Performance Tuning Mechanisms for Data Warehouse: Query cache

Vishal Gour
Department of Computer Application
Govt. Engineering College Bikaner
Bikaner

Dr. S.S.Sarangdevot
Department of Computer Science & Information Technology
J.R.N. Rajasthan Vidyapeeth University,
Udaipur (Rajasthan)

Govind Singh Tanwar
Department of Computer Science Engineering
Govt. Engineering College Bikaner
Bikaner

ABSTRACT:

Data warehousing encompasses architectures, algorithms, and tools for bringing together selected data from multiple databases or other information sources into a single repository, called a data warehouse suitable for direct querying or analysis. In recent years data warehousing has become a prominent buzzword in the database industry, but attention from the database research community has been limited. At the warehouse, queries can be answered and data analysis can be performed quickly and efficiently since the information is directly available, with model and semantic differences already resolved. The primary goal of data warehouse is to free the information locked up in the operational database so that decision makers and business analyst can make queries, analysis and planning regardless of the data changes in operational database. As the number of queries is large, therefore, in certain cases there is reasonable probability that same query submitted by the one or multiple users at different times. Each time when query is executed, all the data of warehouse is analyzed to generate the result of that query.

In this paper we will try to find the common problems faced. These kinds of problems are faced by Data Warehouse administrators which are minimizes response time and improves the efficiency of data warehouse overall, particularly when data warehouse is updated at regular interval. The overall performance of the system and provide good strategies that make superior Data Warehouse.

Keyword - Data warehouse; Operational database; Response time; performance;

1. INTRODUCTION

As data warehousing is an emerging area, a lot of problems are found in the system. One of the major problems faced by the industry today is data warehouse maintenance.

Data warehouses require a very high level of maintenance.[18] Any reorganization of the business processes and the source systems may require the data warehouse to change. Updates are often needed in these situations. Keeping in view this problem we have decided to do some research on maintenance of data warehouses. Experts say that more resources are required for maintenance of a data warehouse rather than its development [1]. Analysts use the data warehouse to answer unlimited variety of questions, which may be very difficult to answer in operational database. Data warehouse contains a number of databases regardless of the number of sources and volume of data. The resulted warehouse is more homogeneous compared to operational data repository [15]. The primary goal is creating strategic planning resulting from long tern data analysis. We can create reports, projection, and business model and can forecast by these analysis [16]

According to [2] there are many ways for a data warehouse project to fail. The project can be over budget, the schedule may slip, critical functions may not be implemented, the users could be unhappy and the performance may be unacceptable.

2. PERFORMANCE TUNING MECHANISMS

While the implementation of a specific phase of the data warehouse may be completed, but the data warehouse program needs to be continued [3]. Progress monitoring needed to be continued against the agreed-on success criteria. The data warehouse team must ensure that the existing implementations remain on track and continue to address the needs of business. Performance issues in data warehousing are centralized around access performance for running queries and incremental loading of snapshot changes from the source systems [4, 5, 6]. The following concepts can be considered for a better performance:

2.1 Communication and Training.

After the employment of data warehouse one needs to address the issues of establishing and maintaining ongoing communication and training, providing help support services, and managing technical infrastructure updates concerning updated versions of hardware, software and services [4]. Communication and training are two interrelated activities. A communication and training program is helpful in keeping the business community and IT community within the organization informed on the current and proposed future developments in the data team to measure progress and warehousing environment. A communication process also offers the data warehouse identify and resolve issues before they become a serious problem.[18] The communications program provides the business components with increased capabilities and functions. Training expands the communication process by maintaining a level competence in both the business and IT community as to the tools and mechanisms of the data warehouses. Training of data

warehouse users is significant and provides the desired output [7]. In most computing projects, management identifies the need for training, but does not always fund training. With every new database there is a need for another training course, complete with reference materials. Every enhancement or change to the warehouse must be documented and communicated to warehouse users. While training users is essential, it distracts from future warehouse development unless new resources are allocated. A continuous education and training program is always required for the data warehouse[3].

2.2. Help Desk and Problem Management

While training reduces the number of data warehouse questions, a support infrastructure is the key to handling other support needs. [7] In order to ensure success one needs to develop a support structure and plan an approach When people are using the system, the questions will flow [8]. If there are no questions than it is likely that no one is using the system. The question asked could be about the validity of the data itself, how to use the data, what calculations make sense and what levels of aggregation are valid, how to pick a report, how report measures were calculated, how to use the applications, how to change an application, and how to build an own application etc. User support is crucial immediately following the deployment in order to ensure that the business community gets hooked [3]. For the first several weeks following user education, the support team should be working proactively with the users. It can't sit back and assume that no news from the community is good news. If there is nothing heard from the business users it means that no one is using the data warehouse. In such a case the support professionals should turn up to the business community so that the users of the data warehouse have easy access to support resources. If problems with the data or applications are uncovered, immediately try to rectify the problems. [18]

2.3. Network Management

If there is a heterogeneous group of platforms for the data warehouse implementation, network management is going to be one of the most demanding tasks [6]. Not only are users coming constantly on-line, but users and equipment are invariably moving to new locations. The networking hardware is proliferating with LANs, WANs, hubs, routers, switches and multiplexers. Leaving behind all this is the next stage – users wanting to access internet based data sources along with the corporate data, requiring even greater bandwidth and network management resources. Managing this environment is one big challenge, capacity planning for the future is another. .[18] If the data warehouse team is not quite good in networking technology than there should be at least one person in the organization who understands technology.

2.4. Capacity Planning

Capacity planning refers to determining the required future configuration of hardware and software for a network, datacenter or web site. According to [9] capacity planning enables the determination of sufficient resources so that user satisfaction can be maximized through timely, efficient and accurate responses. Capacity planning is important when starting a new organization, extending the operations of an existing business, considering additions or modifications to product lines, and introducing new techniques, equipment and materials [9]. In business, capacity is the maximum rate of output for a process. This means that capacity is the work that the system is capable of doing in a given period of time. .[18] The goal of capacity planning is to meet current and future demand with a minimal amount of waste.

2.5. Data Loading Performance

To load a data warehouse, regular loading or propagation of data from operational systems is needed. Factors affecting data loading performance include [4]:

i. The decreasing batch window (time available for an intensive batch processing operation such as a disk backup) to load data from more and more sources, coupled with increased usage of the data warehouse.
ii. The frequency and size of these loads.
iii. The changing natures of these loads as source systems change or are replaced.
iv. The increasing demand for more metadata regarding the data to be loaded.
v. If load performance remains an issue, consider maintaining a synchronous replica of the full database to source data and to the warehouse.

All these problems mean that the data warehouse team must plan for and examine performance on an ongoing basis. .[18] Creating a performance management system that logs all relevant data or selecting a product that provides this functionality will arm you with a critical tool in the fight to keep on top of the growing data management problem.

2.6. Query Management

To have an efficient query management system most of the predefined and ad hoc queries should access summary data instead of detailed data [4]. One needs to employ query navigators to redirect base table queries to the aggregate and summary table level and examine which tables and columns were accessed and the number of rows retrieved. Check response time for these queries and any effects these have for e.g. paging or locking.

Break down the predefined queries into smaller queries for processing. Consider doing most of your resource intensive processing away from the current level of detail and try to run large queries during off-peak hours. This will give more processor time to smaller queries and will aid in getting quick results. . [18] Try to push query processing up from the client to the application server level. As part of end user workstation design, consider the employment of thin clients, forcing Query processing and scheduling up to the server level.

In a data warehouse data should be distributed in a hierarchical manner where the most common information has the least amount of distribution and the least common information has the highest level distribution. Minimize the amount of cross network data retrieval and combination of data from different locations. Allow information access users to follow a hierarchical path when searching for data.

2.6.1. Query cache

For the fast access the database we use the query cache. Query cache will store all record of executed query.Query cache will keep record of newly executed queries. The major goal of the query cache is to reduce the response time of query. It will increase the brainpower of data ware house so that system will memorize the latest work it has performed. This memory will be used afterward for answer the result of queries which has been earlier performed by the users. The cache will maintain two state valid and invalid state. When any query submitted by the user, the cache memory is first examined to check whether requested query is already store in the cache. If the query is stored, then check the state is valid or invalid. If state is valid then data can be access and if state is invalid then data can't be access. but If user send a query of insert, update, delete and drop then data will be alter in database and state of related query will be invalid. Now invalid state data and query can't be access by user. This can save important time and improve data warehouse performance by not reevaluating the queries which are already stored in the cache.

One of analyst place a query to show me the employee of a company, who working under the manager_id is 100,101,201. The query will look like as follows: -

SELECT emp_id, name, salary, manager_id

FROM employees

WHERE manager_id IN (100, 101, 201);

When the query is submitted, query cache will be examined to check whether this query is available or not and state is valid or invalid. If it is not available, query will be evaluated and result will be store in the query cache. The results of the query are shown in the table1 –

Table 1 – output of the above query

Emp_id	Name	Salary	Manager_id
202	Mukesh	6000	201
200	Mohan	4400	101
205	Sohan	12000	101
101	Rohit	17000	100
102	Sanjay	14000	100

If any other user submitted the same query the result will be retrieved from query cache because that query is already stored in the cache. We will call this Query1.

Let suppose another user wants to the employee of a company, whose salary greater than equal to 10000 **AND** manager_id is 100,101. The query will look like as follows:

SELECT emp_id, name, salary, manager_id

FROM employees

WHERE manager_id IN (100, 101) and Salary>=10000;

When the query is submitted, cache memory is examined. Same query is stored in the cache memory and state is valid then we can get the result of Query 2 as shown in Table 2.

Table 2 – output of the above query

Emp_id	Name	Salary	Manager_id
205	Sohan	12000	101
101	Rohit	17000	100
102	Sanjay	14000	100

Now result of Query 2 will be generated from the Query 1 result set instead of going through from all the data stored in the data warehouse. This process will save lot of time and effort required to go through all the records.

Queries against complex view definitions must be answered very fast because users engaged in decision support activities require fast and quick answers to their questions. Even with sophisticated optimization and evaluation techniques, there is a limit to how fast one can answer such queries. The main objective of a materialized view is to improve query performance [14].

However, when a warehouse is updated especially due to the changes of remote information sources, the materialized views must also be updated. While queries calling for up-to-date information are growing and the amount of data reflected to data warehouses has been increasing, the time window available for making the warehouse up-to-date has been shrinking. Hence, an efficient view maintenance strategy is one of the outstanding issues in the data warehouse environment. This can improve the performance of query processing by minimizing OLAP queries down time and interference.[13]

2.6.2. Query cache state

- **Invalid –**if query is not stored in query cache then state will be invalid. If data is updated by user by any these query insert, update delete the state will be invalid.

- **Valid –** if query is stored in query cache and not updated in database from any these query insert, update delete the state will be valid.

Our problem is that we have a query and query result stored in the cache. But if the warehouse is updated with the new data the cache query result will reflect to old data. We will create a mechanism of state;

Query 1 is submitted by the user and his result is stored in the query cache. When next user submit the same query on updated data warehouse the query cache will check the state if state is invalid, it means the data warehouse is updated with new data. Now the query doesn't have to go through from all of the records. It will get the last index of the query result stored in the query cache. Then it will start searching the records which meet the query criteria from onward to that index. This can save lot of time and effort required to search the large amount of data.

2.7. The Problem Management Process

Communication processes for data warehouse project and program support deals with issues like corresponding with internal and external groups of the data warehouse project, scheduling deliverable reviews and conducting meetings to get status of the project and the issues arising in the project, helping quality assurance representatives to manage quality and notifying the data warehouse group and others in the organization when project milestones are reached. The scope of the selected communication program identifies the people who should be contacted, the main messages to pass, and the type of communication and its frequency [4].

This can be achieved by giving a detailed and scheduled program of education and training for developing and supporting vision clarity for the data warehousing environment. The members of the team should establish a list of recommended technology components and standards that should be used. The team further reviews and approves the proposed approach, functional architecture and technology components and standards.

2.8. The Problem management Process Development

This process specifies how to collect, document, answer and/or escalate calls, requests, and queries related to issues with the data warehousing environment. Problem documentation can be completed either by the help desk representative and/or in conjunction with a form completed by the end user or IT support person requesting a service or action.

All inquiries, no matter how trivial should be logged, especially during the start of a new data warehouse or mart. These bits of information can form clues to taking proactive action to bigger problems before they emerge. Having a production ready data warehouse means support must be expedited in an efficient, responsive, and businesslike manner. At stake is the ability of the business to stay competitive if the business information the warehouse contains is not current, accurate, timely and available when needed.

This thought must be kept in mind by all help desk personnel as they strive to answer those nagging questions: why queries don't run the way or as fast as they expect. Ongoing checkpoint reviews are a key tool to assess and identify opportunities for improvement with prior deliverables [3]. Data warehouses most often fall off track when they lose their focus on serving the information needs of the business users.

2.9. Network Management

If the data warehouse is implemented using a heterogeneous group of platforms, network management will be one of the most difficult and tough tasks [5]. New users will continuously come online and users along with equipment are invariably moving to new locations as well. The networking hardware is always increasing in numbers with LANs, WANs, hubs, switches, routers and multiplexers. Users always want to access internet based Data sources along with the corporate data, requiring even more bandwidth and network management resources. There should be some knowledgeable person in the organization who could handle these issues.

2.10. Software and Hardware Issues

According to [7] client/server technology is less reliable, secure, and timely than its mainframe predecessor. Data access tools are just beginning to mature. Networks add new layers of complexity, and monitoring performance and tuning of servers is imperfect.

The Results are gaps in available technology and software, leaving users frustrated and their needs unmet. To overcome these problems warehouses needed to get their software and hardware updated in a timely manner to avoid any shortcomings in performance. [18] Three strategies are available to make changes to this technical layer depending upon the scope, timeframe and criticality of the data warehouse environment. These strategies include:

i. Installing new software releases, patches, hardware components or upgrades, and network connections (logical and physical) directly in the production environment.
ii. Installing new software versions, hardware upgrades, and network improvement tasks in a temporary test environment and migrates or reconnects to production once certification testing has concluded.
iii. Installing technical infrastructure changes into a permanent test or maintenance environment and migrate the production environment once certification testing has concluded.

2.11. Extract, Transform and Load (ETL)

ETL is a data integration function that involves extracting data from outside sources (operational systems), transforming it to fit business needs, and ultimately loading it into a data warehouse [10] To solve the problem, companies use extract, transform and load (ETL) technology, which includes reading data from its source, cleaning it up and formatting it uniformly, and then writing it to the target repository to be exploited . The data used in ETL processes can come from any source: a mainframe application, an ERP application, a CRM tool, a flat file or an Excel spreadsheet. [11].

According to some industry experts approximately 60-80 percent of a data warehousing project effort is spent on

this process alone. In today's high volume, client/server environment data acquisition techniques have to coordinate staging operations, filtering, data hygiene routines, data transformation and data load techniques in addition to cooperating with network technology to populate the data warehouse and operational data stores.

It's more than just in the data acquisition process. While data acquisition is the predominant process using the ETL tools, the data delivery process and movement of data from the analytical functions to the ODS or operational systems use ETL processing as well. The full-blown set of ETL operations must combine into a cohesive, integrated system. A system that ensures each process will fit into the overall effort efficiently, determines how the tool will be used for each component and synchronizes all ETL events. There should be ETL expert in the organization who ensures that the ETL processes have strength and endurance [12].

This requires an overarching view and control over the entire environment and is the job of an ETL architect. The ETL architect ensures program efficiency by creating a cohesive ETL architecture to ensure that the various ETL functions form one cohesive system. Taking the time to properly architect a highly integrated set of processes and procedures up front is the fastest way to achieve a smoothly running system that is maintainable and sustainable over the long haul. To accomplish an efficient, scalable and maintainable process, the ETL architect must have the following roles and responsibilities [12].

i. The ETL architect should have a close eye on the needs and requirements of the organization. He/she must understand the overall operational environment and strategic performance requirements of the proposed system. The architect must interact with the source system operational and technical staff, the project database administrator (DBA) and the technical infrastructure architects to develop the most efficient method to extract source data, identify the proper set of indexes for the sources, architect the staging platform, design intermediate databases needed for efficient data transformation and produce the programming infrastructure for a successful ETL operation.

ii. An ETL programmer should not only see his or her single threaded set of programs. The architect must see the entire system of programs, He/she must ensure the technical team understands the target database design and its usage so that the transformations which convert the source data into the target data structures are clearly documented and understood. The ETL architect oversees each and every one of the ETL components and their subcomponents.

iii. The ETL process is much more than code written to move data. The ETL architect also serves as the central point for understanding the various technical standards that need to be developed if they don't already exist. These might include limits on file size when transmitting data over the company intranet, requirements for passing data through firewalls that exist between internal and external environments, data design standards, standards for usage of logical and physical design tools and configuration management of source code, executables and documentation. The ETL architect must also ensure that the ETL design process is repeatable, documented and put under proper change control.

iv. A key consideration for the ETL architect is to recognize the significant differences that the design and implementation methods for a business intelligence system have from an online transaction processing (OLTP) system approach.

v. The role of the ETL architect also extends to that of consultant to the programming effort. The architect works closely with the programmers to answer questions and plays a key role in problem resolution. Depending on the size of the programming effort and the project organization, the ETL architect may also supervise the development of the programming specifications. In any case, the ETL architect plays a key role as a reviewer and approver during the peer review process.

vi. One last role for the ETL architect must be to ensure that the various software tools needed to perform the different types of data processing are properly selected

ETL is one of the most important sets of processes for the sustenance and maintenance of Business Intelligence architecture and strategy [12].

3. CONCLUSION

A major reason for data warehouse project failures is poor maintenance. Without proper maintenance desired results are nearly impossible to attain from a data warehouse. Unlike operational systems data warehouses need a lot more maintenance and a support team of qualified professionals is needed to take care of the issues that arise after its deployment including data extraction, data loading, network management, training and communication, query management and some other related tasks. To carry out all these functions and processes a qualified team of full time skilled professionals is required who can efficiently and constantly take care of all the data warehouse maintenance issues in a timely manner. the first and the most important part of a data warehouse maintenance program is the training of its users. The training program gives the users of data warehouse an insight into the qualities and capabilities of a data warehouse and teaches them the methods to benefit from it. Often the data warehouse projects fail because the users don't know how to use it according to the business needs. The communication process also continues along with the training program. The communication process keeps the business users and IT users in contact with each other to have exchange of views, suggestions and any guidance towards enhanced performance of a data warehouse.

The help desk and problem management play an important role in taking valuable output from the data warehouse. Some of the processes like ETL are carried out during the night, which require presence of support staff to rectify any problem.

The process defines necessary routines and instructions to counter any problem found in the warehouse. If the problems found in the data warehouse are not addressed at the right time, this leads to performance shortfalls, and usability and availability issues in near future. Thus help desk and

problem management play a key role in improving data warehouse performance and getting the desired out put from it. Network management also plays its part in improving data warehouse performance. we concluded that by having a fast and reliable network user queries get a much shorter response time especially in a distributed data warehouse.

The hardware and software resources for the data warehouse are compulsory for taking maximum output from it.

Query Cache technique is to store queries and their corresponding results. If similar query is submitted by any other user the result will be obtained using cache memory.

ETL functions needed to be carried out by a competent and trained ETL team. The ETL team is headed by an ETL expert. It's the responsibility of ETL architect to devise a comprehensive and effective ETL process to load the data warehouse. The ETL architect/expert ensures that the ETL processes have strength and endurance. The ETL architect works in close coordination with the business users and identifies which data and at what level of detail is required. View materialization is a strategy used to provide fast answers to user queries.

4. REFERENCES

[1] System Analysis and Design. 2nd Edition. 1999. Elias M. Awad.

[2] The evolving data warehouse market: Part1. Charlie Garry copyright 2004 Meta Delta.

[3] The data warehouse toolkit. 2nd edition. Ralph Kimball, Margy Ross. 2002 Wiley computer publishing.

[4] Data Warehouse Management Handbook by Richard Kachur. 2000 Prentice Hall

[5] Building, using, and managing the data warehouse. Ramon Barquin, George Zagelow, Katherine hammer, Mark sweiger, George Burch, Dennis Berg, Christopher Heagele, Katherine Glassey-Edholm, David Menninger, Paul Barth, J.D. Welch, Narsim ganti, Herb Edelstein, Bernard Boar, Robert Small. Data warehousing institute series from Prentice Hall.

[6] Data warehouse, Practical advice from the experts. 1997. Prentice hall by Joyce Bischoff & Ted Alexander

[7] Lessons from a successful data warehouse implementation. Dr. John. D Porter and john. J Rome. Arizona State University.

[8] Building a data warehouse for decision support. 1996 Prentice Hall. By Vidette Poe with contributions from Laura L. Reeves.

[9] Wikipedia, the web's free encyclopedia http://en.wikipedia.org/wiki/Data_warehouse

[10] The concise technical dictionary http://www.thetechdictionary.com/ term/etl_%28data_integration%29

[11] The computer world magazine http://www.computerworld.com/databasetopics/businessintelligence/datawarehouse/story/0,10801,89534,00.html

[12] Fundamentals of database systems. 4th Edition. Persons international and Addison Wesley. Ramez Elmasri and Shamkant B. Navathe

[13] Ideal Strategy to Improve Datawarehouse Performance by Fahad Sultan & Dr. Abdul Aziz. (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 409-415

[14] Efficient incremental view maintenance in data warehouses. Ki Yong Lee, Jin HyunSon, Myoung Ho Kim. Korea Advanced Institute of Science and Technology.

[15] R. Winter and B. Strauch. A method for demand-driven information requirements analysis in data warehousing Projects. In Proc. HICSS, pages 1359–1365, Hawaii, 2003.

[16] Albrecht, J.; Hümmer, W.; Lehner, W.; Schlesinger, L.: Using Semantics for Query Derivability in Data Warehouse Applications, appears in: Proceedings of the 4th International Conference on Flexible Query Answering Systems (FQAS'00, Warsaw, Poland, October 25 - 25), 2000.

[17] Adiba, M.E., and Lindsay, B.G. "Database Snapshots," Proceedings of the 6thInternational Conference on VLDB, pp. 86-91, 1980

[18] Strategy to make superior Data ware house by Vishal Gour in International Conference on advance computing and creating entrepreneurs Feb2010.