IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 9, PP 77-82

A Comparative Study of Data Migration Techniques

Ajeet Ghodeswar[1], Trupti Shah[2], Amruta Mhatre[3], Santosh Dodamani[4]
[,1,2,3]*Department Of Computer Engineering, Atharva College Of Engineering,Mumbai, Maharashtra, India*

Abstract: *This Paper Gives The Overview Of The Data Migration And Its Basic Concepts. The Need For The Data Migration In Terms Of Business Needs, Different Environments And The Criticality Of The Legacy Databases Are Discussed. The Data Migration Process Generally Carried Out In Three Steps: Plan, Migrate, And Validation. An Automated ETL (Extract-Transform-Load) Process/Tools Is Use To Move The Data From Source To Target Database. ETL Is The Process Of Data Cleaning, Transforming And Finally Loading It Into A New Target Database. Existing Methodologies Like Line Of Code, Sampling Techniques And MINUS Queries Are Briefly Discussed In The Paper. During The Process Of The Data Migration (DM), Too Much Data Is Extracted, Transformed, Structured, And Loaded (ETL) From Legacy/Source Database Into A Newer Structure That Is The Target Database. This Process Leads To Various Types Of Changes In Data, Known As Data Inconsistencies /Quality Issues*

Keywords- *Data Migration, ETL, Data Mapping, Data Sampling*

I. Introduction

Many Established Businesses Have Legacy Databases That Are Costly To Maintain, Risky To Modify; Affecting On Business's Competitiveness, Reputation, And Outcomes. Solution To Tackle With Problem Is `Data Migration'. Regarding Information Technology (IT) Maintenance, More Than Ever Companies Are Confronted With The Challenge Of Migrating Data From At Least One Source To One Target Business Application [2]. Since The Success Of The Application Replacement As A Form Of IT Maintenance Is Dependent On The Underlying Data Migration Project, It Is Vital To Achieve The Migration On Time And Budget. During Actual Data Migration, Organizations May Face To Shrinking Budgets, Smaller Teams, And Tighter Deadlines, Which Often Force Them To Sacrifices Quality. Migration Does Not Mean Simply Moving Data. In Reality, Data Migration Is Always A Subset Of A Larger Application Initiative Like CRM, ERP, BI, And Many More. It Often Involves A Custom-Built Application Consolidation, Migration, Or Upgrade.
Modern Companies Consider Their Data As A Precious Asset. However, Any Unplanned And Rough Movement Of This Asset In The Form Of An Unprincipled Migration Process Exposes That Company To A Higher Risk. Thus, It Is Very Important To Follow A Stringent And Stepwise Approach Which Eliminate Data Migration Risks At All Stages Of Moving.

Data Validation Is The Key Operation Need To Be Perform After Migration Process Over. It Ensures That Data Migration Is Properly Processed And That It Will Not Corrupt The Target System Or Loose Important Information. Goal Of Validation Phase In Migration Process Is To Get The Exactly Same Copy Of Source Data And Must Retain Same Functional Correspondence With The Source One. A Single Migration Process May Involve Thousands Of Records, Making It Suspicious For Defects.

Database Migration: It Is The Process Of Moving Data From The Old Database(S) To A New Database(S). We Called Old Database As A Legacy Database Or The Source Database And This Source Database Is Migrated To The New Database, Called As Target Database. Data Migration Is A One-Time Process. It Involves The Re-Structuring Of Data In Some Way: This May Mean Fields Being Merged, Or Formats Being Changed, Or The Data Being Transformed In Various Other Ways. If No-Restructuring Takes Place Then We Would Call This As Simple Data Movement.

Data Migration Phases

The Data Migration Process Generally Carried Out In Three Steps: *Plan, Migrate, And Validation*

In *Planning*, Special Attention Is To Be Given For Old And New Systems, Databases Configurations, As Well As Hardware Specifications. Considering These Factors, Complexity, Risks At Every Stage Of Migration Process Is To Find Out.

In *Migration* Step, Data Is Actually Moved From Older To Newer Database. Effective Data Migration Procedure Is Achieved Through The Mapping Of Old System To The New System Providing A Design For *Data Extraction* And *Data Loading*. The Design Is Constructed Such That It Relates Old Data Formats To The New System's Formats And Requirements. The Data Can Be Migrated Manually Or Using An Automation Tool And May Involve Many Phases But Minimally It Have*data Extraction* Where Data Is Read From The Source Database And *Data Loading* Where Data Is Written To The Target Database.

An Automated ETL (Extract-Transform-Load) Process/Tools Is Use To Move The Data From Source To Target Database. ETL Is The Process Of Data Cleaning, Transforming And Finally Loading It Into A New Target Database. As A Standard Practice The Data Transformations Are Managed In The Data Mapping Document, Which Forms The Base For Development As Well As Testing. In Addition To Mapping The Old Database Structure To The New One, The ETL Tool May Integrate Certain Business-Rules To Enhance The Quality Of Data Moved To The Target Database.

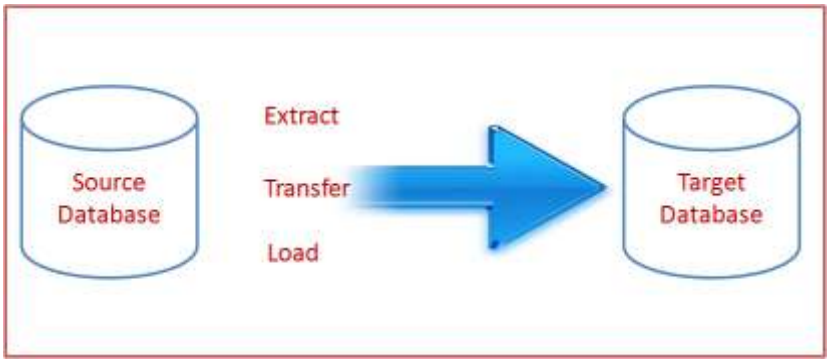

Fig. 1 An ETL Process For Databases

Broadly *Data Loading* Can Be Classified Under Two Types:

1. *Direct Loading* -

The Direct Loading Is Nothing Coping The Data From Source Database To Target Database. Checking Is To Be Performed To Confirm Whether Complete Data Is Moved By Counting And Verifying Number Of Rows From Source To Target.

2. *Incremental Loading*-

In Incremental Loading, Some Constraints Are Imposed While Transformation Of Data. That Is, Cleaning The Data Or Loading The Data With Specific Conditions.

Last But Not Least Step Of Data Migration Is *Validation*. Validating Data Indicates The Comparison Of Original Data Set With The Resulting Target Data Set By Applying A Manual Or Automated Comparison, Or Combination Of Both.

Mapping Document Is One Of The Outcomes Of Data Migration Process. It Consists Of Table-To-Table, Column-To-Column Level Mapping Data Between Source And Target Database. Table Showing Below Is A Sample Of Mapping Document. The Column Fields Of Mapping Document Can Change Based On Database And Its Inner Configuration. Before Data Validation Actual Starts, The Correct Fields Need To Be Match Between Source And Target Databases.

Table 1: Mapping Document

Mapping ID	Key Indicator	Source Table Name	Source Field Name	Source Data Type	Source Field Length	Business Rule	Target Table Name	Target Field Name	Target Data Type	Target Field Length
M001	PK	Emp	Emp_Id	Number	10	Direct Mapping	CUSTOMER_INFO	CUST_ID	Number	15
M002	NA	Emp	Emp_Fname	Text	50	Direct Mapping	CUSTOMER_INFO	CUST_F_NME	Varchar	255

II. Need Of Data Migration

Data Migration Is The Process Of Moving A Data From Existing/ Legacy Source Database To Targeted Newer Database. Though Data Migration Term Sounds Simple And Easy To Understand; It Is The Most Complicated, Risky, Time And Budget Overflowing But Unavoidable Process. There Are Many Reasons For Migrating Data From Old To New Database. Each Is Motivated By The Need To Find New Efficiencies, Better Manage Risk, And Stay Competitive.

- Systems Consolidations: Firms Are Seeking To Reduce Structural Costs By Standardizing On Modern, Cost-Effective Platforms And Technologies; And By Decommissioning And Retiring Inflexible And Hard To Maintain Legacy Applications.
- M&A Activity: Means Infrastructure, Business Processes, And Supporting Systems And Applications Must Be Streamlined. Huge Merger And Acquisition (M&A) Activities Over The Last Few Years 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. Its Replacement Induces The Migration Of The Contained Data [4], [5]. Where Customizations And Quirks Associated With An Earlier Version Of The Underlying Software Mean A Valid Migration Will Not Be Straightforward.
- Ever Changing Legal Regulations Require That Companies To Replace Their Business Applications From Time-To- Time.
- Technological Progress And Upgrades: E.G. Migration To More Commoditized Platforms Or Off-Premise Data Storages [6]– [8].
- In The Industry Like Banking, Automobile, Pharmaceutical Etc. Almost Every Six Months, The Database Get Full And Thus Data Needs To Be Transfer To Another Database.

III. Review Of Existing Methodologies

This Section Will Light On Various Existing Methodologies Of Data Validation Testing. Some Approaches Are Manual While Some Are Automated. Ultimately Data Validation Is A Process To Measure The Data Migration Quality And Will Assure That The Application Will Have The Same Functional Behavior After Migration.

Lines Of Code:

- This Is Simplest And Easily Performed Data Validation Process. Simple SELECT Query Blended With COUNT Operator Is Executed On Individual Datasets Will Work For It. This Validation By Row Counting Is Like Primary Testing. If Row Counts Of Source And Target Databases Are Same, Then It Assumes That Migration Is Successfully Done.
- SELECT Count(Emp_Id) From Emp_Source;
- SELECT Count(Cust_No) From Cust_Target;
- This Method Is Beneficial In Terms Of High Speed, Especially When COUNT Operation Is Performed On Tables Which Have Primary Or Unique Keys. Though Method Is Simple, But Doesn't Verify Data Values, Especially In Migration Which Is Accompanied By The Data Type Transformation.

Sampling Techniques:

- This Technique Assumes That Error Is Uniformly Distributed. Thus, Randomly Any Record Is Peaked From Source DB And Checked Against In Target DB. Since Method Do Not Test All Records, Fails To All Types Of Errors. Even These Comparisons Are Time Consuming With Limited Coverage.

MINUS Queries:

- This Technique Uses `SELECT' Query Along With `MINUS' Operation. The Query Is Executed On Both Source And Target Separately, To Detect What Source/Target Has That Target/Source Lacks. Thus `MINUS' Queries Needs To Be Executed Twice (Source-To-Target And Target-To-Source).
- SELECT Cust_Id
- FROM Cust_Source;
- MINUS
- SELECT Cust_No
- FROM Cust_Target;
- Along With Consumption Of Time And Resources, It Does No Other Validation Like Data Type Mismatch, Null Values, Data Corruption Etc.The Methodologies Described Above Does Not Guarantee 100% Data Coverage, Does Not Detect Various Types Of Data Inconsistencies; Even Process Is Time Consuming.

IV. Literature Survey

The Various Research Papers Are Studied As A Part Of Literature Survey. The Main Focus Of This Study Is:

1. Getting Deeper Knowledge Of Migration Process
2. To Know Which Type Of Testing Specially Can Help To Eliminate Risks In The Process Of Data Migration
3. To Get The Knowledge Of Data Quality Issues And Inconsistencies In Data; And Methods To Eliminate These Errors
4. To Know Currently Used Tools And Techniques In Testing Process Of Data Validation

Following Are The Studied Papers, Really Help In Getting More And More Detailed Knowledge Of Process Of Migration:

[1] Florian Matthes, Christopher Schulz, Klaus Haller, "Testing & Quality Assurance In Data Migration Projects," 27th IEEE International Conference On Software Maintenance (ICSM), 2011

The Paper Is Thorough Study Of Data Migration Process. Authors Elaborate Various Reasons Of Data Migration Projects. The Main Objective That All Migration Projects Have In Common Is To Permanently Move The Data From Source Business Application Into The Target In Implementing A Specific Process.

The Migration Process Model Explain In Paper Consists Of Four Main Stages I.E. *Initiation, Development, Testing, Cutover*; Which In Turn Contain Fourteen Detailed Distinct Phases. The Author Discussion Focuses On The Actual Process Of A Data Migration Project Including Key Deliverables. The Elaborate Data Migration Process Forms A Solid Foundation Allowing For Risk Identification And Discussion Of Appropriate Mitigation Techniques.

There Are Three Levels Of Risks To The Migration Project: The Business Level, The IT Management Level And The Data Migration Level. Combining Data Migration Specific Experience, General Best Practices From The Ever Growing-Body Of Literature In Project Management And Software Development, As Well As Testing Techniques Of The Previous Section Helps To Assure A High Quality And Mitigate The Risks Typically Arising In Data Migration Projects. The Paper Discusses Practice-Based Testing And Quality Assurance Techniques To Reduce Or Even Eliminate Data Migration Risks.

According To The Authors "*83% Of Data Migrationsfail Outright Or Exceed Their Allotted Budgets And Implementation Schedules".*

[2] 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

The Paper Comes Up With The Idea Of Sampling Technique To Do Quality Checks For Huge Data Base Migration. Basically, A Sample Is A Group Of Records/Data Selected From Large Dataset, Which In Turns Represent The Dataset. Then The Samples Are Studied And Based On Them The Judgments Are Made About The Larger Dataset. The Paper Elaborate On Types Of Sampling Like Random Sampling, Systematic Sampling, Stratified Sampling, Cluster Sampling. The Proposed Methodology Take A Random Sample From A Dataset And Then To Use The Information From The Sample Like Mean, Standard Deviation Etc. To Make Conclusions About Particular Dataset. The Methodology Is Simple To Understand And Implement.

The Paper Discusses On Quality Assurance Metrics Used To Data Migration Testing Like Timeliness, Accessibility, Completeness, Integrity, Accuracy, Validity Etc. The Paper Describes As Easily Implemented, Non-Disruptive, Scalable And Comprehensive Foundation For Capturing Data Quality In The Data Migration Work.

[3] Iakov Kirilenko, Eduard Baranov, "Automation Of QA In The Project Of DB Migration From SQL Server Into Oracle", Syrcose Software Engineering Colloquium, 2012

The Paper Is Actual Experience Of QA Process For The DB Migration Project From MS SQL Server 2005 To Oracle 11gr2. The Main Purpose Of The Migration Project Is To Result In New Database, Which Contains The Same Data And Has A Functional Correspondence With The Initial One. The QA Process For Project Is Executed In Main Directions: Generated Code Testing And The Data Migration Process Testing. The Migrated Code Is Tested For Code Functionality, By Trace Recording Method And For Coverage Completeness.

Authors Had Work On Large Database Schema Which Contains More Than 2000 Tables And At Almost 10000 Columns; Some Tables Contain Tens Of Millions Of Records. For Validating Schema, A Row Counting Method Is Used Where Number Of Rows In All Tables Is Count And Results Are Compare. This Method Is Simple To Implement And Work At High Speed; But Method Doesn't Verify Objects Values. Thus Author Proposed A Method Based On Hash Keys Comparison. In The Initial DB Hash Keys Are Calculated For All Columns In Every Table And Results Are Saved In A Separate Table. The Validation Method Using Hash Keys Comparison Has Helped To Improve Data Migration Process And It Can Provide Rather High Probability Of The Data Migration Correctness.

[4] Manjunath T N, Ravindra S.Hegadi And Mohan H. S., "Automated Data Validation For Data Migration Security", International Journal Of Computer Applications (0975-8887)Volume 30- No.6, September 2011

The Paper Gives The Practical Approach Towards Data Validation. The Methodology Is To Automate The Data Validation For Data Migration From Mainframe Machine To DB. The System Developed By Authors Shows

Extra Records In Source/Target Database Using Simple SQL 'SELECT' Query And 'MINUS' Operator; Also Depict A Query To Find Duplicate Records. This Method Is Very Simpler And Fast. The Captured Automation Process Resultsshow 11% Efficiency Than Manual Process. The Benefits From This Process Are Speed, Accuracy And Timeliness, Precision And Other Quality Factors Can Be Achieved.

V. Data Quality Issues

During The Process Of The Data Migration (DM), Too Much Data Is Extracted, Transformed, Structured, And Loaded (ETL) From Legacy/Source Database Into A Newer Structure That Is The Target Database. This Process Leads To Various Types Of Changes In Data, Known As Data Inconsistencies /Quality Issues. Below Table Is Survey On Various Types Of Data Quality Issues That Can Happen In DM Projects.

Table 2: Data Quality Issues

Data Inconsistency	Description	Example
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.
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 Typemismatch	Dissimilarity In Source And Target Data Types.	Source Data Field For Interest Rate Was Float; However, Target Data Field Is Set To Int.
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.
Misplaced Data	Source Data Fields Not Being Transformed To The Correct Target Data Field	A Source Data Field Was Supposed To Be Transformed To Target Data Field 'Last_Name'. However, The Development Team Unintentionally Mapped The Source Data Field To 'First_Name'
Extra Records	Records Which Should Not Be In Target Database; Are In The Target Database	Target Database May Have Records Or Data Values Which Are Not Present In Source Database.

Typically, Data Migration Is A Part Of Various Big Projects, Where Migrating Or Upgrading Database Is A Key Requirement; And Hardly Perform As A Single Project. Even, Regularly Data Is Not Moved As It Is; Rather Many Business Rules, Normalization Concepts Are Applied.

VI. Conclusion

The Above Comparative Study Suggests For The Room For An Automated Tool Which Is Robust, Agile And Platform Independent Which Could Be Used To Configure, Design And Execute Test Cases.The Tool Should Support Data Validation Testing Between The Database Data Sources Like Mysql, Oracle, SQL Server, MS Access; Also File Data Sources Such As Delimited File(CSV), Fix Length File. Applying Primary Data Validation Filter Before Data Comparator Will Be The Optimized Solution To Lessen TC Execution Time. The Objective Of Data Validation Testing Should Be To Compare Target Database Against Source; And Retain The Exactly Same Copy Of Source Data Into Target Database, Along With Applied Business Rules. Source And Target Database Records Are Compared Based On Key (Record Identifier). Such System Can Handle The Data Quality Issues And Other Issues Of The Existing Data Migration Systems More Efficiently.

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] G. Schroder,¨ "Automatisierte Migration Von Legacy Daten," Diplomar-Beit, Fakultat¨ Fur¨ Informatik, Technischeuniversitat¨ Munchen,¨ Garchingbeimunchen,¨ Germany, 2006. R. E. Sorace, V. S. Reinhardt, And S. A. Vaughn, "High-Speed Digital-To-RF Converter," U.S. Patent 5 668 842, Sept. 16, 1997.
[6] C. Burry And D. Mancusi, "How To Plan For Data Migration," 2004.
[7] IBM, "Best Practices For Data Migration - Methodologies For Assessing, Planning, Moving Andvalidating Data Migration," Somers, NY, USA, P. 16, 2009.
[8] K. Haller, "Data Migration Project Management And Standard Software – Experiences In Avaloq Implementation Projects," In *Data Warehousingconference - DW2008: Synergiendurch Integration Und Informationslo-Gistik*, St. Gallen, Switzerland, 2008, Pp. 391–406.