PERFORMANCE ENHANCEMENT OF DATA WAREHOUSE USING STATISTICAL DISTRIBUTION FUNCTIONS WITH HEURISTIC

Bikramjit Pal [1], Dr. Mallika De [2]

[1] Assistant Professor, Department of Computer Application
JIS College of Engineering, Kalyani, Nadia – 741235, West Bengal, India
[2] Head, Department of Engineering and Technological Studies
University of Kalyani, Kalyani, Nadia – 741235, West Bengal, India

ABSTRACT

The sole motive of this paper is to show that how the performance of a data warehouse can be improved by implementing statistical tools and heuristics on the functionality of data warehouse so that the performance can be enhanced. While this entire process of data creation for analysis moves on, the actual data warehouse uses a heuristic to store the needed data from a buffer which can be viewed as a storage device. The end data warehouse user will able to view the updated information from a temporary storage area.

INTRODUCTION

The proposed work is mainly a multiprocessor system where data from various databases are put on to the ready queue for processing in the ETL. The data first in the queue will be processed first in the ETCL. The processed data i.e. information from the ETCL will be moved onto a temporary storage area. The processed data will remain there until all other data from the database get processed and get collected in the temporary storage. Finally these processed data will be moved to the Real-Time data warehouse.

ETL: **Extract, transform and load** (**ETL**) is a process in database usage and especially in data warehousing that involves:

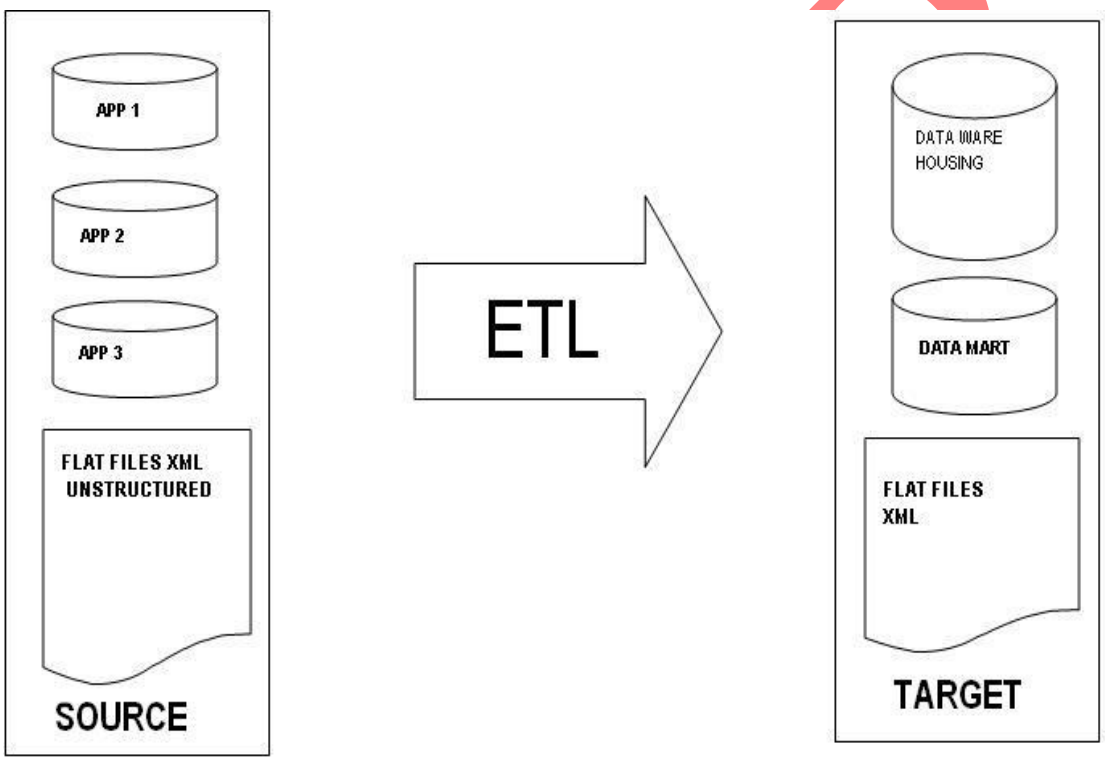
- Extracting data from outside sources
- Transforming it to fit operational needs (which can include quality levels)
- Loading it into the end target (database or data warehouse

Extract: The first part of an ETL process involves extracting the data from the source systems. In many cases this is the most challenging aspect of ETL, as extracting data correctly will set the stage for how subsequent processes will go.

Transform: The transform stage applies a series of rules or functions to the extracted data from the source to derive the data for loading into the end target. Some data sources will require very little or

even no manipulation of data. In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the target database.

Load: The load phase loads the data into the end target, usually the data warehouse (DW). Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information, frequently updating extract data is done on daily, weekly or monthly basis. Other DW (or even other parts of the same DW) may add new data in a historicized form, for example, hourly. To understand this, consider a DW that is required to maintain sales records of the last year.



HEURESTIC SEARCH TECHNIQUE: Heuristic search is an **AI search** technique that employs heuristic for its moves. *Heuristic* is a rule of thumb that probably leads to a solution. Heuristics play a major role in search strategies because of exponential nature of the most problems. Heuristics help to reduce the number of alternatives from an exponential number to a polynomial number. Common heuristic techniques are:

1. Simulated annealing algorithm, invented in 1983, uses an approach similar to hill-climbing, but occasionally accepts solutions that are worse than the current. The probability of such acceptance is decreasing with time.

2. Tabu search extends the idea to avoid local optima by using memory structures. The problem of simulated annealing is that after "jump" the algorithm can simply repeat its own track. Tabu search prohibits the repetition of moves that have been made recently.
3. Swarm intelligence was introduced in 1989. It is an artificial intelligence technique, based on the study of collective behavior in decentralized, self-organized, systems. Two of the most successful types of this approach are Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO). In ACO artificial ants build solutions by moving on the problem graph and changing it in such a way that future ants can build better solutions. PSO deals with problems in which a best solution can be represented as a point or surface in an n-dimensional space. The main advantage of swarm intelligence techniques is that they are impressively resistant to the local optima problem.
4. Evolutionary Algorithms succeed in tackling premature convergence by considering a number of solutions simultaneously.
5. Neural Networks are inspired by biological neuron systems. They consist of units, called neurons, and interconnections between them. After special training on some given data set Neural Networks can make predictions for cases that are not in the training set. In practice Neural Networks do not always work well because they suffer greatly from problems of under fitting and over fitting. These problems correlate with the accuracy of prediction. If a network is not complex enough it may simplify the laws which the data obey. From the other point of view, if a network is too complex it can take into account the noise that usually assists at the training data set while inferring the laws. The quality of prediction after training is deteriorated in both cases. The problem of premature convergence is also critical for Neural Networks.
6. Support Vector Machines (SVMs) extend the ideas of Neural Networks. They successfully overcome premature convergence since convex objective function is used; therefore, only one optimum exists. Classical divide and conquer technique gives elegant solution for separable problems. In connection with SVMs, that provide effective classification, it becomes an extremely powerful instrument. Later we discuss SVM classification trees, which applications currently present promising object for research. Description and comparative analysis of simulated annealing, tabu search, neural networks and evolutionary algorithms can be found in.

NORMAL DISTRIBUTION: Normal distribution is a continuous distribution that has a bell shaped probability density function given by

$$F(x) = [\exp(-(x-\mu)^2/ 2\sigma^2] / \sigma \sqrt{2\pi} \text{ ---------- (i)}$$

where μ is the mean and σ^2 is variance.

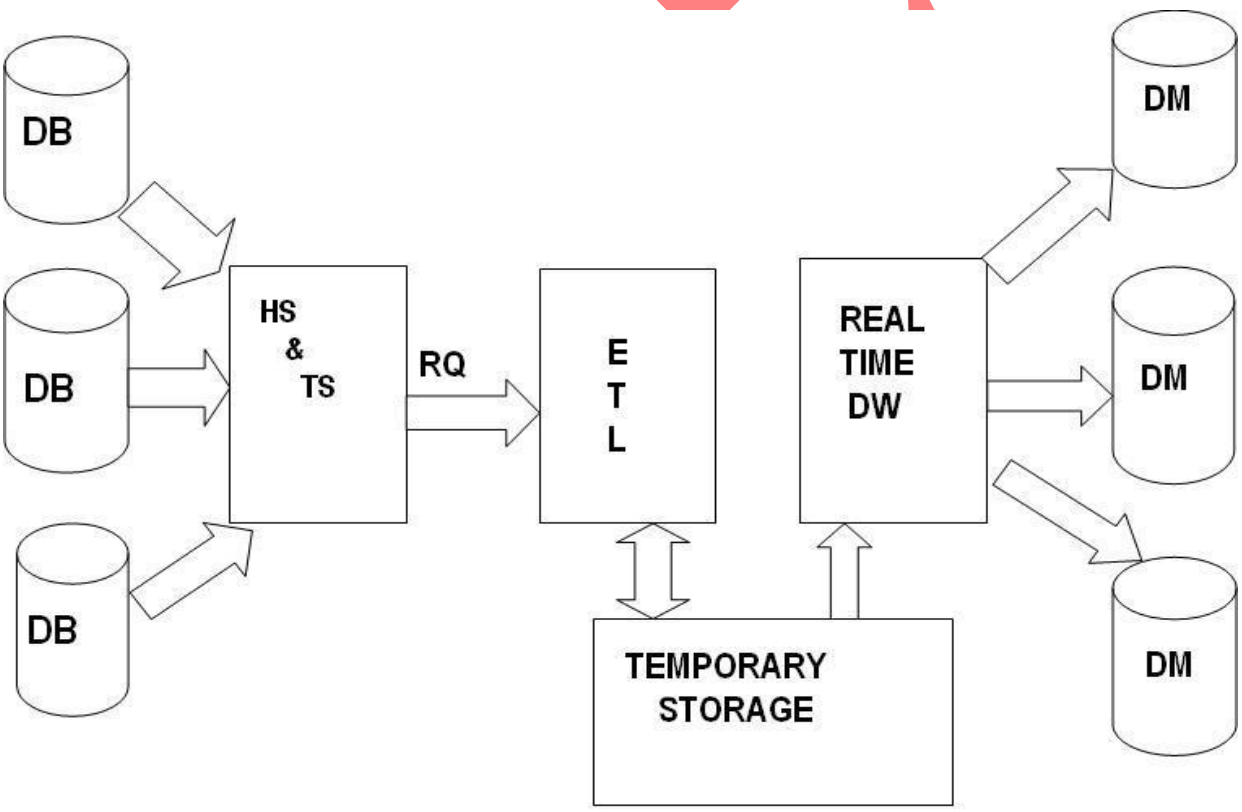
POISSON DISTRIBUTION: If a count x has a Poisson distribution and if the expected value of the count is λ then the probability that x is chosen is given by

$$F(x) = (e^{-\lambda} \lambda^{x}) /! \; x \text{ ---------- (ii)}$$

PROPOSED WORK: The Proposed work is an implementation of Normal and Poisson Probability distribution to the data set coming from various data bases so that activities in the staging area can be performed in a more sophisticated way where ETL processing of fresh data and simultaneously viewing the old and new information from data mart and a temporary storage respectively takes place. We have considered two approaches for the data arriving from the databases.

CASE – I --→ For continuous stream of data coming from various data bases for ETL we implemented Normal distribution as given by eqn. (i). That is, data follow normal distribution. The advantages of using Normal Distribution are that it is tractable and is analytically sound, thus generating data for good analysis purpose. This gives a real time view of the system.

CASE – II -→ For discrete flow of data from data bases to the staging area we implemented Poisson Distribution to determine the functional rate as given by eqn. (ii). Poisson distribution is used to model situation where the data population is very large but the probability of a particular event at each trial approaches zero. We have implemented this in the batch mode of processing.



Data in both the cases enters the temporary storage and then to the data warehouse. We have implemented a heuristic search technique, Swarm Intelligence, in the warehouse so that only timely needed data will enter the warehouse. The implementation is as follows:

Let the data warehouse be in a distributed environment. Data can be able to interact with each other directly or indirectly in that distributed environment. We have implemented altruism algorithm which is based on Hamilton's rule of kin selection. Altruism is the property where one entity saves the life of other or group of entities for the benefit of the entire group. Here, since the data is completely distributed over a certain region, implementation of altruism algorithm what we think will definitely remove data starvation process by changing the priorities. Therefore, if two or more tables are related through foreign key, each table will access the fitness of other tables by using Hamilton's rule.

CONCLUSION

We have tried to theoretically establish that the data coming from various sources can follow normal and poison distribution as per their arrival and also given the concept of a temporary storage from where data can be sent to the data warehouse. The warehouse then uses a heuristic search and implements altruism technique to change priorities among data stored in various tables. This process will definitely enhance the performance by proper scheduling data in various tables in terms of time required for analytical searching. Implementations of various heuristics can be a further scope of this work.

REFERENCES

1. Pal Bikramjit, Rout Laxmi, Shah Priyadarshani, Dr. De Mallika, "Logical Data Warehouse Design Using Distributed Schema Architecture", International Journal of Computer Science and Technology, Vol. 3, Issue 2, April – June 2012, pp. 191 – 194

2 Chowdhury R., Pal B.,"Proposed Hybrid Data Warehouse Architecture Based on Data Model", International Journal of Computer Science and Communication, Vol. 1, No. 2, 2010, pp. 211–213.

3 Pal B.,"Comparison of Data Warehouse Architecture Based on Data Model", International Journal of IT and Knowledge Management, 2010, No. 2, pp. 303 – 304 .

4 Thomas Stöhr, Holger Märtens, Erhard Rahm, "Multidimensional Database Allocation for Parallel Data Warehouses", Proceedings of 26th International Conference on Very large Databases, Cairo, Egypt, 2000.

5 Smith J., David Van Dyken J., Zee C. Peter, "A Generalization of Hamilton's Rule for the Evolution of Microbial Cooperation" Science 25 June 2010: Vol. 328 no. 5986 pp. 1700-1703

6 http://www.cs.utexas.edu/~dana/Ch11.pdf

AUTHOR PROFILE

Mr. Bikramjit Pal is currently working as Assistant Prof. and Head in the Department of Computer Application, JIS College of Engineering, Kalyani, Nadia, West Bengal from October 2006. He has an overall experience of more than 16 years in teaching post graduate and under graduate students. He is an MCA from Berhampur University, Berhampur, Orissa of 1998 batch and has obtained B. Sc. (Hons.) degree in Statistics from B. H. U., Varanasi in 1995. Mr. Pal has been involved in active research for the last four years and presently pursuing Ph. D. from Department of Engineering and Technological Studies, University of Kalyani, West Bengal.