MULTI OWNER DATA SHARING AND DUPLICATE DETECTION

P.Bhavani[#1] V.Karpagavalli [*2] S.Thamilselvi[*3] and R.K.Shyamala[*4]

[#1]*Assistant Professor, Department of Computer Science and Engineering, Manakula Vinayagar Institute of Technology, Madagadipet, Pondicherry, India.*

[*2*,3*,4] *Under Graduate Student, Department of Computer Science and Engineering, Manakula Vinayagar Institute of Technology, Madagadipet, Pondicherry, India.*

bhavani.mit@gmail.com
vidhya.mithun1221@gmail.com
sanju.sriniv18@gmail.com
shyamala141995@gmail.com

***Abstract*— An infrastructure build in the data mining and communication platform which is reliable to challenge the commercial and non-commercial IT development communities of data streams in high dimensional data cluster modeling. Duplicate detection is the technique to detect multiple replications and redundancies of practical world entities. Duplicate content needs to be caught in shorter time for huge databases in shortest time possible; to maintain the integrity of the databases available is really a difficult task. We present here two progressive duplicate identification technique techniques which are efficient enough to find the duplicity in as minimum time possible with us. Our technique is efficient enough to increase the database efficiency and increase the gain of the overall process. It takes small time compared to traditional processes. The proposed technique for our system is the ETL(Extract, Transform and Load) Tool is the SSIS(Sql Server Integrated Services) which performs best as per our result.**

***Index Terms*—Data Cluster, Modeling, Duplicate, Database.**

I. INTRODUCTION

Data mining is the analysis step of the "information identification in databases" process, or KID. The term is a misnomer, because the goal is the extraction of patterns and identification from bigger databases, not the extraction (mining) of data itself. Data mining is an intra-domain subfield of computer science. It is the computational process of discovering patterns in bigger database involvings.

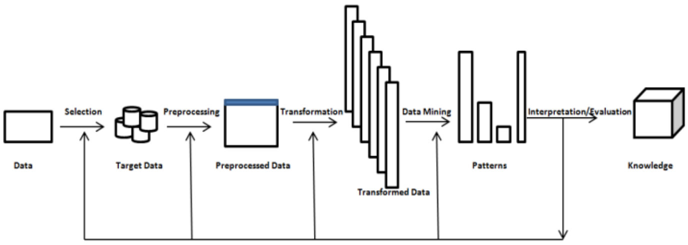


Fig.1: Overview of the KDD Process

at the intersection of artificial intelligence, machine learning details and database detailing. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics , complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "Knowledge Identification in Database" process, or KID.

II. PROPOSED WORKING PROCEDURE

A. ETL (Extract, Transform and Load) Tool is the SSIS (Sql Server Integrated Services)

In the proposed system, We suggested that the ETL (Extract, Transform and Load) Tool is the SSIS (Sql Server Integrated Services).

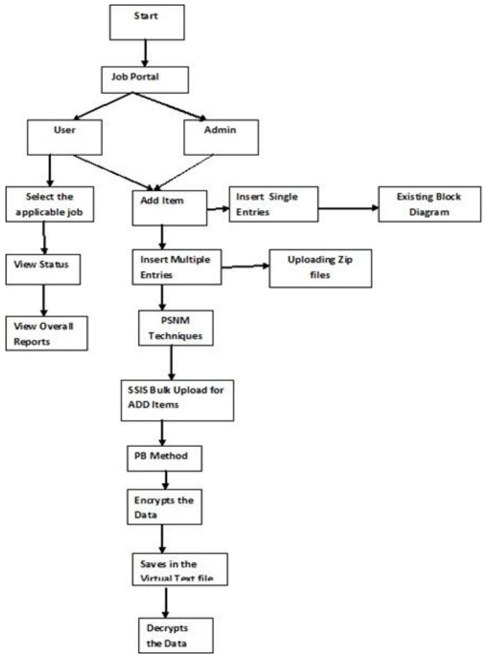


Fig.2: System Flow Diagram

It performs best on high level and almost to clean datasets, and avoiding redundant to verify each and every record in the database, which performs best on large and very dirty datasets. This technique is to improve the efficiency of duplicate detection even on very large datasets. The duplicate detection workflow consists of the three actions pair-selection, pair-wise evaluation, and clustering. For a modern work-flow, only the first and last phase needs to be customized. Therefore, we do not examine the evaluation of the phase and recommend techniques that are separate of the high quality of the likeness operate. Our techniques developed normally used techniques, organizing and preventing, and manipulating the detection are required to be categorized near to one another or arranged in same fields, respectively.

The load for my data warehouse has recently been moved from a series of stored process running in a SQL to running in an SSIS package to take advantage of the parallelism to the training the instructor indicated that for most tasks SQL can be quicker and more flexible.

B. Module Descriptions

System Modules for our proposed system are as follows

1. Web Services
2. SSIS Bulk Upload
3. Performance Optimization
4. Data Automatic Extraction(DAE)
5. Extracting Details (SSRS Report)

Web services

We had been proved that several actual life Web solutions allow asking for only one discussion of a connection, but not for the other. We have suggested using details removal to think bindings for the feedback factors and then confirm these bindings by the Web support. Through this strategy, a whole new type of concerns has become tractable. We have proven that offering inverse features alone is not enough. They also have to be prioritized accordingly. We have applied our program, PROGRESSIVE, and revealed the credibility of our strategy on actual information places. We believe that the attractiveness of our strategy can be found in the successful symbiosis of details removal and Web solutions.

SSIS Bulk UPLOAD

The Bulk Insert task can transfer data only from a text file into a SQL Server table or view. To use the Bulk Insert task is used to track and switch from other database management systems (DBMSs), you must export the data from the source to a text file and then import the data from the text file into a SQL Server table or view.

The destination must be a table or content of the SQL databases. If the destination table or view already contains data, the new data is appended to the existing data when the Bulk Insert task runs. If you want to replace the data, run an Execute SQL task that runs a DELETE or TRUNCATE statement before you run the Bulk Insert task. For more information, see Execute SQL Task.

You can use a format file in the Bulk Insert task object. If you have a format file that was created by the **bcp** utility, you can specify its path in the Bulk Insert task. The Bulk Insert task supports both XML and non-XML format files. For more information about format files, see Format Files for Importing or Exporting Data (SQL Server).Only members of the sys-admin fixed server role can run a package that contains a Bulk Insert task.

Performance Optimization

To optimize performance, consider the following:

If the text file is located on the same computer as the SQL Server database into which data is inserted, the copy operation occurs at an even faster rate because the data is not moved over the network.

The Bulk Insert task does not log error-causing rows. If you must capture this information, use the error outputs of data flow components to capture error-causing rows in an exception file.

Data Automatic Extraction

Data Automatic Extraction (DAE) is involved with getting organized information from records. DAE methods experience from the natural imprecision of the removal process. Usually, the produced information is way too loud to allow direct querying. PROGRESSIVE triumphs over this restriction, by using DAE completely for finding applicant organizations of interest and providing these as information into Web service phone calls. Known as Progressive Duplicate Deletion (PSNM) techniques aim to identify exciting organizations in written text records. They can be used to produce applicants for PROGRESSIVE. The first strategy mentioned in this document suits noun words against the titles of organizations that are authorized in a understanding – a simple but effective technique that circumvents the disturbance in learning-based PSNM techniques.

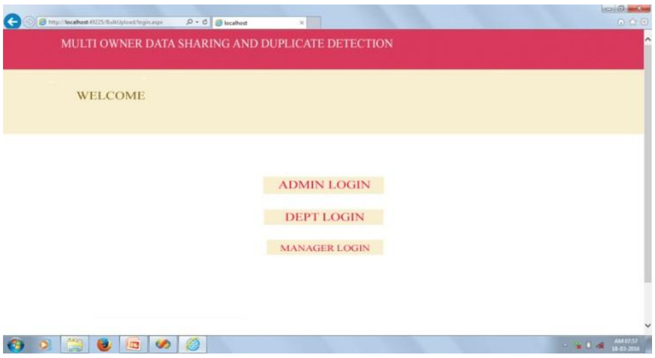
Extracting Details (SSRS Report)

Once the Web pages have been retrieved, it remains to extract the candidate entities. Information extraction is a challenging endeavor, because it often requires near-human understanding of the input documents. Our scenario is somewhat simpler, because we are only interested in extracting the entities of a certain type from a set of Web pages.

The load for my data warehouse has recently been moved from a series of stored process running in a SQL to running in an SSIS package to take advantage of the parallelism to the training the instructor indicated that for most tasks SQL can be quicker and more flexible.

III. SAMPLE OUTPUT SCREENSHOTS

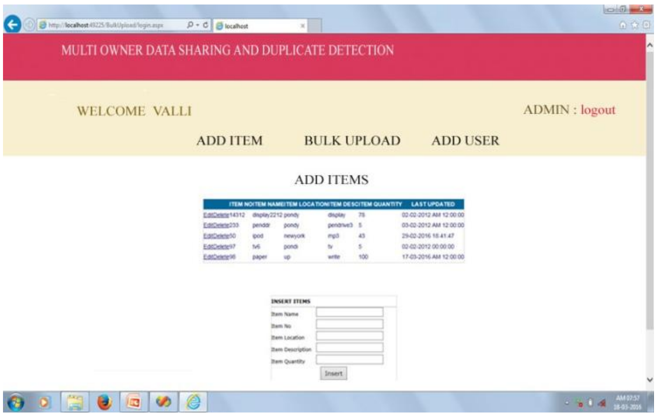
A. *Figures and Tables*



MULTI OWNER DATA SHARING AND DUPLICATE DETECTION

WELCOME

Fig.1: Login Page



MULTI OWNER DATA SHARING AND DUPLICATE DETECTION

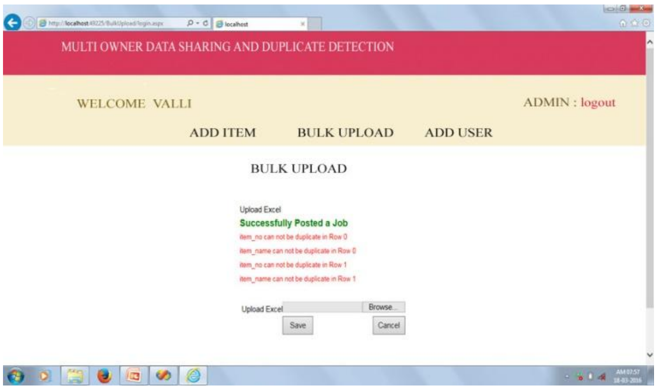
WELCOME VALLI ADMIN : logout

ADD ITEMS

	ITEM NO	ITEM NAME	ITEM LOCATION	ITEM DESC	ITEM QUANTITY	LAST UPDATED
EditDelete	14312	display2212	pondy	display	75	02-02-2012 AM 12:00:00
EditDelete	233	penddr	pondy	pendrive3	5	03-02-2012 AM 12:00:00
EditDelete	50	ipod	newyork	mp3	43	29-02-2016 18:41:47
EditDelete	97	tv6	pondi	tv	5	02-02-2012 00:00:00
EditDelete	98	paper	up	write	100	17-03-2016 AM 12:00:00

INSERT ITEMS
Item Name
Item No
Item Location
Item Description
Item Quantity

Fig.2: Admin Login (ADD ITEMS)



MULTI OWNER DATA SHARING AND DUPLICATE DETECTION

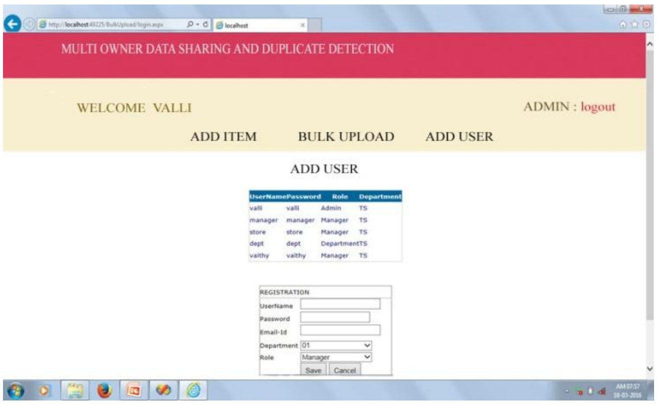
WELCOME VALLI ADMIN : logout

BULK UPLOAD

Upload Excel
Successfully Posted a Job
item_no can not be duplicate in Row 0
item_name can not be duplicate in Row 0
item_no can not be duplicate in Row 1
item_name can not be duplicate in Row 1

Upload Excel

Fig.3: Bulk Upload



MULTI OWNER DATA SHARING AND DUPLICATE DETECTION

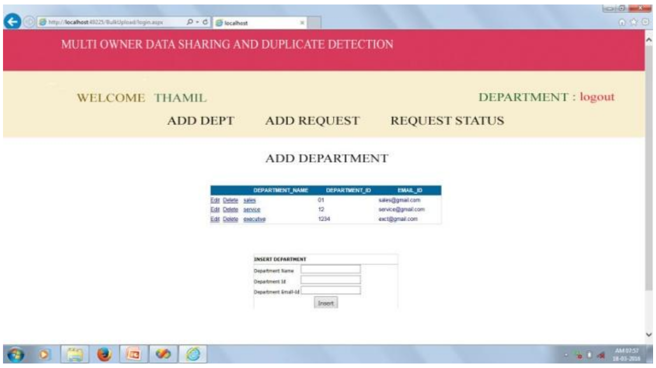
WELCOME VALLI ADMIN : logout

ADD USER

UserName	Password	Role	Department
valli	valli	Admin	TS
manager	manager	Manager	TS
store	store	Manager	TS
dept	dept	Department	TS
vaithy	vaithy	Manager	TS

REGISTRATION
UserName
Password
Email-Id
Department 01
Role Manager

Fig.4: User Creation Module



MULTI OWNER DATA SHARING AND DUPLICATE DETECTION

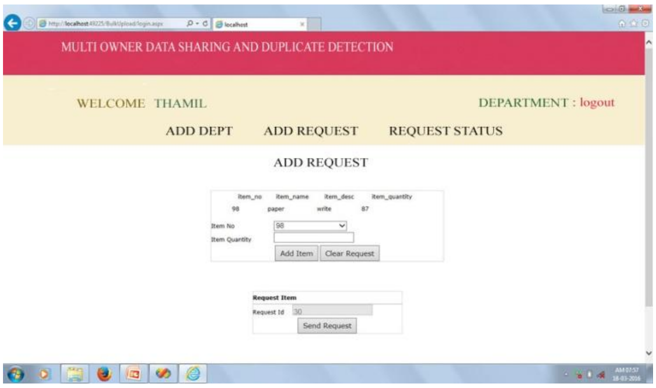
WELCOME THAMIL DEPARTMENT : logout

ADD DEPARTMENT

		DEPARTMENT_NAME	DEPARTMENT_ID	EMAIL_ID
Edit	Delete	sales	01	sales@gmail.com
Edit	Delete	service	12	service@gmail.com
Edit	Delete	executive	1234	exct@gmail.com

INSERT DEPARTMENT
Department Name
Department Id
Department Email-Id

Fig.5: Department Login (Dept Addition)



MULTI OWNER DATA SHARING AND DUPLICATE DETECTION

WELCOME THAMIL DEPARTMENT : logout

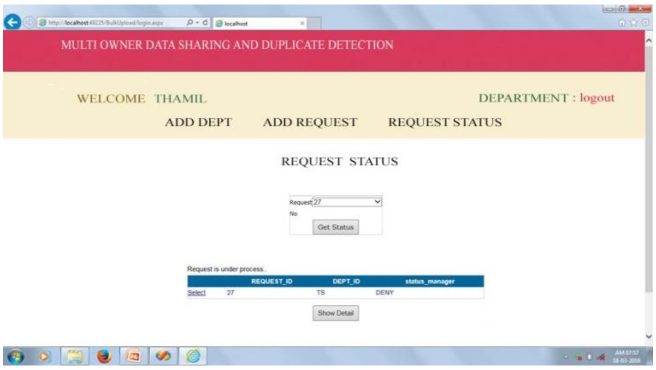
ADD REQUEST

item_no	item_name	item_desc	item_quantity
98	paper	write	87

Item No 98
Item Quantity

Request Item
Request Id 30

Fig.6: Add Request



MULTI OWNER DATA SHARING AND DUPLICATE DETECTION

WELCOME THAMIL DEPARTMENT : logout

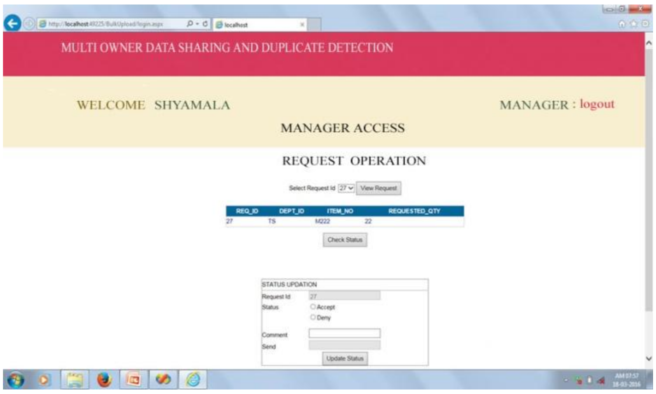
REQUEST STATUS

Request No 27

Request is under process..

	REQUEST_ID	DEPT_ID	status_manager
Select	27	TS	DENY

Fig.7: Request Status



MULTI OWNER DATA SHARING AND DUPLICATE DETECTION

WELCOME SHYAMALA MANAGER : logout

MANAGER ACCESS

REQUEST OPERATION

Select Request Id 27

REQ_ID	DEPT_ID	ITEM_NO	REQUESTED_QTY
27	TS	M222	22

STATUS UPDATION
Request Id 27
Status ○ Accept ○ Deny
Comment
Send

Fig.8: Manager Login (Request Operation)

IV. CONCLUSION

This project is to focus the ETL tool for SSIS Packages to increase the efficiency of duplicate detection for situations with limited execution time. This dynamically changes the ranking of comparison candidates based on intermediate results to execute promising comparisons first and less promising comparisons later. To determine the performance gain of our algorithms, we proposed a novel quality measure for progressiveness that integrates seamlessly with existing measures. We proposed a progressive sorting method, Magpie, a progressive multi-pass execution model, Attribute Concurrency, and an incremental transitive closure algorithm. In future work, we want to combine our progressive approaches with scalable approaches for duplicate detection to deliver results even faster. In particular, Kolb et al. introduced a two phase parallel SNM, which executes a traditional SNM on balanced, overlapping partitions. Here, we can instead use our PSNM to progressively find duplicates in parallel.

Acknowledgment

We thank our **Dr.K.B.Jayarraman HOD, Ph.D (Department of Computer Science and Engineering)** to help us for creating this paper with his sincere guidance and Technical Expertise in the field of communication. The help of our guide **Mrs.P.Bhavani M.E, Department of CSE, Manakula Vinayagar Institute of Technology** is really immense and once again we thank her for her great motivation. We thank Manakula Vinayagar Institute of Technology to provide us such a standard educational environment so that we are able to understand the minute concepts in the field of Engineering.

REFERENCES

[1] S. E. Whang, D. Marmaros and H. Garcia Molina, "Pay-as-you-go entity resolution", *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 5, pp. 1111-1124, 2012.

[2] L. Kolb, A. Thor and E. Rahm, "Parallel sorted neighborhood blocking with MapReduce,"*Proc. Conf. Datenbanksysteme in B¿¿ro, Technik und Wissenschaft*, 2011

[3] J. Madhavan, S. R. Jeffery, S. Cohen, X. Dong, D. Ko, C. Yu and A. Halevy, "Web-scale data integration: You can only afford to pay as you go", *Proc. Conf. Innovative Data Syst. Res.*, 2007

[4] U. Draisbach and F. Naumann, "A comparison and generalization of blocking and windowing algorithms for duplicate detection," in Proceedings of the International Workshop on Quality in Databases (QDB), 2009.

[5] M. A. Hernández and S. J. Stolfo, "The merge/purge problem for large databases," in Proceedings of the ACM International Conference on Management of Data (SIGMOD), 1995, pp. 127-138.

[6] P. Singla and P. Domingos, "Object identification with attribute-mediated dependences," in European Conference on Principles of Data Mining and Knowledge Discovery (PKDD), 2005, pp. 297-308.

[7] S. Yan, D. Lee, M.-Y. Kan, and L. C. Giles, "Adaptive sorted neighborhood methods for efficient record linkage," in Proceedings of the ACM/IEEE-CS joint conference on Digital libraries (JCDL), 2007, pp. 185-194.

[8] P. Christen and K. Goiser, "Quality and complexity measures for data linkage and deduplication." in Quality Measures in Data Mining, ser. Studies in Computational Intelligence, 2007, vol. 43, pp. 127-151.

[9] D. Menestrina, S. Whang, and H. Garcia-Molina, "Evaluating entity resolution results," Proceedings of the VLDB Endowment, vol. 3, no. 1, pp. 208-219, 2010.

[10] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James, "Automatic linkage of vital records." Science, vol. 130, pp. 954-959, 1959.

BIBLIOGRAPHY

Bhavani.P was born in Pondicherry on 4th September 1987. She received her Bachelor of Engineering in computer science and engineering from Sri Manakula Vinayagar Engineering College, Pondicherry University, Pondicherry, India in 2009. Master of Engineering in computer science and engineering from Sri Manakula Vinayagar Engineering College, Pondicherry University, Pondicherry, India in 2011. She is currently working as Assistant Professor in Manakula Vinayagar Institute of Technology, Puducherry. She has published seven research papers in refereed journals.

V.Karpagavalli pursuing B.E degree in Computer science and engineering from Manakula Vinayagar Institute of Technology, Puducherry. Her research interests include Data Mining.

S.Thamilselvi pursuing B.E degree in Computer science and engineering from Manakula Vinayagar Institute of Technology, Puducherry. Her research interests include Data Mining.

R.K.Shyamala pursuing B.E degree in Computer science and engineering from Manakula Vinayagar Institute of Technology,Puducherry. Her research interests include Data Mining.