International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4

ISSN: 2454-4248
95 – 97

Approaches to Avoid Traditional Multidimensional Data Cube: A Survey

Prarthana A. Deshkar
Ph.D. Scholar, CSE Department,
G. H. Raisoni College of Engineering, Nagpur, India
prarthana.deshkar@gmail.com

Dr. Parag S. Deshpande
Supervisor, CSE dept,
G. H. Raisoni College of Engineering, Nagpur, India
psdeshpande@cse.vnit.ac.in

Prof. A. Thomas
HoD, CSE department
G. H. Raisoni College of Engineering, Nagpur, India

***Abstract*—** Data analysis is the growing need of the current era. Data analysis is not only restricted to business domain only. Advancement in the technology opens the door of technology to the every common person and hence data generation is increasing exponentially day by day. Incorporating such huge amount of data in the data analysis system is the big challenge. Handling variety of data is also the difficult issue. The numerous options are coming out to solve these problems. To support the decision making system, Online Analytical Processing (OLAP) is the more suitable option. OLAP uses the multidimensional data analysis approach of data analysis. OLAP includes the analysis of current data as well history data and also the aggregated or summary data. To handle the aggregated data traditionally data cube is used. This paper focuses on the various research techniques to enhance the performance of the data cube.

***Keywords*-** *OLAP, data analysis, multidimensional data analysis, aggregation, summary data.*

*****

I. INTRODUCTION

Decision making systems are using all the aspects of the data, also all the levels of the data in which it can be represented. Modeling the data in the multidimensional model, facilitates the user in analyzing the data with respect to the dimensions of the data. Dimension and measure value or fact value are the components of the multidimensional data. Measure or fact value is the value on which analysis can be done to make the decisions. Dimensions are the entities which describes the fact value.

Eg. Sale of coffee is Rs. 5000 in Mumbai in January 2018.

In this example sale is the fact value and the components describing it, i.e. coffee, Mumbai, January 2018 are the dimension values. Dimension entities show various relationships between the entities like, hierarchical relationship, sequential relationship, and dependency relationship. To perform the OLAP analysis, data need to be absorbed in this model.

As the dimension structures are carrying hierarchical structure, data can be represented at various levels of granularity. The current or actual data is present at the most granular level, and the data need to be aggregated at the higher levels of granularity. E.g Sale data of the product in the retail industry is collected at minute level and can be aggregated at, day, week, month, year level in the time dimension. Such aggregated data plays very important role in the decision making system. With the help of analysis done on aggregated data, decisions are made to enhance the current system.

Multidimensional data analysis is not restricted for the commercial domain; such type of analysis is required in almost all types of domains. Cube is the most popular and accepted architecture to store the aggregated multidimensional data. A cube is used to generate aggregation of multiple dimensions or multiple combinations of multiple dimensions. Cubes are generated for the faster data retrieval as the aggregations are already performed and stored. Cube can perform complex calculations quickly since all possible calculations are pre-aggregated at the time of building a cube. Hence cube gives the faster results i.e. it reduces the query execution time. But this is done with the side effect of space and time overhead.

For example, network traffic analysis software records network information per seconds. If we want to analyze network traffic information for last three months, data would contain 7776000 rows for only one entity say IP address. Now if want to build cube on this with two parameters, it will generate 7776000^2 rows. Here cube formation will takes months, during which new data would have been generated with new network traffic trends. Invalidating all aggregates in the cube since those calculations were for old network uses. It restricts us from having recent data analysis.

To resolve this issue, various approaches are proposed with its own pros and cons. This paper aims to discuss some of them. The paper is organized as section I tells the approaches

 ISSN: 2454-4248
95 – 97

proposed to optimize the cube generation and storage process. Section II discusses some systems which follows the multidimensional data analysis system with some variation in the cube architecture to resolve the challenge of time and storage in the cube generation process.

I. Cube handling techniques

a. Dynamic Cubing

Real time applications are generating the data which needs to be refreshed in real time in the data warehouse. Managing such real time data warehouse is the challenge in front of the data analyst. The dynamic cube is the technique which is mainly focusing on this need of handling dynamic data warehouse.

Important aspect of data warehouse generation process is the data modeling part. While designing the dynamic cubing architecture, hierarchical dimensional structure is considered. This technique can handle naturally ordered or non – ordered dimensional entities. Materialization of view is proposed to optimize the retrieval of huge volume of data. The data model proposed in this work is conceptualized as hierarchical hybrid multidimensional data space. To handle the dynamic nature of the incoming data, the data space axes are kept unordered. A special data grouping structure is considered to partition the data. Various operators, metrics, and relations are defined to make optimal use of the data partitions, which are analogous to the OLAP technology [6].

Partitions are organized in the special tree structure. This tree structure is going to index the partitions. Nodes of this special tree structure are holding the partition space whereas leaves are the fact or measure values and the aggregated values. The nodes which are carrying the data space with the aggregated values are representing the materialized sections of the cuboids. The complete tree structure representation can be considered as the data cube which is handled the partially materialized view. Hence the proposed tree structure to store the multidimensional data space is a cube structure. The nodes of the tree, which are the materialized views, are generated at the run time. The runtime generation of the materialized view is guided by the metrics proposed. Algorithms are also proposed to handle the insertion of the measure values and extraction of the values. Data can be extracted from the tree structure, using the range queries, point queries and the group by queries [6].

b. Generation of Data cube in parallel

In the OLAP analysis response time is crucial and hence cube is having pre-computed data. When data volume is very large, performance may be compromised due to time required to generate the cube for such data. Hence the cube generation can be done be in parallel processing approach. Extendible multidimensional array can be used to construct the cube. The multidimensional data array can be dynamically extendible across all the dimensions. The multidimensional data array is the index array. The main feature of this approach is the data array can be extendible without relocating the data. Parallel cube generation process also can be done by fixed sized array also. This approach uses the shared memory technique to construct the cube. This approach uses, the layered parallel processors and hence improves the performance with minimum processors [12].

II. Systems having different approach to analyze multidimensional data

There are various commercial as well as research efforts to create the multidimensional analysis which are trying to enhance the performance of the cube architecture. Here we discuss few such research systems which had proposed the different techniques to handle the cube and the overall multidimensional analysis environment.

The first system which we are discussing here is the system which is considering big data as a data source to perform the multidimensional data analysis [4]. The framework provided is using the multidimensional cube to perform the analysis. The system is mainly focusing on the scientific data. It follows the server client architecture. The framework provides the layered architecture.

The system stores the data cube as a separate layer in the system. Users are provided various application interfaces to handle the data cube in the system. The system allows the in-memory data cube to reduce the query execution time [4]. This system handles the data cube like the file system, which gives the complete information of the data cubes, including its list, size ownership, etc.[4]. This is going to facilitate the user to handle the data cube effectively. The separate analytical layer is responsible to perform all the cube related operations. It allows to perform the OLAP operations like, slicing, dicing, rollup, etc. This layer is responsible for other analytical operations like, time series analysis, merge, split, aggregations, etc. the system applies various primitives to do all these tasks. The system specifically uses the array based primitive system. One layer is dedicated to manage the workflow in the system. System can handle the requests from various users at a time. To handle the requests put by the users, scheduling need to be done. This scheduling responsibility is carry out by this layer. User need to submit the workflow in the form of the query, written in the query language designed by the system. This layer manages the execution of job by distributing it to various cubes [4]. Thus

 ISSN: 2454-4248
95 – 97

the query execution time can be reduced. At the last front end layer is designed to communicate with the system. The user is going to communicate with the system through the workflows via the front end layer. The front end is designed to handle the complex user scenarios. The system builds various operators to handle the different data mining operations [3]

The other research approach is the multidimensional data analysis system designed as software as service. The architecture of the system is layered architecture. The data layer is to handle the data repository required for data mining algorithms and other data sources. Second layer is the enterprise layer, which is having functionality components. It includes the functional components for the task of preprocessing, data transfer and communication with the other components [11].

The system provides the predefined templates which consist of data and the patterns possessed by the data, which in turn solve the user queries. These templates are having appropriate data mining algorithm which solves the user query. The input values in terms of parameters required to execute the algorithm are decided by the system itself. Rigorous analysis is done to form the templates. Various functions and operators are defined in the templates. The concept of session management is applied to handle the security issues which are critical in case of the online commercial applications [11].

The system is having limited set of data mining algorithms. As the system is providing the predefined templates services also may in limited number.

III. Conclusion and Future Scope

Multidimensional data analysis requires the cube structure to have the aggregated and summarized data for the analytical operations. Such type of analysis is going to help the organizations to have the improved performance of their systems. Researchers also benefited by the multidimensional analysis to evaluate the techniques and result they obtained. The cube architecture comes with its unwanted side effect of increased time and storage complexity. To avoid this unwanted feature of the cube many techniques are proposed and used by the data analysis system designers. Here we have tried to discuss some techniques which handle the cube technology differently, like parallel execution of cube or handling dynamic cube with the new type of data structure. Also we have discussed some system which uses cube but trying to optimize its performance. Though all the approaches are managing to overcome the drawbacks of cube technology, but increases the operational complexity. The technique is required which can balance these two aspects optimization of cube architecture and minimum operational complexity.

References

[1] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, Neoklis Polyzotis “SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics”, Proceedings of the VLDB Endowment, Vol. 8, No. 13 Copyright 2015 VLDB Endowment 2150-8097/15/09.

[2] Data Modeling Guide, IBM Cognos Analytics Version 11.0.0, Copyright IBM Corporation 2015, 2017.

[3] Sandro Fiore, Alessandro D’Anca, Donatello Elia, Cosimo Palazzo, Ian Foster, Dean Williams, Giovanni Aloisio, “Ophidia: a full software stack for scientific data analytics”, 978-1-4799-5313-4/14/$31.00 ©2014 IEEE

[4] S. Fiorea, A. D’Ancaa, C. Palazzoa,b, I. Fosterc, D. N. Williamsd, G. Aloisioa, “Ophidia: toward big data analytics for eScience”, 2013 International Conference on Computational Science, doi: 10.1016/j.procs.2013.05.409, 2013

[5] Architecture for Enterprise Business Intelligence, an overview of the microstrategy platform architecture for big data, cloud bi, and mobile applications

[6] Usman AHMED, “Dynamic Cubing for Hierarchical Multidimensional Data Space”, PhD thesis, February 2013

[7] Muntazir Mehdi, Ratnesh Sahay, Wassim Derguech, Edward Curry, “On-The-Fly Generation of Multidimensional Data Cubes for Web of Things”, IDEAS ’13 October 09 - 11 2013, Barcelona, Spain

[8] Yang Zhang, Simon Fong, Jinan Fiaidhi, SabahMohammed, “Real-Time Clinical Decision Support Systemwith Data StreamMining”, Hindawi Publishing Corporation Journal of Biomedicine and Biotechnology Volume 2012

[9] Sandra Geisler, Christoph Quix, Stefan Schiffer, Matthias Jarke, “An evaluation framework for traffic information systems based on data streams”, 2011 Elsevier Ltd. All rights reserved.

[10] IBM Cognos Dynamic Cubes, October 2012

[11] Marta Zorrilla, Diego García-Saiz, “A service oriented architecture to provide data mining services for non-expert data miners”, Elsevier 2012.

[12] Dong Jin, Tatsuo Tsuji, “Parallel Data Cube Construction Based on an Extendible Multidimensional Array”, International Joint Conference of IEEE TrustCom-11, 2011.