International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Interrelated Document Warehouse Report Visualization

V. Vivek[1], N. J. Subashini[2]

[1]M. Tech Cloud Computing, Department of Information Technology, SRM University, SRM Nagar, Kattankulathur, Tamil Nadu, India, 603203

[2]Assistant Professor, Department of Information Technology, SRM University, SRM Nagar, Kattankulathur, Tamil Nadu, India, 603203

Abstract: ***Data warehouse or an enterprise data warehouse , is the kind of data storage used for reporting and analyzing data's from one or more different sources and creates a central common repository of data.In the existing system, a report will be generated for the general facts and records of user's information based on the data centric mining operation where data's have been fetched from the repository of data warehouse storage area. In our proposed system , we integrate the data from multiple sources like xml, excel , csv files, etc by enabling the central view is the core projection of our project. Interrelating the various different formats of files like pdf, excel, csv files will be stored in the common data warehoused repository storage where it will be mined under data mined knowledge cubes.***

Keywords: Information visualization, OLAP report, user interface, data and knowledge visualization.

1. Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data processing software package is one in every of variety of analytical tools for analyzing knowledge. It permits users to research knowledge from many various dimensions or angles, categorise it, and summarize the relationships known.

Technically, data mining is that the method of finding correlations or patterns among dozens of fields in massive relative databases. We integrate the information from multiple sources like xml, excel, csv files, etc by enabling the central read is that the core projection of our project. Interrelating the assorted completely different formats of files like pdf ,excel ,csv files are keep within the common knowledge warehoused repository storage wherever it'll be mined below knowledge mined data cubes.

A report are going to be generated for the general facts and records of user's data supported the info central mining operation wherever data's are fetched from the repository of information warehouse cargo area. In our planned system we have a tendency to integrate the info from multiple sources like xml, excel , csv files, etc by sanctioning the central read is that the core projection of our project. Interrelating the assorted totally different formats of files like pdf ,excel ,csv files are going to be hold on within the common knowledge warehoused repository storage wherever it'll be strip-mined below knowledge strip-mined data cubes.

Hierarchy on dimension attributes of reticular information has mount projected in our approach wherever it lets dimensions to be viewed at totally different levels of detail that forms a information cube in a very three-d generalization of a cross-tab memory to store information cubes.

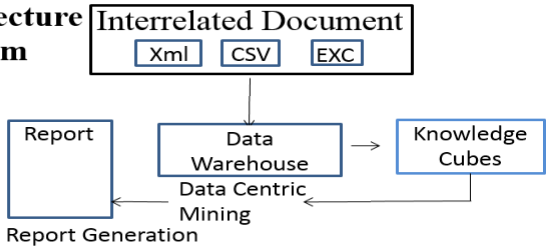
2. Implementation

2.1 Proposed Architecture

Interactive analysis of data has been proposed, by allowing data to be summarized and viewed in different ways (Pdf,CSV,Excel,etc) with negligible delay. Data that can be modeled as dimension attributes and measure attributes are called multidimensional data where it stores the entire data's in the form of inter-related storage in the data warehoused repository.

The interrelated document will be stored in data warehouse now *Mediator based data integration* algorithm has been implemented for generalization & suppression of data's in the data warehoused paradigm. Depending upon user request knowledge Cube will be generated,Mining process will be carryout to obtain specific result from that finally report will be generated.

Architecture Diagram



Licensed Under Creative Commons Attribution CC BY

2.2 Multi Attributable Documents Insertion - Different Columns And Different Data Structure:

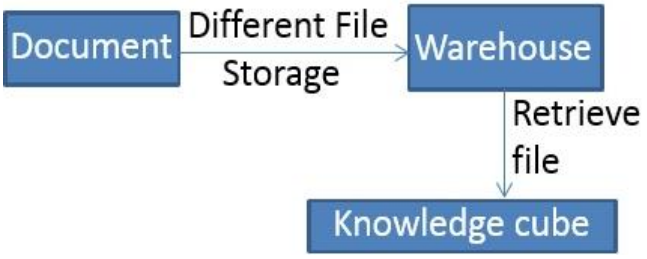
This column-oriented has advantages for data warehouse, used in customer relationship management (CRM) systems. Column-oriented organizations are more efficient when new values of a column are supplied for all rows at once, because that column data can be written efficiently and replace old column data without touching any other columns for the rows. Row-oriented organizations are more efficient when many columns of a single row are required at the same time, and when row-size is relatively small, as the entire row can be retrieved with a single disk seek.

2.3 Warehouse Creation

The data warehouse database schema is often quite simple compared to those of OLTP databases or the data preparation area. A star schema consists of a single fact table and a number of dimension tables. A snowflake schema adds secondary dimension tables.

Designing a data warehouse is very different from designing an online transaction processing (OLTP) system. In contrast to an OLTP system in which the purpose is to capture high rates of data changes and additions, the purpose of a data warehouse is to organize large amounts of stable data for ease of analysis and retrieval. Because of these differing purposes, there are many considerations in data warehouse design that differ from OLTP database design.

Data warehouse data must be organized to meet the purpose of the data warehouse, which is rapid access to information for analysis and reporting. Dimensional modeling is used in the design of data warehouse databases to organize the data for efficiency of queries that are intended to analyze and summarize large volumes of data. The data warehouse schema is almost always very different and much simpler than the schema of an OLTP system designed using entity-relation modeling.

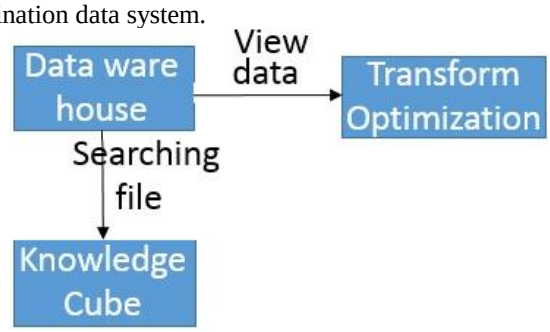


2.4 Data cleansing and transformation optimization

As we accomplish many data transformations during the process of extracting data from the source systems. However, there are often additional tasks to complete before we can load data into the data warehouse. For example, we must reconcile inconsistent data from heterogeneous data sources after extraction and complete other formatting and cleansing tasks. we should also wait until after the extraction process to incorporate surrogate keys. Some transformations that we might technically accomplish during the extraction process may interfere with the performance or operation of the online source system; we should defer these tasks until after extraction is complete.

After extraction from the source systems, the data should reside in a data preparation area where the cleansing and transformations can be completed before the data is loaded into the data warehouse. The data preparation area can be a separate database or separate tables in the data warehouse database. During the cleansing and transformation phase, we can execute procedures to validate and verify data consistency, transform data into common formats, and incorporate surrogate keys

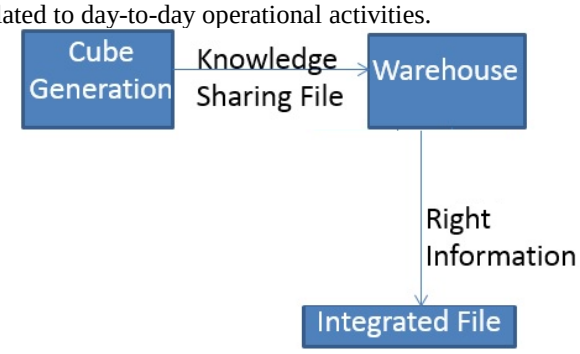
Data cleansing also called as data scrubbing. In this module the process of amending or removing data is carried out in a database that is incorrect, incomplete, improperly formatted or duplicated. In metadata and data warehouse, data transformation module converts a set of data values from the data format of a source data system into the data format of a destination data system.



2.5 Cube Generation

In this section we provide a conception of knowledge cube, and of knowledge types, kinds and qualities and convert the data into knowledge cubes. Informal knowledge sharing file in storage and retrieving. Knowledge management is fundamentally active. As our base notion knowledge is understood as justified true belief, which is bound to the human being, with a dimension of purpose and intent, identifiying patterns in its validity scope, brought to bear in action and with a generative capability of new information.

It is a perspective of "knowledge-in-use" because of the importance for its utilisation in companies and for knowledge management. In contrast, information is understood as data in relation with a semantic dimension, but is lacking the pragmatic and pattern-oriented dimension, which characterises knowledge. Knowledge management needs to be proactive, tightly integrated file and integrally related to day-to-day operational activities.



Licensed Under Creative Commons Attribution CC BY

2.5 Multi Dimensional Data Construction

The multidimensional data construction is an integral part of On-Line Analytical Processing, or OLAP. Because OLAP is on-line, it must provide answers quickly. The multidimensional data model is designed to solve complex queries in real time.

"The central attraction of the dimensional model of a business is its simplicity and it is the fundamental key for the users to understand the databases, and allows software to navigate databases efficiently."

The multidimensional data construction is composed of logical cubes, measures, dimensions, hierarchies, levels, and attributes. The simplicity of the model is inherent because it defines objects that represent real-world business entities. Analysts know which business measures they are interested in examining, which dimensions and attributes make the data meaningful, and how the dimensions of their business are organized into levels and hierarchies.
In this module it will Show the relationships between multiple variables and then finds the clusters of similar data types. Data of information has been consolidated in the common repository of storage.

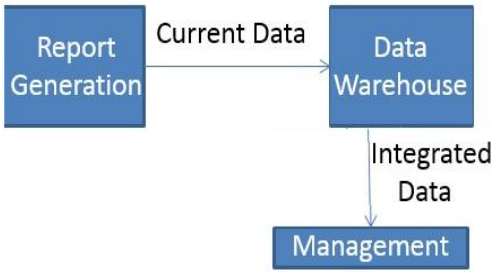
2.6 Report Generation and Visualization

A **report generator** is a computer program whose purpose is to take data from a source such as a database, XML stream or a spreadsheet, and use it to produce a document in a format which satisfies a particular human readership.

Report generation functionality is almost always present in database systems, where the source of the data is the database itself. It can also be argued that report generation is part of the purpose of a Spreadsheet. Standalone report generators may work with multiple data sources and export reports to different document formats.

Information Systems theory specifies that information delivered to a target human reader must be Timely, Accurate and Relevant. Report generation targets the final requirement by making sure that the information delivered is presented in the way most readily understood by the target reader.

In this module Data warehouses store current data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons. Integrate data from multiple source systems, enabling a central view across the enterprise. This benefit is always valuable, but particularly so when the organization has grown by merger.



3. Conclusion

In this project we have to summarize about big data mining process in the currency conversion. And also show how the data will be flow of data mining process and implemented Meditator based data integration algorithm, knowledge cube for generalization of data in the data warehouse paradigm.

4. Future Work

Oracle R Enterprise, a component of the Oracle Advanced Analytics Option, makes the open source R statistical programming language and environment ready for the enterprise and big data. Designed for problems involving large amounts of data, Oracle R Enterprise integrates R with Oracle Database. R users can run R commands and scripts for statistical and graphical analyses on data stored in the Oracle Database. R users can develop, refine, and deploy R scripts that leverage the parallelism and scalability of the database to automate data analysis. Data analysts and data scientists can run R packages and develop and operationalize R scripts for analytical applications in one step—without having to learn SQL. Oracle R Enterprise performs function pushdown for in-database execution of base R and popular R packages. Because it runs as an embedded component of Oracle Database, Oracle R Enterprise can run any R package either by function pushdown or via embedded R while the database manages the data served to the R engines.

Reference

[1] P. Ciuccarelli, M.I. Sessa, and M. Tucci, “Code: A Graphic Language for Complex System Visualization,” Proc. Italian Assoc. for Information Systems (ItAIS), 2010.
[2] J. Bertin, Semiology of Graphics: Diagrams, Networks, Maps/Jacques Bertin; Translated by William J. Berg., Univ. of Wisconsin Press, 1983.
[3] S.K. Card, J.D. Mackinlay, and B. Shneiderman, Readings in Information Visualization: Using Vision to Think. Morgan Kaufman Publishers Inc., 1999.
[4] E.R. Tufte, The Visual Display of Quantitative Information, second ed. Graphics Press, May 2001.
[5] A. Unger, P. Muigg, H. Doleisch, and H. Schumann, “Visualizing Statistical Properties of Smoothly Brushed Data Subsets,” Proc.Fourth Int’l Conf. Information Visualisation, pp. 233-239, 2008.
[6] J. Heer and G. Robertson, “Animated Transitions in Statistical Data Graphics,” IEEE Trans. Visualization and Computer Graphics,vol. 13, no. 6, pp. 1240-1247, Nov. 2007.
[7] E. Papageorgiou, “Review Study on Fuzzy Cognitive Maps and Their Applications during the Last Decade,” Proc. IEEE Int’l Conf. Fuzzy Systems (FUZZ), pp. 828-835, 2011.
[8] D. Iakovidis and E. Papageorgiou, “Intuitionistic Fuzzy Cognitive Maps for Medical Decision Making,” IEEE Trans. Information Technology in Biomedicine, vol. 15, no. 1, pp. 100-107, Jan. 2011.
[9] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, third ed. Prentice Hall, 2009.

Licensed Under Creative Commons Attribution CC BY

[10] M. Risi, M.I. Sessa, G. Tortora, and M. Tucci, "Visualizing Information in Data Warehouses Reports," Proc. 19th Italian Symp. Advanced Database Systems (SEBD), pp. 246-257, 2011.

Licensed Under Creative Commons Attribution CC BY