International Journal of Database Theory and Application
Vol. 11, No.1 (2018), pp. 21-28
http://dx.doi.org/10.14257/ijdta.2018.11.1.03

Logical Data Integration Model for the Integration of Data Repositories

Waseem Jeelani Bakshi[1,*], Rana Hashmy[2], Majid Zaman[3]
and Muheet Ahmed Butt[4]

[1,2,4]Department of Computer Sciences, University of Kashmir
[3]Directorate of IT&SS, University of Kashmir
[1]waseembakshi@gmail.com, [2]ranahashmy@gmail.com,
[3]zamanmajid@gmail.com, [4]ermuheet@gmail.com

Abstract

Physical integration of heterogeneous data sources is too high a price for small and medium level enterprises. This is primarily because of the variance in data base management system (MySQL, MsSQL, Oracle, etc.), underlying operating systems (Linux, Windows, etc.), besides numerous legacy sources (DBase, Foxplus, etc.). Enterprises are spending huge amounts for establishing data warehouses while as the data of interest increases gradually [Harmeek Bedi]. Over the years enterprises have accumulated numerous data sources along with legacy repositories which have variant data, schema and context and their integration is increasingly difficult. In this paper we propose a LDIM (Logical Data Integration Model) which will describe data, schema and context for integration of heterogeneous data sources including legacy repositories.

Keywords: *Heterogeneous Data, Legacy Sources, Logical Data Integration Model*

1. Introduction

With the advent of new information systems in recent times, enterprises are facing a big challenge of coping up with these different but coexisting information systems. The information contained in these diverse information systems is required for decision making which is vital for realizing business opportunities in the highly competitive markets. The information being scattered across different systems needs to be integrated and presented to the end users in a unified view as if it were a single information system. Inspite of the fact that the information is physically distributed over heterogeneous data sources the users should be provided with a homogeneous logical view of the data.

The heterogeneity problem arises when the concerned databases are autonomous with respect to their design which may involve differences in terms of terminology, domains, data types, scope, *etc*. To overcome this heterogeneity and integrate such diverse data sources we need to represent all the relevant data using a unified global data model which can describe the data, schemas and the contexts and at the same time support a mediated approach for integration of databases and legacy systems.

Information around us is growing in volume at a very fast rate. This is due to the fact that today the use of database applications has become very common in almost every enterprise. Every enterprise, small, medium or large uses databases to store and retrieve their relevant data owing to the ease with which they can perform these tasks using database applications. However, the problem being faced by most of the enterprises is that the data resides in multiple disparate sources. This may be due to many reasons including

Received (February 12, 2018), Review Result (March 6, 2018), Accepted (March 15, 2018)

[*] Corresponding Author

addition of new independently developed data systems to an existing enterprise, acquiring different data sources as a result of mergers and acquisitions, upgradation to a new system in terms of platform, software, *etc*. The result is that the exchange of data between these diverse heterogeneous data sources becomes very difficult. In order to use this data efficiently and judiciously it is important to find a solution to the problem of integration of this heterogeneous data.

Moreover, this data is useful only if it is easily and timely available to the end user. However, the end users, who are generally naive users, are dependent on a handful of people like designers, programmers and top-level managers to get their desired information from this rich source of information. As such a mechanism is required which will allow naive users to easily derive useful and relevant information from the otherwise complex collection of data sources. Usually the search engines used on the web retrieve information using the keyword(s) resulting in the generation of a list of relevant documents ranked on the basis of relevance. A similar mechanism could be provided to the users in an enterprise for transparently retrieving relevant information from the information sea stored in diverse databases.

The heterogeneity in data exists in the form of system, syntax, structure and semantics. Over the years there have been multiple approaches to solve the information integration challenges posed by these heterogeneous data environments. The focus of these solutions has shifted from the multi-database integration (Federated Databases) to the heterogeneous data integration. As far as the heterogeneity based on system, syntax and structure is concerned many solutions have been devised including CORBA, DCOM, MULTIBASE, OBSERVER, SIRUP, *etc*. However, a lot of work needs to be done in the area of semantic heterogeneity.

2. Heterogeneous Data

Most of the enterprises have disintegrated data sources, having different DBMS (Oracle, MsSQL, MySQL, *etc.*,) and varying development tools (JSP, PHP, ASP) and thus in this process most of the enterprises end up having multiple autonomous heterogeneous data sources and of course massive data sets. Now data is all around us, most of the time structurally different and in numerous data sources including in files and legacy sources. Every passing day is adding more data and its management is becoming complicated for enterprises. Almost all enterprise users are directly or indirectly at the mercy of programmers to view/extract data of their interest [6][7].

According to Alon V. Halevy [11] there are many potential circumstances where semantic heterogeneity may arise, including:

- Enterprise information integration
- Querying and indexing the deep web
- Merchant catalog mapping
- Schema vs. data heterogeneity
- Schema heterogeneity and semi-structured data.

These along with many other sources in simple schema use and versioning create mismatches. Halevy further states that the possible drivers in semantic mismatches can occur from world view, perspective, syntax, structure, versioning and timing:

- one schema may express a similar “world view” with different syntax, grammar or structure
- one schema may also be a new version of the other
- multiple shames may be derived from the same source schema

- there may be many sources modelling the same aspects of the underlying domain (“horizontal resolution” such as for competing trade associations or standard bodies), or
- there may be many sources that cover different domains but overlap at the same (“vertical resolution” such as between pharmaceuticals and basic medicine).

With numerous heterogeneous data sources all around enterprises, industry ended up having data integration tool commonly known as Data Warehouse. Data Warehouse is a central data repository of integrated data from numerous heterogeneous autonomous disintegrated disparate data sources. It stores historical data and every day data is extracted from operational sources into one single place and is used by end users of the enterprise for knowledge extraction, analysis and decision making [8-10].

Despite being the cofounders of the data warehouse, Kimball and Inmon gave different designs of the data warehouse. Inmon proposed the dependent data mart structure whereas Kimball gave the data warehouse bus structure.

3. Data Warehouse

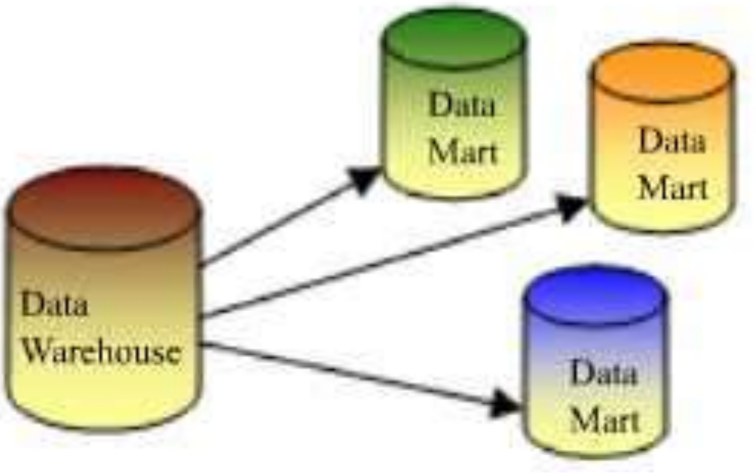
The term data warehouse was coined by William H. Inmon, who is known as the father of Data Warehousing. Inmon described a data warehouse as being a subject-oriented, integrated, time-variant and nonvolatile collection of data that supports decision-making process of the management. Typically, a data warehouse is housed on an enterprise mainframe server or in the cloud. Data from various online transaction processing (OLTP) applications and other sources is selectively extracted for use by analytical applications and user queries. A data warehousing aggregates structured data from different sources so that it can be efficiently utilized for making strategic decisions thereby enhancing business intelligence. Data warehouses enable users to correlate broad business data for better analysis of the available data that enhances corporate performance.

The design of data warehouses is different from standard operational databases. They are designed to give a long-range view of data over time. Data warehouses are employed to do the analytical work, leaving the transactional databases free to take care of the transactions. The data warehouses can analyse data from multiple sources. However, data warehouses are not able to handle raw, unstructured, or complex data.

In a data warehouse the data is obtained from different operational systems relevant to an enterprise. This data passes through various stages before it can be used for reporting. Generally, a typical data warehouse involves a staging layer that stores raw data extracted from each data source, an integration layer that integrates this disparate data using transformations and stores the transformed data in data warehouse database which stores the data in the form of hierarchical groups called dimensions, facts and aggregate facts. The access layers are used for retrieving the data which is used by managers and business analysts for data mining, market research, online analytical processing and decision making.

4. The Dependent Data Mart Structure

The main idea behind the proposal put forward by Inmon was that we should be able to transfer data from diverse OLTP systems into a centralized place thereby enabling its use for analysis. This required the data to be organized into subject oriented, integrated, non-volatile and time variant structures. The data should be accessible at detailed atomic levels by drilling down or at summarized levels by drilling up. The data marts are the sub sets of the data warehouse. Each data mart relates to the data pertaining to an individual department. It is optimized for analysis requirements of the particular department for which it is created.



Data Warehouse & Data Marts

In the top down OLAP environment the flow of data begins with the extraction of data from operational data sources. This data undergoes validation and consolidation at the staging area so that a level of accuracy is ensured. This validated and consolidated data is then transferred to the Operational Data Store (ODS). This step is though optional and may be avoided by loading data into the data warehouse in a parallel process. The inclusion or exclusion of the operational data store depends upon the business requirements, *i.e.*, if there is a need for detailed data in the data warehouse then the Operational Data Source is considered.

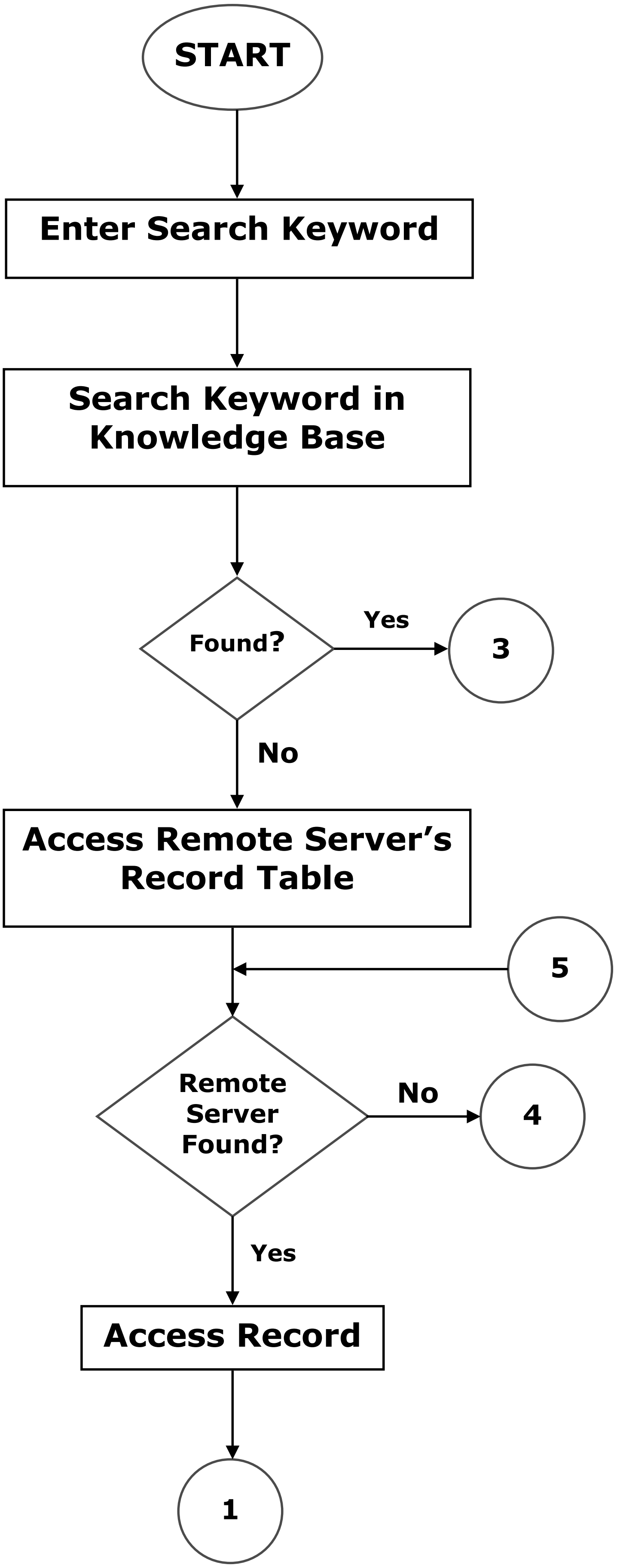
The data undergoes aggregation and summarization and is then extracted and loaded into the Data warehouse. As the Data warehouse aggregation and summarization processes are complete, the data is extracted from the Data warehouse and placed in the staging area to perform a new set of transformations on them. This ensures that the data is organized in particular structures required by the data marts. After this the data marts can be loaded with the data and the OLAP environment becomes available to the users.

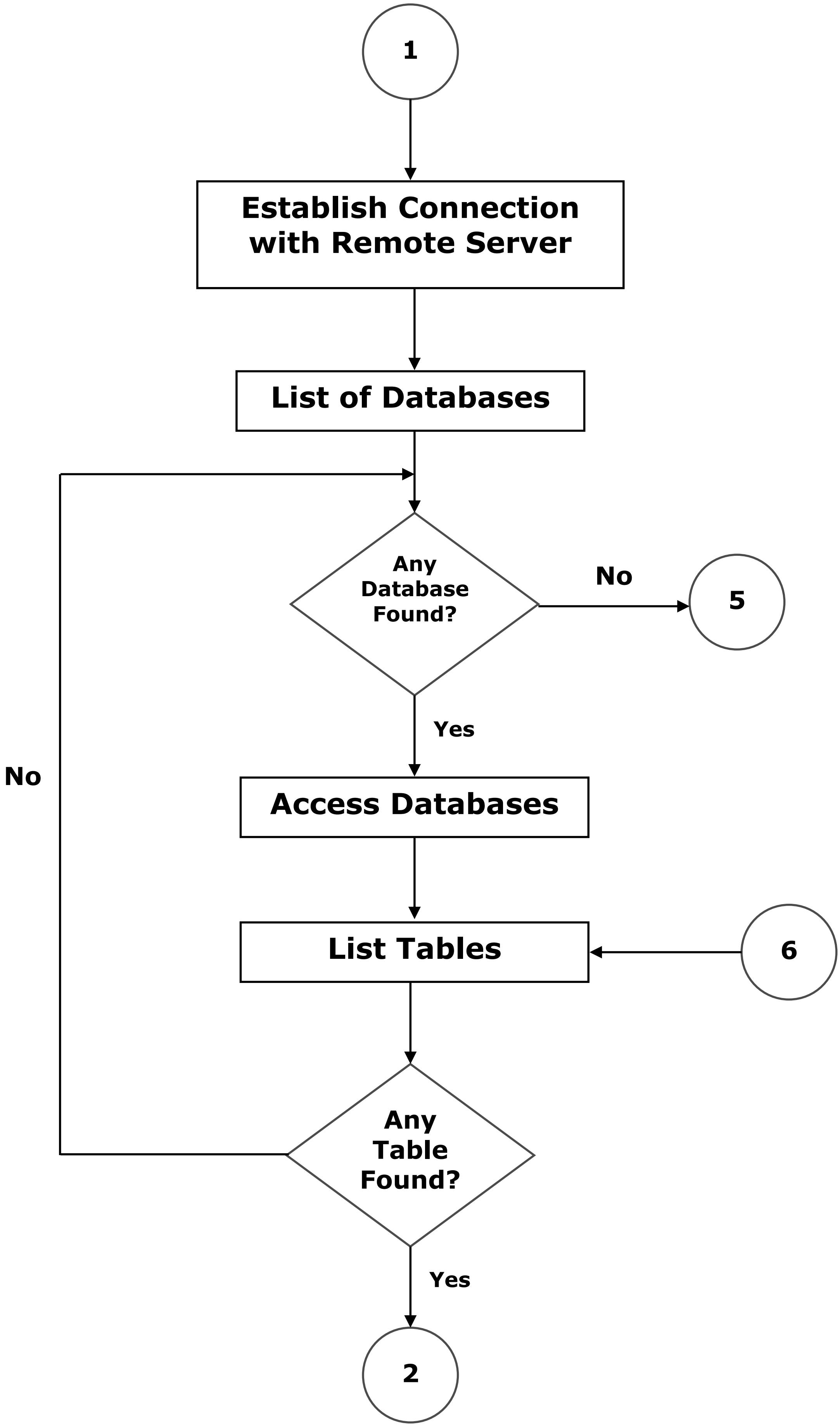
5. Logical Framework

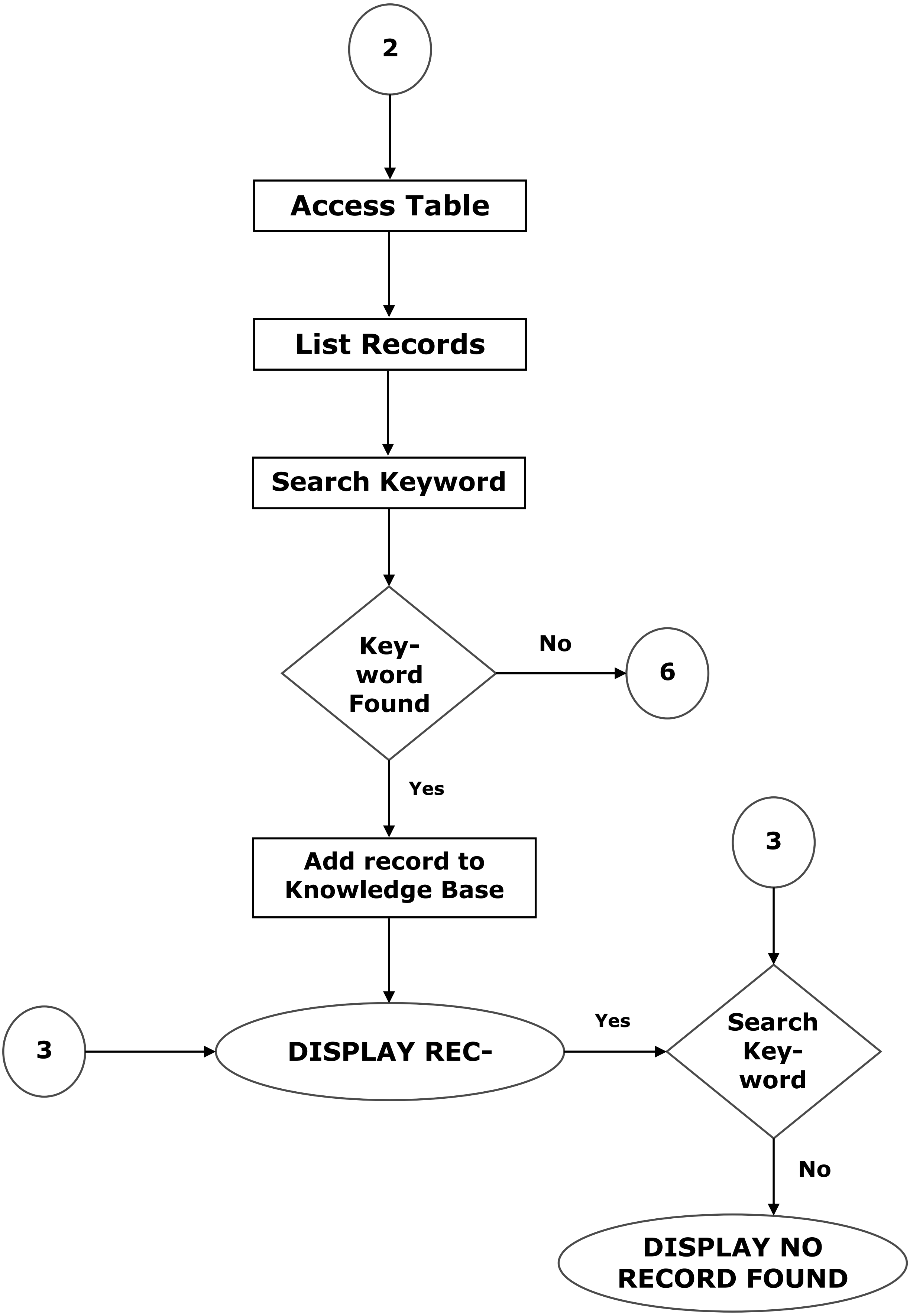
The proposed model is graphically shown below. This model assumes heterogeneity of data sources which are geographically distant however, accessible via IP address. The model allows administrator to identify and add information (IP address, type, port, *etc.*,) pertaining to heterogeneous data source.

Google like interface is presented to (enterprise) user who can search for any keyword by typing it in the search bar. User does not have to write a query to search for a keyword in different databases; rather he/she needs to write the keyword only. The Model generates query for every data source which in turn means for every Database Server identified by admin, for every database within Database Server wherein every table of database their shall be query generated and executed by the model and accordingly result is generated and displayed for the user.

In this user gets the list of all the tables, of all the databases, from all the heterogeneous data sources (servers) where the searched keyword has been found. Further user can click on any of the desired links which will redirect user to the page where the whole instance of that table where the keyword is found, gets displayed.


START
Enter Search Keyword
Search Keyword in Knowledge Base
Found?
Yes
3
No
Access Remote Server's Record Table
5
Remote Server Found?
No
4
Yes
Access Record
1


1
Establish Connection with Remote Server
List of Databases
Any Database Found?
No
5
Yes
No
Access Databases
List Tables
6
Any Table Found?
Yes
2


2
Access Table
List Records
Search Keyword
Key-word Found
No
6
Yes
Add record to Knowledge Base
3
3
DISPLAY REC-
Yes
Search Key-word
No
DISPLAY NO RECORD FOUND

6. Conclusion

In this paper Logical Data Integration Model is proposed for describing data, schemas and contexts for the integration of Heterogeneous Data Sources. This model overcomes heterogeneity at all levels including physical and semantic. User can access desired information without having to have any expertise in SQL. Further the user does not need to physically go over to the location where data is physically stored.

References

[1] http://www.datacenterjournal.com/understanding-costs-data-warehouses/.

[2] M. Zaman, S. M. Quadri and M. A. Butt, "Generic Search Optimization for Heterogeneous Data Sources", International Journal of Computer Applications, vol. 44, no. 5, **(2012)** April, pp. 14-7.

[3] M. Zaman, M. A. Butt and S. M. Quadri, "User Desired Information Translation", Journal of Global Research in Computer Science, vol. 3, no. 6, **(2012)** July 10, pp. 51-3.

[4] M. Zaman and M. A. Butt, "Enterprise Management Information System: Design & Architecture", International Journal of Computational Engineering Research (IJCER), ISSN. 2250:3005, May **(2013)**.

[5] M. Zaman, S. M. Quadri and E. M. Butt, "Information Integration for Heterogeneous Data Sources", IOSR Journal of Engineering, vol. 2, no. 4, **(2012)** April, pp. 640-3.

[6] M. A. Butt, S. M. Quadri and M. Zaman, "Data warehouse implementation of examination databases", International Journal of Computer Applications, vol. 44, no. 5, **(2012)** April, pp. 18-23.

[7] E. M. Zaman, S. M. Quadri and E. M. Butt, "Data Warehouse Implementation of Examination Databases", In Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS) 2012 Jan 1 (p. 1). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[8] M. A. Butt and M. Zaman, "Assessment Model based Data Warehouse: A Qualitative Approach", International Journal of Computer Applications, vol. 62, no. 10, **(2013)** January 1.

[9] M. A. Butt and M. Zaman, "Data quality tools for data warehousing: enterprise case study", IOSR Journal of Engineering, vol. 3, no. 1, **(2013)**, pp. 75-6.

[10] M. Ahmed Butt and M. Zaman, "Data Quality Tools for Data Warehousing: Enterprise Case Study", International Journal of Computer Applications, vol. 62, no. 10, **(2013)** January, pp. 22-4.

[11] A. V. Halevy, "Why Your Data Won't Mix", ACM Queue, vol. 3, no. 8, **(2005)** October.

[12] S. Biffl, W. Danar Sunindyo and T. Moser, "Semantic Integration of Heterogeneous Data Sources for Monitoring Frequent-Release Software Projects", International Conference on Complex, Intelligent and Software Intensive Systems, **(2010)**.

[13] R. Ashok Kumar and Y. Rama Devi, "Efficient Approaches for Record level Web Information Extraction Systems", Published in International Journal of Advanced Engineering & Application, **(2011)** January, pp. 161-164.

[14] L. Tu Tari, P. Hakenberg, J. Chen, Y. Son, T. Gonzalez and G. Baral, "Incremental Information Extraction Using Relational Databases", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, **(2012)** January.

[15] U. Masermann and G. Vossen, "Design and Implementation of a Novel Approach to keyword searching in Relational Databases", In proceedings ADBIS-DASFAA Symp. On 'Advances in Databases & Information Systems', Prague, Czech Republic, **(2000)** September 5-8.

[16] U. Masermann and G. Vossen, "SISQL: Schema Independent Database Querying (on and off the Web)", In proceedings of IDEAS 2000, Yokohoma, Japan, **(2000)** September 18-20.

[17] Md. S. Shahriar and J. Liu, "Constraint-Based Data Transformation for Integration: An Information System Approach", International Journal of Database Theory and Application, vol. 3, no. 1, **(2010)** March, pp. 85-92.

[18] J. C. Shafer and R. Agrawal, "Continuous Querying in Database Centric Web Applications", In Proceedings of 9th Intl. WWW Conference, Elsivier Science, **(2000)** May 15-19, pp. 519-531.

[19] R. Goldman, N. Shivakumar, S. Venkatasubramanian and H. Garcia-Molina, "Proximity Search in Databases", Proceedings of VLDB 98.

[20] M. Toyama and T. Nagafuji, "Dynamic and Structured Presentation of Database Contents on the Web", In proceedings EDBT'98 (Spain, 1998), LNCS # 1377, Springer-Verlag, pp. 451-465.

[21] L. Tu Tari, P. Hakenberg, J. Chen, Y. Son, T. Gonzalez and G. Baral, "Incremental Information Extraction Using Relational Databases", Knowledge and Data Engineering, IEEE Transactions on Issue:99, **(2010)** October 28, pp. 25-35.

[22] Md. S. Shahriar and J. Liu, "Constraint-Based Data Transformation for Integration: An Information System Approach", International Journal of Database Theory and Application, vol. 3, no. 1, **(2010)** March, pp. 85-92.

[23] L. Jinga, D. An-Ronga and Mao Qi-Zhia, "The Study of Integration of Multi-Sources Heterogeneous Data Based on The Ontology", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXVII. Part B2. Beijing, **(2008)**.