ISSN:0975-9646

Shabia Shabir Khan et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (5) , 2012,5164 - 5168

Expedition for the Exploration of Apposite Knowledge

Shabia Shabir Khan[#], Dr. Mushtaq Ahmed Peer[*]

[#] *Department of computer science,*
University of Kashmir, Srinagar, India
[*]*Chairman J&K BOPEE, India*

***Abstract*— Once we have the organized collection of data, typically in digital form, we try to move from operational aspect towards the analytical one wherein the users who are the actual Knowledge workers and take the strategic decisions are of more importance, in contrast to the users who work on the operational aspect of a big system and take the tactical decisions. Now, if the application domain is changed from operational data management to data management for strategic purposes, we need to change the data base design as well, keeping the dynamic nature into consideration. So, Information providing service is the first point of concern for knowledge extraction. The ever increasing interest in the knowledge discovery made us to move the data from the statistical aspect towards designing creative algorithms in an attempt to find out the possible efficient solution so as to make strategic decisions as proper as possible. The aim of the paper is to present a Data Warehouse architecture that focuses on recent updation in the OLTP systems, in addition to the basic ELT/ETL process. Further, the paper discusses a wide range of knowledge discovery techniques used to extract fresh knowledge out of the data present in such a Real-Time Data Warehouse. The future research direction in the field has also been mentioned.**

***Index Terms*— Data Warehouse, Relational database, Statistics, Knowledge based systems, Artificial Intelligence, Fuzzy logic, Data Mining**

I. INTRODUCTION

The problem of lack of certainty or uncertainty, wherein we have the limited knowledge and is impossible to exactly describe the existing state or predict future events, can be resolved by the introduction of the information. One of the best Information providing services is Data Warehousing that involves the retrieval and consolidation of data periodically from the OLTP source systems into a dimensional or normalized data store using the technique of ETL (Extract, Transform, Load) or ELT(Extract, Load, Transform). Traditionally, the so-called Data Warehouse stores the historical data and is mostly updated periodically (typically monthly, weekly or daily) in batches, not every time a transaction happens in source system.

So, this type of Data Warehouse seems to be sort of static in nature and is thus known by the name "Static Data Warehouse" in which the most recent operational records are not included. However, for the applications like stock brokering, online telecommunications etc, where new and most recent data is quite important to analyze and react in a near real-time manner, the static Data Warehouse is not enough [2, 3, 5, 7].

Here, what we really require is a warehouse that is dynamic in nature or in other words, an Active / Dynamic Data Warehouse for which we need an appropriate data flow architecture wherein the fresh data can be loaded into the warehouse each time a transaction occurs in the OLTP source systems.

II. ARCHITECTURE FOR DYNAMIC DATA WAREHOUSE

Figure1 below represents one such attempt to provide the historical as well as the current data through Data Warehouse.

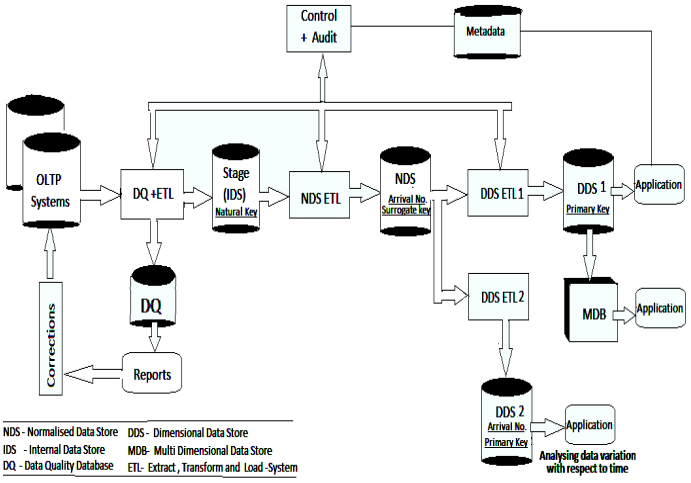

Figure 1: Configuration of Data Stores within an Active Data Warehouse System

A. *Description*

In figure 1, Configuration of data stores shows the arrangement of how the data flows from the OLTP source systems towards the analytical applications (OLAP, Data Mining, Reporting, Data Querying etc). It includes the ETL System that integrates (combining the same record from several different source systems into one record), transforms (converting, calculating, or modifying the data to suit the target database) and then loads the data into a data store (internal /external) that stores the integrated data into a single format (different from the varying formats used in OLTP systems) and thus makes it more suitable for analysis. In

addition, the architecture has a 'Control system' that automates the arrangement, coordination and management of the ETL or ELT system, an 'Audit system' that monitors and logs the operational statistics of the ETL processes and 'Data Quality Rules' through which data in the data stores is passed in order to ensure Data Quality wherein the inconvenient data is put into the Data Quality (DQ) database to be reported and then corrected in the source systems, automatically or manually.

A data store which is an important component of such a configuration can be one or more databases or files containing Data Warehouse data wherein the data is arranged in a particular format depending on the application requirements. The architecture includes a stage which is an internal data store (i.e. not accessible by the end user or the end-user applications) for transforming and preparing the data obtained from the source systems, before the data is loaded into a normalized data store (NDS). A normalized data store also is an internal data store in the form of one or more normalized relational databases and is a better format to integrate data from various source systems, especially in third normal form and higher. Apart from this purpose, NDS has also the ability to load data into several DDSs. Taking these two benefits into consideration, this kind of data store is appropriate one when it comes to continuous online transactions in the source systems that need to be reflected in the active or dynamic Data Warehouse because it uses the snow flake schema (normalized schema) wherein there is only one place to update without data redundancy. NDS ETL is the only application that is able to update NDS. We can avoid the data to be staged to disk first, rather, directly load the entities to NDS wherein the ETL Server is used to perform the data Integration and transformation online [2, 21].

In addition to this we have an operational data store (ODS) which is an internal as well as user facing data store in the form of one or more normalized relational databases.ODS contains only the most recent or current version of master data, that is useful for transactional queries. As far as updation in the warehouse data is concerned, the request for the updated version of the data is sent back by the ODS towards the source systems. In that case, the details in the fact table, such as 'quantity', 'delivery report 'or 'status', can be overwritten, however the original dimensions are not changed. In addition to this, ODS does not make use of any Surrogate Key (identifier of each record). Thus, we cannot use ODS when the historical versions of the data are also of concern to us [2, 3, 11].

B. Keys supporting the architecture:

Normalized data store (NDS) is the master data store that contains the complete data sets, including all historical transaction data and all historical versions of master data. There are two types of tables in the NDS: transaction table and master table, similar to OLTP source systems. The NDS transaction table and master table act as the source of data for the DDS fact tables and dimension tables respectively.

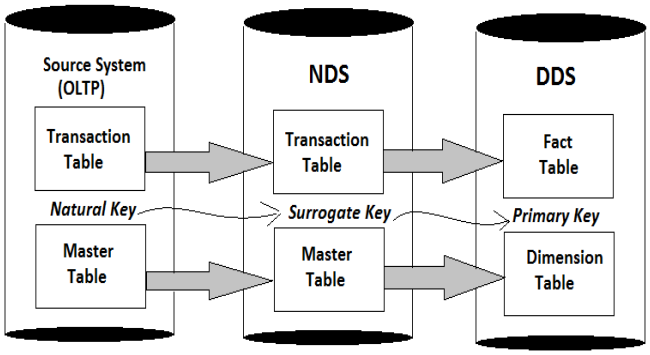

Figure 2: Translation of keys as we move from source to warehouse (NDS acting as source of data for DDS)

Figure 2 shows the tables in the DDS populated from the tables in the NDS wherein the data is loaded into DDS in dimensional format to make it accessible to users for querying etc.

As the data is loaded from stage to NDS, the natural Key (identifier of the master data row in source system) is translated to the Data Warehouse surrogate key. The NDS tables use this surrogate key as an identifier of the master data row within the Data Warehouse. In the DDS, the surrogate key is used as the primary Key of dimension tables. The Surrogate key is used to link a fact table and the dimension tables together especially in snow flake schema.

C. Advantage behind the architecture:

The main advantage of the architecture in figure 1 is that it provides good flexibility creating new or smaller DDSs that are populated from the NDS depending on the user's analytical requirements. NDS is the master data store, containing a complete set of data including all the historical and current versions of data and DDS may not contain the whole set of data as DDS ETL is parameterized. The second advantage is that it is easier to maintain master data in NDS and publish it from there because it contains little or no data redundancy and data updates are performed more easily and quickly compared to the dimensional master data store as we need to update only one place within the data store.

D. Mechanism behind the architecture:

The main important point of concern is the continuous updation in the source data that is to be reflected in the Active Data Warehouse. The solution starts with a data store that is normalized and uses a surrogate key i.e. NDS. As far as updation in NDS is concerned, if there is a change in master data, the attributes are not overwritten by new values but the new values are inserted as a new record keeping the old version (the old row) also in the same table and under the same surrogate key. In addition to this we use a Time Sequence Generator that generates numeric counter value starting from '1'. Such "Arrival Number" or "Version Number" is recorded under another field in NDS in an attempt to maintain the latest version of the record. So, depending on the arrival time, each arriving record is assigned

its own 'Arrival No' or 'Version No'. Here, if the source record is new then the Arrival No. for that record is 1. And if the record is updated version of any previous arrived record and thus having the same surrogate key then the updated version will be saved as a new record in the same NDS but under the different Arrival No. that is calculated by incrementing the Arrival No. of its previous version by one. So, the Arrival No. will be incremented each time a new updated form or version arrives. Finally, the question is whether to send all the arrivals or versions of each record forward towards the data store (that analyses the data variations on the basis of time) or simply forward the most recent version of the data to the dimensional data store. Since NDS can forward the data towards several DDSs, so we have used two - 'DDS ETL 1' and 'DDS ETL 2' systems in the architecture (Figure 1). 'DDS ETL 1' will extract the latest version of the record based on the Surrogate key and largest Arrival No of each record. Thus, the most recent data will be available to us in an Active Data Warehouse through 'DDS1'. And 'DDS ETL 2' will extract all the versions of the records in the NDS and loads them in DDS2 that can be used for analyzing the variations in each transaction for future prediction etc.

E. Example:

The figure3 below presents an example to show the necessity of the architecture (figure1) wherein the data is usually updated every hour and hence the Data Warehouse also needs an hour-based updation.

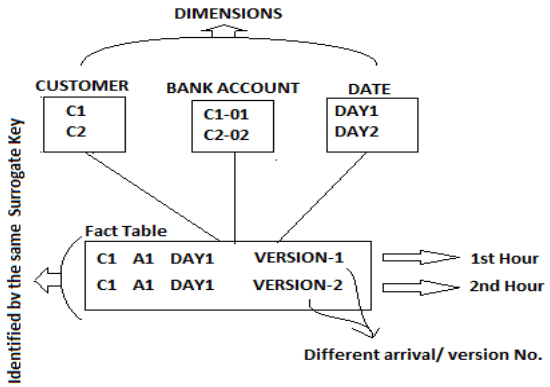

Figure 3: Analysing the frequent changes in the account balance at different hours in a particular day

Such an Active Data Warehousing technique represents an ideal vision of maintaining a fresh central repository of data in an attempt to provide timely competitive information to business users for proper strategic decision making.

III. TOWARDS THE ANALYSIS PHASE

(Exploration of the Active Data Warehouse architecture)

Once architecture has been provided, we get the most recent updated data. However the major concern is the utilization of such fresh data in order to extract the real time knowledge. Implicitly, we assume that the users of the database are clerical whose main focus is the addition and retrieval of the data elements from the database as efficiently as possible so that the overall consistency of the database is kept intact. Such kind of database users is involved in the operational aspect of the larger system. However, there is another kind of users who take strategic decisions for analytical issues in contrast to the tactical decisions that are taken for operational issues. Strategic decisions are taken in order to answer some dynamic questions (of analytical nature) e.g. "Which is the best location to open a counter to sell the products?", "Is it financially viable to continue our manufacturing unit in Kashmir?" A simple SQL query is really not enough to answer such questions. For answering this question, we first need to go through a central repository having strategic importance i.e. Data Warehouse consisting of the whole set of data elements captured from different operational data sources.

Next, we need a design paradigm that would represent a way to seek information out of the physical data store. One such paradigm is OLAP which is used to easily navigate and visualize the data. It is all about summation, aggregation, summarization of the information storing it in a multi-dimensional format scenario when they offer data mining solutions at the database level. OLAP and data mining seem to be the same due to the perception one holds of their analytical function. However the difference lies in the type of analysis they are concerned with. As far as type is concerned, analysis can be retrospective (which focuses on the issues of the past and present events) or predictive (which focuses on predicting certain events or behavior based on the historical information). The retrospective analysis is possible through OLAP easily. So, the answer to certain questions such as "What has been going on?" can be provided by the Active Data Warehouse and multidimensional database technology (OLAP). However, the predictive analysis is not possible wherein the answer to the question like "What next?" cannot be provided by OLAP. Among several techniques that can help us in answering such complex questions, we have the Statistical Concept which is based on classical mathematical methods and is concerned with probabilistic theory including Bayes Theorem, thus providing effective ways of processing uncertain information.

Once the fresh data is ready for the analysis through Active or Dynamic Data Warehouse, we can use it in certain problems where we want to simulate the human intelligence, well known by the name Artificial Intelligence. This is an attempt to emulate the brain working with programming methods, for example building a program that plays chess. As far as the future prediction is concerned, we use the concept of Heuristics in rule based Intelligent Expert-System and for the 'almost natural' prediction we try to incorporate the fuzzy logic as well. Fuzzy logic provides a way of processing uncertain information wherein a degree of membership is associated to each element of the fuzzy set, in contrast to the normal set in which each element has associated limited value i.e. either TRUE or FALSE.

In addition to this, we have optimization techniques such as Genetic algorithm that is based on the concept of natural evolution and works on the basic principle i.e. "survival of the fittest" under which crossover and mutation play an important role to find the hidden unknown solution of any problem e.g. 8-Queen problem can be easily solved using the Genetic Algorithm. The knowledge representation models or mechanisms are often based on Logic or Rules. The aspects of Artificial Intelligence bear a very close relationship to formal logic. Logic is the philosophical study of valid reasoning wherein Artificial Intelligence begins with the proposition calculus that deals with the statements with values 'true' and 'False' and is concerned with analysis of proposition .In order to manipulate the expressions in Propositional logic , we have the rules of Inference (Modus Ponens or modus Tollens)[3,15].

Artificial Intelligent system is a knowledge based system that can provide us the ability to solve the complex problems, however, such systems rely on experts who possess specialized knowledge of some problem domain. No doubt, even a brilliant expert is only a human and thus can make mistakes which can finally lead to wrong decision making. In addition to this, the more data shows the more changes which further needs the creation of more rules that can lead to complex rule based system.

Apart from this, above techniques (specially the statistical inference) can prove to be useful when we want an answer to "who", "what", "where" and "when". However, as far as strategic decision making is concerned, we are still left with some unanswered questions like "How" or "why".

IV. ADVANCEMENT TOWARDS DATA MINING:

The next advancement in the knowledge extraction is when a machine attempts to learn its own rules. Data mining, considered to have been originated from three branches of artificial intelligence-- neural networks, machine-learning and genetic algorithms leads us to such advancement. Data mining is actually the process of uncovering the fluctuating hidden patterns or trends in the data that is not immediately apparent by just summarizing the data. It is used to predict the future (predictive analytics) in addition to explain the current or past situation (descriptive analytics). After the interpretation of the information, knowledge can be extracted by identifying relationships among patterns. This can provide the answer to "How". The principles of such relationships describe the patterns and can provide the answer to "why". Now following represent the overall steps for performing the process of data mining:

1. Building a fresh Data Warehouse system using the process of ETL or ELT.
2. Loading the Data Warehouse data into multidimensional databases.
3. Data Analysis (using application software).
4. Knowledge presentation (using graph , table , association rules etc)

Actually, for many cases several concepts from Statistical Inference have been incorporated into data mining but the fundamental difference lies in the way they perform the function. In case of statistics, once a conceptual model is built (null hypothesis), we go for the validation of that hypothesis which leads us to the final acceptance or rejection of the null hypothesis. In contrast, data mining works almost in an opposite way wherein the first step does not start with the null hypothesis. Rather we just have a data set and we don't really know what and which pattern we are looking for. So, here we start by applying the interestingness criteria (notion) over the dataset in an attempt to get some interesting patterns forming the basis of the hypothesis thus the name "Hypothesis discovery". The data in the data set can be of any type from temporal to multimedia and the interestingness criteria can be frequency, rarity, correlation, consistency or periodicity etc. Two concepts rule the process of data mining i.e. 'Item Set' and 'Association' that are managed by the thresholds 'Support' and 'Confidence' respectively. We have certain data mining algorithms like Apriori algorithm and brute-force algorithm for frequent item set and association rule mining. Apart from mining for associations, we try to discover classification trees and clusters within a given data set, well known by the process of Classification and clustering respectively. The two techniques intuitively seem to do the same thing but they differ from each other if we observe closely. While clustering (unsupervised--no training data is provided to train) groups the data into different clusters based on similarity, the classification process (supervised—training data is provided to train) groups data into different classes on the basis on difference between the data elements. The technique of classification provides us different forms of representation:

Decision tree: Tree-shaped structures that generate rules for the classification of a dataset. Specific decision tree techniques include Classification and Regression (CART) and Chi Square Automatic Interaction Detection (CHAID) wherein CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits.

1. Artificial Neural Network: Non-linear predictive models resembling biological neural networks in structure that comprises of links, associated weights and nodes (input, hidden, output) and learn through training. For attempting to build a perfect neural network we go for the back propagation of errors from the output nodes through the hidden layers to modify the link weights. However, such network is less able to explain the internal processing.
2. IF-THEN rules: under such rule the antecedent can be any condition on the data elements and the consequent will be the class group to which the data elements satisfying the corresponding antecedent belong.

As far as clustering is concerned we have several techniques such as following:

1. Nearest neighbor method: A technique that finds the nearest neighbor of a data element by taking into consideration the maximum distance (acting as threshold) that can exist between elements of cluster.

2. Iterative partitional clustering: Under such technique, we are given fixed number of clusters, not knowing how and what elements exist in each cluster and we try to rearrange elements in clusters by computing cluster centroid.

Further, efforts are being made towards the exploration of knowledge by providing improved scalable interactive methods. The main aim is to be able to find certain patterns or trends in the data and forecast the future values of the data. Investigating data mining process, user interface issues, database topics, or visualization has always been a point of concern in the research area [7]. However the emphasis should now be on the issue related to the most current and recent changes of voluminous data necessary to act upon the results of intelligent analyses.

V. CONCLUSION AND FUTURE INSIGHTS

Knowledge is hidden behind the voluminous data available through the OLTP systems. We need to analyze this data in an effective manner using a Data Warehouse. In this paper we presented architecture for the Data Warehouse that, in addition to the historical data, can provide us the most recent data as well. Further, we examined in a systematic way a wide range of techniques that are available to us for extracting the recent information / knowledge out of the data present in such a real-time Data Warehouse. Selection of a suitable technique depends on the type of query we are concerned with. The paper provides a flow in an attempt to always provide an up-to-date knowledge. However this does not end up problems behind the knowledge exploration. The data mining technique can prove to be more powerful when, in addition to the most active data, the streaming data can also be handled easily where the two critical technological issues are size of the database and the query complexity.

REFERENCES

[1] Shweta Pandya and Bhaumik Shroff , *Data Warehouse : Intelligent Management Decision Support*, 2nd International CALIBER-2004, New Delhi, 11-13 February, 2004

[2] Vincent Rainardi , *Building A Data Warehouse, With examples in Sql Server,* Apress- (2008)

[3] Alex Berson, Stephen J. Smith , *Data Warhousing, Data Mining and OLAP* , Tata McGraw-Hill Education, 01-Mar-2004.

[4] V.Mallikarjuna Reddy, Sanjay K Jena, *Active Data Warehouse Loading by Tool Based ETL Proce*dure, Dept. Of Computer Science, National Institute of Technology Rourkela, India

[5] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*,2nd edition (reprint 2011)

[6] Clifton, Christopher , *Encyclopædia Britannica: Definition of Data Mining* -(2010) http://www.britannica.com/EBchecked/topic/1056150/data-mining.

[7] U.M.Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy(eds.), *Advances in Knowledge Discovery and Data Mining,* AAAI Press 1996, p. 1-34.

[8] Hand D.J., Mannila H., and Smyth P., *Principles of data mining,* MIT Press (2001)

[9] David J. Hand ,*Statistics and Data Mining: Intersecting Disciplines* ACM SIGKDD Explorations, June 1999

[10] Hoffmann F., Hand D.J., Adams N., Fisher D., and Guimaraes G. (eds) , *Advances in Intelligent Data Analysis*, Springer- (2001)

[11] Karakasidis, A., Vassiliadis and P., Pitoura , *Extraction Transformation Loading-A ETL Queues for Active Data Ware*- housing, Proceedings of the 2nd International Workshop on Information Quality in Information Systems (IQIS'2005), New York, USA, ACM Press 28-39,2005

[12] Dr. R.S.Chhillar, BarjeshKochar , *Road to Data Warehouse* , 2nd National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries, sep 26-28,2008.

[13] Ricardo Jorge Santos and Jorge Bernardino, *Real-Time Data Warehouse Loading Methodology* , IDEAs'o8, ACM - 2008.

[14] R. Agrawal, T. Imielinksi and A. Swami, *Mining association rules between sets of items in large database,* The ACMSIGMOD Conference, pp. 207-216, Washington DC, USA, 1993.

[15] John L.Gordon , *From logic to Fuzzy Logic*, 2005 , AKRI Ltd. Applied Knowledge Research & Innovation.

[16] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, *Mining optimized association rules for numeric attributes*, The ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 182-191, 1996.

[17] ACM SIGKDD, *Data Mining Curriculum* , 2006-04-30. http://www.sigkdd.org/curriculum.php.

[18] Jiawei Han and Jing Gao , *Research Challenges for Data Mining in Science and Engineering* ,University of Illinois at Urbana-Champaign – 2008

[19] Manole VELICANU, Academy of Economic Studies, Bucharest Gheorghe MATEI, Romanian Commercial Bank, *Building a Data Warehouse step by step* , Informatica Economică, nr. 2 (42)/2007

[20] Saida Aissi1, Mohamed Salah Gouider , *Towards the Next Generation of Data Warehouse -Personalization System-A Survey and a Comparative Study*, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, May 2012.

[21] Lori Bowen Ayre, *Data Mining for Information Professionals,*June 2006

[22] Ziyu Lin1, Dongzhan Zhang1, Chen Lin, Yongxuan Lai, and Quan Zou, *Performance Optimization of Analysis Rules in Real-time Active Data Warehouses,* School of Information Science and Technology, Xiamen University, Xiamen, China, Springer-Verlag Berlin Heidelberg 2012

[23] Joseph Guerra & David Andrews, *Why You Need a Data Warehouse,* Copyright Andrews Consulting Group, Inc. 2011.

[24] Philip Russom, *Next generation -Data Warehouse Platforms,* fourth quarter 2009, TDWI best practices Report.