http://dx.doi.org/10.21172/ijiet.93.12

ITDA: An Efficient Analysis And Visualization Solution

Prarthana A. Deshkar[1], Dr. Parag S. Deshpande[2] & Prof. A. Thomas[3]

[1]*Ph.D. scholar, CSE dept. G.H.Raisoni College of Engineering, Nagpur, India*
[2]*Supervisor, CSE dept, G. H. Raisoni College of Engineering, Nagpur, India*
[3]*HoD, CSE department, G. H. Raisoni College of Engineering, Nagpur, India*

Abstract: Multidimensional analysis plays a very important role in any type of decision making system irrespective of the domain. Target users of multidimensional analysis system can be data analysis professionals, business community, researchers, etc. It has been observed that many already existing systems, tools are either expensive or requires much technical expertise. This system is trying to overcome both the challenges. Integrated Tool for Data Analysis (ITDA) offers the web based complete software solution for multidimensional data analysis with interactive interface. Its main advantage is it offers the cube less architecture, which reduces the time and storage overhead, occurred due to cube generation.

Keywords: multidimensional data analysis, visualization, report, cube.

I. INTRODUCTION

Efficient, effective and real time decision making system is the need of current life. Irrespective of the domain, data analysis plays the important role in effective working of the any organization. Because of the advancement in the technology, the data volume is increasing exponentially as well as formats of the data are also changing rapidly. Because of this changing scenario, data analysis needs and target users of data analysis system is also changing.

The data analysis tools available in market are adopting these changes in the market. But it is observed that to use the data analysis tools which are present in the market requires the expertise in technology. Some tools may require a bit of programming knowledge also if user requires the advanced data analysis. Price of data analysis system is again an issue; research community may not always afford such expensive analysis systems. Also many systems does not support the multidimensional analysis, which is the key aspect of modelling if data possess different kind of relationships [2].

In response to all these aspects, the system is formulated named as Integrated Tool for Data Analysis (ITDA). The system follows the cube less architecture and generates the aggregation queries on – the – fly without generation of cube. The system extracts the data, transform it according to the analysis needs and load the data in the system [1].

In response to all these aspects, this work describes the basic functionalities available for multidimensional data analysis using a system, named as, Integrated Tool for Data Analysis (ITDA). The actual results of the basic functionality are discussed here with the help of sample dummy data. The complete discussion includes

ITDA: An Analysis and Visualization Platform

ITDA is a web based project. The design and implementation of this system is focusing on the high end needs of the multidimensional analysis. Simultaneously the design also takes care of minimizing the technical expertise required to perform the data analysis. This case study will explain how ITDA can be used effectively to fulfil the multidimensional analysis needs and efficient visual representation. The complete operating mode of the ITDA can be categorized into several stages as, first stage is; creation of multidimensional data model, which is termed as environment in the ITDA terminology, next stage is utilizing the created model or environment, then generation of user events for analysis, next is generation of reports or visualization of results. The paper is arranged in the same way as these mentioned stages of operating mode of the ITDA.

II. IMPLEMENTATION DETAILS

A. Creation of Multidimensional data model in ITDA

ITDA provides the flexible and interactive interface for modelling the user data. Identification of correct dimensions and fact value is the critical and most important part for any multidimensional data analysis solution. This initial phase of analysis requires the domain knowledge as well as technical expertise. ITDA implemented this phase of population of data in the multidimensional model through the process of extraction transformation and load [1].

The complete process is divided into main two stages like, proper selection of data required for analysis. This may include selection of appropriate data and for that ITDA generates the query on the selected data source and allow the user to edit it so that required fields can be accessed. After finalizing the data source, next stage is mapping of dimensional model to the actual data source. This stage will generate the metadata which is further used for the on – the – fly query generation.

ITDA allows user to mention all the characteristics possessed by the dimension in the specific domain. This mapping phase support the special property of time dimension like, hierarchical property and sequential property.

At the end of the creation of model phase, ITDA creates the metadata repository for the specific model called as 'environment' in the ITDA. One user can create any number of environments with the same or different data sources based on the analysis needs.

B. Use environment

The environment which is nothing but the model of the data along with its metadata is further used for the analysis. This ITDA system is based on the principal of data analysis for the users having less or no expertise in information technology. Hence data scientists who are not the technical experts should also use the system for the data analysis.

ITDA provides the user friendly interface to use the environment; user can edit it and also can update the environment.

C. User Event Generation

Data analysts are not always suppose to write the analytical queries to perform the multidimensional analysis and doing it repeatedly may degrade the analysis performance. ITDA provides the interactive and easy way to handle this. ITDA system will automatically generate the query based on the user requirements which are termed as user events.

ITDA gives the tree selection UI as dimension possesses the hierarchical structure. The basic selection method is provided by lazy load as the dimension data can be further organized as hierarchical level and hierarchy can be extended at any level of granularity. All dimensions are initially loaded at root level in a collapsed state and user can populate the dimension tree level by level on clicking expand at each level. Figure 1 shows the initial interface for generation of user events.

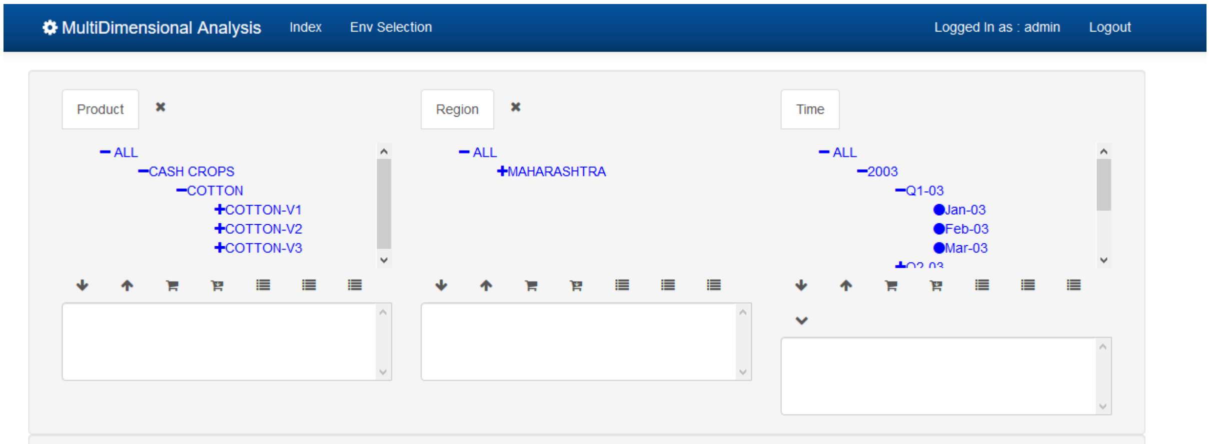


Figure 1. ITDA interface to generate user events

This interface provides features so that user can easily select and navigate the data. Consider the data for cotton seeds which is a cash crop collected for one year from the Maharashtra region. The dimension which stores information of cotton is having hierarchy of level 4, named as attributes, crop type, crop, variety, and item. The dimension region is of level 3 and time is represented as hierarchical dimension of level 3. All the dimensional hierarchy can be shown as figure 2.

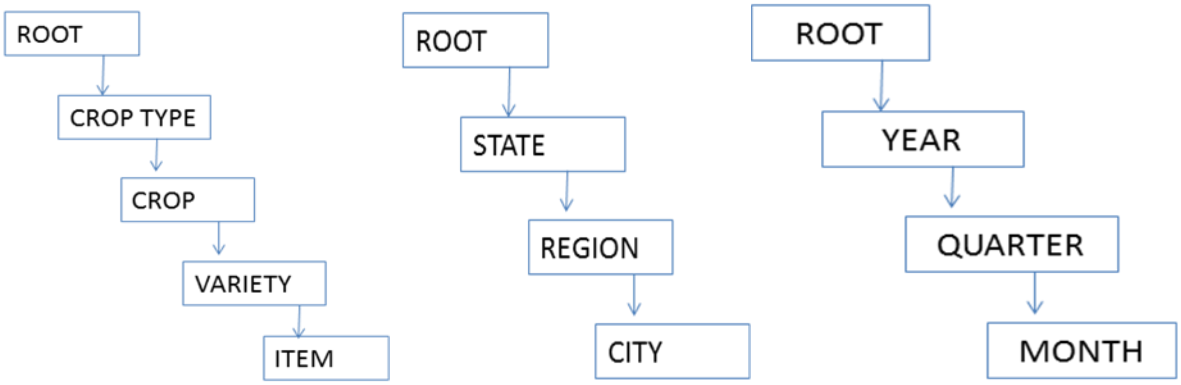


Figure 2. Hierarchical tree structure of dimensions

Based on the analysis need user will select the level of dimension and the actual value which is required. There is no restriction on the number of values user can select. To handle the complex and long hierarchy structure, this interface provides some features like extraction of all children of the selected parent node, extraction of all the elements from the same level of hierarchy, etc.

This interface gives special treatment to the time dimension. Some functions are provided to help the analyst. The functions are:

QTD Quarter-to-date measure will calculate cumulated measure, say sales, from the start of required quarter to that month. Required quarter is the quarter to which the given month belongs. For example, Quarter-to-date(May-03) will give aggregated measure, say value, of the months April-03 and May-03 since April is the start of quarter to which May belongs.

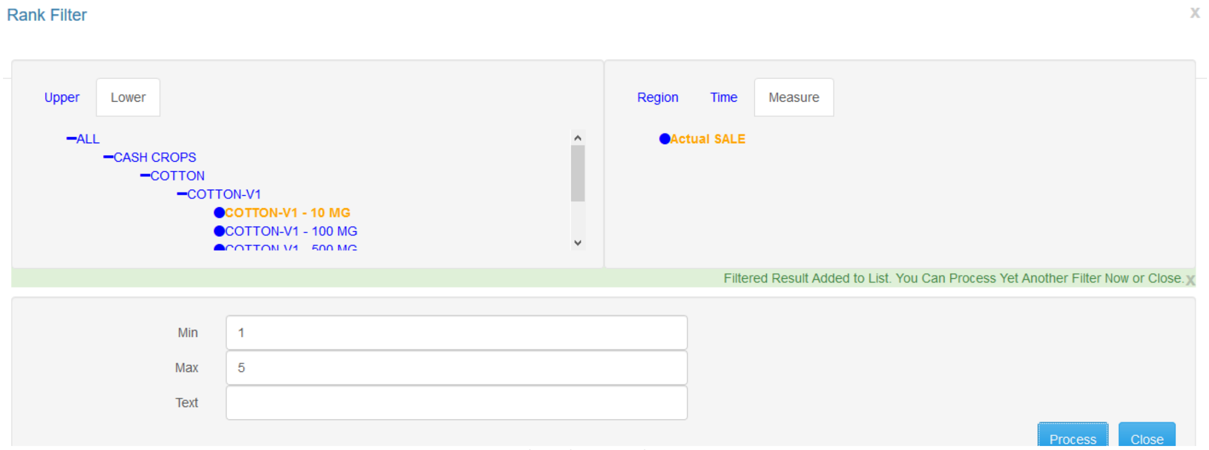
YTD Year-to-date measure will calculate cumulated measure, say value, from the start of required year to that given month. Required year is the year to which the given month belongs. For example, Year-to-date(May-03) will give aggregated measure, say sales, of the months Jan-03 up to May-03 since Jan-03 is the start of year to which May-03 belongs.

QRS Quarter-rolling-sum measure will calculate cumulated measure, say value, from the start of previous quarter up to given month. Previous quarter is taken with respect to the given month. For example, Quarter-rolling-sum(Jul-03) will give aggregated measure, say value, of the months April-03 and Jul-03 since April is the start of previous quarter with respect to May's quarter.

YRS Year-rolling-sum measure will calculate cumulated measure; say sales, from the start of previous year up to given date. Previous year is taken with respect to the year of the given date/month. For example, year-rolling-sum(Jul-03) will give aggregated measure, say sales, of the months Jun-02 up to Jul-03 since Jun-02 is the start of previous year with respect to Jul-03, a difference of 11months + present month of consideration.

While selecting the dimension data analyst can filter the selection to have the effective analysis. There are two ways to filter the data. One is termed as rank filter and other is measure filter.

Rank Filter allows the selection of the dimension values from the whole based on the rank of the fact vale. The interface is shown in the figure 3.



−ALL
−CASH CROPS
−COTTON
−COTTON-V1
●COTTON-V1 - 10 MG
●COTTON-V1 - 100 MG
●COTTON-V1 - 500 MG

●Actual SALE

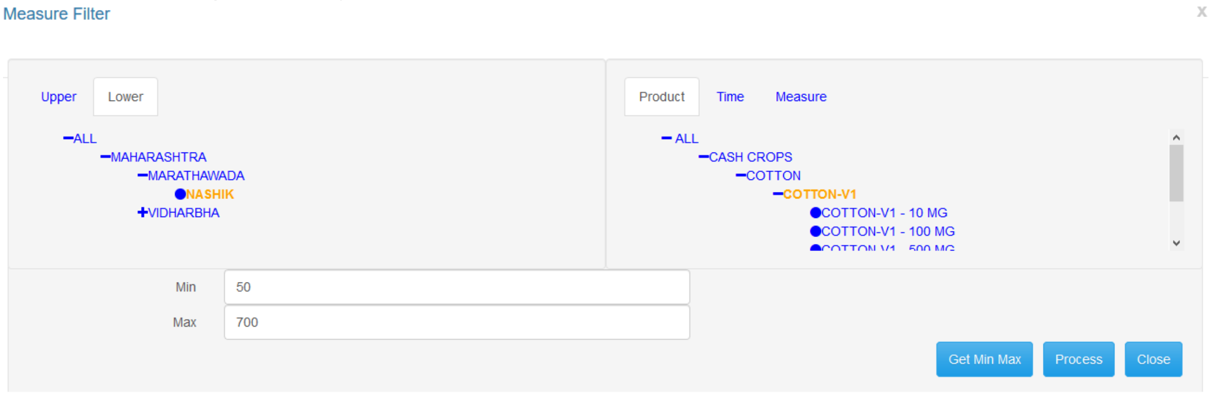
Filtered Result Added to List. You Can Process Yet Another Filter Now or Close. x

Min 1
Max 5
Text

Figure 3. Interface for rank filer

For example user will select the product dimension data from the 4th level whose measure value, i.e. sale rank from 1 to 5 for the Nagpur region.

Measure Filter is similar to rank filter. In this case user is going to select the input value based on measure value's range instead of rank. Figure 4 shows the interface for the measure filter. For example, Let's consider product dimension. User would be interested in adding the variety of cotton, i.e.' COTTON-V1' whose fact (sale) value range from 50 to 700 in 'Nashik' region for the year '2003'.



Measure Filter

- ALL
  - MAHARASHTRA
    - MARATHAWADA
      - ●NASHIK
    - +VIDHARBHA

- ALL
  - CASH CROPS
    - COTTON
      - COTTON-V1
        - ●COTTON-V1 - 10 MG
        - ●COTTON-V1 - 100 MG
        - ●COTTON-V1 - 500 MG

Min	50
Max	700

Figure 4. Interface for measure filter

Customized formulas can be used to perform the preliminary analysis of the data. It facilitates with additional analytical query functions such as:

Dense Rank The data can be graded based on ordering of measure value with respect to a given dimension. Dense rank would give same rank for a particular value and there would not be any gap between ranks.

Cumulative Distribution is used to get the fraction of number of rows below the current row, including the current row, according to the rank and the total number of rows.

Ntile(N) Ntile function can be used to calculate quartiles. Based on the rank, the data would be distributed into N buckets.

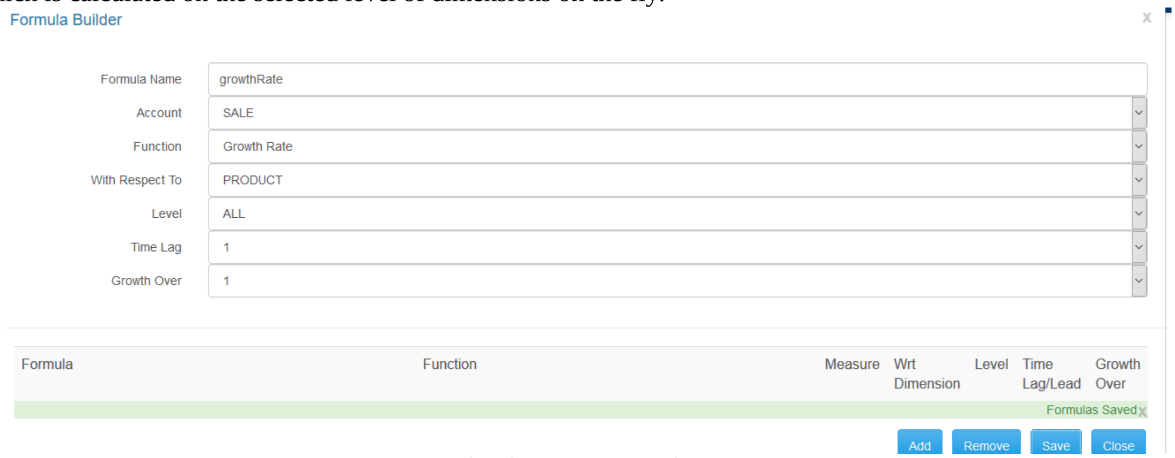
Percent Rank To one for a given dimension, it gives the percentage of corresponding measure value required to reach rank 1. This will be useful to visualize the target to be achieved in a business scenario.

Market Share Market share gives the ratio of measure value for a given dimension compared to its measure value in parent hierarchy.

Growth Rate Provides the percentage growth in the measure value for a particular dimension compared its own value at a previous time.

Each function can be specified as either Fixed or Relative function. Fixed function will take the parent as absolute parent level hierarchy i.e. root level for each dimension. Relative function will take the nth parent from the present level for comparing ranks or ratios.

The value for this customized formula will be calculated run time and displayed in the report along with the measure value which is calculated on the selected level of dimensions on the fly.



Formula Builder

Formula Name	growthRate
Account	SALE
Function	Growth Rate
With Respect To	PRODUCT
Level	ALL
Time Lag	1
Growth Over	1

Formula	Function	Measure	Wrt Dimension	Level	Time Lag/Lead	Growth Over

Formulas Saved

Figure 5. Interface for the customized formula

Aggregate functions: As the system is not storing the aggregations, user is having full flexibility to select any of the aggregation function at the time report generation.

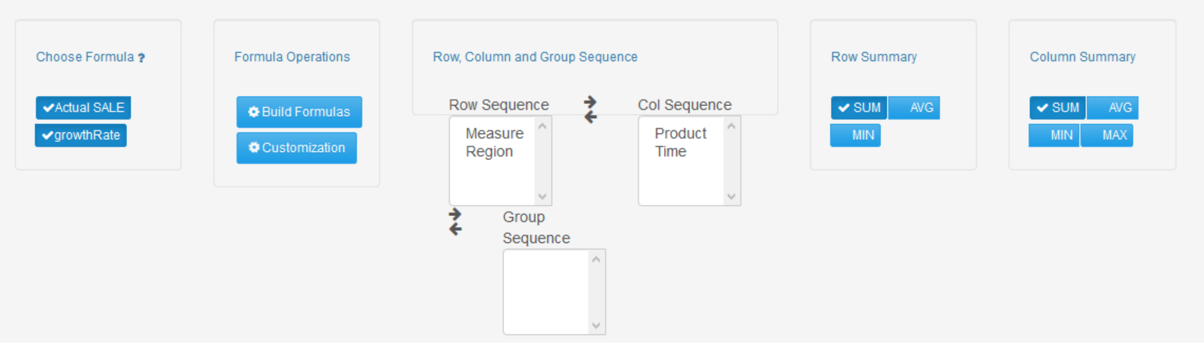


Figure 6. Interface for selection of aggregate function

Display options: Multidimensional analysis report generated by the ITDA can be stored for further use and presentation. Reports can be stored in either csv or html format. User can have graphical representation of the data for effective presentation and to make the analysis more understandable.

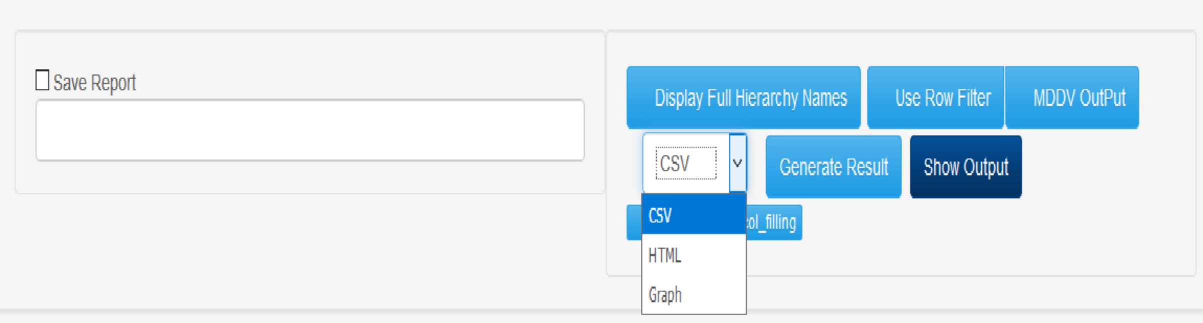


Figure 7. Interface for report visualization options.

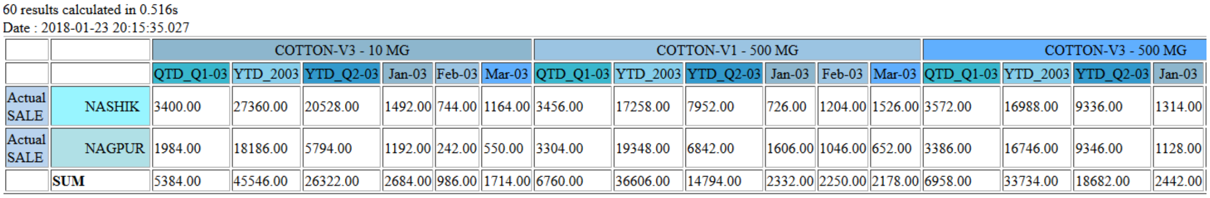
D. *Report generation*

Output matrix generation: Once the row and column sequences are specified, the outputs from the execution of queries are stored in a RxC matrix where;

R is product of number of inputs for all dimension specified in row sequence

C is 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 creates overlapping plot.



60 results calculated in 0.516s
Date : 2018-01-23 20:15:35.027

		COTTON-V3 - 10 MG						COTTON-V1 - 500 MG						COTTON-V3 - 500 MG			
		QTD_Q1-03	YTD_2003	YTD_Q2-03	Jan-03	Feb-03	Mar-03	QTD_Q1-03	YTD_2003	YTD_Q2-03	Jan-03	Feb-03	Mar-03	QTD_Q1-03	YTD_2003	YTD_Q2-03	Jan-03
Actual SALE	NASHIK	3400.00	27360.00	20528.00	1492.00	744.00	1164.00	3456.00	17258.00	7952.00	726.00	1204.00	1526.00	3572.00	16988.00	9336.00	1314.00
Actual SALE	NAGPUR	1984.00	18186.00	5794.00	1192.00	242.00	550.00	3304.00	19348.00	6842.00	1606.00	1046.00	652.00	3386.00	16746.00	9346.00	1128.00
	SUM	5384.00	45546.00	26322.00	2684.00	986.00	1714.00	6760.00	36606.00	14794.00	2332.00	2250.00	2178.00	6958.00	33734.00	18682.00	2442.00

Figure 8. Output representation in html format

Visualization module provides facility to plot different graphs/charts. The chart types include line chart, spline, pie chart, area chart, area-spline, polar, gauge. The plotting is done at the client side avoiding the cost of sending a new chart every time user selects a new chart type. Client doesn't request the server for new chart unless data is changed.

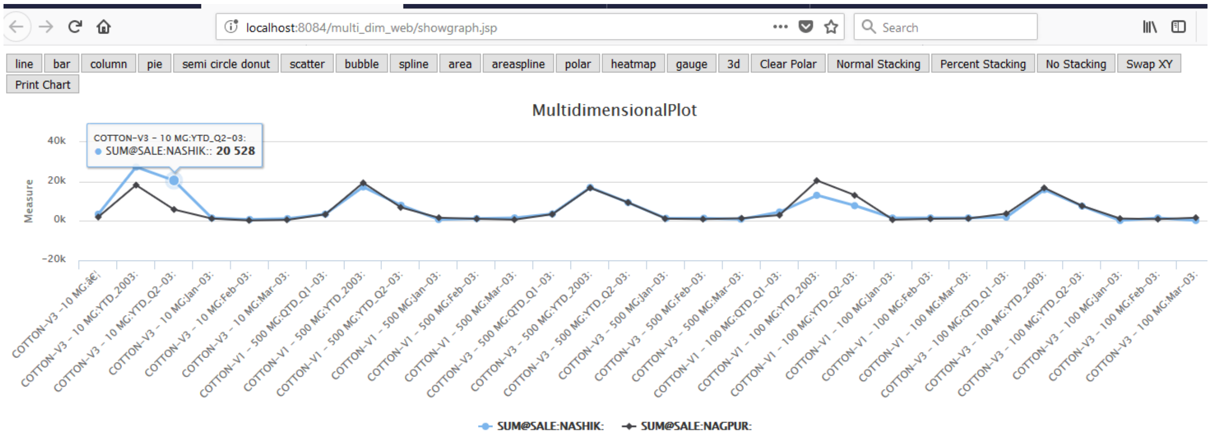

localhost:8084/multi_dim_web/showgraph.jsp
Search
line
bar
column
pie
semi circle donut
scatter
bubble
spline
area
areaspline
polar
heatmap
gauge
3d
Clear Polar
Normal Stacking
Percent Stacking
No Stacking
Swap XY
Print Chart
MultidimensionalPlot
COTTON-V3 - 10 MG:YTD_Q2-03:
SUM@SALE:NASHIK:: 20 528
40k
20k
0k
-20k
Measure
COTTON-V3 -10 MG:â€¦
COTTON-V3 - 10 MG:YTD_2003:
COTTON-V3 - 10 MG:YTD_Q2-03:
COTTON-V3 - 10 MG:Jan-03:
COTTON-V3 - 10 MG:Feb-03:
COTTON-V3 - 10 MG:Mar-03:
COTTON-V1 - 500 MG:QTD_Q1-03:
COTTON-V1 - 500 MG:YTD_2003:
COTTON-V1 - 500 MG:YTD_Q2-03:
COTTON-V1 - 500 MG:Jan-03:
COTTON-V1 - 500 MG:Feb-03:
COTTON-V1 - 500 MG:Mar-03:
COTTON-V3 - 500 MG:QTD_Q1-03:
COTTON-V3 - 500 MG:YTD_2003:
COTTON-V3 - 500 MG:YTD_Q2-03:
COTTON-V3 - 500 MG:Jan-03:
COTTON-V3 - 500 MG:Feb-03:
COTTON-V3 - 500 MG:Mar-03:
COTTON-V1 - 100 MG:QTD_Q1-03:
COTTON-V1 - 100 MG:YTD_2003:
COTTON-V1 - 100 MG:YTD_Q2-03:
COTTON-V1 - 100 MG:Jan-03:
COTTON-V1 - 100 MG:Feb-03:
COTTON-V1 - 100 MG:Mar-03:
COTTON-V3 - 100 MG:QTD_Q1-03:
COTTON-V3 - 100 MG:YTD_2003:
COTTON-V3 - 100 MG:YTD_Q2-03:
COTTON-V3 - 100 MG:Jan-03:
COTTON-V3 - 100 MG:Feb-03:
COTTON-V3 - 100 MG:Mar-03:
SUM@SALE:NASHIK:
SUM@SALE:NAGPUR:

Figure 9. Graphical representation of multidimensional analysis

III. CONCLUSION AND FUTURE SCOPE

The basic functionality for multidimensional analytics in ITDA is described here. The ITDA system follows the cube less architecture. It avoids the space and time overhead of the cube generation process. One of the important motives of this research is the ease of use of the data analytics system. ITDA offers the user friendly and interactive interface which helps the researchers or data analysis professional who are not technical experts. This system also gives effective visualization options, report generation and maintenance options, which makes it a complete multidimensional data analysis system.

Future work is related to adding more and more data mining and machine learning options, so that this system will support the advanced data analytics.

IV. REFERENCE

[1] Prarthana A. Deshkar, Parag S. Deshpande, A. Thomas, "A software Infrastructure for Multidimensional data Analysis: A Data Modeling Aspect", International Journal of Computer Science and Information Security, Volume 16, No. 1, January 2018

[2] Prarthana A. Deshkar, Parag S. Deshpande, A. Thomas, " Multidimensional Data Analysis Facilities and Challenges: A Survey for Data Analysis Tools", International Journal of Computer Applications (0975 – 8887), Volume 179 – No.13, January 2018

[3] 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.

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

[5] 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

[6] 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

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

[8] Usman AHMED, "Dynamic Cubing for Hierarchical Multidimensional Data Space", PhD thesis, February 2013

[9] 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

[10] 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

[11] 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.

[12] IBM Cognos Dynamic Cubes, October 2012

[13] Marta Zorrilla, Diego García-Saiz, "A service oriented architecture to provide data mining services for non-expert data miners"