HEURISTIC APPROACH FOR CUSTOMER DATA INTEGRATION

Seema Lute[1] and Prakash R Devale[2]
[1]Scholor and [2]Professor & Head,
Department of Information technology, Bharati Vidyapeeth Deemed University, Pune, India

ABSTRACT

In today's world maintaining users or customers data by giant industries like telecom, Insurance etc. is a very cumbersome task and more over allowing some different domain to access this data is even a hard task than the prior one. So integrating these data on common platform so that it can be shared by the different data requesters and provider is a quite an appreciation task. However, maintaining and integrating quality customer data is one of the greatest challenges it executives face—and this challenge only gets more daunting as businesses both grow and become more complex. Economic slowdown has fuelled customer data integration (CDI) of different enterprises under a single roof. Today, more or less all enterprises have CDI solution, which interacts closely with the ERP, CRM systems. Hence CDI is very critical and will act as catalyst to ERP, CRM system integrations to a large extent. Success in CDI will also ensure reduced impact on customer –the top most priority during the integration. In our approach of CDI batched stream processing (BSP) is incorporated, where it is a distributed data processing paradigm that models recurring batch computations on incrementally bulk-appended data streams. The model is inspired by our empirical study on a trace from a large-scale production data-processing cluster at the web server end; it allows a set of effective query optimizations that are not possible in a traditional integration model. And also selective encryption model to preserve the privacy of the customer data makes the system more effective and enriched.

KEYWORDS: *Customer Data Integration (CDI), Batch Stream Processing (BSP), Selective Encryption. Privacy Preserving.*

I. INTRODUCTION

As yet, there is no universal set of CDI standards. Today's CDI standards are a function of individual organizational needs. CDI may be characterized as "the combination of processes, controls, automation and skills necessary to standardize and integrate customer data originating from different sources." And "a comprehensive set of technology components, services, and business processes that create, maintain, and make available an accurate, timely, integrated and complete view of a customer across lines of business, channels, and business partners." CDI is technically a subset of MDM (Master Data Management) which comprises a set of processes and tools which consistently define and manage the non-transactional data entities of an organization. CDI and MDM however share a common logical approach. Both integrate data from across different sources. Both document data lineage and data evolution over time. Both strive to achieve single "golden" records which consolidate data and eliminate duplication of information. MDM is often perceived as covering a broader spectrum of data.

However, in reality, although initially focused on customer data, CDI solutions can cover much of the same ground. The essential construct is the same—a truly robust CDI solution can be readily expanded to include larger MDM applications by moving beyond customer data to include that of other key parties.

While the primary reason most companies pursue a CDI strategy is profitability related, CDI often provides other important benefits as well.

1.1 Compliance

Many enterprises are subjected to different levels of regulatory. All of these initiatives require a solid data foundation. These more stringent requirements mean that businesses often need to retain more extensive customer data and to have better data access and control. Today's requirements often mean that companies need to increase:

- Data accuracy and timeliness
- Traceability of transactions for audit trails
- Point-in-time accountability

1.2 Fraud Detection

CDI enables improved customer analysis, and with that, the potential to better protect the organization from fraud. By using an actively managed, central store of customer data, organizations can gain real-time insight into the identity of new applicants—and better detect behavior patterns indicative of possible issues.

1.3 Privacy Preserving

CDI enables organizations to enact programs that enhance customer loyalty and improve retention. Through cooperative integration across business units, it enables the organization to eliminate duplication of efforts, expedite data updates, and present a more unified and effective face to its customers. Effective CDI thus requires a willingness of all business units to share in customer ownership: the goal is to create and maintain a single, comprehensive record for each customer that reflects the entirety of each customer's relationship with the corporation.

1.4 Reducing costs

Recent surveys indicate that repetitive, one-off integration efforts absorb approximately 30% of development projects' budgets. By reducing data and systems redundancies, CDI makes data processing and application systems more efficient—and helps to significantly reduce these costs. That, however, is just the beginning. In the process, CDI provides efficiencies that often translate to dramatically faster time-to-market for new products and services. Once organizations integrate customer data well and continue to follow up on that integration to keep records current and complete can avoid time-consuming project-by project integration efforts.

In our proposed system we developed a CDI system which is a web service. It is designed and assumed to receive the data from many sources like passport department, Bank, RTO and Police department etc. Our system integrates the customer data on a single respiratory system which comes through various data sources using batch stream processing technique, so that the data requesters can get the access in less time. System maintained data privacy by the selective encryption using reverse cycle cipher technique and CDI privacy policies which enriches the concept of CDI to a good extent.

The rest of the paper is organized as follows. Section 2 discusses some related work and section 3 presents the design of our approach. The details of the results have conducted on this approach are presented in section 4 as Results and Discussions. Sections 5 provide the conclusion and future scope to our system.

II. RELATED WORK

Data integration is a big challenge in database industry. It faces two major problems which are the structural and semantics diversities of source schemas that to be merged. Semantic similarities and differences are difficult to recognize and resolve, it needs to understand the intended meaning of a concept for each element ([1], [2]). The semantic conflict of data during the data integration can be in terms of naming conflict (homonym and synonym), type conflict, key and cardinality conflict, and etc. Most of the previously developed integration systems, such as Tukwila [3], DIXSE [4], LoPix [5], LSD [6] and DIKE [7], aim to handle integration of XML databases.

Even though these systems have achieved certain results, however they still have some limitations. For instance LSD [6] used neural network technique to integrate the schemas from different data sources. DIKE [7] only focused on the structural different of the source data. Most of the existing approaches are not focused on solving the problems of naming conflict during the integration.

The workflow is (for present purposes) assumed to be a sufficient and complete representation of the transformation of data inputs to data outputs of CDI. It consists of the following kinds of representational constructs:

- Files and fields – a file consists of records; records consist of fields; fields are typically named and can be thought to have a data type as well as contents (a fixed width or variable number of bytes).[8] Notices that a field name could identify a virtual field in a virtual record, which is built by a workflow.
- Auto-generated Workflow – an earlier research project [9-13] demonstrated, given an input set of fields, and output set of fields, and a set of operators, how to automatically generate all possible workflow graphs that connect the inputs to the outputs via the operators. The prototype, though interesting, was limited in several ways – it was limited to one file input and default options of operators.
- An operator, input fields, output fields, parameters an operator takes as input a collection of input fields and returns a collection of output fields. Generally, CDI operators return additional fields adding to the input fields the newly computed output fields. Parameters can govern the exact behavior of an operator even including which fields are to be generated. Operators must be defined somehow – they can be defined in a general purpose programming languages.
- Workflow and macros – a workflow can be viewed as a graph of operators (specifying in detail which parameterized operator operates on which fields) so that input fields (the original data) are operated on by a sequence (graph) of operators until final output fields are generated. If we treat each operator as a basic data transformation instruction, macros can be viewed as subroutines – they are similar to operators in that they can take parameters but they are also similar to (mini) workflows in that they can contain graphs of operators and macros.

And we studied some CDI systems like “The SPARK Solution Developer Program” is designed to help solution providers quickly build data integration capabilities between their solutions and CRM, as well as any other application or endpoint on Scribe Online. Using Scribe’s latest offering, SaaS independent software vendors (ISVs) who offer integration to more than one CRM vendor can extend their presence in multiple CRM markets. As customers expand the scope of CRM in their businesses, integration can readily incorporate the SaaS ISVs’ offerings with connections both to CRM and to other complementary applications.

Whereas Scribe Software is the leader in CRM data integration solutions, helping businesses maximize their investments in CRM, ERP, industry applications, and other data assets. With over 12,000 customers and 1,000 partners worldwide, Scribe is a proven provider of cost-effective, reliable data solutions that give a competitive advantage to businesses large and small. With a range of offerings covering cloud, hybrid, and premise integration needs, Scribe has solutions that span a wide array of industries including Financial Services, Life Sciences, Manufacturing, and Media & Entertainment.

III. PROPOSED SYSTEM

In this section, we describe our approach of Customer data Integration System with a heuristic approach for the steps shown in figure 1. As shown in figure there are 5 main steps in our approach.

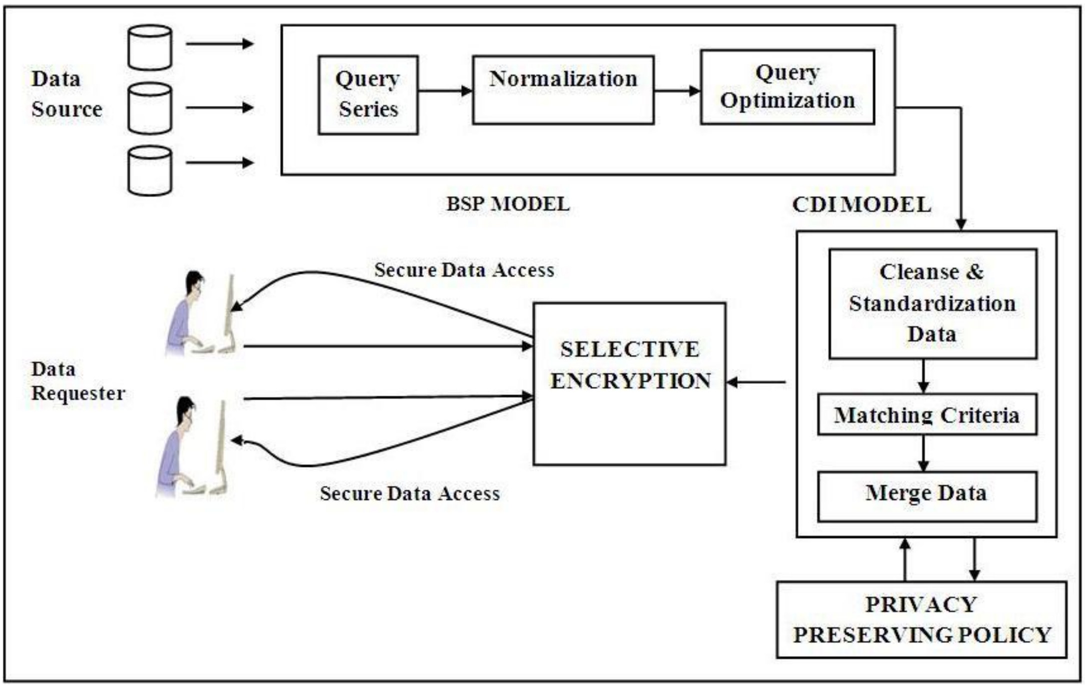

Figure 1. Overview of our approach

Step 1: Here in this step many data sources with different attribute field are submitted for CDI to our system. Where these data is preprocessed and then passed to BSP system.

Step 2: BSP system actually takes the data queries in linear manner and finally provide a batch of data having same characteristics to commit the transaction to reduce the time drastically with the following steps

- Query Series: Here the Queries are collected in many vectors to form a complex query handling structure.
- Normalization: All the vectors are normalized with a single or multiple characters with the tags like *insert, update, delete* and *select*
- Query Optimization: All the normalized vectors are batched and tagged to different recursive procedures to conclude the committing transactions.

Step 3: Here CDI is done with the following steps

- **Cleanse & Standardize Data:** When analyzing data, consider source attributes that would get benefit from data cleansing and standardization rules. Cleanse lists are intended to facilitate data conversion during the staging process to ensure that the data that ends up in the staging table is in a standardized, consistent format.
- **Matching Criteria: -** During data analysis, identify which columns are appropriate for matching. For example, if a gender column is null 80% of the time, then this column is probably not a column to use in a match rule.
- **Merge Data:** Here system automatically aggregates one or more records into a single record. Auto merge rules are often based on exact matches across key fields and near matches on secondary fields.

Step 4: Here a privacy policy is applied on the integrated data to maintain the confidentiality of the true data owner based on the data requester's type.

Step 5: Here selective encryption is implement based on the privacy policy using reverse cycle cipher encryption [14] Reverse Circle Cipher uses 'circular substitution' and 'reversal transposition' to exploit the benefits of both confusion and diffusion. This scheme uses an arbitrarily variable key length which may even be equal to the length of the plaintext or as small as a few bits coupled with an arbitrary reversal factor.

IV. RESULTS AND DISCUSSIONS

Our system was made using Java Language and the NetBeans 6.9.1 IDE with Apache Tomcat server and MySQL server as database. And in our proposed system we used 5 different data sources each of having1000 records, which collectively fed the data to our CDI system in a given instance. As the system is enriched by the BSP system it drastically reduces the time to integrate records in a common database respiratory. The performance of our system has shown below in the figure no 2. Where X – axis indicates the number of data records in thousands and Y-axis represents the time required to integrate in millisecond. As graph is clearly declares our system is quietly enriched in performance time.

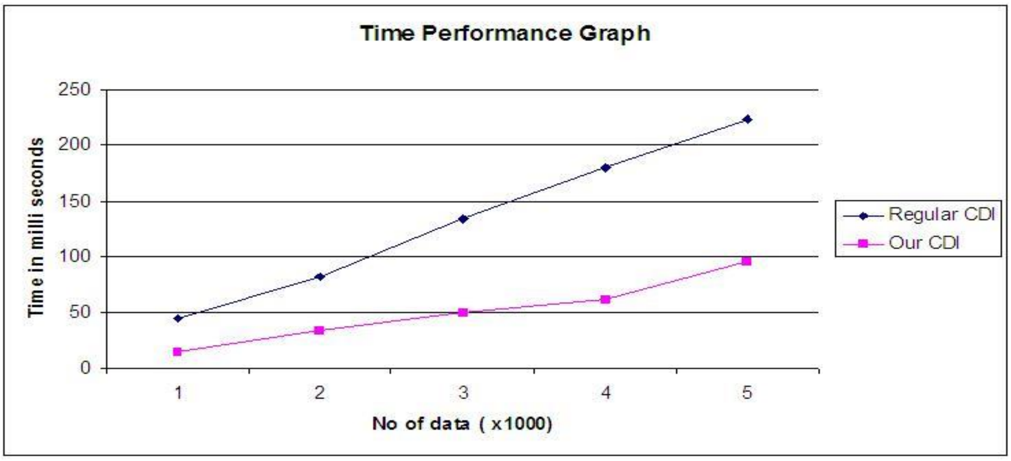

Figure 2. Time Performance Graph

V. CONCLUSIONS AND FUTURE WORKS

Our system of customer data integration successfully integrates the data of different domain fields on a common platform to provide accurate and deeper data tuple for the data requesters and data providers.

To get more descriptions, our system first collects the customer data from five different assumed domains like passport, insurance, telecom, bank and law enforcement .Where each domain is considered as a department which actually enters (one thousand data record) the data using common CRM. Our system collects this data and store in a common respiratory system using BSP model. Privacy of the data is successfully preserved by the selective encryption using reverse cycle cipher technique for the data requesters.

As a future work for this CDI concept we are planning to integrate the data of the different journal publications or of the different colleges with the live data tuple on a common platform. Where we can provide excellent data privacy, smart access to the data with the greater accuracy.

ACKNOWLEDGEMENTS

I express my deep sense of gratitude and my research guide Prof. P. R. Devale for his continuous inspiration and valuable guidance in throughout my project work.

REFERENCES

[1]. O. Unal and H. Afsarmanesh,(2009) "Schema matching and integration for data sharing among collaborating organizations," Journal of Software, vol. 4, p. 248.

[2] O. Unal and H. Afsarmanesh, (2010)"Semi-automated schema integration with SASMINT," Knowledge and Information Systems, vol. 23, pp. 99-128.

[3] R. Pottinger and A. Levy. (2000) “A scalable algorithm for answering queries using views”. Proc. of 26th VLDB Conference, Cairo, Egypt, 484–495.

[4] P. Gianolli, J. Mylopoulos. (2001) A semantic approach to XML based data integration. Proc. Of the 20th. International Conference on Conceptual Modelling (ER), Yokohama, Japan.

[5] P. Bellstrom.(2005) Using Enterprise Modeling for Identification and Resolution of Homonym Conflicts in View Integration. *Information Systems Development: Advances in Theory, Practice and Education*: 265-276.

[6] A. Doan and A. Halevy, (2005) "Semantic integration research in the database community: A brief survey," AI magazine, vol. 26, p. 83.

[7] I. Palopoli, G. Terracina, and D. Ursino.(2003) A System Supporting The Semi-Automatic Construction of Cooperative Information Systems From Heterogeneous Databases. *Softw. Pract. Exper*: 847-884.

[8] T. Talley, "Research Proposal :(2006) Developing a Domain-Specific Modeling Language (DSML) for Customer Data Integration (CDI) Tasks," ALAR RFP Problem Description.

[9] C. Thompson, W. Li, C. Bayrack, (2004) "A Framework to Automate the Generation of BCDI Process Flows," *Applied Research in Information Technology,* Acxiom Lab for Applied Research (ALAR), Little Rock, AR. Acxiom champion: Bob Ludwig

[10] Z. Xiao, C. Thompson, W. Li,(2006) "A Practical Data Processing Workflow Automation System in Acxiom Grid Architecture," *International Workshop on Workflow Systems in Grid Environments*(WSGE06) , Changsha, China.

[11] Z. Xiao, C. Thompson, W. Li, (2006) "Automating Workflow for Data Processing in Grid Architecture," *International Conference on Info and Knowledge Eng.* (IKE'06), Las Vegas, NV.

[12] C. Thompson, W. Li, Z. Xiao, (2007) "Workflow Planning on a Grid," Architectural Perspectives column, *IEEE Internet Computing*, January-February.

[13] C. Thompson, (2006) "A Framework to Automate the Generation of BCDI Process Flows," ALAR Project Continuation.

[14] Ebenezer R.H.P. Isaac, Joseph H.R. Isaac and J. Visumathi,"Reverse Circle Cipher for Personal and Network Security".

BIOGRAPHY

Seema Lute, PG Scholar in information Technology at Bharati Vidyapeeth Deemed University, Pune. Her fields of interest are Customer Data Integration, and Database.

Prakash Devale, presently working as a professor and Head department of Information Technology at Bharati Vidyapeeth Deemed University College of Engineering, Pune. He received his ME from Bharati Vidyapeeth University and pursuing Ph.D degree in natural language processing.