Improvement of Query Processing Speed in Data Warehousing with the Usage of Components-Bitmap Indexing, Iceberg and Uncertain Data

Uma Pavan Kumar
Kethavarapu
Associate Professor, IT Dept.
AIMS Institutions, Bangalore

Lakshma Reddy, Ph.D
Bhavanam
Principal BCC College,
Bangalore

Sreedevi.S.Erady
Assistant Professor, IT Dept.
AIMS Institutions, Bangalore

ABSTRACT

Data warehousing is a huge collection of data sources meant for handling strategic decisions with historical and current data. The recent trend in information technology market is big data analytics. Huge amounts of data available with the companies but the proper information as per the requirements is still a challenging issue. The current article is dealing with the implementation of iceberg queries, slowly changing dimensions and uncertain data processing. The idea behind the integration of these aspects is processing subsets of the data from the huge amounts of the data. In case of iceberg querying the main theme is aggregate query processing such as count, sum, average, maximum and minimum kind of calculations. In most of the cases the analysis of the data and comparison of the performance aspects between aggregations only, so usage of iceberg querying will improve the processing speed of the data warehousing. The second component we are considering to process data warehousing is slowly changing dimensions. The dimension tables are the basic things of the data warehousing construction, usually dimensions will change slowly not frequently. If we are able to track the changes done to the dimensions such as maintenance of historical and current data, with the tracking of data we can get the subset of data which got modified or which we need to process, which will greatly reduce the number of records to process. The third component we are considering is uncertain data processing. The data which is not having any structure and no information about the format is known as uncertain data, now a days the data population of the data like reviews, likes or shares in social media, MS response all are recorded in uncertain format. The processing of uncertain data with softest computing will give the identification of missing data, parameterization aspects are possible.

General Terms

Data Warehousing, Query Processing, ETL, OLTP, OLAP, Repository, Data Marts

Keywords

Slowly Changing Dimensions, Ice Berg queries, uncertain data, Bitmap Indexing, Soft set computing

1. INTRODUCTION

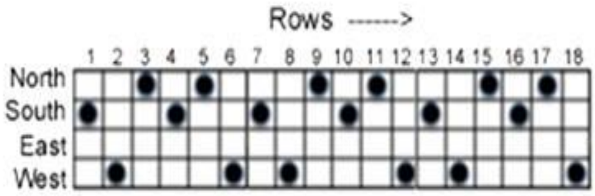
We Data warehousing environment allows processing of various source data (.doc, XML, ERP, Java, Relational etc.) by Extraction, Transformation and Loading (ETL) process. ETL the data will be identified in the source known as Online Transaction Processing (OLTP) which is day to day data and having clerical type of entries. After ETL processing the data will be populated to a repository known as data warehousing or sometimes some portion of data can be redirected to data marts depending on the approaches provided by inmon and Kimball· The important point here is the format of the data in the data warehouse or data mart is relational because management of that kind of data is easy when compared with other categories. Once the data is populated into the repository management of the strategic decisions is easy, which is technically known as Online Analytical Processing (OLAP), through which is possible to generate the reports and data analysis aspects. Data warehousing users are ranging from top management to end user so which is known as Techno-Functional module. Technical aspects involving identification of source data processing up to the level of repository. Functional aspects involving management of strategic decision after getting the reports. Some of the ETL tools are Informatica, Abinitio, Data stage with parallel jobs and Oracle Warehouse Builder, OLAP tools are Business Objects, Cognos Report net, and Cognos Impromptu. The organization of this paper is in section 2 the processing of data warehousing data is described, in section 3 the component bitmap indexing importance and significance in data warehousing is explained in section 4 the importance and relativity to data warehousing in section 5 description of uncertain data processing, in section 6 how all these components are integrated is explained Section 7 explains the usage of slowly changing dimensions in the data processing.

2. WORK FLOW IN DATA WAREHOUSING

In the development of information technology applications especially in case of bulk data processing data warehousing environments are elected. The data warehousing involves the techno-functional aspects of the application context and user context. Various user levels are supported by data warehousing environment such as ETL developers, Strategic analyzers, design engineers, Report developers. The flow of the data starts from source gathering the source data might be Flat files (.Doc, .Xls, COBOL …), XML files, ERP models, relational models. The main advantage of data warehousing is support of various categories of source data. The source data is refereed as online transaction processing, the source data is then Extracted transformed and loaded into the target repository. The repository may be data warehousing or data mart, warehousing consists of the entire data within it where as in data mart a specific portion of the data or a specific department data is available. So with the usage of ETL processing the data will be populated to the repository. Later the data will be used to generate the reports known as online analytical processing. The report generation will helpful to the users of the company for the strategic decisions, which will decides the market value and improvement of productivity of the organizations. The user levels are either functional or technical, the technical users in data warehousing environment having the functionalities of ETL processing and OLAP handling. The functional users concentrate on consideration of various contexts of the data, putting queries towards to the technical users. The most commonly used ETL tools in software industry are Informatica, Abintio, Data stage, Oracle warehouse builder.OLAP tools like Business Objects, Cognos are helpful to generate the reports based on the requirements of the functional users of the company. The following representation shows the general flow of data in data warehousing environments. First step is to design the source qualifier by importing various source data into the ETL process. With the help of source qualifier various transformations of the data like Routing, Rank, Lookup of data, Update Strategy, Aggregator, Expression, Joiner, and Filter can be used so as to get the output records to the target data. Once this process is done then save the entire design into the repository. The next step is workflow creation through which we can create the task so as to materialize the logical design into the data repository. The final step is workflow execution. The common diagrammatic representation of data warehousing environment data flow will be represented in fig 2.The data warehouse repository consists of Meta data which describes data about data, summary data which consists of analyzed data in the form of aggregation format, raw data contains the raw facts which are available in non-processed format. In the reporting side online analytical processing allows the user so as to generate the data reports as per the requirements

3. BITMAP INDEXING USAGE IN DATA WAREHOUSING

Indexing is a way of labeling the data with some pointers so as to locate the data in the faster manner. Various indexing mechanisms are available so as to support the faster data retrieval in the data warehousing environment. The most commonly used and best method of indexing in data warehousing is bitmap indexing.

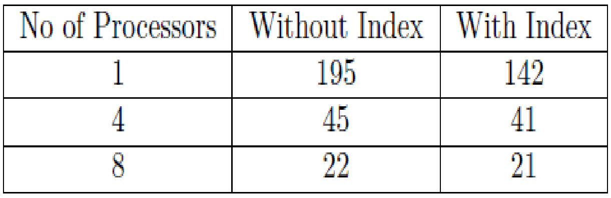


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
North			●		●				●		●				●		●	
South	●			●			●			●			●			●		
East																		
West		●				●		●				●		●				●

Fig 1: Bitmap Indexing example

The Bitmap indexing is best suited for handling of Boolean kind of the data such as gender, result of examination, True or false category of the data. Bitmap indexing with the association of data warehousing will produce better results because the bitmap indexing is mainly depends on 0’s and 1’s kind of the data, which will directly processed by the CPU that is no need of conversion of the data items into another format which will greatly reduce the processing time of the records.

Table 1. Time factors with and without data processing



No of Processors	Without Index	With Index
1	195	142
4	45	41
8	22	21

Various bitmap indexing techniques are available to handle the bulk data for instance simple bitmap index, encoded bitmap indexing, enhanced bitmap indexing and scatter bit map indexing. Depending on the requirement the corresponding indexing technique can be used. Data warehousing data usually consists of bulk amounts of the data; the main requirement of the data warehousing environment is to get the data in stipulated time factor which is a big challenge. The best method of reducing time factors while processing the bulk data is indexing. In case of data warehousing bitmap indexing is most suitable type of indexing mechanism so as to process the bulk amounts of the data. To handle current and historical data formats for generation of the strategic decisions bitmap indexing associated data processing is affordable.

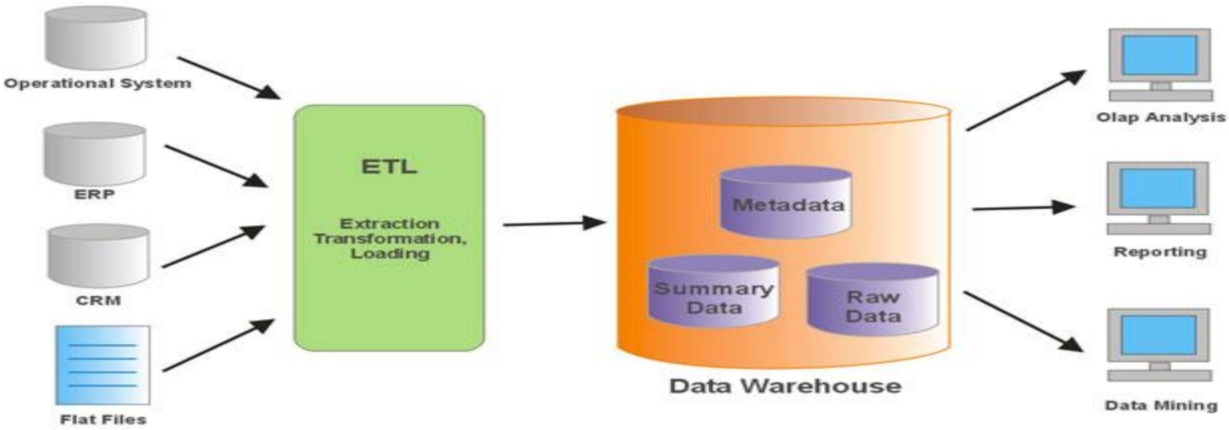


Fig 2.Data warehousing Environment

4. DATA WAREHOUSING SCOPE OF APPLICATIONS

Business Intelligence (BI) applications are systems designed to help an organization make intelligent business decisions based upon the results of analyzing massive amounts of data. Companies may architect their business intelligence environment in various ways. However, most BI implementations contain three main functional tiers. These include an Extract, Transformation and Load tier (ETL) Data warehouses can be described as *decision support systems* in that they allow users to assess the evolution of an organization in terms of a number of key data attributes or dimensions [1]. By exploiting multi-dimensional views of the underlying data warehouse, users can "drill down" or "roll up" on hierarchies, "slice and dice" particular attributes, or perform various statistical operations such as ranking and forecasting. This approach is referred to as Online Analytical Processing or OLAP. The progresses of computer technology have enabled the development of new kinds of large database applications such as data warehouses and scientific databases. Unlike the traditional databases systems, these data sets are characterized by the high dimensionality, and are mostly read-only and append only systems. The query Data warehouses can be described as *decision support systems* in that they allow users to assess the evolution of an organization in terms of a number of key data attributes or dimensions. By exploiting multi-dimensional views of the underlying data warehouse, users can "drill down" or "roll up" on hierarchies, "slice and dice" particular attributes, or perform various statistical operations such as ranking and forecasting. This approach is referred to as Online Analytical Processing or OLAP. The progresses of computer technology have enabled the development of new kinds of large database applications such as data warehouses and scientific databases. Unlike the traditional databases systems, these data sets are characterized by the high dimensionality, and are mostly read-only and append only systems. The query performance in such environments is a critical challenge as most of the associated queries are complex and ad hoc in nature and require huge volumes of data to be processed.

5. CONTEXT OF UNCERTAIN DATA

The web based data sources like twitter, social media generating huge amounts of the data in the form of likes, shares and comments. All these activities are not structured and not having any predefined formats of the data and data types. To process structured, semi structured like relational models and XML kind of data we are having associated editors and packages. To handle uncertain data models first the subset of data need to be converted into a specific format and later the processing of those data items is possible. The concepts like probability model and fuzzy logics are available to handle the uncertain data types, but they are suffering from lack of parameterization aspects and generalization issues. The best way of handling the uncertain data is soft set computing through which we can achieve the parameterization along with generalization of the subsets of uncertain data. To achieve the conversion process the soft set computing is providing some notations along with the notations we can achieve the common model to perform parameter based solution of the data. With the methods of missing values and approximation principles we can generate algebraic structures and interval-valued fuzzy sets. The soft set computing also depends on Boolean type of the data as that of bitmap indexing and iceberg querying. The uncertain data may be populated in the form of reviews, SMS over mobiles, online voting and interviews .All these populated items are not having common format and certain, but there is a mandatory requirement to handle and process those data subsets generated by various means of sources. The following example explains low, high, good, bad, loss, profit but all are represented in common format of binary which are easy to integrate with iceberg and bitmap indexing formats as they are also represented in binary formats. After getting the binary formats of all the data sets the applicability of operations like Bitwise OR, AND, NOT, XOR are helpful in the generation of required result set. Processing of uncertain data is a big challenge in data warehousing environments because the data categorization, classification and utilization all are needed which is tedious in general. The softest computing is best solution for uncertain data processing.

6. ICEBERG BASICS

Ice berg query is a special class of aggregation query, which computes aggregate values above a given threshold. Provide dynamic pruning and vector alignment algorithms to efficiently compute iceberg queries using compressed bitmap indices. If iceberg queries are integrated with bitmap indexing we can expect better performance. The processing of iceberg queries along with bitmap indexing is important for the following reasons.

1. Implementation of iceberg queries for processing of aggregate functions like Sum, Maximum, Minimum.

2. Processing of iceberg queries for the average computing without anti-monotone property.

3. Prediction of number of iceberg queries and optimal order of the attributes for getting better outcome in less amounts of time.

7. INTEGRATION OF BITMAP, ICEBERG AND UNCERTAIN PROCESSING

The main aim of our research work is to get benefit of integration of bitmap indexing, iceberg query processing and uncertain data processing. The connecting factor between them is binary data format. The main basis of bitmap indexing is data encoding in the format of 0's and 1's, through which the entire indexing of the data and data processing and result generation is done. The iceberg querying is mainly concentrates on aggregation of the data items such as count, sum, maximum, minimum and avg of the data. The aggregation of the records is most helpful for getting the strategic decisions. The way of converting the iceberg resultant into 0's and 1's is through the threshold value. If the generated value is accepted by the limit value we can notify that as 1 otherwise it would be 0.So the conversion of iceberg generated result will be converted into 0's and 1's which is required by the bitmap formatting. Uncertain data processing is done by soft set computing which is based on the concept of parameterization such as identification of missing values, pattern based data processing. In the context of parameterization the mathematical format will be replaced by 0's and 1's and the same will be integrated with iceberg and bitmap indexing which we have already done. The following sequence of steps to be followed while handling the above mentioned context of the data processing.

ALGORITHM (BITMAP, ICEBERG, UNCERTAIN)

INPUT: BITMAP_VECTOR,
ICEBERG RESULT_SET,
UNCERTAIN_DATASET

OUTPUT:
RSULTANT_RECORDSET

Step 1: Create Bitmap indexing for the given record set and notify the generated indexing as Bi [R1, R2, R3……Rn]

R1…Rn denotes the record set selected by the bitmap indexing.

The outcome of this step is bitmap vector with 0's and 1's combination. We can denote this generated resultant as B [v1, v2, v3……..vn] where v1, v2….vnє {0, 1}

Step2: Create Iceberg query for the data set which involves the aggregate functions such as Maximum, minimum, count, sum or average.

We are denoting that with IQ (Aggregate operators) which will give the aggregation of record set.

IQ(Group by{R1,R2,……Rn},Count{R1,R2,….Rn}……..).

The outcome of this step is Iceberg query generates the resultant records with various aggregate functions.

Step3: Applying Threshold Value (TV) on the resultant iceberg resultant records which will give the resultant records which are meeting the specified value.

The mechanism is TV checking in {IQ_ResultSet}

The resultant of this step, is the set of records meeting the TV value, which we can be denoted as {TVIQR1, TVIQ2 …TVIQn}

Step 3: The uncertain data processing with soft set computing which will identify the set of uncertain data values from the available data sets.

With softest computing parameterization concept the missing values in the data sets, pattern based data processing can be achieved.

SSCOMP {Uncertain_SET} →{Missing_value, Pattern_Ids}.

Step4: Integration of all the above generated resultant record sets.

Such as B[v1,v2…..vn] OP IQ[R1,R2,…..Rn] OP SSCOMP{Uncertain_SET} where OP may be in OR,AND,XOR,Union,Intersection.

Step 5: STOP

8. CONCLUSION AND FUTURE SCOPE

Data warehousing environment is meant for processing the current and historical data handling so as to generate the strategic decisions. The main challenge of data warehousing is processing of huge amounts of the data in stipulated time factors. For this we are depending on bitmap indexing which is very much helpful in the faster data processing. The iceberg querying is useful for the aggregation of the records based on the threshold value as a background of various aggregate functions. The third component we considered is uncertain data which is needed because the real time data nowadays generated in the uncertain formats. For the uncertain data processing the main key aspect we selected is softest computing. The conclusion of this paper is the integration of bitmap indexing, iceberg querying and uncertain data processing with the data processing in data warehousing will definitely give the better resultant data processing with in the specified amount of time factors while processing bulk amounts of data in data warehousing environments. The future work of this research is to generate the time factors for individual data processing and integration of all these three components time factors to show that the data management in the context of these components is better than normal data processing in data warehousing.

9. ACKNOWLEDGEMENTS

The current research work is carried out in Pondicherry engineering college, under the guidance of Dr.S.Saraswathi, HOD IT department, Pondicherry engineering college,puducherry,I am very thankful to my guide.

10. REFERENCES

[1] Bhosale"Efficient Indexing Techniques on Data Warehousing" International Journal of Scientific & Engineering Research vol 4, Issue 5, May-2013, ISSN 2229-5518.

[2] NaveenGar,PhDscholar,SNUniversity,Jharkhand,"Bitmap Indexing technique for data warehousing and data mining", International Journal of Latest trends in engineering& Technology vol 2.Issue 1,January 2013,ISSN:2278-621X

[3] Zanab qays abdulhadi, school of information systems and engineering, central south university, china. "Bitmap index as effective indexing for low cardinality column in data warehouse", International Journal of computer applications, vol 68, April 2013, ISSN: 0975-8887).

[4] Jesus Camacho- Rodriguez "Web data indexing in the cloud: Efficiency and Cost reductions", ©ACM 2013, March 18-22.

[5] Naveen Garg, PhD Scholar, "An Efficient Approach for data indexing in DWH&DM", International Journal of Innovations in engineering and Technology, vol-1, Issue 4, Dec 2012.

[6] Biyramjit paul, Asst.Prof Dept. of CA,Westbengal "Comparative study of various Bitmap Indexing techniques used in Data warehouse" , ,International Journal of Emerging trends & Technology in computers,ISSN 2278-6856,Vol-1,Issue-3,Sep-2012.

[7] Amorntep Keawpibal "Enhanced Encoded Bitmap Index For equality Query", Thailand, IEEE, 2012.

[8] T.P.Latchoumi, "Multi Agent Systems In Distributed Data warehousing", International Conf. on Computer & Communication Technology

[9] Andrea Campagna,"Frequent Pairs in Data Streams: Exploiting Parallelism and Skew", 2011 11th IEEE International Conference on Data Mining Workshops

[10] Gehad Galal ,"Exploiting Parallelism in Knowledge Discovery Systems to Improve Scalability", ,1060-3425/98 (c) 1998 IEEE

[11] Marco Vieira, Henrique Madeira" Integrating GQM and

Data Warehousing for the Definition of Software Reuse Metrics", ,2011 34th IEEE Software Engineering Workshop

[12] Munawar " Towards Data Quality into the Data Warehouse development", 2011 Ninth IEEE International Conference on Dependable,

[13] Abdolreza Hajmoosaei, "Autonomic and Secure Computing978-0-7695-4612-4/11© 2011 IEEEDOI 10.1109/DASC.2011.1941200-2011 IEEE Ninth International Conference on Dependable,

[14] COMPARISON PLAN FOR DATAWAREHOUSE SYSTEM ARCHITECTURES Data sheet© 2011 Microsoft Corporation

[15] Satkaur , Research scholar, S.K.I.E.T. ,Kurukshetra, Haryana "International Journal of Advanced Research in computer Science and Software Engineering" , , Volume 3, Issue 5, May 2013 ISSN: 2277 128X .

[16] Jens Dittrich JorgeArnulfo,Quian´eRuizInformation Systems Group Saarland University "Efficient Big Data Processing in Hadoop Map Reduce, Proceedings of the VLDB Endowment", Vol. 5, No. 12Copyright 2012 VLDB

[17] Lizhe Wang, School of Computer, China University of Geosciences, "G-Hadoop: Map Reduce across distributed data centers for data-intensive computing, Future Generation Computer Systems", the international journal of grid computing and esciences 2012 Elsevier.

[18] Bo Dong, Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China, "A Novel Approach to Improving the Efficiency of Storing and Accessing Small Files on Hadoop: a Case Study by PowerPoint Files", , 2010 IEEE International Conference on Services Computing.

[26]

[19] Muhammad Inayat Ullah, Gomal University, "Transformation of Flat File into Data Warehouse", Global Journal of Computer Science and Technology, Volume 11 Issue 13 Version 1.0 August 2011.

[20] Sheetal ganu, Punjabi university, "Improved Extraction mechanism in ETL process for building of a Data Warehouse", IEEE international conference, Mumbai.

[21] Ranjit Singh, Research Scholar, University College of Engineering (UCoE), Punjabi University, "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010 41ISSN (Online): 1694-0784

[22] Md AL Mamun, "Performance improvement techniques for customized data warehouse", IOSR-JCE, April 2013.

[23] Bin He "Efficient Iceberg query evaluation using compressed bitmap index", , IEEE Transactions,Sept,2012

[24] Chuyang Wei, "Efficient Cube computing on an extended multi-dimensional model over uncertain data," IEEE 2012

[25] http://www.tgc.com/dsstar/01/0109/102533.html