International Journal of Electronics, Communication & Soft Computing Science and Engineering
ISSN: 2277-9477, Volume 3, Issue 3

Data Staleness Minimization in Data Stream Warehouse

Archana R. Ugale Prof. Wankhade N.R

Abstract — **Data Stream warehouse is a combination of Data warehouse & Data stream management system which combines the feature of these two system together & maintains a joined view of current & historical data .The proposed system works for the scheduling of updation in streaming data warehouses by minimizing data staleness over time & maintains the consistency in all tables along with minimizing the different problems faced by a stream warehouse like view hierarchies ,priorities, data consistency, heterogeneity caused due to different inter-arrival time ,data volumes & transient overload.**

Key Words — **Data consistency, Data staleness, Stream warehouse ,Data Warehouse, Data consistency ,Data Mining**

I. INTRODUCTION

Data mining is the process of extraction of useful information from a large relational database by using various techniques. Data mining also known as knowledge discovery in which databases includes nontrivial extraction of implicit previously unknown & useful information from data.. A data warehouse is a relational database, a central repository of data which stores historical data derived & current data from various transactional & real time applications. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. In addition to it includes an extraction, transformation, and loading (ETL) solution, an online analytical processing (OLAP) , data analysis tools, and many applications that manage the process of gathering data and delivering it to business users. Traditional data warehouses are updated during downtimes & maintain the data in various views which concerns with some specific applications .Data warehouses specially works on the historical data rather than current data. Apart from this Data Stream Management System support simple analyses on recent arrived data in real time. Data stream warehouses combine the features of these two systems by maintaining a combined view of current & historical data. It enables the real time decision support system to work for various business critical applications which leads to increased profits, customer satisfaction & prevention of serious problems that can occur. It can be applicable to various applications like:

1.Online stock trading
2.Credit card or telephone fraud detection
3.Network data warehouses maintained by Internet Service Providers
4.Financial Organization
5.Scientific Research
6. Government Organization
7. Military and Defense
8. Disaster Management
9. Weather Forecasting

The goal of a streaming warehouse is to reflect the new data in all relevant tables & views as earliest. On the arrival of new data the applications, views& triggers defined on the data warehouse can take quick action. Currently streaming warehouses has focused on speeding up the Extract-Transform-Load process. It also work on supporting various warehouse maintains policies like immediate, deferred & periodic. It works on the mechanism which data should be updated next on arrival of new data & number of tables can be updated at a time. Scheduling of real time applications in stream data warehouses faces the various scheduling metrics like hard real time system which restricts to complete job within deadline. In a firm real time system jobs can miss their deadlines & if they do they are discarded .However ,a streaming warehouse must load all the data arrived in system therefore no updates can be discarded In a soft real time system ,late jobs are allowed to stay in the system and the performance metric is lateness which can be defined as difference between the completion times of late jobs & their deadlines. Rather than concerned about properties of update jobs it works on the Data Staleness. Data Staleness can be defined as the difference between the current time & time stamp of the most recent record[1].

II. DATA DESIGN ISSUES IN DATA STREAM WAREHOUSE

Data stream warehouse faces the various scheduling challenges [4] like hard real time system where the jobs must be completed before their deadlines. In a firm real time system jobs get discarded if they miss their deadlines .so the performance ratio in this system is low .In a soft real time late jobs are allowed to stay in the system & performance metric is lateness which calculate the difference between completion time of late jobs & their deadlines.

A. Data Consistency

To ensure that each view reflects a "consistent "state of its base data [5], [6] though different base tables are scheduled for updates at different times. Current work on streaming

warehouse has centered to maintain the data's freshness & speed up the Extract-transform-Load process [7],[8].

The work is going on supporting various warehouse maintenance policies such as immediate where views updates whenever the base data changes. Deferred update views generates only when queried & periodic which updates data periodically. However the main issue is choosing the table that are now out of date due to the arrival of new data & which should be updated next[1].It requires a scheduler that limits the number of simultaneous jobs update & determine which jobs to schedule next to update.

.

B. Heteroginity & Nonpreemtbility

The different inter-arrival times & data volume varies according to different streams. This makes kind of heterogeneity make real time scheduling difficult[1].

III. SCHEDULING APPROACHES

There are number of systems worked for Data stream warehouses in different ways. Various scheduling algorithms are used for loading data feeds in to real time data warehouses which can be used in applications such as IP network monitoring, online financial trading credit card fraud detection system etc. Lukasz Golab, Mohammad Bateni,Howard Karloff, Mohammad TaghiHajiaghyayi [3]suggested the algorithm to schedule the updates on one or more processors in a way that minimizes the total staleness .To limit the maximum stretch between the updation time length of updation.They proved that any on line no preemptive algorithm that is never idle achieves a constant competitive ratio with respect to the total staleness of all tables in the warehouses, provided that the processors are sufficiently fast. Brian Babcock, ShivnathBabu MayurDatar Rajeev Motwani[9] worked over fluctuates data rate of burst data warehouses. They suggested the strategy for processing large volume streams ,overload scheduling of queries to minimize resource utilization during times of peak load. They present Chain scheduling for data stream system that is near optimal in minimizing run time memory usage for any collection of single stream queries involving selections, projections, foreign key joins with stored relations. Urmila,K.SivaRamaKrishna, P.RajaPra ashRao[10] The problem of scheduling updates in a real time streaming warehouse. They presented the notation of averages staleness as a scheduling metric presented scheduling algorithms designed to handle complex environment of a streaming data warehouse. Though number of techniques used by various people on variety of applications but mostly all they worked on the properties of tables but still the problems is pending related to maintain Data consistency, priorities, Number of table's updates at a time etc. The existing systems are not fully efficient to maintain data warehouses. The main problem exist with this mechanism is that new data may arrive on multiple streams, but there is no technique to limit the numbers of tables that can be updated concurrently.

IV. PROPOSED WORK

A. System Architecture

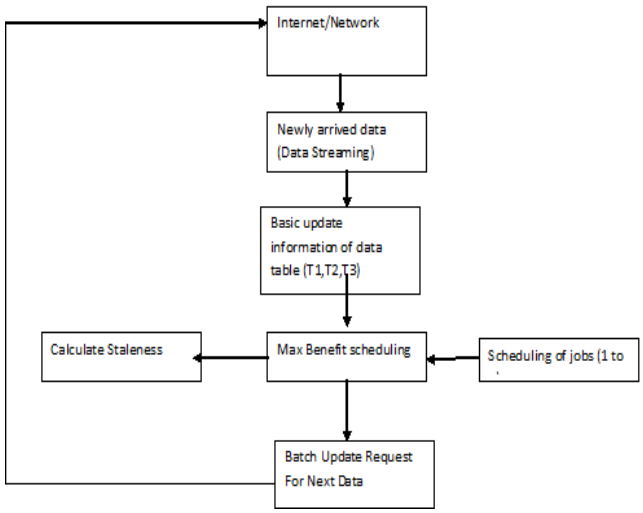

Fig 1 :System Architecture

B. Mathematical Model

Scalable scheduling updates in streaming data ware house approach illustrate mathematical model that contain scheduling of updation in streaming data warehouses where the main purpose is of job concern with newly arrived data in tables and aims to minimize data staleness accompanied by minimizing various problems associated by stream warehouse.

F1 GU:-GENERATE UPDATE REQUEST.
F2 CS:-CALCULATE STALENESS.
F3 CD:-CALCULATE DELAY VALUE.
F4 AR:-ACCUMULATED UPDATE REQUEST.
F5 BA:-BATCH UPDATE REQUEST.
F6 ST:-SCHEDULE TABLE.
S={RD,GU,CS,CD,AR,BA,ST}
Let ,RD:{rd1,rd2,…..rdn} where RD consists of receive data.
Let, GU:{gu1,gu2,….gun} where GU consists of contain generated update request.
Let ,CS:{cs1,cs2,…csn} where CS contain calculate staleness values.
Let, CD:{cd1,cd2,….cdn} where CD consist of calculate delay value.
Let, AR:{ar1,ar2,…arn} where AR contain updated request.
Let, BA:{ba1,ba2,….ban} where BA consist batch updated request.

Let, ST:{st1,st2,….stn} where ST consists of final scheduled request.

Function F1 : return the generated updated request F1(RD)->GU FOR GENERATING UPDATE. EG.F1(RD)->{gu1,gu2,…gun} FOR GENERATING REQUEST.
Function F2: return the calculate staleness. F2(GU)->CS FOR CALCULATE STALENESS. EG,F2(GU)->{cs1,cs2,…csn} belongs to CS.
Function F3: returns delay value F3(CS)->CD EG.F3(CS)->{cd1,cd2,….cdn} belongs to CD.
Function F4: return Augmented request. F4(CD)->AR EG.F4(CD)->{ar1,ar2,….arn} belongs to AR.
Function F5: returns batch update request F5(AR)->BA EG.F5(AR)->{ba1,ba2,…ban} belongs to BA.
Function F6: return final updated table request. F6(BA)->ST EG.F6(BA)->{st1,st2,…….stn} belongs to ST.

Table I :Functional dependency table of functions in a system

Functions	F1	F2	F3	F4	F5	F6
F1	1	0	0	0	0	0
F2	1	1	0	0	0	0
F3	0	1	1	0	0	0
F4	0	0	1	1	0	0
F5	0	0	0	1	1	0
F6	0	0	0	0	1	1

C. Experimental Setup:

Proposed system accepts live data streaming from "twitter" website & fetches tweets arriving at different rates which forms a master table. Derived table are generated from base table. Jobs/Process are scheduled on derived tables .Jobs are independent but synchronized over table access named as reader & writer processes or jobs & varying ratios of these jobs can be applied for table update .Window size allows derived table to be updated as batch update.

V. RESULT & DISCUSSION

Proposed system is tested and compared with greedy heuristic method of scheduling and Fig 2 shows improving performance by minimizing staleness of data update. Fig 3 indicates no of jobs versus staleness of data update on varying batch size updates. Relative lateness on varying window size is observed to indicate improved performance of proposed method in plot Fig 4.

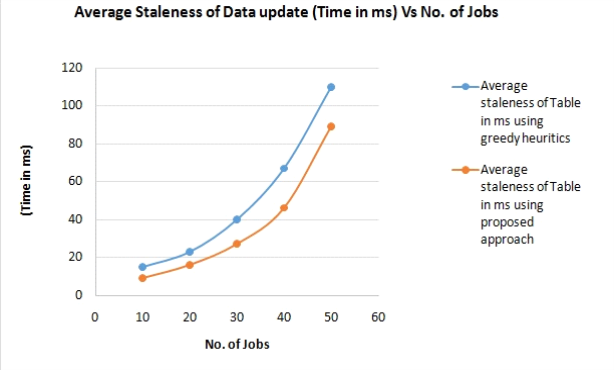

Fig 2: Average Staleness of data update vs. No. of jobs

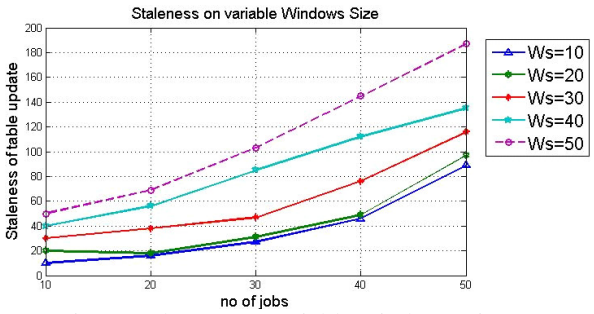

Fig 3 :Staleness on variable windows Size

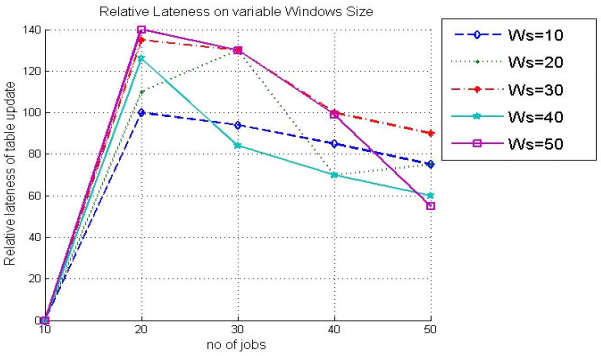

Fig 4: Relative Lateness on variable windows size

CONCLUSION

Thus the presented system minimizes data staleness by optimal job scheduling to obtain data freshness & maintains the consistency of data .The main feature of this implementation is efficient resource utilization which efficiently work with large volumes of data & enables the real time system to work for various business application.

REFERENCES

[1] Lukasz Golab,Theodore Johnson & Vladisav Shkapenyuk:Scalable scheduling Of updates In Streaming Data Warehouses,IEEE transaction on knowledge & Data engineering ,vol 24.no 6,jun 2012

[2] L. Golab, T. Johnson, J.S. Seidel, and V. Shkapenyuk, "Stream Warehousing with Datadepot," Proc. 35th ACM SIGMOD Int'lConf. Management of Data, pp. 847-854, 2009.

[3] MohammadHossien Bateni,Lukasz Golab,Howard Karloff, MohammadTaghi Hajiaghyayi:Scheduling to minimize Staleness & Stretch in Real Time Data Warehouses ACM 2009.

[4] A. Burns, "Scheduling Hard Real-Time Systems: A Review,"Software Eng. J., vol. 6, no. 3, pp. 116-128, 1991.

[5] L. Colby, A. Kawaguchi, D. Lieuwen, I. Mumick, and K. Ross,"Supporting Multiple View Maintenance Policies," Proc. ACMSIGMOD Int'l Conf. Management of Data, pp. 405-416, 1997.

[6] Y. Zhuge, J. Wiener, and H. Garcia-Molina, "Multiple ViewConsistency for Data Warehousing," Proc. IEEE 13th Int'l Conf.Data Eng. (ICDE), pp. 289-300, 1997.

[7] N. Polyzotis, S. Skiadopoulos, P. Vassiliadis, A. Simitsis, and N.-E.Frantzell, "Supporting Streaming Updates in an Active DataWarehouse," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 476-485, 2007.

[8] C. Thomsen, T.B. Pedersen, and W. Lehner, "RiTE: Providing On-Demand Data for Right-Time Data Warehousing," Proc. IEEE 24thInt'l Conf. Data Eng. (ICDE), pp. 456-465, 2008.

[9] Brian Babcock,Shivnath Babu ,Mayur Datar & Rajeev Motwani Chain:Operator scheduling for memory minimization in Data Stream system,SIGMOD 2003,253-264.

[10] P.Urmila ,K.Siva RamaKrishna,P.Raja Prakash Rao :Scheduling Of Updates in Data warehouses ISSN 2230-9624 Vol 3,issue 3,2012 pp 362-367.

AUTHOR'S PROFILE

Archana R. Ugale is post graduate student of computer engineering at LGN sapkal college of engineering ,Nashik under University of Pune.She complete her undergraduate cource of engineering from Amravati University..Her areas of interest include Data mining and Data Warehouse.

Prof. Wankhade N.R completed his post-graduation from Bharti vidyapit ,Pune,Maharashtra .Presently he is working at LGN sapkal college of engineering ,Nashik,Maharashtra,India as a professor and head of computer engineering department .He has presented papers at National and International conferences and also published paper in national and international journals on various aspect of the computer engineering and networks.His research of interest include computer networks, network security, wireless sensor network.