ISSN: 2321-7782 (Online)

Volume 3, Issue 11, November 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

The Data Migration Testing Approach

Madhu Dande
TCS,
Bangalore, India

Abstract: The main objective of this paper is about the migration of database from source database server to target database server based on the conditions, before moving to target database if any Transformation/Manipulation has to be done. In this process based on the client requirements, the business analyst will develop the Business requirement specification documents, Entity-Relationship document, mapping documents whereas developers will design and develop the High level and Low level design document as well develop the code and testers has to identify the bottlenecks and process to resolve the issues in data warehouse testing. Testing team will design and develop the test scenarios, test cases and traceability matrix documents. Testers will execute the test cases on the test environment of the database (Source & Target) to make sure the source and target data is matching as expected.

This document details the testing process involved in BI which is ETL and Reporting testing the requirement and increases the test coverage. By following this approach and process will provide a defect free product with quality deliverables within the timeline.

Keywords: SDLC phases, Environments and Servers, Data Mapping sheet, ER-Diagram, ETL Testing, System Testing, Reporting Testing.

I. INTRODUCTION

In recent years the amount of data stored in databases has increased in petabytes/zeta bytes in size. Data are closely related to the applications like Business, Customers, Scientific, etc are stored in the database in bytes. So there is a huge data dump where there are possibilities for data redundancy, data constraints, data integrity and data inconsistency in an organization i.e. enterprise level.

BI – Data Warehouse (DWH) and Reports testing is one of the key areas. Companies tend to use different database servers or systems in building a data warehouse which helps them in developing efficient reports for best decision making process.

This paper contains four problems within sections and considering the fourth problem and providing the solution to improve the effective utilisation of the network resources based on the weightage and usage on Phases and Environment which would define the importance of the network maintenance.

Section-1 defines about the network which exists in the one of the largest retail chain stores in UK.

Section-2 defines the systems which are connected the network and this section mainly concentrated on the man power resources who work in the IT section of the business.

Section-3 define the different Environments in the organisation for develop and implement the work in the form of Project, which improves the customer service increases the productivity of the business.

Section-4 define the different types of the projects in the organisation as follows New Business, New enhancements, modification of the existing project, Migration projects, decommissioning the projects and data centre consolidation which consist of the Hardware, software and Networks.

II. Data Migration

Data migration is a process which involves the migration of data from an existing database to a new database (i.e. moving or transforming the data from legacy source system to target database systems). This includes one to one and one too many mapping and movement of static and transactional data. Migration also relates to the physical extraction, transmission and loading data from legacy to target systems.

Data migration activity includes everything with respect to data. It ensures that the new database is up and running without any issues. It specifically contains all the legacy data (data that is present in the existing database), and the data has been migrated to correct table(s) and column(s).

a) Direct Data Migration

In this type of migration the data that are present in the source database is directly transferred to the target database.

Migration of database from one database to another database without impacting the functionality of the application and the content of the data should not be changed. As well enhancements of the tables and columns within the database schemas either by creating new tables and columns or updating the tables and columns as well the column types i.e. data type and data sizes.

E.g.:- legacy data to the new database, file system to database system, technology change, version upgrade etc

b) Data Consolidation

Consolidation of different data bases (e.g.: Oracle, Flat files, SQL Server, DB2 to Oracle) in the same organization to build target database as single source database system. So we can achieve the Enterprise level consolidation which reduces the cost of the hardware, software, increases the security and last but not the least maintenance cost will be reduced drastically.

c) Data migration with changes to target

This migration requires some transformation of the source data before it is migrated to the new database.

For e.g., a field ‘Priority’ in the source database may allow only one out of the three values High, Medium and Low. But in the target database the same field allows one out of the three values 1, 2 and 3. Thus the High value in the source database will be transformed to a value 1 and will then be migrated to the target database.

Migration of the data from legacy system to the target system on daily, weekly, monthly as well periodically is also called as batch processing. Eg: “End of business day” closedown.

d) Road Map technologies

In this section explains about the different types of migration like Operating System changes, installing new applications, and patches/build software applications. As well organization or enterprise level decision has to change the systems and its configuration like installing new monitors, increase the size of the memory or Hard disk or Processor speed etc.

In this data migration process whenever an organization decides to upgrade or change the Hardware/Software(Database) of the server, it will need to transport the existing data to the new database or to change the architecture to improve the performance.

Note: Most of the time everyone in the project is concentrated on their own work but they never ever try to know who is owner and location of the database server.

e) *Data Volumes*

As a company grows the volume of data also grows respectively, in turn the database size also getting increased tremendously. These collected data to be tested or verified by the database owner to reduce the number of data redundancy, data inconsistency, Data accuracy and increase the data efficiency.

III. Dependencies

Dependencies are more important to reduce the risk in the project.

a) Environmental Dependencies

When a project is executed on an onshore-offshore model then user needs to concentrate on the connectivity. Offshore teams will be working in Remote systems for Security reasons and in certain cases even Onsite team will have limited access.

b) Support Dependencies

» Training (legacy & target applications) testing team

» Business Analysts -provide expert knowledge on both legacy and target systems

» Operations - Hardware / Software / Database Analysts - facilitate system housekeeping when necessary

» User Acceptance Testers - chosen by the business

» Business Support for data cleanse

c) Data Dependencies

» Translation Tables - translates legacy parameters to target parameters

» Static Data / Parameters / Seed Data (target parameters)

» Business Rules - migration selection criteria (e.g. number of months)

» Entity Relationship Diagrams / Transfer Dataset / Data Dictionary/Schemas (legacy & target)

» Sign Off / User Acceptance criteria - within agreed tolerance limits

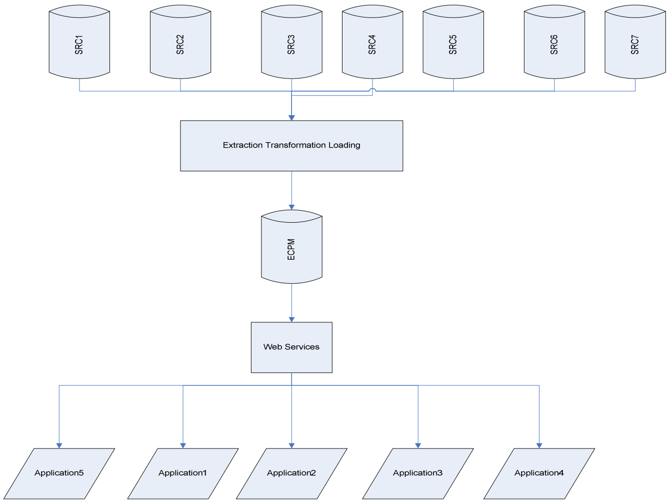

Fig (1): Enterprise level Architecture

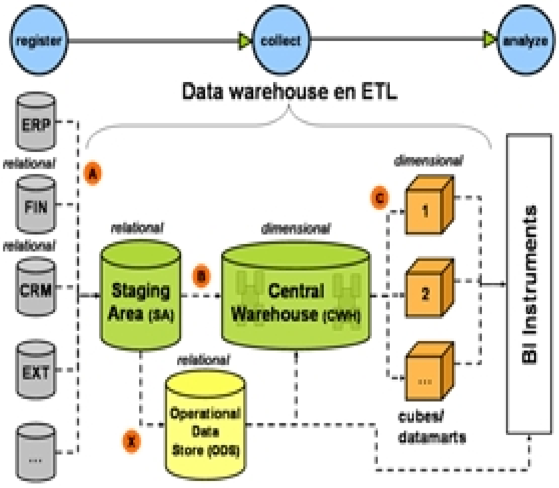

Fig (2)

Considering one of the largest banks in USA with different portfolios has Loans (Auto, Housing, and Personal etc), Personal banking and Money Savings or investment banking. The different databases from the different portfolios which are extracted the customer data and pushed into single database as “Enterprise Customer Profile Management” (ECPM). Where the customers are having the accounts in different portfolios and their address are not in sync, which is a major challenge for the banking to track their details. So the banking directors decided to build a system with Customer information which is to be called as ECPM System. Any Portfolio in that bank should not store the customer information in their database. Any customer information details will be pushed into ECPM system. Similarly when the portfolios need the customer information, they can create/update/view the data through web service with different types of access & permissions. With the help of ECPM shown in the Fig (1) & (2), bank has customer data in safe and secured as well there is no ambiguity of the customer details.

IV. SDLC Phases

a) Business Requirements Gathering

In this phase, collecting the business requirements from the client by the business analyst and it should be reviewed by the subject matter expert to make sure the requirement is clear and understandable to the project team and requirements should not have any ambiguity.

b) Analysis & Design

Analysis & Design involves the understanding of the requirements of data migration which includes estimating the size of the database, complexity and identifying the different types of migration.

1. The size of the source database (also called the existing database). However for testing purpose the size of the database may not be equal to the physical size of the database, because it is not necessary that all the data that is present in the source database is going to migrate to the target database (new database). Data migration may perform some data cleanup on the source database so while migration the corrupted data or irrelevant data will be removed automatically to the target database.
2. Estimating the number of fields (columns) and the tables that are included in the scope of data migration.
3. Identifying the fields (table names and fieldnames i.e. column names) that has maximum number of records.
4. Identifying the field that has the most complicated migration logic. The requirements for data migration generally come in form of spreadsheet(s) that contain a field level mapping between the source database and the target database.
5. Mapping explains the destination (of a field from source database) in the target database shown in below Fig (3). Mapping could be ‘one -to one’, or ‘many-to one’. In ‘one-to-one’ mapping all the data from say sourcedb.table1.field1 is migrated to targetdb.table1.field1. In ‘many-to-one’ mapping data from more than one field in the source database is migrated to only one field in target database. The documents will also specify the different types of transformations that are being performed on the source data before the migration. A thorough analysis of the requirement documents (or the mapping document(s)) is the activity to be performed in this step. This step would also help to identify the fields which can be automated and the fields that can be performed manually for testing

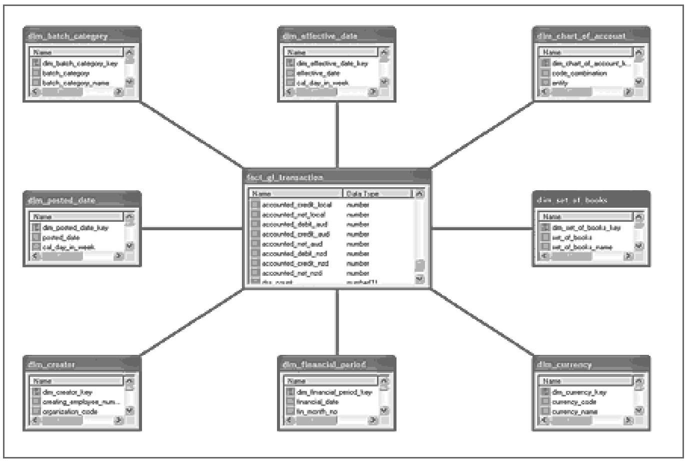

Fig (3)

Database Design to be taken care by the below team:

Role	Responsibilities
DBA Lead	1. Identify the Database Server 2. Monitor the capacity and performance of the database 3. Providing the access(user rights)
DB Lead	Needs to be taken care the data modeling and database architecture to increase the performance of the data retrieving
ETL Lead	Needs to understand the database modeling and the business rules

Table-2

1. *Data mapping*

This process describes the mapping data from the legacy to target database schemas. It also takes care the reformatting of data if needed. This would normally include the derivation of translation tables used to transform parametric data. Which needs to be updated the static/seeded data into the target database based on the data mapping sheet.

2. *Specifications*

These designs are produced to enable the developer to create the Extract, Transform and Load (ETL) modules. In data migration specification: Functional (e.g. migration strategy) Detailed Design (e.g. data mapping document)

3. *Iterations*

The refinement of the migration code i.e. increases data volume and decrease exceptions

» Continual identification of data cleanse issues

» Confirmation of parameter settings and parameter translations

» Identification of any migration merge issues

» Reconciliation

From the experience the majority of the data will conform to the migration rules and as such take a minimal effort to migrate ("80/20 rule"). The remaining data, however, is often highly complex with many anomalies and deviations and so will take up the majority of the development time.

c) Testing

This includes development of test approach and test scripts for all types of migration that are in scope. The content of the fields to be tested automatically is extracted from the source and target databases using SQL queries for comparison.

Following points to be remembered by the Testing Team

1. Before the testing team starts to develop their approach, the requirements (in case of data migration), Data mapping (transformation rules and logic) and ER document (mapping of attributes) in the source and target database has to be frozen.
2. The testing team should be provided with the latest requirements (mapping documents) else it will lead to major rework in the approach, SQL queries and automating programs.
3. Testing team may not understand the transformation rules or business rules correctly in case of functional transformations.

4. The testing team must have good knowledge in understanding the relationship between different attributes else the SQL queries and automating programs will lead to incorrect results and eventually result in improper defect reporting.

a. *Test Strategy*

This document will define the Process of testing, the Scope of testing, and estimated effort with Timelines of the testing. The process of Test Planning can be initiated once the Test Strategy is signed off by the Customer.

b. *Test Planning*

The test planning phase deals with the creation of a test approach for each field. This includes both automated testing and manual testing at field level.

- Activities during planning of data migration testing.
- Understanding the Requirements.
- Understanding the databases involved.
- Estimating the size of the data that is under the scope of data migration.
- Identifying the 'Entity Relationship' in the existing and the new databases.

c. *Manual Testing Techniques*

While in manual testing we need to start with the counting number of records in the source and target and compare the results.

In few cases where the minimal test data is set up for the first time in databases, a manual comparison would be a better option by using the Excel spreadsheet.

In database testing, users prefer to test the sample data from the source and the target. Testing techniques can be implemented to compare the values in source and target (expected and actual values) and count the number of records between source and target through Excel.

The limitation would be the number of records that can be compared manually, because with the increase in the number of records to be compared, the chances of human errors are more. For the data that is being setup from spread sheets, the validation can be done manually. The manual approach is preferred because the data is already available in spreadsheets and the data from the database can be imported to another spreadsheet and comparison can be done. For the comparison some of the functions available in excel can be used.

The drawback in this method would be the limit on number of records that can be imported to Spreadsheets at a time. The records from the database need to be sorted in the same order as in the source spreadsheets, before comparison can be done.

d. *Test Execution*

Test execution step involves execution of automated test scripts and also execution of manual test scripts.

e. *Result and Defect Analysis*

All test scripts that fail during the test execution phase result are defects. The defects that are reported during the test execution phase should be associated with one or more test scripts. It ensures that all defects are properly tracked at all times during the test execution phase. Defect analysis results in a test execution summary report. The objective of this report is to indicate a trend in defect reporting exercise. The report indicated those areas of data migration that are prone to defects and also the areas which are comparatively defect free. This helps in deciding another cycle of test execution.

Horizontal reconciliation (number on legacy = number on interim = number on target)

Vertical reconciliation (categorization counts (i.e. Address counts by region = total addresses) and across systems).

Note: Check list has to be maintained in all stages (legacy, interim, and target).

Create the test cases and SQL queries based on the below phases.

Testing can be divided into three phases

Phase 1: Extraction

1. Extract the data from the Sources (different file formats of data and different databases)
2. Validation of different formats of the data sources

Note: Sources means different types of formats (Text file, Excel file, Mainframe format file, Oracle, DB2, Etc….)

3. Data could be extracted based on the required fields.
4. Extraction scripts are granted security access permissions to the source systems.
5. Source to Extraction destination is working in terms of completeness and accuracy.
6. Extraction is getting completed within timeline.

Phase 2: Transformation

1. Transaction scripts are transforming the data as per the expected logic based on the data mapping sheet.
2. Detailed and aggregated data sets are created and are matching.
3. Transaction Audit Log and time stamping
4. Validate the Business Rules and Data Integrity
5. Clean the invalid records, rejected and error records
6. Entity Integrity
    a. Referential Integrity
    b. Domain Integrity
7. Check for data Consistency
    a. Validity, Accuracy, Usability and Integrity of related data between applications
8. Check for data Redundancy
    a. Duplicate Records
    b. Repeated data records
9. Check for aggregate Calculation
10. Transformation is getting completed within timeline

Phase 3: Loading

1. Data sets in staging tables to load into destination tables.
2. Both incremental and total refresh can be done.

3. Validation of timely Batch jobs
   a. One time batch run
   b. Weekly/Monthly batch run
   c. Whenever needed, we need to run the batch based on the work
4. Validation of data
   a. Data type
   b. Size of the data
   c. Data values (Test Data)

Design and development of Automated Test Scripts This step is carried out in parallel with the test planning. As the approach for a field that would undergo automated testing is developed and is reviewed internally. The TESTING team writes a program for the field that brings data from two different database onto a single platform (in this case that will be the program itself) and the program then performs a comparison at data unit level between the data fetched from the two databases. One program can actually perform automated testing for more than one field if the fields are from the same table and are undergoing similar type of migration. Next step is to identify the tests to be automated and the tests that can be automated.

	AS-IS	TO-BE(KISS Model)
Requirements Gathering	BRS	BRS
Design Docs	Universal Design Document, Detailed Technical Documents, Mapping Document	Universal Design Document, Detailed Technical Documents, Mapping Document, ER-Diagram
Testing	1.Test Scenarios is considered as one test case 2.In all stages data validation is done	1. Each test condition to be validated 2. Test scenario divided into based on the test conditions 3. Each test case can be divided into the test steps
Testing process	1.Validating the Source, Lookup, Target and error tables 2.Validating the mapping tables 3.Data validation between source and taparget	1. Validating the source, by counting the number of records 2. It is not needed to validate completely on each record (where we don't have the requirements clear) 3. Need to make sure to validate the target (validation based on requirements on the specific point) 4. Sharing of workload will be easy i.e. to execute the test cases. 5. Validation of the test results will be easy 6. Effort calculation, test metrics easy to maintain.
SQL	Huge SQL Query needs to be executed	1.Write a SQL query for each column 2. Count the number of records for each table. 3. Validate the each column in the table 4. Validate the business rules 5. Validate the data values 6. Check the data integrity, consistency & redundancy
Sharing the workload	Single person should be assign to work on the specific query	It is easy to share the workload within the team
Debugging	While debugging the SQL query, it will be taken too much time to identify the issue	Simple queries debugging is very easy
SQL Knowledge	Mandatory need the SQL Expertise to write or to execute the Query	Basic knowledge is sufficient to create or to execute the Query.
Dependency	Resource dependency will be there!	There is no dependency of the resource to work
Execution Time	If the SQL query is ready to execute, then time taken is very less Note: SQL Debugging may take more time)	If the SQL query is ready to execute, then time taken is very less. Note: Create and update the SQL query
Test result Comparison	Needs more effort to compare the test results	Less effort to compare the test results (Expected = Actual)
Enhancements	Need huge effort to update the SQL query	less effort to update the SQL query

Table-2: Traditional Vs Typical approach

Test Automation

When testing large databases, manual testing of the migration may prove to be in-efficient. Sampling of data can be done through manual verification, but this can lead to in-efficient testing as a larger part of the database may be left untested. Thus automated tests should be performed to ensure that an error free migration is performed. Automation involves using scripts which will do the job of comparing the data in the source and the target databases. For an automation would be those fields where the number of records involved are large and also those where there is functional transformation taking place when the data is being migrated from the source to the target database.

Automation can be done by writing test scripts that compare the data present in the two databases. The script may be written in any programming language that is comfortable to the tester and which has the functionality to access both the source and target databases. The automated script merely has to retrieve the data present in the two databases and compare the values/records of each column/table of the source and target database with each other. The scripts will contain SQL queries that will retrieve the data present in the database. The complexity of these queries will depend on the database and also on the type of migration being performed. Data retrieved from the source and target databases should be sorted in the same manner. After the script retrieves the data, the script should perform a one to one comparison between the data elements retrieved from the source and target databases. The details of the comparison should be written into a file thus allowing the tester to verify if any errors were encountered during the comparison.

V. Conclusion

In General Data warehousing Testing Approach which can be database testing, ETL Testing and data reporting testing should follow Keep It Short & Simple (KISS) Model to get defect free quality product with efficient way of doing the testing and reducing the complexity of the testing. This approach is simple to implement and makes testing job is easy and very efficient. Increase the testing efficiency, frequency of the test cycle execution by using automating it.

References

1. Ralph Kimball, “The Datawarehouse Toolkit”
2. William H Inmon, “Building the Datawarehouse”
3. Matteo Golfarelli, Stefano Rizzi “A Comprehensive Approach to Data Warehouse Testing” 2009 ACM 978-1-60558-801-8/09/11
4. Harry M. Sneed, “Testing a Datawarehouse - An Industrial Challenge” ISBN:0-7695-2672-1

Author(s) Profile

Madhu Dande, received the M.Tech degree in Computer Science & Engineering from Visvesvaraya Technological University during 2001 - 2003, completed his B.Tech in Electrical & Electronics Engineering from Sri Venkateswara University during 1995 - 1999. He is member of IETE and CSI. He has worked in Centre for Development of Telematics (C-DOT), Bangalore, where he has implemented Voice of Ethernet project successfully. Later he has joined TCS as Automation Test Lead in 2003. He has shown his skills design and development of the Automation Framework. He had patent on Generic Unit and Integration Automation Framework Testing granted by USPTO. In the process of filing a patent on Demon Web UI Utility to validate the broken and orphan links at run time of the business critical applications. Designed and developed a Framework on Robust Automation Testing tool for functional testing. Currently working on Human Vs System Resources utilization in Production Environment to reduce the utilization of power, Maintenance Cost and increase the availability.