A Hybrid Data Warehouse Model to Improve Mining Algorithms

Kadhim B.S.AlJanabi
Department of Computer Science
Faculty of CS and Mathematics
University of Kufa
Kadhim aljanabi@uokufa.edu.iq

Rusul Kadhim
Department of ComputerScience
Faculty of CS andMathematic
University of Kufa
rusul.kadhim1992@gmail.com

Abstract:

The performance of different Data Mining Algorithms including Classification, Clustering, Association, Prediction and others are highly related to the approaches used in Data Warehouse design and to the way the data is stored (lightly summarized, highly summarized and detailed).Detailed data is important to get detailed reports but as the amount of data is huge this represents a big challenge to the mining algorithms, on the other hand, the summarized data leads to better algorithms performance but the lack of the required knowledge may affect the overall mining process.

Knowledge extraction and mining algorithms performance and complexities represent a big challenge in data analysis field, hence the work in this paper represents a proposed approach to improve the algorithms performance throughout well designed warehouse and data reduction technique.

The work in this paper presents a hybrid warehouse galaxy model that stores data in three different formats including detailed, summarized and highly summarized data. The time and space complexity are the major criteria in the proposed approach.

Real data was collected about schools, students and teachers from different AlNajaf AlAshraf cities, the data was preprocessed, reduced mainly through concept hierarchy and then converted into dimensions and fact tables (Warehouse Galaxy Model) which in turn are converted into multidimensional cubes. Roll up and drill down queries were highly used to get the required information.

The resultant data cubes and in turn the corresponding warehouse model presented in this work showed a reasonable improvement in knowledge extraction algorithms for the data under discussion.

The results of the queries showed better performance of different roll up and drill down queries compared to detailed data queries.

Keywords: Data Warehouse, Data Cube, Data Mining, Summarization, Data Reduction

1. Introduction

Building good Data Warehouse(DW) will lead to good reports extracted using Data Mining(DM) algorithms and techniques. A DW can be of different models and types, it may be RDMS designed specifically to meet the needs of transaction processing systems with some modification. It can be loosely defined as any centralised data repository which can be queried for business benefit. Data warehousing can be expressed as is a new powerful technique making it possible to extract historical data and overcome inconsistencies between different data formats. DW can be used to integrate data throughout an organization, regardless of data format, site or location, or communication requirements it is possible to incorporate additional or expert information[1,2,3,4].

In other words the DW provides data that is already transformed from one format into another

and summarized, therefore making it an appropriate environment for more efficient Decision Support Systems applications.

According to Bill Inmon, author of Building the DW, he specified the characteristics of DW in four different concepts[1,2]:

- DW is Subject-Oriented: In operation systems, data is organized according to the required transactions and application, whereas in DW data are organized according to subject. When data is organized according to the subjects, this means that it contains only the information necessary for DS processing.
- DW is Integrated: Data is integrated from different sites, locations and operations systems. Inconsistency represents a big challenge in DW modeling due to the origin of data.
- DW is Time-Variant: DW is a repository for storing historical data that are five to 10 years old, or older, to be used for comparisons, trends, and forecasting. These data is not updated, new data can be added to the warehouse.
- DW is Non-Volatile: Data can be transformed from one format to another, or it can be summarized in different ways but is not updated or changed once they enter the DW, but it can only loaded, accessed and processed.

Data cleansing represents an important issue of creating an efficient DW in that it is the removal of certain aspects of operational data, such as low-level transaction information, which slow down the query times when data is huge. Data should be extracted from operational sources at prespecified intervals and stored centrally but the cleansing process has to remove duplication and manage the differences between various styles of data collection[1, 5, 6,7].

Data Mining (DM), sometimes called knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful knowledge. This knowledge can be used in different levels of decision making to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to preprocess and analyze data from many different dimensions, classify, categorize cluster and identify the relationships[2,7,8].

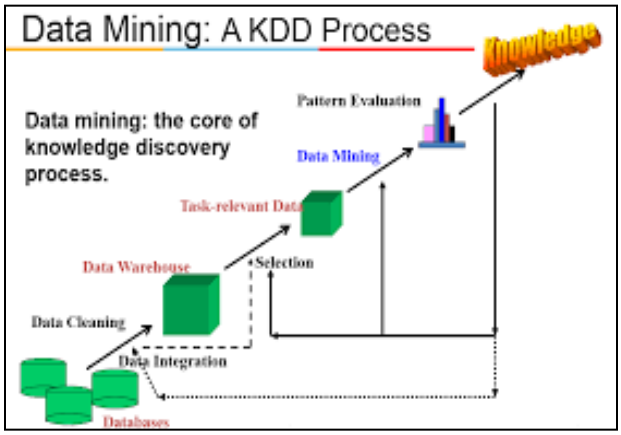
In DM field, information represents an important term. For example, analysis of student achievements at the university like associations between achievements and educational environment data can provide *information*. On the other hand when the information can be converted into another higher level of representation, this in called knowledge. For example, summary information on students outcomes can be analyzed to get knowledge about the attributes highly required in getting jobs after graduation. Figure(1) shows the different levels of these concepts.

Figure(1). Different Data levels With Values and Volumes[1,2].

Dramatic changes in data capturing, processing capabilities, data transmission through networks and medias, and storage capabilities are enabling organizations to integrate their various databases into *DW*. Data warehousing is the process performed on centralized data management and data retrieval. The new and dramatic emerging technologies in information and communication technologies help in the adaptation of DW in different organization strategies [2,3, 10,11].

DM techniques can be classified into five different fields; classification, clustering, association, prediction and link analysis. Each of these techniques has its own algorithms and applications in business context. Mining employment data set represents a crucial factor for different organizations in both public and private sectors. Many works have been carried out in this field.

The whole KDD and DM process is shown in figure(2).



Figure(2). Knowledge Discovery in Databases (KDD) Process[1,2].

2. Data Collection

Real data containing tens of tables each with hundreds or even thousands records about schools, students and teachers were collected. Tables I and II show sample of records from the logged data.

3. Problem Statement

As the collected historical data becomes huge, both the storage and applied algorithms performance represent a big problem in the whole knowledge extraction. Data reduction approach through warehouse design and data summarization is applied to get better knowledge extraction algorithms performance in both time and space complexities.

Most of data mining algorithms highly depend on number of records in the data set, number of attributes representing the data and number of distinct values in each attribute.

Given a dataset $R=\{R_1, R_2, R_3, \ldots, R_n\}$ where n is the number of records in the dataset(e.g. number of transactions), and let $A=\{A_1, A_2, \ldots, A_m\}$ where m is the number of attributes(fields or columns), and let $A_1=\{a_{11}, a_{12}, \ldots, a_{1k}\}$, $A_2=\{a_{21}, a_{22}, \ldots, a_{2t}\}$, $A_m=\{a_{m1}, a_{m2}, \ldots, a_{mr}\}$ where $a_{11}, a_{12}, \ldots, a_{1k}$represent the distinct values in A_1.

Time complexity of different mining algorithms applied on the dataset and data cube operations(roll up and drill down) representing by Big O notation is a function of n, m, k, t, r and so on.

The work in this paper tends to reduce(summarize) the dataset by applying concept hierarchy which tends to reduce the number of distinct values in the attributes and hence reducing the number of records in R.

4. Proposed Approach

A data cube with *m* attributes(fields) can be stored as an array with m dimensions. Each element of the array contains the measure value(s), such as count, total amount, number of students, and number of schools. The array itself can be represented as a 1-dimensional array. For example, a 2-dimensional array of size (*x* x *y)* can be stored as a 1-dimensional array of size *x**y*, where element (*i*,*j*) in the 2-D array is stored in location (*y***i*+*j*) in the 1-D array. The disadvantage of storing the cube directly as an array is that most data cube elements contain zero elements(sparse), so the array will contain many empty elements (zero values)[5,6,8,9,10].

The proposed approach is shown in the following algorithm:

Proposed Algorithm

Step 1: The data from different schools and cities was collected.
Step 2: The collected data was converted into subject oriented tables (i.e. Schools, Students,..)

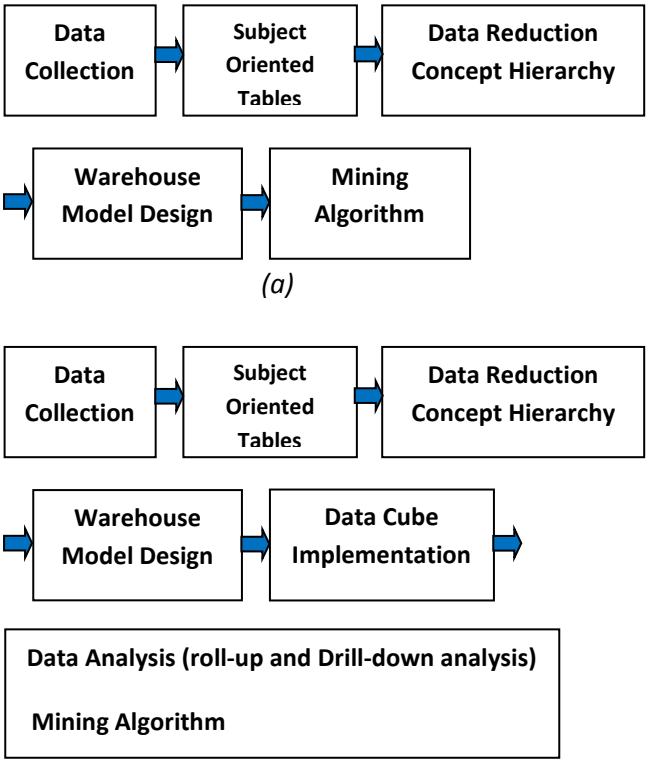
Step 3: Data reduction technique using Concept Hierarchy was applied to get reduced data sets.

Step 4: The reduced data from has been converted into dimensions and fact tables (Warehouse Galaxy Model).

Step 5: Creating the required data cubes

Steps 6: Applying Roll up and Drill down analysis and also mining algorithms.

Step 7:Analyze the results.



Figure(3).Phases of . (a) Traditional approach, (b) proposed approach

Phase one:

As the complete data was collected, it was then converted into tables of clear subjects(Subject oriented) each with feasible attributes and attribute types as shown in tables I and II.

Table(I). Dimensions and Keys(Detailed) in a, b and c .

School Type key	*School Type Description*
1	Primary
2	Intermediate
3	Preparatory
4	Secondary
5	Preparation Teachers Institute
6	Institute of Fine Arts
7	Professional industrial
8	Professional agricultural
9	Professional commercial
10	Professional applied arts
11	Professional Computer and Information Technology

(a)

Specialization key	*Specialization*
1	Islamic education
2	Arabic language
3	English language
4	Kurdish language
5	Mathematics
6	Geography
7	History
8	Biology
9	Social
10	Physics
11	Chemistry
12	Public
13	Science
14	Sociology
15	Management

(b)

School Level key	*Level*
1	First primary
2	Second primary
3	Third primary
4	Fourth primary
5	Fifth primary
6	Sixth primary
7	Seventh Intermediate
8	Eighth Intermediate
9	Ninth Intermediate
10	Fourth scientific
11	Fourth literary
12	Fifth scientific
13	Fifth literary
14	Sixth scientific
15	Sixth literary
16	First Institute
17	Second Institute
18	Third Institute
19	Fourth Institute
20	Fifth Institute
21	First professional
22	Second professional
23	Third professional

(c)

Table(II). Samples of 20 Records for Detailed Fact Tables(fact 1, fact 2 and fact 3 in a, b and c respectively).

	Year key	City key	School Type key	Gender key	Level key	number of students	number of Passed students	number of failed students	number of cancelled students
2	1	1	1	1	1	12005	9010	2301	518
3	1	2	1	1	1	1221	943	226	39
4	1	3	1	1	1	7	6	1	0
5	1	4	1	1	1	3922	2876	623	87
6	1	5	1	1	1	1745	1295	346	60
7	1	6	1	1	1	668	460	148	11
8	1	7	1	1	1	722	511	112	15
9	1	8	1	1	1	1608	1163	280	53
10	1	9	1	1	1	1634	1265	292	32
11	1	10	1	1	1	927	685	183	23
12	1	1	1	2	1	11121	8208	1778	271
13	1	2	1	2	1	959	723	150	40
14	1	3	1	2	1	8	6	0	0
15	1	4	1	2	1	3594	2680	480	62
16	1	5	1	2	1	1498	1257	267	44
17	1	6	1	2	1	600	422	99	16
18	1	7	1	2	1	658	517	84	14
19	1	8	1	2	1	1489	1012	283	154
20	1	9	1	2	1	1453	1058	241	26
21	1	10	1	2	1	763	617	135	25

(a)

1	Year key	City key	School Type key	School Gender key	number of schools	number of students	number of teachers
2	1	1	1	1	99	58968	2443
3	1	2	1	1	6	2167	94
4	1	3	1	1	0	0	0
5	1	4	1	1	47	18172	1130
6	1	5	1	1	12	4145	199
7	1	6	1	1	6	1860	82
8	1	7	1	1	8	3266	195
9	1	8	1	1	12	5267	210
10	1	9	1	1	21	6277	392
11	1	10	1	1	10	2625	145
12	1	1	1	2	101	52552	2622
13	1	2	1	2	5	1799	113
14	1	3	1	2	0	0	0
15	1	4	1	2	42	16008	1056
16	1	5	1	2	11	3544	181
17	1	6	1	2	4	1462	72
18	1	7	1	2	9	2856	232
19	1	8	1	2	12	4889	231
20	1	9	1	2	20	5393	396
21	1	10	1	2	8	2008	144

(b)

1	Year key	City key	School Type key	Specialization key	Gender key	number of teachers
2	1	1	1	1	1	97
3	1	2	1	1	1	48
4	1	3	1	1	1	2
5	1	4	1	1	1	60
6	1	5	1	1	1	69
7	1	6	1	1	1	20
8	1	7	1	1	1	5
9	1	8	1	1	1	50
10	1	9	1	1	1	56
11	1	10	1	1	1	49
12	1	1	1	1	2	286
13	1	2	1	1	2	33
14	1	3	1	1	2	0
15	1	4	1	1	2	117
16	1	5	1	1	2	47
17	1	6	1	1	2	16
18	1	7	1	1	2	40
19	1	8	1	1	2	43
20	1	9	1	1	2	49
21	1	10	1	1	2	22

(c)

Phase two:

Concept hierarchy is applied on the data to get reduced and summarized version of the data as shown in figure(4).

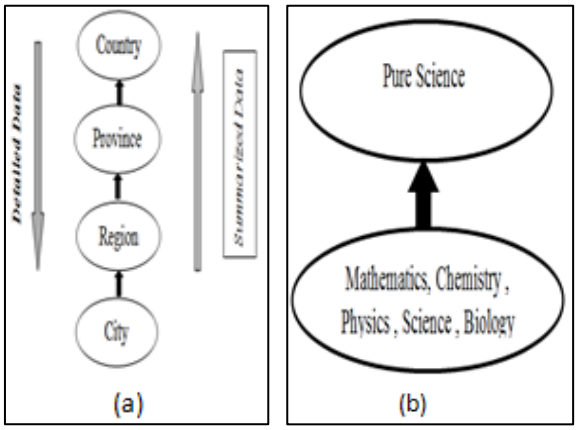


Figure (4). Concept Hierarchy. (a) Location Hierarchy. (b) Specialization Hierarchy

Different dimensions, keys and key descriptions for the summarized data are shown in table III.

Table(III). Attribute Summarization. (a) School Type, (b) Specialization, (c) School Level.

School Type key	School Type Description
1	primary
2	Intermediate
3	Preparatory
4	Secondary
5	Institutes
6	professional

(a)

Specialization key	Specialization
1	Islamic education
2	Languages
3	Pure Science
4	Humanities and Social Sciences
5	Art education
6	physical education
7	Arts
8	Agricultural
9	Industrial
10	Commercial
11	Other

(b)

School Level key	Level
1	First primary
2	Second primary
3	Third primary
4	Fourth primary
5	Fifth primary
6	Sixth primary
7	Seventh Intermediate
8	Eighth Intermediate
9	Ninth Intermediate
10	preparatory Fourth
11	preparatory Fifth
12	preparatory Sixth
13	First Institute
14	Second Institute
15	Third Institute
16	Fourth Institute
17	Fifth Institute
18	First professional
19	Second professional
20	Third professional

(c)

Samples of 20 records for the summarized dataset are shown in table(IV).

Table(IV). Samples of 20 Records for Summarized Fact Tables

	Year key	City key	School Type key	Gender key	Level key	number of students	number of Passed students	number of failed students	number of cancelled students
2	1	1	1	1	1	13233	9959	2528	557
3	1	2	1	1	1	6335	4631	1117	158
4	1	3	1	1	1	4891	3624	867	123
5	1	1	1	2	1	12088	8937	1928	311
6	1	2	1	2	1	5692	4359	846	122
7	1	3	1	2	1	4363	3204	743	219
8	1	1	1	1	2	11585	9391	1812	250
9	1	2	1	1	2	5642	4557	872	82
10	1	3	1	1	2	4415	3627	737	60
11	1	1	1	2	2	10367	8641	1205	192
12	1	2	1	2	2	4986	4187	612	82
13	1	3	1	2	2	3816	3104	530	200
14	1	1	1	1	3	10982	8857	1615	217
15	1	2	1	1	3	5442	4154	791	85
16	1	3	1	1	3	4149	3262	582	66
17	1	1	1	2	3	9735	8260	954	163
18	1	2	1	2	3	4652	3744	483	88
19	1	3	1	2	3	3523	2799	411	152
20	1	1	1	1	4	10305	8022	1622	252
21	1	2	1	1	4	5023	3831	843	116

(a)

	Year key	City key	School Type key	School Gender key	number of schools	number of students	number of teachers
2	1	1	1	1	105	61135	2537
3	1	1	1	2	106	54351	2735
4	1	1	1	3	27	7475	390
5	1	2	1	1	65	24177	1411
6	1	2	1	2	57	21014	1309
7	1	2	1	3	41	13034	564
8	1	3	1	1	51	17435	942
9	1	3	1	2	49	15146	1003
10	1	3	1	3	45	11555	624
11	2	1	1	1	110	63097	2568
12	2	1	1	2	109	56227	2758
13	2	1	1	3	30	8584	451
14	2	2	1	1	64	25430	1407
15	2	2	1	2	58	22101	1328
16	2	2	1	3	42	13914	575
17	2	3	1	1	52	18284	964
18	2	3	1	2	50	15988	1006
19	2	3	1	3	44	11507	594
20	3	1	1	1	112	64704	2560
21	3	1	1	2	111	59777	2814

(b)

	Year key	City key	School Type key	Specialization key	Gender key	number of teachers
2	1	1	1	1	1	147
3	1	2	1	1	1	149
4	1	3	1	1	1	160
5	1	1	1	1	2	319
6	1	2	1	1	2	180
7	1	3	1	1	2	154
8	1	1	1	2	1	279
9	1	2	1	2	1	247
10	1	3	1	2	1	279
11	1	1	1	2	2	761
12	1	2	1	2	2	459
13	1	3	1	2	2	302
14	1	1	1	3	1	177
15	1	2	1	3	1	166
16	1	3	1	3	1	176
17	1	1	1	3	2	566
18	1	2	1	3	2	275
19	1	3	1	3	2	215
20	1	1	1	4	1	57
21	1	2	1	4	1	66

(c)

And the overall attributes distinct values after reduction are shown in table(V)

Table(V). Attribute Distinct Values.

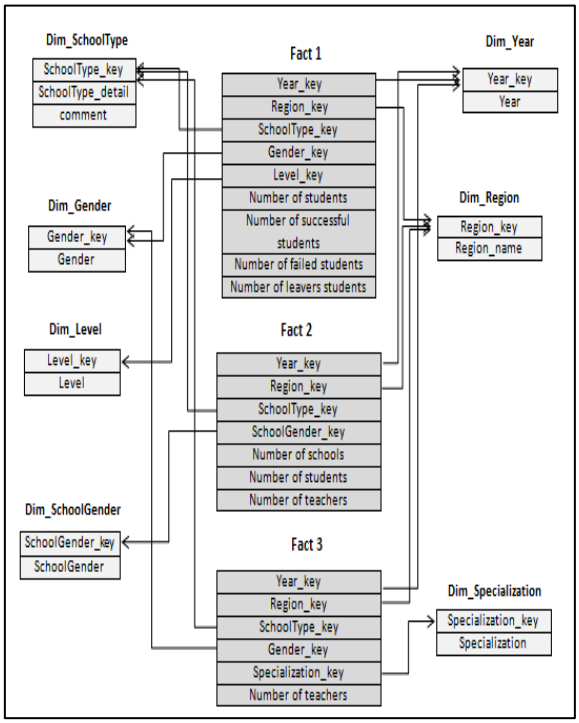
Attribute	Number of distinct values	
	Before Summarization	After Summarization
City	10	3
School Type	11	6
Specialization	75	11
Level	23	20

Attribute	Number of distinct values
Year	6
Gender	2
School Gender	3

Phase three:

Data Warehouse Galaxy model is used in this work since it reflects the historical data in real world and can be used easily to enforce data ownership and security. The model consists of three fact tables containing the required keys and

measures and seven dimensions containing the detailed key information as show in figures (5)



Figure(5). Proposed Data Warehouse Galaxy Model.

where Fact1 contains the measures(number of students, number of passed students, number of failed students and number of cancelled students), Fact2 contains the measures(number of schools, number of students and number of teachers) and Fact3 contains the measure (number of teachers).

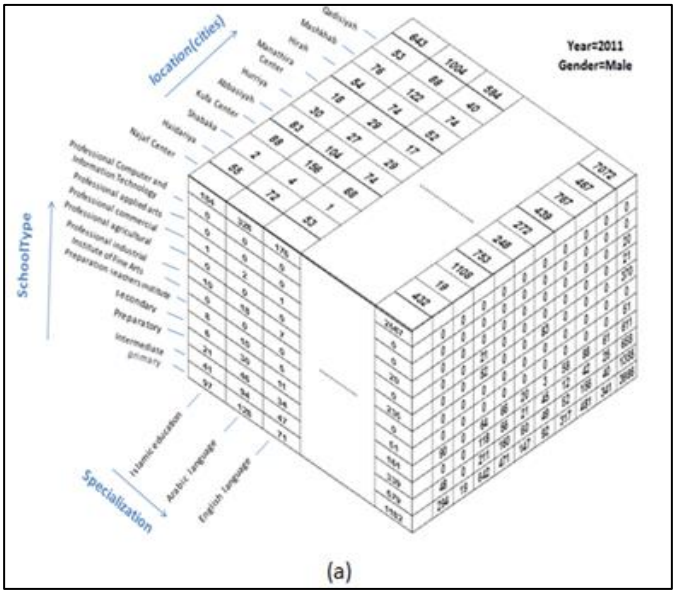
Phase four:

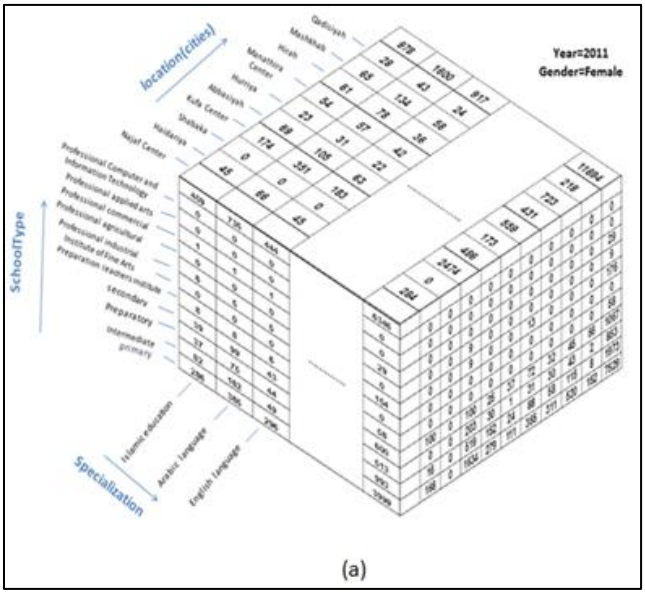
In this phase, all the tables (Dimensions and facts) were converted into data cubes. It is obvious that the minimum number of cube dimensions is three (x, y, z), however in the fact tables show in figure(5) the numbers of dimensions required for Fact1, Fact 2 and Fact 3 are 5, 4 and 5respectively. The number of cubes depends on the number of distinct values of each dimension. The three dimensions with larger number of distinct values represent x, y and z for the cubes and the dimension(s) with minimum number of distinct values represents the number of cubes, as shown in table (VI).

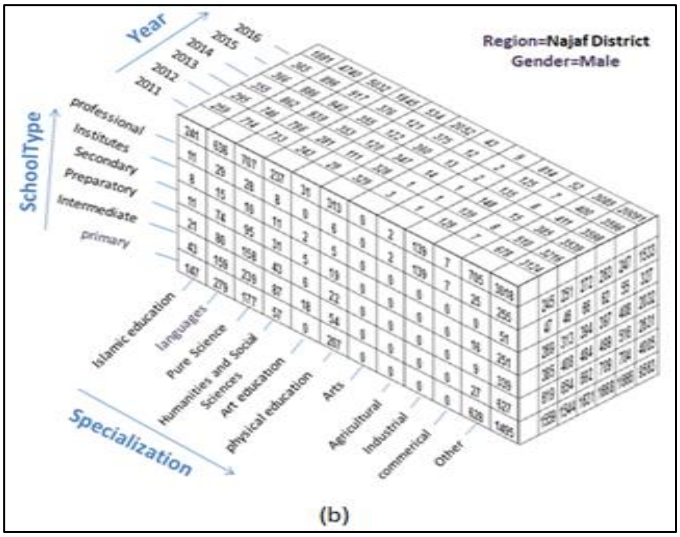
Table(VI). Number of Cubes Representing the Fact Tables Before and After Reduction.

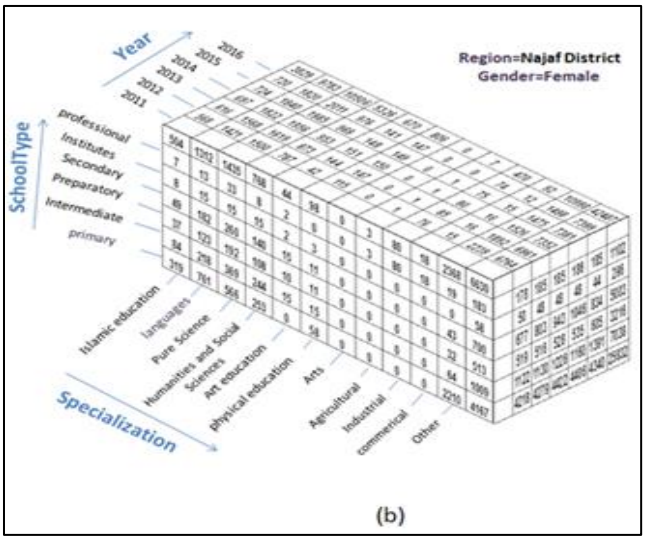
Fact Table	Dimensions	Number of Distinct values		Number of Cubes		Dimensions representing x, y, z	
		Before Reduction	After Reduction	Before Reduction	After Reduction	Before Reduction	After Reduction
Fact 1	Year Region School Type StudentGender Level	6 10 11 2 23	6 3 6 2 20	12	6	Region, School Type, Level	Year, School Type, Level
Fact 2	Year Region School Type School Gender	6 10 11 3	6 3 6 3	3	3	Year, Region, School Type	Year, Region, School Type
Fact 3	Year Region School Type Teacher Gender Teacher Spec.	6 10 11 2 75	6 3 6 2 11	12	6	Region, School Type, Teacher Specialization	Year, School Type, Teacher Specialization

different data cubes were mentioned here representing different keys and measures are shown in figures 6 and 7.





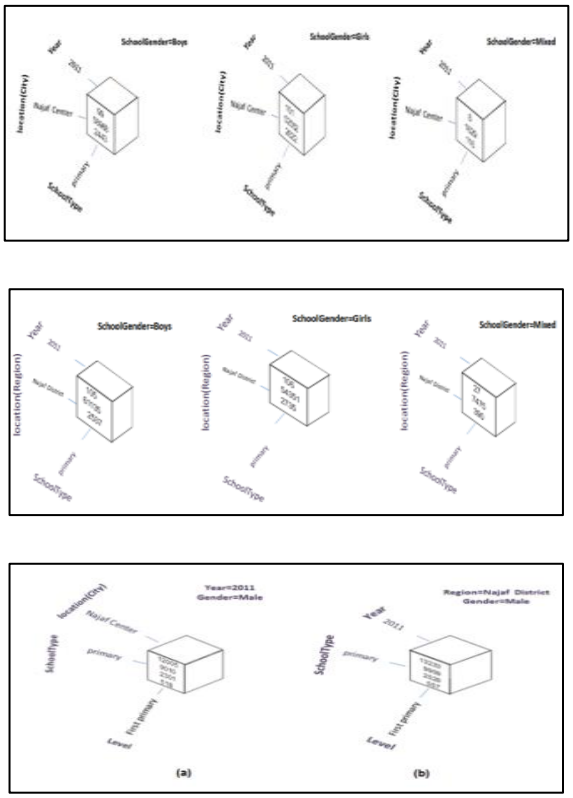




Figure(6). Data Cube Representation of the Warehouse Model for Male Gender.(a) Detailed data, (b) Reduced Data

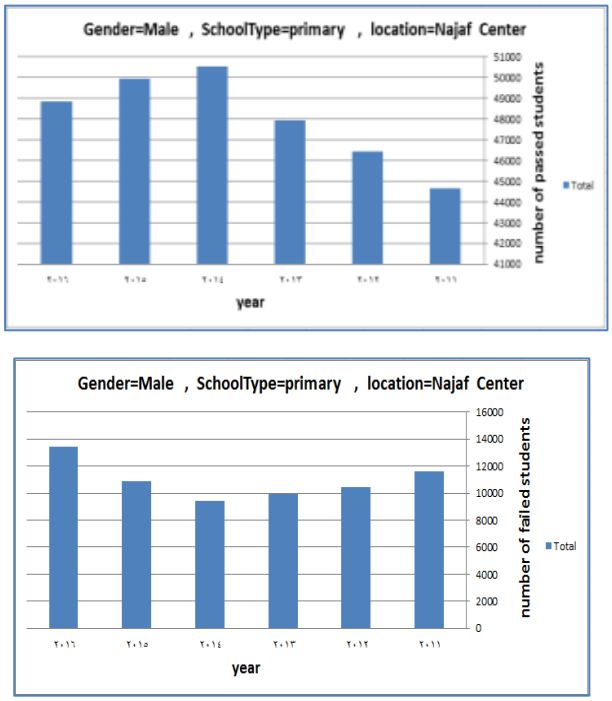
Figure(7). Data Cube Representation of the Warehouse Model for Female Gender. (a) Detailed data, (b) Reduced Data.

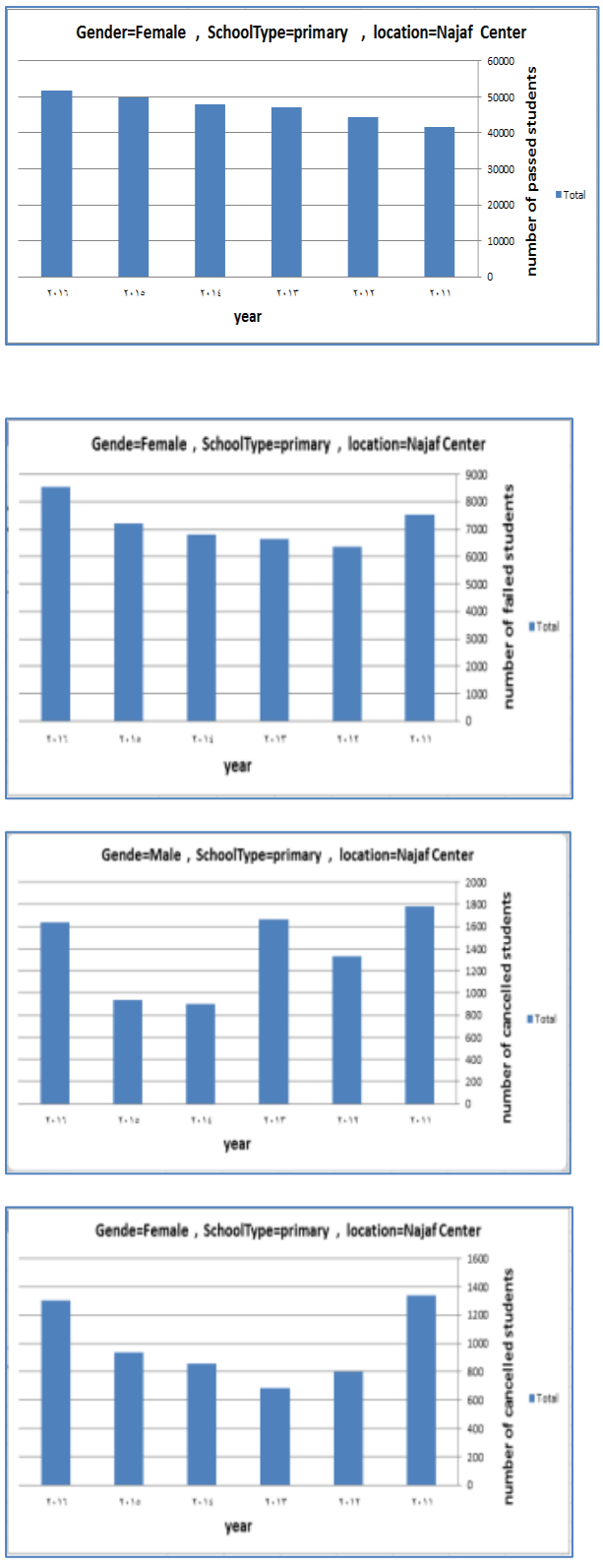
The data cubes shown in figure 6 and 7 represent a single measure cubes. However, all the fact tables of the data warehouse have multiple measure values representing number of students, schools, passed and failed students and others. Samples of data cuboids are shown in figure (8).



Figure(8). Sample Cuboids for Multi Valued Fact Tables.

Data distribution using histograms for the data tables is shown in figure(9).





Figure(9). Logged Data Distribution.

5. Discussion and Conclusions

As data increased dramatically and as different organizations in different sectors try to make use of the data(historical and up to date), the storage, processor speeds and other technologies tend to solve such problem in storing, retrieving and processing the data. Emerging of Data Warehouse Technologies represent a step is dealing with the huge amount of data.

Data can be stored either as detailed or summarized depending on the knowledge required. Different DW models including Star, Snowflake and Galaxy can be used to represent the data. Each of these models has its advantages and applications.

The work in this paper presents and approach for solving data mining and data cube processing algorithms obstacles due to the huge amount of data under consideration. The collected data used in this work consists of data about regions, cities, schools, teachers and students where tens of tables each with hundreds or even thousands of records, which means that cube operations (Roll up and Drill down) will face difficulties in both time and space complexities, and hence data reduction and summarization represent a feasible solution.

Tables I and II and figure 9 show samples of the logged detailed data that were used to study the approach effectiveness. Concept hierarchy was used for data summarization and hybrid warehouse model was used to store both detailed data for detailed data analysis and knowledge extraction(better results with high time and space complexities) and summarized data that gives summarized knowledge with better complexities.

Both time and space complexities are functions of number of data items(records) in the data tables, number of attributes of each table and the distinct values for each attribute. The concept hierarchy was used to reduce the number of distinct values in many attributes which in turn gives aggregate and summarized data.

From all the notes mentioned above, it is clear in figure (7) that the data cube of the original (detailed data) has the dimensions (75*11*10*6*2) whereas the data cube after reduction has the dimensions(11*6*6*3*2) which leads to feasible improvement in both time and space complexities.

Tables III, IV and V show the processed data after concept hierarchy was applied(reducing the distinct values in many attributes). High data reduction was received which in turn improves both time and space complexities as show in table VI and figure 6, 7 and 8.

6. References

[1] Jiawei Han and M. Kamber "Data Mining: Concepts and Techniques" 3rd Edition.,
Morgan Kaufmann, 2010.

[2] M. Steinbach, P.-N.Tan and V. Kumar, Introduction to Data Mining, Addison-Wesley,
2006. ISBN: 0-321-32136-7.

[3] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2002.

[4] D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001.

[5] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques
with Java Implementations, Morgan Kaufmann, 2nd ed., 2005, ISBN 0-12-088407-0

[6] Hong Cheng, Xifeng Yan, Jiawei Han, and Chih-Wei Hsu, "Discriminative Frequent
Pattern Analysis for Effective Classification", in Proc. 2007 Int. Conf. on Data Engineering
(ICDE'07), Istanbul, Turkey, April 2007.

[7] J. Han, J. Pei, Y. Yin and R. Mao, "Mining Frequent Patterns without Candidate
Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery,
8(1):53-87, 2004.

[8] X. Wu · Vipin Kumar · J. Ross Quinlan · J. Ghosh · Q. Yang · Hiroshi. "Top 10 Algorithms in Data Mining", (2008) 14:1–37 DOI 10.1007/s10115-007-0114-2, © Springer-Verlag London Limited 2007.

[9] Venky H., Anand R., Jeffrey D. U1lman "Implementing Data Cubes
Efficiently", SIGMOD '96 6/96 Montreal, Canada @ 1996 ACM 0-89791 -794-
419610006.

[10] Edward H., David W. C., Ben K., "Optimization in Data Cube System Design", Journal of Intelligent Information Systems, 23:1, 17–45, 2004 c 2004,Kluwer Academic Publishers. Printed in The United States.

[11] ParvanehShabanzadeh, RubiyahYusof, "An Efficient Optimization Method for Solving
Unsupervised Data Classification Problems", Computational and Mathematical Methods in
Medicine, Volume 2015 (2015), Article ID 802754, 9 pages