A REVIEW MODEL ON QUERY OPTIMIZATION USING OPEN SOURCE DATABASE

Navneet Mehta
Student (Mtech)
Department of CSE
Lovely Professional University
Jalandhar, Punjab, India

Abstract
Query optimization is used to optimize the efficient result with less time and minimum cost. The query formatted according to particular search engine and then the translated query is passed to the search engines. HTML screen scrapping is used to parse the data from each individual search engine. Then these results are combined to make the single result. The purpose of this method is ensuring that no duplicate results are displayed to users. This will improve the quality of search engine and reduce the unnecessary noise before displaying to the user. The filter function is used to remove the unnecessary results. The SPARQL protocol and RDF graph are used. In this paper, the query is optimized in case of cloud computing. Due to increase the business technology, the customers move towards cloud computing. In this paper, the various approaches are reviewed for query optimization.

Keywords
Query Optimization, Constraint based pruning, Federation, SPARQL query, RDF graph
Introduction
Queries are executed either in relational database xml database. Because of limited storage space available we use the cloud for executing the query. Cloud computing provides software and hardware which is available on the network. The user use it according to their requirement by paying money per use. Cloud computing is just a new name given to   grid computing   cloud computing is a technology that allow user to use application without installation and access personal files at any computer.
According to NIST definition,

" A cloud Computing is a pay per use mpdel which available, convenient and on demand network access to a shared pool of resources

minimal management efforts and service provider interaction."
Cloud Computing consist of three deployment model
1. Infrastructure as a Service.
2. Platform as a Service(PaaS)
3. Software as a Service(SaaS).
Basically, users use the software as a service in which user does not know from which location software is accessed. They just require the applications for running that software.
Cloud Computing transform a high level query language on a database into an equivalent and efficient low level query language on a relational fragments. Map reduced framework and SPARQL query language are used. RDF database was proposed by W3C to encode the information in a machine readable form. It consists of three triple i.e. subject, predicate and object. Each triple encode the binary relation predicate between subject and object and represent a single knowledge. SPARQL query language extracts the information from RDF Graph.
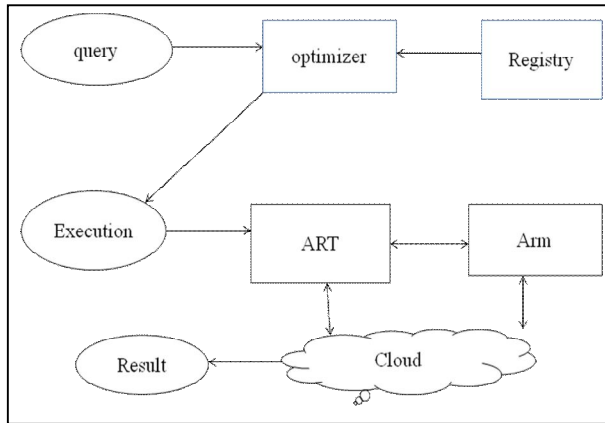
Optimizer are characterized by three dimensions
1. Searching space for execution.
2. Strategy for exploring it.
3. Cost model to compare with each other.

Example
1. Consider a query in which user want to find the name of employees with identification number who work in IT department and whose salary is less than 50000 per month.

Now in this query, first we perform a selection operation on department and salary. Then a join operation is performed and at last projection is performed.

$$\Pi\ ename, empid$$

σ Dept = IT σ Sal < =50000

department   employee
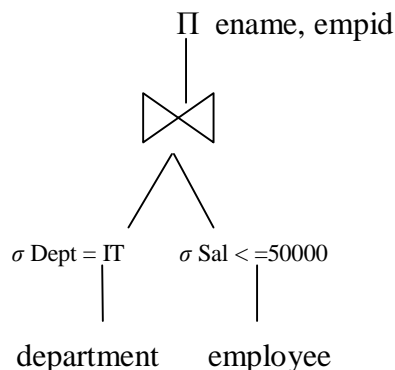
2. Consider a query in which user want to find how many years he/she born i.e.years, months, days, hours, minutes.

Solution:
a.   Enter the date of birth. The syntax is
Define &dob= to_date(' 21-nov-1988');
b.   Select&dob, trunc(months_between(sysdate, &dob/12,0)//'years'//trunc(months_between(sysdate,add_months(&dob,trunc(months_bet ween(sysdate,&dob/12)*12,0)),0)// 'months'//add_months(&dob,trunc(months_b etween(sysdate,add_months(&dob,trunc(mont

hs_between(sysdate,&dob)/12)*12,0)),0))// 'days'// to_char(sysdate, 'HH24')// 'hours'// To_char(sysdate,'MM')//'minutes'// to_char(Sysdate, 'ss') from dual;

Review of literature
Cost Based Optimization Technique
This technique is used for execute the optimized query with low communication cost and processor. We know that the cost of the query execution changed according to the order of the processing steps. Thus, the cost depends upon the size and complexity of query. In this technique, the query optimizer creates a binary tree. All dependent variable are set to the branch tree. Weights are assigned to each variable. Thus, the cost is calculated by summing the weights. The weights are assigned according to computational time of operation. This technique consist of two algorithms
1. Create a graph
2. Store the result of graph.
Heuristic approach is used for storing the result of query in storage space. In any organization or a system same query are executed again after a certain period of time. This increases the cost and reduces the performance. This problem is reduced by heuristic approach.

Semantic Based Query Optimization technique
This technique is use integrity constraints such as primary key, foreign key, unique key etc. the views are created from original table so that unauthorized person or user can only access the available data. However, this technique does not store the result of query. So, every time query is executed. It increases the cost and reduce the efficiency.
Semantic based optimization consists of two algorithm.
1. Chase
2. Back Chase
    This two algorithm is used for optimize the queries.
Federation
As the data grow day by day in cloud computing, it is difficult to retrieve the result of complex

query since data is distributed into various parts and stored in different space. This problem can be solved by divided the query into multiple queries and each query is executed at different location.

Approaches to federation

1. Warehousing: the data are stored in advance at the central location. Each query is evaluated through this central storage.
2. Distributed query processing: the complex query is divided into multiple queries. Then, the query is distributed among different location for execution.

Alternatives of federation

Federation consist of three alternatives

1. Integration in single repository: All data is physically loaded into a central location. Queries are evaluated against central location i.e. single repository. The disadvantage of this alternative is that it need to download the complete database for executing the query. However, unnecessary data are downloaded instead of small sets of data.
2. Federation over multiple single repositories: The data is managed by each data sources in single repositories. The result are then merged together to produce the single result. Single repositories are federated over the single federation layer. For end user, this approach is completely transparent that is no query rewriting is needed.
3. Federation over multiple SPARQL endpoints: Multiple repositories are accessed over single federation layer. The access is realized using SPARQL endpoints available to the data provider. An adhoc federation can be built by simply add the additional SPARQL endpoint to the federation. However, the disadvantage is that multiple result produced at each stage than a normal execution. Such results produce irrelevant result.

Constraint based Pruning Approach

This approach is used for pruning the irrelevant results. The probability and selectivity are used at each execution stage. The selectivity select the result based on threshold value which is determined from the available Meta data. The probability selects the high scores at each result. Then this score is compared with threshold value and results are pruned. This approach is an improvement over the existing approaches method.

Various Issues for retrieving the data

The data are retrieved in cloud computing using various techniques with respect to searching techniques and within respect to time when query is optimized.

Types of query optimizer with respect to searching technique

1. Exhaustive search technique: It is cost based and provides optimal results.
2. Heuristic search technique: It performs the selection and projection operation. It uses join and semi join of executing the complex query. But it does give optimal results.

Type of query optimizer with respect to time: It include

1. Static: Query is optimized before execution.
2. Dynamic: Query is optimized at run time.
3. Hybrid: First, the query is executed by using static. If the error is occurred, then query is recompiled at run time.

Conclusion

The query optimization is used for fetch the efficient result with minimum time and cost. The various approaches or techniques are used
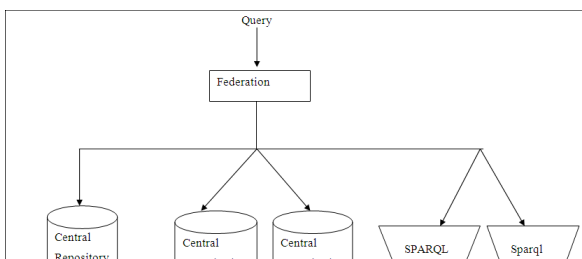
in cloud computing for optimizing the results. The best technique is Constraint based pruning techniques in which some constraints are applied to query and pruning are done to get the meaningful result. The query is executed by different search engines and the results get from these engines are merged to get the single results which is displayed to users.

References
[1]   N. Samatha , K. Vijay Chandu,P. Raja Sekhar Reddy.(Sep 2012) "Query optimization Issues for data retrieval in Cloud Computing" International Journal Of Computational Engineering Research Vol. 2 Issue 5,.
[2] Saurabh Kumar, Gaurav Khandelwal, Arjun Varshney, Mukul Arora, (Sep2011), "Cost Based Query Optimization with Heuristics", International Journal of Scientific & Engineering Research Volume 2, Issue 9,.

 [3]   Alin Deutsch, Lucian Popa, Val Tannen, 2001. "Chase & Backchase: A Method for Query Optimization with Materialized Views and Integrity Constraints", University of Pennsylvania.

[4] Danh Le-Phuoc, Josiane Xavier Parreira, Michael Hausenblas, Manfred Hauswirth, (Sep 2010). "Continuous Query Optimization over Unified Linked Stream Data and Linked Open Data".

[5] Olaf Hartig, Christian Bizer2, and Johann-Christoph Freytag, (2011) "Executing SPARQL Queries over   the Web of Linked Data".

[6] Peter Haase, Tobbias Methab, Michael Ziller, (2009). "An Evaluation of Approaches To Federated Query Processing over Linked Data"

[7] Michael Schmidt, Michael Meier, Georg Lausen, (2009). "Foundation of SPARQL query optimization".

 [8] Prof. Bhaumik Nagar1, Prof. Ashwin Makwana2, Chirag Pandya3, (March 2012). "Optimizing query execution over the linked data", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 3.

[9] Pawel Jurczyk and Li Xiong, " Dynamic Query Processing for P2P Data Services in the Cloud".

 Books
[1] Mourad Benchikh,   Elmasri & Navathe , Ramakrishnan & Gehrke, Silberschatz & Korth "Oracle 9i Documentation Query processing".
[2] Anurag Basu, 3rd edition 2010"Relational Database management system"