Automated Data Validation Framework for Data Quality in Big Data Migration Projects

V. Rathika [1], Dr. L. Arcokiam[2]

[1] Assistant Professor Idhaya College for Women, Kumbakonam
[2] Asso. Professor, St.Joseph's College, Trichirappalli,

ABSTRACT: *The process of moving vast amount of data from one place to another is called big data migration. Huge volume of data is extracted, transformed, structured and loaded from legacy data base into a newer structure in the process which leads to data corruption. Data validation testing is essential after data migration process over to ensure data quality. To perform efficient data migration process, source data are mapped to new system which handles all the data formats. It is important when upgrade and relocation of existing systems. Businesses are creating significant data management challenges by increasing volumes of data. They should be able to access and organize volumes of data stored in a variety formats. Compare to manual process, automate data validation improve data quality in less time and cost with good quality. This paper emphasis on proposing model to do automatic quality checks for huge volume of data migrations.*

Keywords: *Architecture, Data Migration, Data Quality, Framework, Validation Testing*

I. INTRODUCTION

Big data migration refers transferring large volume of data between computer systems, storage types and formats. This is applied in the switching of single or multiple old systems to a new system when historical data are migrated into new ones [1]. Government organizations and companies are invested substantial amounts of time and money in their attempts to voyage legacy data for use in new information intensive applications. Customer Relationship Management (CRM), Business Intelligence (BI), and Enterprise Resource Planning (ERP) projects will be succeeded depends heavily on the quality of data that emerges from the data migration process. Unexpectedly data migrations

won't run smoothly always, as the process of migrating data from legacy systems for more data demanding applications unearths serious data quality problems.

All organizations are practiced the fact of "extract, load and explode" where legacy data has proved to be singularly unfit for use in new business applications. Receiving the data right from the origin is necessary if a data migration project is to succeed [2]. Data quality is progressively more serious problem for organizations big and small. It is essential to all data integration initiatives. Before data can be used successfully in applications, it \desires to be analyzed and cleansed [3]. To guarantee high quality data is continued, organizations need to apply enduring data cleansing processes and procedures, and to check and track data quality levels over time. Otherwise poor quality will lead to bigger costs, breakdowns in the supply chain and poorer customer relationship management. Data quality has the following features such as accuracy, validity, integrity, completeness, timeliness and accessibility [4].

II. RELATED WORKS

In paper [4], authors proposed a model to do quality checks for huge database migrations using random sampling techniques. In paper [5], authors discussed practice based testing and quality assurance techniques to reduce or even reduce data migration risks. Concerning Information Technology continuation, more than ever companies are confronted with the challenge of migrating data from at least one source to one target business application. Data migration understood as a tool supported one time process which aims at migrating converted data from a source structure to a target data source whereas both structures differ on a theoretical and / or methodological level. While the first level refers to business object types which significant conceptualizations of objects in the business world are contract, product, and customer account. The second level denotes their procedural

comprehension within databases. In paper [6], authors discussed and analyzed possible set of causes of data quality issues from complete survey and discussions with SMEs. In paper [7], authors proposed automation of data migration validation testing process to ensure quality assurance and risk control across industries.

In paper [8], authors listed out some of the following quality factors are definition conformance, accuracy, validity, non duplication, completeness, accessibility and timeliness. In paper [9], authors proposed an automated tool for data validation testing in migration project. It automates the comparison of all data from source and target that ensures quality. In paper [10], authors explained data quality issues at source database will lead to create errors in the ETL process. In paper [11], Corporate events like mergers and acquisitions or carve outs guide to company wide data integration and consolidation activities. Achievement of novel business models and processes brings along new functional and non functional necessities no longer supported by the existing application. Its exchange induces the migration of the contained data. New constitutional and dictatorial requirements demand for adapted business processes only supported by up to date business applications.

Decommissioning the accessible application comprises the migration of data. These demand the reoccurring alternate or consolidation of existing business applications as a type of upholding. As a consequence, data migration projects represent an everlasting although infrequently performed discipline.

In paper [12], Modern Companies believe their data as a valuable asset. However, any unexpected and rough movement of this asset in the shape of an unprincipled migration project exposes that company to higher risk. It is therefore fundamental to follow a hard and step wise approach which openly addresses data migration risks. Definitive goal of a data migration is to move data, a lot of truthful planning needs to happen prior to the move in order to ensure a successful migration. If the migration attempt does not properly 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 twist up with poor quality. It has the issues such as costs associated with error detection, costs associated with error prevention, costs associated with error rework, costs associated with delays in processing, time delays in operations, difficulty and faulty decision making, and enterprise wide data inconsistency.

III. AUTOMATED DATA QUALITY CHECKING PROCESS

The proposed framework "Fig.1" will handle the quality issues automatically and produce good quality data migrations between data warehouses. Each and every step in this model can concentrate about data quality. This examines both source and destination data warehouses to ensure the maximum possibility of quality migration. After completion of migration, the entire process will switch over to target data ware house only. Data quality has completeness, consistency, validity, integrity, conformity, and accuracy as factors to ensure the quantity of quality. Expected availability of data is called completeness. Synchronization of data is called consistency. It won't provide conflict information. Correctness and reasonableness of is called validity.

Unaltered data is referred as integrity. It ensures data should not be altered in the migration process. Format of data is called conformity and it ensures data has same format in both source and destination. The proposed model will explain about these quality factors. It consists of data assessment, mapping document, data extraction and validation, migration validation, and reconciliation process – post migration. These stages are explained in the following sub sections.

3.1 Data Assessment

Data sources are identified in this stage. System extracts and queries are run to conduct user interviews on the data migration process. Migration scope and validation strategy are reviewed here to develop work plan. Migration scope document, migration validation strategy document, and work plan document are the outputs of this stage.

3.2 Mapping Document

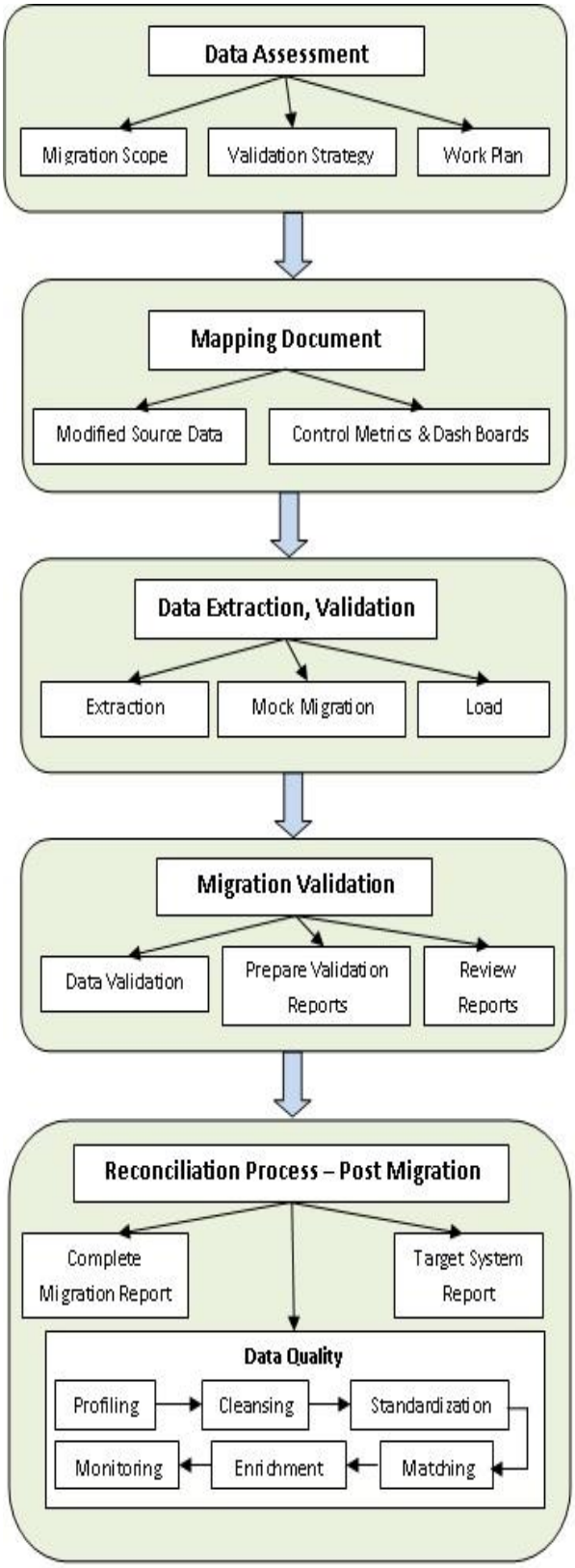
This is based on the work plan document of previous stage. This prepares the environment of source and destination. Validation and transformation rules are prepared by this stage only. This gives output as modified data source, and control metrics and dashboards.

3.3 Data Extraction, validation and loading

Data element mappings, tables, scripts jobs to automate the extraction are created by this stage. After the creation it verifies them also. Data are extracted from the source system and mock migration will be conducted to verify the

environment of the migration process. After this it sends source data into the new system using ETL tools. This conducts internal data validation checks like business rules and referential integrity checks and performs data validation. The outputs are extracts from source system, data jobs and scripts, application loaded with converted data, exceptions, alerts and error handling control points migration modules.

Figure 1. Framework of Automated Data Quality Checking



3.4 Migration Validation

In this stage only data is moving (pilot migrations) from source data warehouse to destination data warehouse. This executes specific customizations on target database and application. This performs data validation and prepares migration validation reports and data movement metrics which are reviewed here. This records count verification on the new system. This provides signed off migration validation document.

3.5 Reconciliation Process – Post Migration

This is the important stage of the proposed model. Data quality has 6 steps that are profiling, cleansing, standardization, matching enrichment and monitoring. This completes data migration reports and cross reference files. This prepares target system reports. The output will be exception reports and cross reference files.

Profiling determines that how existing data sources meet the quality standards of the solution. This can identify issues that require immediate attention and avoids the unnecessary processing of unacceptable data sources. So, this will reduce execution time. This is having the tasks like column statistics, value and pattern distribution. Column statistics can identify invalid details of a cell. Value may identify normal and outlier values in a column. Pattern Distribution can identify invalid strings or irregular expressions.

After the success of profiling, cleansing can do de-duplication to ensure that all business rules are properly met. Standardization restructures data into a common format to build more consistent data. This stage identifies and standardizes patterns of data across variety of data sets like tables, rows and columns.

Data matching finalizes data sets into identifiable groups. This can merge related records. Enrichment enhances the values of data to attract customers. This adds extra and additional information from various sources. This can provide a better understanding of customers. Monitoring helps organizations immediately

recognize and correct issues before the quality of data declines.

IV. CONCLUSION

The proposed framework has successfully automated the data validation to ensure quality in the migration process. This framework will fit for any kind of source data. This checks quality in all the steps and prepares the report. If there is any controversy, it rectifies immediately in the same step itself. So, this ensures data quality in the entire process. The benefits of this framework, one can save time, money, and quality will be very high.

REFERENCES

[1] Shinde Anita Vitthal, Thite Vaishali Beban, Roshini Warade and Krupali Chaudhari.: Data Migration System in Heterogeneous Database, *in International Journal of Engineering Science and Innovative Technology*, 2(2), pp. 88–92, (March 2013).

[2] Klaus Haller.: Towards the Industrialization of Data Migration: Concepts and Patterns for Standard Software Implementation Projects, *in Springer*, pp. 63–78, (2009).

[3] Vishnu B, Manjunath T N and Hamsa C.: An Effective Data Warehouse Security Framework, *in International Journal of Computer Applications Recent Advances in Information Technology*, pp. 33–37, (2014).

[4] Manjunath T N, Ravindra S Hegadi and Archana R A.: A Study on Sampling Techniques for Data Testing, *in International Journal of Computer Science and Communication*, 2(1), pp. 13–16, (June 2012).

[5] Florian Mathhes, Christopher Schulz and Klaus Haller.: Testing & Quality Assurance in Data Migration Projects, In *27th IEEE International Conference on Software Maintenance*, Williamsburg, pp. 25–30, (2011).

[6] Manjunath T N, Ravindra S Hegadi and Mohan H S.: Automated Data Validation for Migration Security, *in International Journal of Computer Applications*, 30(6), pp. 41–46, (September 2011).

[7] Priyanka Paygude, Devale P R.: Automated Data Validation Testing Tool for Data Migration Quality Assurance, *in International Journal of Modern Engineering Research*, 3(1), pp. 599–603, (February 2013).

[8] Manjunath T N, Ravindra S Hegadi and Ravikumar G K.: Analysis of Data Quality Aspects in Data Warehouse Systems, *in International Journal of Computer Science and Information technologies*, 2(1), pp. 477–485, (2011).

[9] Priyanka Paygude, Devale P R.: Automation of Data Validation Testing for QA in the Project of DB Migration, *in International Journal of Computer Science Engineering and Information Technology Research*, 3(3), pp. 15–22, (August 2013).

[10] Ranjith Singh, Kawaljeet Singh.: A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing, *in International Journal of Computer Science Issues*, 7(3), pp. 45–51, (May 2010).

[11] Haller K.: Towards the Industrialization of Data Migration: Concepts, *in 21st International Conference on Advanced Information Systems Engineering*, Netherland, pp. 70–78, (2009).

[12] Atsa Etoundi Roger, Abessolo Alo'o Ghisiain and Simo Bonaventure Joel.: Migration of Legacy Information System based on Business Process Theory, *in International Journal of Computer Applications*, 33(2), pp. 27–34, (November 2011).