Information Translation: A Practitioners Approach

Majid Zaman, S. M. K. Quadri and Muheet Ahmed Butt

Abstract - 20th century resulted in accumulation of two things- wires and data, while both brought enormous success to organization in specific and information technology in general, 21st century is all about management. Industry realized need to get rid of wires and integrate/manage data present everywhere around us. Fiber & Wi-Fi is replacement to wires; however data integration/management is still challenge at large because of varying underlying structure, format, operating system etc. In this paper we propose/introduce various methods of data transformation at application level without having to modifying underlying structure of data storage.

***Index Terms*— Data, Information, Data Warehouse, ISL**

I. INTRODUCTION

With the advent of computerization primary goal of organizations across the globe was automation of working system, this resulted in massive collection of data irrespective of organization business logic and process, not much was thought about integration of application and data. Once a blessing became huge problem in organizations. Data sources spread across organization is difficult to manage and result in data inconsistency. Globally organizations spent lot of money for data management and to overcome inconsistency in data.

With the introduction of Data Warehouse which is used to integrate data from many heterogeneous data sources which includes the working and archive data pertaining to the organization. It also includes multiple subject areas and is typically implemented and controlled by a central organizational unit such as the corporate Information Technology (IT) group often called as central or enterprise data warehouse. Data Warehouse integrates data from heterogeneous/homogeneous data sources however data translation still remains challenge at large [1].

Manuscript received July 31, 2012; revised August 12, 2012

Er. Majid Zaman is working as Scientist in Directorate of Information Technology & Support Systems, University of Kashmir, Srinagar, J&K, India: zamanmajid@rediffmail.com

Dr. S. M. K. Quadri is working asHead & Director, PG Department of Computer Science, University of Kashmir, Srinagar, Srinagar, J&K, India : quadrismk@hotmail.com

Er. Muheet Ahmed Butt is working as Scientist in Directorate of Information Technology & Support Systems, University of Kashmir, Srinagar, J&K, India: ermuheet@gmail.com

On internet there is a huge data explosion going, according to Eric Schmidt, Google CEO "every two days now we create as much information as we did from the dawn of civilization up until 2003, something like five Exabyte of data" he says [2]. In 2011 300 million website were added making total number of websites to 555 million(December 2011)[3], thus resulting numerous data sources each having its own structure and schema, user desired data presentation still remains issue at large and needs to be understood and covered at the earliest.

II. DATA & INFORMATION

Data refers to the lowest abstract or a raw input which when processed or arranged makes meaningful information. It is the group or chunks which represent quantitative and qualitative attributes pertaining to variables. Information is usually the processed outcome of data. More specifically speaking, it is derived from data. Information is a concept and can be used in many domains.

Data can be in the form of numbers, characters, symbols, or even pictures. A collection of these data which conveys some meaningful idea is information. It may provide answers to questions like who, which, when, why, what, and how.

The raw input is data and it has no significance when it exists in that form. When data is collated or organized into something meaningful, it gains significance. This meaningful organization is information [4].

III. FILE FORMATS & DATABASE

Some file formats are designed for very particular types of data: PNG files, for example, store bit mapped images using loss less data compression. Other file formats, however, are designed for storage of several different types of data: the Ogg format can act as a container for many different types of multimedia, including any combination of audio and/or video, with or without text (such as subtitles), and metadata. A text file can contain any stream of characters, encoded in one of many kinds of character encoding schemes, including possible control characters. Some file formats, such as HTML, Scalable Vector Graphics, and the source code of computer software are also text files with defined syntaxes that allow them to be used for specific purposes[5][8].

On the other hand a database is a collection of data that is organized so that it can easily be accessed, managed, and updated. In computing, databases are sometimes classified according to their organizational approach. The most prevalent

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

approach is the relational database, a tabular database in which data is defined so that it can be reorganized and accessed in a number of different ways. A distributed database is one that can be dispersed or replicated among different points in a network. An object-oriented programming database is one that is congruent with the data defined in object classes and subclasses [6][10].

IV. PROBLEM DEFINITION

Most of the internet and intranet users are not well versed with technology. It has been observed that even top level managers are dependent on technical support of the organization for carrying out there day to day tasks.

Data in the organization may be present in database however user wants the same data as hardcopy, or as in most cases written text in the website is copied and pasted on Microsoft word. User wants part of the image but does not understand if it is possible to edit the picture or not.

The problem is that there is no single generic data format available, information is present in different formats requiring different tools for its mining.

In prevailing circumstances user is required to have comprehensive knowledge of system/database/file formats in order to use information the way he/she wants to. The target system needs to be built which hides the technology from users and provides him with the information in desired format.

V. SYSTEM ASSUMPTION & SOLUTION

Heterogeneous data spread across multiple sources having varying underlying structure and data format which are extracted, transformed and loaded into single Data Warehouse. We assume Warehouses/Marts depending on enterprise architecture are created as such data is centralized [7][9].

Solution is conversion of result in user desired format i.e. user query is executed on warehouses and generated result is converted into user desired format, (excel/odt/pdf etc.).User can also describe feature of his/her file format i.e he/she wants Verdana 12 as font size in word 2007 format.

VI. PROPOSED ALGORITHM

- ISL- INTELLIGENT SOFTWARE LAYER is placed between user and warehouses, user input is received by and converted into query by ISL and same is executed on warehouse, auxiliary information is saved for later use e.g. font type size format, thus user input is received by ISL.
- User is provided with GUI so that he/she can input his/her query(Google sought) along with desired format in which user wants his/her result e.g.(.Microsoft word. determine extension of the said format[11][12][13].
- Result generated as a result of execution of query is not passed on to user but is received by ISL for converting it into user desired format.
- ISL receives result from warehouse, creates new text file and saves result in this newly created text file(.txt), file name is based on time stamping principle e.g. 1545220412.txt where 15 is hours, 45 is minutes, 22 is day 04 is month and 12 is year.
- ISL converts it into user desired file format, translation requires
  a. User desired format is already know.
  b. Create new file, with the same name as that of text file but with user desired extension i.e. if use wants output in word format then 1545220412.docx is created.
  c. User requirement such as font, size, margins etc. is taken care of at the time of file creation, e.g. word file is created in which font size, and margins etc. become integral part of this newly created file.
  d. Data saved in text files is read char by char and appended into the newly created file.
- File created is passed on to the user, and both text file and application file are deleted, as such disk storage is not an issue.
- ISL does not need to buy application license such as Microsoft office, pdf, etc.
- ISL can be initially tested for few formats e.g. word, pdf etc. before support for all formats can be extended.

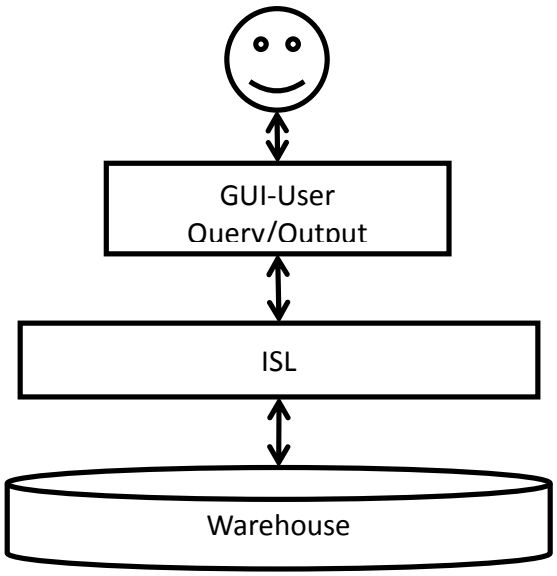


Fig: 1: Diagrammatic Representation of Algorithm

VII. CONCLUSION

User over the years has become more demanding; he/she does not only need information but wants it in specific format which should be complete and correct. Globally centralization was prioritized because of collection of massive data in heterogeneous data sources, however much was not thought for naive user and information system was still at the mercy of technocrats time has now come to stress more upon user demands so as to meet user demands and make user dependency on technocrats minimal.

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

REFERENCES

[1] http://docs.oracle.com/html/E10312_01/dm_concepts.htm. 13 May, 2012, Oracle's official documentation library for all of its products and of all versions, contains links to the most current documentation for products, including the former Sun products.

[2] http://techcrunch.com/2010/08/04/schmidt-data. 15 May, 2012; TechCrunch is a web publication that offers technology news and analysis, as well as profiles of startup companies, products, and websites.

[3] http://royal.pingdom.com/2012/01/17/internet-2011-in-numbers. 12May,2012, Pingdom makes it easy to monitor the uptime and response time of websites and servers on the Internet

[4] http://www.differencebetween.net/language/difference-between-data-and-information.8May,2012, Difference Between is an online knowledge base that analyses the differences between anything and everything.

[5] http://en.wikipedia.org/wiki/File_format.25 May, 2012, Wikipedia is largely open encyclopedia.

[6] http://searchsqlserver.techtarget.com/definition/database. 5 May,2012, This gives information on Microsoft SQL Server 2012, from pricing and licensing to features around high availability, business intelligence and cloud computing, besides blogs.

[7] Mohammad Ghulam Ali, “Object Oriented Approach for integration of heterogeneous databases in a multidatabase system and local schemas modifications propagation”, international journal of computer sciences and information security, vol 6, No. 2, 2009.

[8] Md. Sumon Shahriar and Jixue Liu, “Constraint-Based Data Transformation for Integration: An Information System Approach”, International Journal of Database Theory and Application Vol. 3, No. 1,pp 85-92, March, 2010.

[9] Stefan Biffl, Wikan Danar Sunindyo, Thomas Moser, “Semantic Integration of Heterogeneous Data Sources for Monitoring Frequent-Release Software Projects”. International Conference on Complex, Intelligent and Software Intensive Systems, 2010.

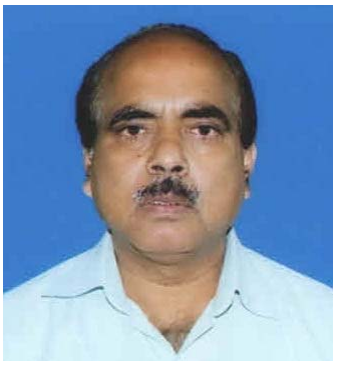
[10] Marc Van Cappellen, Wouter Cordewiner, Carlo Innocenti, “Data Aggregation, Heterogeneous Data Sources and Streaming Processing: How Can XQuery Help? Bulletin of the IEEE Computer Society, Technical Committee on Data Engineering, 2008.

[11] S. Agarwal, S. Chaudhary, and G. Das. 'Dbxplorer, “A system for keyword based search over Relational Databases”. In proceedings of ICDE 2002

[12] F. Song, W.B. Croft, “A general language model for information retrieval”. In proceeding of SIGIR 1999.

[13] Bhalotia, A. Hulgeri, C. Nakhe, S. Chakarbarti and S. Sudarshan, “BANKS: Browsing and keywords searching in Relational databases”. In proceedings of ICDE 2002

Er. Majid Zaman
is working as Scientist in Directorate of Information Technology & Support Systems, University of Kashmir, Srinagar, J&K, India: zamanmajid@rediffmail.com



Dr. S. M. K. Quadri
is working as Head & Director, PG Department of Computer Science, University of Kashmir, Srinagar, Srinagar, J&K, India : quadrismk@hotmail.com

Er. Muheet Ahmed Butt
is working as Scientist in Directorate of Information Technology & Support Systems, University of Kashmir, Srinagar, J&K, India: ermuheet@gmail.com

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)