World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:1, 2007

Two-Phase Optimization for Selecting Materialized Views in a Data Warehouse

Jiratta Phuboon-ob, and Raweewan Auepanwiriyakul

***Abstract*—**A data warehouse (DW) is a system which has value and role for decision-making by querying. Queries to DW are critical regarding to their complexity and length. They often access millions of tuples, and involve joins between relations and aggregations. Materialized views are able to provide the better performance for DW queries. However, these views have maintenance cost, so materialization of all views is not possible. An important challenge of DW environment is materialized view selection because we have to realize the trade-off between performance and view maintenance. Therefore, in this paper, we introduce a new approach aimed to solve this challenge based on Two-Phase Optimization (2PO), which is a combination of Simulated Annealing (SA) and Iterative Improvement (II), with the use of Multiple View Processing Plan (MVPP). Our experiments show that 2PO outperform the original algorithms in terms of query processing cost and view maintenance cost.

***Keywords*—**Data warehouse, materialized views, view selection problem, two-phase optimization.

I. Introduction

A data warehouse (DW) can be defined as subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decision [1]. It can bring together selected data from multiple database or other information sources into a single repository [2]. To avoid accessing from base table and increase the speed of queries posed to a DW, we can use some intermediate results from the query processing stored in the DW called materialized views. Although materialized views speed up query processing, they have to be refreshed when changes occur to the base tables. Therefore, materialized view selection involved query processing cost and materialized view maintenance cost. So, many literatures try to make the sum of that cost minimal. For all of operation, i.e., select, project, join, order, group-by and aggregation operation; join operation has the most impact on query processing cost. In addition, some researchers consider only join order optimization or aggregation operation, or both. The existing algorithms solving query optimization, multiple query optimizations, and materialized view selection can be classified into four categories, i.e., deterministic algorithm, randomized algorithm, evolutionary algorithm and hybrid algorithm [3]. In our previous work [4], we analyzed and compared only three types of algorithm; deterministic algorithm, evolutionary algorithm and hybrid algorithm. However in this paper, we explore the application of randomized algorithm. We address Two-Phase Optimization (2PO) algorithm, which is a combination of Simulated Annealing (SA) and Iterative Improvement (II), to the materialized view selection problem with MVPP techniques. In the experimental study we show that, comparing to [5] and [10], our method provides a further improvement in term of query processing cost and view maintenance cost.

The rest of the paper is organized as follows. Section 2, we describe Multiple View Processing Plan (MVPP). Section 3, we focus on Iterative Improvement and Simulated Annealing. Section 4, we propose our Two-Phase Optimization approach which aimed to solve the materialized view selection problem. Section 5, deal with our experimental studies, and conclude in section 6.

Manuscript received November 30, 2006. This work was supported in part by the U.S. Department of Commerce under Grant.

Jiratta Phuboon-ob study at School of Applied Statistics, National Institute of Development Administration (NIDA), Bangkok, Thailand , on leave from Mahasarakham university, Thailand (e-mail: jiratta.p@ msu.ac.th).

Raweewan Auepanwiriyakul is an Assistant Professor with School of Applied Statistics, National Institute of Development Administration (NIDA), Bangkok, Thailand (e-mail: raweewan@as.nida.ac.th).

II. Multiple View Processing Plan (MVPP)

According to [5], they use multiple query processing technique (MQP) to build multiple views processing plan (MVPP) in order to identify views to be materialized.

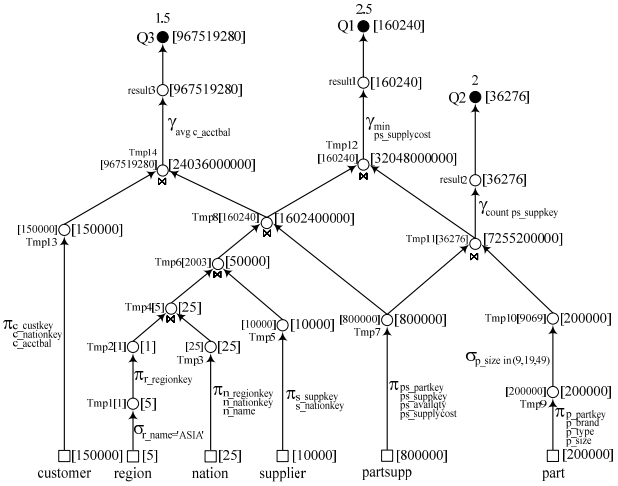

Fig. 1 An example MVPP plan

An MVPP is a directed acyclic graph that represents a query processing of DW views. An example MVPP is shown

in Fig. 1. The leaf nodes correspond to the base relations, and the root nodes represent the queries. Any vertex which is an intermediate or a final result of a query is denoted as a view. The cost for each operation node is labeled at the right side of the node. The query access frequencies are labeled on the top of each query node.

III. Iterative Improvement and Simulated Annealing

A. Iterative Improvement (II)

[7] exposed II algorithm to the large join query optimization problem. II is a simple hill-climbing algorithm. This algorithm performs a large number of local optimizations. A local optimization starts at a random state, and seeks minimum cost point using a strategy-like hill-climbing. At the starting point, a random neighbor is selected. If the neighbor's cost is lower than current's cost, the move is carried out and a new neighbor with the lower cost is sought. II performs random series of move and accepts only downhill ones until it reaches a local minimum. This algorithm is repeated until a time limit is exceeded or a predetermined number of starting points is processed, then the lowest local minimum encountered is the result. The II algorithm present in Fig. 2.

1. random starting point
2. at the starting point, a random neighbor is selected
3. if the neighbor's cost is lower than current's cost then move is carried out and a neighbor is sought
4. performs random series of move and accepts only downhill ones until it reaches a local minimum

Fig. 2 Iterative Improvement (II) algorithm

1. random starting point
2. at the starting point, a random neighbor is selected
3. compare neighbor's cost and current's cost
   3.1 if the neighbor's cost is lower than current's cost then move is carried out and a neighbor is sought
   3.2 otherwise moving to neighbor's state or staying in current's state with probability
4. performs random series of move and accepts both of downhill and uphill until it reaches a local minimum

Fig. 3 Simulated Annealing (SA) algorithm

B. Simulated Annealing (SA)

[8] invented SA algorithm, and used it on traveling sale man problem. SA follows a procedure similar to II, but it accepts uphill move with some probability, while II performs only downhill move. At each step, the SA considers neighbor's cost of the current's cost, and probabilistically decides among moving the system to neighbor's state or staying in current's state. The probabilities are chosen, so the system ultimately tends to move to states of the lower cost. This step is repeated until the time becomes zero, or until the system reaches a state which is good enough for the application. [9] applied this algorithm to the optimization of some recursive queries.

In [10], they introduced a new approach for materialized view selection based on SA in conjunction with the use of a MVPP. Fig. 3 shows SA algorithm.

IV. Our Approach: Two-Phase Optimization for Selecting Materialized View

Ioannidis and Kang [6] inspired Two-Phase Optimization (2PO) algorithm to the optimization of project-select-join queries. Our approach is designed based on 2PO with MVPP for solving the materialized view selection problem. 2PO combines both SA and II. It begins by running II to find a good local minimum, before applying SA to search for the global minimum from the state found by II. Our algorithm present in Fig. 4.

1. Input a MVPP represented by a DAG
2. Use width-first searching method to search through all of the nodes in the DAG and produce an ordered sequence of these nodes into a binary string
3. Call Iterative Improvement algorithm
4. Call Simulated Annealing algorithm
5. Present set of views to materialized with minimum cost

Fig. 4 Our Two-Phase Optimization (2PO) algorithm

In the following subsection, we give the details of representation of solutions and define the cost model of materialized view selection.

A. Representation of Solutions

Output from MVPP is a DAG. We do not directly use a DAG as the input of our 2PO algorithm. Instead, we first map it into a binary string. For example, search through the DAG, shown in Fig. 1, using width-first, we obtain the mapping array, i.e. [result3,0], [result1,0], [result2,0], [tmp14,0], [tmp12,0], [tmp11,0], [tmp13,0], [tmp8,0], [tmp10,0], [tmp6,0], [tmp7,0], [tmp9,0], [tmp4,0], [tmp5,0], [tmp2,0], [tmp3,0], [tmp1,0]. A binary string of {0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0} indicates that none of node is materialized. A binary string of {0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0} implies that nodes {tmp13, tmp7} are materialized, but others are not.

B. Cost Model of Materialized View Selection

According to [5], a linear cost model is used to calculate the cost of query Q. The cost of answering Q is the number of rows in the table that query Q_i used to construct Q.

Let M be a set of materialized views, $C_{q_i}(M)$ be the cost to compute q_i from the set of materialized views M, $C_m(v)$ be the cost of maintenance when v is materialized, and f_q, f_u are query and updating frequency, respectively. Then the total query processing cost is $\sum_{q_i \in Q} f_{q_i} C_{q_i}(M)$. The total maintenance cost is $\sum_{v \in M} f_u C_m(v)$.

The total cost of the materialized views M is

$$\sum_{q_i \in Q} f_{q_i} C_{q_i}(M) + \sum_{v \in M} f_u C_m(v)$$

Our goal is to find the set M, if the members of M are materialized then the value of total cost will be minimal among all feasible sets of materialized view.

V. Experimental Studies

In our experiment, we employ MQP technique to all of three algorithms and use the same cost model proposed by [5] to compute query processing cost, materialized view maintenance cost and total cost. We do not consider any constraints. We use the TPC-H database of size 1GB as a running example throughout our paper. It has 22 read-only queries. Most of them are large and complex, and perform different operations on the database table. For more details on this benchmark refers to www.tpc.org. In this paper, we cover all of regular aggregate functions; MIN, MAX, SUM, AVG, COUNT, VARIANCE and STDDEV as listed in Fig. 5.

Query 1 (MIN)
```
select  min(ps_supplycost)
from    part, partsupp, supplier, nation, region
where  p_partkey = ps_partkey
        and s_suppkey = ps_suppkey
        and s_nationkey = n_nationkey
        and n_regionkey = r_regionkey
        and r_name = 'ASIA';
```

Query 2 (MAX)
```
select  max(o_totalprice)
from    customer, orders, lineitem, nation, region
where c_custkey = o_custkey
        and o_orderkey = l_orderkey
        and c_nationkey = n_nationkey
        and n_regionkey = r_regionkey
        and r_name = 'ASIA'
        and o_orderdate >= '1994-01-01'
        and o_orderdate < '1995-01-01';
```

Query 3 (SUM)
```
select  n_name, sum(l_quantity)
from    orders, lineitem, supplier, nation, region
where l_orderkey = o_orderkey
        and l_suppkey = s_suppkey
        and s_nationkey = n_nationkey
        and n_regionkey = r_regionkey
        and r_name = 'ASIA'
        and o_orderdate >= '1994-01-01'
        and o_orderdate < '1995-01-01'
group by n_name;
```

Query 4 (AVG)
```
select avg(c_acctbal)
from    partsupp, supplier, customer, nation, region
where ps_suppkey = s_suppkey
        and c_nationkey = s_nationkey
        and s_nationkey = n_nationkey
        and n_regionkey = r_regionkey
        and r_name = 'ASIA';
```

Query 5 (COUNT)
```
select  count(ps_suppkey)
from    partsupp, part
where p_partkey = ps_partkey
        and p_brand <> 'Brand#45'
        and not p_type like '%BRASS%'
        and p_size in (9, 19, 49);
```

Query 6 (VARIANCE)
```
select variance(ps_supplycost)
from    supplier, partsupp, part
where s_suppkey = ps_suppkey
        and p_partkey = ps_partkey
        and p_brand <> 'Brand#45'
        and not p_type like '%BRASS%'
        and p_size in (9, 19, 49);
```

Query 7 (STDDEV)
```
select  stddev(l_tax)
from    customer, orders, lineitem
where c_custkey = o_custkey
        and o_orderkey = l_orderkey
        and o_orderdate >= '1994-01-01'
        and o_orderdate < '1995-01-01';
```

Fig. 5 Our Queries

We observe that these queries are defined over overlapping portion of the base data or intermediate query result. For example Query 5 and Query 6 can share the intermediate result of joining part and partsupp. We assume that base table update once a time and the frequencies of Query 1 to Query 7 are 4, 6, 7, 2, 5, 9, 3 respectively.

We use the algorithm to generate MVPP proposed by [5]. Fig. 6 gives accessing plan for each of the above queries, denoted as op1, op2, op3, op4, op5, op6, op7 respectively. As a consequence, first of all, we push up all select, projects, and group-by operation. Second, we create a list of query and order them based on the result of query access frequency multiplied by query processing cost. Therefore the initial list is {op4, op7, op3, op2, op1, op6, op5}.Third, we pick up op4, and merge the rest of the queries with it in the order of that in the list. Then we get the first MVPP. Fourth, the first element of list is moved to the end of the list, so the list becomes {op7, op3, op2, op1, op6, op5, op4}. We generate the second MVPP by using third step. We repeat the third and fourth step to generate all 7 MVPPs. Next, we push down select, project, and group operations respectively for all of MVPPs. Finally, we select the cheapest one; shown in Fig. 7. In our MVPP, we assume that methods for implementing select and join operation are linear search and nested loop approach. Before comparing the cost, we compute query processing cost, materialized view maintenance cost and total cost of all-virtual-views and all-materialized-view, demonstrated in Table I. Table II gives the selected views and their cost from each algorithm. We compare these costs between three algorithms as following:

In deterministic algorithm, given a MVPP, we execute the view selection algorithm proposed by [5] to select materialized view. The view selected are Tmp11, Tmp15, Tmp17, Tmp21 and Tmp24. Based on these results, it would be benefit to materialize them, reducing the cost from 7,688,720,739,017 to 6,184,919,079,222.

According to simulated annealing algorithm, we use an existing simulated annealing package [11] and define this problem based on [10]. We first map the solution of view selection problem into a binary string of 1s and 0s. For example, search through the DAG, shown in figure 7, using width-first, we obtain the mapping array, i.e. [result2,0], [result7,0], [result3,0], [result4,0], [result1,0], [result6,0], [result5,0], [tmp15,0], [tmp25,0], [tmp11,0], [tmp23,0], [tmp19,0], [tmp22,0], [tmp21,0], [tmp14,0], [tmp8,0], [tmp17,0], [tmp16,0], [tmp20,0], [tmp13,0], [tmp6,0], [tmp24,0], [tmp18,0], [tmp12,0], [tmp4,0], [tmp5,0], [tmp7,0], [tmp10,0], [tmp2,0], [tmp3,0], [tmp9,0] and [tmp1,0]. According to our search method, A binary string of {1,1,0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1} means that nodes named result7, result2, tmp9 and tmp1 are materialized, but others are not. We set SA parameters like [10] the results are Tmp5, Tmp11, Tmp15, Tmp17, Tmp18, Tmp21 and Tmp24. Based on these results; it would be benefit to materialize them, reducing the total cost from 7,688,720,739,017 to 6,184,918,609,222.

For our two-phase optimization algorithm, we map a DAG into a binary string using the same method as used by SA. Then we run II and then followed by SA. The selected views are Tmp5, Tmp11, Tmp12, Tmp15, Tmp17, Tmp18, Tmp21 and Tmp24. Based on these results; it would be benefit to materialize them, reducing the cost from 7,688,720,739,017 to 6,184,918,159,222.

Table II compares our 2PO algorithm result to the deterministic algorithm and SA algorithm for materialized view selection. This table shows that our 2PO algorithm approach provides a better result than deterministic algorithm and SA algorithm. Although our maintenance cost is the most expensive, however, our query processing cost is the cheapest one. So our total cost is minimal. Consider the result of

deterministic algorithm, its maintenance cost is the cheapest, however, its query processing cost is the most expensive this lead to its total cost the most expensive. For SA algorithm, its query processing cost and maintenance cost is moderate, so its total cost is moderate too.

VI. Conclusion

In this paper, we introduce Two-Phase Optimization (2PO) algorithm, which is a combination of Simulated Annealing (SA) and Iterative Improvement (II), to materialized view selection with Multiple View Processing Plan (MVPP) proposed by [5]. Comparing to deterministic algorithm proposed by [5] and Simulated Annealing proposed by [10], our approach provides a better result than Deterministic algorithm and Simulated Annealing algorithm. Two-Phase Optimization finds the best solution, and avoids unnecessary large uphill moves at the early stages of Simulated Annealing.

Acknowledgment

We are grateful to Wanida Prapavasit, Roozbeh Derakhshan and Dyke Stiles for helping by way of discussion and provide material for this paper.

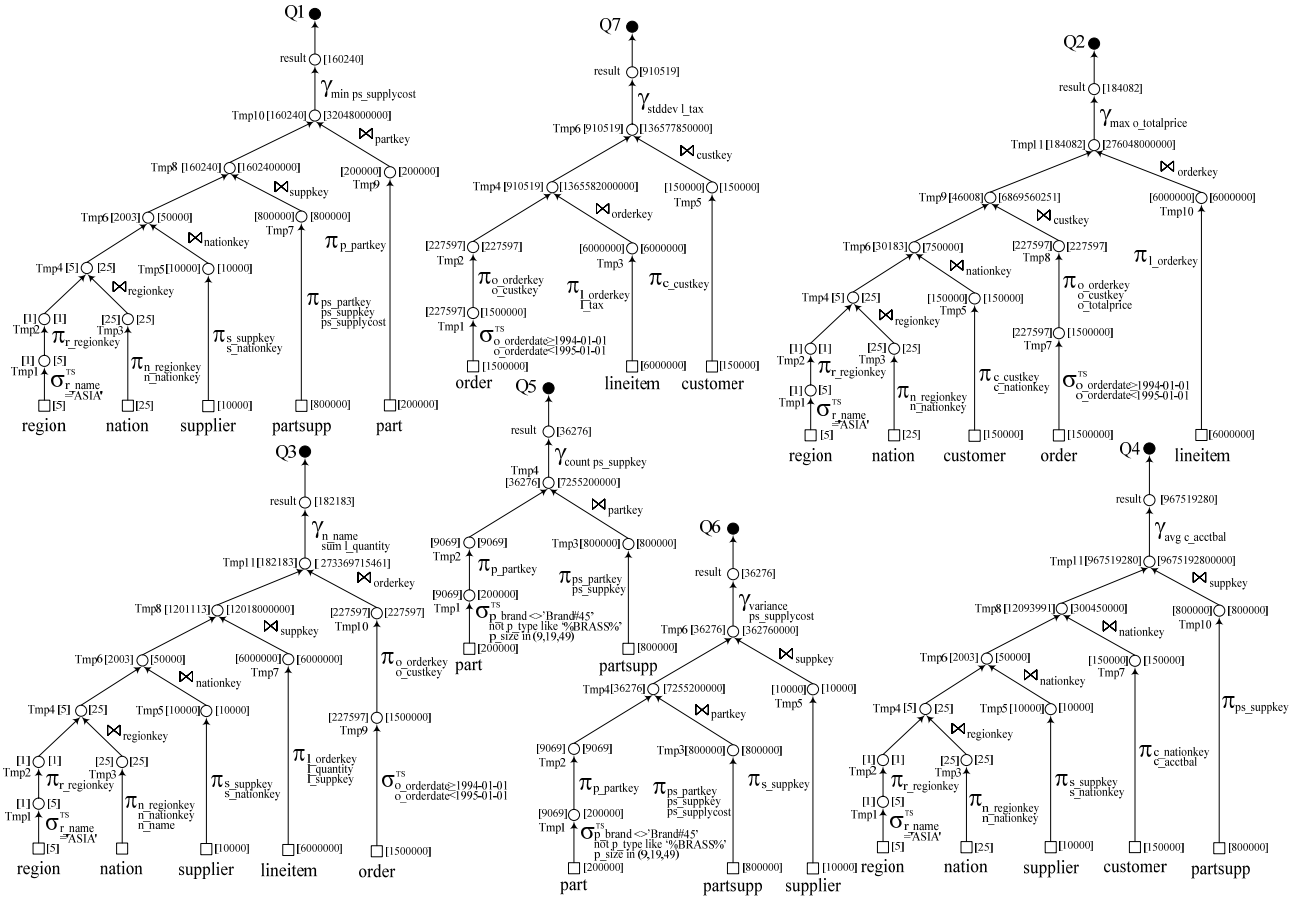

Fig. 6 Individual Optimal Query Processing Plan

TABLE I
The Query Processing, Maintenance and Total Cost

	Cost of query processing	Cost of maintenance	Total Cost
All-virtual-views	8,494,509,321,063	-	8,494,509,321,063
All-materialized-views	1,941,298,714	7,686,779,440,303	7,688,720,739,017

TABLE II
The Query Processing, Maintenance and Total Cost for Each Algorithm

Algorithm	Selected views	Cost of query processing	Cost of maintenance	Total Cost
Deterministic	Tmp11, Tmp15, Tmp17, Tmp21, Tmp24	591,205,328,714	5,593,713,750,508	6,184,919,079,222
Simulated Annealing	Tmp5, Tmp11, Tmp15, Tmp17, Tmp18, Tmp21, Tmp24	591,204,438,714	5,593,714,170,508	6,184,918,609,222
Two-Phase Optimization	Tmp5, Tmp11, Tmp12, Tmp15, Tmp17, Tmp18, Tmp21, Tmp24	591,203,688,714	5,593,714,470,508	6,184,918,159,222

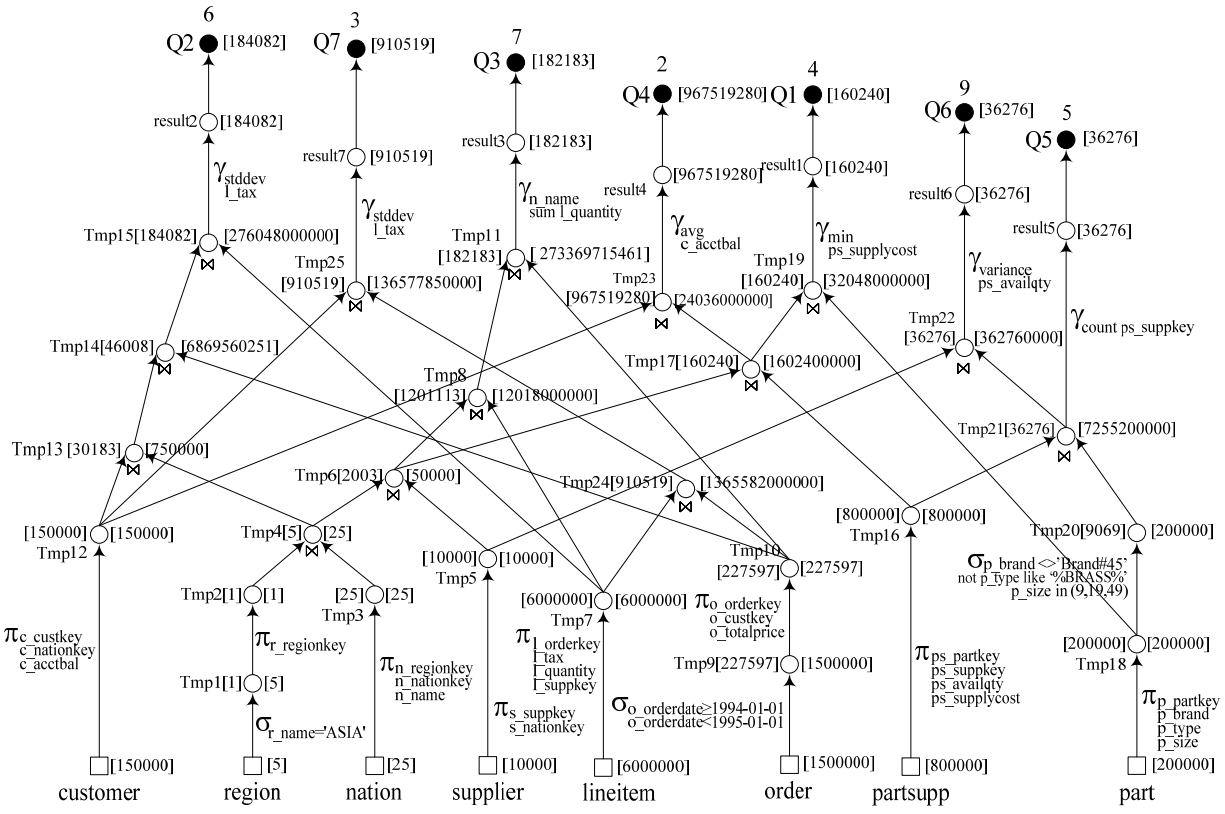

Fig. 7 An MVPP

REFERENCES

[1] W.H.Inmon, Building the Data Warehouse, John Wiley and Sons, 2002.
[2] J. Widom, "Research Problems in Data Warehousing," Int. Conf. on Information and Knowledge Management, 1995, 25-30.
[3] C. Zhang, X. Yao, and J. Yang, "An Evolutionary Approach to Materialized Views Selection in a Data Warehouse Environment." IEEE, 2001,31, 282-294.
[4] J. Phuboon-ob, and R. Auepanwiriyakul, "Analysis and Comparison of Algorithm for Selecting Materialized Views in a Data Warehousing Environment" APDSI, 2006, 392-395.
[5] J. Yang, K. Karlapalem, and Q. Li, "Algorithms for Materialized View Design in Data Warehousing Environment," VLDB Conference, 1997, 136-145.
[6] Y. E. Ioannidis and Y. C. Kang, "Randomized Algorithm For Optimizating Large Join Queries, " *ACM SIGMOD*, 1990, 312-321.
[7] S. Nahar, S. Sahni, and E. Shragowitz "Simulated Annealing and Combinatorial Optimization," *Design Automation Conference*, 1986, 293-299.
[8] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by Simulated Annealing," Science, 1983 ,671-680.
[9] Y. Ioannidis, and E. Wong, "Query optimization by simulated annealing," *ACM SIGMOD*, 1987, 2-22.
[10] R. Derakhshan, F. Dehne, O. Korn and B. Stantic, "Simulated Annealing for Materialized View Seletion in Data Warehousing Environment," DBA, 2006, 89-94.
[11] D. Stiles, "A 'C' Robust Simulated Annealing Package". Available: http://www.engineering.usu.edu/ece/research/rtpc/projects/comb/robust, 2006.

Jiratta Phuboon-ob received the B.Sc. (Hons.) degree in Statistics from Srinakharinwirot University, Mahasarakham Campus, Thailand in 1992 and the M.S. degree in Applied Statistics from School of Applied Statistics, National Institute of Development Administration (NIDA), Bangkok, Thailand in 1997.

She is currently pursuing the Ph.D. degree in Computer Science at NIDA, Bangkok, Thailand. Her research interests include databases, data warehouse and data mining.

Raweewan Auepanwiriyakul received the B.Sc. degree in Radiological Technology from Mahidol University, Thailand, in 1982 and the M.S. and Ph.D. degree in Computer Science from University of North Texas, U.S.A., in 1985 and 1989, respectively.

Currently she is an Assistant Professor with School of Applied Statistics, National Institute of Development Administration (NIDA), Bangkok, Thailand. Her research interests include databases, object-oriented analysis and design, and data warehouse.