DESIGN AND SELECTION OF MATERIALIZED VIEWS IN A DATA WAREHOUSING ENVIRONMENT

MR. G. JAGAN NAIK[1], DR. A.GOVARDHAN[2], DR. P C RAO[3]

[1]Ph.D Scholar, Department of Computer Science & Engineering, JNTU-Hyderabad, Telangana-500085, gjnaik1106@gmail.com

[2]Professor, Department of Computer Science & Engineering, JNTU-Hyderabad, Telangana-500085, govardhan_cse@jntuh.ac.in

[3]Professor, Department of Computer Science & Engineering, IARE-Hyderabad, Telangana-500043.

ABSTRACT: In this script, we depict the arrangement of a data warehousing system for a planning association 'R'. This system plans to help customers in recuperating data for business examination beneficially. The helper diagram of this data warehousing system uses the dimensional showing thoughts of star and snowflake designs. In addition, frequently got to estimation keys and characteristics are secured in various summary sees (showed up observes) remembering the ultimate objective to confine the inquiry planning cost. A cost demonstrates was made to engage the evaluation of the total cost and preferred standpoint related with picking each rose view. Using the cost examination framework for appraisal, a balanced insatiable estimation has been realized for the assurance of developed sees. This computation considers most of the cost factors related with the showed up observes decision method, including request get to frequencies, base-data invigorate frequencies, question get to costs, see upkeep costs and the availability of the structure's accumulating. The figuring and cost show have been associated with a game plan of authentic database things removed from association 'R'. By picking the most fiscally keen arrangement of showed up blueprint sees, the total cost of the help, accumulating and request treatment of the system is enhanced, thusly achieving a compelling data warehousing structure

Key Terms: Materialized View, Query Processing Cost, Dimension Keys.

I. INTRODUCTION

A data circulation focus is an information base that stores a huge volume of isolated and delineated data for On-Line Analytical Processing and Decision Support Systems [1]. The basic designing of a data warehousing structure given in [2] is showed up in Figure 1. To diminish the cost of executing all out request in a data warehousing condition, as regularly as conceivable used sums are as often as possible pre-considered and showed up along with framework sees so future inquiries can utilize them clearly. Indeed, developing these outline points of view can confine request response time. In any case, if the source data changes as regularly as would be prudent, keeping these developed sees invigorated will achieve a high help cost. In addition, for a structure with confined storage space or possibly with countless sees, we may have the ability to rise only a little division of the points of view. As needs be, different parameters, including customers' request frequencies, base association revive frequencies, question costs, see bolster costs and the availability of the structure's accumulating, should be viewed as to pick a perfect game plan of abstract points of view to be developed.

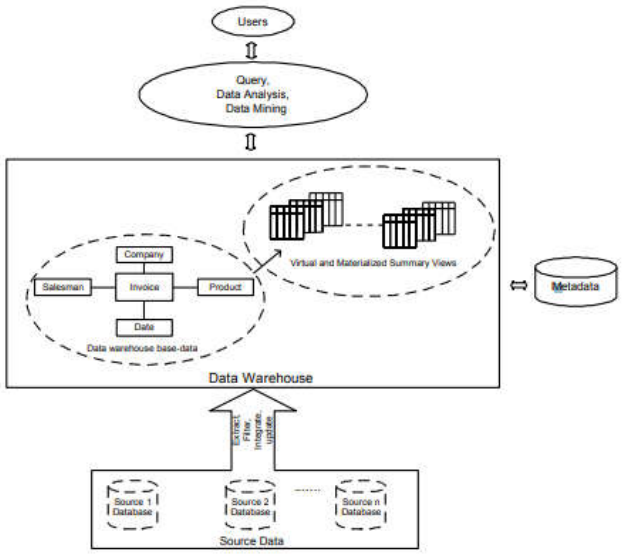

Fig 1: architecture for data warehousing system

To goad the trading of data appropriation focus plan and rose see decision, consider a data stockroom which contains the going with assurance and estimation tables: INV (Co_no, Inv_no, Inv_date, P_no, Qty, Amt) CO (Co_no, Co_name, R_no) PD (P_no, P_name, Mfr_no, Type_no, Cat_no).

Expect the sizes of the truth and estimation tables 'INV', 'CO' and 'PD' are 114B, 12B and 6B, independently, where B demonstrates the data square size which is 2K in the database structure (e.g., Oracle). Given a subset of ordinary customer's request [3] and the inquiry repeat between each revive time interval. By then we can figure the total cost Ctotal and each cost part

(i.e. request planning, upkeep and limit costs) for the going with three view appearance frameworks: • the each virtual-see method • the all-rose sees procedure • the picked showed up observes strategy Table 1 shows the calculation results, from which we specify the going with target actualities: The each virtual-see system requires the most shocking inquiry taking care of expense yet no view support and limit costs are vital. The all-showed up observes methodology can give the best request execution since this system requires the base inquiry getting ready expense. In any case, its total upkeep and limit costs are the most hoisted. The picked materialized views method requires a fairly higher inquiry getting ready expense than the all-showed up observes procedure, anyway its total cost Ctotal is the smallest.

	Total query processing cost $Total(C_{qp})$	Total maintenance cost $Total(C_{mt})$	Total storage Cost $Total(C_{storeT})$	C_{total}= $Total(C_{qp})$ $Total(C_{mT})$ $Total(C_{store})$
All-virtual-views	10920	0	0	10920
All-materialized-views	949	2829	709	4487
Selected-materialized-views	1200	2184	240	3624

Table 1: The query, maintenance and storage costs for three view materialization strategies.

In light of the above cost examination, plainly, the selected materialized-sees technique is the best with respect to both inquiry execution and bolsters cost of data warehousing systems. Starting late, showed up observe decision issue has begun intense talk in the database look at system. Harinarayan, Rajaraman and Ullman [4] showed a ravenous estimation for the decision of showed up observes with the objective that inquiry appraisal costs can be streamlined in the excellent example of "data 3D squares". Regardless, the costs for view support and limit were not tended to in this bit of work. Yang, Karlapalem and Li [5] proposed a heuristic figuring which utilizes a Multiple View Processing Plan (MVPP) to get a perfect developed see decision, with the ultimate objective that the best blend of good execution and low upkeep cost can be expert. Nevertheless, this figuring did not consider the structure storing necessities. Gupta [6] also developed a greedy computation to join the upkeep cost and limit prerequisite in the decision of data stockroom rose sees. "And also" see charts were familiar with address all the possible ways to deal with make circulation focus points of view to such a degree, to the point

that the best inquiry way can be utilized to improve request response time. In this paper, we discuss our experiences in delineating and picking fitting showed up observes for data warehousing systems. For our circumstance consider, the fundamental blueprint of this data warehousing system uses the dimensional exhibiting thoughts of star and snowflake plots as presented in [3]. The insatiable figuring presented by Gupta [6] has been grasped and adjusted for the decision of showed up observes. A cost show was delivered to enable the appraisal of the total costs and focal points drew in with picking each showed up view. We associated the computation and cost model to a game plan of real database things isolated from this association. In perspective of the cost examination, a game plan of showed up observes are propelled the total cost (i.e. the request, support and limit costs), with the objective that the best blend of good execution and low upkeep cost can be refined. Diverse view rise methodology are poor down and their shows are attempted [7]. Whatever is left of the paper is created as seeks after. The expense show and balanced avaricious figuring for the decision of developed sees are presented in portion 2. Guidelines for the blueprint and assurance of rose sees for data warehousing systems are discussed in region 3. Zone 4 completes the paper with a compact discourse of future work.

II. TECHNIQUES IMPLEMENTED

Materialized Views Selection: We as of now continue forward to address the related issue of data appropriation focus plan for our relevant examination, to be particular, the decision of layout points of view to be secured in the data stockroom. Points of interest of developing framework sees particularly have been clarified in the composition [6, 8]. For our relevant examination, a cost demonstrate is developed to enable the evaluation of request cost, upkeep cost, storing cost and focal points (i.e. assets in all around question costs) related with seeming each rundown find in the data dissemination focus. A balanced anxious estimation using the cost examination strategy for appraisal is then shown for picking a perfect game plan of showed up observes.

Cost model: The evaluated inquiry, upkeep and capacity costs in the accompanying depictions will be ascertained as far as information square size B. For effortlessness, different factors, for example, the computational expense and

correspondence cost are overlooked in our estimation. The point by point clarification of the cost figuring is displayed.

Query processing cost for selection, aggregation and joining: The examination expect that there is no file or hash enter in any of the rundown sees, thusly direct hunt and settled circle approach are utilized for the choice and join tasks, separately. The aggregate inquiry cost Total(Cqr) for preparing r client's inquiries between each refresh time interim is

$$Total(Cqr = \sum_{i=1}^{r} \text{fqi} * \text{Cq(qi)}$$

Information distribution center upkeep cost Assume that re-calculation of every rundown see Vi requires determination, collection and joining of its precursor see Vai with n measurement tables. On the off chance that there are j synopsis sees in the distribution center which are appeared, the aggregate upkeep cost 'Total(Cm)' for these emerged sees is at that point

$$Total(Cm) = \sum_{i=1}^{j} \text{fui} * \text{Cm(vi)}$$

(f_{ui} = 1 in our case study, since we assume that all sales summary views are updated once within a fixed time interval.)

Adapted greedy algorithm for materialized summary view selection: Let T be the set of all sales summary views grouped by various dimension key attributes. Based on the greedy algorithm of [6], we develop an adapted greedy algorithm for determining the optimal set of materialized summary views L, a subset of T, such that the total cost Ctotal is minimized. The algorithm is based on the cost model.

Materialized views selection algorithm:

1. Determine the optimum query and maintenance paths for computing all summary views in the data warehouse;

2. Calculate the Net(Bi) and hi of each summary view in the query paths. Let T be the number of summary views possibly chosen as materialized views.

for i =1 to T do Calculate the Net(Bi) of each summary view

Vi:

Storage effectiveness of summary views:

hi = Net(Bi) / S(Vi);

3. List summary views in descending order according to the value of their storage effectiveness such that those views with the best storage effectiveness will be chosen first;

4. Calculate the Ctotal for each view:

i = 1;

Ctotal = Total(Cqall) - Net(Bi); for i = 2 to T do Ctotal = Ctotal - Net(Bi); find the Min(Ctotal) as the optimal cost for materialized view selection;

5. Select the best materialized view set L

i = 1; Ctotal = Total(Cqall) - Net(Bi);

while Ctotal > Min(Ctotal)

i = i + 1;

while S(L) < S Select Vi from the summary view set TL with the highest storage effectiveness;

S(L) = S(L) + S(Vi);

endwhile

Ctotal = Ctotal - Net(Bi);

endwhile

return L.

Cost analysis: The rundown perspectives to be emerged are arranged in plunging request as indicated by the comparing stockpiling adequacy 'hey' recorded. The main thirty-four synopsis sees recorded in this table are the arrangement of ideal appeared sees L. The aggregate cost Ctotal, and its cost segments versus capacity size of the appeared sees are plotted in Figure 2.

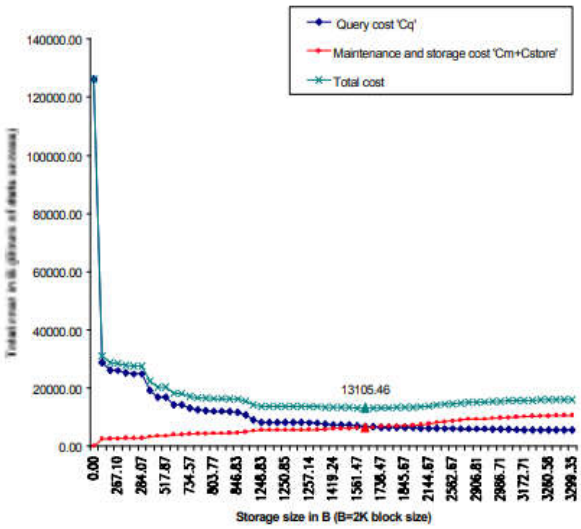

Figure 2: Total costs Ctotal, total query processing cost and the sum of maintenance and storage costs vs. storage size of the materialized views.

We see that the Ctotal is ruled by the Total(Cqr) before achieving the ideal point. This ideal point happens at an expense of 13105.46B and is assigned as the base aggregate cost Min(Ctotal). The Total(Cqr) drops definitely in the wake of appearing the

main rundown see 'CO-P-DAY', diminishing by over 75% while using just 15% of the aggregate storage room required by the arrangement of ideal emerged sees L. In this way, emerging outline see 'CO-P-DAY' is exceptionally savvy for enhancing the inquiry execution of the information stockroom.

After this first view has been picked, there is little decrease in the Total(Cqr) when more rundown sees are appeared. The total of aggregate support and capacity costs,

Cm (Vi)+Cstore(Vi), increments directly as the quantity of emerged rundown sees increments. In any case, its greatness is generally little contrasted and the Total(Cqr) before achieving the ideal point Min(Ctotal). In the wake of achieving this ideal point, Ctotal is ruled by the total Cm (Vi)+Cstore(Vi). This is on the grounds that emerging extra synopsis sees (i.e. outline sees with negative net advantage Net(Bi)) past the ideal point Min(Ctotal) can't diminish question cost, however expands the capacity and support costs. In this way, it isn't financially savvy to appear extra perspectives in the wake of achieving Min(Ctotal). In the event that all the rundown perspectives of the information distribution center are appeared, inquiry execution can be enhanced. In any case, this strategy requires the most noteworthy upkeep and capacity cost. For an information distribution center with restricted hard plate storage room and little upkeep window, emerging a couple of outline sees which have the best stockpiling adequacy greetings (i.e. 'CO-P-DAY' for this contextual analysis) can successfully diminish inquiry reaction time since they yield the best advantage yet require minimal measure of storage room and support costs. In the circumstance of an information distribution center which can be taken disconnected for view support and can have huge circle space accessible for the capacity of emerged sees, putting away the arrangement of ideal appeared sees L can limit inquiry and upkeep cost while accomplishing great question execution.

III. GUIDELINES FOR WAREHOUSE SCHEMA DESIGN AND MATERIALIZED VIEWS SELECTION

Our encounters picked up from this contextual analysis can be outlined into the accompanying rules for the two information distribution center plan and emerged see determination. On Data Warehouse Design. Utilize the littlest size of number or numerical qualities for the key ascribes in measurement tables to limit storage room and question handling time. Standardize measurement tables with huge measure of records and chain of importance levels to accomplish littler measurement tables. Consequently, the capacity size and joining cost can be lessened considerably Denormalize measurement tables with generally few records and credits to limit the quantity of joins required. On a level plane parcel the reality table, which has a ton of records, into littler outline sees as indicated by its measurement key credits to enhance inquiry execution, and further empower clients to choose different synopsis sees for appearance in light of the question get to recurrence. Store outside keys of measurement tables in the synopsis sees, particularly those measurement tables that are oftentimes gotten to help enhance the inquiry execution. Moreover, information in these synopsis perspectives can likewise be effortlessly utilized by different questions. Store much of the time got to measurement properties (e.g. Co_name and P_name for our situation contemplate) in the outline sees, particularly for the measurement tables which have a lot of records, to limit the quantity of joins and question preparing costs.

On Materialized Views Selection: Appear outline views that are oftentimes gotten to by clients. Emerge those generally shared perspectives which are utilized for producing other rundown sees. Emerge those perspectives whose sizes have been significantly diminished from their predecessor's perspectives. At the point when the capacity factor is little (i.e. a lot of circle stockpiling is accessible), emerging an arrangement of ideal appeared sees 'L' by the determination strategy as showed in Section above can accomplish the best mix of good question execution and low support cost.

IV. CONCLUSION

For this circumstance consider, strategies for sketching out a profitable data warehousing structure in perspective of the application necessities of a building association 'R' have been investigated. A mutt development was proposed for this data dissemination focus by applying dimensional showing thoughts. A cost show was created to register the costs and focal points related with seeming each

datum stockroom see. The total cost under five test conditions, made out of different request precedents and frequencies, were surveyed for three unmistakable view appearance procedures: 1) each and every virtual-see system, 2) all-developed sees strategy, and 3) picked rose sees method. The total cost evaluated from using the picked developed sees system was ended up being the humblest among the three approachs in all cases. Further, an examination was coordinated to record particular execution times of the three systems in the estimation of a settled number of inquiries and upkeep frames. Yet again, the picked rose sees methodology requires the most restricted total taking care of time. A balanced energetic estimation using the cost examination system for evaluation was delivered for developed sees assurance. This view decision logic was attempted both deductively and likely and wound up being especially pragmatic for the change of the data conveyance focus. General standards for data stockroom plan and showed up observes decision in perspective of this work are displayed and a model of the data appropriation focus structure was executed using a monetarily open data warehousing programming "Prophet Discoverer". The cost appraisal method and points of view decision figuring made for this circumstance study will be associated in the execution of other data warehousing applications, for instance, stock, age and purchasing examinations, et cetera. In like manner, appropriation focus see self-bolster methodologies other than the view re-estimation methodology gotten by this work will moreover be investigated, to also diminish structure upkeep cost and achieve data stockroom change.

V. REFERENCES

[1] J.Yang, K. Karlapalem, and Q. Li. “A framework for designing materialized views in data warehousing environment”. Technical Report HKUST-cs96-35, 1996. IEEE Int’l conference on Distributed Computing Systems (ICDCS ‘97), Maryland, U.S.A., May 1997.

[2] H. Gupta. “Selection of Views to Materialize in a Data Warehouse”. Proceedings of 23rd VLDB Conference, Athens, Greece 1997.

[3] G.Chan. A case study for the design and selection of materialized views in a data warehousing environment. MSc Dissertation, The Hong Kong Polytechnic University, Hong Kong, 1998.

[4] J.Yang, K. Karlapalem, and Q. Li. "Algorithms for Materialized View Design in Data Warehousing Environment". Proceedings of 23rd VLDB Conference, (Athens), Greece 1997, P.136-145.

[4] Oracle Discoverer 3.0 User's Guide, Oracle. [10] Oracle Discoverer 3.0 Administration Guide, Oracle

[5] S. Chaudhuri and U. Dayal. "An Overview of Data Warehousing and OLAP Technology". SIGMOD Record, 26(1):65-74, 1997.

[7] J. Hammer, H. Garcia-Molina, J. Widom, W. Labio, Y. Zhuge. "TheStanford Data Warehousing Project". IEEE Data Engineering Bulletin, June 1995.

[8] G.Chan, Qing Li, Ling Feng. Design and selection of materialized views in a data warehousing environment: A case study. 1999. Http://www.cs.cityu.edu.hk/~csqli/ papers /DOLAP99.ps.gz.

[9] V. Harinarayan, A. Rajaraman, and J. Ullman. "Implementing data cubes efficiently". Proceedings of ACM SIGMOD 1996 International Conference on Management of Data, Montreal, Canada, June 1996, pages 205--216.

[10] N. Huyn. "Efficient View Self-Maintenance". Proceeding of ACM Workshop, Montreal, Canada. 1996.