International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 13 (2018) pp. 11316-11323

Performance Analysis of improvement in Clustered Data Migration using Machine Learning Approaches

S.Venkata Prasad
Research Scholar, Manonmaniam Sundaranar university, Tirunelveli, Tamilnadu, India.

Dr.K.G.Srinivasagan
Professor, Department of Computer Science & Engineering,
National Engineering College, Kovilpatti, Tamilnadu, India.

Dr.V.Vivek
Assistant Professor (SG), Department of Computer Science & Engineering,
Sethu Institute of Technology, Virudhunagar, Tamilnadu, India.

V.Manimaran
Assistant Professor (Senior Grade), Department of Information Technology,
National Engineering College, Kovilpatti, Tamilnadu, India.

Abstract
In this digital world, every digitalized organization wants to maintain old, complex and precious customer data. No organizations want to leave old customer data in the old system and start new system with empty data, during migration. Maintaining historical data is the pillar of organizations strategic growth and Analytics. Organizations need to go for data migration when changing new application system or replacing storage unit or moving data to third party cloud. Migration of historical and live data to new system is the fundamental for the success of organization. Data loss is one of the risk during data migration. In a recent survey, 66% of IT professionals stated that data loss or corruption was the greatest fear when migrating data. Enigmatic victims may act on stealing the customer confidential data during migration, this illegal copy could be business confidential data or customer's personal data. This study provides method of secured data migration without data loss by defining clustering model and developing machine learning artificial intelligence algorithm [1].

Keywords: Database, Artificial intelligence, Data migration, Technology upgrade, Acquisition, Applications upgrade, Performance improvement, machine learning.

INTRODUCTION

The world is a dynamic place. Standards and technology are evolving continuously. In this energetic proceeding change of business. Company Merger and acquisition, technologies upgrade and applications upgrade are the frequent movement in market. Daily many applications are being introduced in the market. These new applications have several advanced features when comparing to previous versions. New releases have features like performance improvement, user interface improvement, advance data base technology, web interfaces etc. So the enterprises need to transformation themselves in this competitive business world for financial business growth and to gain customer satisfaction. Data need to be migrated to new application or database migration, to adapt these changes for organizational growth. Data loss and data theft are the major threat during traditional migration [2]. Our approach is to develop a system to perform data migration between any system without data loss and data theft. The new system will also learn or gain knowledge by itself during migration by adopting machine learning methodology.

DATA MIGRATION FRAMEWORK

Data migration is the process of transferring data between data storage systems [3]. When organization plans for data migration, the basic expectation from the enterprise is, same set of data existed in the source system need to be transferred successfully to target system. Organization plans for data migration for the following reasons, an upgrade of existing hardware, transfer to a completely new system, hardware platform upgrade, upgrading a database or migrating to new software, company-mergers when the parallel systems in the two companies need to be merged into one. In the manual data migration, Team manually move disk files from one folder (or computer) to another, database insert queries, developing custom software, or other methods. The specific method used for any particular system depends entirely on the systems involved and the nature and state of the data being migrated. In traditional method of data migration Figure.1, Source data is prepared in text file format, by using source code [4]. The same text file is interpreted by the designated system and data will be moved to respective tables. This method is commonly used by various systems. In this method various steps have been involved in each stage for successful data migration. Initially data migration architect, need to explore the existing source system and target system architecture, to identify the necessary data for transfer. Next the quality of data need to be analyzed like duplication, incorrect data etc., which is called as Data profiling. Then Migration structure has to finalized and build the migration file for data transfer. The intermediate flat file is moved or transferred to new system which is called data migration execution. In the final transition stage, new system will interpret the data and move to respective tables.

Table 1. Field structure in sending and receiving system

Source System	Value	Receiver System	Value
First Name	Suresh	Cardholder Name	Suresh Raja
Last Name	Raja	Card Type	VISA
Card Type and Card Number	VISA 0123 456 8910 1234	Card Number	0123 4567 8910 1234
Expiration Date	May-12	Expiration Date	May-12

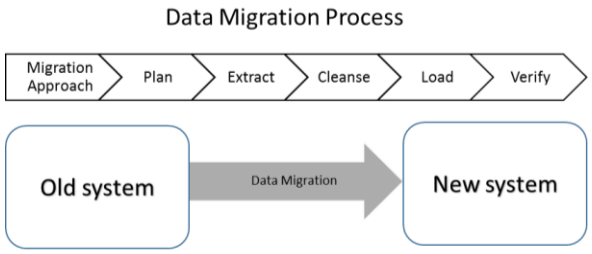

Figure 1. Data Migration Process

Data transformation strategy – Machine learning

Data will be stored in different databases formats across the enterprise as shown in Figure 2. Data need to transform from one application to other applications, during application or date base migration as shown in figure 3. Data transformation [5] is necessary to ensure data from one application or database is transformed to receiver applications in receiver databases format. Source and receiver system will have different data structure and it may vary between applications. Data structure in both systems need to be determined for data mapping as a first process in data transformation. Data mapping for both structure and unstructured data, from source system to receiver system has to be done either manually or by using programming language.

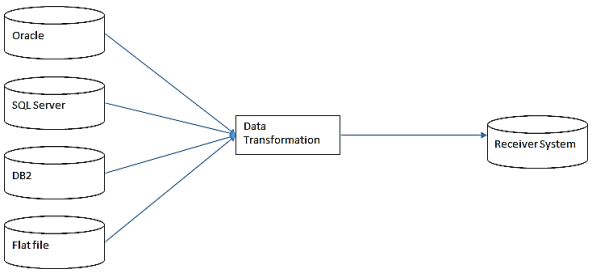

Figure 2. Data Transformation Methodology

Machine learning Data transformation algorithm learn from the data transfer as shown in Figure 4. This algorithm need to train by providing right format of Data. Data preparation for machine learning involves exploration, analysis and iterative process. Machine learning algorithm performance will improve based on the quality of data provided to algorithm. Initially existing data structure and new system data structure need to analyze for preparing the structure of date for transfer.

Data to be transferred from the existing system need to be cleaned, formatted for feeding into data transformation machine learning algorithm. Set of protocols trained to machine learning algorithm will process the data and perform data transformation to new system. Data migration machine learning algorithm will automatically learn and upgrade protocols based on the data provided for transformation.

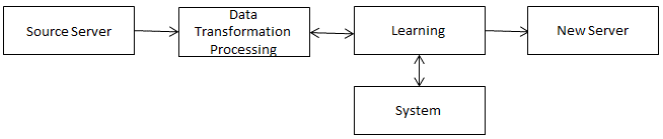

Figure 3. Machine Learning during Data Transformation

Analysis on Data Migration Tools

To Migrate unprecedented growth of enterprise data, various tools and application are available. Each and every tools will have its own protocol migration strategy. Some tool focus on data base specific, and few tools give priority for volume of data and so on. Some applications are released by database vendor, which will be specific for their supplied data base.

Machine Learning

Machine learning is the science of developing and training the system to learn automatically and gain more knowledge from the experience without programming or data feed. The initial process of machine learning is observations, Rule database will be built with effective machine learning algorithm [6], which can be used to analytics and taking better decision. The primary goal of machine learning is to allow the system to learn, gain knowledge automatically without human intervention or assistance and adjust actions accordingly. Machine learning algorithm can be categorized as supervised machine learning algorithms, unsupervised machine learning algorithms, Semi-supervised machine learning algorithms, Reinforcement machine learning algorithms.

Machine learning over manual rule creation

Machine learning and rules-based systems are used to building expert system. Rules-based systems are from artificial intelligence which use a series rules to guide a system to reach a conclusion or recommendation [7]. A rules-based system is designed by set of facts which is known as knowledge base, set of rules which is known as the rules engine. Rule-based systems are constructed from the domain experts. The domain experts build all the protocols for making decision and to take special scenario decision with expert knowledge. As the set of rules well defined by the domain expert, design rule based system stress-free.

Maintaining of few application is financial burden or additional expense for enterprise. Below table listed the details study of advantages and disadvantages of various tools available in the market.

Table 2. Comparison of various Data migration software's

Product Name	Company	Advantages	Disadvantages
Alooma	Alooma , USA	Partnerships with data warehouse providers	No on-premise option for companies requiring that deployment option. Costly for some basic extract and load needs.
IBM (Information Server Infosphere platform)	IBM	progress towards common metadata platform	Difficult learning curve. Long implementation cycles. Became very heavy (lots of GBs) with version 8.x and requires a lot of processing power.
Informatica PowerCenter	Informatica	Focus only on B2B data exchange	Several partnerships diminishing the value of technologies. Limited experience in the field.
Microsoft (SQL Server Integration Services)	Microsoft	Standardized data integration	Problems in non-Windows environments. Takes over all Microsoft Windows limitations. Unclear vision and strategy.
Oracle (OWB and ODI)	Oracle	Tight connection to all Oracle applications	Tools are used mostly for batch-oriented work, transformation rather than real-time processes or federation data delivery.
SAP Business Objects	SAP	SAP Business Objects created a firm company determined to stir the market	Uncertain future. Controversy over deciding which method of delivering data integration to use (SAP BW or BODI). Business Objects Data Integrator (Data Services) may not be seen as a stand-alone capable application to some organizations.
SAS	SAS , USA	Great support for the business-class companies	Misplaced sales force, company is not well recognized. SAS has to extend influences to reach non-BI community. Costly one.
Sun Microsystems	Sun Microsystems	Single-view' services draw together data from variety of sources; small set of vendors with a strong vision	Relative weakness in bulk data movement limited mindshare in the market support and services rated below adequate.
Sybase	Sybase	Broad partnerships with other data quality and data integration tools vendors	Falls behind market leaders and large vendors gaps in many aspects of data management.
Syncsort	Syncsort , USA	Easy implementation, strong performance, targeted functionality and lower costs	Struggle with gaining mind share in the market. Lack of support for other than ETL delivery styles. Unsatisfactory with lack of capability of professional services.
Tibco Software	Tibco Software Inc	Support for federated views; easy implementation, support and performance	Scarce references from customers; not widely enough recognized for data integration competencies. Lacking in data quality capabilities
ETI	ETI software Solutions	one of the earliest vendors on the data integration market; support for SOA service-oriented deployments;	Relatively slow growth of customer base rather not attractive and inventive technology.
iWay Software	iWay Software, Co.	Reasonable ease of implementation effort	Gaps in specific capabilities relatively costly - not competitive versus market leaders.

Pervasive Software	Pervasive Software	Good use of metadata	Inconsistency in defining the target for their applications. No federation capability. Limited presence due to poor marketing.
Open Text	SKYVIA	Easy licensing for business solutions	Limited federation, replication and data quality support; rare upgrades due to its simplicity. Weak real-time support due to use third party solutions and other database utilities.
Pitney Bowes Software	Pitney Bowes	Ease of use, fast implementation, specific ETL functionality	Rare competition with other major companies, repeated rebranding trigger suspicions among customers. Narrow vision of possibilities even though. Data Flow comes with variety of applications. Weak support, unexperienced service.

Machine learning algorithm has few advantages over rules-based methods, instead of using defined rules from expert, machine learning methods learns outcomes from the experts. Machine learning is probabilistic and uses statistical models rather than defined rules.

In Machine learning learned knowledge and result are based on the historical real time learned world. In Machine learning, few scenarios like unavailability of real data or lack of training to system or lack of data provided to the system for learning may result to incorrect decision.

Clustered Structure											
Table	Data for Learning Analysis						Data for reporting				
Field Values	Sender	Receiver	User	Authorization	Source Table	Designation table	No of records	Processed	To process	Failed	Reason

Figure 4. Modeled Cluster used for Data Migration

Clustered Data for Migration

Most of the database migration project success ratio is less, when compared to other project completion status. More human efforts are involved in data migration. The data migration without data loss and with less human intervention is need in the industry.

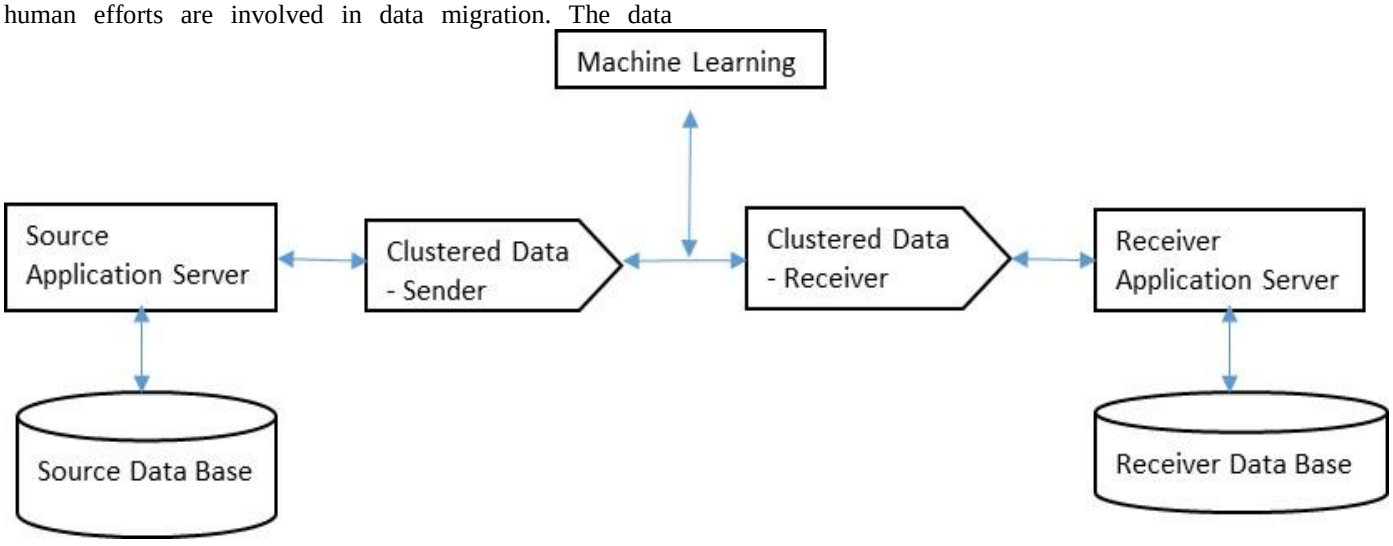

Figure 5. Architecture of Clustered Data Migration using Machine learning (CDMML)

Many studies are ongoing for the successful completion of data migration project. Even after demonstrating research results in data migration, still organization hesitate to initiate data migration process. This paper proposes a method named Clustered Data Migration using Machine learning (CDMML) Figure 5 which will enable to migrate data between any two types of applications irrespective of database, database type, platform and application with less human intervention using machine learning algorithm [8]. Figure. 5 shows the proposed architecture of data migration using CDMML. In this approach as a first step, database architecture in the target system need to be analyzed for forming cluster structure. This analysis includes structures, data type, mandatory fields and incremental fields. After the analysis is completed, the clustering format of data need to be finalized. This clustering data will be used to transfer data between source and target systems.

This molded cluster structure will contain table data, user control information for machine learning and transfer status

records. This clustered structure is created in both source and target systems will play interface role to transfer data between the systems. Source and Receiver date structure is shown in the Figure 6. The defined cluster structure is shown in below Figure 7.

To initiate migration, data in the source table need to be populated in the cluster structure. This structure has source table values, learning details like sender, receiver, quantity, quality of data, validation details etc., Data transfer will be initiated only if learned or trained rule matches with source and receiver system. If mismatch in both systems, then transfer will not be initiated, and update with error details in the status record. All this validation logic need to be implemented in both the systems, while training or learning rules. This status details holds migration status like success and failure details of the document. In the receiver system, cluster with same structure need to be created.

Connectivity between two systems need to be established as shown in the below Figure 8. Initiate data transfer between the systems by sending clustered records as shown in below Figure 8. During data transfer or after completion of transfer status records need to be monitored to identify the failure or success rate. Failure reasons can be identified from the status record and administrator needs to fix the issues to complete migration without loss.

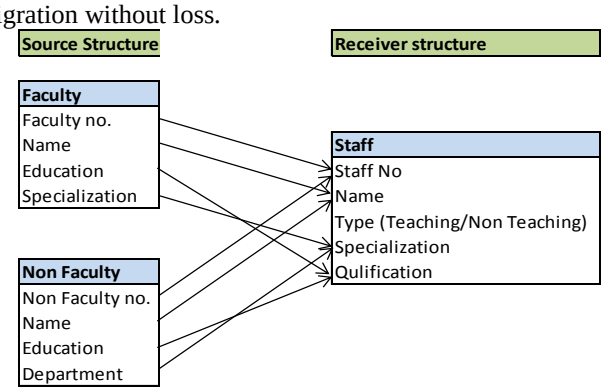

Figure 6. Data Structure in Sending and Receiving system

Table	Learning Data						Status - Reporting				
Field Values	Sender	Receiver	User	Authorization	Source Table	Designation table	No of records	Processed	To process	Failed	Reason
Faculty no. Name Education Specialization	Admin_Old	Admin_New	Principal	All	Faculty	Staff	110	60	50	5	Spl Characters
Sending System											
↕											
Receiver System											
Staff No Name Type (Teaching/Non Teaching) Specialization Qulification	Admin_Old	Admin_New	Principal	All	Faculty	Staff	110	60	50	5	Spl Characters
Field Values	Sender	Receiver	User	Authorization	Source Table	Designation table	No of records	Processed	To process	Failed	Reason
Table	Transferred - Machine learning algorithm						Status - Reporting				

Figure 5. Clustered Data Migration from Source to Receiver System

To successfully implement machine learning algorithm for data transfer, data migration process owner need to set the system for learning with required details. Initially set of rules need to be created and need to give training to the algorithm for learning. Expert team need to frequently review the rules learned by the machine to ensure the correctness. Rules learned in the system are dependent to this process, this cannot be transferred to another data migration process as shown in Figure 9.

Data Set

College system decided to migrate to new system, which has more advanced features. New proposed CDMML methodology is planned to be implement for this migration. After detailed analysis of both data base format, cluster structure need to be developed in source and receiver system as shown in Figure 10. In the old system, data planned for migration need to be populated into the cluster structure by using the application. Similar cluster need to be prepared in the receiver/target system. The cluster populated need to be transmitted to new system via established connection; this cluster used for transfer will be interpreted by the intermediate machine learning algorithm and authorize the date transfer to new system and application in the new system move the received data to respective tables in the new system.

Cluster structure shown in below Figure 10 need to be prepared in the source system and transmitted to receiver system. In Figure 10 cluster data is prepared with data in table Faculty from source system. The data populated in cluster structure in Figure has first record in table. Similarly cluster need to be populated for all records in the table. The field values will be change for each record. Intermediate machine learned migration algorithm will interpret the cluster data and authorize the data transfer. Status records will be updated automatically based on the migration status for each cluster.

In Figure 11, formatted document is prepared with values or data in table Non Faculty. Once all the data from first table reached the target system, we need to check the status record, if any error updated in the record, it needs to be fixed and initiate re-migration the data, once data migration completed for first table, similar migration need to be initiated for second table. In the receiver system application place the received

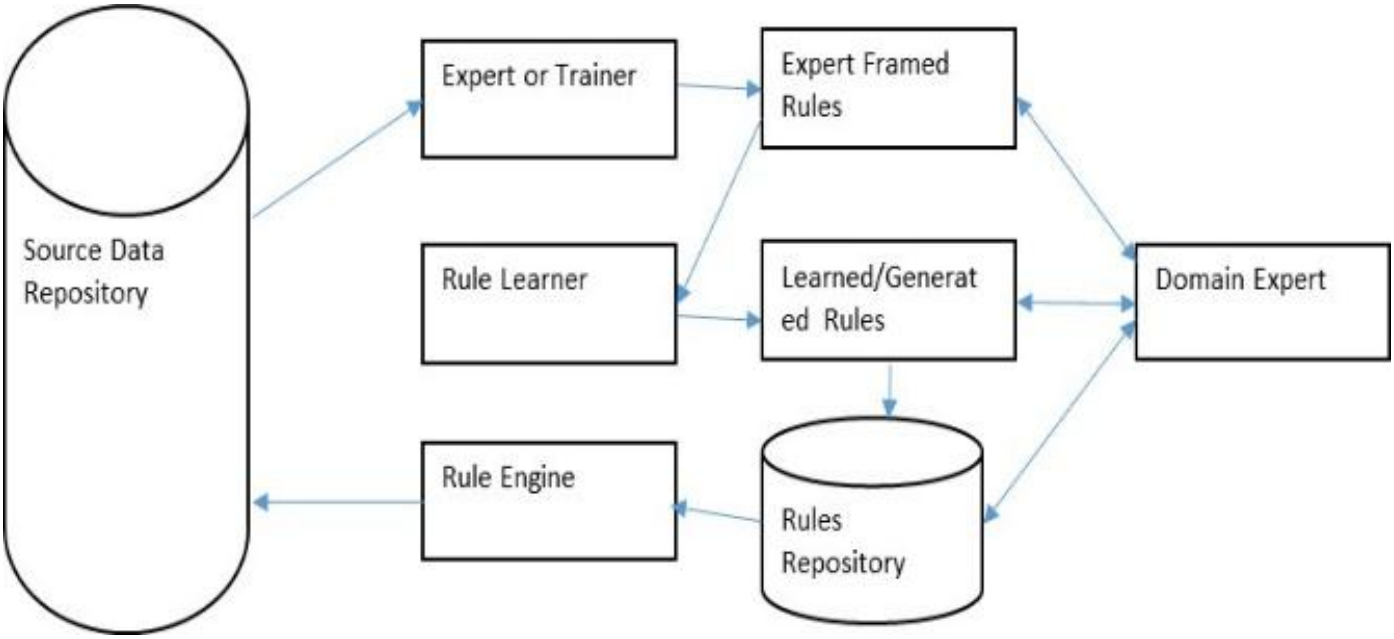

Figure 8. Machine learning during Data Migration Process

data in the respective tables. In this method all the data will be migrated successfully without any data loss. Records migrated will in control by the machine learned algorithm designed in The rule created by machine learsning algorithm as shown in figure 12 need to be reviewed frequently by data migration process owner. System should allow data migration process this methodology. Error records can be tracked and reprocessed successfully.

owner to edit the rules if required. This scenario will occur if any incorrect entry in the learned data. Below table listed the machine learned set by using algorithm.

Table	Learning Data						Status - Reporting				
Field Values	Sender	Receiver	User	Authorization	Source Table	Designation table	No of records	Processed	To process	Failed	Reason
1 Mark Tech Maths M.Sc	Principal	officeAdmin	Principal	Administrator	Non Faculty	Staff	4	4	0	1	Auth Fail
Old Application Non Faculty Table - Sending System											
New Application Staff Table - Receiver System											
1 Mark Tech Maths M.Sc	Principal	officeAdmin	Principal	Administrator	Non Faculty	Staff	4	4	0	1	Auth Fail
Field Values	Sender	Receiver	User	Authorization	Source Table	Designation table	No of records	Processed	To process	Failed	Reason
Table	Transferred - Machine learning algorithm						Status - Reporting				

Figure 9. Cluster mapping between Source 2nd Table and Receiver system

Machine learned Rule Set					
Sender	Receiver	User Type	Data Type	Authorization	Authorization Rule
Office Admin	Office Admin	Admin	FEES	Y	IF text conv to Num
Chairman	Chairman	Admin	ADMISION	Y	All rule
Principal	DATAEntry	Entry	FEES	N	Not allow revenue report
HOD	HOD	HOD	Mark	Y	Allow mark, Student info
AsstPROF	AsstPROF	HOD	Mark	Y	Allow mark, Student info
PROF	DATAEntry	Entry	Mark	N	Not allow Mark report
DATAEntry	DATAEntry	Entry	Student	Y	Allow only Entry

Figure 10. Machine Learned Rule

EXPERIMENT RESULT

New proposed experimented CDMML method is compared Various parameters like manual intervention, Security, data loss are taken into consideration for comparing with other algorithms. In this new proposed method, due to adaption of machine learning concept, human involvement is very much reduced. In this method, as authentication implemented in transfer status, data theft or incorrect transfer is completely stopped and due to reprocessing method data loss during migration is minimized.

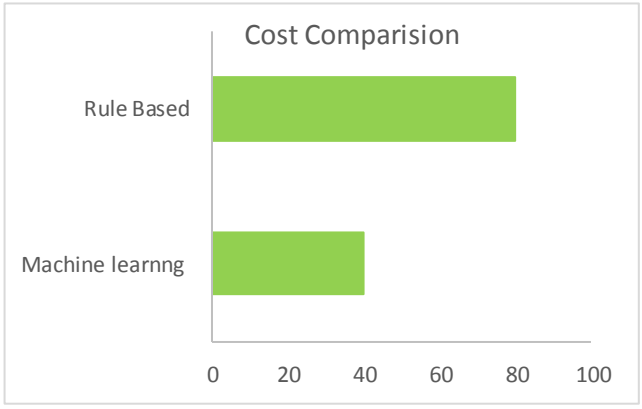

Figure 6. Cost Comparison between Rule based and Machine Learning Data Migration Process.

Machine learning algorithm against rule based reduces the manual efforts involved in data migration. To build the set of rules for data migration, the migration team need to set up an expert team to build the set of rules and migration strategy. In this machine learning concept, the rules are framed by the machine learning algorithm, which is a real time rules. The salary paid for the consultant is expensive is reduced in this method as shown in figure 13.

Developing algorithm is one-time investment and not a repetitive expense. Progress of rule learned is shown in figure 14.

with other data migration algorithms available in the market, across various database platforms.

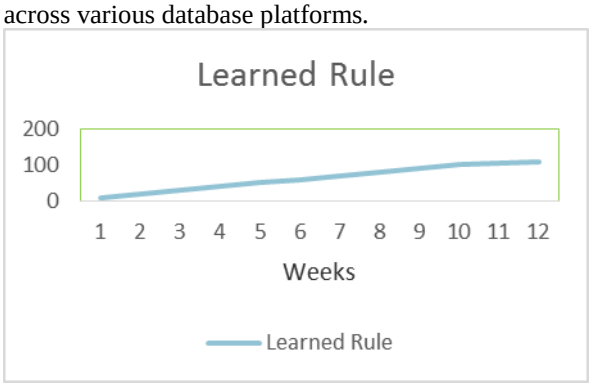

Figure 7. Machine Learning Progress

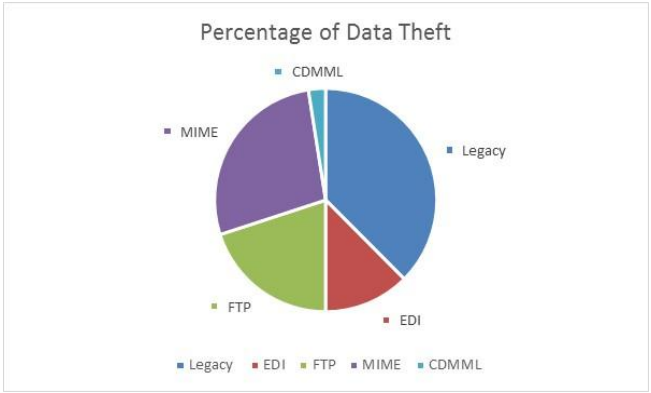

Figure 8. Comparison of Data Theft during Data Migration

During data migration security is one of the most important threat during migration, customer sensitive data like personal need to be transferred with more caution. The legal requirement need to follow during data migration. More stipulated algorithm need to be developed for data migration. Extensive authenticated check implemented during transfer and receiving will ensure the secured data migration. Study of Data theft of various algorithms is shown in figure 15.

Data migration team need to minimize the data loss and data corruption during data migration. Source system and receiver system may have different data format in data base and fields, so some times data might not be received by receiver system. To avoid these data loss data migration testing need to executed to avoid this data loss, is again an additional expense to the project team. Status and reprocess method implemented in proposed algorithm, will reprocess the failure the data and data loss during migration is completely reduced as shown in figure 16.

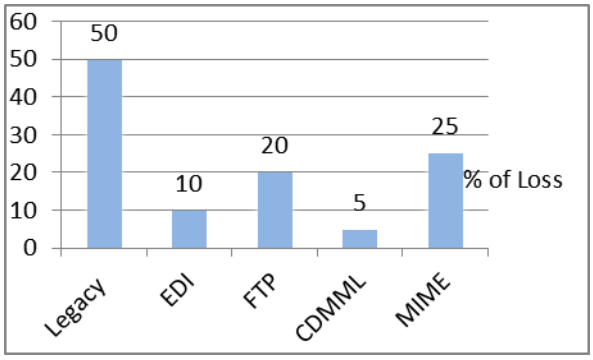

Figure 9. Comparison of Data Loss during Migration

CONCLUSION

Various methods and algorithm has been analyzed with various parameters like cost effectiveness, performance, security, data loss. Proposed model will be cost effective as we proposed machine learning algorithm, cost for data migration consultant is removed to form data migration rule set. Authentication check suggested in sender and receiver system will enable secured transfer between sender and receiving system. Status reports will help to resend the failed data to minimize the data loss. In this paper we analyzed and compared the proposed methods and other methods and provided the result in experiment section.

In the future analysis, we will try to migrate the data available in the hard copies, this method provides solution for transferring data from one database to another database. Processing hard copy and converting to data base required Image processing concept to be included at the source data base analysis stage. We will attempt this proposed solution with converting hard copy image to designation database.

REFERENCES

[1] Kazakov, D. and Kudenko, D. (2001). Machine learning and inductive logic programming for multi-agent systems. In Luck, M., Marik, V., Stepankova, O., and Trappl, R., editors, Multi-Agent Systems and Applications. Springer-Verlag.

[2] DATA MIGRATION BEST PRACTICES etApp Global Services January 2006.

[3] Alan, Robert. (2002). The Serials Data Migration Dilemma. Technical Services Quarterly 20 (4), 29-38.

[4] Manikandan S. Preparing to analyse data. J Pharmacol Pharmacother. 2010;1:64–5

[5] Bland JM, Altman DG. Transforming data. BMJ. 1996;312:770.

[6] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1, 81-106.

[7] H. Liu, A. Gegov and F. Stahl, Unified Framework for Construction of Rule Based Classification Systems. In: W. Pedrycz and S. M. Chen (eds.) Information Granularity, Big Data and Computational Intelligence, Studies in Big Data 8, Springer, pp.209-230, 2015.

[8] Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2, 139-172.