International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Humming Bird (01st March 2014)

RESEARCH ARTICLE OPEN ACCESS

Scheduling to Minimize Time Staleness and Execute Job using Max Benefit Algorithm

Simi M Cyril1,S Gayathri2

1Simi M Cyril. Author is currently pursuing M.E (Software Engineering) in Vins Christian College of Engineering. e-mail:simimcyril@gmail.com.

2S Gayathri.Asst. Professor of the Department of IT in VINS Christian College of Engineering.

Abstract:
Streaming warehouse is to propagate new data across all the tables and views as quickly as possible. The new data can be loaded,the applications and triggers defined on the warehouse can take immediate action. This helps business field, may lead to increased profit,improved customer satisfaction and prevention of serious problems. It solves the problem of non-preemtively scheduling updates in a real time streaming warehouse.The problem of this approach is that new data may arrive on multiple streams but there is no mechanism for limiting the number of tables that can be updated simultaneously.This may lead to CPU-Cache thrashing,disk-arm thrashing,context switching,etc.This can be motivated by scheduler algorithms that limits the number of concurrent updates, orders the job by their priorities and can break the tie with that data.It also determines which job to schedule next and also reduces the old data.

Keywords—Data Warehouse Maintenance, Data Staleness, Online Scheduling.

I. INTRODUCTION

Data Warehouses are updated during downtimes and stores layers of complex materialized views over terabytes ofhistorical data. It supports simple analysis on recently arrived data in real time. Streaming Warehouse is to propagate a new data.This enables a real-time decision support for business-critical applications that receive streams of append-only data from external sources.

Applications include :

Business field: may lead to increased profit, improved customer satisfaction,and prevention of serious problem.

Credit card or telephone fraud detection: calls are collected in nearly real-time and compared with past customer behavior.

Streaming warehouse is to propagate new data across all the relevant tables. Continuously someinformation can be collected from multiple sources. These jobs can be updated into tables at a particular period of time. It can be prioritize and delete the old data by prioritized EDF and Max Benefit algorithm. Streaming warehouses has focused on speeding up the Extract-Transform-Load (ETL) process. In[4], there has also been work on supporting various warehouse maintenance policies,such as immediate ,deferred and periodic.

The data arrive on multiple streams ,but there is no mechanism for limiting the number of tables that can be updated simultaneously. This may lead to some error like thrashing.This may motivated by scheduling algorithms that limits the number of updates and determines which job to schedule next..Update job is modeled as an atomic non-preemptive task.The purpose of an update scheduler is to decide which of the released update jobs to execute next.It cannot load a batch of data for the table that has not yet been loaded.

The transient overload, low priority jobs are deferred in favor of high priority jobs.When the period of transient overload is over, the low-priority jobs will be scheduled for execution.A solution to this problem is to “chop up” the execution of the jobs that have accumulated a long freshness delta to a maximum period times.This technique introduces a degree of preemptively into long jobs, reducing the chance of priority inversion[10].The recent work on streaming data warehousing, including system design [3], real time ETL processing, and continuously inserting a streaming data feed at bulk-load speed[8].Data Warehouse are typically refreshed in a a fashion: the updates from data sources are buffered during working hours, then loaded through the ETL process when the warehouse is overnight. The concept of recent works haveintroduced the novel issues. All update to produce systems area propagated immediately to the warehouse and incorporated in an on-line fashion.

Scheduling challenges that must be simultaneously dealt with a Streaming Warehouses. Data consistency and Transient overload are some challenges in scheduling. Streaming warehouses are inherently overloaded and a network problem will generally lead to a significantly increased volume of data flowing into the warehouse.The integration of data from multiple sources can be distributed to many modern information systems and business applications. Materialized views are one of the critical technologies applied in such data integration services that aim to store the integrated data to ensure efficient access ,reliable performance, and high availability.

II. RELATED WORKS

A. DATA STREAMS

Data stream is generated by an external source, with a batch of new data, consisting of one or more records, being pushed to the warehouse with period p. The period of a stream is unknown or unpredictable ,choose a period with which the warehouse should check for new data. A streaming warehouse maintains two types of tables:base and derived. Each table may be stored partially or wholly on disk. A base table is loaded directly from a data stream. A derived table is materialized view defined as an SQL query over one or more tables. When new data arrive on stream i ,an update job J is released whose purpose is to execute tasks,load new data into the corresponding base table T, update the indices.

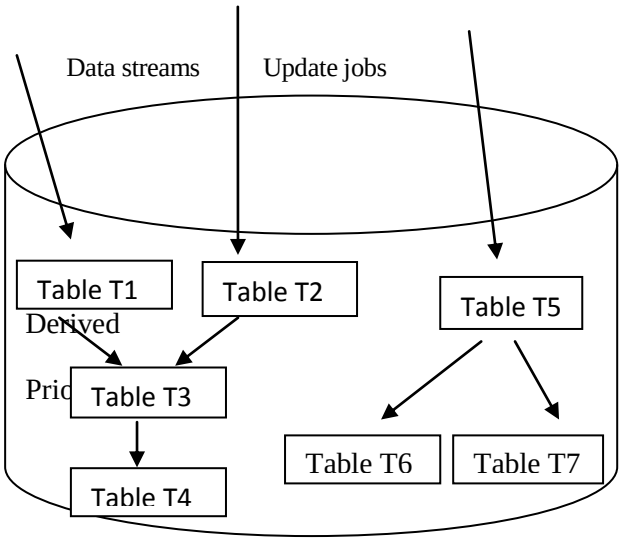

Fig: 1 A Streaming Data Warehouse.

Updated job is modeled as a atomic , non-pre-emptible task. The purpose of update scheduler is to decide which of the released update jobs to execute next. Update several tables together for instance the base table is a direct source of a set T.

B DATA STALENESS

The Data Staleness S_i (t), the staleness of table T_i at time t, to be the difference between t and the freshness of T_i. The staleness function of time for a base table T.

$$S_i(t) = t - F_i(t) \quad (1)$$

Staleness occurs linearly until the completion of the first update job at time t.Data staleness as a function time for a base table. T. The batch combines the process the new data can be arrives at a time.Fig.2 and Fig .3 shows the staleness of a base table T_i.

C SCHEDULING ALGORITHM

Atrack logically represent the fraction of the computing resources into the tracks.The basic scheduling algorithm prioritize jobs to be executed on individual tracks, and will serve as building blocks of our multitrack solutions.

(i)**FCFS**the track should have the finishing time is same. First arrives the table can be update first based on this algorithm.

(ii) **Prioritized EDF**orders the job$J_i$by their priorities .If some data can be break the ties by deadlines. The job define its release time r_i. Using scheduled algorithm it will prioritize jobsP_i.Deadline can be estimated by the eqn (2),

$$Ji = r_i + P_i \quad (2)$$

(iii) **Max Benefit** it can reduce the old data. It have a maximum efficiency.Using this algorithm can estimate which job is to be execute next.The benefit of executing a job J_i is its benefit of weighted freshnessdeltaΔF.Scheduler is to minimize the weighted staleness.Benefit of executing a job is calculated by eqn (3),

$J_i = P_i \Delta F_i$(3)Eqn (3) means, its priority weighted freshness data canbe executed. The qoal of the scheduler is to minimize the weighted staleness using the eqn(3).Marginal benefit of executing a job$J_i$is its benefit per unit of a execution timeE_i using a eqn (4),

$$J_i = p_i \Delta F_i / E_i(\Delta F_i) \quad (4)$$

Use max benefit can be periodic and aperiodic update job.Jobs are assumed to be periodic then the useful information about release times of future jobs.The weighted staleness which will orders the job as priorities. The useful information can be updated into the scheduled process.

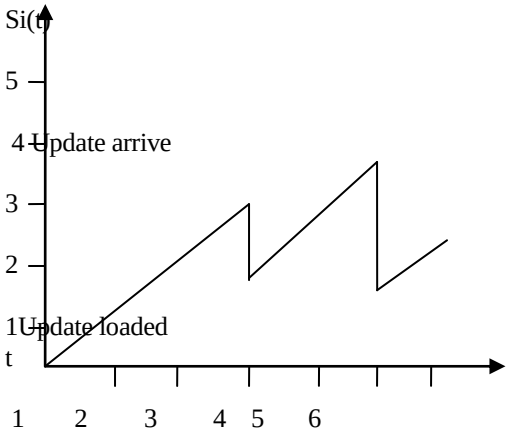

Fig 2: Staleness of a base table.

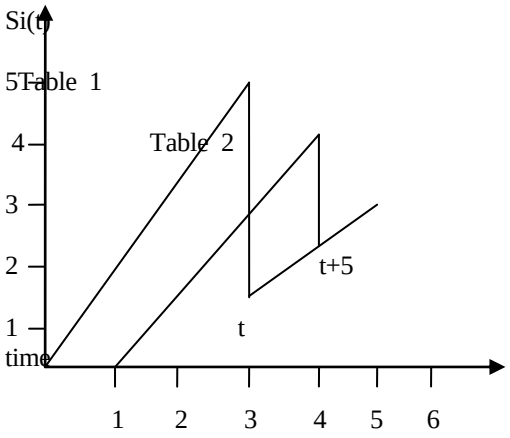

Fig :3 Staleness of two tables

III PROPOSED ALGORITHM

The proposed method presented hereuses scheduler algorithms that limits the number of updates and determines which job to schedule next. This helps to prioritize and delete the old data. Prioritized EDF Algorithm can orders the job by their priorities. It can prioritize only the single track simultaneously. So using the scheduler algorithmMax Benefit with lookahead can be choose which job is to execute next and prioritize multiple track simultaneously.It has maximum efficiency. B[S] represents the marginal benefit of running all the tasks in set S followed by a job can be estimated by an freshness delta by execution time $E_{k,}$

$$B[S]=\left(\sum (p_m \Delta F_m)+_{pk}\Delta F_k\right)\Big/\left(\sum(E_m) +E_k\right)$$

(4)Schedule the job with highest marginal benefit from set S. Main innovation is the multitrack proportional algorithm for scheduling the large and heterogeneous job sets encountered by a streaming warehouse.It can also monitor data loads, queries and reports, data archives and recoveries. Additionally an update chopping to deal with transient overload. It will shows the complexity of minimizing the staleness of a set of base tables in a streaming warehouse.

Derived table relationships are similar to precedence constraints among jobs, a derived table cannot be updated until its sources have been updated. The work closest to discuss the interaction of queries and updates in a firm real-time database, is to install updates to keep the data refresh but also ensure that read transactions meet their deadlines. The work has also shown in the web databases, which aims to balance the quality of service of read transactions into the old data can be deleted. It assumes a snapshot model rather than our append-only model .

AALGORITHM DESCRIPTION

Detailed steps of the algorithm are as following:

Step I: Sort the released jobs by the scheduling algorithm.

Step II: Jobs can be arrived by FCFS algorithm.

Step III:It can orders the job by their priorities. Update multiple track simultaneously.

Step IV: Max Benefit will reduce the old data.
Step V: Choose which job is to be execute next.

Step VI:Schedule the process at time.

Step VII: Data Warehouse monitoring.

IV. EXPERIMENTALRESULTS

The best way toschedule updates of tables and views in order to maximize data freshness.Max Benefit algorithm is to maximize the data freshness. The interesting scheduling algorithm problems can be hard to proximate the process. The base table jobs all have a priority of one and the derived tables have priority. Using update chopping to limit the amount of data loaded at the job period.Update chopping is an effective strategy. With the help of an algorithm it has no transient slowdowns.

No Slowdowns1
2
3
Utot

Fig :4 No slowdowns,uniform jobs, tracks

Line 1 shows the Max Benefit, 2 shows EDF, 3 shows the lookahead.It identifies there has no slowdowns. The relative lateness for uniform jobs with identical parameters.

Previous work on scheduling with precedence constraints focused on minimizing the total time needed to compute a set of nonrecurring jobs. When a derived table should be scheduled as a freshness delta to determine.It isaability to reserve resources for shot jobs that can be refresh the tables and monitor data warehouse.

V. CONCLUSION

This paper can be motivated, formalized and solved the problem of nonpreemptively scheduling updates in a real time streaming warehouse . To extend the framework with new basic algorithms.Also plan to fine-tune the proportional algorithm.The interesting problems for future work involves choosing the right scheduling "granularity" when it is more efficient to update multiple tables together. For more efficiency and the tables are updated a jobs in multiple track, another algorithm can be used.

REFERENCES

[1] NavdeepKanwal, AkshayGirdhar, "Supporting Streaming Updates in a active Data Warehouse," IEEE conf,2007.

[2] Hassan, N. Y. and Aakamatsu, N., "Contrast nhancement technique of dark blurred Image",International Journal of Software Engineering, Vol. 6, No. 2, 2006, pp. 223-226.

[3] 2.Chen, Soong-Der and Ramli, Abd. Rahman, "Stream warehousing with datadepot,"", IEEE Transactions on Consumer Electronics, Vol. 49, No. 4, 2003, pp. 1310-1319.

[4] N.Polesel, , G. Mathews, V.J. Tele Media Int. Ltd., Frankfurt, "Supporting Streaming Updates in an active Data warehouse", IEEE Transactions, Vol. 9, Issue. 3, 2002, pp. 505-510 .

[5] Zimmerman, J. B., Pizer, S. M., Staab, E.V., Perry, J. R., McCartney, W. and Brenton, B. C., "RItE: providing On Demand Data for Right-Time Data warehousing" t", IEEE Transaction on Software Eng,, Vol. 7, No. 4, 1988, pp. 304-312.

[6] Rangayan, R.M., "Tradiness Bounds under Global EDF Scheduling", Proc. ACMSIGMOD Int'l Conf. Management of Data , 2005, pp. 338, ISBN 0-8493-9695-6.

[7] Morrow, W.M., Paranjape, R.B., Rangayyan, R.M., Desautels, J.E.L.,"Scheduling Hard Real-Time Systems:AReview,",SoftwareEng.j. Vol. 11, No. 3, 1992, pp. 392-406.

[8] Rangayyan R.M., Alto H., and Gavrilov D., "RiTE:ProvidingOnDemand Data for Right-Time Data warehousing" Journal of Software Eng, 2001, 10:804813.

[9] Wei,Liyang, Kumar, Dinesh,Khemka, Animesh, Turlapti, Ram, Suri, Jasjit S. "An Over view of real-time database system ", Advance in Real-Time Systems,S.H. Son,ed.,Volume 6914, 2008, pp. 691426-691426-8 .