International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.3, Issue.1, Jan-Feb. 2013 pp-599-603 ISSN: 2249-6645

Automated Data Validation Testing Tool for Data Migration Quality Assurance

Priyanka Paygude[1], P. R. Devale[2]
[1]Research Scholar, [2]Professor
Department of Information Technology, Bharati Vidyapeeth University College of Engineering, Pune, India

Abstract: *Data migration has become one of the most demanding proposals for IT company managers. Even though these projects earn high business benefits, such as reduced costs, improved productivity, and data manageability, they likely to involve a high level of risk due to the huge volume and criticalness of moved data. In order to reduce risk and guarantee that the data has been migrated and transformed successfully, it is essential to employ a thorough Quality Assurance (QA) strategy in migration projects. Testing is a key phase of migration project for delivering a successful migrated data and addressing any issues prior and after the migration process.*

Manual testing for data validation process is time consuming and inaccurate; so automated data validation assure data quality with highly reduced time, cost and maintaining good data quality. The paper proposed automation of data migration validation testing process for quality assurance and risk control across industries.

Keywords: *Automation Testing, Data Migration, Data Quality, Data Validation, ETL*

I. INTRODUCTION

Data is a precious asset for any company. So, any unplanned transfer of data can be very risky for company. In reality, planning is the top most success factor for any data migration project, independent of underline complexity. Appropriate thorough planning reduces the business impact such as application downtime, overall performance degradation, and technical incompatibilities, risk for example, completeness risk, semantic risk, data corruption/loss.

Each reason for migrating data is motivated by the need to find new efficiencies, better manage risk and stay competitive, as follows:

- Systems Consolidations: Firms are looking for reducing structural costs by standardizing on modern, cost-effective platforms and technologies; and by retiring inflexible and hard to continue legacy applications.
- M&A Activity: merger and acquisition (M&A) activities has created large organizations with a wide range of technologies that require complex IT integration programs to support merged business entities [3].
- System Upgrades: Implementation of novel business-models and processes brings along new functional and non-functional requirements no longer supported by the existing application [4].
- Ever changing legal regulations, technological progress and upgrades.

Many companies are using Business Intelligence (BI) for making managerial strategic decisions in the expectation of gaining a competitive lead in today's hard business platforms. Mostly firms uses sampling technique to test data which covers far less than 10% of data under test. Therefore, remaining at least 90% of data is untested. Thus decisions typically fail due to incorrect, untested data, which will cost their firms millions of dollars. [6]
The objective of paper is to propose an automated approach for data migration validation testing and data quality assurance.

II. DATA MIGRATION OVERVIEW

Data transfer can be of two types: first, a simple data movement that is moving data from source database to target database without restructuring and second, data migration. Data migration is the process of transferring data between computer storages, types, formats, or computer system. It is the process of moving data from the old database(s) to a new database. We called old database as a legacy or source database and this database is migrated to the new database, called as target or destination database. The data migration process becomes a difficult challenge when source and target databases are different in their internal structures. So, simple import/export procedures will not work. Thus data migration process is better to perform using automated ETL (Extract – Transform - Load) tools than doing manually.

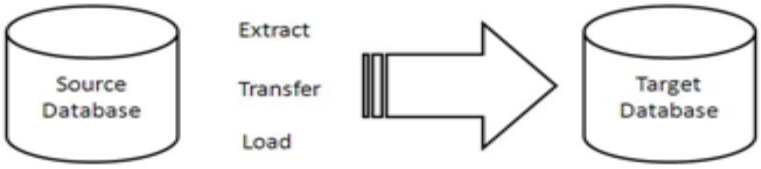

Fig. 1 Data Migration Overview

Data migration is a one-time process. It involves the re-structuring of data such as fields being merged, or formats being changed, or transforming data in various other ways. If no-restructuring takes place then we would call this data movement [2].

III. LITERATURE SURVEY

This segment sheds light on work published in the area of testing, quality assurance and data quality issues in data migration projects.

Authors has undergone literature review stage and evolved with the problem statement with the help of work, has published till today in the area of data quality and data validations in data migration projects.

Florian Matthes, Christopher Schulz, Klaus Haller, "Testing and Quality Assuarance in data migration projects," 2011 - discusses practice-based testing and quality assurance techniques to reduce or even eliminate data migration risks.

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

Lixian Xing, Yanhong Li, "Design and Application of Data Migration System in Heterogeneous Database", 2010 - paper is based on database migration project and methodically introduces technique issues of data migration involving manual work which may contribute to organizations that have data migration demands.

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.

Manjunath T N, Ravindra S Hegadi and Archana R A, "A Study On Sampling Techniques For Data Testing"(2012) - This paper emphasis on proposing model to do quality checks for huge database migrations using random sampling techniques.

Manjunath T.N, Ravindra S Hegadi Ravi kumar G.K (2011) - Discussed and analyzed possible set of causes of data quality issues from exhaustive survey and discussions with SMEs.

This paper is proposing the method of automating the data validation testing for data migrations for quality assurance and risk management in migration process, resulting in effort and cost reduction with improved data quality parameters.

IV. ONGOING METHODOLOGIES

Designing and implementing the successful migration of high volume data, unstructured content is always challenging. And testing, validating, or otherwise quality assuring results adds greatly to its complexity, cost, risks, and the time required for completion. After the migration process completes, the process of data validation testing starts for assuring user about the integrity of the migrated data.

Various methods are used for data migration validation testing:

1. Sampling Technique:

Sampling technique assumes that error is uniformly distributed, which is not true in real scenario. A sample is a group of units selected from a larger dataset of population. Valid conclusions can be drawn, by studying the samples. Random sampling is mostly used sampling method. Sampling is the process of selecting a small number of elements from a larger defined target dataset such that the information gathered from the small dataset will allow judgments to be made about the larger datasets [7].

Cons:

- Highly inefficient, error-prone process
- Requiring major manual involvement
- Time consuming comparisons of source and target systems.
- Ad-hoc procedures with limited coverage
- Final results are not 100% reliable

2. Writing 'MINUS' queries:

In this, an individual 'SELECT' query executed on both the databases/tables and then, 'MINUS' operation is used between the source and target select query result. Then the output contains records of source which are not contained in the target.

But result only shows the extra rows that are in the Source but the target lacks and not the extra rows that are in target but the Source lacks. Thus 'MINUS' queries needs to be executed twice (Source-to-Target and Target-to-Source).This double query execution, consumes more time and resources utilization [9].

```
SELECT emp_id
FROM emp_table;
MINUS
SELECT cust_id
FROM cust_table;
```

Cons:

- Double query execution, consumes more time and resources utilization.

- Only provides n of rows not present in target database/ table, no other validation like data type mismatch, null values, data corruption etc.

Traditional data validation testing techniques are highly time and resource consuming, with limited data coverage leading errors that may go undetected. These limitations can be address using an automation testing approach.

Using an automated approach for data validation testing, will make the testing process deterministic rather than procedural and can validate 100% of the migrated data along with taking care of business constraints used during transformation.

V. PROPOSED SYSTEM

The proposed tool will automate the entire testing process, from scheduling to execution to comparison to reporting across multiple database platforms that helps companies eliminate risks associated with migrations process.

A. Automated Proposed Testing Model

Mapping data is the key document for any migration project, which contains a mapping relation between source and target database. This document is a logical data map between source and target database along with transformation constraints. User will map source database with target database along with the input of mapping document that is created in migration process.

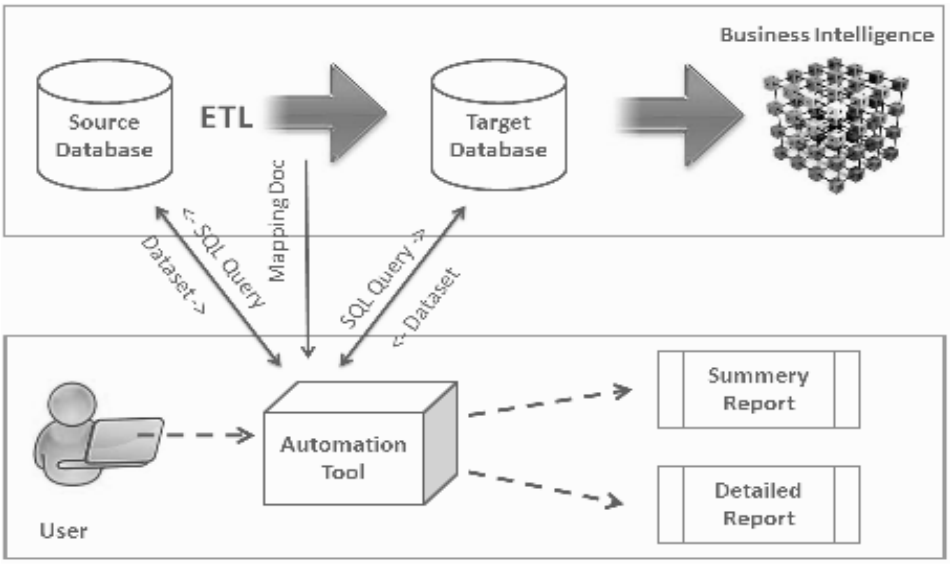

Fig. 3: Architecture for Automating Data Validation

Query to be fire on databases, will be different based on the underline database platforms. User will write either own query or chose a query from snippets. This query will automatically map to source and target database by considering underline database platforms. These queries will be fire on individual database, which in turn return a dataset. Datasets, return from source and target database will be compared by automation tool. Based on comparison result, data mismatches will be logged into summery and detail report.

B. Modules of Proposed System

The proposed methodology consisting of four modules, which are explained with the associated features in the following sub-sections. The modules are divided as

1) Test Design Library

It takes total control of test design. Test design is the foundation of any powerful testing.

- Reusable Query Snippets – Brings flexibility and reduce time in the process of query design. Snippet libraries consist of various basic query fragments that one can use to modularize queries, helping to speed up the process.
- Allow to paste queries created using yours favorite editor, to execute on respective databases.

2) Test Scheduling

Allow user to Schedule testing by time for maximum productivity

- Simplify the process by scheduling tests for the specific times when the underline architecture is available, or for a window of time when other activities will have least impact.
- Build Scenarios for scheduling your execution runs at specific dates and times

3) Live Dashboard
GUI showing live status of test execution process

- Drill-down into data as processes it to examine results as they become available during execution
- Real-time statistics for each executed Queries and for the Scenario execution as a whole
- Export detailed results in Excel, CSV or XML formats to share, store

4) Reporting
Use reports to share both high-level and detailed views of testing. Two Types of Reports will be generated:

- High-Level report gives view of testing success from high end, like total records affected, time taken, total no of defects etc.
- Detail-level report gives record level detailing.
- Export reports as PDF, XML, HTML files to share within organization or for future audit needs

C. Process of Automated Data Validation Testing
Once the process of migrating data from source to target database is done with the help of ETL tool, data validation testing can be started.

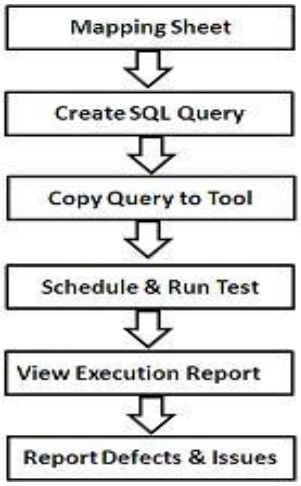

Fig. 4: Stages of automated data validation testing

Prerequisites to start automated data validation testing are:

A. Read only access to source and target schemas/tables along with database connectivity.

B. Data mapping document, created for ETL process of data migration.

Steps discuss below for data validation testing:

1. Review mapping sheet: This sheet is a logical data map that describes the mapping between target and source database, along with information of transformation rules.

2. Create SQL query: User can create SQL queries for both source and target database using any SQL editor.

3. Copy Queries to tool: User will copy created SQL queries for source and target database to tool.

4. Schedule & Run Test: Automation testing tool will allow user to schedule test for particular date, time and then run test on scheduled time.

5. View Execution Report: Automation data migration testing tool will generate two reports of run test: summery and detail report, showing result of test run.

6. Reports Defects & Issues: Detected defects in run test are logged to defect repository.

VI. ASSURING DATA QUALITY

Below table describe some common defects that proposed automated testing tool will find in data migration projects, to avoid the adverse effect that any of these anomalies can cause data migration project fail, and ultimately Business Intelligence [5].
Following table shows various test cases that tool will test:

Table 1. Data inconsistencies

Data inconsistency	Description	Example
Data Truncation	Loss of data due to truncation of data field	Source data field value "Mumbai City" is being truncated to "Mumbai C" It happens because target data field is having less or incorrect length to capture the entire source data field.
Data Type Mismatch	Dissimilarity in source and target data types.	Source data field for Interest Rate was float; however, Target data field is set to int.
Missing Data	Values of some data fields missing in either source or target databases.	Missing data values while transferring data from source to target database or data value is not completely transfer to target database.
Duplicate Records	Records which are similar to two or more records, called as duplicate records.	Record of employee with unique employee id is repeated more than one.
Transformation Logic Errors	Transformation logic is not followed causing errors in data values	Default integer data value of source database is to be transfer into target database in percentage format, which is not done properly in migration process causing bad data.
Null Translation	Incorrect transformation of source NULL values to target database.	NULL values of source data field are supposed to transform by default value in target data field. However, due to incorrect logic of implementation, it results in the target data field containing NULL values.

VII. CONCLUSION

Data migration is a tough project with high level of risks like time overruns, budget etc. Use of quality ETL tool will minimize the risk of defects in data of target database. Even though, testing of data for its validation is important and can't be overlooked. Testing validity of data using manual or just writing 'MINUS' querise, are not the effective way causing risk with data quality. Here, the proposed system assure data quality using standardize way of data testing in migration projects across the enterprise, multiple platforms, and applications. Using proposed solution, one can save time, cost, and manual efforts; along with data quality assurance.

References

[1] Florian Matthes, Christopher Schulz, Klaus Haller, "Testing and Quality Assuarance in data migration projects," 2011 27th IEEE International Conference on Software Maintenanace(ICSM)

[2] P. Howard and C. Potter, "Data migration in the global 2000 - research, forecasts and survey results," London, United Kingdom, p. 29, 2007

[3] Sagar Khandelwal, Kannan Subramanian and Rohit Garg, "Next Generation Cross Technology Test Data Solution for M&A", 2011 27th IEEE International Conference on Software Maintenance (ICSM)

[4] Endava, "Data Migration - The Endava Approach," London, United Kingdom, p. 11, 2007

[5] John Hess, "Dealing With Missing Values in The Data Warehouse" *A Report of Stonebridge Technologies*, Inc-1998

[6] C. Burry and D. Mancusi, "How to plan for data migration," 2004

[7] Manjunath T N, Ravindra S Hegadi and Archana R A, "A study on sampling techniques for data testing", International Journal of Computer Science and Communication, Vol. 3, No. 1, January-June 2012, pp. 13-16

[8] IBM, "Best practices for data migration - Methodologies for assessing, planning, moving and validating data migration," Somers, NY, USA, p. 16, 2009

[9] Manjunath T N, Ravindra S Hegadi, Mohan H S, "Automated Data Validation for Data Migration Security", *IJCA Online*, **30**/number 6/3642-5088: ISBN: 978-93-80864-89-0.