Volume 3, No. 6, June 2012

Journal of Global Research in Computer Science

ISSN-2229-371X

TECHNICAL NOTE

Available Online at www.jgrcs.info

USER DESIRED INFORMATION TRANSLATION

Majid Zaman [*1], Muheet Ahmed Butt [2] and Dr. SMK Quadri [3]

zamanmajid@rediffmail.com[1]

ermuheet@gmail.com[2]

quadrismk@hotmail.com[3]

Abstract:With the advent of computerization primary goal of organization across the globe was automation of their system, this result in massive collection of data in respective 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 all over the organization was becoming difficult to manage and inconsistency of data resulted in creation of team not meant for development but data management.

Many organizations have started reinvesting in data management in the form of creation of Data Warehouse and again organisation across the globe are not stressing upon user needs and demands but only focusing on integration of heterogeneous data sources with goal of making data centralised and consistent by creating Warehouse.

21st century user has needs, he/she not only needs data but needs refined and cleaned data, he/she wants data in his/her desired format. While data integration is paramount need of the hour is to make data stored in such flexible manner so that user can be provided information in his/her desired format. In this paper we introduce methods of data transformation at application level without need to modify underlying structure.

INTRODUCTION

20th century resulted in accumulation of two things-wires and data. While both brought enormous success to organisation in specific and information Technology in general, 21st century was all about management. Industry realised we need to get rid of wires and integrate as well as manage data present every were around us, getting rid of wires seems to be easy(fibre & wifi) how ever data integration and management is still a challenge at large because of varying underlying structure, format, operating system etc.

With the introduction of Data Warehouse- A data warehouse, deals with multiple subject areas and is typically implemented and controlled by a central organizational unit such as the corporate Information Technology (IT) group. Often, it is called a central or enterprise data warehouse. Typically, a data warehouse assembles data from multiple source systems [1]. Data Warehouse integrates data from heterogeneous/homogeneous data sources however data translation is still challenge at large.

On internet there is data explosion, 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’s 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 understood and covered at the earliest.

DATA & INFORMATION

Data refers to the lowest abstract or a raw input which when processed or arranged makes meaningful output. 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].

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].

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 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].

PROBLEM

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.

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

The problem is that their is not one generic data format, information is present in different formats requiring different tools to use such information. The problems are not only with data formats but with different operating systems, were in users find it extremely difficult to manage data and use information in desired format.

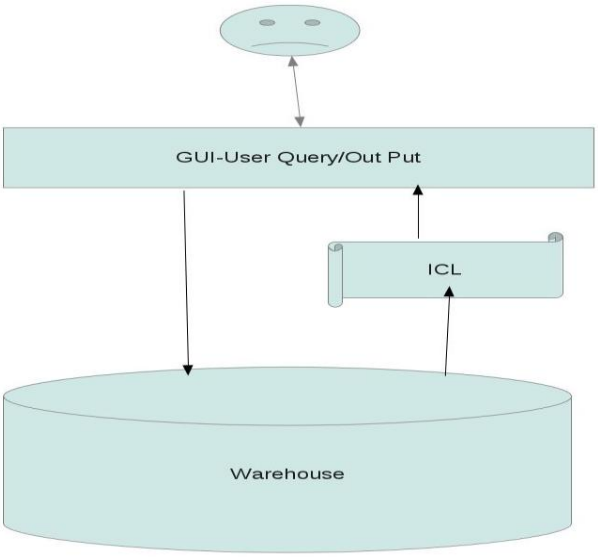
In prevailing circumstances user is required to have system knowledge of system/database/file formats in order to use information the way he/she wants to, were in system needs to be built which hides technology from users and provides him with information in desired format.

PROPOSED SYSTEM

a. The solution is twofold, one is centralisation of data-Data Warehouses/Mart is standard defacto for centralisation of enterprise data. Heterogenous data spread across multiple sources having varying underlying structure and data format are extracted, transformed and loaded into single Data Warehouse.We assume Warehouses/Marts depending on enterprise architecture are created as such data is centralized.
b. Second part of the soultion 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,

Algorithm:

a. Warehouse/Mart is already created.
b. 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,Excel/odt/pdf). User can also describe feature of his/her file format i.e he/she wants Vardana 12 as font size.
c. User input is converted into query and same is executed on warehouse.
d. Result generated as a result of execution of query is not passed on to user but passed onto ISL-Intelligent Software Layer
e. ISL is placed between user and warehouses, however it comes into work only when warehouses has generated its output.
f. ISL recieves result from warehouse, creates new text file and saves the same in newly created text file(.txt), file name is based on time stamping princple e.g 1545220412.txt where 15 is hours, 45 is mins, 22 is day 04 is month and 12 is year.
g. ISL converts it into user desired file format, translation requires
   a) determine user desired format
   b) determine extention of the said format
   c) create new file, with the same name as that of text file but with user desired extention i.e if use wants output in word format then 1545220412.docx is created.
   d) file creation is done in such a manner that user requirement such such as font, size etc is saved at the time of creation of file, i.e such information is made part of file as is done by all file formats this includes .jpg, pdf etc
   e) data saved in text files is read char by char and saved into the newly created file.
   f) File created is passed on to the user, and both text file and application file are deleted
   g) ISL does not need to buy application lisence such as microsoft office, pdf, etc but only need to file format.



CONCLUSION

User over the years has become more demanding, he/she does not only need information but wants it in specific format. Globally centralization was prioritized because of collection of massive data in heterogeneous data sources, how ever 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 use demands and make user dependency on technocrats minimal.

REFERENCES

[1]. http://docs.oracle.com/html/E10312_01/dm_concepts.htm 13 May,2012

[2]. http://techcrunch.com/2010/08/04/schmidt-data 15 May, 2012

[3]. http://royal.pingdom.com/2012/01/17/internet-2011-in numbers/ 12 May,2012

[4]. http://www.differencebetween.net/language/difference-between-data-and-information/

[5]. http://en.wikipedia.org/wiki/File_format

[6]. http://searchsqlserver.techtarget.com/definition/database

[7]. R. Ashok Kumar, Dr Y. Rama Devi, "Efficient Approaches for Record level Web Information Extraction Systems". Published in International Journal of Advanced Engineering & Application, pp 161-164, Jan 2011

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

[9]. Ramakrishna Srikant, Sugato Basu, Ni Wang, Daryl Pregibon, "User browsing models: relevance versus examination". In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 223-232, 2010.

[10]. 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.

[11]. E. Alfonseca, K. Hall, and S. Hartmann, "Large-scale computation of distributional similarities for queries". In Proceedings of NAACL-HLT, Association for Computational Linguistics, pp 29-32, 2009.

[12]. Bo Yang and Manohar Mareboyana, "Progressive Content-Sensitive Data Retrieval in Sensor Networks". Journal of Computer Science 5 (7):pp 529-535, 2009.

[13]. 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.

[14]. 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.

[15]. Alon Halevy, "Information Integration". In Encyclopedia of Database Systems, 2009.

[16]. Peter Pach, Attila Gyenesei, and Janos Abonyi, "Compact fuzzy association rule based classifier".