Test Set for Different KPI's and dataset

Decision Tree Construction for different test set based on the training set

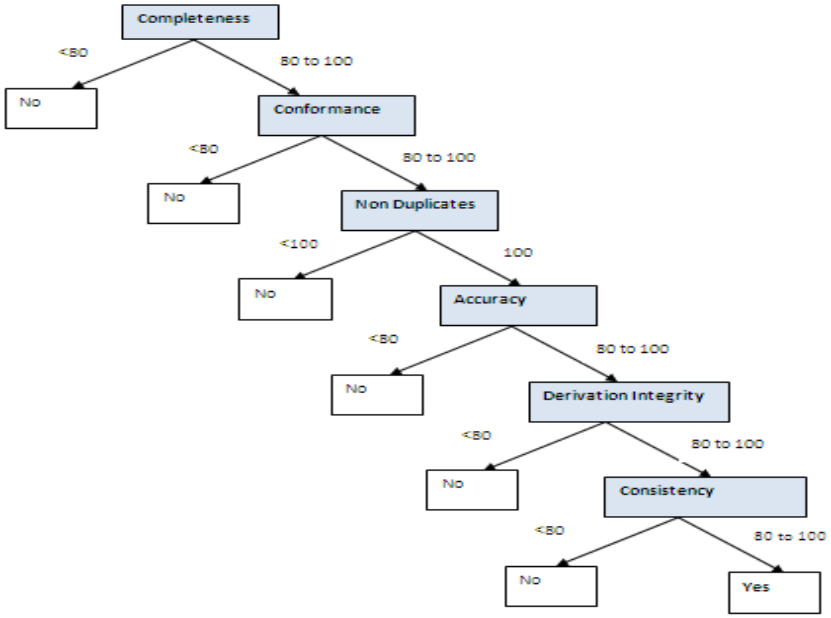

Figure-10: Test Set for Different KPI's and dataset

VI. CONCLUSIONS

In current information technology, process improvement is vital part of any businesses, in this regard data migration is essential from legacy systems to newer system or data warehouses for their business decisions, in this connection there is an essence of developing data quality assessment model to assess the underlying data in the decision databases. The Proposed data quality assessment model evaluates the data at different dimensions to give confidence for the end users to rely on their businesses. Author extended to classify various data sets suitable for decision making. The results reveal the proposed model is performing an average of 12.8 percent of improvement in evaluation criteria dimensions with respect to selected case study.

REFERENCES

[1] Alex Berson and Larry Dobov [2007]. "Master Data Management and Customer Data Integration for a Global Enterprise", Tata McGraw-Hill Publishing Company Limited.
[2] Allen Dreibelbis, Eberhard Hechler, Ivan Milman, Martin Oberhofer, Paul van Run, Dan Wolfson [2008]. "Enterprise Master Data Management: An SOA Approach to Managing Core Information", Dorling Kindersley (India) Pvt. Ltd.
[3] Jack E. Olson [2003]. "Data Quality: The Accuracy Dimension", Elsevier.
[4] Ralph Kimball and Joe Caserta [2004]. "The Data Warehouse ETL Toolkit", Wiley Publishing, Inc.
[5] Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications - Batini, Scannapieco - 2006
[6] A Data Quality Management Maturity Model, Kyung-Seok Ryu, Joo-Seok Park, and Jae-Hong Park, ETRI Journal, Volume 28, Number 2, April 2006
[7] Manjunath T.N., Ravindra S. Hegadi, Ravikumar G.K."Analysis of Data Quality Aspects in Data Warehouse Systems". (IJCSIT) International Journal of Computer Science and Information Technologies, 2 (1), 2011, 477-485.
[8] Manjunath T.N., Ravindra S. Hegadi, RaviKumar G.K.,"Design and Analysis of DWH and BI in Education Domain",IJCSI International Journal of Computer Science Issues, 8(2),March 2011 ISSN (Online): 1694-0814.545-551.
[9] Manjunath T.N., Ravindra S. Hegadi and Mohan H.S.Article: "Automated Data Validation for Data Migration Security". International Journal of Computer Applications, 30(6), 41-46, September 2011. Published by Foundation of Computer Science, New York.
[10] G. John and P. Langley, "Static versus Dynamic Sampling for Data Mining", Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 367-370, AAAI Press, Menlo Park, CA, 1996.
[11] Microsoft® CRM Data Migration Framework White Paper by Parul Manek, Program Manager Published: April 2003.
[12] "Data Migration Best Practices NetApp Global Services", January 2006.
[13] Liew, C. K., Choi, U. J., and Liew, C. J. 1985. "A Data Distortion by Probability Distribution,"ACM Transactions on Database Systems (10:3), pp.395-411.
[14] Xiao-Bai Li, Luvai Motiwalla BY "Protecting Patient Privacy with Data Masking" WISP 2009.
[15] Domingo-Ferrer J., and Mateo-Sanz, J. M. 2002. "Practical Data-Oriented Microaggregation for Statistical Disclosure Control," IEEE Transactions on Knowledge and Data Engineering (14:1), pp. 189-201.
[16] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, and S. Paraboschi. Designing data marts for data warehouses. ACM Transactions on Software Engineering Methodologies, 10(4):452{483, 2001}.