International Journal of Soft Computing 8 (5): 340-351, 2013
ISSN: 1816-9503


# Statistical Data Quality Model for Data Migration Business Enterprise

[1]T.N. Manjunath and [2]Ravindra S. Hegadi
[1]Bharathiar University, Coimbatore, Tamil Nadu, India
[2]Department of Computer Application, Solapur University, Solapur, Maharastra, India

**Abstract:** In current information trends, state of decision making is one of the important deeds for any organization or enterprise to identify them in the business market in this connection, the data which is present in data warehouse or decision databases should be very accurate and help them to give proper decisions. When the organizations or enterprises undergo a merger/takeover demands the data migration from legacy systems to modern systems/decision databases, i.e., target systems. If a target/decision databases is very large to ensure quality assurance of the decision database is tedious. The resource utilization required to conduct full data verification is exorbitant. This research proposes a mathematical model using deterministic statistical methods to reduce resource utilization and assures greater data quality. The proposed method validated using various data sets and volumes against man effort, CPU time, defects raised and cost. It also ensures comfortable confidence for end users to rely on the data quality for decision making.

**Key words:** Data quality, statistical methods, data migration, business enterprise, CPU time

## INTRODUCTION

New business models, constant technological progress as well as ever-changing legal regulations require that firms replace their business applications from time to time due to compatibility aspects. As a side effect, this demands for migrating the data from the existing source system to a target system as depicted in Fig. 1. Since, the success of the data migration as a form of IT maintenance, it is crucial to accomplish the migration in time and on budget. This however, calls for a stringent data migration process model combined with well-defined data quality assurance measures. Data migration is the process of transforming legacy data to a new system. This can involve moving disk files from one location to another. Legacy data means the data which is ready to migrate from current storage to another place and current storage include database records, spreadsheets, text files, scanned images and paper documents. All these data formats can be migrated to a new system (Manjunath *et al*., 2010, 2011). If a source systems containing millions of data has to migrate towards target/decision databases to ensure quality assurance of the decision database. To achieve high data quality decision database is complex in nature. It demands a lot of effort due to high volume of data transformations vary from simple to complex in nature. Various kinds of rejections and corrections as the target database schema and design basis is different and does not want to have

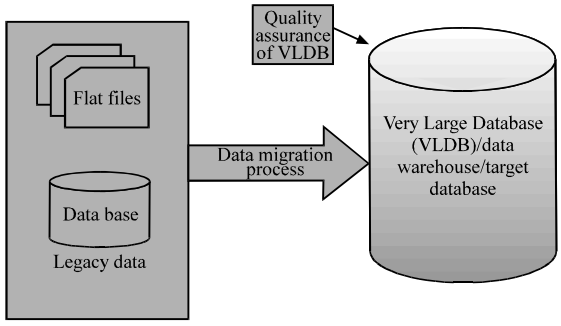


Fig. 1: Data migration process to VLDB

data problems of source databases and various kinds of reconciliations required to synchronize data between source and the decision databases.

After the process of data migration, there is an essence of performing quality assurance of the target system/VLDB since business decisions are rely on the target decision database, it is tedious task to validate all the records in the target system/VLDB (TDWI, 2006). To move massive data from legacy system to decision database or business intelligence applications, some specialists argue that automated algorithms with faster processing times justify processing the entire database (Potts, 1997; Manjunath *et al*., 2011). However, no single approach solves all post data migration quality assurance problems. Instead, processing the entire database offers both advantages and disadvantages depending on the data domain. Drawback of processing very large

**Corresponding Author:** T.N. Manjunath, Bharathiar University, Coimbatore, Tamil Nadu, India

databases affects various aspects of the data domain process including the points. Inference, quality of the findings, speed and efficiency.

Firms that have achieved significant Return On Investment (ROI) in data migration applications have done so by performing data analysis on business domains. To ensure the quality assurance on the decision database typically using sampling methods have major advantages. ROI is the final justification for data migration and most often, the return begins with a relatively small sample. Exploring a representative sample is easier, more efficient and can be as accurate as exploring the entire database. After the initial sample is explored, various preliminary models can be fitted and assessed (English, 1999). However, it is likely that the initial modelling generates additional, more specific questions and more data exploration is required. Advantages of the processing a sample database are speed and efficiency, visualization, generalization and economy.

**Benefits of data quality in business enterprise domain:** Providing the high quality data for end user gives more confidence in making appropriate decision, benefits of data quality in business enterprise domains are (Fields *et al*., 1986).

**Lower costs:** Data quality helps improvement in pervasive, drastically reducing costs by simplifying the process for identifying all data quality issues, measuring all data quality levels and cleansing data in all applications for all data domains.

**Reduce risks:** Data quality helps firms to identify, resolve and prevent data quality problems, ensuring that enterprise data is trusted and authoritative, so you can address compliance needs with confidence, maintain competitive advantage and avoid losing customers proactive monitoring and enforcement of data quality, consistently and in all applications across the globe, help enable data governance, increasing confidence in making decisions about business processes and operations.

**Drive efficiency:** Business and IT collaborate more efficiently to complete the defined task in days rather than months. Business analysts and data supervisor can manage data quality tasks themselves. Line of business managers can access authoritative and trustworthy information at the point of use within the applications they use day to day (Ravikumar *et al*., 2011).

**Improve productivity:** Data quality allows IT organizations to access all data quickly for faster decisions. Agile data architecture delivers reusable and compliant data services, allowing IT to be more responsive to the needs of the business. Data tasks are accelerated by building data quality mappings within a familiar development environment that provides data profiling and prebuilt rules for matching and address cleansing, enabling quick validation of data quality transformations using midstream profiling (Fields *et al*., 1986; English, 1999). Data quality assurance is to understand the state of data quality of the decision database. In this context, a small and random sample of records can provide an accurate image of the data quality (Firth, 1996; Manjunath *et al*., 2010). By adopting deterministic statistical sampling techniques and statistics parameters can be used to achieve data quality. As per Statistical Quality Control (SQC) for measuring performance, identifying unacceptable variance and corrective actions and Statistical Process Control (SPC) minimum data samples required are 50 in numbers. Researcher presents, first, a field data quality process model for data migration business enterprise to reduce effort, cost with greater data quality. In this connection, there is no model which suggests performing data validation which reduces cost, effort with high quality of data for end-users to rely on the data after migrations for decision making.

The proposed mathematical model gives comfortable confidence to the end user to rely on the data quality for business decisions. The model emphasizes on the usage of deterministic sampling methods to draw quality and size of samples to ensure required confidence on the data quality in turn reduction in effort and cost of testing cycle time, it also ensures high data quality and further extended to validate the model using statistical central limit theorem. This process is useful for the data testing in any business domain which gives accurate decisions for the external world and management as well. Proposed model can be generalized for any application which is undergoing migrations from old storage devices until building data warehouses/VLDB which will handle heterogeneous data formats after Extract Transform and Loading (ETL) process.

**Literature review:** A recent survey done by Gartner in 2011 on poor data quality reveals organizations lose $8.2 million annually through poor data quality (Friedman and Smith, 2011). Furthermore, 140 companies surveyed and presented 22% estimated their annual losses resulting from bad data at $20 million. Much of this loss is due to lost productivity among workers who realizing their data is incorrect is forced to compensate for the inaccuracies

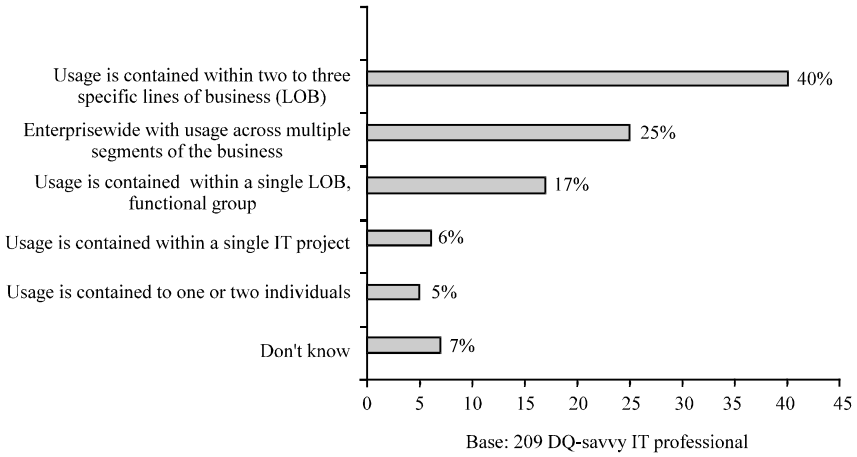


Fig. 2: Data quality importance in different business enterprise domains: A survey by Global Data Quality Online Survey in 2010 (March 2010 Global Data Quality Wave Customer Forrester Reference Online Survey and August 2010 Global Data Quality Online Survey)

or create workarounds when using both operational and analytic applications. Reducing costs and risks for data migrations by Patrick Allaire, Justin Augat, Joe Jose and David Merrill hitachi data systems in 2010 (Allaire *et al*., 2012). Emphasized organization need to innovate in data migration data quality, it is intended to provide IT decision makers with information on costs, risks and considerations regarding the migration of data from old storage platforms to new storage to reduce this significant area of operating expenses. In 2007 Bloor research surveyed, 84% of projects were running over time with budget, 70% data migration domains applied data cleansing methods to refine the data. A data quality Management Maturity Model by Ryu *et al*. (2006) showed empirically that data quality improves as data management matures (Howard, 2011). In addition, 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. Another ramework is based on semantics to evaluate the quality of the form, meaning and use of the data (Price and Shanks, 2004). In Arizona State University during 2004 in data quality workshop, father of data quality Larry English highlighted. The business costs of non-quality data including irrecoverable costs, rework of products or services, workarounds and lost and missed revenue may be as high as 10-25% of revenue of total budget of an organization which is very high. A considerable amount of data quality research involves investigating and describing various categories of desirable attributes (or dimensions) of data. In 2002, the USPS and Price waterhouse coopers released a report stating that 23.6% of all US mail sent is incorrectly addressed (TDWI, 2006). Data reduction via adaptive sampling by XIAO-BAI in 2002 (data reduction is an important issue in the field of data mining and data migration applications (Li, 2002). One framework seeks to integrate the product perspective (conformance to specifications) and the service perspective (meeting consumers' expectations) (Kahn *et al*., 2002). One industry study estimated the total cost to the US economy of data quality problems at over US$600 billion per annum (Eckerson, 2002). Incorrect data which includes invalid and outdated information can originate from different data sources through data entry or data migration and conversion assignment. Larry English tell management that the best costs to cut are the costs and wastes of resources caused by poor quality information (English, 1999). Telco clients saw their entire IQ team laid off during the Telco overcapacity crisis a few years ago-even after the team had discovered they were not invoicing some $50 million of services (Manek, 2003). The 3 years later, the leader of the IQ team saw me at an IQ conference and told me that they were all brought back in. Within a year, the team had recovered another $8-9 million per year by improving and error-proof the processes to prevent future failure. Figure 2 shows a survey done by global data quality in 2010 and highlighted the importance of data quality in business domains. No literature found on the developing data quality model using deterministic statistical methods for data migration business enterprise.

**Preliminaries (Statistics for data quality assurance):** Probability theory and statistical theory are employed to guide data quality assurance method for data testing. In most of the business and data analysis research, sampling is widely used for gathering information about a population. One fundamental aspect of study design for

data quality assurance model and its concepts in support of best practices, methods, most popular sampling techniques and sample size determination using Chi-Square Method and its advantages for survey system research. Factors such as time, cost and how many resources are actually available are constraints that have taken in to account when designing a study (Fields *et al*., 1986; Badri *et al*., 1995).

**Sampling as a best practice in data migration quality assurance:** SAS Institute Inc. (1998) pointed out, extracting data from a database for the purpose of data mining and migration is based on the sampling techniques routinely used in surveys. This is similar to the way in which statistical sampling traditionally has been performed on large populations of observations. When appraisers acquire information from a general population, they sample only enough population to get a good approximation. One need not have to identify every single occurrence of a pattern within a data set in order to infer that the pattern exists. Once you lock onto a pattern you can get a feel for how extensive the pattern is throughout the entire data set through alternative analytical methods. One should not feel to process the entire data set at one time. There will usually be more than enough results from the clustered data. The study addresses several common apprehensions about how best to use sampling in data migration business domains (Allaire *et al*., 2012; Gupta, 1997). In particular, this study addresses the following questions:

- What needs to be done to prepare data for sampling?
- How should the size of the sample to be determined?
- What are the common types of sampling methods applied for data selection?
- Should multiple samples to be taken to preserve data characteristics?
- Benefits of using sampling for quality assurance for business enterprise?

**Preparing the data for sampling:** Prior to sampling data and analyzing business problems, most of the businesses will allow the data transformation before loading into decision databases/data warehouse. Business users from various departments will expect to access the data they need quickly and easily. Decision databases/VLDB/Data warehouses enable many groups to access the data, facilitate updating the data and improve efficiency of checking the data for reliability and preparing the data for analysis and reporting (Allaire *et al*., 2012; Ryu *et al*., 2006). For example, if the business problem is to profile customers then all of the data for a single customer should be contained in a single record. If you have data that describes a customer in multiple records then you could use the decision database/VLDB/data warehouse to rearrange the data, prior to sampling.

**Determining the sample size:** Determining the appropriate size of a sample given a particular database table, formulated the formulae using the statistics theory. The formulas are designed to help the beneficiary to select the optimal sample size by addressing questions as follows:

- How to point out the error variable in each column of a database table?
- Which records should be in the model?
- What is the functional form of the model?
- What is an acceptable level of accuracy for the results?

If the answers to these questions are known then sampling theory may be able to provide a reasonably good answer to the required sample size. The less confidence you have in the answers to these questions, the more you are into exploration of the data and iterating through the proposed model (Manjunath *et al*., 2010; Allaire *et al*., 2012).

**Common types of sampling:** Sampling in combination with data engineering research is one of the most popular approaches to data collection and analysis in engineering stream (Han and Kamber, 2006). The different types of sampling techniques are given.

**Random sampling:** Statistical sampling method in which every record within the target database has an equal likelihood of being selected with equal probability. Use random-number generator in an extract tool or data analysis tool or query tool. Determine the required DSS (Data Sample Size) and number of records in the target database. Program the random-number generator to calculate a number between 1 and the total number of records in the target database. Select records where the number generated is less than or equal to the number of records required for the sample plus 1 or 2.

**Systematic sampling:** Sampling in which every nth record is selected. Select a ratio based on the ratio of required DSS (Data Sample Size) to total number of database records. Randomly select the first record to be chosen based on the ratio used. Systematic sampling is appropriate when the data population is ordered in a truly

random sequence and there is no bias in its ordering. Use this approach when random sampling of the desired database is not feasible.

**Stratified sampling:** Sampling a database that has two or more distinct groupings or strata in which random samples are taken from each stratum to assure the strata are proportionately represented in the final sample. Use stratified sampling when the database being sampled is a distribution in records such that a small number of records exist for one subtype.

**Cluster sampling:** Sampling a database by taking samples from a smaller number of subgroups (such as geographic areas) of the population. The sub-samples from each cluster are combined to make up the final sample. For example, in sampling sales data for a chain of stores, one may choose to take a sub-sample of a representative subset of stores (each cluster) into a cluster sample, rather than randomly selecting sales data from every store. This technique may be used to represent all retail sales data only when the clusters truly represent the same relative kinds of data and the same relative process consistency.

**Two-stage sampling:** Sampling from multiple files/database of the same data type and then conducting a sample from the combined subset of group of data. This technique is used when collecting data randomly from distributed data files/databases or from different time periods and then randomly selecting a final sample from the merged samples. This is appropriate when there are many different files/databases to sample and the size of the merged samples is larger than necessary for assessment. The first stage of sampling assures an adequate representation of data from each file/database or time period. Each sample of the first stage should be proportionate to its subpopulation (of records) so on group inordinately biases the final sample. The second stage of sampling assures adequate representation of the combined samples to represent the complete distribution of data.

**Inference from random samples:** It tells the distribution of data when researchers randomly select records from a database tables and obtain a sample of sufficient size, the sample reflects characteristics of the database as a whole. The sample size gets larger, the distribution (shape) of the sample reflects the distribution of the entire database. The data have a symmetric and bell-shaped distribution. This is a commonly occurring and well-understood type of shape and is referred to as a normal distribution. Not all

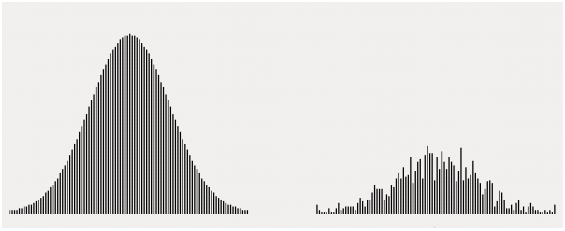

Fig. 3: Samples reveals data characteristics

data are distributed normally. Figure 3 shows the representations of databases often reveal that the data are skewed in one direction or the other (SAS Institute Inc., 1998; Knight and Burn, 2005).

**Chi-Square Method:** The analysis of data normally involves the fraction of acceptable values in a given population. This may consist of estimating a single proportion. A simplified chi-square criterion is proposed for measuring the goodness-of-fit between the distributions of the reduced and full data sets. Given a large dataset, the objective is to find a reduced dataset whose frequency distribution is as close to the true distribution as possible where the true distribution refers to the frequency distribution of the original dataset. The size of the reduced set is expected to be substantially smaller than that of the full set. A classical measure of the closeness of (or distance between) the actual and expected distributions is the chi-square goodness of fit statistic, given by:

$$\chi^2 = \sum \frac{(n_i - m_i)^2}{m_i} \quad (1)$$

Where:

$n_i$ = The frequency of the actual (or observed) distribution

$m_i$ = The frequency of the expected distribution and subscript i runs over all possible category (record) combinations

When the data values are of the continuous type, they are grouped into certain intervals before applying Eq. 1. The statistic $\chi^2$ follows asymptotically a $\chi^2$ distribution with appropriate degrees of freedom. To apply Eq. 1 to the data reduction problem, let N and n be the sizes of the original and reduced datasets, respectively and call p = n/N the sampling proportion. Let J be the number of attributes and Kj (j = 1, ..., J) be the number of categories (intervals) in the jth attribute. Call each category combination a sample. Then, the total number of sample can be as $\prod_{j=1}^{J}$ Kj large as the actual number of samples, C may be smaller since some of the

samples may not appear in the data. Let $N_i$ and $n_i$ (i = 1, ..., C) be the frequencies of the ith pattern in the original and reduced datasets, respectively. Then, $\chi^2$ can be computed by:

$$\chi^2 = \sum_{i=1}^{c} \frac{(n_i - pn_i)^2}{pn_i} \quad (2)$$

Now, it appears that the data reduction problem described at the beginning of this study can be translated to finding a reduced dataset that minimizes the $\chi^2$ value in Eq. 2.

**Confidence interval:** Given a simple random sample of size n from a population, the number of "True's" X divided by the sample size n provides the sample proportion $\hat{p}$ an estimate of the population proportion p. This proportion follows a binomial distribution with mean p and variance (p(1-p))/n. Since, the binomial distribution is approximately normal for large sample sizes, tests of significance and confidence intervals for a single proportion use a z statistic (Allaire *et al*., 2012; NetApp, 2006). To find a confidence interval for a proportion, estimate the standard deviation $s_p$ from the data by replacing the unknown value p with the sample proportion $\hat{p}$ giving the $S_p$:

$$\text{Standard error } s_p = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (3)$$

An approximate level C confidence interval for p is:

$$\hat{p} \pm Z^{\oplus} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (4)$$

where, $z^*$ is the upper (1-C)/2 critical values from the standard normal distribution. To test the null hypothesis $H_o$: $p = p_o$ against a one or two sided alternative hypothesis Ha, replace p with $p_o$ in the test statistic:

$$Z = \frac{p - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \quad (5)$$

The test statistic follows the standard normal distribution (with mean = 0 and standard deviation = 1). The test statistic z is used to compute the p-value for the standard normal distribution, the probability that a value at least as extreme as the test statistic would be observed under the null hypothesis. Given the null hypothesis that the population proportion p is equal to a given value $p_o$, the p-values for testing $H_o$ against each of the possible alternative hypotheses are:

- P (Z>z) for $H_a$: $p>p_o$
- P (Z<z) for $H_a$: $p<p_o$
- 2P (Z>|z|) for $H_a$: $p\neq p_o$

**Sample size:** An increase in sample size will decrease the length of the confidence interval without reducing the level of confidence. This is because the standard deviation decreases as n increases (NetApp, 2006). The margin of error m of a confidence interval is defined to be the value added or subtracted from the sample proportion which determines the length of the interval:

$$M = Z^{\oplus} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (6)$$

Given a guessed value $p^*$ for the proportion p, substitute $p^*$ for p to calculate m. Solving for n gives the expression $n = (z^*/m)^2 p^*(1-p^*)$. The margin of error is maximized when $p^* = 0.5$ in which case $n = (z^*/2m)^2$.

**Central limit theorem:** Statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population. Furthermore, all of the samples will follow an approximate normal distribution pattern with all variances being approximately equal to the variance of the population divided by each sample's size as depicted in Fig. 4.

**Statistical Data Quality Model for data migration:** Data Quality Model for data testing after data migration business enterprise, data migration logic involves 1:1 column mapping and derived column mapping varies from

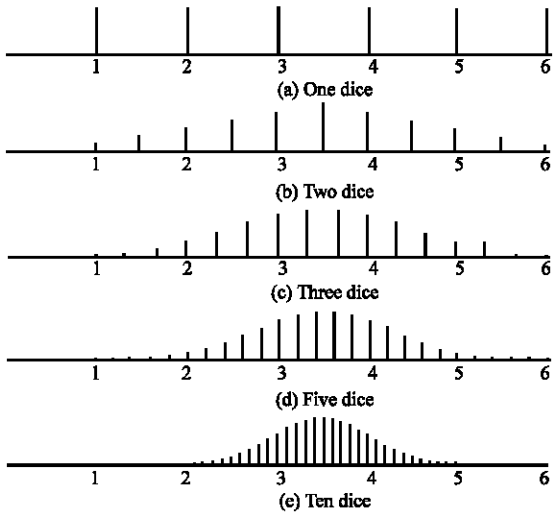


Fig. 4: Bell shaped curve with different samples

simple to complex logic which involves composite lookups and transformations (Miles and Huberman, 1994). To build the Data Quality (DQ) Model, initially as per Statistical Quality Control (SQC) 50 samples are selected from target system/decision system. An error in the sample system is recognized by comparing with the source system using SQL scripts and key concept. Equation 7 represents the Error field (E) in the target database/decision system:

$$\text{Error field}(E) = \sum_{i=1}^{m}\sum_{j=1}^{n} E_{ij} \quad (7)$$

Where:

$E_{ij}$ = The error field in the target database table after migration

m, n = The number of rows and columns in the database table

$$\text{No. of records}(n) = \sum_{i=1}^{n} n_i \quad (8)$$

where, n is the total sample records selected represented in Eq. 8 and i varies from 1 to n:

$$\text{Standard deviation}(S) = \sqrt{\frac{\sum d \times d}{n\text{-}1}} \quad (9)$$

S is the standard deviation of the error field in the sample records, represented in Eq. 9:

$$\text{Sample size}(N) = \left(\frac{Z_c \times S}{B}\right)^2 \quad (10)$$

Where:

$Z_c$ = Confidence constant

S = Standard deviation of the error represented in Eq. 10

B = Bound level

The sample size N is selected from entire population using sampling viz. simple random samples, cluster, two-stage, systematic, stratified sampling methods. To select n sample records out of N records such that each record has an equal chance of being selected, Eq. 11 represents the same:

$$\text{Simple random sample}(n_i) = NC_n \quad (11)$$

Where:

N = Total population

C = The combinations

n = The sample records

To select the sample records using Cluster Method is represented by Eq. 12 and 13:

$$\text{Cluster}(C_i) = \frac{N}{R_c} \quad (12)$$

Where:

N = The total population

$R_c$ = The required number of clusters

$C_i$ = The clusters of entire population

$$\text{cluster sample record}(csr_i) = \text{SRS}(C_i) \quad (13)$$

Where:

SRS ($C_i$) = The simple random samples in each cluste

$csr_i$ = The cluster sample records

To select the data via systematic sampling method, first find the systematic interval number using Eq. 14 and 15:

$$\text{Systematic interval size}(k) = \frac{N}{n} \quad (14)$$

Where:

N = The total population

n = The interval number

k = The interval size

$$\text{Systematic Sample Records } (SSR_i) = \sum k \quad (15)$$

Where:

k = The sample record selected at kth interval size

$SSR_i$ = The Systematic Sample Records

To select the data via stratified sampling method, first divide the population into different non overlapping strata using Eq. 16 and 17:

$$\text{Strata}(N_i) = \sum_{i=1}^{n} N_i \quad (16)$$

where, $N_i$ is the non overlapping groups derived from total population:

$$\text{strata field sample records } (stsr_i) = \text{SRS}(N_i) \quad (17)$$

Then, select sample record randomly from each group $stsr_i$ is the stratified sample records. To select the data via two stage sampling method, first divide the population into two halves using Eq. 18 and randomly select the data using Eq. 19:

$$\text{Two stage numeral}(f) = \frac{N}{2} \quad (18)$$

Where:

f = The two stage numeral

N = The total population

$$\text{Two stage sample records } (Tssr_i) = \text{SRS}(f) \quad (19)$$

Where:
$Tssr_i$ = The Two stage sample records
SRS (f) = The Sampled Records in two Stages

After applying the above formulations for selecting the data from entire population then comparison between source database and target database is done using below algorithm.

**Algorithm compare attribute pairs:**
Begin algorithm
For each record in the target database TD (1,..., N)
For each attribute $A_{ij}$ in TD (i rows and j columns)
For each record in the source database SD (1,..., M)
For each attribute $B_{ij}$ in SD (i rows and j columns)
Compare the values in $A_{ij}$ and $B_{ij}$
If ($A_{ij} = B_{ij}$) then
Update the comparisons array as zero
Else update the comparisons array as one
End for
End for
End for
End for
Output the comparisons array
End algorithm

**Case on data sample size calculation:** Sample size calculation is very much required to select the subset of the population from entire population of data set, it is calculated based on the statistics theory standard error of the variable with the largest variance this is done using ISR (Initial Sample Records) for large datasets as per SQC minimum 50 samples has been drawn from entire population to draw any conclusion. To perform data quality of target system is tedious task and lot of man power and resources in terms of CPU utilization, cost and effort. Figure 5 illustrates 50 ISR from target database with error fields (Badri *et al*., 1995; Fields *et al*., 1986):

| Rec # | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | No. of errors | SD(d) = (x-X) | d×d |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Record1 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record2 | | | | | | | | | E | | | | | 1 | 0.5600 | 0.3136 |
| Record3 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record4 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record5 | | | | E | | | | E | | | | | | 2 | 1.5600 | 2.4336 |
| Record6 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record7 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record8 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record9 | | | E | | E | | | | | | E | | | 3 | 2.5600 | 6.5536 |
| Record10 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record11 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record12 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record13 | | E | | | | | E | | | | | | | 2 | 1.5600 | 2.4336 |
| Record14 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record15 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record16 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record17 | | E | | | | | | | E | | | | | 2 | 1.5600 | 0.4336 |
| Record18 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record19 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record20 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record21 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record22 | | | | | | E | | | | | | | | 1 | 0.5600 | 0.3136 |
| Record23 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record24 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record25 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record26 | | | E | | | | E | | | | E | | | 3 | 2.5600 | 6.5536 |
| Record27 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record28 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record29 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record30 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record31 | E | | E | | | | | E | | | | | E | 4 | 3.5600 | 12.6736 |
| Record32 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record33 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record34 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record35 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record36 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record37 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record38 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record39 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record40 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record41 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record42 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record43 | | E | | | | | E | | | | | | | 2 | 1.5600 | 2.4336 |
| Record44 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record45 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record46 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record47 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record48 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| Record49 | | | | | E | | | | | | E | | | 2 | 1.5600 | 2.4336 |
| Record50 | | | | | | | | | | | | | | 0 | 0.4400 | 0.1936 |
| | | | | | | | | | | | | | | 22 | | 46.32 |

Fig. 5: Initial sample records from target database with error fields

- Number of fields containing an error = 22
- n = Number of records in the data sample = 50
- Mean ($\bar{x}$) or average number of errors per record = 22/50 = 0.440

$$\text{Standard deviation (S)} = \sqrt{\frac{\sum d \times d}{(n-1)}} = 0.9723$$

Mean = 0.4400, let the bound be + or -10% of mean = B = 0.0440, standard deviation of the sample = S = 0.9723, desired confidence level = 95%, the z constant = 1.9600, data sample size = N = $((Z \times S)/B)^2$, 1875.7686, the actual sample size (rounded) = 1876.

Therefore, to have a 95% confidence level that the mean number of errors of the data population is within 10% of the mean number of errors of the sample, researchers would need to select 1876 records from the target database.

## MATERIALS AND METHODS

In this study, researcher proposes a new methodology to perform the data quality assurance via data reduction by the sampling methods for data migration business enterprises which gradually reduces effort, cost and reduces test cycle time in turn which reduces software development life cycle time and gives end users a comfortable confidence on the data quality.

Data quality has become very important for any business decisions across industries, so there is very much need in developing the quality assurance model for evaluation of the quality of the target system in less time and effort. Researchers have used properties of statistics to formulate the model using sampling techniques highlighted in study 4. The main objective of this research is to study and develop the data quality assurance model for data warehouse system and data migration business enterprises using sampling techniques to ensure comfortable confidence for the business users/decision makers when they are making critical decisions. Figure 6 depicts the overall proposed system with different components.

**Proposed algorithm:** After the data migration process, assurance of the data quality is very important for any business decisions, algorithm describes the data quality model using statistical theory, sampling methods and central limit theorem.

**Algorithm statistical data quality model:**

Begin algorithm
For each table in the Target Database TD (1,..., N)
 For each attribute $A_{ij}$ in the target database table
  Determine data sample size
  Extract data sample records using sampling methods from target database
  Central limit theorem validation for sample and full data volume
  Extract same data from source database SD (1,..., M) using keys
  Compare Attribute Pairs ( )
  If pattern NOT holds
  Then mark each attribute as possible error
  Else repeat for all the data files
  End for
  Compute time, man-hour, cost, defect and CPU response parameters for datasets
  Select and output the listed parameters
 End for
End algorithm

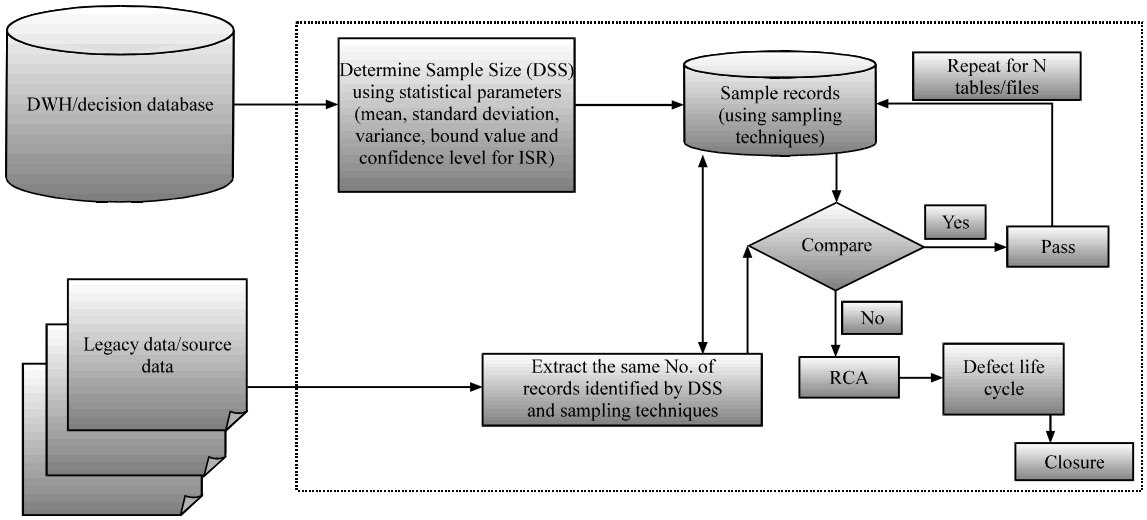


Fig. 6: Data quality model for data migration quality assurance

## RESULTS AND DISCUSSION

Experiments were conducted on different data volumes on engineered data sets. A set of data with 56 attributes and 100,000 records was arbitrarily generated. Each attribute had a known division and variety. Then, errors were simulated for analysis.

In Fig. 7, the X-axis represents data volume and Y-axis represents time for data validation (Table 1). For 10,000 data volume manpower required is 10 h for full validation where as to validate the sample data manpower required is 2 h, similarly for other data volumes.

In Fig. 8, the X-axis represents data volume and Y-axis represents defects raised while using sampling methods. For full data validation of 10,000 data volume defects raised were 5. Table 2 where as for sample data validation defects raised were 5. Experiments reveal full validation and sample validation represents same number of defects, data used for sample validation includes all types of underlying data.

Figure 9 represents the mapping of time value of money incurred on the man power supply. The X-axis represents data volume and Y-axis represents cost incurred on man power. To validate 10,000 volume, manpower cost incurred amounts to $248 where as to validate the data on sample basis the manpower cost incurred amounts to $50. The experiment reveals 80% cost savings on manpower by using sampling theory for testing (Table 3).

In Fig. 10, the X-axis represents data volume and Y-axis represents CPU execution time. For full data validation of 10,000 data volume CPU execution time is 36,000 sec where as for sample data validation CPU

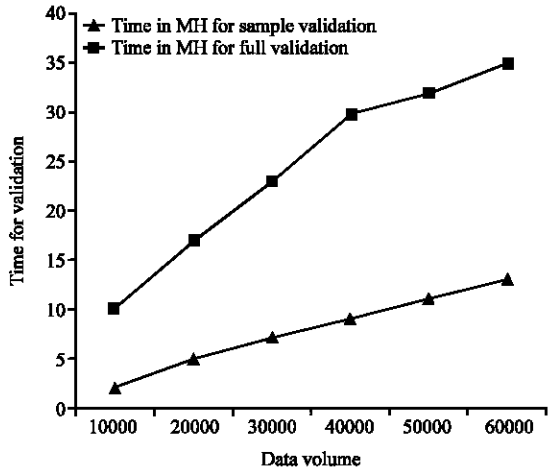


Fig. 7: Graph representing relationship between data volume and time

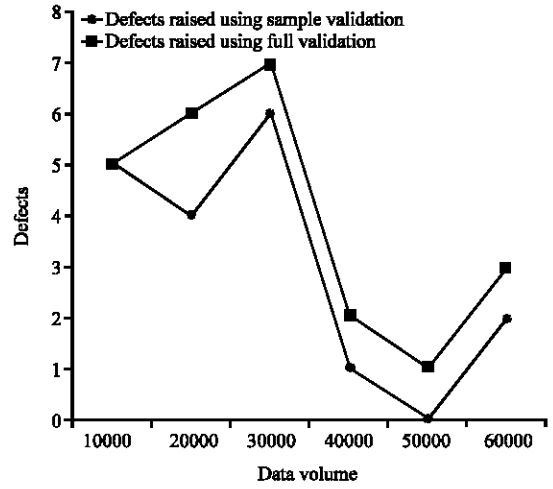


Fig. 8: Graph representing data volume vs. defects

Table 1: Data volume vs. time in MH

| Data volume | Time in MH (Man Hours) for full validation | Time in MH (Man Hours) for sample validation* |
|---|---|---|
| 10000 | 10 | 2 |
| 20000 | 17 | 5 |
| 30000 | 23 | 7 |
| 40000 | 30 | 9 |
| 50000 | 32 | 11 |
| 60000 | 35 | 13 |

*Sample validation-samples from large dataset based of sample size determination with 95% confidence and 10% bound value. Man Hours (MH) includes scripting, validation and defect monitoring

Table 2: Data volume vs. defects raised

| Data volume | Defects raised using full validation | Defects using sample validation |
|---|---|---|
| 10000 | 5 | 5 |
| 20000 | 6 | 4 |
| 30000 | 7 | 6 |
| 40000 | 2 | 1 |
| 50000 | 1 | 0 |
| 60000 | 3 | 2 |

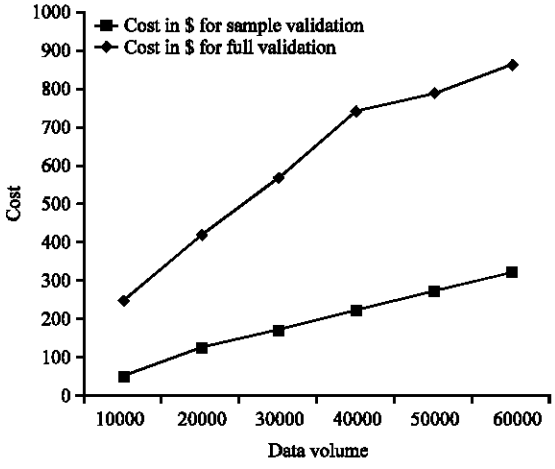


Fig. 9: Chart representing data volume vs. cost

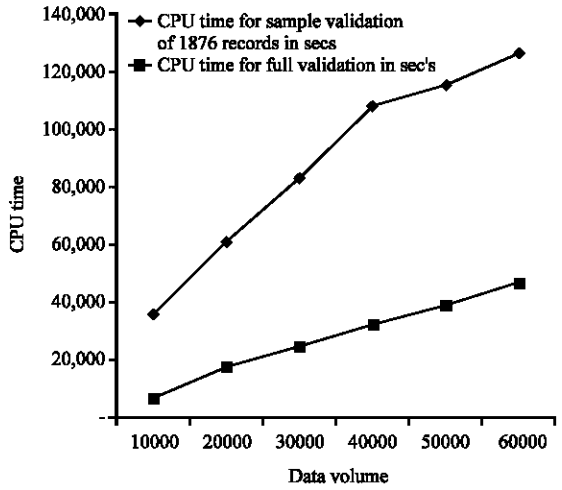


Fig. 10: Chart representing data volume vs. CPU seconds

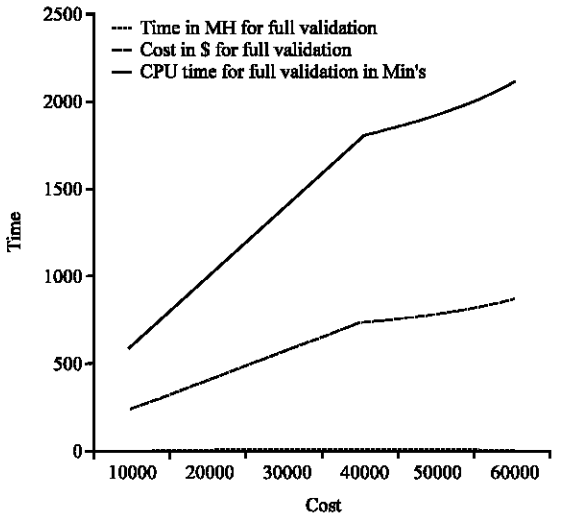


Fig. 11: Console of cost and time for full validation

Table 3: Data volume vs. cost in $

| Data volume | Cost in $ for full validation | Cost in $ for sample validation |
|---|---|---|
| 10000 | 247.50 | 49.50 |
| 20000 | 420.75 | 123.75 |
| 30000 | 569.25 | 173.25 |
| 40000 | 742.50 | 222.75 |
| 50000 | 792.00 | 272.25 |
| 60000 | 866.25 | 321.75 |

execution time is 7,200 sec. Experiments reveal there is 80% savings in time Table 4. In Fig. 11, the X-axis represents data volume and Y-axis represents CPU execution time, man-hours and cost incurred. The experiment reveals manpower and CPU execution time moves in line-up with cost.

In Fig. 12, the X-axis represents data volume and Y-axis represents CPU execution time, Man-hours and

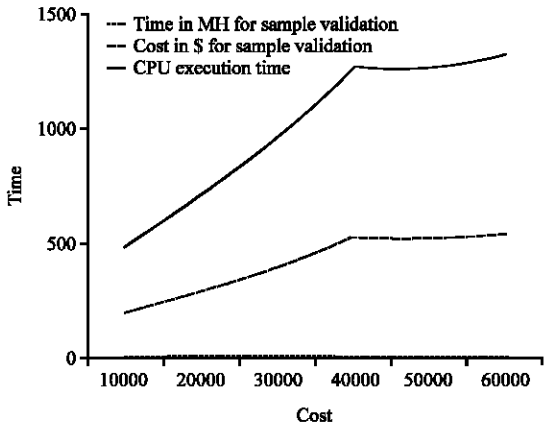


Fig. 12: Console of cost and time for sample validation

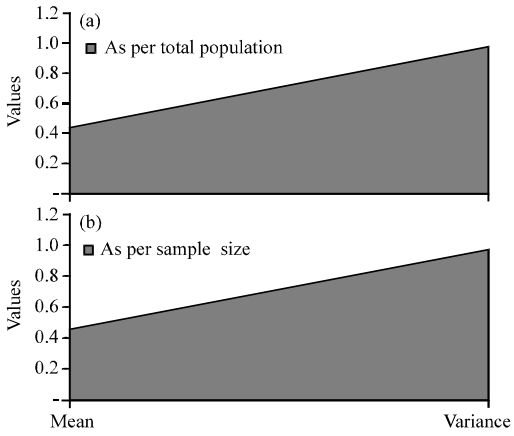


Fig. 13: Central limit theorem illustration on sample and entire population of data sets

Table 4: Data volume vs. CPU time

| Data volume | CPU time for full validation in sec's | CPU time for sample validation of 1876 records in secs |
|---|---|---|
| 10000 | 36000 | 7200 |
| 20000 | 61200 | 18000 |
| 30000 | 82800 | 25200 |
| 40000 | 108000 | 32400 |
| 50000 | 115200 | 39600 |
| 80000 | 126000 | 46800 |

cost incurred. The experiment reveals manpower and CPU execution time moves in line-up with cost for sample validation. Figure 4 depicts the mean and variance of sample population is approximately equal to mean and variance of entire population (Fig. 13).

## CONCLUSION

Data quality is very important for post data migration process to validate the entire data set is tedious task due to high volume of data, complex required to conduct full

data verification is exorbitant. Hence, need an optimal method to assess state of data quality for the decision database after ETL/migration within manageable limits of cost and time without compromising the quality of data. The proposed mathematical model using deterministic statistical methods on various data sets is ensured 68.02% reduction in man effort, CPU time and cost which ensures comfortable confidence for end-users to rely on the data quality of the decision databases. The results reveals the proposed method is effective with respect to effort, cost, defects raised and CPU utilization time with high data quality on different data sets.

## REFERENCES


Allaire, P., J. Augat, J. Jose and D. Merrill, 2012. Data migration best practices and nondisruptive migration service capability for enterprise storage, Hitachi data systems. HITACHI. http://www.hds.com/assets/pdf/white-paper-reducing-costs-and-risks-for-data-migrations.pdf.

Badri, M.A., D. Davis and D. Davis, 1995. A study of measuring the critical factors of quality management. Int. J. Qual. Reliab. Manage., 12: 36-53.

Eckerson, W., 2002. Data warehousing special report: Data quality and the bottom line. Applications Development Trends. http://www.estv.ipv.pt/PaginasPessoais/jloureiro/ESI_AID2007_2008/fichas/TP06_anexo1.pdf.

English, L.P., 1999. Improving Data Warehouse and Business Information Quality. 1st Edn., John Wiley Sons Inc., New York, USA., ISBN-13: 978-0471253839, pp: 10.

Fields, K.T., H. Sami and G.E. Sumners, 1986. Quantification of the auditor's evaluation of internal control in data base systems. J. Inf. Syst., 1: 24-77.

Firth, C., 1996. Data quality in practice: Experience from the frontline. Proceedings of the Conference of Information Quality, October 25-26, 1996, Singapore, pp: 25-26.

Friedman, T. and M. Smith, 2011. Measuring the business value of data quality. Gartner Inc., Stamford, CT., USA.

Gupta, V.R., 1997. An introduction to data warehousing. System Services Corporation. http://system-services.com/dwintro.asp.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann, San Francisco, CA., USA., ISBN-13: 9781558609013, Pages: 770.

Howard, P., 2011. Data migration. White Paper by Bloor Research, UK., May 2011.

Kahn, B.K., D.M. Strong and R.Y. Wang, 2002. Information quality benchmarks: Product and service performance. Commun. ACM., 45: 184-192.

Knight, S.A. and J. Burn, 2005. Developing a framework for assessing information quality on the world wide web. Inform. Sci. J., 8: 159-172.

Li, X.B., 2002. Data reduction via adaptive sampling. Commun. Inform. Syst., 2: 53-68.

Manek, P., 2003. Microsoft® CRM data migration framework white paper. Microsoft Corporation.

Manjunath, T.N., R.S. Hegadi and G.K. Ravikumar, 2010. A survey on multimedia data mining and its relevance today. Int. J. Comput. Sci. Network Secur., 10: 165-170.

Manjunath, T.N., R.S. Hegadi and G.K. Ravikumar, 2011. Analysis of data quality aspects in datawarehouse systems. Int. J. Comput. Sci. Inform. Technol., 2: 477-485.

Miles, M.B. and A.M. Huberman, 1994. Qualitative Data Analysis: A Source Book of New Methods. Sage Publications, Thousand Oaks, USA.

NetApp., 2006. Data migration best practices. Network Appliance Global Services, January, 2006. http://www.sqaforums.com/attachments/575862-Data_migration.pdf.

Potts, W.J.E., 1997. Data Mining Using SAS Enterprise Miner Software. SAS Institute Inc., Cary, NC., USA.

Price, R.J. and G. Shanks, 2004. A semiotic information quality framework. Proceedings of the IFIP International Conference on Decision Support Systems, July 1-3, 2004, Prato, Italy, pp: 658-672.

Ravikumar, G.K., J. Rabi, T.N. Manjunath, R.S. Hegadi and R.A. Archana, 2011. Design of data masking architecture and analysis of data masking techniques for testing. Int. J. Eng. Sci., 3: 5150-5159.

Ryu, K.S., J.S. Park and J.H. Park, 2006. A data quality management maturity model. ETRI J., 28: 191-204.

SAS Institute Inc., 1998. SAS institute white paper, from data to business advantage. Data Mining, The SEMMA Methodology and the SAS® System, SAS Institute Inc., Cary, NC., USA.

TDWI, 2006. The taking data quality to the enterprise through data governance: A report. The Data Warehousing Institute.