ITDA: Cube-Less Architecture for Effective Multidimensional Data Analysis

Prarthana A. Deshkar and Parag S. Deshpande

Abstract Recent developments in real-time applications, sensor technology, and various online services are responsible for generating large amount of data which can be used for analysis. Performing multidimensional data analysis on such type of data requires aggregation at various levels which is generally done using data cubes. Generation of data cubes involves lot of storage and time overheads which make such approach practically less feasible if aggregation involves lot of hierarchies in dimensions. The Integrated Tool for Data Analysis (ITDA) project aims to provide a data analytics solution, under single Web-based platform to address the issue of generating the cube for high volume data by proposing the 'on-the-fly aggregation' architecture. This paper presents the ITDA which aims to provide the support for absorption of data, modeling it in multidimensional model, analyzing the absorbed data, and producing effective visualization. Target users can do analysis on their data without relying on costly tools or any prior knowledge in programming. In this paper, detailed architecture of ITDA software with its operating mode is discussed.

Keywords Multidimensional data analysis · Data mining · Cube
Cube-less architecture

P. A. Deshkar (✉)
Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, India
e-mail: Prarthana.deshkar@gmail.com

P. S. Deshpande
CSE Dept., G. H. Raisoni College of Engineering, Nagpur, India
e-mail: psdeshpande@cse.vnit.ac.in

P. S. Deshpande
Computer Science and Engineering, Visvesvaraya National Institute of Technology, Nagpur, India

M. L. Kolhe et al. (eds.), *Advances in Data and Information Sciences*, Lecture Notes in Networks and Systems 38, https://doi.org/10.1007/978-981-10-8360-0_16

1 Introduction

In recent years, multidimensional analytics tools are becoming guide for data researchers as these tools give them cutting edge over their counterparts in marketplace. Due to increased frequency of data generation, data under consideration of analysis is also increasing tremendously. The large size of the data and complexity in data analysis demands an easy platform so that researchers and target users can do analysis on their data without the hard-core knowledge of information technology.

Ad hoc querying or ad hoc reporting is the main need of data analysis. To achieve this, data modeling is essential task if the system wants to facilitate the variety of domains. Multidimensional data modeling is the way to provide facility to perform ad hoc analysis. Analyzing multidimensional data is of growing need to extract the knowledge and hence to enable the decision making in various domains. Data analysis process, which leads to the enhanced decision making, combines various techniques like statistical techniques, data mining algorithms, and machine learning techniques. With all these techniques, presentation of analysis output with attractive visuals is a key part of popular analytics systems. Most of the current multidimensional systems rely on data cubes which are very much resource- and time-intensive. In this context, ITDA architecture is proposed to give the solution for multidimensional analysis with the reduced memory and time overheads as compared to the existing systems. The proposed system is providing analysis without the generation of cube.

Section 2 of this paper explains the preliminary concepts which are used in this paper. Section 3 provides the review of the systems which are already available. Section 4 proposes the new architecture. Section 5 elaborates the operating mode of the proposed architecture. Section 6 discusses the data security and privacy measures taken by the proposed architecture, and the last section is the conclusion of the proposed work.

2 Preliminaries

Multidimensional data is a type of data which talk about the fact which is associated with various entities called dimensions. Measure carries information of operational values such as sales, quantity. It represents the value of parameter on the basis of which multidimensional analysis is performed. Dimensions are informational entities such as product, region, and time, which are used to analyze the data. Generally dimensions have hierarchical structure. Dimensions may have additional properties like sequential relationship or more complex relationship.

For example, region dimension may have hierarchical relationship such as country, state, region, city. Time dimension has hierarchical as well as sequential properties. It can have days, weeks, months, and year as hierarchical levels. Sequential property indicates that each month is having value of the previous month and the next month. These relationships in the data play an important role in the analysis. The hierarchical

Table 1 Sample x values

Dimension	#Unique	Symbol	Value
Product	2	$X1$	CD, PEN DRIVE
Region	1	$X2$	PUNE
Year	2	$X3$	2015, 2016

relationship allows comparing fact values at different levels of hierarchy and can be used to design the market strategy for advertising requirements, supply chain management, to find out market share and so on. Sequential relationship helps to keep track of comparative progress among dimensions and within dimensions. For example, ratio of sales with previous month sales gives us growth over last month. Dependency relationship helps in the future prediction. For example, economists might base their predictions of the annual gross domestic product (GDP) on the final consumption spending within the economy. They use a dependency between GDP and final consumption spending.

Multidimensional data model is having multiple dimensions, and while analyzing it often the data at higher level of dimension is required; e.g., data generated at the seconds or hours level and for analysis data may be required at month, quarter, or year level. Multidimensional data analysis often requires such aggregated data, and hence these aggregations are stored in the form of cube to have faster data retrieval. A cube is used to generate aggregation of multiple dimensions or multiple combinations of multiple dimensions. As the number of rows and columns in the base data table having dimensions and facts increases, the cube generation may suffer from combinatorial explosion. Suppose there are d dimensions each with N unique rows, then the size of cube is given by N_d.

Let x1, x2, x3, … be the number of unique entries in each column except the measure column. In general, the number of rows in cube will be (x1 + x2 + x3 + … + xn) + (x1x2 + x2x3 + … + xnx1) + (x1x2x3 + x2x3x4 + … + xnx1x2) + … + (x1x2 …) + 1. The extra addition of 1 is to include the final aggregation of fact (grand total). The following example represents the same with sample data (Table 1).

In the above example, number of rows in the cube can be calculated as:

$$(x1 + x2 + x3) + (x1x2 + x2x3 + x1x3) + (x1x2x1) + 1 = (2 + 1 + 2) + (2 * 1 + 1 * 2 + 2 * 2) + (2 * 1 * 2) + 1 = 18.$$

Need of Cube-Less Architecture

To perform the analysis in multidimensional environment, cube architecture is used to store the aggregates. Aggregated values facilitate the analysis process and reduce the response time of the system. But forming a multidimensional model for high-speed streaming data and generating data cube for analyzing it is a big challenge from the perspective of space and performance.

Let us take an example table containing seconds data to analyze the feasibility. The table under consideration would contain, say, 3 columns only. One is time column where data is recorded by a sensor which gives out values every second, and the second is the value/reading given by the sensor corresponding to each second. Third might be modes for those of which we are interested to see the value in each second. Even if we consider only 30 days data, it would result in table containing 5,184,000 rows. Further, running cube query on this table for a single aggregation would result in

$$2 * 2,592,000 + 25,920,002 = 6,718,469,184,000$$
$$= 6.72 * 1012, \text{rows.}$$

Hence, using cube architecture one has to wait for months to do some basic processing on such a high granular data. This is one aspect in the generation of cube. One more aspect is all these calculations are based on consideration that we are storing aggregations for only one aggregation function. But in real-life problems, analysts may require aggregations on the multiple aggregation functions like average, minimum, maximum. One more challenge in cube generation is relationship between the dimensions. Consider a scenario where two levels of a hierarchy, product and state, are present in the table. It is possible that multiple products are manufactured in a single state and multiple states manufacture a single product. So there is a many-to-many relationship between product and state. Cube generation has to be done twice, one considering a hierarchy of state under product and the other with a hierarchy of product under state. Again, it is inefficient in both space and time.

The proposed system is the research effort to overcome big challenge in handling the multidimensional stream data while generating the cube. It aims to develop a system which will address the storage overhead of cube architecture for stream data from any domain and will try to incorporate maximum statistical and data mining and machine learning algorithms for analysis. So the main objective of this system is to provide a single platform for multidimensional reporting, statistical processing, data mining, machine learning and visualization. The proposed system is designed on the basis of cube-less architecture, where aggregations are performed at query level and calculated on-the-fly, hence trying to reduce the time and storage overheads which are the main side effects of the cube architecture. The system is targeted expert as well as non-expert data miners.

3 Related Work

The proposed system aims to provide a data analytics, or decision support system works on the huge volume of data which comes from any domain. Many existing commercial products and research projects are working to facilitate the data analysts but with different approaches. Due to advancement in the technology, data generation

speed is increasing exponentially and the communities want the knowledge from this data in very less time so that they can strengthen the decision support system.

Usman AHMED (2013) in the paper [1] proposed the technique to reduce the time required to get the analysis result from large amount of data, rightly said as 'analysis latency is pre-aggregation of data in a cube'. But cubing of data gives rise to problem like complexity in calculations and storage of data. In the same paper, to handle real-time data loading and thus avoiding the storage overhead of cube, they have proposed one approach to create blank tables with the same structure as that of source tables, then data is copied to it, then data is loaded in the data warehouse, and those temporary tables are removed.

B. Janet, A. V. Reddy (2011) in the paper [2] proposed the approach to manage or overcome the storage space issue. The approach is to use the subset of materialized view.

Authors Konstantinos Morfonios, Yannis Ioannidis (2006) in the paper [3] proposes another approach is by avoiding storage of unwanted and redundant aggregations. They have proposed a ROLAP cubing method called Cubing Using a ROLAP Engine (CURE) that computes whole data cube over very large data space constituted of hierarchical dimensions. CURE uses an efficient algorithm for partitioning fact table that helps improving the cube computation speed. One more approach is performing cube construction process in parallel [4].

Sandro Fiore, Alessandro D'Anca, Donatello Elia, Cosimo Palazzo, Ian Foster, Dean Williams, Giovanni Aloisio (2014) in the paper [5, 6] propose the decision support system for big data. Even if the systems are targeting big data as their data source, they are following the traditional OLAP structure to store the multidimensional data, i.e., cube architecture to store the data. Also some programming knowledge is required to customize the analysis, and hence system can focus on very limited set of data mining algorithms.

Many decision support systems are very much problem specific. Hence, the architecture and set of analytical algorithms are specific to that domain only, for example, analysis of text data [4], analyzing the stream data of clinical domain for classification. Zhang et al. [7] proposed 'on-the-fly' cube generation for sensors data [8], for analysis of the traffic data [9].

IBM research team (2008), in document [10], provides a dynamic cube architecture for a very popular commercial analytical system IBM Cognos. It is also creating a cube, using in-memory caching to support large database. They have implemented the approach by storing the once retrieved data from cube in caches, and if required retrieve from caches.

From the above discussion, it seems that the decision support system which handles the storage overhead of cube architecture and not restricted to particular domain is the need. Also it is observed that the systems which are developed carry very specific analytical features, or may require the expert knowledge in the data analysis.

This proposed architecture is addressing the storage overhead by avoiding the generation of cube, and the aggregations are done at query level or on-the-fly. Further optimization of the query structure is also planned to enhance the performance of the query execution. The proposed system is also offering the complete analytical

processing including multidimensional reporting, statistical processing, data mining, machine learning, and visualization under single platform. Also the focus of system is not restricted to any specific domain or problem.

4 Proposed Architecture

The proposed system is basically designed to facilitate the researchers and data analyst with the complete package of multidimensional reporting, statistical processing, data mining, machine learning and visualization. This is achieved by the Web-based system with user-friendly and secure environment for the data analyst. Proposed system is functionally independent; this means it does not require any additional external component or system to complete the task. Also, the components of this proposed system are integrated and there is no need to install any of the components separately, which is often common for most of analytics tools.

Proposed system architecture is mainly divided into two parts, data modeling part and data analysis part. As the system is modeled as a Web application, user can access it as a client and the processing part is handled by the server. Client will act as a data provider, and the processing algorithms are residing at the server side. The main objective of the proposed cube-less architecture is to reduce the time and storage overhead which are the side effects of the cube architecture; hence, data storage on server is avoided (Fig. 1).

Proposed system consists of two main parts containing various components. First is data absorption from different data sources, collection of metadata, and formation

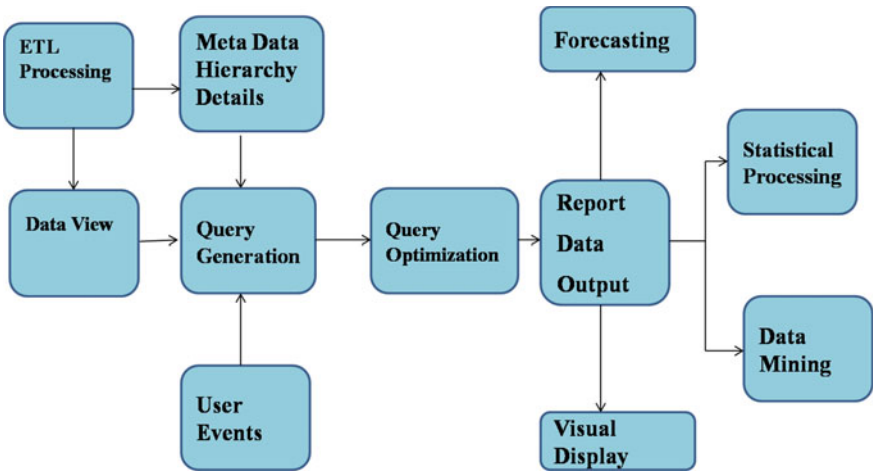

Fig. 1 Cube-less architecture

of multidimensional model, and second is multidimensional analysis on modeled data which further extends to perform statistical analysis and data mining.

Data modeling functionality mainly includes the extraction, transformation, and loading (ETL) process. Source data is given to the ETL process, and it produces the ready to analyze data. ETL process is responsible to extract the data residing on various sources and in variety of formats. It also performs cleansing and customization of data according to the analysis needs. This process is also responsible to generate the metadata of the ready to analyze data. The proposed system is not going to store the data and the aggregations; hence metadata is having crucial role in this system. Aggregations can be generated on-the-fly by using the metadata.

User events are nothing but the data requirements of the analyst to perform the MDDM. Target users of the system can be expert or non-expert data miners, and hence system provides the graphical user-friendly interface to select the data items. System will convert the user selection into user events. With the help of these user events and metadata, system will automatically generate the query. Queries are generated on-the-fly even for the aggregated data. Calculation of aggregated values is done at the runtime. And hence time and storage overhead are avoided by the system.

One query is generated to achieve the value of one cell in the multidimensional output. Though the system achieves the reduction in storage and time overhead, queries are generated in high number. To optimize the number of queries generated, system will further introduce the query optimization engine to have reduced number of queries.

Multidimensional data analysis report is generated by the system. This report is in grid form as well as it provides the attractive visualization techniques to represent the multidimensional output. This output can further be analyzed using different statistical techniques, data mining algorithms, forecasting, and machine learning algorithms. All these components are integrated into the system itself. All these components also provide the reports with grid values and the visual representation.

5 Operating Mode

Expert or non-expert data miners can access the system by creating client users. User will upload the data through these clients. The system is able to extract the data from flat files or from databases. For flat files system supports excel, csv, and txt file formats; for databases it supports MS-Access and Oracle formats. Using the graphical user-friendly interface, user can perform transformations on data to achieve the customization according to organization needs for analysis. The ETL process ends by generating the metadata of the final customized data prepared after transformation. During this entire process, one log file is maintained to record the activity log of user. It contains all status messages and stack trace in case of some exceptions and errors. During each step of execution of ETL algorithm, logged user first records its status and then performs the step. If there are some bad rows in the given dataset, then the process takes care of putting it to bad file which is also

maintained by the ETL process. Along with these files, metadata information is also stored in the file, which is then used for further analysis. Metadata contains description of the hierarchy of the dimension, formats followed by the data items, database credentials, and customized formulas created by user to analyze the data. While deciding metadata parameters time dimension is handled separately. Time dimension requires the special treatment as it can have sequential as well as normal dimension properties. If it is normal time dimension, then its hierarchy and bounds for uppermost levels can be directly considered; i.e., if uppermost level for time is year, then we need to record lower and upper bound for year so that we can give interactive interface to user for time dimension data selection during analysis. If it is sequential dimension, then it is needed to have lower and upper bound for each level. In the terminology of ITDA, the proposed architecture, all these files are stored in the 'environment' of the client user. Environment is the conceptual area given to user along with the multidimensional model for the single data view. One user may have any number of environments for the absorption of data. Environment is in the form of folder on the client machine. Data is accessed with the help of environment. Because of the environment structure, user can handle the many-to-many relationships present in the dimension, separately. At the time of implementation, the metadata information which is stored in file is passed to server in a form having smaller grammar and easy to map with the data structures used by the modern programming language. Only transfer of metadata instead of complete data will reduce the transfer and storage overhead.

After creation of the environment, now system is made ready to perform the analysis. To generate the multidimensional output, user needs to provide the data which is to be analyzed. In ITDA terms, these user requirements for analysis are termed as user events. The proposed system provides different mathematical facilities to prepare the data for analysis. Users are allowed to specify the aggregation function on which data aggregation is required. User is free to select more than one aggregation function at a time as the aggregations are not going to be stored in the system, so no need to worry about time and storage required for the aggregations. Along with the aggregation functions, the system provides various filters to filter the data which is to be analyzed.

Rank Filter Rank filter provides an easy way to add inputs based on the rank of input values.

Measure Filter Measure filter is similar to rank filter. Here, the user would want to select inputs based on the measure value's range instead of rank value.

In the multidimensional report, along with the aggregated values for dimension values, user also can have some analytical functions for the preliminary analysis. The analytical functions which are provided in the system are dense rank, cumulative distribution, ntile, percent rank to one, market share, growth rate, etc.

With the help of all the information provided through the user events, query builder engine generates the queries automatically. Query builder module takes the input selections for each of the dimensions in row and column sequences and builds a single query for each combination.

NR_{di} be the number of selection from dimension *i* present in row sequence.
NCdi be the number of selection from dimension *i* present in column sequence.

$$\text{Total number of queries fired} = \text{p i NRdi X p i NCdi}$$

The query complexity can be given by N X k where N is number of rows in the table and k is number of cells in the output matrix.

Once the row and column sequences are specified, the outputs from the execution of queries are stored in a $R \times C$ matrix where R is product of number of inputs for all dimension specified in row sequence and C is the product of number of inputs for all dimension specified in column sequence.

There is a specific mapping from output matrix to the graph being plotted. The measure is plotted along the *y*-axis. Each one of the rows in the matrix is mapped to *x*-axis. The columns in matrix are mapped as legends which create overlapping plot.

6 Security and Privacy

The proposed architecture is maintaining the security and privacy by giving different privileges to users according to their need. As the proposed system is going to store only the metadata for the environments created by user on the client side, no other user can access the metadata. To maintain the privacy, the system is managed by the authorized login, i.e., admin login. Admin is responsible to allow any user to create their account. No other user can access the data from the other user.

To use this system, user first needs to create the account. According to level of understanding and experience in the data analysis domain, there are four privileges given to user while creating the account. If the user is expert data miner, then the user is having right to upload the data in the system, can create the multidimensional data model by its own, and then can perform multidimensional analysis on it. Professional analysts having data but wants to avoid the technicality of creation of multidimensional model, then they can just upload their data, and admin is going to create the environment for them and notify them after successful creation of environment. Some users can have only right to perform analysis using statistical and data mining algorithms on the environments created and assigned by the admin. Here in this case, user is not having access to source data. And for the non-expert users, there is facility that admin will assign some predefined reports to study the analysis.

Every user can create the environments based on the analysis need. N number of environments can be created for a single dataset. These environments are stored on the client side, hence not accessible to any other user.

7 Conclusion and Future Work

In this work we have presented the system ITDA, a complete solution for data analysis. This model follows the cube-less architecture to overcome the side effects of cube architecture like storage requirement and time to build a cube. In this paper, we talk about the architecture of the system and its components. Also we focus on the operating mode of the system so that system capabilities can be explored. We propose a system which can perform multidimensional reporting, statistical processing, data mining, machine learning and visualization under single platform. In future, the query optimization part of the system needs to be developed so that time required to generate on-the-fly queries can be optimized.

References

1. Ahmed U (2013) Dynamic cubing for hierarchical multidimensional data space. Ph.D. thesis
2. Janet B, Reddy AV (2011) Cube index for unstructured text analysis and mining. In: ICCCS'11, 12–14 Feb 2011, Rourkela, Odisha, India
3. Morfonios K, Ioannidis Y (2006) CURE for cubes: cubing using a ROLAP engine. In: VLDB'06, 12–15 Sept 2006, Seoul, Korea
4. Jin D, Tsuji T (2011) Parallel data cube construction based on an extendible multidimensional array. In: 2011 International Joint Conference of IEEE TrustCom-11
5. Fiore S, D'Anca A, Elia D, Palazzo C, Foster I, Williams D, Aloisio G (2014) Ophidia: a full software stack for scientific data analytics. 978-1-4799-5313-4/14/$31.00 ©2014 IEEE
6. Fiore S, D'Anca A, Palazzo C, Foster I, Williams DN, Aloisio G (2013) Ophidia: toward big data analytics for eScience. In: 2013 international conference on computational science. https://doi.org/10.1016/j.procs.2013.05.409
7. Zhang Y, Fong S, Fiaidhi J, Mohammed S (2012) Real-time clinical decision support system with data stream mining. J Biomed Biotechnol
8. Mehdi M, Sahay R, Derguech W, Curry E (2013) On-the-fly generation of multidimensional data cubes for web of things. IDEAS'13 09–11 Oct 2013, Barcelona, Spain
9. Geisler S, Quix C, Schiffer S, Jarke M (2011) An evaluation framework for traffic information systems based on data streams. Elsevier Ltd. All rights reserved
10. IBM Cognos Dynamic Cubes, Oct 2012