



Sudan University of Science and Technology  
College of Graduate Studies

**Study on the Application and Admission Process to Sudanese Universities  
Using Data Mining Techniques**

دراسة في عملية التقديم و القبول للجامعات السودانية باستخدام تقنيات تنقيب البيانات

By

**Atifa Elssir Khalid Elgimari**

The dissertation is conducted under the supervision of  
**Professor Eltayeb Salih Abuelyaman**

A Thesis Submitted for the Degree of Doctor of Philosophy in  
Computer Science

September

2014

## **Dedication**

*To my Mom: Bit wahab Yousif Elhassan*

*To the spirit of my father Elssir Khalid Elgimari....my true love*

*To the spirit of my dearest cousin: Suaad Younis Eldisoogi*

## **Abstract**

This study aims to use modern Data Mining techniques to analyze admission data of Sudanese universities. The current methodology contains two data warehouse structures; a star and a snowflake. A star data warehouse structure was used to develop several Association Rules Mining models. The snowflake data warehouse structure was used to create a smart user interface based on OLAP (On Line Analytical Processing) technique as the first system for OLAP in Sudan. However the Association Rules Mining technique is often used in the business intelligence field, and has limited applications in other fields. We could apply it in higher education field (s) to investigate relationships between attributes, if we developed several Association Rules Mining models. To improve the efficiency and effectiveness of data analysis tasks, the accuracy of those mining models was compared. Through our developed user interface based on OLAP system, end users could easily find the answer of their queries in seconds about data regards to different dimensions. Through this study we supported that Data Mining techniques have huge potential benefits in terms of multidimensional analysis and can help to solve Sudan's education need for skilled analysts. We could discover some hidden patterns what have been done and what is the result is not clear. Moreover, we could solve our typical research questions, such as: What percent of faculty choices do students usually choose on the application form? Is there a need for all these number of faculty choices offered to students on their application forms? Are there any associations between the variation of students' faculty preferences and students' geographical locations, in residential provinces?

## المستخلص

هذه الدراسة تهدف إلى استخدام تقنيات تنقيب البيانات الحديثة في تحليل بيانات القبول للجامعات السودانية. الطريقة المستخدمة تحتوي علي نوعين من مستودعات البيانات: نجمي و ندقات الثلج. مستودع البيانات النجمي قد استخدم في تطوير عدد من نماذج تنقيب قواعد الارتباط. مستودع البيانات ندقات الثلج استخدم في انشاء واجهة مستخدم ذكية بناءً على تقنية المعالجة التحليلية المباشرة كأول نظام للمعالجة التحليلية المباشرة في السودان. بالرغم من أن تقنية تنقيب قواعد الارتباط غالباً تستخدم في مجال ادارة الأعمال الذكية، وله تطبيقات محدودة في المجالات الأخرى. فقد استطعنا تطبيقه في مجال التعليم العالي لدراسة العلاقات بين الصفات، عندما طورنا عدد من نماذج تنقيب قواعد الارتباط. لتحسين كفاءة و فاعلية مهام تحليل البيانات، فقد قورنت دقة تلك النماذج. من خلال تطويرنا لواجهة المستخدم بناءً على نظام المعالجة التحليلية المباشرة فقد تمكن المستخدمين بسهولة من ايجاد اجابة لأسئلتهم في ثواني عن البيانات في مختلف الأبعاد. من خلال هذه الدراسة فقد أكدنا ان تقنيات تنقيب البيانات لها فوائد عظيمة في مفاهيم التحليل متعدد الأبعاد ويمكنها ان تساعد في حل مشاكل تعليم السودان والتي تحتاج الى محللين مهرة. فقد استطعنا اكتشاف بعض الأنماط الخفية والتي يمكن أن تكون غنية بالمعلومات وفي غاية الأهمية. أكثر من ذلك فقد استطعنا ان نجيب على أسئلة البحث التي وضعناها مثلاً: ما هي نسبة الكليات التي يختارها الطلاب عادةً في استمارة التقديم؟ هل هناك حاجة لكل هذا العدد من رغبات الكليات المتاحة للطلاب في استمارة التقديم؟ هل هناك أي ارتباطات بين تباين الكليات المرغوبة للطلاب والمواقع الجغرافية للأقاليم التي يقيمون فيها؟

## **Acknowledgements**

I would like to thank my supervisor Professor. Etayeb Salih Abuelyaman for his patience during the research progress; and for his inspiration and guidance provided me during these years of research. If I have gained any skill as a researcher, it is because of his lessons and his advice and his incredible patience. He consoled me and built my self-confidence. He was appreciating my ideas, even if I thought they were trivial. Also he gave me professional assistance, remarkable insights, and invaluable instructions especially about how to write a scientific paper for international audiences and for publication in international journals. Without him, it would have been impossible for me to have accomplished this task.

I would like to send special thank to Dr. Mohammed El hafiz Mustafa Musa, Dean of College of Computer Science and Information Technology- Sudan University for Science and Technology, for his long time outstanding exploration of the Data Mining field, and for helping me take the first steps in this research area. I appreciate great support he gave me in dealing with a completely new research area. I would like to thank him for allowing me to attend the machine learning and Data Mining courses.

Getting data for research and analysis purposes is a very difficult task in Sudan due to different information security reasons. I would like to thank Professor. Ezz Eldain Mohamed Osman, College of Computer Science and Information Technology- Sudan University for Science and Technology, for helping me in this regard to secure educational data from the Ministry of Higher Education and Scientific Research in Sudan. He is benevolent and well known for supporting and encouraging researchers even though he doesn't know any one of them. He always returned my calls and those of the unknown caller as well.

I would like to thank Tahani Hafiz – an officer in the Ministry of Higher Education and Scientific Research, for providing me with the soft copies of my required data. She made very helpful explorations about the data. I wish to especially thank Professor Ahmed Hassan Eljack, School of Management Studies – Ahfad University for Women, for encouraging me to obtain my Ph.D degree. He has also given me the key of how to write a scientific research. I would also like to thank Dr. Sumaia Elzain, School of Management Studies – Ahfad University for Women, for her support, and facilitation of my work leave request from the School of Management Studies. Especial thanks goes to my close friend. Nour Mohammed Osman, for her generosity of time spent with me; for the friendly relationship she helped to create between me and SQL server. I gratefully thank my close friend Israa Shazerwan for her encouragement, support, and useful advice. I wish to express my sincere thanks to Alhafeed Library officers at Ahfad University for Women, for customizing a study room for me at the library for several months. I would like to thank my nephew, Ali Abuidress, for his support. I appreciate the great efforts he made to obtain the hard copies of my collected data. I would specially like to thank my closest brother Mohamed Elsir Khalid for providing me with valued reference books, and for his encouraging words. I want to thank my brother Yasir Elsir Khalid for supporting me all the time, and for providing me with high storage external hard drive.

For this dissertation I would like to acknowledge my friend, Amina Ali, the American University in Cairo – Egypt, for her support. I appreciate the considerable amount of time and effort she spent to edit this research.

Lastly, I'd like to thank all my family for their moral support during the writing of this research, for their endless love, for their encouragement to me in this challenging period.

I should give the most appreciation to my kindhearted mother because she was asking God to help me all the time; I was encouraged by hearing her kind voice at long nights asking God to give me success in all my life, for her providing me with a convenient environment for studying, and for exempting me from all of the home obligations.

# Table of Contents

Abstract.....	i
Abstract in Arabic.....	ii
Acknowledgements.....	iii
List of Figures .....	ix
List of Tables.....	xiv

## Chapter1

<b>Introduction.....</b>	<b>1</b>
1.1 Research Problem.....	4
1.2 Objectives.....	5
1.3 Scope.....	6
1.4 Research Questions.....	6
1.5 Organization of the Research .....	7

## Chapter2

<b>Literature Survey.....</b>	<b>8</b>
2.1 Introduction to Data Mining technique .....	9
2.2 Data Preprocessing .....	11
2.3 Data warehouse and OLAP.....	12
2.4 Creating a data cube.....	15
2.5 Association Rules Mining Technique.....	19
2.5.1 Computing of the support and confidence of a discovered rule.....	31
2.5.2 Combining the Association rules mining technique with OLAP technique .....	37



2.5.3 Association rules mining technique applications .....	43
2.6 Data mining applications .....	44
2.6.1 Data mining applications in the higher education domain .....	44
2.6.2 Data mining applications in developing a system.....	46
2.7 An analysis of the presented literature survey.....	48

### **Chapter3**

<b>Data Preprocessing.....</b>	<b>51</b>
3.1 Data cleaning.....	55
3.2 Data Integration.....	57
3.3 Data transformation.....	60
3.4 Data reduction.....	64
3.4.1 Data discretization and concept hierarchy .....	65
3.4.2 Data generalization .....	67
3.5 Applying Data formation for OLAP .....	68

### **Chapter 4**

<b>Research Methodology.....</b>	<b>82</b>
4.1 Data Collection.....	84
4.1.1 The original source of the collected data.....	86
4.1.2 Describing the collected data.....	87
4.2 Data warehouse.....	92
4.2.1 What is the importance of a data warehouse? .....	93
4.2.2 Data warehouse structure.....	96
4.3 OLAP technique .....	99
4.3.1 OLAP (Cube) structure.....	100

4.3.2 OLAP programming.....	112
4.4 Data mining Process.....	116
4.4.1 The Mining Structure.....	121
4.4.2 The Mining Model.....	123

## **Chapter 5**

<b>Evaluation and Conclusions.....</b>	<b>140</b>
5.1 Results.....	141
5.2 Evaluation and Discussion.....	145
5.3 Conclusions.....	157
5.4 Recommendations and future work.....	160

## List of Figures

Figure 2.1: Architecture of a typical data mining system .....	10
Figure 2.2: Data preprocessing as a step in the process of knowledge discovery.....	11
Figure 2.3: Cube implementation in Microsoft SQL Server Analysis Services .....	17
Figure 2.4: Multidimensional database display with star schema.....	39
Figure 2.5: Browsing OLAP cube according to measure .....	39
Figure 2.6: Frequent items according to MIN_SUPPORT from OLAP cube ...	40
Figure 2.7: Association rules from university cube .....	40
Figure 3.1: we usually use some techniques for cleaning data.....	56
Figure 3.2: The integration is the process of combining data from multiple sources into a single data store.....	58
Figure 3.3: Sort transformation data .....	59
Figure 3.4: Integrating distinct high school names .....	60
Figure 3.5: Data transformation participates in the ETL process.....	62
Figure 3.6: SSIS package for colleges' transformation.....	63
Figure 3.7: Data reduction.....	64
Figure 3.8: An example of a high school field descritization .....	66
Figure 3.9: An example of a faculty field descritization.....	67
Figure 3.10: An example of a generalization process for a state source field	68
Figure 3.11: Student table (Master).....	69
Figure 3.12: School look up table attributes .....	73
Figure 3.13: Major table .....	74
Figure 3.14: State look up table attributes .....	74

Figure 3.15: Direction table attributes .....	75
Figure 3.16: Country table attributes .....	75
Figure 3.17 Department table attributes .....	75
Figure 3.18: Program type table attributes .....	76
Figure 3.19: College table the attributes.....	76
Figure 3.20: University table attributes.....	76
Figure 3.21: Date table attributes.....	77
Figure 3.22: Scores table attributes.....	77
Figure 3.23: Choices table attributes .....	77
Figure 3.24: The proposed structure of the relational database.....	78
Figure 4.1: Research Methodology Overview.....	83
Figure 4.2: Multidimensional snowflakes warehouse.....	98
Figure 4.3: implementation of Location Hierarchy.....	101
Figure 4.4: implementation of Location dimension with drilling down to Sudan country.....	101
Figure 4.5: Implementation of Higher education dimension.....	102
Figure 4.6.a: Implementation of the cube by querying the Higher education dimension about scores average percent.....	103
Figure 4.6.b: Implementation of the cube by querying the Higher education dimension about filled choices percent.....	104
Figure 4.7.a: Implementation of the cube by querying the Location dimension about scores average percent per province for Sudan country.....	105
Figure 4.7.b: Implementation of the cube by querying the Location dimension about filled choices percent per province for Sudan country.....	105
Figure 4.8.a: the High School dimension is filtered by major and queried by scores average percent .....	106

Figure Fig 4.8.b: the High School dimension is filtered by major and queried by filled choices percent .....	106
Figure 4.9.a: the High School dimension is filtered by School's ownership and queried by scores average percent .....	107
Figure 4.9.b: the High School dimension is filtered by School' ownership and queried by filled choices percent.....	107
Figure 4.10.a: the High School dimension is filtered by School's learning methods and queried by scores average percent .....	107
Figure 4.10.b: the High School dimension is filtered by School' learning methods and queried by filled choices percent.....	108
Figure 4.11.a: Measuring of the high schools by score average percent according to their type.....	108
Figure 4.11.b: Measuring of the high schools by filled choices percent according to their type.....	108
Figure 4.12.a: Female Vs Male are measured by scores average percent.....	109
Figure 4.12.b: Female Vs Male are measured by filled choices percent.....	109
Figure 4.13.a: The maximum scores average of students per province.....	110
Figure 4.13.b: The maximum scores average of students per state.....	110
Figure 4.14.a: The maximum scores average of students per admission type	111
Figure 4.14.b: The maximum scores average of students per high school's major.....	111
Figure 4.14.c: The maximum scores average of students per high school's ownership.....	111
Figure 4.14.d: The maximum scores average of students per high school's type.....	111
Figure 4.15: The system uses FAC to measure admissions for Sudan at 2005	113

Figure 4.16: Minimum Scores Average for female students who admitted in the University of Khartoum- faculty of Science at 2009.....	114
Figure 4.17: Minimum Scores Average for male students who admitted in the University of Khartoum- faculty of Science at 2009.....	115
Figure 4.18: The proposed cube that built based on a star data warehouse structure.....	118
Figure 4.19: one-to-many relationship- each student can apply to many faculty choices and has many scores average.....	119
Figure 4.20: Associating each student with his\her faculty choices and scores	120
Figure 4.21: A mining Model Structure.....	122
Figure 4.22: An OLAP mining structure.....	123
Figure 4.23: A relational view of an OLAP mining model.....	124
Figure 4.24: The proposed case structure.....	124
Figure 4.25: A Mining Model of Student_Admission_Model.....	125
Figure 4.26.a: The found itemsets for the Middle province.....	126
Figure 4.26.b: Other found itemsets for the Middle province.....	126
Figure 4.26.c: The rest of the found itemsets for the Middle province.....	127
Figure 4.26.d: The discovered rules for the Middle province.....	127
Figure 4.27.a: The found itemsets for the North province.....	128
Figure 4.27.b: The discovered rules for the North province.....	128
Figure 4.28.a: The found itemsets for the West province.....	129
Figure 4.28.b: Another found itemsets for the West province.....	129
Figure 4.28.c: The discovered rules for the West province .....	130
Figure 4.29.a: The found itemsets for the East province.....	130
Figure 4.29.b: The discovered rules for the East province.....	131
Figure 4.30.a: The found itemsets for the South province.....	131

Figure4.30.b: The discovered rules for the South province.....	132
Figure 4.31.a: The dependency net view of Student_Admission_Model.....	132
Figure 4.31.b: The dependency net view shows that The East province has no association with any college.....	133
Figure 4.32: The dependency net view of admissions based student's sex....	134
Figure 4.33: A Mining Model of both High_School_Major_Sex and Major_Province Models.....	134
Figure 4.34: The found itemsets for the of High_School_Major_Sex Model...	135
Figure 4.35: The discovered rules for the of High_School_Major_Sex Model	135
Figure 4.36: The dependency net view of High_School_Major_Sex Model...	136
Figure 4.37: The found itemsets for the of Major_Province Model.....	137
Figure 4.38: The discovered rules for the of Major_Province Model.....	137
Figure 4.39: The dependency net view of Major_Province Model .....	138
Figure 5.1: The developed OLAP system.....	149
Figure 5.2: The predictive models' task.....	150
Figure 5.3: The lift chart of the Student_province_Model, where 90 % of the target can be captured using 75 % of the data.....	155
Figure 5.4: The lift chart of the Student_Sex_Model, where 90 % of the target can be captured using 86 % of the data.....	156

## List of Tables

Table 2.1: A cube with two dimensions.....	35
Table.3.1: The contained fields in the final integrated faculty table.....	60
Table 3.2: Description of student table attributes .....	69
Table 3.3: student’s choices.....	70
Table 3.4: Choices measures.....	71
Table3.5: Scores average by subject for each student’s choice.....	71
Table3.6: Student’s scores average.....	71
Table 3.7: Description of school table attributes .....	73
Table 3.8: Description of state table attributes.....	75
Table 4.1: ADMFR05.....	89
Table 4.2: ADMFR05_ch.....	90
Table 4.3: FAC05.....	92
Table 5.1: Measuring of scores average percent along several dimensions...	142
Table 5.2: Measuring of maximum scores average along several dimensions	143
Table 5.3: Itemsets for Student_Admission_Model.....	144
Table 5.4: The generated rules for Student_Admission_Model.....	144
Table 5.5: Itemsets of High_School_Major_Sex model and Major_Province model.....	145
Table 5.6.a: Applying “importance” measure on two items A and B for itemsets .....	151
Table 5.6.b: Applying “importance” measure on two items A and B for rules	151



## **Chapter 5**

# Chapter 1

## Introduction

Data mining is a powerful new analysis tool with great potential in the information system world and database analysis methods. It can be best defined as the task of discovering meaningful knowledge from large amounts of data. This task of discovering or extracting knowledge from large amounts of data cannot be considered an easy process. In some sources authors described it as gold mining -a process by which one must analyze data deeply to extract the valuable knowledge in its content. Accordingly, others suggested that Data mining should have been more appropriately named “knowledge mining rather than Data mining”.

In comparison to other data analysis methods, Data mining is making great potential changes in the world of data analysis methods. With respect to the amount of data that must be analyzed, some data analysis methods such as ordinary statistics, usually deal with less data than Data mining analysis methods; thus, Data mining techniques can analyze a large amount of data sets. If we compare the type of data in each method, we find that other methods are suited for static, categorical and structured data types. However, Data mining methods are suited for static, dynamic, real, and multi-variant data. On the other hand, Data mining methods deal with historical or pre-existing data which is completely different from data generated by other methods. In addition, those traditional analysis methods require complex queries in order to find the best answers for decision makers. The clear difference between Data mining methods and the other data analysis methods such as statistics, is that Data mining methods often follow intelligent methods in order to learn a system. This means that efficiency and scalability of algorithms are very important in Data mining methods.

From all of these comparisons, we can conclude that Data mining techniques can be considered an accumulation of other data analysis methods.

The Data mining science supports many valued techniques such as Classification, Clustering, Association Rules Mining, etc. Each technique could be convenient according to the mining problem. For example, association rule mining technique is designed specifically for association analysis. This technique is typically used for shopping basket analysis, where one can analyze the association between items. For example, if there are two items X and Y, the Association Rules Technique aims to find patterns of the form  $X \rightarrow Y$ , with the intuitive meaning “baskets that contain X tend to contain Y”.

Since the core of the Association Rules is finding relations among items occurring together within the same transaction, it is applied in many business intelligence applications. In contrast, other techniques have been used in different other applications. For instance, the Classification technique has been used largely in the education field to determine whether the student will pass or fail.

Based on the kind of knowledge to be mined, Data mining techniques are divided into two categories: a descriptive Data mining technique; which performs descriptions on the general properties of the data in the database, and predictive Data mining technique, which performs inference on the current data in order to make predictions (Han, Kamber 2006). In view of that, the association rule can be considered as one of the Data mining techniques is regard as a descriptive technique; while others such as Classification technique is regard as a predictive technique.

Nowadays, Data mining techniques are applied in many application domains such as medicine, engineering, science, business, and **education**. However, the applications of Data mining techniques are still limited in different domains.

Currently in Sudan, the education area is an attractive area for research; the number of Sudanese universities was limited several years ago. Most of the universities were concentrated in Khartoum state only, despite the large number of students who were waiting to get a chance to enroll in one of the higher education institutes. This produced a high level of competition because there were a limited number of available seats for each college. The higher education policies in Sudan worked on increasing the number of universities more and more in the last 20 years. For example, in 2005 the total number of college choices was 1457, but this number increased drastically to 1619 in 2009. On one hand, this increase has helped students in getting more opportunities to enroll in Sudanese universities while also reducing the level of competition. However, it produced a large number of options for colleges to which student could apply. On the other hand, that increase has provided us with a huge number of students' datasets which we need to manage and analyze scientifically by using appropriate and modern analysis methods.

Undoubtedly, the educational sector has a great affection by the political instability. The unstable political situation may have forced some students to move from one state to another and discontinue their studies. That occurred, for example, when the southern region of Sudan was separated from the north in 2011. As a result, three universities were moved from the north of Sudan to the south of Sudan. In addition, some cities that fall in areas close to the geographical boundaries between the two countries are becoming major sources of conflict.

Unstable political situations often provide us with changeable students' data that require an effective analysis method to handle. Accordingly, Data mining techniques have been proposed as suitable and modern analysis methods to meet those needs.

Indeed, a qualified analysis based on optimized scientific methods helps higher education institutes to make the right decisions toward enhancing their educational strategies. The educational strategies may include increasing/decreasing the number of universities in a certain state relative to another, increasing a student's promotion rate, increasing a student's retention rate, or decreasing a student's drop-out rate, etc.

In this study the association rule mining technique has been applied in the field of education. It has been applied to a dataset that had been collected from the Ministry of Higher Education and Scientific Research in Sudan. These collected data cover the records of students who applied to the Sudanese universities and were admitted into them within the period 2005- 2009.

To discover knowledge, we will follow two methodologies: analyzing the collected database using OLAP technique, and applying the association rule mining algorithm to find valued associations.

## **1.1 Research Problem**

Every year a large number of students apply to Sudanese universities, and accordingly, a huge number of application forms are printed to be available for students. Each application form contains about 45 choices of student's interest. The processing of applications is extremely costly and much effort is required in the data entry processing. The field of business intelligence often tries to make right decisions that could be successful in finding some ways in order to reduce the cost, time, and effort. To that end, and in our case of mining the educational database, it should be useful to answer some questions such as: Is there need for all these number of choices? How many choices qualify a student to be admitted to his/her preferred college? Are there any relationships between students' demographic

information such as their higher schools, states, and their strategies for applying to Sudanese universities? What percent of choices do students usually fill in the application form in regard to the variation in residential regions? This research intends to benefit from the Data mining technology as much as possible.

Since the data analysis processing using the traditional methods needs to apply complex queries and takes more time, Data mining experts have provided us with a great technology for easy data analysis; this technique is the On Line Analytical Processing (OLAP) technique. With this technique, we can perform many operations such as roll up, drill down, slice, and dice. These OLAP operations, in many cases, have proven to provide excellent solutions of getting fast, smart, and easy methods to analyze data. Furthermore, they have helped decision makers to accomplish respected outcomes on the way of obtaining the right decision.

In addition to benefiting from advantages of the OLAP operations, this research targets to develop mining models based on the OLAP Mining (OLAM) means.

## **1.2 Objectives**

The objectives of this study are to:

- Develop an OLAP system smartly enabling the Ministry of Higher Education and Scientific Research analysts to analyze students' data based on the application and admission processes.
- Evaluate and analyze the dominant factors that affect students' college preferences using the Association Rules Mining technique.

### **1.3 Scope**

- The investigated time period:  
These data represents records of students who had applied and were admitted in the higher education institutes within the period (2005 – 2009).
- The investigated institute(s): The dataset of this study has been obtained as a secondary data source from the Ministry of Higher Education and Scientific Research in Sudan, specifically, from the Department of General Directorate for Admissions, Certificates' Authentication and Accreditation.

### **1.4 Research Questions**

- Do students need for 45 faculties choices to choose?
- Do students from a certain state prefer certain colleges based on their gender, education type, and other students' categories?
- Are certain colleges such as medicine and engineering still more attractive for the most students?
- Is there any analytical significance for the students who were born and grew up in the different states with their success in high school, and their college preference?
- To what extent will this proposed system help decision makers in higher education institutions to make their decisions based on information obtained from it?

## **1.5 Organization of the Research**

This research is organized into five chapters. In Chapter 2 a review of relevant background literature is discussed which provides the actual seeds of the research. In Chapter 3, hard efforts of data preprocessing are documented. This chapter explains how we cleaned, integrated, transformed the collected data, and formation of data to suit the research objective is performed. In chapter 4, the methods used for the research is elaborated including description of the collected data, illustrating the bases of the proposed data warehouse structure, designing of OLAP structure, programming OLAP, and creating a number of OLAM models. Chapter 5 analyzes and discusses the results and performance of the models described in Chapter 4 for OLAP and OLAM followed by a summary and conclusion for the research.



## Chapter 2

### Literature survey

Nowadays, database analysis techniques have attracted a great deal of attention in the information world, due to the need for analysis in diverse fields like industry, agriculture, medicine, physics, biology, chemistry, and environmental science to name a few. This analysis has proved to be very essential to the recent events that changed the world like analyzing the last global climate changes, and the financial crisis of September 2008. Because of events such as these, analysts are encouraged to increase their efforts to improve the database analysis techniques.

The traditional database analysis techniques with probability statements, such as statistics, are limited and do not often meet the wide availability of huge amounts of data. There is an imminent need to analyze such data and turn it into useful information. Moreover, these traditional database analysis techniques, though theoretically sound, oversimplify results based on summary statistics; such as means or variances, rather than on individual uniqueness. Furthermore, the traditional database analysis techniques provide users with aggregated results about data, but they cannot find the relationships between items.

A database analysis technique, *Data mining*, effectively participated in solving such problems in which there is a shortage or limitation of traditional database analysis techniques. If we compare Data mining with other techniques, and for example, in contrast to the use of traditional statistics; *Data mining* uses a wide range of real databases, terabytes in contrast with megabytes in statistics. Another advantage of the Data mining technique is that it integrates other techniques from

other disciplines such as database and data warehouse, statistics, machine learning, pattern recognition, and neural networks.

In this chapter we discuss the following topics: a brief introduction to Data mining technique in section 2.1. Data preprocessing is one of data mining process steps, in section 2.2 we discuss some concepts that could be related to it. We introduce some concepts about data warehouse and OLAP in section 2.3. How to create a data cube is the topic of 2.4. In section 2.5 we discuss some issues related to Association Rules Mining Technique such as: computing support and confidence count of association rule, Classification technique, Integrating Association Rules Mining Technique with other data mining techniques, Integrating Association Rules Mining Technique with OLAP technique in the term of (OLAM), and Association rule mining applications. In section 2.6 we discuss Data mining applications in the higher education domain and in developing systems. An analysis of the presented literature survey is the topic of section 2.7.

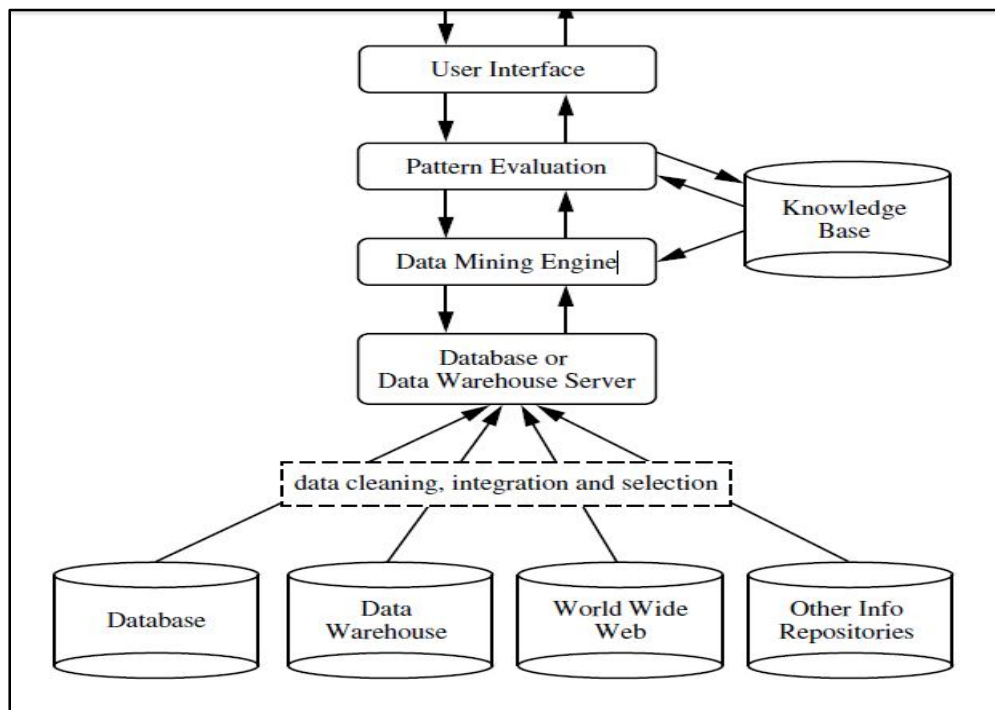
## **2.1 Introduction to Data mining technique**

*Data mining* has been defined in many expressions. In the popular book [Han, Kamber 2006] the authors defined it as a process of extracting knowledge from large amounts of data. In the way of defining a Data mining as a process of discovering hidden patterns in a large database, the author [Luan 2002] went deeply into describing it as a method for delineating systemic relations between items when there is no a priori knowledge about the nature of those relations.

The noted progress of data collection tools and storage media in the last few years has provided analysts with a huge number of database and information repositories; hence data miners have been able to dig deeper into the data to

discover potential knowledge, and then identify underlying relationships and features in data and generate models for better prediction.

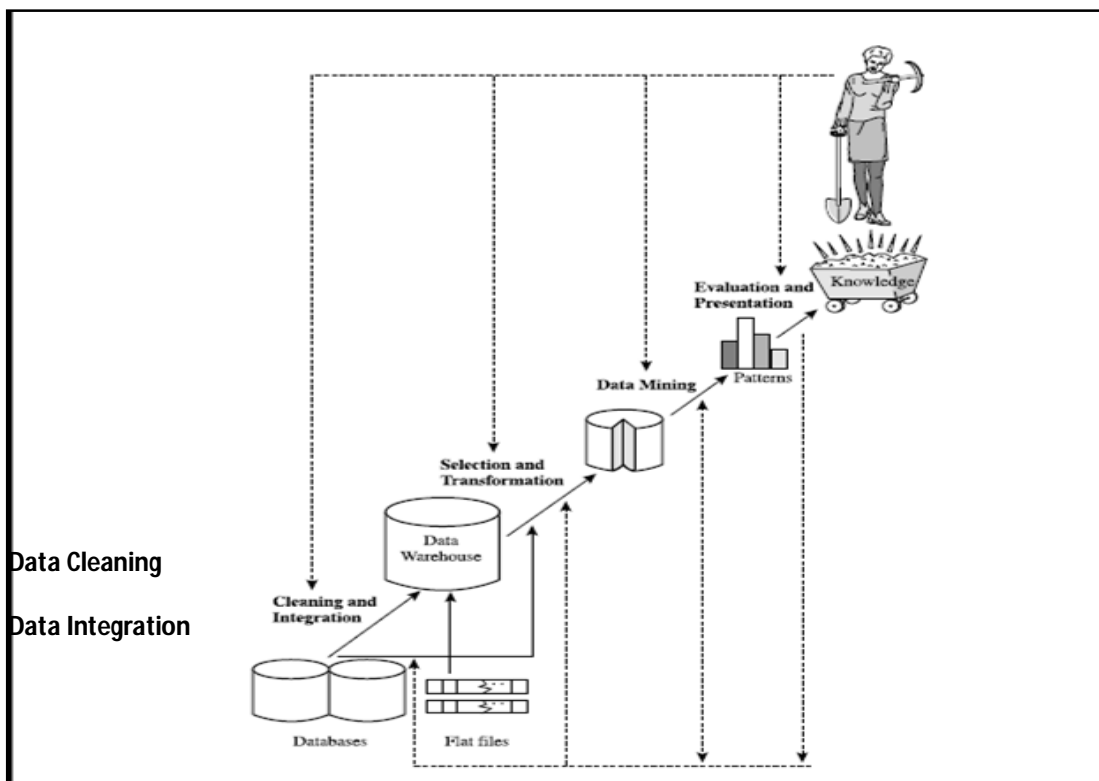
Data mining technique is that it integrates other techniques from other disciplines such as database and data warehouse, statistics, machine learning, pattern recognition, and neural networks. The term Data mining tools includes a number of analytical techniques that are used to extract useful patterns or meaningful knowledge from large datasets. Classification, Clustering, Association Rules Mining are examples of these Data mining techniques [Han, Kamber 2006]. According to [Han, Kamber 2006], Architecture of a typical data mining system is shown in figure 2.1.



**Fig.2.1:** Architecture of a typical data mining system

## 2.2 Data Preprocessing

Referring to [Han, Kamber 2006], the knowledge discovery process is done through the following steps: data preprocessing, data warehouse, Data mining, pattern evaluation, and knowledge discovery presentation. Accordingly, data preprocessing is an important step in the knowledge discovery process as shown in Figure 2.2.



**Fig. 2.2:** Data preprocessing as a step in the process of knowledge discovery.

[Han, Kamber 2006] stated a number of data preprocessing techniques such as: Data cleaning techniques which can be applied to eliminate noise and inconsistencies in the data, merging data from multiple sources into a unified data store, such as a data warehouse, and they suggested the data integration techniques. To applying some operations such as normalization, that may improve the

accuracy and efficiency of mining algorithms involving distance measurements, they presented some data transformations techniques. Moreover, they discussed data reduction techniques which can be used to reduce the data size by aggregating, Clustering, or eliminating redundant features, for instance. The data reduction technique must be used very carefully in order to obtain reduced representation in the data volume but produces the same or similar analytical results. Likewise, [Han, Kamber 2006] introduced data discretization technique as a part of data reduction but with particular meaning, especially for numerical data.

Note that, the above techniques are not mutually exclusive; they may work together. For example, the removal of redundant data may be seen as a form of data cleaning, as well as data reduction at the same time.

### **2.3 Data warehouse and OLAP**

Due to the availability of different kinds of information and data storage repositories, according to [Ponniiah 2010], many organizations thought to collect all data from multiple heterogeneous data sources into a single storage repository. To do so, they had to build a *data warehouse* system in the 1990's, in order to achieve competitive advantages in facilitating the decision making process. By using data warehouse systems, users were able to integrate and store data in a unified schema from heterogeneous data sources such as Excel and Access databases. A *data warehouse* storage repository often contains cleaned and integrated data, additionally; one of the significant advantages of the *data warehouse* is its own structures that differ from the traditional database structures. The traditional database structures habitually generalize and consolidate data into two dimensions in contrast with multidimensional structure as in a *data warehouse* storage repository.

According to the sources [Ponniah 2010; Han, Kamber 2006; Rob, Ellis 2007], a *data warehouse* stores data in the following two table formats: dimensions and facts tables. Dimensions are entities with related to which users want to store their records. For instance, an education institute may create an *admission* data warehouse in order to keep records of students' admission using dimensions *scores, admission year, location, and higher school*. While the term dimension means a concept of distributing all data according to the users need, the term dimension table describes the dimension attributes. For example, a dimension table for *location* may contain the following attributes: *city, state, and country*. The attributes of a dimension table are often used to categorize or summarize facts. In the multidimensional data model all dimension tables are organized around a central representation which is called a fact table. The fact table is a table that measures numerical facts about its related dimension tables. Each dimension table has a key to ally with a fact table; thus the fact attributes are nothing more than measures and dimension tables keys. For example, a fact table for *admission* data warehouse could include: student's scores average as a measure, keys of dimensions *scores, admission year, location, and higher school*.

Generally, according to source, [Han, Kamber, 2006], there are three Data Warehouse models: Star schema, where a fact table in the middle connected to a set of dimension tables through a direct relationship form; Snowflake schema, where another dimension tables can be related to a fact table through an indirect relationship form; and Fact constellations sighted as a collection of stars, where multiple fact tables share dimension tables.

Storing data in a multidimensional structure has had a valuable effect in the database analysis area. It has provided analysts with on- line analytical processing (OLAP), remarkable analysis technique, that is, [Han, Kamber 2006; Jadav, Panchal 2012]. This technique is fashioned on a multidimensional structure of a

*data warehouse*; data can be analyzed with some functions such as summarization, aggregation, and consolidation as well as the ability to view information along different dimensions. Each dimension in the multidimensional model can have a multiple level of abstraction defined by concept hierarchies. Based on the concept hierarchies, OLAP technique in the multidimensional model has eminent operations: roll-up, drill-down, slice and dice, and pivot. For example, if we consider a hierarchy of a *college* dimension, we define it as a total order “department < faculty < university”, Roll-up operation. Users can aggregate the college dimension with respect to its hierarchy by ascending the *college* hierarchy from the level of department to the level of university. The Drill-down operation is the inverse of Roll-up operation in which users can aggregate the college dimension with respect to its hierarchy by descending the *college* hierarchy, from the level of university to the level of department. All tools of data warehouse and OLAP are based on a multidimensional data model. By using this model we can view data in a form of a data cube. Therefore, with a data cube, data can be modeled and viewed in multiple dimensions and facts, as well as with multiple levels of abstraction.

Comparing data mining and OLAP with traditional statistics in the context of analysis methods was the topic of many researches. In general and although theoretically sound, most of these researches concur that traditional statistics with probability statements are limited and, at times, generalizing. Moreover, most analysis results draw conclusions based on summary statistics, such as means or modes, rather than on individual exclusiveness. In contrast, applying OLAP operations, such as: slicing, dicing, drilling down, and rolling up, enabled the user to systematically explore the knowledge space to find out useful knowledge for analysis purposes. Authors, (Han, Kamber 2006), distinguished between the (OLAP) technique and the traditional statistical methods from the respect of their

applications; they agreed that (OLAP) technique has been targeted for business applications; whereas, the statistical methods has been targeted for socioeconomic applications. Then, what about using the (OLAP) technique in the education domain?

The source, (Han, Kamber 2006), also discussed the differences between On- Line – Transaction- Processing (OLTP) and OLAP, and they predicted that these differences will be decreased. [Rob, Ellis 2007] developed a case project that clearly described how to convert OLTP to OLAP for data mining purposes

## **2.4 Creating a data cube**

Creating an efficient data cube is predominant operation in the analysis process. [Rob, Ellis 2007] presented a great work in this area; their objective was to provide students with both the theoretical knowledge as well as practical experience with data warehousing tools and techniques. [Rob, Ellis 2007] applied a number of methods such as: developing a graduate course on data warehousing and Data mining, description of a particular case project that detailed some data preprocessing techniques, and describing the process of development of a data cube as well as application of OLAP tools.

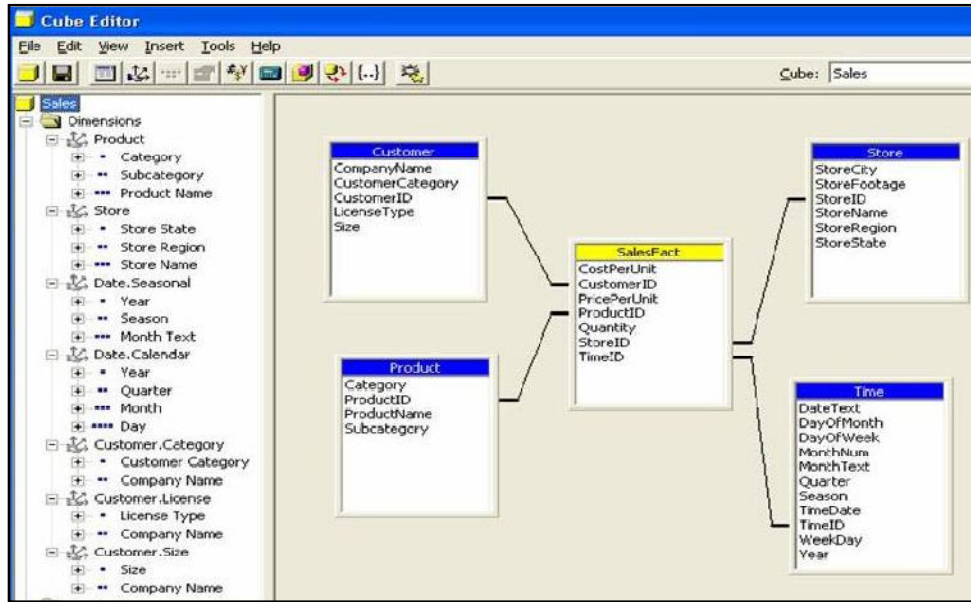
Based on the used OLTP System, the authors, [Rob, Ellis 2007] discussed the disadvantages of that system, and justified the need for developing a data warehouse as a solution. In the discussion of the disadvantages of the used OLTP System, they stated the following points: The OLTP system is painfully impractical when it comes to strategic decision support. The types of information that are requested from management must be dealt with individually by the information technology (IT) staff. Data aggregations are programmed into reports, but any comparisons across time or products must be a manual task. The capability



of storing data of the OLTP system is typically limited to only 2 years, although the company generated computerized data for over 20 years.

In order to handle all these disadvantages of the OLTP System, the authors proposed to develop a data warehouse. Developing a data warehouse was a suitable solution, because it provided: a central repository for historical data; an integrated stage for historical analysis of sales data; and users could apply the operations of (OLAP) techniques by themselves. Therefore, they explained how to convert data from (OLTP) system to OLAP system after developing a data warehouse structure. Through developing their data warehouse structure, first they have designed the dimension tables, their hierarchies, and the facts, after that, [Rob, Ellis 2007] chose a star schema for designing a data warehouse structure. The STAR schema was chosen because the dimension tables were not normalized and the size of these dimensions was not too large, The STAR schema was also chosen as it provides an intuitive design that is easily understood by users. The star schema has been used as a base for an OLAP cube that was used to implement the database. The database implementation includes: dimension implementation and fact implementation.

Following the implementation of dimension and fact tables, Cube and OLAP implementations are shown in Figure 2.3.



**Fig. 2.3:** Cube implementation in Microsoft SQL Server Analysis Services

While [Rob, Ellis 2007] described detailed steps for creating a data cube, [Ivanova, Rachev 2004] depicted a new approach for creating a data cube that approach allows users to view aggregated data through multiple viewpoints. From a conceptual view, the authors described the components of the cube as a base cuboid surrounded by a collection of sub-cuboids. These sub-cuboids are used to compute aggregation of the base cuboid across one or more dimensions.

Some or all of the available cuboids are often required for multi-dimensional analysis. In order to sort the raw data set that produces the required cuboids, the aggregated data has been viewed based on two basic methods for computing a group-by: sort-based and hash-based. On the other hand, the aggregated data has been viewed by presenting the concept of Dynamic Data Cube (DDC). DDC was suited to dynamic growth of the cube which gracefully expands the data cube in any direction. One of the important advantages of this method is that it provides a more effective and proficient performance when handling clustered data.

One of the multiple viewpoints of aggregated data in this research was the conceptual model of data cube classes. The concept of cube classes has been used to describe initial user requirements as the starting point for one of the data-analysis stages. These cube classes basically consists of multiple areas. Namely these areas are:

1. head area, which contains the cube's class name;
2. measures area, which contains the measures to be analyzed;
3. slice area, which contains the constraints to be satisfied;
4. dice area, which contains the dimensions and their grouping conditions to address the analysis;
5. cube operations, which cover the OLAP operations for a further data-analysis stage.

Throughout this portion of the research, the authors concluded that conceptual models of data cubes were very interesting areas, especially when they are used for constructing and using resource information extracted from data cubes. Moreover, they stated two different levels of OLAP tools which were used in order to implement a multidimensional model: Structural and Dynamic. The Structural level refers to structures that form the database schema and the metadata that provides the model's key semantics (i.e. facts, measures, dimensions). The Dynamic level refers to OLAP operations and the definition of final user requirements.

Finally, [Ivanova, Rachev 2004] set some challenged tasks for future work. For example, in the direction of dynamic data cube (DDC), they recommended that researchers develop methods of reducing the space requirements of the dynamic data cube from the full data cube size by deleting unnecessary data. Furthermore, they recommended that more discussion about the DDC properties occurs which

enables more powerful management of the memory space, as well as empty regions of the cube.

The authors recommended develop an appropriate class definition for data cube of the Dynamic level as one of the levels of OLAP tools used to implement a multidimensional model and to add extra properties in class for advance analyzing.

Clearly, the Data mining technique has outperformed the traditional data analysis techniques. It has presented the innovative data warehouse structure, which extracts data from multiple data sources and stores them in a unified storage repository with multidimensional structure as mentioned previously [Rob, Ellis 2007]. OLAP from the perspective of the data warehouse has been viewed. OLAP often receives cleaned, integrated, and transformed data from the data warehouse, and does some calculated operations such as, aggregation, consolidation, and summarization. The multidimensional structure of the data warehouse enables OLAP to perform its calculated operations along different dimensions, and then view its calculated operations with multiple level of abstraction using a data cube. Therefore, the data cube in the OLAP area is also classified as the multi-dimensional database.

## **2.5 Association Rules Mining Technique**

The Association Rules Mining Technique is one of data mining techniques. In fact, each of data mining techniques has its own purpose. While some techniques are used for description; other techniques are used for the purposes of prediction. For instance, Association Rules Mining is classified as one of the Data mining techniques that are used for description purposes. Many researchers have discussed this technique from different concerns: its main task, definition, measures, applications, and its integrations with other Data mining techniques.

In the view of main task of Association Rules Mining technique, several authors stated that, main task of An association rule is to find correlations between items with finding patterns of the form  $X \rightarrow Y$ , with the intuitive meaning “baskets that contain item X tend to contain item Y”. [Lui et al. 2006; Han, Kamber 2006; Hristovskia et al. 2001].

Considering the concept of market basket analysis, X and Y are individual products in the basket which are called *items*. The term set of items is more commonly used as *itemset* in Data mining research literature [Han, Kamber 2006]. An itemset that contains k items is a k-itemset. The set {containing PC, Printer} is a 2-itemset, and the set {with PC, Printer, and Scanner} is a 3-itemset.

In this context *transaction* concept has been defined. [Huang et al. 2007] referred the term *Transaction* to a basket which corresponds to a single visit of a customer to a store. Thus, the core of the association rule is finding relations among items occurring together within the same transaction.

As [Han, Kamber 2006] believed, the occurrence frequency of an item is the number of transactions that contain the itemset. This is also known as the *frequency, support count, or count* of an itemset.

As [Lui et al. 2006; Han, Kamber 2006; Hristovskia et al. 2001] defined Association Rules Mining technique based on its main task, authors [Bogdanova, Georgieva 2005; Allard, et al. 2010; Jiang, Gruenwald 2006; Jiang, Gruenwald 2006] followed a mathematical way for defining Association Rules Mining technique; they affirmed that, Association Rules Mining technique is typically an implication rule, and to form it, Let  $I = \{I_1, I_2, \dots, I_m\}$  be a set of items. Let D be a set of database transactions where each transaction T is a set of items such that  $T \subseteq I$ . Each transaction is associated with the identifier TID. Let X is a set of items. A

transaction T is said to contain X if and only if  $X \subseteq T$ . An association rule is an implication rule that takes the form:

$$X \rightarrow Y [s,c] \dots \dots \dots (1)$$

Where,  $X \subseteq I$ ,  $Y \subseteq I$ , and  $X \cap Y = \Phi$ , rule *support*= s, and *confidence*=c are two measurement parameters of rule interestingness. According to [Han, Kamber 2006; Ben Messaoud et al. 2006], the *support* parameter reflects the usefulness of discovered rules, whereas the *confidence* parameter reflects the certainty of discovered rules.

The expression  $X \cap Y = \Phi$  indicates that X and Y are disjoint items. From that expression, we can conclude the fact that the association rule mining algorithms suppose that the items that are given as inputs to the algorithms are disjoint. Now, what is the nature of relationship between transactions? Based on the answer to this question, association rules have been divided into two types based on the nature of relationship between transactions and their itemsets: intra - transaction association rules and inter - association rules [cited in Huang et al. 2007]. The term intra- transaction association rules is used when association rule mining algorithms are based on a finite set of disjoint transactions, and the rule associates itemsets across the same transaction. In contrast, the term inter-transaction association rules are used when the rule associates itemsets across different transactions records.

Since the general form of an association rule is:

$$X \rightarrow Y \dots\dots\dots (2)$$

For more clarity, we can express it with the form:

$$\text{e.g. } \text{Visits}(X, \text{“Sudan”}) \rightarrow \text{Meets}(X, \text{“very kind people”}) \dots\dots\dots (3)$$

Where X is a variable representing a tourist, and Visits and Meets are predicates. The left hand side of the rule -Visits(X, “Sudan”) - is called the rule antecedent, whereas, the right hand side of the rule - Meets(X, “very kind people”)- is called the rule consequent. Each distinct predicate in rule (3) has been referred to as attribute or dimension; for example, the predicates Visits and Meets are considered dimensions of the rule.

Based on the number of data dimensions involved in the rule, [Han, Kamber 2006] divided association rule mining into two types: a single – dimensional association rule or intra-dimensional association rule and multidimensional association rule.

A single – dimensional association rule or intra-dimensional refers to an association rule that contains only one dimension or a single distinct predicate and it takes the following form where X is a variable representing a tourist:

$$\text{Visits}(X, \text{“Sudan”}) \rightarrow \text{Visits}(X, \text{“Omdurman”}) \dots (4)$$

As we have seen in the association rule (4) there is only one dimension (visits) with multiple occurrences which occurs twice within the rule. Such rules are commonly mined from transactional data. However, a multidimensional association rule refers to an association rule that contains more than one dimensions and it takes the form:

$\text{Visits}(X, \text{“Sudan”}) \wedge \text{interested in } (X, \text{“History”}) \rightarrow \text{Visits}(X, \text{“Omdurman”}) \dots \dots \dots (5)$

Regarding to the association rule in (5), there is more than one dimension (Visits, interested in) that can be noted. This rule relates what a tourist visits as well as the tourist’s interest. Here additional information regarding the tourist, who visits Sudan, such as tourist’s interest, tourist’s nationality, may also be considered.

A k-predicate set is a set containing k conjunctive predicates. For instance, a set of predicates {Visits, interested in} from rule (5) is a 2-predicate set. Furthermore, Rule (5) contains multiple occurrences of some predicates such as (visits). In this case we can say rule (5) has repeated predicates and such rules are called hybrid- dimensional association rules. In contrast, multidimensional association rules with no repeated predicates are called inter-dimensional association rules.

Alternatively, a multidimensional data model is a base of data warehouse and OLAP. This model views data in a form of a data cube. By using a data cube we can model and view data in multiple dimensions. So the data cube can be defined by dimensions or attributes. Data cubes are n-dimensional such as 2-dimensional cube, 3-dimensional cube... n-dimensional cube.

Although multidimensional association rules methods have added a potential knowledge in Data mining applications, but its mining quality was a challenging issue that exhibited in many sources. To deal with this topic [Li et al. 2006] presented a multi- tier granule mining approach and provided a foundational framework that was used to represent multidimensional association rules. Through that approach, they divided attributes into some tiers and then compressed the large



multidimensional database into granules at each tier. In order to illustrate associations between these tiers, they built association mappings.

Based on the types of the handled values in the rule, [Han, Kamber 2006] categorized association rule technique into two categories: a Boolean association rule and quantitative association rule. A Boolean association rule involves association between the presence and absence of items. When a rule describes associations between quantitative items, then it is called a quantitative association rule. Some researchers' work has been performed based on using Boolean (qualitative) data. Other work has been done based on using quantitative data, and other work has been done based on mixed data (quantitative and qualitative). [Lui et al. 2006].

As many researchers meant to discuss the association rules technique through a conceptual view, others meant to discuss it through a technical view. For instance, authors [Han, Kamber 2006; Bogdanova, Georgieva 2005; Qaddoum 2009] have emphasized that association rule mining can be viewed as a two-step process: First, finding all frequent itemsets; second, generating strong association rules from the frequent itemsets. For the first step, [Han, Kamber 2006] defined principally *Frequent items* as patterns (such as subsequences, a set of items, or substructures) that appear in a data set frequently. They stated that, an itemset such as a computer device, antivirus software, and printer that appears frequently together on a transaction data set is called a *frequent itemset*. Moreover, an item *I* is considered a frequent itemset if its relative support satisfies a pre-specified minimum support threshold. At this stage, the order of appearance of each item is ignored, and basically, users do not care about determining whether the computer device or the antivirus software item appears first. They only care about their appearance together in each transaction. In other cases, ordering of items may be

important, for such cases [Han, Kamber 2006] introduced a concept called *sequential patterns*; this type of frequent patterns is used when items occur in a sequence form. In this type users often care about the order or sequence of items; for example subsequence patterns, such as buying first a PC, followed by a digital camera, and then a memory card. A frequent occurring of subsequence patterns is called *frequent subsequence patterns*.

[Han, Kamber 2006] introduced the concept of a substructure pattern; this pattern refers to different structure forms, such as, graphs, trees, or lattices. These forms can be combined with itemsets or subsequences. It is called a *frequent substructure pattern* if a substructure occurs frequently.

Often *Frequent items* are used as elements in further analysis. [Han, Kamber 2006] proposed the concept of *frequent itemset mining* which is used to discover association among items in large transactional or relational data sets. Discovery of interesting relationships among large data sets could be beneficial for many sectors. In the business sector, managers can benefit from the discovery of interesting correlating relationships in many business decision-making processes from catalog design and cross-marketing to customer shopping behavior analysis. However, a major challenge in mining frequent itemsets from a large data set is that such mining techniques often generate a huge number of itemsets satisfying the minimum support (*min-Sup*) threshold specifically when *min-Sup* takes a low value. Also when an itemset is frequent, each of its subsets is frequent as well. To solve this problem, the concepts of closed frequent itemset and maximal frequent itemset have been introduced [Han, Kamber 2006].

When we think of representing all the above types of frequent patterns in the form of association rules we don't care about the differences between them, and we use the same association rule form. [Han, Kamber 2006] considered the

Boolean variable concept to represent each item in the basket when dealing with the Market Basket Analysis. They supposed that if we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together. They concluded that these patterns can be represented in the form of association rules; as shown in the following example:

Computer  $\rightarrow$  Printer [support = 2%, confidence = 60%]..... (6)

It expressed that customers who purchased computers also tend to buy printers at the same time.

Actually, many researchers have benefited greatly from the concept of market basket analysis when they applied association rule mining techniques in their developed applications. [Huang et al. 2007; Hristovskia et al. 2001]

An association rule has the potential to generate millions of patterns or rules. However, not all the generated rules are interesting. Some of them may be trivial, whereas others may be interesting rules. Indeed, we think of keeping the interesting rules and dropping the trivial ones. However, before we think about how to drop the trivial ones, we should know how to distinguish between them, i.e. to decide which ones should be dropped and which ones should be kept. [Han, Kamber 2006; Bogdanova, Georgieva 2005; Huang et al. 2007] described the interesting patterns that are easily understood by humans. These patterns indicate a high level of validation with some degree of certainty and are useful and novel.

Towards this direction, support and confidence are used in association rule to measure its interestingness and then to drop the trivial rules.

Considering the association rule in rule 1:

$$X \rightarrow Y [s,c] \dots\dots\dots (1)$$

Where,  $X \subseteq I$ ,  $Y \subseteq I$ , and  $X \cap Y = \Phi$

And consistent with [Han, Kamber 2006; Jadav, Panchal 2012], the support is denoted by (s) and confidence is denoted by (c). These are two parameters used to measure the association rule. They are usually associated with each discovered rule to measure its interestingness. Typically, we consider association rule to be interesting if its support and confidence satisfies both a minimum support threshold and a minimum confidence threshold. Fortunately, such a threshold can be set by users or domain experts. For a dataset D, the support S represents the percentage of transactions from a transaction database that the given rule satisfies. This is taken to be the probability,  $P(X \cup Y)$ ; where  $X \cup Y$  indicates that a transaction contains both items X and Y. We can express support S by the union of itemsets X and Y.

The itemset support with the form  $P(X \cup Y)$  is sometimes referred to as relative support; whereas the occurrence frequency is called absolute support. The relative support of an itemset was defined by [Han, Kamber 2006] as shown in rule 7:

$$\text{Support}(X \rightarrow Y) = P(X \cup Y) \dots\dots\dots (7)$$

The confidence  $c$  in the transactions in dataset  $D$  is the percentage of transactions in  $D$  containing  $X$  that also contain  $Y$ . This is taken to be conditional probability,  $P(Y|X)$ .

$$\text{Confidence } (X \rightarrow Y) = P(Y|X) \dots\dots\dots (8)$$

The conditional probability  $P(Y|X)$  according to Bayes theory takes the form:

$$P(Y|X) = P(X \cup Y)/P(X) \dots\dots (9)$$

But as we have seen in (7)  $P(X \cup Y)$  represents the support of the itemset  $I$ . Therefore, we can express confidence as follows:

$$\text{Confidence } (X \rightarrow Y) = P(Y|X) = \text{support } (X \cup Y) / \text{support } (X) = \text{support count } (X \cup Y) / \text{support\_count}(X) \dots\dots\dots (10)$$

Support count of an itemset mentioned in equation (10) represents the occurrence frequency of that itemset in all transactions or the number of transactions that contain it. Equation (10) shows that the confidence of rule  $X \rightarrow Y$  can be easily derived from the support counts of  $X$  and  $X \cup Y$ .

However, [Han, Kamber 2006] expressed the support and confidence by using conditional probabilities, but [Huang et al. 2007] have taken another way to express them as they have considered the concept of mega-transaction itemsets instead of traditional transaction itemsets. And they have described support and confidence as follows:

$$\text{Support} = T_{xy}/S \dots\dots\dots (11)$$

$$\text{Confidence } T_{xy}/T_x \dots\dots\dots (12)$$

Where  $T_{xy}$  is the set of transactions that contains both items, X and Y or  $X \cup Y$ ,  $T_x$  is the set of transactions that contains X, and S is the total number of all transactions in the transaction database. Equation (11) has described *support* as number of the transactions containing both itemsets X and Y among the total number of all transactions in the transaction database. Nevertheless, equation (12) has described *confidence* as the number of the transactions containing both itemsets X and Y among the set of transactions that contains X.

As we have seen in equations (11) and (12) support and confidence can be expressed in a percentile view. To better understand the representation of support and confidence by percentile measurements, consider the following example, where X is a variable representing a tourist:

$$\text{Visits}(X, \text{“Sudan in Ramdan”}) \rightarrow \text{Interested\_in}(X, \text{“Hilomurr”}) [0.5\%, 90\%] \dots \dots \dots (13)$$

\*Hilomurr (Apray) is Sudanese drink. Usually Sudanese people drink it during the month of Ramadan only.

This rule expresses that a tourist who visits Sudan in Ramadan has a 90% chance of being interested in Hilomurr, and 0.5% of the all tourist’ data belong to this category.

[Han, Kamber 2006; Qaddoum 2009] stated that, an association rule is said to be strong if it satisfies both of the minimum support minsup and the minimum confidence minconf thresholds. While [Han, Kamber 2006] explained the role of support and confidence measures in solving the problem of rule interestingness, [Lui et al. 2006] proposed and developed a novel approach to find the interesting

rules among many others that were often generated. All the possible generated rules could be obtained by setting both minsup and minconf thresholds to 0. However, this causes a problem of combination explosion, and a large amount of context information, which increases difficulty of rule analysis. To solve this problem, [Lui et al. 2006] decided to work on a data set of a major application for Motorola, One attribute of this data set was classified as a class attribute with discrete values. It indicated the final nature of the call such as *failed during setup*, *dropped while in progress*, and *ended successfully*.

As there was no prediction or Classification needed for this application, the objective of this work was to be aware of the data and to diagnose causes of some problems in order to solve the problems. In that work, the problems were *failed during setup* and *dropped while in progress*.

Since the data set was a typical Classification data set, rules that characterized product problems were of the following rule:

$X \rightarrow y$ , where X is a set of conditions and y is a class, e.g,

Such rules usually belong to class association rule mining techniques; which are a special type of association rules with only a class on the consequent of each rule.

Typically, by applying this rule on the applied application, y takes one of the following values: *failed-during-setup*, *dropped-while-in-progress* and *ended-successfully*. During that work, such rules have been used to help users to identify interesting knowledge.

The authors of that work examined several interestingness techniques and outlined lack of contexts as the following main shortcoming: Most techniques follow an individual treatment with rules; however, a rule is only interesting in a

meaningful context and in comparisons with others. A single rule is seldom interesting by itself and its support and confidence values will be meaningless. In this context, disability of finding generalized knowledge from rules, general impressions knowledge is much more useful than individual rules because they may reveal some hidden underlying principles; and lack of knowledge exploration tools, however these tools are important for the user to explore the rule space in order to find useful knowledge, but, many existing techniques for visualizing rules, they typically also treat and visualize rules individually.

To handle such inadequacies of examined interestingness techniques, [Lui et al. 2006] proposed and developed an OLAP approach to deal with all these problems. Through this approach, a major part of rule exploration was directed as an OLAP problem based on rule cubes. OLAP operations such as slicing, dicing, drilling down and rolling up, would be used to explore rules in order to systemically discover useful knowledge. Moreover, the OLAP approach has been proposed to assist the user in getting general impressions, and it has been proposed to provide the user with a natural way for visualization with contexts.

### **2.5.1 Computing of the support and confidence of a discovered rule**

Currently, rule interestingness is one of the biggest challenges that users of Data mining techniques face, especially in such techniques that generate a large number of rules like Association Rules Mining technique. Since support and confidence count is used in measuring the usefulness of discovered rules, computing it is considered an important issue.

To know how to compute the support and confidence of a discovered rule simply; let us consider the following example 1:



### Example 1:

Suppose that are 300 customers (or 300 transactions) have come out of a store. Their purchasing process is detailed as follows:

- 100 customers out of them purchased a computer.
- 50 purchased a printer.
- 150 customers purchased the both items, a computer and a printer.

Our goal in this example is to find out the percentage of customers who purchased a computer and also bought a printer in all transactions that involved computers (confidence), and the percentage of customers who belong to this category out of all datasets (support).

The information that customers who purchase computers also tend to buy printers at the same time is expressed in the following association rule form:

Computer  $\rightarrow$  Printer (support (%), confidence (%))

Using the given information of the detailed purchasing process, we can get:

- The total number of all transactions under analysis = 300
- The number of transactions that contain both items = 150
- The number of transactions that contain computer item = 250

Note: the number 250 has been computed by adding 100 to 150 because the computer item is included in both categories of customers who purchased a computer only (100 customers) in addition to customers who purchased both a computer and a printer (150 customers).

Now by applying equations (11) and (12) where:

$$\text{Support} = T_{xy}/S \quad (11)$$

$$\text{Confidence } T_{xy}/T_x \quad (12)$$

If we substituted  $S = 300$ ,  $T_{xy} = 150$ , and  $T_x = 250$ , then we can get

Support =  $150/300 = .5 \%$  and (**Note:** we compute the support count by getting the percentage of both items out of the entire dataset or transactions)

Confidence =  $150/50+150 = 150/250= 60 \%$  (**Note:** we compute the confidence count by getting the percentage of both items out of only transactions that contain computer).

Regarding to association rule:

$$\text{Computer} \rightarrow \text{Printer (support (\%), confidence (\%))}$$

In the process of computing support and confidence, we consider both sides of the association rule which represent the category of customers who purchased a computer and a printer together. In the support count computation, we relate this category (items purchased together) to the total number of all transactions. Conversely, in calculating the confidence count, we relate it to the transactions that contain the item(s) in the rule antecedent (left hand side to the rule).

We can compute the two measurements, support and confidence for the previous example using equations (7) and (10).

To compute support we use the rule 7:

$$\text{Support } (X \rightarrow Y) = P(X \cup Y) \quad \dots\dots\dots (7)$$

Remember that support S is the percentage of transactions in the dataset D that contain both items X and Y.

Also we can find the confidence count for the previous example using equation (10):

$$\text{Confidence } (X \rightarrow Y) = \text{support count } (X \cup Y) / \text{support count}(X) \dots\dots\dots (10)$$

To illustrate this example using the support formula:

$$\text{Support } (\text{computer} \rightarrow \text{Printer}) = P(\text{computer} \cup \text{Printer})$$

To compute confidence, remember that the confidence in the transactions in dataset D is the percentage of transactions in D containing X that also contain Y.

$$\text{Support count } (X \cup Y) = P(X \cup Y) = 150/300 = 50 \%$$

$$\text{Support count}(X) = (100+150)/300 = 250/300 = 83 \%$$

$$\text{Confidence} = \text{support count } (X \cup Y) / \text{support count}(X) = .5 / .83 = 60 \%$$

The computing process of support and the confidence count with an OLAP cube is of great concern to many researchers such as [Lui et al. 2006]. They applied the computing process of support and the confidence count of an association rule when they handle a multidimensional cube. To explain their method of such computing, we consider the following example2;

**Example 2:**

Suppose we have a data set with two dimensions, A and C. The dimension C is a class dimension, which has two values, *yes* for a tourist who is Interested in Hilomurr and *no* for a tourist who is not interested in Hilomurr. The dimension A is a month dimension and has three values, R (Ramadan), S (Suffer), and M (Moharrem). Assume that the data set has **1000** data points, and the 2-dimensional cube appears as shown in table 2.1:

**Table 2.1:** A cube with two dimensions

C	no	10	0	5
	yes	90	30	22
		R	S	M
		A		

In fact, when we deal with multidimensional cubes we follow the same way of computing support and confidence count as in traditional cases.

This is the 2-dimensional cube which gives us the following 6 (2 x3) one - conditional rules:

1.  $A = R \rightarrow C = \text{yes}$ ,

The cell value that contains  $A = R$ , and  $C = \text{yes}$  is 90.

By applying equations (11) and (12):

$$\text{A confidence count} = 90 / (90+10) = 90/100 = .9 = 90 \%$$

This rule expresses that a tourist who visits Sudan in Ramadan has a chance of a  $c = 90\%$  of being interested in Hilomurr. And 0.09 of the all tourist' data belong to this category according to the following computing of support count:

A support count in this rule is  $90/1000 = .09 = 9\%$

2.  $A = R \rightarrow C = \text{no}$

The cell value that contains  $A = R, C = \text{no}$ , is 10

A confidence count =  $10/100 = .1 = 10\%$

This rule expresses that a tourist who visits Sudan in Ramadan has a chance of a  $c = 10\%$  of not being interested in Hilomurr. And .01 of the all tourist' data belong to this category according to the following computing of support count:

A support count in this rule is  $10/1000 = .01 = 1\%$

3.  $A = S \rightarrow C = \text{yes}$

The cell value that contains  $A = S, C = \text{yes}$ , is 30

A confidence count =  $30 / (30 + 0) = 1 = 100\%$

This rule expresses that a tourist who visits Sudan in Suffer has a chance of a  $c = 100\%$  of being interested in Hilomurr. And .03 of the all tourist' data belong to this category according to the following computing of support count:

A support count in this rule is  $30/1000 = .03 = 3 \%$

Now can you compute the confidence and support counts for the rest of rules?

4.  $A = S \rightarrow C = \text{no}$
5.  $A = M \rightarrow C = \text{yes}$
6.  $A = M \rightarrow C = \text{no}$ .

As we have seen above, the computing process of support and confidence count of multidimensional cubes follows quite straightforward the same way as in traditional cases. With good understanding of dealing with support and confidence measures, we can turn out clear and comprehensible results using the Association Rules Mining technique.

### **2.5.2 Combining the Association rules mining technique with OLAP technique**

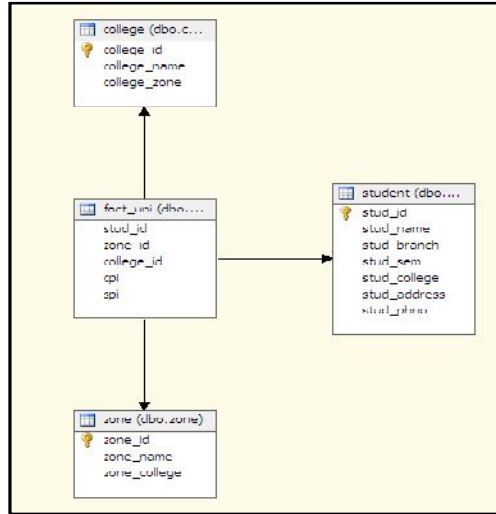
Many researchers integrated the Association Rules Mining technique with other Data mining techniques; others integrated it with OLAP. OLAP is one of the Data mining techniques that typically involve group-by and aggregation operators; its application enables data analysts and organizational managers to acquire the ability to understand the performance of an enterprise and help to make right decisions.

However, the fact that both OLAP techniques and the other Data mining techniques are used to analyze data, then, the results produced are used to support decision makers, [Han, Kamber 2006] argued that both techniques differ in their functionalities. While OLAP techniques are targeted towards simplifying and supporting interactive data analysis using summarization/ aggregation tools, other Data mining techniques are targeted to automate the process of discovering hidden

patterns and interesting knowledge as much of possible. According to this view, the authors maintained that Data mining techniques involve deeper analysis and cover much boarder applications than simple OLAP techniques because they perform not only data summary and comparison but also Classification, association, Clustering, prediction, and other tasks.

Nevertheless, a considerable amount of work by [Jadav, Panchal 2012; Bogdanova, Georgieva 2005; Dehkordi 2013] found that OLAP techniques define different operations on data cube but they cannot give the relationship between data. As a result, they combine OLAP with Data mining techniques to form the concept called OLAP mining (OLAM)

Data mining techniques are one of the OLAP outputs beside the other outputs such as database reporting tools, and data analysis tools. Data mining techniques such as Classification, association rule mining are integrated with OLAP technique to introduce a powerful paradigm i.e, on-line analytical mining (OLAM) (also called OLAP mining). Some authors were specifically interested to mining association rules in data cubes. [Jadav, Panchal 2012] used OLAM in the way of applying association rule mining technique on OLAP cube. For designing an OLAP cube, they proposed three dimension tables: *STUDENT*, *COLLEGE* and *ZONE*. All have primary key and different attributes of different dimensions. Then they designed *UNIVERSITY* table as a fact table that contains all the reference keys of dimension tables and measures that are numeric attributes according to which data will be store in to data cube. The considered measure attributes were CPI and SPI. The designed OLAP cube is displayed in Fig 2.4. With star schema is represented with one fact table; others are dimensions tables.



**Fig.2.4:** Multidimensional database display with star schema

In figure 2.5 OLAP cube was generated and different attributes in dimensions were browsed according to student name, branch semester, college name etc.

OLAP cube according to measure CPI and SPI																					
Zone		All																			
		College Name				Stud Sem				Stud Branch				Grand Total							
		CCET		KITRC		KITRC		KITRC		KITRC		KITRC		KITRC		KITRC					
		4		4		4		4		2		2		2		2					
		computer		Total		Total		Total		computer		Total		electrical		electronics					
Stud Name	Stud College	Cpi	Spi	Cpi	Spi	Cpi	Spi	Cpi	Spi	Cpi	Spi	Cpi	Spi	Cpi	Spi	Cpi	Spi				
jigna	KITRC							9	9	9	9					9	9				
	Total							9	9	9	9					9	9				
pinkal	KITRC	9	8	9	8	9	8										9	8			
	Total	9	8	9	8	9	8											9	8		
shaily	CCET							7	8	7	8					7	8	7	8		
	Total							7	8	7	8					7	8	7	8		
swati	KITRC											8	9			8	9	8	9		
	Total											8	9			8	9	8	9		
taru	KITRC													9	9	9	9	9	9		
	Total													9	9	9	9	9	9		
Grand Total		9	8	9	8	9	8	16	17	16	17	8	9	9	9	17	18	33	35	42	43

**Fig.2.5:** Browsing OLAP cube according to measure



Different frequent items and rules from the data cube have been obtained as a result of applying association rule mining technique on data cube according to  $\text{min\_sup}=4$  as shown in Figure 2.6:

Support	Size	Itemset
4	3	1 = Z1, 1 = KITRC, Spi = 8 - 9
4	2	1 = Z1, Spi = 8 - 9
1	2	Cpi >= 7.6299722448, Spi = 8 - 9
4	3	Cpi >= 7.6299722448, Spi = 8 - 9, Stud College = KITRC
4	2	Cpi >= 7.6299722448, Stud College = KITRC
4	3	1 = 8 - 9, 1 = Z1, Spi = 8 - 9
1	3	1 = 8 - 9, 1 = Z1, 1 = KITRC
4	2	1 = 8 - 9, Spi = 8 - 9
4	2	1 = 8 - 9, 1 = Z1
4	2	Spi = 8 - 9, Stud College = KITRC
1	2	1 = Z1, 1 = KITRC
4	3	1 = 8 - 9, 1 = KITRC, Spi = 8 - 9
4	2	1 = 8 - 9, 1 = KITRC
4	2	1 = KITRC, Spi = 8 - 9

**Fig. 2.6:** Frequent items according to MIN\_SUPPORT from OLAP cube

The generated rules were shown as in figure 2.7:

Pr...	Importance	Rule
1.000	0.336	Stud Name = jigna, 1 >= 7.6299722448 -> 1 = 8 - 9
1.000	0.239	Stud Name = jigna, 1 >= 7.6299722448 -> Spi = 8 - 9
1.000	0.336	Stud Name = jigna, Stud Sem = 4 -> 1 = 8 - 9
1.000	0.239	Stud Name = jigna, Stud Sem = 4 -> Spi = 8 - 9
1.000	0.336	Stud Name = jigna, 1 = Z1 -> 1 = 8 - 9
1.000	0.336	Stud Name = jigna, 1 = KITRC -> 1 = 8 - 9
1.000	0.336	Stud Name = jigna, Cpi >= 7.6299722448 -> 1 = 8 - 9
1.000	0.336	Stud Name = jigna, Stud Branch = computer -> 1 = 8 - 9
1.000	0.336	Stud Name = jigna, Stud Address = snagar -> 1 = 8 - 9
1.000	0.336	Stud Name = jigna, Stud College = KITRC -> 1 = 8 - 9
1.000	0.239	Stud Name = jigna, 1 = Z1 -> Spi = 8 - 9
1.000	0.239	Stud Name = jigna, 1 = KITRC -> Spi = 8 - 9
1.000	0.239	Stud Name = jigna, Cpi >= 7.6299722448 -> Spi = 8 - 9
1.000	0.239	Stud Name = jigna, Stud Branch = computer -> Spi = 8 - 9
1.000	0.239	Stud Name = jigna, Stud Address = snagar -> Spi = 8 - 9
1.000	0.239	Stud Name = jigna, Stud College = KITRC -> Spi = 8 - 9
1.000	0.336	Stud Branch = electrical -> 1 = 8 - 9
1.000	0.239	Stud Branch = electrical -> Spi = 8 - 9

**Fig. 2.7:** Association rules from university cube

Other studies also addressed the issue of limitation of the OLAP technology and its inability to provide users with automatic tools to explain relationships and associations within data. Therefore, some researchers applied OLAM in different applications and realized its advantages. [Jadav, Panchal 2012] applied OLAM to discover association rules in OLAP data cube with students' data. In contrast, [Bogdanova, Georgieva 2005] applied it to discover association rules in OLAP data cube using a Web based client/server system.

[Jadav, Panchal 2012; Bogdanova, Georgieva 2005] used integration of OLAP with the association rules technique to discover interesting correlation relationships among OLAP data cube; others directed their studies to apply this integration to improve the mining process of Association Rules from data cubes. For example, [Ben Messaoud et al. 2006a; Ben Messaoud et al. 2006b] proposed a general framework for mining inter-dimensional association rules from data cubes. With inter-dimensional meta-rule, users were allowed to target the mining process in a particular portion in the mined data cube.

Typically, the needed aggregate values for discovering association rules are already pre-computed and stored in the data cube. Since the dimension COUNT cell of a cube often stores the frequency of corresponding multidimensional data values, the authors explained that calculating the values of the support and the confidence of association rules is a simple process with some summary measures such as the COUNT measure. But, in an analysis process, simple frequencies are less attractive to users compared with investigating multidimensional data cube and their associations according to measures. Throughout that work, they computed support and confidence measures according to a sum-based aggregate measure where users were not restricted to analysis associations that were only driven by the traditional COUNT measure. However, support and confidence are used to evaluate interestingness of mined rules, [Ben Messaoud et al. 2006a; Ben

Messaoud et al. 2006b] proposed two additional measures; Lift and the Loevinger [cited in Ben Messaoud et al. 2006]. Finally, in order to handle the multidimensional structure of data, an efficient bottom-up algorithm was developed.

In contrast to [Ben Messaoud et al. 2006a; Ben Messaoud et al. 2006b], in which they used some OLAP cube measures to compute support and confidence values of association rules, [Allard et al. 2011] used measures and dimensions of OLAP cube to discover association rules among the most common rules. They stated that the most common rules include: Functional Dependencies (FDs); Conditional Functional Dependencies (CFDs); and Association Rules (ARs). Through that study, they tried to solve the problem of handling long lists of rules that could be displayed with many tools, and such rules are difficult to be inspected by users. Therefore, [Allard et al. 2011] proposed a new means to display and navigate through those rules. With that means they used On-Line Analytical Processing (OLAP) to display all FDs, CFDs and ARs in artificial view, through the number of values in each cell. Each set of rules was presented as a cube, where dimensions correspond to the rule antecedent and measures correspond to the rule consequent. Then, they benefited from the similarity between the form antecedent  $\rightarrow$  consequent of the dependency rules, and the form dimensions  $\rightarrow$  measure of the rules. They designed a cube to be a representation of a subset of all rules that can be extracted from a relation. In order to see the rules at several levels of granularity, they made cubes to reflect the hierarchy that exists between FDs, CFDs and ARs. Finally, [Allard et al. 2011] used a lattice of OLAP to navigate the selected rules; where nodes are OLAP views, and edges are OLAP navigation links. With that lattice of navigation, users can be guided from cube to cube, to add or to remove dimensions or to change the granularity levels.

In that work, the navigation links showed that some operators have a predictable behavior about the appearance or disappearance of rules.

### **2.5.3 Association rules mining technique applications**

Nevertheless the concept of OLAM, which integrates OLAP with Data mining techniques, still plays an important role in enhancing the Data mining process as we have seen in the previous applications. There is a tremendous need for more enhancement of the Data mining process since the Data mining techniques are used in numerous applications. The Association Rules Mining technique has been used in many domains such as medicine, economic, engineering, science, business, and many other industries. For example, in the medical domain [Hristovskia et al. 2001] used the association rule mining technique to discover the relationships between medical concepts in order to explore the new concepts.

As mentioned above, the association rule mining has a valuable role in the medical domain. It also has made a great contribution in the climate domain and has received much attention recently.

The availability of climate data has increased lately as well. Satellite and radar has made it important to find accurate and effective tools to handle the large volume of data. The Association Rules Mining technique has been used in many meteorological applications such as multi-station atmospheric data analysis. Additionally; a vast literature has been targeted to find the relationship between climate items such as the weather elements and natural events, weather and disaster prediction. For example, the association rule mining technique had been used to extract previously unknown patterns of abnormal ocean salinity and temperature variations [Huang et al. 2007]. The goal of that study was to discover temporal and

spatial variation patterns that could be used to predict ocean current variations in waters surrounding Taiwan.

In that study, in order to illustrate the strength of the discovered association rules in analyzing global climate change, [Huang et al. 2007] compared them with traditional association rules. The experimental results verified that their used model was effective in predicting the occurrence of salinity and temperature variations.

In another work using the Association Rules Mining technique to support meteorological applications, [Kohail, El-Hales 2011] applied Association Rules Mining technique among other Data mining techniques such as: outlier analysis, Clustering, prediction, and Classification. After each mining technique, they presented the extracted knowledge and described its importance in the meteorological domain.

The essential goal of that work was to extract useful knowledge from daily weather historical data collected locally at Gaza city.

## **2.6 Data mining applications**

Data mining techniques are more attractive analysis tools in many applications. These techniques provide data miners with extensive knowledge about their data. In section 2.6.1 we overview how data mining techniques have been applied in higher education domain, section 2.6.2 discusses applying data mining techniques in developing a system.

### **2.6.1 Data mining applications in the higher education domain**

As the association rules technique is applied in the meteorological domains, it has also been applied in the educational domains beside the other Data mining

techniques. [Oladipupo, Oyelade 2010] derived a method that identifies students' failure patterns using association rule mining technique which identifies hidden relationship between the failed courses and suggests relevant causes of the failure to improve the performances of low capacity students.

In other works, the association rules technique has been used in a form of an Associative Classification technique such as in [Qaddoum 2009]. The objective of that project was to discuss and evaluate a modeling approach for student evolution. [Qaddoum 2009] developed that project as a component of an adaptive achievement system.

In other works similar to [Qaddoum 2009], Data mining techniques have proven be a pioneering choice for many researchers in the higher education area in the recent years, which solved various educational problems. Students' enrollment process and predicting the academic paths of students are topics for several researchers [Nandeshwar, Chaudhari 2009; Chang 2006]. In that direction, [Vialardi et al. 2007] developed a recommendation system to provide support for the student through recommendations to better choose available courses to in which to enroll. These recommendations were based on the experiences of previous students with similar academic achievements. For that purpose a Classification technique was used on real data corresponding to seven years (2002-2008) of student enrollment.

In other applications of higher education domain, the Data mining techniques have been used for studying and predicting students' performance, For instance, [Suresh, Mahale 2011; Kotsiantis, Pintelas,2003; Al-Radaideh et al. 2006; Quadri, Kalyankar 2010; Kovačić 2010] described student performance in different university courses. In research [Al-Radaideh et al. 2006] the collected data in that

research was restricted to those students who studied the C++ course in Yarmouk University in the year 2005.

Many studies have been made in order to help students directly in making their decisions in different academic issues such as: choosing a course, prediction of students' success, students' major counseling, etc. Other works have been suggested to help higher education institutions enhance their decision making process by applying Data mining techniques to maintain some problems in the higher education domain. Some of these issues include: college transfer data, students' retention, and prediction of drop-out rates.

[Delvarai et al. 2008] proposed a new guideline in order to improve the performance indicators by developing their main educational processes for planning, evaluation, and counseling. Throughout that work, they contributed widely to the discussion on how the various Data mining techniques could be applied to the set of educational data.

Some other models have been developed to help enhance higher education institutions. [Luan 2002] introduced a case study in which he developed a model to discuss the transfer student issues. The aim of that model was to provide a profile of the students who transferred and predict which other students might transfer; thus enabling the college to provide these students with all possible assistance before they decided to transfer.

## **2.6.2 Data mining applications in developing a system**

Everyday, researchers and scientists think about how to benefit from knowledge discovered from applying Data mining techniques. Indeed, knowledge would be useless unless its benefits, understanding and use are known. Developing a system is considered an ideal use of new knowledge which plays an intermediary role between knowledge experts and end users. In this direction, some systems

have been developed in many application domains. For example, a Data mining system called “Opportunity Map”, [Lui et al. 2006], built an innovative system based on a class association rule and four basic ideas. The first idea was that the traditional mining paradigm hinders rule analysis. The second was that rules cannot be analyzed in context by using traditional mining. The third was that using OLAP operations must be used in performing rule analysis. The fourth was the mining of general impressions. Rule cubes and OLAP provide a general framework for exploration of rules in context to help the user find useful knowledge. The interesting fact is that the system has been deployed since 2006, and is used daily by Motorola.

As we have seen, Motorola applies the Data mining system, “Opportunity Map”, in its manufacturing industry and communication field.

Due to the fact that medical discoveries are increasing continuously and every day we receive a new concept in the world of medicine, [Hristovskia et al. 2001] developed an interactive discovery support system for the field of medicine to benefit medical researchers. They developed a system based on association rule mining of the Medline bibliographic database. The goal of the system was to discover new, potentially meaningful relations for a given starting concept of interest with other concepts that have not been published in medical literature before.

Developing a system based on discovered knowledge also includes some applications in the education domain. For example, [Al-Radaideh et al. 2006] built a system that facilitates the use of generated rules from a decision tree as a Classification method. They mentioned the benefit of developing such systems to help the decision makers utilize necessary actions needed to enhance the quality of the educational system.



The authors, [Vialardi et al. 2007], through their system developed to help students decide which courses to enroll in, agreed that using discovered knowledge in developing a system had a unique place in the education domain in terms of contributing to the enhancement of the educational environment.

## **2.7 An analysis of the presented literature survey**

The presented literature survey showed that Data mining techniques have been applied in many application domains such as medicine, weather, science, business, and **education**. However, the applications of Data mining techniques are still limited in different domains.

In the term of OLAP technology, most of the previous studies mainly have focused on the three storage modes: ROLAP (Relational OLAP), MOLAP (Multidimensional OLAP), and HOLAP (Hybrid OLAP). Through those studies the authors have provided us with extensive knowledge about: types of data that could be hosted in each mode, the physical location of the storage modes regarding to the data warehouse architecture, how the storage modes are associated with Data mining tools or other analysis tools, and making a comparison between them based on different factors such as memory space, speed, the required hardware.

Since a memory space is considered a headache topic for many system developers, most of the previous studies have focused on comparing ROLAP with MOLAP based on the appropriate use of each one in different applications in order to reduce the required memory space. Thus, our approach differs from theirs in terms of the goal. Our approach targets applying one of the Data mining techniques called the Association Rules Mining technique. This technique is based on a data source with formats of OLAP cubes. Integrating Association Rules Mining Technique with OLAP technique produces the term of OLAP-Mining (OLAM). Since Association Rules Mining technique often finds the associations between

items, through this approach we can solve our typical research questions such as: Are there any associations between the variation of students' faculty preferences and students' locations, in residential provinces? Do students from a certain state prefer certain colleges based on their gender, education type, and other students' categories?

Among several innovations in recent technology, Data Mining is making comprehensive changes in the field of higher education. However, using Data Mining techniques in the education domain is still limited, and its applications have been carried out for limited educational purposes. For example, the developed system in [Al-Radaideh et al. 2006] was proposed to predict the student's performance in only one course in one class. The researchers of that system have concluded from their results that the Classification accuracy for the system algorithms was not so high. That could indicate that the collected data was not sufficient to generate a Classification model of high quality.

From all the previous works and in term of using Data Mining techniques in the higher education domain, we can note that some techniques as the Association Rules Mining technique have a limited application compared with the other techniques such as Classification technique. On the other hand, the OLAP technique has never been used in any application in the higher education domain.

Although most researches in Data Mining area focus on mining data in the context of applications, Data Mining techniques has a limited range of applications in the education domain. Thus, there were few prior works that have been presented on developing a system using Data Mining techniques in the higher education domain.

We studied Data Mining related applications to draw the concepts, and then proposed an OLAP system to match the need for rapid analysis. Our proposed system is expected to be used by the Ministry of Higher Education in Sudan in order to help them in the decision making process. Since OLAP cubes often solve questions of a statistical nature, our proposed OLAP system is supposed to solve our typical research questions such as: Is there a need for all these number of faculty choices offered to students on their application forms? What percent of faculty do students usually choose on the application form? To what extent will this proposed system help decision makers in higher education institutions to make their decisions based on information obtained from it?

The literature review carried out helped us to understand the growing importance of the use of Data mining techniques in the field of higher education.

## **Chapter3**

### **Data preprocessing**

Currently, we are standing on the edge of a new age, the information age. The world is full of data. The rapid growth of data has greatly improved the smart analytical techniques which solve many problems by scientific approaches. However, accurate analyzing results depend on data quality. After we collect the data, we must enter it into computer programs such as: Access, SPSS, Excel, or any other program. During this process, whether we have entered data by hand or by a computer scanner, no matter how carefully the data has been entered, we will be almost guaranteed that there will be incorrect padding, incorrect reading of written codes, inconsistent formatting, missing data, inaccurate data, and other errors. Data are often preprocessed in order to help improve the quality of the data and, consequently, of the analyzing results.

Data preprocessing is a term used to prepare data to be ready for the Data mining process. It takes about 60% of the total time and efforts of data miners to preprocess. Data miners may need to revisit it during all knowledge discovery process steps since data contain factors of dirty data, such as: inconsistent data, missing values or other errors. They may also need to preprocess data to solve problems that can be derived from integration such as occur in redundant or inconsistent data. Such problems happen when data miners attempt to integrate multiple data from heterogeneous sources. Data miners may need to convert data from one format to another or apply various operations on data such as normalization and aggregation. They perform these tasks using one of data

preprocessing methods called data transformation. They may also use another data preprocessing method called data reduction to obtain a reduced representation of the data set that is much smaller in volume. Moreover, data miners make great efforts to restructure data that are in a form not suitable for Data mining models.

Nevertheless, data preprocessing serves as a good companion for the data miners throughout all the necessary steps in order to discover knowledge. They believe that, it has a crucial role in finding valued results of any database analysis. For example, getting good mining results often depends on the good reforming of data preprocessing. In other words, they trust that, having quality data will lead to having quality mining results.

Even so, applying data processing techniques before mining will typically improve the overall quality of the items mined, and/or reduce the required time for the actual mining.

As soon as the researchers think to develop any project or application, they ask themselves a couple of questions: what is the problem and why is it to be searched? What do they hope to achieve and how? What information is available? To start the practical part of their research, they often try to collect useful information about their problem. The main purpose of collecting data is to answer questions whose answers are not immediately observable. Practically, the researchers cannot collect data that does not require data preprocessing. Some collected data contain values that are not consistent with policy or common sense. Others contain redundant rows. Therefore, sooner or later, we find ourselves having to preprocess the collected data. Consequently, they cannot ignore the role of the data preprocessing process in finding valued results.

Data miners, on the other hand, regularly do not collect data in a suitable form for data mining models. In order to form this data in a convenient structure for data mining models and for further mining process, they often make a great effort to get data with high quality through applying a number of data cleaning techniques, and then try to store it in a data warehouse form suitable for data mining process. Therefore, data preprocessing is an essential step for any data mining process. Obtaining good mining results often depends on the proper reforming of data preprocessing. In other words, quality data leads to having quality mining results.

One of the most effective solutions for data preprocessing is Business Intelligence Development Studio, generically known as BIDS. It is a full-featured development environment for building solutions that include: Analysis Services, Integration Services, and Reporting Services projects. Such projects are specific to SQL Server business intelligence. Typically, Business Intelligence Development Studio was based on the Microsoft Visual Studio development environment, but multiple features and designing tools had been added to it so as to make it compatible with the SQL Server services (mentioned above) and project types, including controls, tools, ETL data flows, OLAP cubes and data mining structure. In this application we used some of the Business Intelligence Development Studio solutions to preprocess our collected data.

The collected data in this application contained several attributes with demographical and academic information of the students; it contained a real data that was obtained from the Ministry of Higher Education in Sudan. These historical data represent numerous records of the students who had applied to Sudanese universities and were admitted into one within the period 2005- 2009.

Although there are different types of admissions in Sudan, the collected data in this application targeted the general admissions type and reflected most of the facts about the other types. Moreover the mainstreams of students were under the umbrella of this type. The collected data were originated into two forms: hard copy of data source which included: Students' directory book and Student's application form, and soft copy of data source which include: a number of Access database files and the Ministry of Higher Education and Scientific Research web site. The collected data are described in more details in chapter 4, Data Collection (section 4.1).

Collecting data, that include: undesired formats, dirty, and inconsistency data, is an expected event for any collected data, the same as ours. To preprocess our collected data, we first cleaned them. To reduce the size some attributes have been selected to be removed. Since we targeted an educational sector, using some of data discretization and concept hierarchies' strategies is an expected issue. As some fields need to be discretized, others need to be generalized. Our collected data as mentioned above are obtained from heterogeneous data sources. Moreover, the digital data (Access database files) are organized in separated tables, using some of the integration techniques as an essential. As this is the first time for using OLAP in the Ministry of Higher Education in Sudan, data formation, compatible with OLAP and data mining techniques, is a crucial work.

This chapter describes how we can apply some of data preprocessing techniques to our case project.

In conclusion we explain how we applied some of these data preprocessing techniques on our collected data, and we establish that good knowledge about these data preprocessing techniques helps us to obtain trusted analysis results through choosing some of them that could be convenient to our case studies. Through this

application, some software, such as: Excel, database Access, and Microsoft visual studio 2008, have been used.

To obtain trusted analysis results with high quality data, we need to discuss in depth the major tasks in data preprocessing. In the following section 3.1, we focus more thoroughly on some useful data preprocessing techniques for cleaning up data. Data integration is the topic of section 3.2. We discuss data transformation in section 3.3 Data reduction is an essential for downsizing data which produces an effective environment for running of trouble-free programs and fast answering of queries, we discuss it in section 3.4. Preparing data to OLAP stage requires good knowing of structures of measures and dimensions of any fact table, we discuss this issue in section 3.5.

### **3.1 Data cleaning**

Data cleaning is a crucial part of data analysis methods, particularly when we collect data with many sorts of errors such as inconsistencies data, incorrect coding, incorrect sensing of blackened marks, missing data, and so on. Such errors contribute to reducing data quality and continue to make problems to the expected results of data analysis. So far, we usually look for some techniques that could be convenient for handling these errors (Figure 3.1). What is more, we cannot ignore the importance of the data cleaning technique in solving the problem of duplicated records. Data cleaning is the process of detecting and correcting these regular errors. Fortunately, each type of these errors has special techniques for handling it.





**Fig. 3.1:** we usually use some techniques for cleaning data

Missing data is one of the features that make data dirty and In need of cleaning. Our collected data contained an example of this type of missing data; the number of records of each table of a certain year has to be identical with the number of records of its associated table in the same year. When we have checked this issue, using Microsoft SQL Server Management Studio 2008 tools, we have found that, the number of records of table ADMFR09 table is 95112 while the number of records of ADMFR09-ch table is 95111, meaning that there was a missed record for several attributes. Then we detected and removed the extra record, since we couldn't obtain the missed record of its associated table. To check the number of records in each table, we used the SQL command called "COUNT" to detect the record that existed in one table and not in the other one. We used the SQL command called "EXCEPT".

In consistence data is one of the greatest challenges that we faced in this data preprocessing application; each college was denoted by a unique symbol, but unfortunately, this symbol changed throughout all years. For example, if the symbol 113 had been assigned for School of Management Studies at Ahfad University for Women last year, that same symbol may be used to represent College of Computer Science and Information at Sudan University for Science and Technology in the following year. Even supposing in case of the student's application form number, it was in shared ownership however it must be a unique number that identify each student. student's application form number is recycled

every year, for example if the application form number 22003 has been indicated for the student whose name was Atfa Elgamari in the year of 2005, that same number may be used to indicate the student whose name was Saadah Mohammed in the year of 2006. Unfortunately, a huge number of students had shared the same application form number during the period 2005 – 2009; for example there were 43492 students who had the same application form number within the only period 2005 – 2006. Updating all these numbers of records is multifaceted work and may generate a large number of errors.

One of the effective solutions for such problems is using the concept of composite primary key which uses more than one key column.

One of the effective solutions for such problems is using the concept of composite primary key which uses more than one key column instead of one.

### **3.2 Data Integration**

Some data analysts, such as data miners, often need to integrate data from multiple disparate data stores to reside in a single physical location, as in data warehousing. The integration is the process of merging data from multiple sources into a coherent data store. These sources may include various forms of data, such as: multiple databases, data cubes, or flat files, with different data format, such as: a database file, XML document, word document, or Excel sheet. Moreover, the data sources can be integrated from different physical locations such as: laptops, cellphones, and document files as shown in Figure 3.2.



**Fig. 3.2:** The integration is the process of combining data from multiple sources into a single data store

Integrating data from multiple sources with different formats is one of the biggest challenges facing data miners. The rest of the challenges of data integration include: Data collected in different bearings such as: structured, semi-structured, and unstructured, Data quality i.e. Data feeds from source systems arriving at different times, huge data volumes i.e. transforming the data into an appropriate format that is meaningful to data miners.

Matching up equivalent real-world entities from multiple data sources is referred to as the entity identification problem. The entity identification problem is a crucial challenge that may face data miners. For example, how can the data miners be sure that: `student_id` in one database and `st_index` in another refer to the same attribute?

For our collected data, the numbers and types of high school names in Sudan were changed last years, some high schools were canceled, some of them converted to secondary schools, and others were established To create a unified list of high schools, we had to integrate all of high schools names during the period 2005 – 2009, and then we had to find a unified list with distinct values of high school names. To integrate all the high schools names, we have used `INSERT INTO` command of SQL server. Basically, SQL Server Business Intelligence

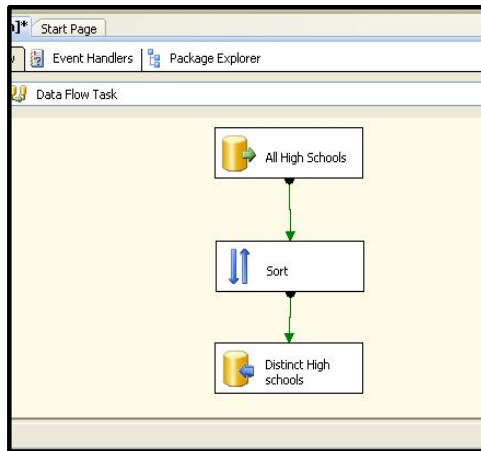
Development Studio (BIDS) have been frequently used in this section since it provides a great environment for developing SQL Server Integration Services (SSIS) packages. Each package consists of data task and Data Flows which consist of: Data Flow Sources, none or Data Flow Transformations, and Data Flow Destinations. Data Flow Sources Define the source for data for the data flow. There are a variety of Data Flow source types, with configuration options that could be appropriate to the source of the data: OLE DB source, Excel Spreadsheet source, Flat file source, and other file source.

To filter all the high schools names in order to get only a list of distinct values, we thought of a way to remove any duplicate records, so, a package of SQL server integration services have been developed (SSIS). Through (SSIS), we can retrieve the distinct records from table without needing to use an SQL Query. It provides a component of Data Flow transformations, which is called "Sort" as shown in figure 3.3, helps to implement such tasks.



**Fig.3.3:** Sort transformation data flow

The main purpose of sort transformation is sorting record based on ascending and descending order as well as Order By operation in SQL. Additionally, it does the same task of distinct operation in SQL which removes duplicated records. Figure 3.4 shows a SSIS package that we developed to catch the distinct names among all the high schools names'. In that package, all high schools table is considered Data Flow Source, is connected to Data Flow Transformations (sort), the sort Transformation sent the distinct school names to the distinct high school table which is considered a Data Flow Destination.



**Fig.3.4:** Integrating distinct high school names

The same operations have been applied on the faculty tables (FAC); all the faculty tables over a five-year period of time have been integrated and stored in a single storage location as a cohort of student’s faculty dataset. Each faculty was indicated by a unique symbol for each year. But unfortunately, when we integrated all the faculty tables in a single storage table, we found they became quite confused; more than one faculty has the same symbol along the specified period, and one faculty could take more than one symbol. To solve that problem we used a composite primary key for that table; a composite primary key for a table often consists of one or more than one column in a table. Therefore, we chose Date and Faculty’s symbol to be the composite primary key for the faculty table. Table 3.1 shows the contained fields in the final integrated faculty table.

**Table.3.1:** The contained fields in the final integrated faculty table

Symbol	Date_ID	Department_Name	Universoty_ID	University_Name	Program_Type_ID	Program_Type	College_ID	College_Name

### 3.3 Data transformation

In general, Data transformation is a process of mapping the entire set of values of a given attribute to a new set of replacement values in which each old

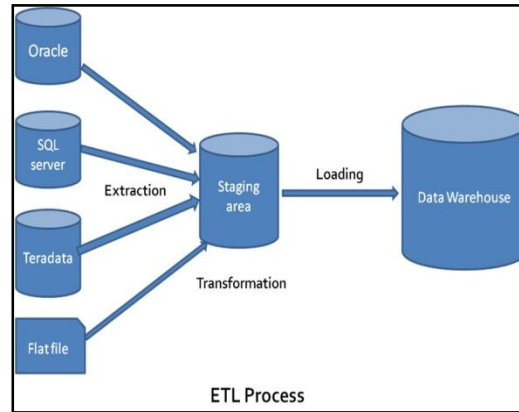
value can be identified with one of the new values; for example, to replace the string “gender” by “sex”. There are several methods for data transforming including: smoothing; which is used to remove noise from data, aggregation; which is used to summarize data and construct cubes, generalization; which considers the concept hierarchy climbing, normalization; which scales data to fall within a small, specified range, and includes the Min-max normalization, Z-score normalization and normalization by decimal scaling.

During the data cleaning process as a stage in the knowledge discovery process, some data transformations may introduce more discrepancies; alternatively, some nested discrepancies may only be detected after others have been fixed which in many cases lead data miners to revisit this stage more and more. For example, a typo such as “2002” in a year attribute may only surface once all date values have already been converted to a uniform format. In general cases, data transformation is often done as a batch process while the users wait without any feedback. As soon as the transformation is complete, the users can go back and check that no new anomalies have been created by mistake. Typically, numerous iterations are required before the users are satisfied.

Data transformation can be specified graphically or by providing examples. In some systems such as SQL server system, results are often shown immediately on the records that are visible on the screen. Luckily, the users are able to undo the data transformation, so that the latest transformed view of the data, that introduced additional errors, can be handled.

Since a data transformation maps a set of data values from the data format of a source data system to the data format of a destination data system, in the ETL (extraction/transformation/loading) process, data transformation plays an intermediary role between data source and data destination. Moreover, in the data

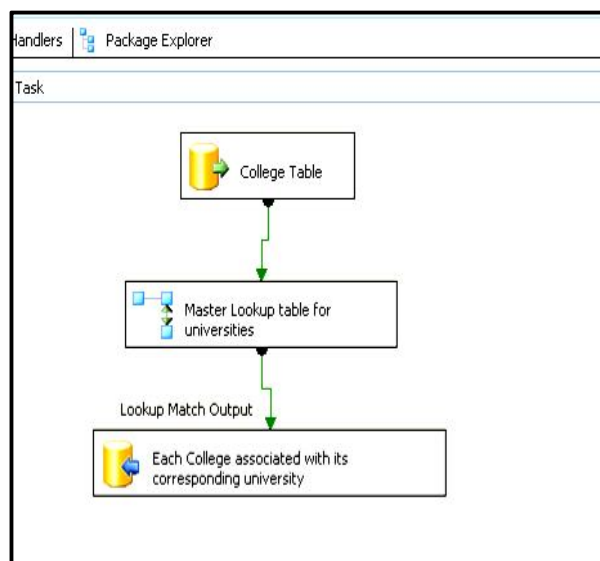
transformation process, users have a possibility to take many inputs from different data sources versus only one output/ destination as shown in Figure 3.5.



**Fig.3.5:** Data transformation participates in the ETL process

In our target application some tables were originals and already existed; some tables were derived from the original ones and added to them. How can we perform a task of joining all these tables together to create an effective database source? For example, “Major table” should be connected to” Schools table” to associate each school with its corresponding major. In another example, each college should be connected to its corresponding university and each department should be connected to its corresponding college and so on. That task might be easy and done manually if we had a limited number of records in each table. But in our case we had a large number of records, so, we developed a set of SSIS packages to transform and join all those tables together. Each package contained a Lookup Transformation tool; the basic task of Lookup Transformation is matching each record in the input table with its corresponding record in the out table. For example, to associate each college with its corresponding university we had three tables: the university table which contained fields of university ID and university

name, the master table, considered a reference table, which contained all information about all universities and their colleges and departments, and the college table which contained fields of college ID, college name, and university ID; the third field of the college table (university ID) associated each college with its corresponding university. To fill in that field, we used the Lookup Transformation which performs a lookup using columns from the input flow (the university table) and a lookup table or reference table (the master table) to retrieve additional columns from the lookup table for the output flow (the college table), either replacing input columns or selecting the additional columns to add to the output. Where records failed to be matched, the Lookup could be sent along a different output path (no matching). Figure 3.6 shows how we transformed all university's name to be joined with its corresponding college by developing SSIS package. Typically, the same process had been applied to join school tables, state tables.



**Fig.3.6:** SSIS package for colleges transformation



### 3.4 Data reduction

Recent data storage technologies such as a database or data warehouse may store terabytes of data. Such huge amounts of data can take a long time and result in complex or impractical data analysis. Moreover, difficulty of the data analysis and mining process will increase according to data enlargement. Data reduction is the process of minimizing the amount of data while producing the same or similar analytical results. Data reduction often follows effective approaches such as Data cube aggregation and Dimensionality reduction in order to downsizing data. It needs very smart users to be careful in reducing the data volume without reducing the data value figure 3.7.



**Fig.3.7:** Data reduction

We applied some data reduction techniques in our project, for example, in our collected data; each database table contained four attributes that were specified for student's full name (first name.....last name as NAME1, NAME2, NAME3, and NAME4). These attributes are pointless to our analysis purposes; consequently we have selected them to be removed. In section 3.4.1 we discuss Data discretization and concept hierarchies. We discuss data generalization in section 3.4.2.

### **3.4.1 Data discretization and concept hierarchies**

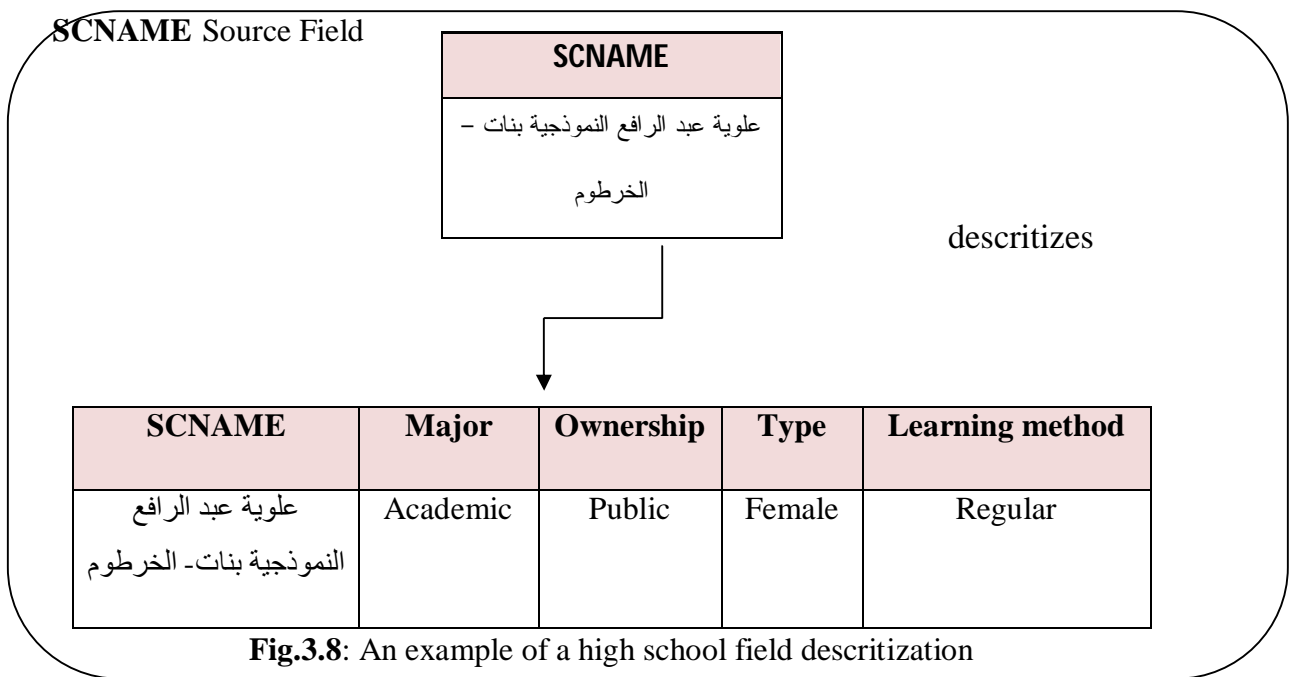
Data discretization techniques are often used to obtain a reduced representation of the data set that is much smaller in volume, under the condition of producing the same (or almost the same) analytical results. They can be used to discretize continuous data types. Numerous values of a continuous attribute leads to long-winded, difficult-to-use. With the purpose of reducing the number of values for a given continuous attribute, Data discretization techniques divide the range of the attribute into intervals. And then interval labels are used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels or higher conceptual levels thereby reduces and simplifies the original data. Therefore it is considered one of the numerosity reduction forms.

Since data discretization and concept hierarchy can be applied on the most of the attributes data types, such as: Continuous attributes, Ordinal attributes, and Categorical attributes, we have greatly benefit from it discretizing some of our categorical attributes for purposes of analyzing data in multiple levels of abstraction.

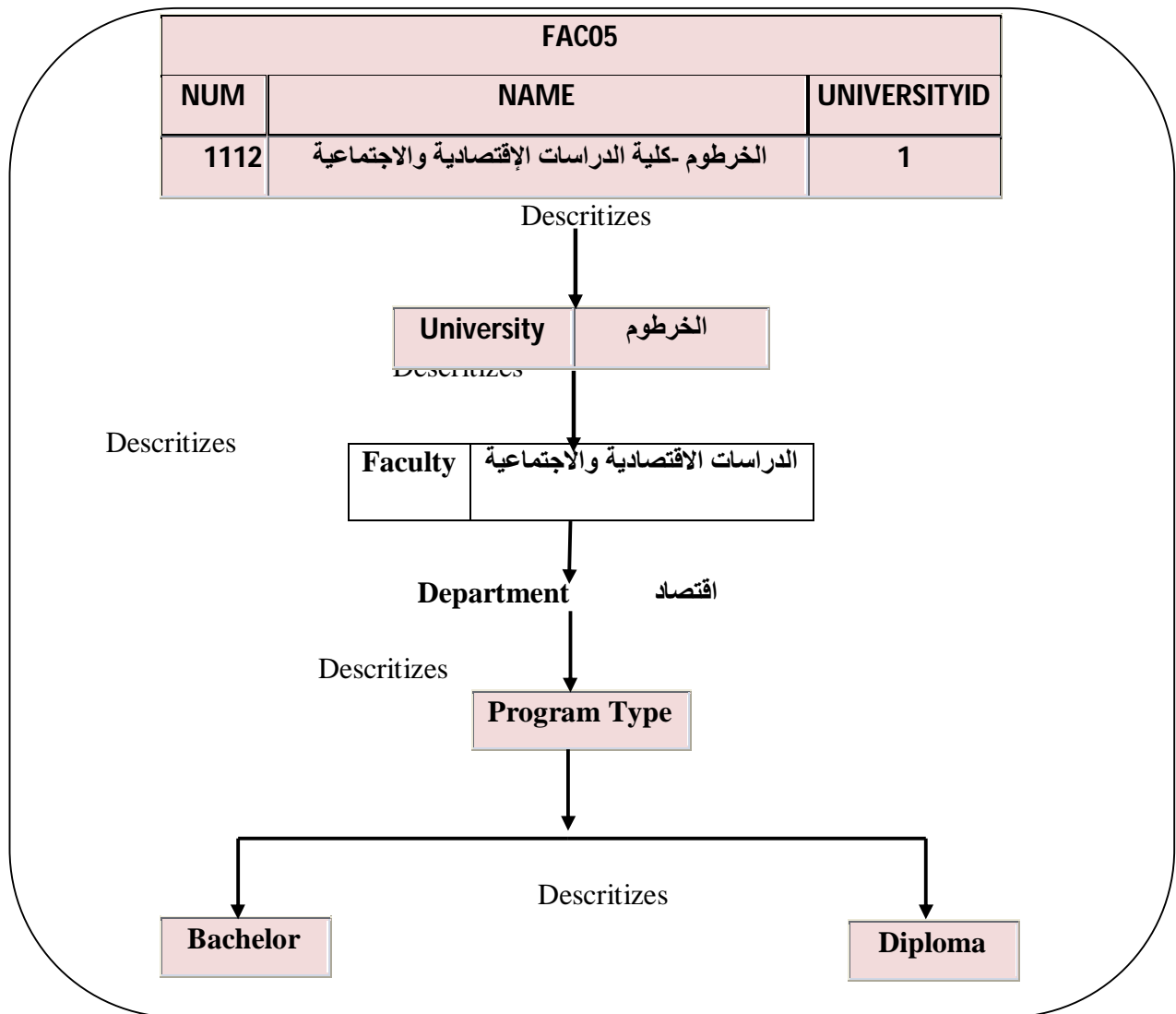
One of the great efforts that were done in the data preprocessing stage was using some fields to derive extra fields for analysis purposes. For example, the collected real data contained two fields that provided us with information about the students' high schools, the high school major program, and the high school full name (SCNAME). the field of the high school full name provided us with only full name of the high school; from this name we derived another field that contained some properties of the high school such as: the school type to determine whether it is for female, male, or mixed students, was the school a private or a public school, and the way of a student's enrollment to each school was determined; and whether

student had been enrolled to a certain school as regular, teachers schools, or home student. Finally, the field of high school name gave an explanation of the geographic location of each high school.

Figure 3.8 illustrates how a new table for a high school has been created as a result of descrittized fields from the SCNAME source field.



The same processing has been done to the faculty name field (FAC) in the Faculty tables, in which extra fields have been derived from it to create the university, program type, faculty, and department fields. These descrittized fields have been derived in order to provide us with a hierarchy concept; thus, each university has a set of faculties, each faculty has a set of departments, and each department is divided into two types of program: often Bachelor and Diploma. An example of descrittization process for faculty field is shown in figure 3.9.

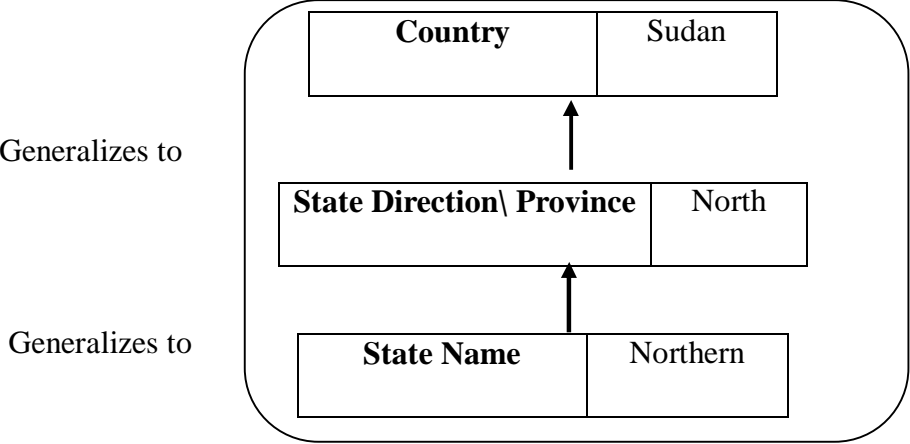


**Fig.3.9:** An example of a faculty field descritization

### 3.4.2 Data generalization

We have derived extra fields from the high school field and faculty field through descritization processes, as well as having benefited from the state source field (STAS) that represents the state's name in where student lives; from this field we have derived state direction and country fields through a generalization process.

Figure 3.10 shows an example on how this process has been done. The generalization process has been applied only on a case of students, who belong to the country of Sudan, for who belong to other countries; all the fields, Country, State direction, and State took the same name, because we never mind about state of any country other than Sudan.



**Fig.3.10:** An example of a generalization process for a state source field

### 3.5 Applying Data formation for OLAP

One of the imperative justifications that could enforce us to make a great effort in preprocessing data is that the collected data are often not in the appropriate form for our analysis purposes; however they are clean and free of noise. For example, data miners need data to be in appropriate form for OLAP. Preparing data to OLAP stage requires taking care of structures of measures and dimensions of any fact table. In this section we prepared a relational database structure to be the initial data source for OLAP.

The collected source data was already found in a relational database form, but subsequent to preprocessing data, we recreated it to produce a new database

structure. That new structure is in the best form for creating a data warehouse structure, and then for an OLAP structure.

The developed database structure contains one master table and other look up and detailed tables. Each table has been explained in details, and the relationship between them has been illustrated.

Master Table:

That table was called student table and consisted of the following attributes which have shown below in Figure 3.11:

Student_Table	
Student_ID	Primary Key
FRMNO	
Sex	
School_ID	
Department_ID	
Date_ID	
State_ID	
St#ScoreAverage	
FilledChoice	
FilledPercent	

**Fig.3.11:** Student table (Master)

The student table's attributes have been described as in table 3.2:

**Table 3.2:** Description of student table attributes

No	Attribute symbol	Attribute Description	Possible values
1	Student_ID	Student's application form number	Int.
2	sex	Student's sex	Female (f), Male (m)
3	State_ID	Foreign key for state table	Int.
4	Admission Year	Admission Year	2005 -2009
5	School_ID	Foreign key for school table	Int.

6	Department_ID	Foreign key for department table	Int.
7	Scores_Average	Student's scores average	Decimal
8	Noof_filled columns	Number of filled columns	Numbers <= 45
9	Filled Choices_percent_	Percentage of filled columns	Decimal

In our case study, it seems that the data mining process starts from the data preprocessing stage; since we have been mining data to discover methods of how to configure the collected data for analysis purposes. In doing so, the 45 fields of students' choices have been used to generate three extra calculated fields: Student's scores average, Number of filled columns, Percentage of filled columns, as shown in table 3.2. Those fields were suggested to be measures for the proposed fact table where it is crucial for OLAP. To illustrate how we generated such fields, we started from table 3.3 which shows fields of a student's choices to apply to Sudanese colleges; these choices start from the choice number1 (CH1) up to the choice number45 (CH45).

**Table 3.3:** student's choices

ADMFR05_ch						
FRMNO	CH1	CH2	CH3	CH4	CH5	.....CH45
12345	1233	1123	1143	0000	0000	0000

That table explained that each student has a right to choose and decide the number of filled columns or choices: however, not all the filled choices are matched with their associated required student's scores to be admitted in one

college. Thus, via table 3.3, table 3.4 has been created to calculate some factors of the student's application form: Number of filled columns, Percentage of filled columns, and then they were joined to the master table as shown in figure 3.11.

**Table 3.4:** Choices measures

Number of filled Choices	Percentage of filled Choices
3	10%

In the tables below, a new calculated field called scores average has been added to table 3.6 based on information in table 3.5. The added calculated field in table 3.6 calculates the scores average for each student, where table 3.5 contains the scores average by subject for each student's choice.

**Table3.5:** Scores average by subject for each student's choice

ADMFR05_ch						
FRMNO	AVS1	AVS2	AVS3	AVS4	AVS6	..... AVS45
12345	80	75	90	0000	0000	0000


**Table3.6:** Student's scores average

ADMFR05_ch							
FRMNO	AVS1	AVS2	AVS3	AVS4	AVS6	..... AVS45	scores average
12345	80	75	90	0000	0000	0000	<b>88.95</b>



Beside creating measure fields essential for OLAP structure, data type is essential in creating any database structure as shown in table 3.2. The data types and possible values for student table's fields is described; for example, some were integer and other was decimal. In general, data types are divided into two main categories: numeric and nominal data. The student's scores average were collected in the original data and were in a form of float data type, which contained a large number of digits. We had to convert those data to decimal data type. Decimal data is one of the numeric data types. It is just numbers that include decimal points. Decimal data contain two parameters: precision and scale. The precision is the total number of digits can be stored to both sides of the decimal, to the left as well as to the right. The scale is the maximum number of digits that can be allowed to be stored to the right of the decimal. For example, 4 is the precision and 2 is the scale of a decimal (4, 2), that means for the decimal (4, 2), only 4 digits are allowed to be stored with 2 of the digits to the right of the decimal, as in the student's scores average (88.95).

Calculated fields were existed in the master table in addition to primary keys of other tables; those primary keys directly connect the master table with other tables, and consequently other tables directly were related to other tables. We will discuss such relationships later in this section. The tables were directly related to the master table included: School table, State table, Department table, and Date table. School table is a look up table for school\_Id in the master table, and consists of the following attributes which have shown in Figure 3.12:

Schools_tbl	
	School_ID
	School_Name
	School_Major
	School_Type
	School_Ownership
	[School_Learning methods]

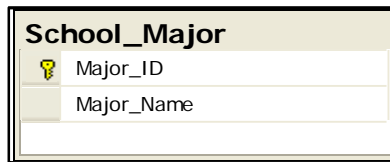
**Fig. 3.12:** School look up table attributes

Its attributes have been described as in table 3.7:

**Table 3.7:** Description of school table attributes

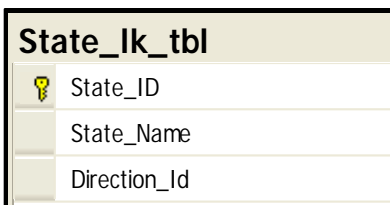
No	Attribute symbol	Attribute Description	Possible values
1	Sch_ID	School's identity number	numbers
2	Sch-Name	School's name	names
3	Major_Id	Foreign key for major table	Academic(1), Commercial (2), Industrial(3), Agricultural(4), Hafazah (5), Scientific Institutes (6), Nisweya (7), Arabic Certificates (8), Aedeen (9).
4	Sch_type	School's type	Female (f), Male(m), Mixed
5	Sch_ownership	School's owner	Public(P), Private(V)
6	Learning_methods	Learning_methods	Regular(R), teacher schools(T), home(H)

The school table was related to look up table called Major table which specify the major of each school. Major table fields are shown in Figure 3.13



**Fig. 3.13:** Major table

Other table is State table, that table is a look up table for state\_Id in the master table, and consists of the following attributes which have shown in Figure 3.14:




**Fig.3.14:** State look up table attributes

State table's attributes have been described as in table 3.8:

**Table 3.8:** Description of state table attributes


No	Attribute symbol	Attribute Description	Possible values
1	State_ID	State's identity number	numbers
2	St-Name	State's name	names
3	St_Dir_Id	Foreign key for directory table	North(1), middle (2), west(3), East(4), South (5)

This table was related to other look up table that associates each state with its corresponding direction, called Direction table, Figure 3.15 shows Direction table's attributes

State_Direction_Ik_tbl	
	Direction_ID
	Dierction_Name
	Country_ID


**Fig. 3.15:** Direction table attributes

Sudanese students can take their high school examinations at different countries; therefore, country table was crucial. Figure 3.16 shows country table attributes.

Country_Ik_tbl	
	Country_ID
	Country_Name

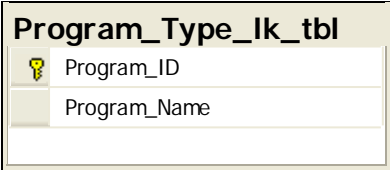
**Fig. 3.16:** Country table attributes


On the other hand, Department table has taken a place among other tables that directly related to the master table. Figure 3.17 shows its attributes.

Department_IK_tbl	
	Department_ID
	Department_Symbol
	Date_ID
	Department_Name
	Program_Type_ID
	College_ID

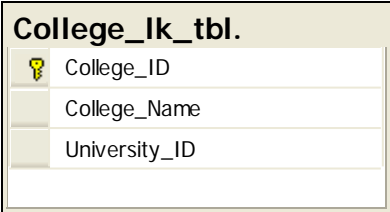
**Fig. 3.17:** Department table attributes


As shown in figure 3.17, the department table was related to two other tables: college table and program type table. Whereas the program type table attributes are shown in figure 3.18, the college table attributes are shown in Figure 3.19.



Program_Type_Ik_tbl	
 Program_ID	
Program_Name	

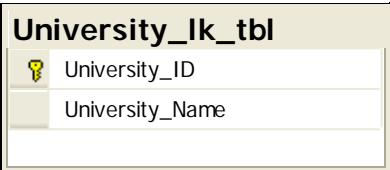
**Fig. 3.18:** Program type table attributes




College_Ik_tbl.	
 College_ID	
College_Name	
University_ID	

**Fig. 3.19:** College table the attributes

The college table was related to university tables which its attributes are shown in Figure 3.20.




University_Ik_tbl	
 University_ID	
University_Name	

**Fig. 3.20:** University table attributes


Since we developed this project for using OLAP, great efforts have been done to preprocess data to be in a form that could be appropriate to OLAP. Accordingly, creating a DATE table that could be converted to dimension was very

essential to compare the analysis results along years; we created a date table to cover the data analysis within the period 2005 – 2009. That table indicated the student’s admission year. Therefore, the old field in the data source (CDATE), where student’s year of application was indicated, has been updated to be suitable for date format. Figure 3.21 shows the date table attributes.


Date	
 Date_ID	
Admission_Year	

**Fig. 3.21:** Date table attributes

As we explained before, the calculated fields based on students’ scores average and students’ choices fields were created; we created two tables, scores and choices to be sources for later measures. Figures 3.22 and 3.23 show the attributes of the scores and choices tables respectively.

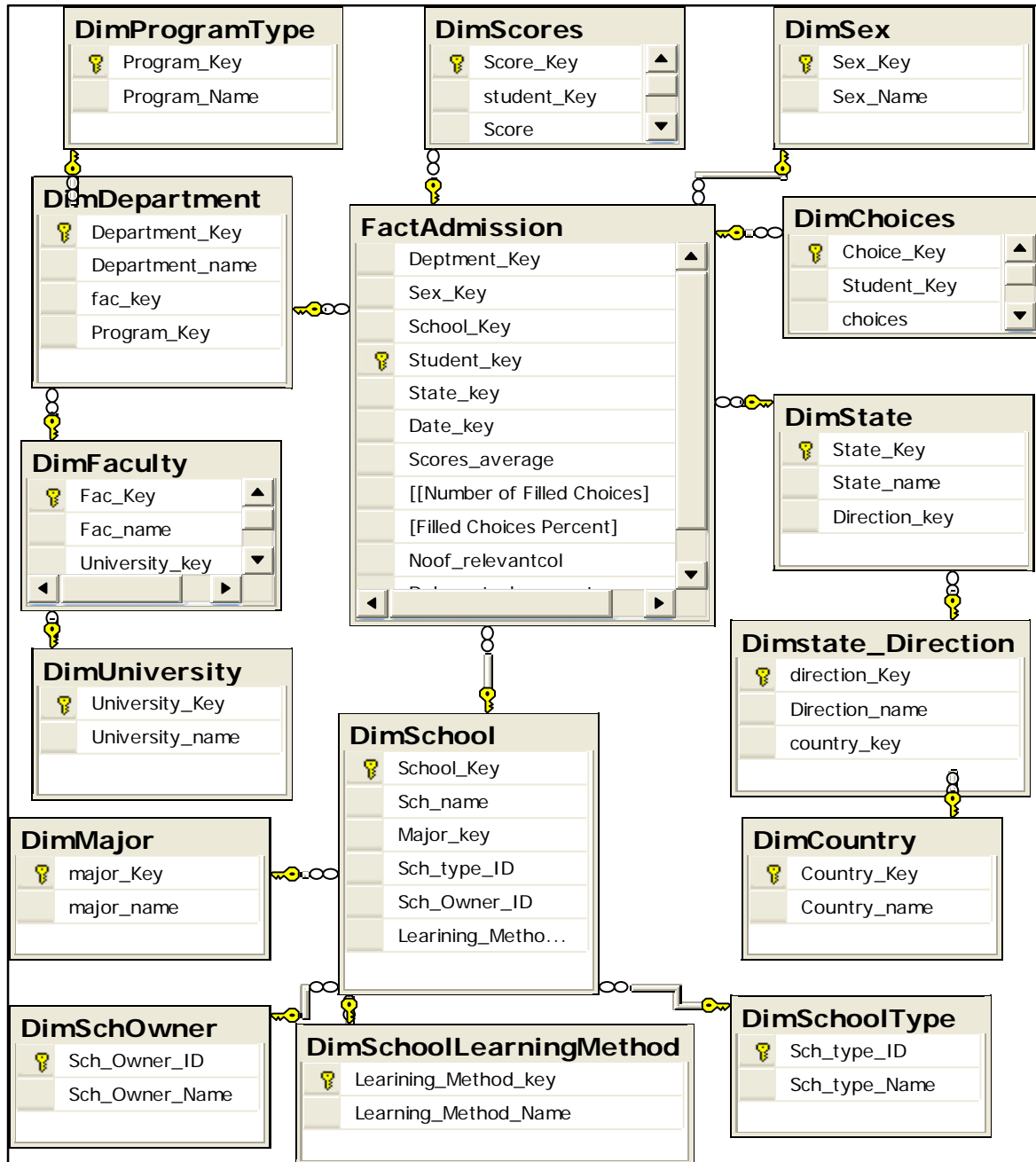
Student_Scores	
 Score_ID	
Student_ID	
FRMNO	
Score	

**Fig. 3.22:** Scores table attributes

Student_Choices	
 Choice_ID	
Student_ID	
FRMNO	
Student_Choices	

**Fig. 3.23:** Choices table attributes

The proposed relational database structure is shown in figure 3.24:



**Fig.3.24:** The proposed structure of the relational database

All the relationships between the master table and other tables is a one-to-many relationship, as well as the relationships with the scores and choices tables,

where each student can apply to more than one faculty choice and each score is associated with score.

In this chapter we provided an overview of some features of a data preprocessing stage. We outlined numerous reasons for missing values which may include: equipment malfunction; inconsistent with other recorded data and thus deleted; i.e. data not entered due to misunderstanding, not be considered important at the time of entry or history or changes of the data not registered.

Through that data preprocessing stage, we performed some preprocessing techniques including: data cleaning, reduction, integration, descritization, generalization, transformation, and formation. We converted Access tables to SQL server databases, joined all the data source tables to create a single table; created single tables that contained distinct values associated with their lookup tables. We used common sense and domain knowledge to remove some attributes such as Students' full names, and mined only selected fields. Finally we selected the final admissions tables which included: student's high school score, high school major program, student's high school, student's gender, student's state, admission year, and student's college.

We performed basic operations; those operations included: modification of some tables, addition of some tables, and choosing of some fields. In the next step we prepared data in a form that could be appropriate for OLAP. Therefore, we developed the relational database structure, and made it ready for use in developing a data warehouse structure and then an OLAP model. Thus, the data will be ready to use as a data source for the next data mining stages such as the mining process.



Although numerous methods of data preprocessing have been developed, the preprocessing takes the greatest efforts of data miners in the data mining process. Moreover, data preprocessing remains an active area of research, due to the huge amount of inconsistent or dirty data and the complexity of the problem, and we cannot ignore the fact that good data preprocessing is a key to producing valid and reliable models and then leading to quality mining results.

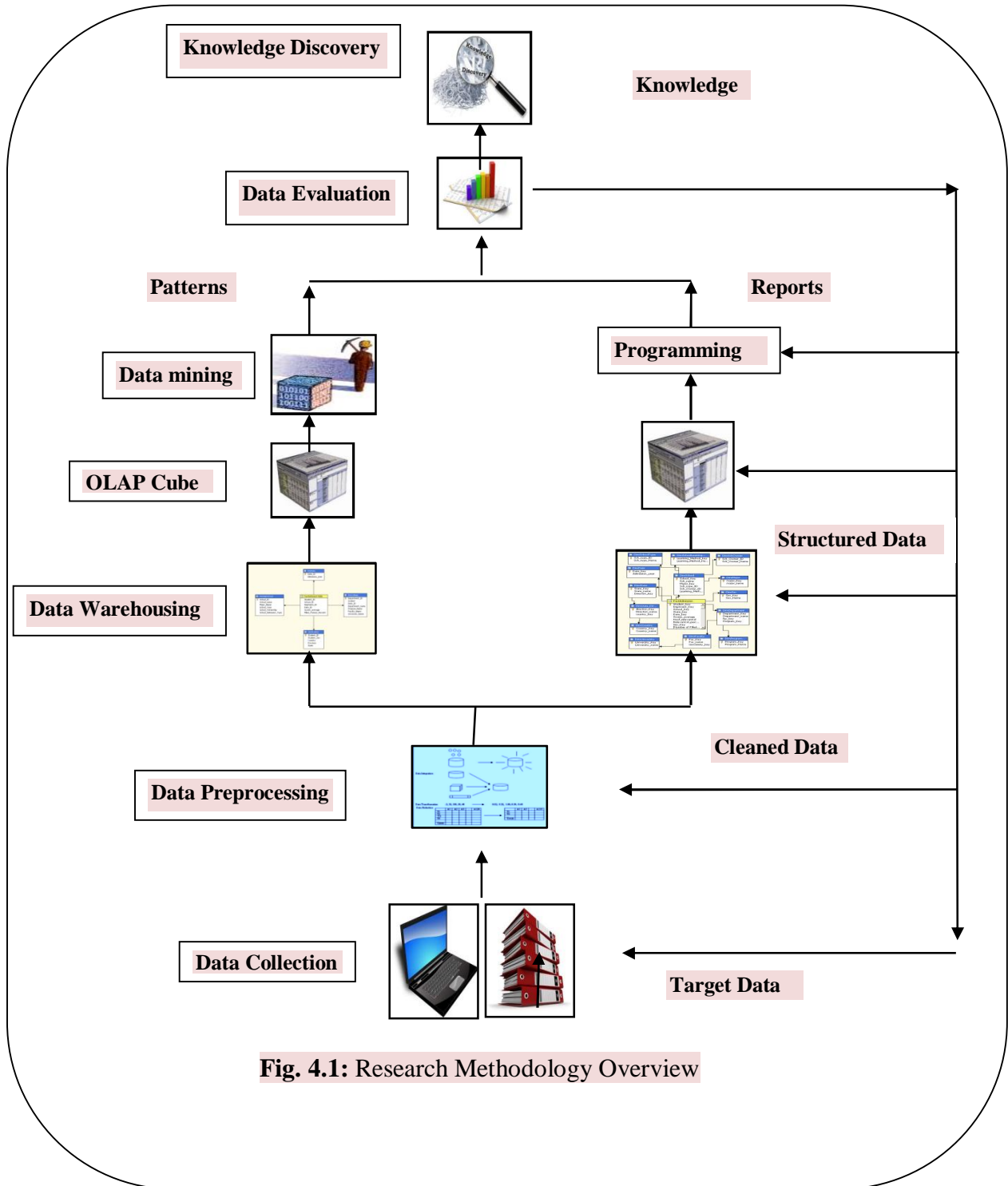
## Chapter 4

### Methodology

After the in-depth study of the topic pertaining to the data mining and its application in the higher education domain, the methodology in this research is adapted. Figure 2.2 describes the steps of knowledge discovery process. Figure 4.1 shows how we hierarchically described our adapted research methodology, which covers all the important steps that we used in the practical part of the research. Referring to figure 4.1 the methodology starts by: collecting data from two types of data sources, hard copies and soft copies, preprocessing the collected data using different tools of data preprocessing (which is described in more details in the previous chapter i.e. data preprocessing), using the preprocessed data to build two data warehouse structures, snowflakes and star, and using both data warehouse structures to develop two separated OLAPs. The OLAP that developed by star data warehouse structure was associated with the association rules mining technique for data mining purposes, which provides data miners with patterns of rules. The OLAP that developed by snowflakes structure was programmed to a smart user interface which provides users with quick reports. All reports and patterns were evaluated to discover knowledge.

This chapter consists of two sections: data collection and description of the followed methods in order to analyze the collected data. We introduce a brief description of collected data in section 4.1. Data warehouse structure mostly is used as a base for OLAP structure; we discuss some issues related to it such as its importance in the analysis process and its proposed design structure in section 4.2. The targeted topic of section 4.3 is OLAP where we confer about the proposed

OLAP structure and how we programmed OLAP to end user. Some data mining structures and models are discussed in section 4.4.



**Fig. 4.1:** Research Methodology Overview

## 4.1 Data Collection

Data Collection is considered the first step of the practical stage of any database analysis research. In the scientific research process of most research fields such as computer sciences, economical and social sciences, chemistry, business, humanities ., researchers often define their research problem through their heavy reading scan in the literature review. Regularly, they define their research problem when they find any gaps in the reviewed literature. Once they specify their research problem, they start the stage of data collection. Data collection is merely gathering and measuring useful information on variables of interest. Researchers often plan to collect useful data that enables them to answer stated research questions, test hypotheses, and evaluate outcomes. However, they do not ensure that they will obtain relevant or specific enough data to fit their research purposes.

The data collection methods are variants based on the research field. The main data collection methods include: Registration, Questionnaires, Experimental measurements, Interviews, Direct observations, Internet websites, and Reporting. Despite the fact that these methods diverge by discipline, ensuring accurate, appropriate and honest data collection remains the same. There is not any doubt that, a high quality of data collection is essential to maintaining the integrity of research and to reducing the likelihood of errors occurring. On the other hand, getting excellent results in database analysis area necessitates good knowledge of how and what information has been collected before data is processed, and understanding of the nature of that data to facilitate later analysis process.

For higher education sectors in Sudan, the Ministry of Higher Education and Scientific Research, the state's higher education governing body, is

responsible for collecting, and analyzing higher education statistics from educational institutions such as universities and other institutions for public and private educational sectors. For application and admission processes, students usually apply to Sudanese universities using a standard application form prepared by the Ministry of Higher Education. And then, to improve timely and accurate data collection, those application forms are often collected from northern, eastern, western, and southern states and branches outside the country. Data entry is usually performed manually. After executing many statistical processes, the admission results for applicant students are released on the web site of the Ministry of Higher Education, and other statistical results about percentages of admissions are disseminated annually in a directory book of Higher Education for admissions.

In the last years, database managers of the Ministry of Higher Education, computerized all their collected data and made great efforts to update it as much as possible. They stored their collected data in a form of Ms. Access software.

In general, a complete set of individual records would contain comprehensive information on students' demographic such as: gender, age, region, high school name, and so on, academic factors include: student's enrollment data students' admissions index number, college name/college symbol, and student certificate information such as high school scores. On the other hand, information about students, who had applied to Sudanese universities but were not admitted, is excluded.

In this section, we identify the original source of the collected data by explaining from where the applicable data were collected in section 4.1.1. We present a detailed description of the collected data along with their original formats in section 4.1.2.

#### **4.1.1 The original source of the collected data**

There are different types of admissions in Sudan, such as: general admission, special admission, special admission for academic staff, and special admission for Darfourian students, as well as other categories. This study concentrates on the general admission type because it involves the majority number of students. In addition, this type could be considered as representative of the other types of admission; without ignoring the fact that all other types of admission are affected by it.

These data represent records of students who had applied to Sudanese colleges and records of students who were admitted yearly between the periods of (2005- 2009). Among the total number of students who had applied to the Sudanese colleges yearly, a large number of students were admitted into these colleges. However, a huge number of students also did not find a place in these colleges for different reasons, or they did not qualify to meet the required scores for admission in the desired colleges. So their names are automatically dropped from the list of the admitted students. As a result, the collected data represent the actual number of students who were admitted as well as the total number of all students who had applied to Sudanese colleges within the period 2005 - 2009.

In this study, a large dataset has been collected as a secondary source of data from the Department of General Directorate for Admissions, Certificates' Authentication and Accreditation at the Ministry of Higher Education and Scientific Research in Sudan.

#### **4.1.2 Describing the collected data**

Since the collected data were collected as a heterogeneous data source, these data were collected into two types of data sources: a hard copy of data source and a soft copy. Both have been detailed as follows:

A) *The hard copy of the data sources:*

This type has been represented by two types of collected data: students' directory book and students' application form. Each of them has been explained as follows:

*Students' directory books:*

Every year the Ministry of Higher Education and Scientific Research in Sudan, department of general directorate for admission provide students with directory books. The book often consists of about 253 pages, and is divided into 8 chapters. The first chapter contains information about admission regulations for applying to the higher education institutes. The second chapter overviews the available public sector of the higher education institutes. The third chapter contains information about the available private sector of the higher education institutes. The fourth chapter explains the applications steps that each student must follow. The fifth chapter explains the required subjects that qualify each student for applying to corresponding college. The sixth chapter lists names of colleges with their corresponding symbols for general competition. The seventh chapter lists names of colleges with their corresponding symbols for less developed states. The eighth chapter lists names of the technical diplomas and their regulations and symbols. Each book of a current year contains limited information about the required scores for admission in Sudanese colleges in the previous year. For

example, the students' directory book that had been published in 2006 contains the required scores for admission in Sudanese colleges for year of 2005 and so on.

*Student's application form:*

Admissions data were initially designed based on information provided by students themselves in the students' application forms.

The initial dataset in the application form consists of the following data elements:

Application form number, Student's name, Admission type, Gender, Faculty choices number, Year of birth, Place of birth, Religion, State residency, Major type, Family Income size, High school name, High school graduation year Associate marks by subjects, Associate degrees earned

*B) The soft copy of the data sources:*

This type of data source is represented by two categories: Access database files, and the Ministry of Higher Education and Scientific Research web site.

*Access database files:*

This access database file consists of 16 tables that cover the period of 2005-2009, where each year is spread across three tables: ADMFR table, ADMFR\_ch table, and a FAC table, in addition to one table for all years which is a lookup table for States. For example the three tables of the year 2005 are; ADMFR05, ADMFR05\_ch, and FAC05 as shown in tables 4.1, 4.2, and 4.3 below.

**ADMFR table:**

The total number of records within the mentioned period was 4,458,200 records, with each record containing eight fields:



**Table 4.1:** ADMFR05

ADMFR05							
FRMNO	NAME1	NAME2	NAME3	NAME4	CODE	SCNAME	FAC
12345	سعاد	يونس	محمد	الدسوقي	1	علوية عبد الرافع النموذجية بنات	1123

- The first field is called FRMNO. This field represents the student's application form number; the data type in this field is numerical. Since each student has a unique application form number, this field is considered as a primary key field.
- The next four fields indicate the full student's name (first name, middle name, sure name, and the last name) which is broken down into the four fields; NAME1, NAME2, NAME3, and NAME4. The data type in this field is categorical.
- The sixth field is CODE. This field represents the students' high school majors which are coded by numbers. The data type in this field is numerical. There are nine majors of high schools that are coded as follows: Academic = 1, Commercial = 2, Industrial = 3, Agricultural = 4, Hafazah = 5, Scientific Institutes = 6, Nisweya = 7, Arabic Certificates = 8, and Aedeen = 9.
- The seventh field is SCNAME. This field shows all the complete names of high schools from which students had taken their final examinations to qualify for applying to Sudanese colleges. The end the NAME field. The data type in this field is categorical.

- The eighth and last field of the ADMFR table is FAC field. This field contains a number of symbols that indicate to faculty associated with a certain universities which students were admitted. The data type in this field is numerical. Some values in this field have the number zero which points to students who withdrew their admission application and cancelled their actual enrollment for different reasons. For example, some students did not know the state location of the college where they had applied. Others decided to apply late to private universities instead of public universities or vice versa. There are also some students who repeated the academic year.

**ADMFR\_ch table:**

The total number of records in this table within the mentioned period was 4,458,200 records; each record contained 98 fields as shown in table 4.2 and was detailed as follows:

**Table 4.2:** ADMFR05\_ch

ADMFR05_ch						
FRMNO	SEX	CH1 – CH 45	AVS1 - AVS45	AV1	CDATE	OLD_STATE
12345	1	1123	.....	...	5	1

- FRMNO field is considered a primary key, so this field is included in this table and represents the first field of the ADMFR\_ch table. It represents the student’s application form number; the data type in this field is numerical.

- **SEX:** is the second field of ADMFR\_ch table. It represents student's sex and is coded as 1 for female and 2 for male students; therefore the data type in this field is numerical.

The fields from CH1 up to CH45 each represent the choice number for each faculty; there are 45 faculties offered in many universities. Students can choose some or all of them in order according to their desire. The data type in this field is numerical; each number represents a certain faculty associated with a certain university.

Fields from AVS1 up to AVS45 indicate Average of student's grades that met the required percentage for admission in a desired faculty. Each percentage is corresponded to a certain faculty example: AVS1 is the required percentage for faculty choice CH1, AVS2 is the required percentage for faculty choice CH2.....etc, the data type in this field is numerical too.

- **AV1:** Average grades for competitor students when they have the same grade and the same desired faculty, but there is only one offered place in the desired faculty. The data type in this field is numerical.
- **CDATE:** represents student's year of application,
- **STAS:** represents the state's name in which student lives. The data type in this field is categorical.

The number of records of ADMFR09 table is 95112; while the number of records of ADMFR09-ch table is 95111.

### **FAC Table:**

The faculty table consists of the following three fields: NUM, NAME, and UNIVERSITYID.

**Table 4.3: FAC05**

FAC05		
NUM	NAME	UNIVERSITYID
1112	الخرطوم-كلية الدراسات الإقتصادية والاجتماعية	1
1312	كلية الهندسة – بكالوريوس الهندسة الكهربائية- السودان	3

\* The web site of the Ministry of Higher Education and Scientific Research is used as a second soft copy of the data source. Fortunately, we have got most of the useful information from the web site early, because the web site has been updated during the research's years.

After the data were collected from different data sources, they will be ready for integrating them to build a data base in one storehouse, which is a data warehouse structure. Data warehousing is the topic of the next section.

## **4.2 Data warehouse**

Data warehouse is an appropriate dimensional repository of information collected from multiple sources. We need it when the relevant data are spread out over several databases and physically located at dissimilar and numerous stores. The core of data warehouse is to collect data from heterogeneous sources to be

located at a single store and sorted under a unified schema. Data in a data warehouse repository must be preprocessed well, that is, must be cleaned, transformed, integrated, loaded, and refreshed. Collecting data in a data warehouse implements direct querying and facilitates the process of getting the relevant data that requested by the data analysts for the purpose of decision making.

We have argued our need for data warehousing in our case project in section 4.2.1. Section 4.2.2 will introduce the proposed data warehouse structure.

#### **4.2.1 What is the importance of a data warehouse?**

In order to understand the importance of using a data warehouse throughout this research, we overview a brief description of the used database structure that we have replaced with a data warehouse structure; for that, we first have described the admission scenario that is applied by the Ministry of Higher Education and scientific Research.

**Admission scenario:** The application process starts in June, and as soon as the application cycle for higher education begins, the student goes to an enrollment center to get an application form and his/her directory book. This process is very difficult due to a large number of applicants in addition to the lack of enough time to complete applications. The necessity of having to complete the application on time causes students to make the future decision of their education quickly and may be in appropriately. However, these students are rarely able to make wise decisions themselves, and often rely on their parents or relatives to choose which college they should attend.

The application form consists of two semi-separated parts. The first part contains information about the student associated with her/his obtained grades, in addition to an empty table for 45 college choices which the student must fill in

with his/her desired college. The second part of the application form contains the same student's application ID that is specified in the first part and the student's index number. When the student returns to the enrollment center to submit his/her completed application, the admissions officer separates the two parts of the application. The first part remains with the admissions officer and the second part, which contains the student's index number, remains with the student for following up with the admission process. The index number is necessary for finding out into which college/university a student has been admitted. However, unfortunately, this part of the application can be lost when the admission decisions are released.

After students submit their completed application forms, a stage of data entry process starts. The data entry is a process of converting the information in the students' application forms from hard copies into soft copies using computers. This processing is done in two stages. The first stage is done at the Ministry of General Education where part of the student's application form is entered into computers. Often, the student's demographic information and the student's scores breakdown is entered in this stage.

After that, the second stage starts during which the rest of the student's application form will be entered in computers in the Ministry of Higher Education and Scientific Research. In this stage, the number of the student's choices is entered along with other additional information such as the student's average score. At the Ministry of Higher Education and Scientific Research, all the students' records are stored in several separated access files which contain relational databases. It is here that all the required calculation is done in order to release the final results of admission. The admission decisions are released throughout many high schools and are posted on the website of the Ministry of Higher Education in Sudan.

As we can see, to some extent, these procedures are complicated, require costly resources, and much time. The resources involved are costly in the sense that the Ministry is obligated to make numerous copies of directory books to distribute to students, and prints of the application form itself. Furthermore, all these admissions procedures are performed manually and data administrators have to delegate a team of typists for the data entry process.

The importance of applying data warehouse technique comes from the scenario of the admission process, which is a process done in two different physical locations: the Ministry of General Education and the Ministry of Higher Education and Scientific Research. That indicates that the data is distributed in different physical locations and different computer storage programs; such as Excel and Access.

As we have seen before, due to a large number of students records, and the limit capacity of the Access, all the data are stored in a separated file of Access, that means the traditional Access software cannot contain all the data in one storage location.

How can decision makers analyze all these distributed data and then make better decisions faster with centralized feature? Nowadays, business intelligence (BI) systems have a great effect on the analysis world; it is the solution for gathering data from different locations, transforming heterogeneous data into homogenous data to be in consistent forms, storing data in a unified cohort, and then diffuse the information to the decision makers. Business intelligence (BI) system can make all these tasks using the concept of a data warehouse. Data warehousing is a result of applying one of the ETL (Extract-Transform - Load) strategies. By using the ETL tools, we could extract data from different storage structures, such as Excel and Access, and transform them into a cohesive structure.

Then, we can load them into a central storage unit to be ready for use in the next mining steps.

#### **4.2.2 Data Warehouse Structure**

A fact table(s) and other tables called dimensions are components of any data warehouse structure. Each fact table contains two categories: keys of dimensions and measurements. Dimensions can be related to the fact table through a direct and indirect relationship form, while, all measurements must be in the fact table within a direct relationship form.

Consequently, across multiple data warehouse structures there is no data warehouse structure consistency in physical attributes, encoding, measurement of attributes, and so forth. Each data warehouse designer has had free rein to propose his/her own design. His/her only goal is how to create a structure that can effectively handle dimensions & measures for analysis purposes. In this study, one fact table has been created which is called *Fact Admission Table*. This fact table consists of five main dimensions and five measurements. Since the fact table can relate to other dimensions with direct and indirect relationship forms, the term main dimensions indicates dimensions that are within direct relationships with the fact table. All dimensions and measurements are detailed below.

The first dimension is the Date dimension. Since the data warehouse structure is a base for the OLAP (Cube) structure, this dimension is necessary for any data warehouse structure that tends to form an OLAP (Cube) structure in the future. This necessity comes when we need to measure any factor over time. As we want to measure some admission factors over time, our date dimension consists of years within the period 2005 -2009.



The second dimension is the State. This dimension contains information about where student permanently lives. Accordingly, it represents the location of the high school that the student attended. Since Sudan is a large country, each state was associated with its geographical direction in a sub-dimension called state-direction or province. Due to the fact that a large number of Sudanese students reside abroad, another sub-dimension which is called Country has been added to this dimension. Consequently this dimension was hierarchically structured in a form of Country – Province- State.

The third dimension is the Department dimension. This dimension contains information about colleges where students were admitted. This was structured in a hierarchy form of University – faculty- Department- Program Type

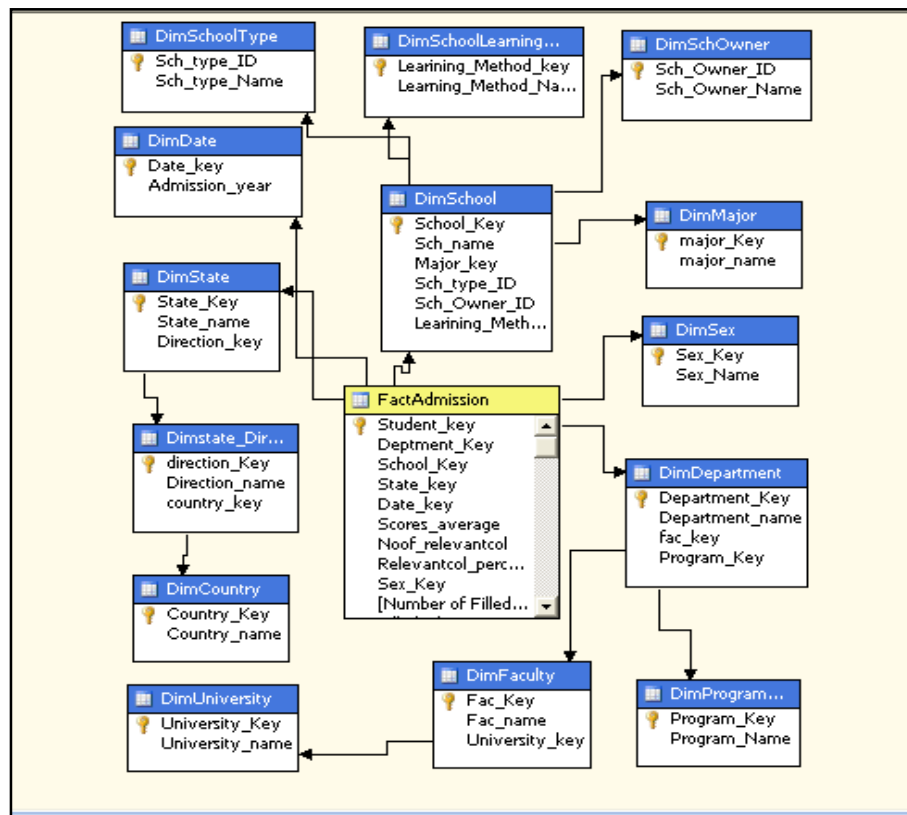
The fourth dimension is Student's High School. This dimension describes the high school where students completed their education and obtained their Sudanese high secondary certificate. Schools were filtered based on: Major (such as academic or agricultural); Designation to categorize schools in Female, Male, or Coed; Ownership to describe whether the school is public or Private; Category to explain the admission type of students in school whether the student is Regular, Home student, or teacher schools.

The last dimension is Student's sex to categorize students i.e. in Female/Male.

A number of measures have been suggested for this data warehouse structure, these measures include: Scores Average. In view of the fact that each student can apply to 45 college choices, each student can has 45 scores average by subjects. Given that, this measure computes all students' scores average by subjects; Number of filled choices, because students vary in number of filled choices, one student fills 3 choices, another fills 20 choices and so on. So, this measure counts the number of filled choices for each student. Using this measure,

the Number of filled choices percent measure has been created to calculate the percentage of each number of filled choices out of the total number of the allowed choices. From this measure other measures have been generated such as Maximum and Minimum of filled choices percent; Fact Admissions Count (FAC which counts rows containing the targeted item in the FAC).

We designed dimension tables, their hierarchies, and the fact measures and we then became ready to create a data warehouse structure. Often, a data warehouse structure is chosen based on the number of created fact tables and their relationships with other dimension tables in the relational database structure. Accordingly, since we have created only one fact table having sub- dimensional tables, the Snowflake structure is convenient to our case. The design of the data warehouse is shown in figure 4.2.



**Fig 4.2:** Multidimensional snowflakes warehouse

Data warehousing has the greatest role in creating OLAP. Since data warehousing represents a perfect method for storing the cleaned data in a central way, it could be a convenient base for OLAP designing. Therefore, the proposed snowflake schema with its multidimensional structure of a data warehouse has been used as a base for OLAP designing as shown next.

### **4.3 OLAP Technique**

OLAP (On-line Analytical Processing) is a Data mining technique to organize the data using a multidimensional data cube. Data in a cube are distributed along attribute dimensions associated with their measurement values. These values are the cell values in the cube [3, 10]. For example, consider data that represents scores average of student admissions across different states in different years. The measurement attribute is the scores average of each state in a particular year. One of the advantages of using a multidimensional data cube (OLAP) is that each dimension of the cube may have a concept hierarchy, which in turn allows the user to see scores average at different levels of state's or year's hierarchy.

In this research, in order to develop an OLAP cube based on data warehouse, we have benefited from the methodology in source [Rob, Ellis 2007] which explained how data can be converted from OLTP to OLAP.

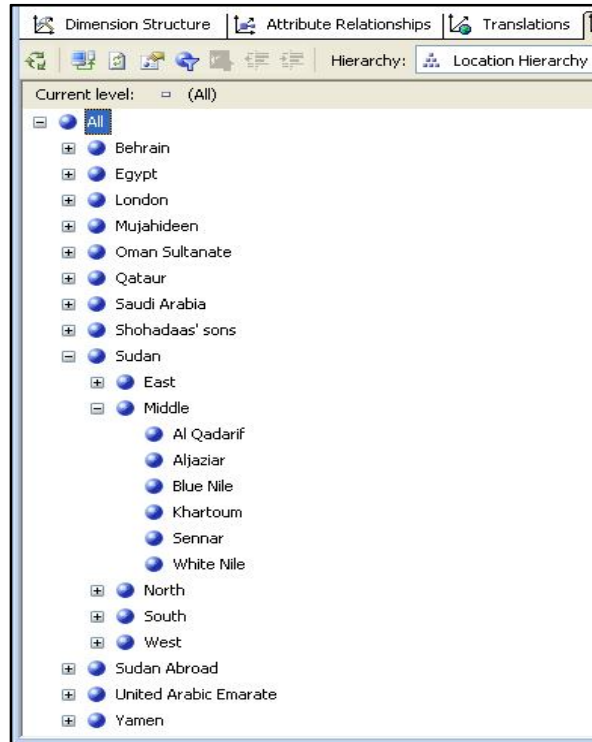
In section 4.3.1, we discuss the proposed OLAP (Cube) structure and its implementations. Programming OLAP to create a smart user interface is the topic of section 4.3.2

### 4.3.1 OLAP (Cube) structure

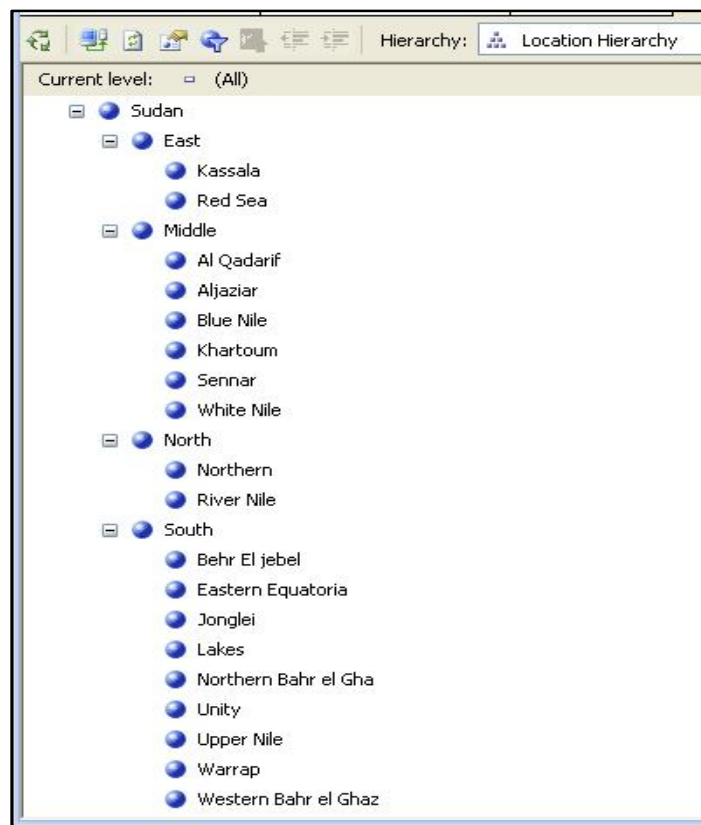
Creating an OLAP structure (cube) using the wizard provided by SQL server analysis services often passes through several steps. The first step is connecting to a warehouse data source. In the second step, a data source view has been created based on the connected data source. The data source view describes how the user looks at his \her data source and checks relationships. After that, a cube has been created using the created data source view. In the creating cube step, we prepared its dimensions with respect to their hierarchies and performed any necessary modifications to the measurements. In the last step, the cube has been deployed and processed to be ready for answering queries.

To build our OLAP structure, we used the data warehouse structure, in section 4.2.2, as a superior data source. Using such data source in analysis services could enable us to process measures along all dimensions. Through this section, we overview implementations of Location and Higher Education dimensions as well as the developed cube implementation.

Figure 4.3 shows implementation of location dimension with its hierarchy i.e. (Country, Province, and State); whereas, figure 4.4 shows implementation of Location dimension when we drill down to Sudan country.



**Fig 4.3:** implementation of Location Hierarchy



**Fig 4.4:** implementation of Location dimension with drilling down to Sudan country

Implementation of Higher Education dimension with its hierarchy is shown in figure 4.5 where the University of Al-Quadarif is drilled down to its faculties.



**Fig 4.5:** Implementation of Higher education dimension

To overview the cube implementation, the following figures show how the cube is implemented and answer different queries based on using different measures across different dimensions in respect to their hierarchies along with date dimension. In this section, we focused on querying the cube dimensions using Scores Average Percent and Filled Choices Percent measures across the period 2005- 2009. Figure 4.6.a shows scores average percent for some universities that belong to the Higher Education dimension and for faculties of the University of Khartoum when we drilled down; whereas, Figure 4.6.b shows filled choices percent for the same dimension with drilling down of University of Shandi.

		Admission Year ▼					
		2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
University ▼	Faculty	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent
الخرطوم	كلية الآداب	72.23%	72.13%	72.17%	72.69%	77.31%	73.00%
	التربية	74.83%	76.70%	77.47%	77.46%	77.72%	76.83%
	كلية العلوم	83.52%	84.02%	84.27%	83.33%	82.43%	83.51%
	العلوم الرياضية	77.33%	78.03%	77.50%	77.08%	77.44%	77.48%
	مدرسة العلوم الإدارية	69.33%	67.04%	70.74%	71.11%	69.44%	69.24%
	القانون	83.25%	79.76%	80.52%	81.35%	81.28%	81.15%
	الهندسة	87.85%	89.05%	88.31%	88.83%	88.25%	88.45%
	الدراسات الاقتصادية والاجتماعية	78.91%	81.39%	80.00%	78.47%	78.94%	79.48%
	الطب و الجراحة	93.45%	93.61%	93.80%	93.12%	93.06%	93.41%
	الطب البيطري	84.04%	84.10%	84.44%	83.78%	83.22%	83.92%
	علوم المختبرات الطبية	89.93%	90.02%	89.99%	89.28%	89.04%	89.66%
	الزراعة	72.84%	73.54%	72.86%	73.00%	73.25%	73.09%
	علوم التمريض	82.95%	83.80%	84.54%	84.58%	84.30%	84.03%
	طب الأسنان	85.23%	86.76%	86.24%	86.98%	86.75%	86.37%
	اللغات	73.98%	74.29%	73.17%	73.80%	72.93%	73.62%
	الصيدلة	91.61%	91.25%	91.40%	91.59%	91.69%	91.51%
الصحة العامة ومعدة البيئة	79.87%	80.26%	81.18%	80.89%	80.53%	80.55%	
الإنتاج الحيواني	78.05%	79.22%	78.31%	78.28%	77.15%	78.20%	
Total		76.35%	75.48%	77.43%	77.43%	78.50%	76.96%
أندمان الإسلامية		70.98%	72.10%	72.77%	73.13%	73.56%	72.47%
السودان		72.00%	72.59%	72.86%	72.54%	73.95%	72.75%
جها		68.08%	70.58%	67.57%	67.52%	66.85%	68.01%
الجزيرة		71.43%	72.57%	73.76%	74.11%	74.33%	73.12%
القرآن الكريم		68.23%	69.29%	70.40%	69.15%	69.47%	69.29%
النيلين		71.07%	72.52%	72.24%	72.15%	72.19%	72.03%
الزعيم الأزهرى		75.75%	77.62%	78.55%	78.98%	79.12%	77.94%
شندي		72.18%	73.31%	73.65%	73.82%	73.94%	73.39%
وادي النيل		68.90%	70.00%	70.24%	70.35%	70.22%	69.98%
		67.70%	66.67%	69.66%	69.70%	69.69%	69.64%

**Fig 4.6.a:** Implementation of the cube by querying the Higher education dimension about scores average percent

		Admission Year ▼					
		2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
University ▼	Faculty	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent
✚	الخرطوم	36.75%	33.33%	31.71%	30.30%	28.31%	32.23%
✚	أمدرمان الإسلامية	41.06%	36.84%	35.56%	33.59%	32.02%	35.95%
✚	السودان	37.78%	34.33%	32.53%	30.29%	29.20%	32.90%
✚	جوبا	32.13%	30.79%	30.64%	27.86%	25.17%	29.28%
✚	الجزيرة	40.48%	35.62%	34.13%	33.17%	31.36%	35.28%
✚	القرآن الكريم	41.79%	38.32%	37.58%	37.85%	34.02%	37.55%
✚	النيلين	37.96%	34.97%	34.39%	32.18%	29.73%	33.87%
✚	الزعيم الأزهري	40.67%	38.29%	35.61%	34.35%	32.85%	36.50%
✚	شندي						
	✚ الآداب	43.90%	40.93%	38.71%	32.93%	31.88%	37.63%
	✚ التربية	34.79%	32.61%	33.02%	29.11%	26.17%	31.14%
	✚ الإقتصاد والتجارة وإدارة الاعمال	37.84%	33.79%	30.45%	26.21%	25.83%	30.41%
	✚ العلوم والتقانة	41.16%	36.69%	35.01%	34.74%	32.15%	35.90%
	✚ القانون	39.31%	34.49%	35.75%	32.17%	32.14%	34.75%
	✚ تنمية المجتمع	48.94%	42.49%	45.94%	44.91%	33.29%	43.20%
	✚ الطب والعلوم الصحية	39.98%	37.26%	35.98%	31.35%	31.46%	35.29%
	Total	39.98%	36.61%	35.30%	31.25%	29.66%	34.48%
✚	وادي النيل	38.41%	34.05%	34.03%	31.46%	28.65%	33.04%
✚	دنقلا	42.35%	35.98%	32.95%	30.41%	28.47%	34.05%
✚	البحر الأحمر	31.78%	28.03%	26.70%	25.08%	22.95%	26.67%
✚	بحر العزال	29.26%	28.56%	31.03%	29.38%	26.81%	28.98%
✚	القضارف	30.43%	27.14%	30.12%	26.87%	28.06%	28.54%
✚	سنار	41.51%	38.19%	37.07%	35.45%	32.51%	36.59%
✚	النيل الأزرق	41.60%	36.40%	37.81%	35.46%	31.41%	36.25%
✚	الإمام المهدي	40.53%	36.47%	35.11%	31.97%	30.61%	35.09%
✚	بغث الرضا	34.36%	31.68%	30.43%	28.88%	26.15%	30.27%
✚	كسلا	33.71%	30.33%	27.58%	25.40%	25.65%	28.52%
✚	كردفان	34.35%	32.01%	28.07%	26.62%	25.89%	28.92%
✚	غرب كردفان	37.37%	31.87%	28.90%	27.45%	26.09%	30.10%

**Fig 4.6.b:** Implementation of the cube by querying the Higher education dimension about filled choices percent

For the Location dimension, we only queried the cube about Sudan country since it contains the majority of students. Its scores average percent is shown in figure 4.7.a; while its filled choices percent is shown in figure 4.7.b.



		Admission Year ▼					
		2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
Country	Direction State	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent
☑ Mujahideen		67.98%	65.97%	58.60%	67.60%	70.63%	67.73%
☑ Sudan	☑ East						
	Kassala	70.76%	71.04%	70.67%	70.89%	71.46%	70.96%
	Red Sea	67.41%	67.57%	68.01%	68.31%	67.90%	67.85%
	Total	69.24%	69.39%	69.41%	69.63%	69.71%	69.48%
	☑ Middle						
	Al Qadarif	69.85%	71.45%	71.36%	71.75%	71.86%	71.28%
	Aljaziar	71.35%	72.37%	72.56%	72.68%	72.62%	72.32%
	Blue Nile	69.05%	68.93%	69.54%	68.29%	68.68%	68.90%
	Khartoum	70.20%	71.50%	71.65%	71.65%	71.89%	71.38%
	Sennar	76.47%	77.37%	77.53%	74.74%	74.30%	76.14%
	White Nile	69.08%	69.72%	70.20%	70.97%	71.18%	70.28%
	Total	70.87%	72.02%	72.20%	72.04%	72.14%	71.86%
	☑ North						
	Northern	68.74%	69.60%	69.82%	70.30%	70.26%	69.77%
	River Nile	70.13%	71.34%	71.17%	71.06%	71.49%	71.05%
	Total	69.59%	70.71%	70.63%	70.76%	71.03%	70.56%
	☑ South						
	Behr El Jebel	69.61%	68.31%	67.27%	66.72%	66.39%	67.35%
	Eastern Equatoria	68.22%	68.11%	69.63%	70.01%	69.07%	68.99%
	Jonglei	73.15%	71.96%	70.28%	66.80%	66.81%	70.19%
	Lakes	69.87%	67.72%	69.10%	68.58%	69.66%	69.02%
	Northern Bahr el Gha	66.55%	66.41%	66.79%	67.21%	69.07%	67.36%
	Unity	67.11%	68.36%	69.41%	68.32%	71.01%	69.03%
	Upper Nile	71.03%	71.28%	68.96%	66.66%	65.47%	68.68%
	Warrap	67.84%	66.68%	67.79%	68.68%	68.18%	67.93%
	Western Bahr el Ghaz	68.34%	67.64%	67.55%	67.20%	67.10%	67.58%
	Western Equatoria	68.42%	66.49%	66.22%	66.48%	65.73%	66.68%
	Total	69.83%	69.28%	68.41%	67.27%	67.05%	68.29%
	☑ West						
	North Darfur	65.08%	65.65%	64.77%	64.68%	65.19%	65.06%
	North Kurdufan	68.59%	68.65%	69.16%	69.63%	70.11%	69.32%
	South Darfur	68.44%	67.28%	66.92%	67.12%	66.72%	67.28%
	South Kurdufan	65.09%	64.56%	64.98%	65.58%	66.30%	65.38%
	West Darfur	64.83%	65.70%	65.16%	65.78%	65.64%	65.50%
	West Kurdufan	65.29%	68.71%	76.79%	69.72%	67.40%	65.77%
	Total	66.95%	66.83%	66.64%	67.00%	67.29%	66.95%
	Total	70.09%	70.86%	70.87%	70.79%	70.87%	70.70%
☑ Sudan Abroad		75.11%	76.77%	75.16%	75.21%	76.68%	75.82%
Grand Total		70.09%	70.87%	70.88%	70.80%	70.88%	70.71%

**Fig 4.7.a:** Implementation of the cube by querying the Location dimension about scores average percent per province for Sudan country

		Admission Year ▼					
		2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
Country	Direction State	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent
☑ Mujahideen		44.59%	19.63%	17.78%	43.70%	34.81%	41.05%
☑ Sudan	☑ East						
	Kassala	36.79%	33.36%	31.84%	28.22%	27.47%	31.58%
	Red Sea	30.04%	25.53%	25.27%	22.80%	21.27%	24.87%
	Total	33.74%	29.63%	28.73%	25.59%	24.42%	28.38%
	☑ Middle						
	Al Qadarif	36.11%	32.06%	31.48%	30.30%	29.60%	31.80%
	Aljaziar	39.68%	35.26%	35.44%	33.67%	31.43%	35.05%
	Blue Nile	42.16%	38.20%	34.82%	33.18%	31.17%	35.74%
	Khartoum	38.17%	34.78%	33.42%	31.26%	29.35%	33.33%
	Sennar	46.46%	40.91%	41.12%	39.32%	38.16%	41.27%
	White Nile	36.72%	33.54%	31.49%	29.81%	28.33%	31.81%
	Total	39.09%	35.25%	34.38%	32.34%	30.48%	34.25%
	☑ North						
	Northern	42.42%	38.80%	36.83%	34.49%	32.16%	36.79%
	River Nile	38.01%	34.13%	32.58%	30.08%	27.59%	32.32%
	Total	39.72%	35.83%	34.27%	31.79%	29.31%	34.03%
	☑ South						
	Behr El Jebel	24.78%	23.68%	23.74%	21.00%	17.94%	21.49%
	Eastern Equatoria	23.42%	24.65%	23.48%	20.68%	18.49%	21.96%
	Jonglei	28.10%	25.73%	25.93%	23.79%	21.00%	25.29%
	Lakes	26.28%	21.85%	22.80%	23.92%	22.67%	23.47%
	Northern Bahr el Gha	19.76%	25.94%	21.16%	22.82%	20.99%	22.22%
	Unity	23.64%	23.18%	25.49%	24.95%	23.92%	24.33%
	Upper Nile	31.24%	28.55%	28.94%	24.43%	24.48%	27.59%
	Warrap	21.65%	24.49%	22.53%	23.54%	22.65%	22.95%
	Western Bahr el Ghaz	29.63%	29.32%	27.66%	27.10%	25.11%	27.84%
	Western Equatoria	22.02%	24.10%	21.43%	18.46%	14.88%	19.95%
	Total	27.15%	26.21%	25.96%	23.26%	21.32%	24.62%
	☑ West						
	North Darfur	37.42%	33.23%	34.22%	31.87%	28.36%	32.79%
	North Kurdufan	34.75%	31.65%	29.76%	27.55%	25.49%	29.29%
	South Darfur	37.81%	36.15%	36.54%	33.04%	30.97%	34.77%
	South Kurdufan	33.36%	32.17%	28.86%	27.55%	26.79%	29.19%
	West Darfur	39.97%	35.89%	35.61%	33.40%	30.83%	34.29%
	West Kurdufan	35.99%	17.55%	29.35%	37.93%	29.52%	35.52%
	Total	36.52%	33.59%	32.95%	30.49%	28.14%	32.04%
	Total	38.02%	34.38%	33.51%	31.28%	29.25%	33.17%
☑ Sudan Abroad		38.57%	35.31%	34.87%	33.02%	29.44%	33.63%
Grand Total		38.02%	34.38%	33.51%	31.28%	29.25%	33.17%

**Fig 4.7.b:** Implementation of the cube by querying the Location dimension about filled choices percent per province for Sudan country

Hierarchy levels were not created for the High School dimension, However, high schools are filtered for major, School' ownership, school's designation (gender type), and school's learning methods.

High schools that filtered based on their major are measured by scores average percent and filled choices percent as shown in figure 4.8.a, and figure 4.8.b respectively

	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
<b>Major</b> ▼	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent
Academic	70.45%	70.48%	70.37%	70.48%	70.52%	70.46%
Commercial	67.63%	67.60%	67.72%	67.67%	67.69%	67.66%
Industrial	67.88%	67.56%	67.34%	67.20%	66.83%	67.35%
Agricultural	70.36%	67.33%	67.77%	65.54%	65.81%	67.24%
Hafazah	77.31%	74.06%	80.74%	74.26%	74.19%	75.40%
Scientific Institutes	73.19%	73.55%	74.33%	74.29%	75.96%	74.32%
Nisweyah	65.98%	69.42%	63.82%	67.59%	70.35%	67.61%
Arabic Certificate	88.05%	88.39%	87.85%	85.75%	83.52%	87.05%
Aedeen from abroad	55.93%	57.60%	82.45%	82.85%	74.90%	58.38%
Grand Total	70.82%	70.84%	70.76%	70.66%	70.70%	70.76%

**Fig 4.8.a:** the High School dimension is filtered by major and queried by scores average percent

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
<b>Major</b> ▼	Filled Choices Percent1	Filled Choices Percent1	Filled Choices Percent1	Filled Choices Percent1	Filled Choices Percent1	Filled Choices Percent1
Academic	34.38%	33.38%	32.92%	32.67%	31.78%	32.99%
Commercial	32.06%	32.03%	31.81%	33.28%	31.57%	32.17%
Industrial	34.55%	34.76%	33.45%	34.34%	33.90%	34.20%
Agricultural	34.07%	27.71%	28.72%	32.11%	24.74%	29.48%
Hafazah	13.53%	18.75%	11.85%	13.42%	15.01%	14.87%
Scientific Institutes	25.50%	26.00%	22.88%	23.11%	21.07%	23.60%
Nisweyah	34.32%	34.81%	35.72%	33.55%	30.07%	33.74%
Arabic Certificate	50.97%	46.21%	45.06%	45.23%	47.85%	47.19%
Aedeen from abroad	6.40%	8.31%	6.66%	11.11%	22.22%	7.47%
Grand Total	34.64%	33.57%	33.12%	32.77%	31.92%	33.17%

**Fig 4.8.b:** The High School dimension is filtered by major and queried by filled choices percent

Querying the High school dimension which filtered by School's ownership is shown in figures 4.9.a and 4.9.b respectively.

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
School Ownership ▼	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent
Public	69.95%	70.82%	70.84%	70.63%	70.70%	70.58%
Private	70.77%	71.05%	71.05%	71.38%	71.47%	71.18%
Grand Total	70.09%	70.87%	70.88%	70.80%	70.88%	70.71%

**Fig 4.9.a:** the High School dimension is filtered by School's ownership and queried by scores average percent

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
School Ownership ▼	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent
Public	38.10%	34.39%	33.59%	31.39%	29.25%	33.30%
Private	37.65%	34.35%	33.21%	30.92%	29.26%	32.71%
Grand Total	38.02%	34.38%	33.51%	31.28%	29.25%	33.17%

**Fig 4.9.b:** the High School dimension is filtered by School' ownership and queried by filled choices percent

School's learning methods, which categorize the student's admitting to the high school are measured as shown in figures 4.10.a and 4.10.b respectively.

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
Learning Method ▼	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent
Regular	71.43%	72.15%	72.28%	71.90%	72.09%	71.97%
Home	68.62%	69.28%	69.05%	69.43%	69.46%	69.14%
Teacher Schools	68.06%	68.76%	68.77%	69.11%	69.13%	68.80%
Grand Total	70.09%	70.87%	70.88%	70.80%	70.88%	70.71%

**Fig 4.10.a:** the High School dimension is filtered by School's learning methods and queried by scores average percent

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
Learning Method ▼	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent
Regular	37.04%	33.84%	32.91%	30.53%	28.75%	32.49%
Home	39.04%	35.24%	34.12%	32.14%	29.82%	34.47%
Teacher Schools	39.54%	35.15%	34.50%	32.50%	29.97%	33.99%
Grand Total	38.02%	34.38%	33.51%	31.28%	29.25%	33.17%

**Fig 4.10.b:** the High School dimension is filtered by School' learning methods and queried by filled choices percent

Last years the numbers of females versus males are rapidly increasing in Sudan. This fact is reflected through the numbers of students in the high schools or in the higher education. Measuring of the high schools that filtered based on female, male, coed is shown in figures 4.11.a and 4.11.b respectively.

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
School Type ▼	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent
Female	69.77%	70.74%	70.76%	70.93%	71.04%	70.66%
Male	69.48%	69.93%	69.90%	70.19%	70.44%	70.00%
Mixed	71.62%	72.57%	72.66%	71.59%	71.32%	71.93%
Grand Total	70.09%	70.87%	70.88%	70.80%	70.88%	70.71%

**Fig 4.11.a:** Measuring of the high schools by score average percent according to their type.

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
School Type ▼	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent
Female	37.35%	33.37%	32.44%	30.41%	28.31%	32.27%
Male	38.58%	35.31%	34.39%	31.99%	30.07%	33.92%
Mixed	38.35%	34.77%	34.10%	31.78%	29.60%	33.63%
Grand Total	38.02%	34.38%	33.51%	31.28%	29.25%	33.17%

**Fig 4.11.b:** Measuring of the high schools by filled choices percent according to their type.

Furthermore, all students who were admitted to the higher education are measured through sex dimension. Results are shown in figures 4.12.a and 4.12.b respectively.

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
Sex Key ▼	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent	Scores Average Percent
Female	70.28%	71.26%	71.31%	71.20%	71.24%	71.06%
Male	69.89%	70.44%	70.41%	70.38%	70.51%	70.33%
Grand Total	70.09%	70.87%	70.88%	70.80%	70.88%	70.71%

**Fig 4.12.a:** Female Vs Male are measured by scores average percent

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
Sex Key ▼	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent	Filled Choices Percent
Female	37.51%	33.60%	32.62%	30.74%	28.61%	32.53%
Male	38.58%	35.25%	34.49%	31.87%	29.91%	33.87%
Grand Total	38.02%	34.38%	33.51%	31.28%	29.25%	33.17%

**Fig 4.12.b:** Female Vs Male are measured by filled choices percent

Through our developed OLAP cube, we also used another measure, maximum scores average, rather than scores average percent and filled choices percent. With that measure, we could indicate the maximum scores average along different dimensions and across our selected years. Figures 4.13.a and 4.13.b show how we could use that measure along the location dimension to measure the maximum scores average per province and per state respectively.

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
Direction ▼	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average
Abroad	94.46	95	93.71	94.29	94.43	95
East	98.83	96.93	95.88	95.43	97.8	98.83
Middle	100	100	100	100	99.1	100
Mujahideen	79.39	72.4	58.6	77.2	75.43	79.39
North	95.41	95.01	95.43	95.43	95	95.43
South	97.11	100	98	92.14	95.47	100
West	97.47	96.41	97.19	95.29	96.1	97.47
Grand Total	100	100	100	100	99.1	100

**Fig.4.13.a:** The maximum scores average of students per province

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
State ▼	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average
Abroad	94.46	95	93.71	94.29	94.43	95
Al Qadarif	93.57	94.83	93.71	94.29	93.14	94.83
Aljazair	97.03	96.86	95.71	96.43	96.48	97.03
Behr El Jebel	87.62	85.86	86	86	86.29	87.62
Blue Nile	96.23	97.57	98.47	99.67	98.98	99.67
Eastern Equatoria	82.46	86.71	90	84.71	84.29	90
Jonglei	97.11	100	98	86	81.53	100
Kassala	98.83	96.93	95	95.43	97.71	98.83
Khartoum	99.48	100	100	98.88	99	100
Lakes	82.5	84.71	83.43	82.14	82.86	84.71
Mujahideen	79.39	72.4	58.6	77.2	75.43	79.39
North Darfur	89.49	90.43	89.86	89.14	91	91
North Kurdufan	93	95.14	92.71	95.29	95.86	95.86
Northern	94.26	92.14	91.86	92.81	94.57	94.57
Northern Bahr el Gha	87	79.14	79	80.86	95.47	95.47
Red Sea	96.22	92.57	95.88	94.14	97.8	97.8
River Nile	95.41	95.01	95.43	95.43	95	95.43
Sennar	100	100	100	100	99.1	100
South Darfur	93.11	92.71	89.29	92.14	94.29	94.29
South Kurdufan	97.47	96.41	97.19	93.53	96.1	97.47
Unity	79.61	82.29	86.43	85	85.57	86.43
Upper Nile	90	90.29	86.86	83.57	88.43	90.29
Warrap	81.43	77.26	78.86	92.14	84.29	92.14
West Darfur	84.5	88.14	83.86	86.57	89.29	89.29
West Kurdufan	91.69	91.62	89.04	86.12	91.44	91.69
Western Bahr el Ghaz	89.29	83.71	81.71	83.57	91.8	91.8
Western Equatoria	84	85.86	80.29	83.43	82	85.86
White Nile	97.58	97.6	99.14	99.71	97.91	99.71
Grand Total	100	100	100	100	99.1	100

**Fig.4.13.b:** The maximum scores average of students per state

We used the same measure along the high school dimension to indicate the maximum scores average of students per admission type, per major, per ownership, and per high school's type (Female Vs Male). Figures (4.14.a – 4.14.d) show that respectively.

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
Learning Method ▼	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average
Regular	100	100	100	100	99.1	100
Home	92.57	93.86	92.57	93.86	93.29	93.86
Teacher Schools	94.1	94.86	94.86	94.43	95	95
Grand Total	100	100	100	100	99.1	100

**Fig.4.14.a:** The maximum scores average of students per admission type

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
Major ▼	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average
Academic	97.32	96.86	96.57	96.48	96.48	97.32
Commercial	84.16	82.71	85.43	86.86	87.14	87.14
Industrial	79.83	82.6	86.57	85.86	85.14	86.57
Agricultural	75	75.4	74.6	76.6	78	78
Hafazah	90	90	90	90	99	99
Scientific Institutes	87	95.29	94.57	93.86	97.43	97.43
Nisweyah	79.29	73.3	72.9			79.29
Arabic Certificate	100	100	100	100	99.1	100
Aedeen from abroad	93	93	88	91.71	79	93
Grand Total	100	100	100	100	99.1	100

**Fig.4.14.b:** The maximum scores average of students per high school's major

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
School Ownership ▼	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average
Public	100	100	100	100	99.1	100
Private	96.12	96	95.86	96.43	96.48	96.48
Grand Total	100	100	100	100	99.1	100

**Fig.4.14.c:** The maximum scores average of students per high school's ownership

	Admission Year ▼					
	2005-07-01	2006-07-01	2007-07-01	2008-07-01	2009-07-01	Grand Total
School Type ▼	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average	Maximum Scores Average
Female	96.12	96.86	96.57	96.43	96.43	96.86
Male	97.32	96.14	95.91	96.48	99	99
Mixed	100	100	100	100	99.1	100
Grand Total	100	100	100	100	99.1	100

**Fig.4.14.d:** The maximum scores average of students per high school's type

However, with our developed OLAP cube, we could use different measures along different dimensions. Then, we could produce significant reports; but unfortunately, such reports have not been designed for end users and cannot be deployed as stand-alone – application. For that, we developed a system that could contribute smartly towards enabling end users to browse data cube with friendly user interface. We have described our developed OLAP system and how end users can query it next.

### **4.3.2 OLAP programming**

The proposed system is an interactive and smart analysis (OLAP) system which enables users to perform statistical operations on any data structures in a multidimensional form. It was developed for the field of education. The intended users of the system are the students who intend to enroll in Sudanese universities. The proposed system is expected to actively help students in choosing the proper university/college among large options of colleges /universities. Moreover, this system providentially empowers enrollment officers of the Ministry of Higher Education with useful knowledge. Such knowledge could help them in the decision making process in line of enhancing education in Sudan.

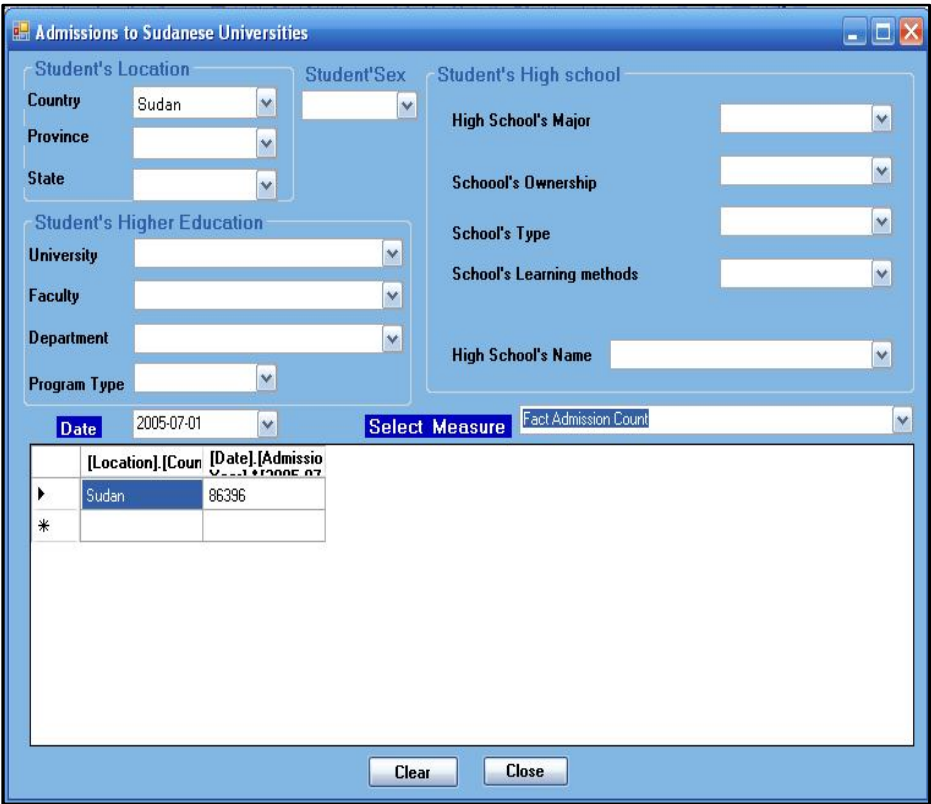
Datasets for data mining applications are often large, real world data, dynamic, prone to updates, and pre-existing compared to datasets for statistical applications which are often limited, static, and user generated.

Our collected databases are numerous, and they consist of real data that was collected from the Ministry of Higher Education in Sudan. These pre-existing data



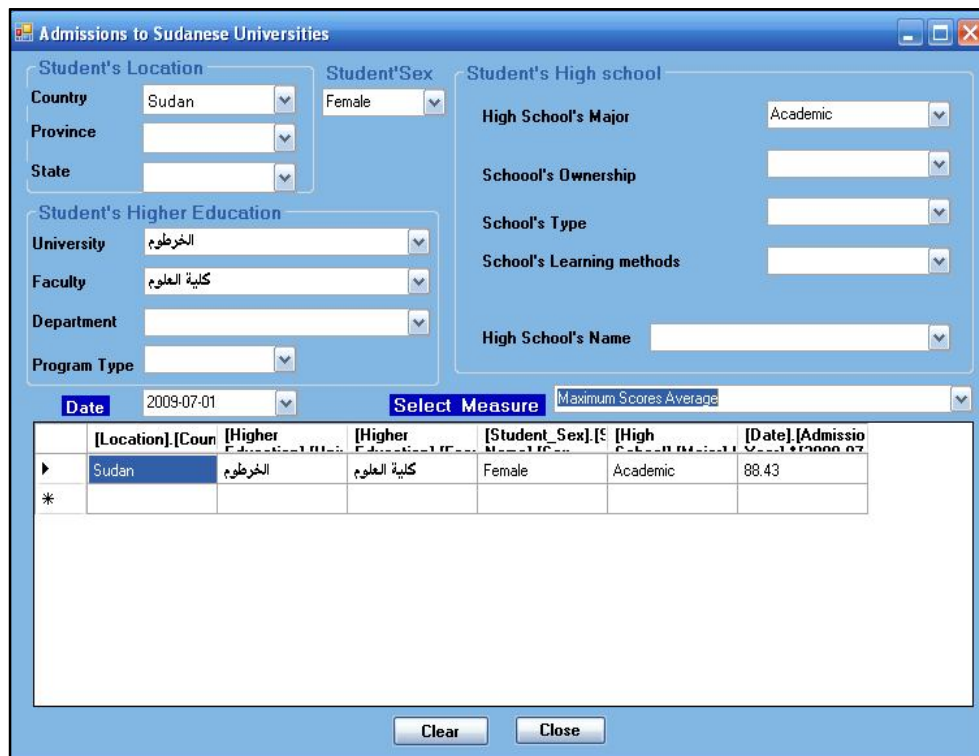
represent records of the students who had applied to Sudanese universities and were admitted into one within the period 2005- 2009.

Users may query OLAP by choosing the dimensions they want to examine/analyze. For example, the users can choose “Sudan” as the “country”, (2005) as the “date”, and (FAC) as the “measure” as shown in figure 4.15. The upper part of the window in the figure shows the pull-down selection menus. The bottom part shows results of the queries. For the selected dimensions in the input pane, the results are shown in the output pane.

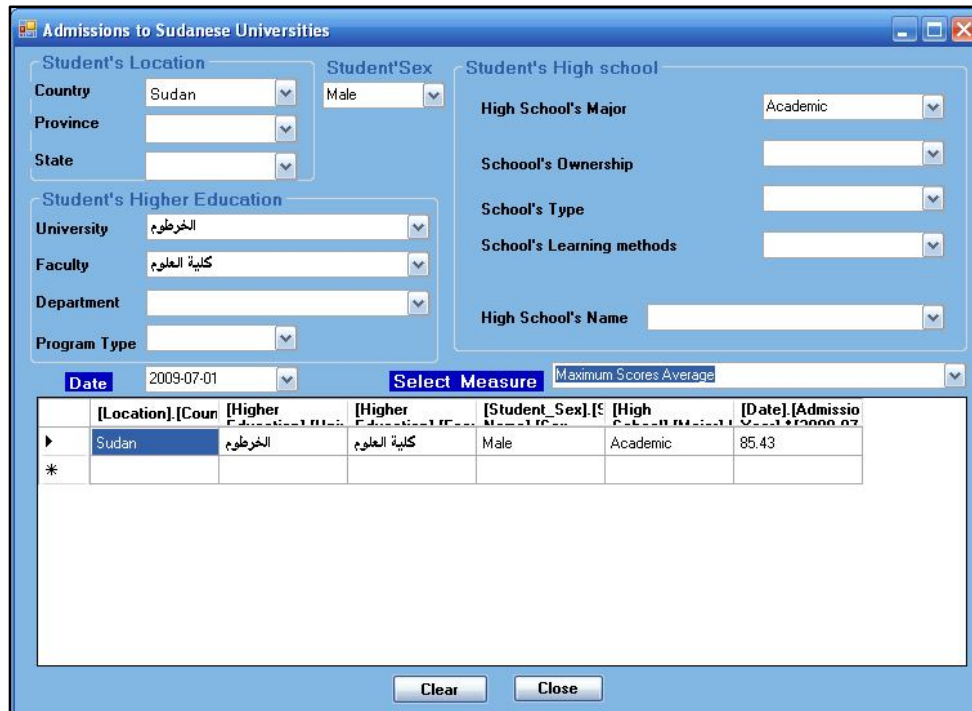


**Fig.4.15:** The system uses FAC to measure admissions for Sudan at 2005

Users can query that system by using another measure. They can drill-down any dimension to retrieve their targeted data. As an example, we used the measure of the (Maximum Scores Average) for female students at 2009. To perform such queries users may select more dimensions; figure 4.16 shows that users can choose the (Academic) major from the High school dimension, (Sudan) as the country, additionally they can choose the University of Khartoum from Student's Higher Education dimension, and then they can drill-down along the same dimension by choosing (Faculty of Science) as the faculty. The response of the system shows a percentage of 88.43. In the same manner, when we executed the same query for male students instead of females, the response of the system shows a percentage of 85.43. Figure 4.17, indicates that, for the chosen date, major, faculty, and location dimensions, female students got higher scores over male students.



**Fig.4.16:** Maximum scores average for female students who chose the academic major in their high school, and they were admitted in the University of Khartoum - faculty of Science at 2009.



**Fig.4.17:** Maximum scores average for male students who chose the academic major in their high school, and they were admitted in the University of Khartoum - faculty of Science at 2009.

A user is expected to follow the hierarchically designed selection system. That is, the user is not allowed to choose a province before a country or a faculty before a university.

One of the advantages of using OLAP is that it analyzes huge records statistically. It also performs analysis at the sub-records level. A sub-record is a version of the record that underwent reduction in the number of dimensions, the measures or both to build an OLAP system with fast response and reduced cost

Several studies have drawn attention to limitations of OLAP technology in the context of the required intelligence for concluding relationships. To confirm the analytical power of OLAP technique, it could be integrated with other Data

mining techniques such as Association Rules Mining and Classification, which produced a new term in Data mining that is called OLAP mining or On-Line Analytical Mining (OLAM) technology. Data mining based on OLAP is discussed in the next section.

#### **4.4 Data mining Process**

Both OLAP and data mining techniques are valuable analytical tools in the knowledge discovery process. Whereas OLAP techniques can answer users' queries that relate to the concept of data aggregation, data mining techniques can answer users' queries that related to the concept of data analysis based on data correlation in order to discover useful knowledge. In the previous section, we could use the OLAP technique to answer some typical questions such as: what is the total number of students who were admitted in the University of Khartoum in the determined years in Sudan country?, what is the filled choices percent of students who applied to Sudanese universities regarding different states and different high schools that students belong to?, what is the maximum and minimum score averages for male and female students, respectively? A Data Mining process is mainly performed based on a restricted question that has been professionally asked. Through this section, we discuss how we could use data mining techniques to answer some typical questions such as: what is the demographic information of students who like to apply to Sudan University or to colleges of Medicine for example?, what percentages of choices should be recommended to these particular students to fill?, what is the estimated score averages for students who completed their education and took the national standard certificate examination from public high schools Vs Private ones?, how does the students' sex affect the admissions

regarding some factors such as: students' location, students' admissions category (Regular Vs Home student)?

A data mining question often consists of a number of concepts and includes: attribute, which is a single piece of data that provides data miners with information about an example,

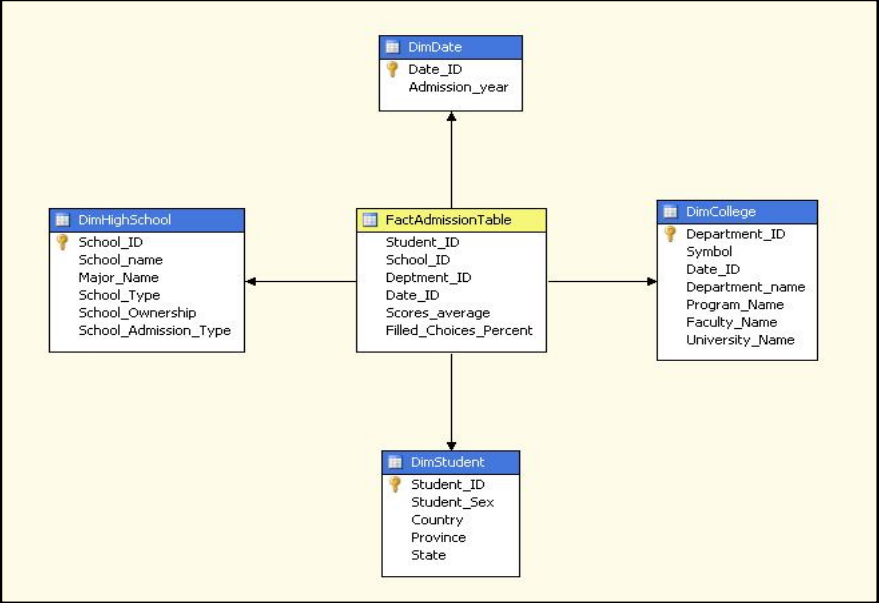
States, which represents all the possible values that can be associated with each categorical attribute, i.e. a school's major attribute may have the states of Academic or Agricultural, and so forth

Case, which represents the entity data miners are mining.

Thorough understanding of the case may well indicate a thorough understanding of the problem to be solved and leads to getting better expected results. Identifying a case depends on determining an anonymous factor (s) about a targeted analysis. Therefore, a case can play the role of attribute and vice versa; for example, data miners may want to know what factors impacted the students' admissions in states. For this situation, the case can very well be the student's state because data miners are interested in attributes about states that impact admissions. Conversely, in a situation where data miners want to examine how the state itself contributes to the students' admissions overall country, the state would be an attribute because it becomes one of the independent factors to be used in the analysis. A case in this situation would be something like a measurement that includes the student's state name, student's scores average, and other attributes.

In the previous sections, we suggested a data warehouse structure that consisted of only one fact table; the fact table was associated with dimensions that included students' location, students' high schools, students' higher education...etc. It also included a few numeric measures such as: students' scores average, percentages of faculty choices that filled by students, maximum and minimum of both scores average and filled choices percentage respectively. Some of these dimensions were

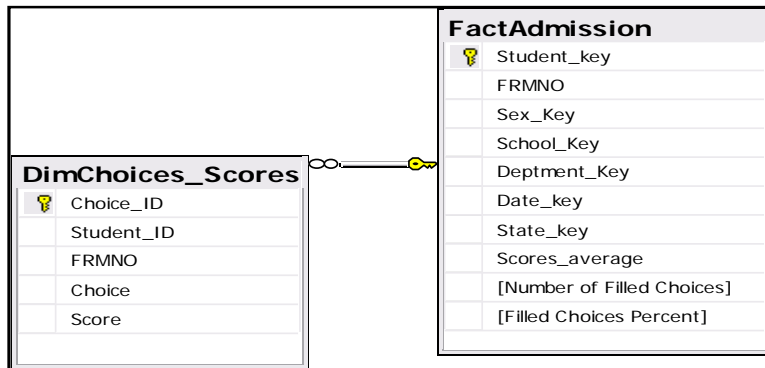
normalized to other related dimensions where they formed snowflakes data warehouse structure. We built our OLAP based on snowflakes data warehouse structure. In this section, we proposed to build a new OLAP based on a Star data warehouse structure, where we could use it in the mining process. The proposed star data warehouse structure consisted of only one fact table in addition to four dimensions; College, Date, Student, and High school. Figure 4.18 shows the proposed cube that built based on a star data warehouse structure, and also shows the proposed attributes for each dimension and the fact table.



**Fig.4.18:** The proposed cube that built based on a star data warehouse structure

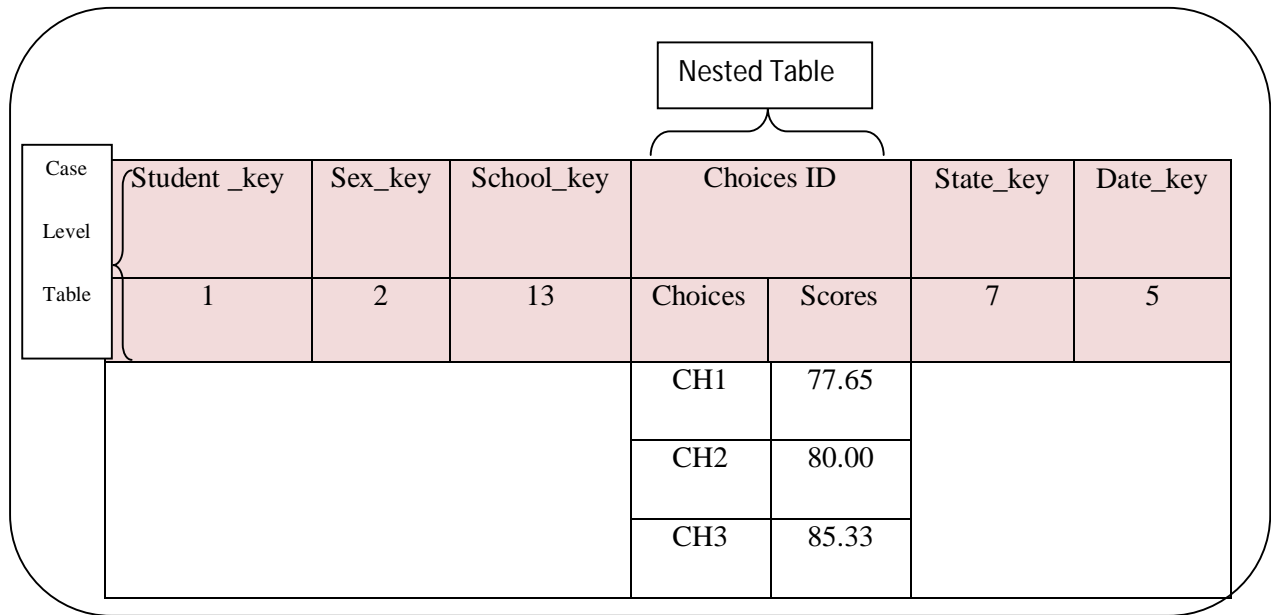
By developing a cube, the data would be ready for starting a mining process to answer some considered questions. In regards to our target questions that can to be answered by building data mining process structure, we had put in mind the required structure of a database for a specific data mining technique that we used for the mining process. Since we have planned to use the association rule mining technique, the association rule mining technique considers the basket market

concept, in which each customer can buy many items. Similarly, in our project, each student can apply up to 45 faculty choices, and each choice was associated with its student's scores average. This generated two concepts; on one hand it generated the concept of one-to-many relationship, where each student can apply up to 45 choices of faculties and has her/his own scores average for each applied choice. This relationship is illustrated in figure 4.19.



**Fig.4.19:** one-to-many relationship- each student can apply to many faculty choices and has many scores average

On the other hand, it generated the concept of Nested table. The nested table is a column in the database with a data type of table. To demonstrate this concept, the two tables in figure 4.19 were combined together to associate each student with her/his choices and scores as shown in figure 4.20. Figure 4.20 shows that the student of key of 1 has applied to three faculty choices CH1, CH2, CH3 and each choice has its associated scores averages 77.65, 80.00, 85.33 respectively. The Choices\_ID column is a nested table that related to the case level table (Fact admission table) through the Student\_key attribute. The concept of nested table is an essential item in applying association rule mining technique where it depends on existence of 1- n relationship type.



**Fig 4.20:** Associating each student with his\her faculty choices and scores

Unfortunately such structure is not a popular structure for mining models built on OLAP, where the case level table is a fact table and the nested table is one of the dimensions. Many questions can be answered by mining models that are created from the mining structure; however, it is difficult to mine data directly from the fact table as in figure 4.20. For example, analyzing the students' scores average at the city level can be extremely intricate for many data mining algorithms because there are too many cities, especially for a large country as Sudan. Though, when data is aggregated to the state/province level, these algorithms may easily discover any hidden patterns.

Probably, the best way to build a mining model based on OLAP is to include the case level table as one of the dimensions, and to have the nested table always come from one of the fact tables using another dimension attribute as the nested key. How we can build such structure is the topic of section 4.4.2.



One of our proposed research questions was, “how do the students, based on their demographics (Sex, Country, State, etc.), affect the admissions process, as well as the scores average percent of students, where it was a measure that contained the aggregated value of each choice for each student.

A Data Mining process mainly can be done through several stages. These stages start with suggesting a mining structure based on a specified problem, then generating a mining model from the suggested mining structure. With the mining model, data miners must define the problem and formulate an object. Also, they can describe what the example data looks like and how they should use the data mining algorithm to interpret the data. In the next stage data miners train the mining model by providing the data examples to the chosen algorithm. The algorithm uses the defined problem that is described by data miners in the mining model in order to examine the data and extract the patterns. These extracted patterns by the algorithm are analyzed to perform predictions or deductions of information. For purposes of testing and validating predictions, data miners provide the chosen algorithm with new data. The new data are formulated in the same way as the training examples.

The mining structure and the mining model are two major objects that are used to manifest these stages, as described in the next two sections.

#### **4.4.1 The Mining Structure**

A mining structure is a list of data columns that will be available when creating models. Those columns are associated with information about them such as their data types and appropriate ways of handling them; some algorithms accept only continuous data, whereas other algorithms accept only categorical data.

Figure 4.21 represents one of our proposed structures; it shows proposed columns that associated with their data types and appropriate ways of handling them.

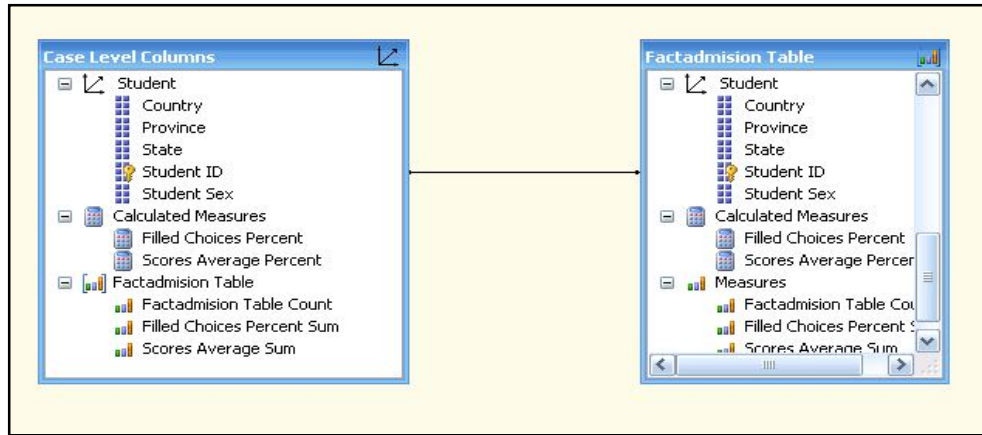
In general a mining structure describes the shape of the problem, additionally; it contains all of the models that are used to analyze the source data of the structure. Since a mining structure defines the domain of a mining problem, it provides models with a superset of columns to create attributes; likewise, it also provides models with a superset of rows to create cases.

To answer questions such as how the students, based on their demographics (Sex, Country, Province, State, etc.), affect the admissions process in certain colleges in regards to student’s province. We suggested developing an OLAP mining structure as shown in figure 4.22. We could partially use this structure to create several mining models.

Mining model structure:		
Columns	Content Type	Data Type
Student ID	Key	Long
Province	Discrete	Text
College		
University Name	Key	Text
Scores Average Percent	Discretized	Double

**Fig.4.21:** A mining Model Structure

Our OLAP mining structure, as shown in figure 4.22, consists of two "tables" for this model. The case "table" on the left represents the case dimension "Student", with all Student properties and a set of measures. The nested "table" on the right represents the fact admission table.



**Fig.4.22:** An OLAP mining structure

#### 4.4.2 The Mining Model

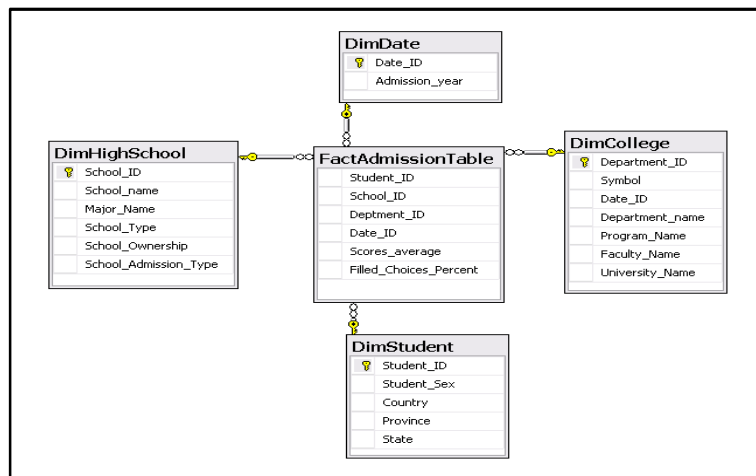
A mining model is the object that transforms rows of data into cases and columns into attributes and then performs the machine learning using a specified Data mining algorithm.

When a mining model is trained, it contains the patterns that the algorithm derived from the data. A model can then be used against new data to predict any output columns that were specified during its creation.

Through this model we planned to analyze student's admissions in universities in regards to her/his province. Figure 4.23 provides a relational view of our developed OLAP mining model. The model analyzes student's admissions based on student's demographics; the list of student's filled faculty choices, scores average, and the associated admission date. In this model we specified DimStudent dimension to be as a case dimension and Student\_ID to be a dimension key. We chose province attribute from DimStudent table as a case level attribute. It contained province name values such as (North, Middle, East, West, and South). It

had a 1-n relationship with dimension key (Student\_ID) where each province can contain many students.

In our proposed case structure, some mining model attributes come directly from relational tables (dimension tables), such as province name, and admission date. Some attributes directly come from nested table (the fact admission table), such as student’s scores average measure, while others come indirectly from nested table that was connected to the fact table, such as university name, where it joined to the fact admission table by the Department\_ID. Our proposed case structure is shown in figure 4.24.

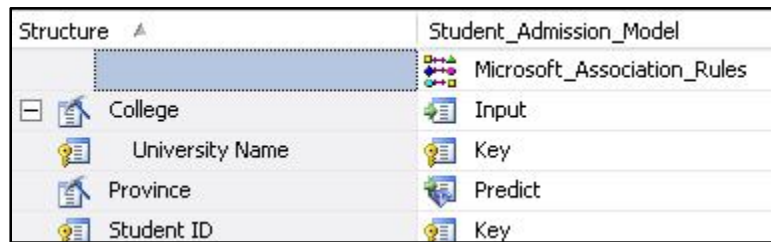


**Fig.4.23:** A relational view of an OLAP mining model

Student_ID	Student_sex	Admission year	Country	Province	State	Admission	
1	Female	2005	Sudan	Middle	Khartoum	University	Scores
						CH1	77.65
						CH2	80.00
						CH3	85.33

**Fig.4.24:** The proposed case structure

In this model, referring to the developed cube on figure 4.18, the case table was the DimStudent dimension, and the nested table was DimCollege dimension which was associated with information of scores average percent from the Fact admission table. We named it Student\_Admission\_Model. Figure 4.25 shows our proposed Mining Model which contains our proposed mining algorithm, plus a list of columns from the mining structure.



**Fig.4.25:** A Mining Model of Student\_Admission\_Model

Since a Mining Model is the application of a mining algorithm to the data in a mining structure, referring to developed mining model as shown in figure 4.25, we used the association rules mining algorithm to build a large association rule based on the previous mining structure. The Microsoft Association Rules algorithm is mainly performed in two steps; the first step is to find frequent itemsets. The second step is to generate association rules based on frequent itemsets. The Student\_Admission\_Model was designed to discover Province based on student's admission. The discovered rules were filtered based on the Province name and cover all the considered admission years. The found itemsets and the representation for visualizing the discovered rules for each province are shown in the next figures (4.26.a, 4.30.b).

Support	Size	Itemset
290868	1	Province = Middle
45710	2	اسم_____تقاله = Existing, Province = Middle
41045	2	أمدرمات الإسلامية = Existing, Province = Middle
39544	2	السودان = Existing, Province = Middle
27673	2	الخرطوم = Existing, Province = Middle
22199	2	النيلين = Existing, Province = Middle
19279	2	الجزيرة = Existing, Province = Middle
14084	2	القرآن الكريم = Existing, Province = Middle
6745	2	بخت الرضا = Existing, Province = Middle
6470	2	الزعيم الأزهري = Existing, Province = Middle
5814	2	الرباط الوطنى = Existing, Province = Middle
5690	2	جوبا = Existing, Province = Middle
5314	2	سنار = Existing, Province = Middle
4813	2	القضارف = Existing, Province = Middle
4666	2	الإمام المهدي = Existing, Province = Middle
3289	2	أمدرمات الأهلية = Existing, Province = Middle
2948	2	النيل الأزرق = Existing, Province = Middle
2919	2	العلوم والثقافة = Existing, Province = Middle
1943	2	شهادة قديمة = Existing, Province = Middle
1874	2	الأحفاد = Existing, Province = Middle
1608	2	استبعاد = Existing, Province = Middle
1455	2	البحر الأحمر = Existing, Province = Middle
1334	2	وادي النيل = Existing, Province = Middle
1327	2	الإمام الهادى = Existing, Province = Middle
1234	2	أكاديمية السودان للعلوم المالية والمصرفية = Existing, Province = Middle

Items: 75

Fig.4.26.a: The found itemsets for the Middle province

Support	Size	Itemset
1203	2	كردفان = Existing, Province = Middle
1193	2	الدلتج = Existing, Province = Middle
1161	2	كسلا = Existing, Province = Middle
1107	2	أفريقيا العالمية = Existing, Province = Middle
990	2	شعدي = Existing, Province = Middle
896	2	علوم الطيران = Existing, Province = Middle
875	2	السودان العالمية = Existing, Province = Middle
848	2	شرق النيل = Existing, Province = Middle
768	2	العلوم الطبية والتكنولوجيا = Existing, Province = Middle
730	2	دقلا = Existing, Province = Middle
679	2	النصر التقنية = Existing, Province = Middle
652	2	أعالى النيل = Existing, Province = Middle
632	2	الخرطوم التقنية = Existing, Province = Middle
571	2	ود مندى الأهلية = Existing, Province = Middle
561	2	كلية الجزيرة التقنية الخرطوم = Existing, Province = Middle
543	2	الحاسبات الآلية = Existing, Province = Middle
501	2	ابناء الشهداء = Existing, Province = Middle
487	2	غرب كردفان = Existing, Province = Middle
426	2	السودان الجامعية للبنات = Existing, Province = Middle
425	2	نبالا = Existing, Province = Middle
397	2	الخرطوم للعلوم الطبية = Existing, Province = Middle
376	2	الخرطوم التطبيقية = Existing, Province = Middle
365	2	المخبريا = Existing, Province = Middle
352	2	بحر الغزال = Existing, Province = Middle
341	2	الفاشر = Existing, Province = Middle

Items: 75

Fig.4.26.b: Other found itemsets for the Middle province

Support	Size	Itemset
334	2	المشرق للعلوم والتكنولوجيا = Existing, Province = Middle
318	2	ودمجني التقنية = Existing, Province = Middle
255	2	جامعة كبرى = Existing, Province = Middle
249	2	كلية الرازي للعلوم الطبية = Existing, Province = Middle
221	2	النيل الأبيض الأهلية = Existing, Province = Middle
208	2	المعهد العالي للعلوم الزكاة = Existing, Province = Middle
207	2	الجريف شرق التقنية = Existing, Province = Middle
206	2	الامارات التقنية = Existing, Province = Middle
196	2	الأردنيةالسودانية = Existing, Province = Middle
172	2	كلية البيان للعلوم والتكنولوجيا = Existing, Province = Middle
162	2	ابوكرعثمان = Existing, Province = Middle
157	2	الكلية الوطنية للدراسات الطبية و التقنية = Existing, Province = Middle
154	2	كلية خيرة العلمية = Existing, Province = Middle
143	2	أكاديمية المنهل للعلوم = Existing, Province = Middle
136	2	الشيخ البدري = Existing, Province = Middle
124	2	كلية الهندسة الكهربائية الاهلية = Existing, Province = Middle
124	2	كلية افريقيا = Existing, Province = Middle
116	2	كتابة التقنية = Existing, Province = Middle
104	2	الكندية السودانية = Existing, Province = Middle
104	2	كلية غرب النيل = Existing, Province = Middle
100	2	أكاديمية السودان للعلوم الاتصال = Existing, Province = Middle
95	2	زالنجي = Existing, Province = Middle
80	2	المصارف = Existing, Province = Middle
78	2	فاردن سنتي = Existing, Province = Middle
71	2	المشرق الاهلية = Existing, Province = Middle

Itemssets: 75

Fig.4.26.c: The rest of the found itemsets for the Middle province

Probability	Importance	Rule
0.951	0.165	النيل الأزرق = Existing -> Province = Middle
0.934	0.158	سنان = Existing -> Province = Middle
0.930	0.149	المصارف = Existing -> Province = Middle
0.929	0.154	أكاديمية السودان للعلوم المالية والمصرفية = Existing -> Province = Middle
0.922	0.148	كلية خيرة العلمية = Existing -> Province = Middle
0.918	0.146	الكلية الوطنية للدراسات الطبية و التقنية = Existing -> Province = Middle
0.916	0.146	المعهد العالي للعلوم الزكاة = Existing -> Province = Middle
0.915	0.156	الخيرة = Existing -> Province = Middle
0.906	0.141	المخربيا = Existing -> Province = Middle
0.903	0.141	ود منجني الأهلية = Existing -> Province = Middle
0.896	0.137	الخرطوم للعلوم الطبية = Existing -> Province = Middle
0.887	0.127	القطبية التقنية = Existing -> Province = Middle
0.887	0.133	العلوم الطبية والتكنولوجيا = Existing -> Province = Middle
0.884	0.123	كلية مكة التقنية لطب العيون = Existing -> Province = Middle
0.878	0.131	القضارف = Existing -> Province = Middle
0.872	0.125	الحاسبات الآلية = Existing -> Province = Middle
0.870	0.126	العلوم والقناة = Existing -> Province = Middle
0.868	0.123	الخرطوم التطبيقية = Existing -> Province = Middle
0.867	0.119	فاردن سنتي = Existing -> Province = Middle
0.866	0.121	ابوكرعثمان = Existing -> Province = Middle
0.861	0.123	بخت الرضا = Existing -> Province = Middle
0.858	0.119	أفريقيا العالمية = Existing -> Province = Middle
0.856	0.117	المشرق للعلوم والتكنولوجيا = Existing -> Province = Middle
0.856	0.120	الرباط الوطني = Existing -> Province = Middle
0.852	0.114	الأردنيةالسودانية = Existing -> Province = Middle
0.850	0.116	الإمام المهدي = Existing -> Province = Middle
0.849	0.112	العلوم والاتصالات = Existing -> Province = Middle

Rules: 34

Fig.4.26.d: The discovered rules for the Middle province







Probability	Importance	Rule
0.979	0.806	نيالا التقنية = Existing -> Province = West
0.966	0.816	النجي = Existing -> Province = West
0.951	0.832	الفاشر = Existing -> Province = West
0.939	0.787	السلام = Existing -> Province = West
0.932	0.821	نيالا = Existing -> Province = West
0.885	0.758	الأبيض التقنية الأهلية = Existing -> Province = West
0.882	0.781	غرب كردفان = Existing -> Province = West
0.856	0.787	كردفان = Existing -> Province = West
0.781	0.731	الدلج = Existing -> Province = West

**Fig.4.28.c:** The discovered rules for the West province

Support	Size	Itemset
22421	1	Province = East
7281	2	البحر الأحمر = Existing, Province = East
4301	2	كسلا = Existing, Province = East
2599	2	Province = East, اســــــــــــــتقالة = Existing
1328	2	Province = East, أمدرمان الإسلامية = Existing
1024	2	Province = East, السودان = Existing
739	2	Province = East, النيلين = Existing
735	2	Province = East, الخرطوم = Existing
500	2	القرآن الكريم = Existing, Province = East
475	2	القضارف = Existing, Province = East
389	2	بورتسودان الأهلية = Existing, Province = East
343	2	الجزيرة = Existing, Province = East
228	2	وادي النيل = Existing, Province = East
204	2	جوبا = Existing, Province = East
203	2	الرباط الوطنى = Existing, Province = East
176	2	كسلا التقنية = Existing, Province = East
169	2	الشرق الاهلية = Existing, Province = East
168	2	الزعيم الأزهرى = Existing, Province = East
113	2	شهادة قديمة = Existing, Province = East
109	2	بورتسودان التقنية = Existing, Province = East
102	2	استبعاد = Existing, Province = East
84	2	دنقلا = Existing, Province = East
76	2	أمدرمان الأهلية = Existing, Province = East

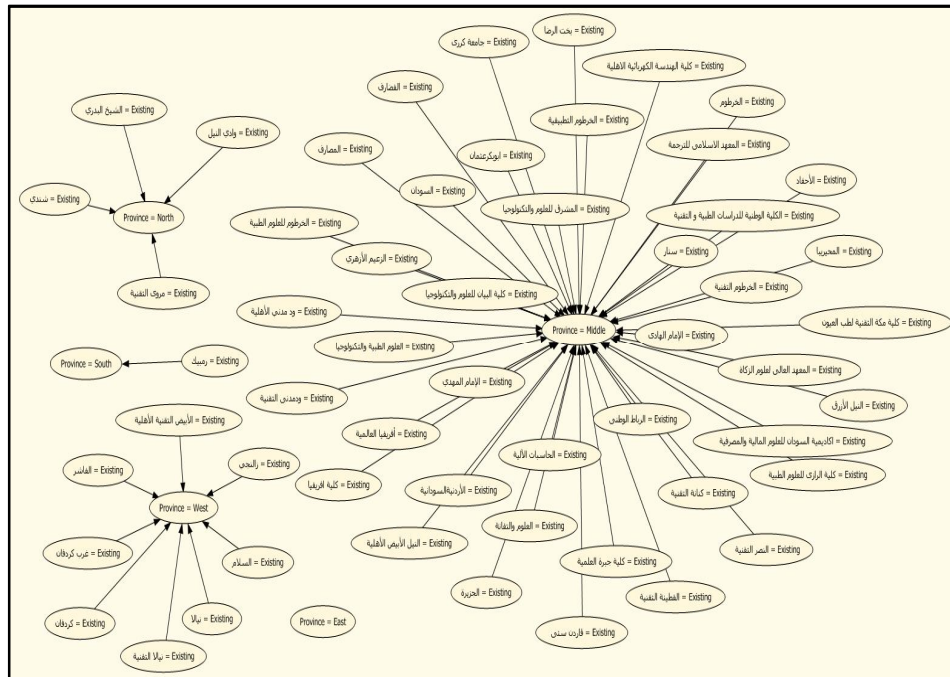
**Fig.4.29.a:** The found itemsets for the East province



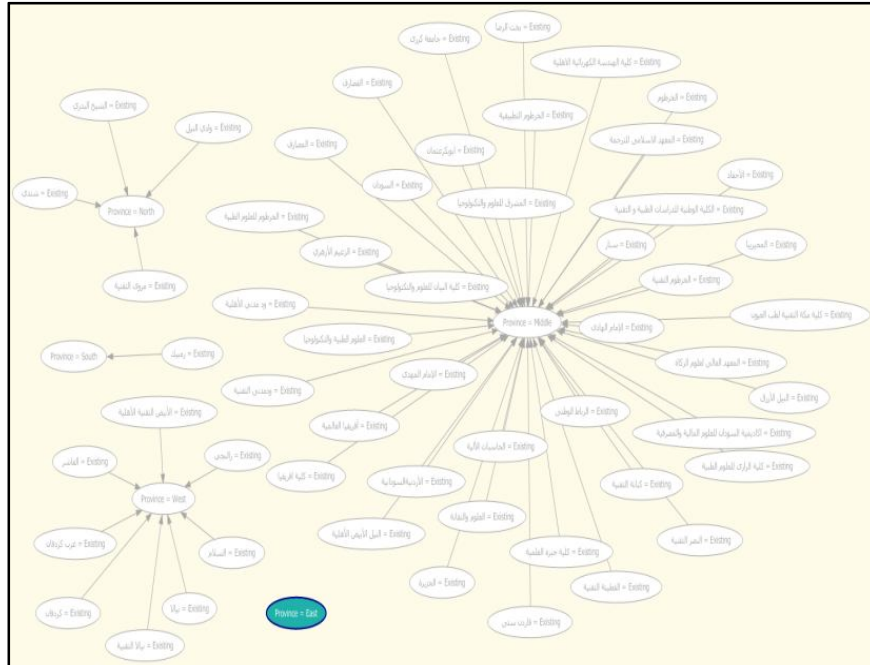
Rules	Itemsets	Dependency Network
Minimum probability:	0.51	Filter Rule: Province = South
Minimum importance:	0.49	Show: Show attribute name and value
<input type="checkbox"/> Show long name		Maximum rows: 2000
Probability	Importance	Rule
0.952	1.341	رمسيسك = Existing -> Province = South
0.667	1.224	بحر الغزال = Existing -> Province = South
0.557	1.143	أعلى النيل = Existing -> Province = South
0.520	1.301	جوها = Existing -> Province = South

**Fig.4.30.b:** The discovered rules for the South province

Another type of Association rules mining algorithm representation for visualizing the results is the Dependency Net view. Figure 4.31.a shows the Dependency Net view for the Student\_Admission\_Model. As you may notice, the relationships are not between model attributes, but instead are between attribute values. Figure 4.31.b shows the East province has no association with any college.

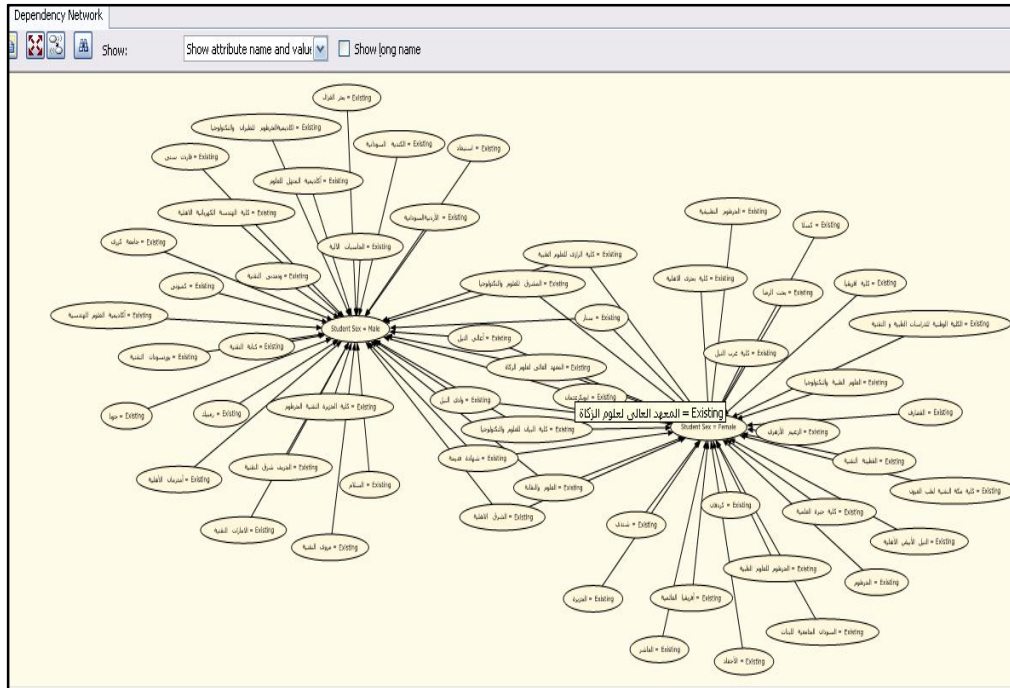


**Fig.4.31.a:** The dependency net view of Student\_Admission\_Model



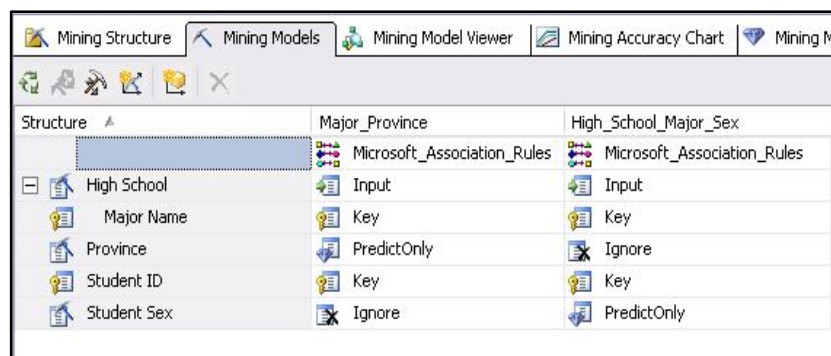
**Fig.4.31.b:** The dependency net view shows that The East province has no association with any college

We can use different algorithms with each model in a structure, or we can use the same algorithm with the same model to build different models with different parameters in the structure. That enables us to answer different questions on the same data set. Through this project, we built another model using the same algorithm with different parameters called Student\_sex. Through that model, we studied the correlations between students' admissions and students' sex in regards to admission years. Figure 4.32 shows its dependency net view.



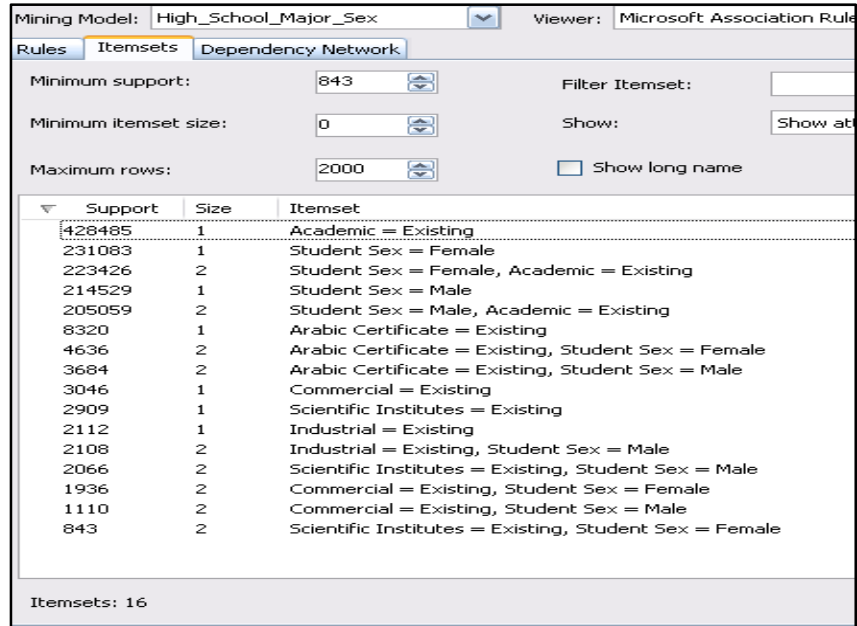
**Fig.4.32:** The dependency net view of admissions based student’s sex

For further mining on data, we investigated the associations between the high school major, student sex, and student province respectively. To do so, we created two models; High\_School\_Major\_Sex and Major\_Province model. Figure 4.33 shows their mining model structure.

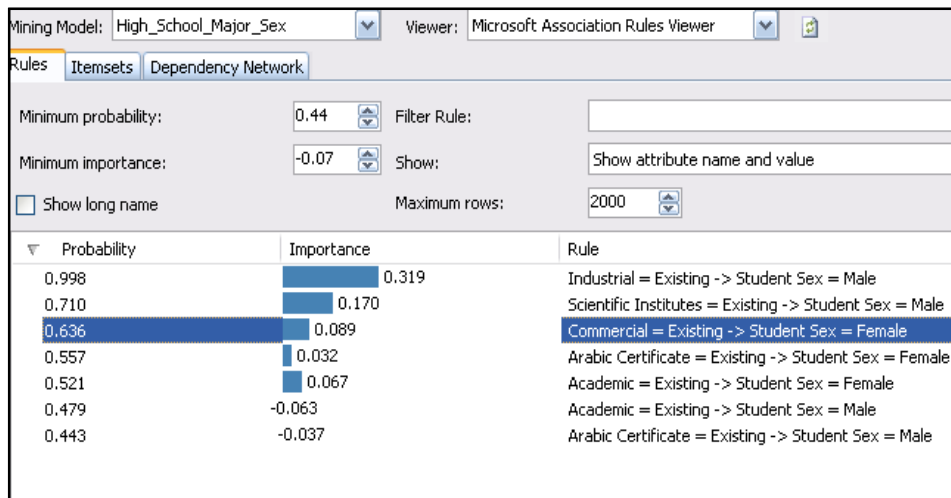


**Fig.4.33:** A Mining Model of both High\_School\_Major\_Sex and Major\_Province Models

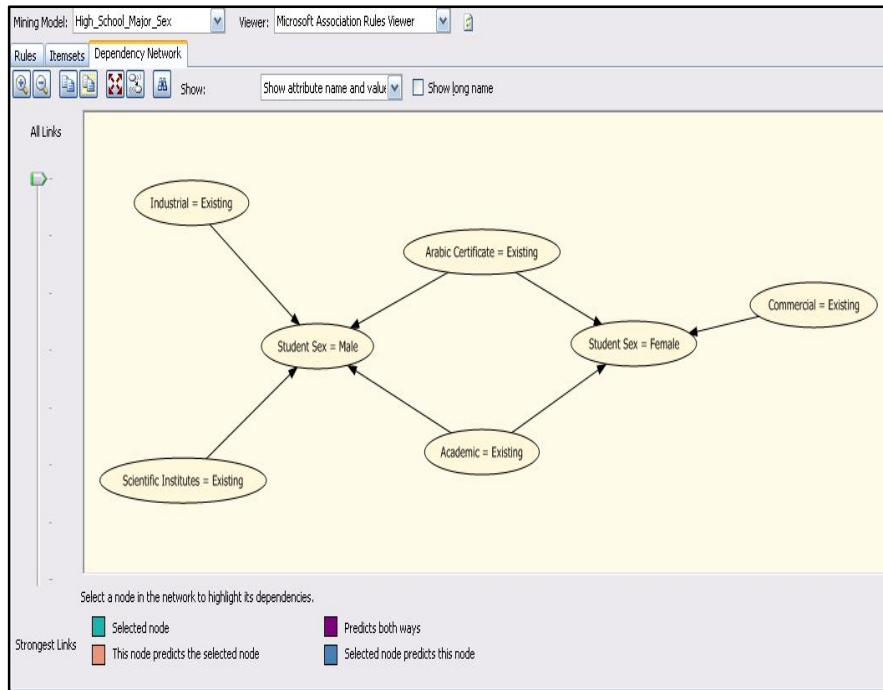
Through High\_School\_Major\_Sex model, we searched for any associations between the high school major and student sex; we found results shown in following figures: figure 4.34, shows our found itemsets, figure 4.35 shows the generated rules, and figure 4.36 shows the dependency net view of admissions based student's sex and student's major



**Fig.4.34:** The found itemsets for the of High\_School\_Major\_Sex Model



**Fig.4.35:** The discovered rules for the of High\_School\_Major\_Sex Model



**Fig.4.36:** The dependency net view of High\_School\_Major\_Sex Model

Through Major\_Province model we looked for any associations between the high school major and student province; we found results shown in following figures: figure 4.37 shows our found itemsets, figure 4.38 shows the generated rules, and figure 4.39 shows the dependency net view of admissions based student's major and student's province



Mining Model: Major\_Province Viewer: Microsoft Association Rules Viewer

Rules Itemsets Dependency Network

Minimum support: 454 Filter Itemset:

Minimum itemset size: 0 Show: Show attribute name and value

Maximum rows: 2000  Show long name

Support	Size	Itemset
428485	1	Academic = Existing
290862	1	Province = Middle
278006	2	Province = Middle, Academic = Existing
68742	1	Province = West
66539	2	Province = West, Academic = Existing
43169	1	Province = North
42189	2	Province = North, Academic = Existing
22423	1	Province = East
21836	2	Province = East, Academic = Existing
19294	1	Province = South
18793	2	Province = South, Academic = Existing
8320	1	Arabic Certificate = Existing
7744	2	Arabic Certificate = Existing, Province = Middle
3046	1	Commercial = Existing
2909	1	Scientific Institutes = Existing
2112	1	Industrial = Existing
1902	2	Scientific Institutes = Existing, Province = Middle
1647	2	Commercial = Existing, Province = Middle
1063	1	Province = Abroad
1063	2	Province = Abroad, Academic = Existing
986	2	Industrial = Existing, Province = Middle
793	2	Scientific Institutes = Existing, Province = West
652	2	Commercial = Existing, Province = West
481	2	Industrial = Existing, Province = West
454	2	Industrial = Existing, Province = North

Fig.4.37: The found itemsets for the of Major\_Province Model

Mining Model: Major\_Province Viewer: Microsoft Association Rules Viewer

Rules Itemsets Dependency Network

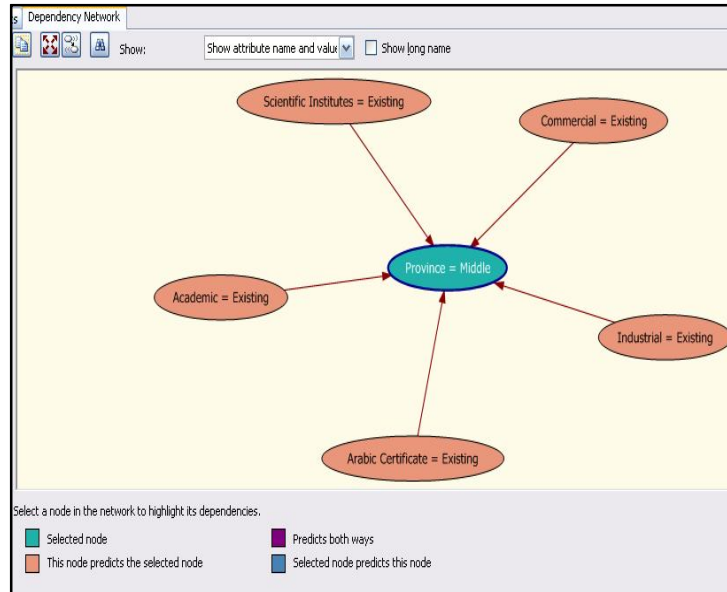
Minimum probability: 0.46 Filter Rule:

Minimum importance: -0.15 Show: Show attribute name and value

Show long name Maximum rows: 2000

Probability	Importance	Rule
0.931	0.158	Arabic Certificate = Existing -> Province = Middle
0.654	0.001	Scientific Institutes = Existing -> Province = Middle
0.649	-0.063	Academic = Existing -> Province = Middle
0.541	-0.082	Commercial = Existing -> Province = Middle
0.467	-0.146	Industrial = Existing -> Province = Middle

Fig.4.38: The discovered rules for the of Major\_Province Model



**Fig.4.39:** The dependency net view of Major\_Province Model

Traditional data analysis methods are often used to write statistical queries about data. Through this chapter, we could slice the data using On-line Analytical Processing (OLAP) tools to answer many queries about an educational data; for example we could find out how many male students admitted in various colleges versus female students, their associated scores average, etc. We could also write a query to see various factors along multi dimensions. However, such traditional methods are very convenient to perform statistical operations on data in multidimensional structures, but they are limited to find the relationship between items and we regularly must write dozens of queries to cover all the possible combinations. In contrast, the data mining approach is the solution to such problems; it reduces the queries number by allowing users to ask a question in terms of data that can support many hypotheses and then explore them with users. Through this chapter, we could use two types of data warehouse structures; snowflakes to develop the OLAP system and star structure for the data mining process.

In general, OLAP and data mining techniques are very effective analytical tools in the data mining world. OLAP is superior at aggregating a large amount of multidimensional data; it enables users to perform some statistical operations on substructures of any data cube structure. Whereas, data mining is superior at finding the hidden patterns of a huge data set; it enables users to analyze correlations among attribute values.

Automatic methodologies are commonly used to analyze data to find hidden patterns in the data mining process. Association rules mining is one of such automatic methodologies. Data mining algorithms that could be applied to a data set analyze data to produce patterns that could be explored for valuable information. These patterns can be in a form of rules and then used for reporting. Through this chapter, we could find many associations among data of students.

Having discussions with various academicians and colleges in different institutions helped us in finding, presenting the techniques, processing and applying data mining in higher education data in Sudan.

## Chapter 5

### Evaluation and Conclusions

In the context of using data mining techniques in higher education domain, we found that most of the published works used data mining techniques other than the association mining technique. They often used predictive techniques such as; classification to predict student success in a certain course, predict student enrollment in a certain college, etc. Through our study, we participated in using the association mining technique in higher education domain to investigate admissions to a number of the Sudanese universities.

Limited works have been undertaken to develop a client application using OLAP structure. With this study, we introduce the first client application in Sudan that was developed based on OLAP structure.

Hard work has helped us to accomplish deeply mined students' data using some of the data mining techniques; the greatest efforts went to data preprocessing stage. The main strategy of this research methodology was divided into two paths: developing an OLAP system and applying the association rules mining as one of data mining techniques. The developed OLAP system provides end users with only one friendly user interface, through that, users can query the OLAP system across different dimensions and get quick answers in a short time. The developed OLAP system strongly helps decision makers of the Ministry of Higher Education in making their right decisions based on integrated information. Through applying the association rules mining technique, we could find many correlations between the students attributes. Such correlations could be exploited in decision making process that relate to application and admissions processes in Sudanese universities.

From our developed OLAP cube, many reports were generated, as well as analyzed based on generated rules from the association rules mining. We discuss those reports in section 5.1. In the previous chapter, we developed some mining models; those models will be interpreted and evaluated in section 5.2. Conclusions of this research are the topic of section 5.3. We offer some recommendations and researches for the future in section 5.4.

## **5.1 Results**

This section of the chapter sets out some results that were produced from implementations of OLAP cube and data mining process.

All reports from OLAP cube, as presented in the previous chapter, were produced by using three measures: scores average percent, filled choices percent, and maximum scores average. Those measures were performed along different dimensions; they covered the period of (2005 - 2009). Furthermore, a grand total column summed results across all years as noted in any report. For measuring filled choices percent, the results were shown in figures 4.6.b, 4.7.b, 4.8.b, 4.9.b, 4.10.b, 4.11.b, and 4.12.b. In general, this measure was tested across all dimensions by observing the grand total column. All reports registered a percentage of < 50 percent. Measuring scores average percent of admissions was done along dimensions of college, location, high school, and student's sex. Table 5.1 summarizes some results produced by observing the grand total column values of each report in figures (4.6.a – 4.12.a).

**Table 5.1:** Measuring of scores average percent along several dimensions.

<b>Dimension Name</b>	<b>Attribute Name</b>	<b>Grand total of scores average percent (2005 - 2009)</b>
Location	East	69.48 %
	Middle	71.86 %
	North	71.56 %
	South	68.29 %
	West	66.95 %
High School	Academic	70.46 %
	Commercial	67.66 %
	Industrial	67.35 %
	Agricultural	67.24 %
	Scientific Institutes	74.32 %
	Hafaza	75.40 %
	Neswyha	67.61 %
	Arabic Certificate	87.05 %
	Aedeen (from abroad)	58.38 %
	Public (Ownership)	70.58 %
	Private (Ownership)	71.18 %
	Regular (Admission)	71.97 %
	Home (Admission)	69.14 %
	Teachers Schools (Admission)	68.80 %
	Female (Type)	70.66 %
	Male (Type)	70.00 %
Coed (Type)	71.93 %	
Student Sex	Female	71.06 %
	Male	70.33 %

Many times data miners may want to facilitate exploring data more thoroughly in order to help decision makers in making right and quick decisions. Therefore we used maximum scores average along different dimensions as well as

in the previous table. Table 5.2 shows some results produced by observing the grand total column values of each report in figures (4.13.a – 4.14.b).

**Table 5.2:** Measuring of maximum scores average along several dimensions.

<b>Dimension Name</b>	<b>Attribute Name</b>	<b>Grand total of maximum scores average (2005 - 2009)</b>
Location	East	98.83 %
	Middle	100.00 %
	North	95.43 %
	South	100.00 %
	West	97.47 %
High School	Academic	97.32 %
	Commercial	87.14 %
	Industrial	86.57 %
	Agricultural	78.00 %
	Scientific Institutes	97.43 %
	Hafaza	99.00 %
	Neswyha	79.29 %
	Arabic Certificate	100.00 %
	Aedeen (from abroad)	93.00 %
	Public (Ownership)	100.00 %
	Private (Ownership)	96.48 %
	Regular (Admission)	100.00 %
	Home (Admission)	93.86 %
	Teachers Schools (Admission)	95.00 %
	Female (Type)	96.86 %
	Male (Type)	99.00 %
Coed (Type)	100.00 %	

One of the greatest advantages of data mining techniques is that they enable us to generate results for both aggregated and individual items. We can develop

model that could be used to find associations between each college and students who belong to a certain province.

Student\_Admission\_Model was designed to predict associations between: student's province and student's preference college; province was filtered based on the province name to give an individual report for each province. Data mining process produced three types of results: itemset, generated rules and dependency net views. We summarized reports of generated Itemsets from that model as shown in figures 4.26.a, 4.30.b and in table 5.3. Generated rules are summarized in table 5.4.

**Table5.3:** Itemsets for Student\_Admission\_Model

Province	Maximum Support	Minimum Support
Middle	Omdurman aleslamiah	Alsharaq Elahleya
North	Wadi Elniel	Elemam Elhadi
West	Kurdofan	Africa Alalemiah
East	Elbaher Elahmer	Omdurman Alahleya
South	Juba	Quraan Kareem

**Table5.4:** The generated rules for Student\_Admission\_Model

Province	Maximum Probability (Confidence)		Minimum Probability (Confidence)	
	University Name	Importance	University Name	Importance
Middle	Elniel Alazraq	<1	Elemam Elmahadi	<1
North	Wadi Elniel	>1	Dungula	<1
West	Niyala Elteghaneia	<1	Eldalang	<1
East	Kassala Elteghaneia	>1	Alsharaq Elahleya	>1
South	Rumbake	>1	Juba	>1

Two other mining models were developed to study how the high school major often is associated with some student's demographics. High\_Schoo\_Major\_Sex model was developed to find the associations between the



high school major and student's sex, whereas Major\_Province model was developed to find the associations between the high school major and student's province. Table 5.5 describes itemsets resulting of the implementations of association rules mining algorithm on the both models. Results were summarized in figures 4.35, and 4.38.

**Table5.5:** Itemsets of High\_Schoo\_Major\_Sex model and Major\_Province model

Model Name	Maximum Support		Minimum Support	
	Item Name	Size	Item Name	Size
High_Schoo_Major_Sex	Academic	1	Industrial	1
	Female	1	Male	1
Major_Province	Academic	1	Industrial	1
	Middle	1	South	1

Many results could be generated from our developed mining models implementations, but which of them could be significant? The next section evaluates those generated results.

## 5.2 Evaluation and Discussion

Through this research, we could approve advantages of using data mining techniques against traditional statistical analysis tools; we could use data mining techniques to handle large data sets where we could analysis 446112 records of real word data. Additionally, we could apply very efficient and scalable algorithms on historical and updatable sort of data. Those data were stored in a centralized data repository called Data Warehouse to be available for decision makers. With data mining techniques, decision makers could easily interpret results without the

need of expert user guidance unlike traditional statistical analysis tools where their results are difficult to understand by end users. Moreover, we could use data mining techniques to measure students' admissions along different angles i.e. Students' admissions per province, per state, per high school major, per student's sex, etc.

We developed an OLAP cube which provided us with many reports of its implementations. By using those reports of OLAP cube implementations, we could investigate some findings that should be discussed. For example, Figures 4.6.a – 4.12.b show that when we measured filled choices percent along different dimensions, all reports registered a percentage of < 50 percent. Those figures also show that students who belong to middle and north of Sudan have the similar percentages of scores average percent and filled choices percent. Additionally, we found that students who belong to south of Sudan have achieved the smallest filled choices percent (24.62 % as shown in figure 4.7.b). Using table1, we noted that students who belong to west of Sudan have achieved the lowest scores average percent comparing with other provinces. The less scores average percent (66.95 % as shown in figure 4.7.a) achieved by students who belong to west of Sudan reflects the bad effects of war and unstable situations on the education sector. Sudan has a large agricultural land, a long Nile; however, students who chose the agricultural major in their high school had achieved the semi lowest scores average percent (67.24 %) comparing with other high school majors. Table1 shows that the lowest scores average percent, when it measured against all the high school majors, was “Aedeen”. That major contains student who came back from abroad; their scores average percent was 58.38 % over all the five years. At that period of years, a small difference appeared in scores average percent between public high schools and private high schools. Public high schools had achieved scores average percent of 70.58 % whereas private high schools had achieved scores average percent of

71.18 %. Students who admitted in high schools as regular, then they repeated their academic year due to their low scores average percent; they were often admitted again in high schools as teacher schools or home students. The category of teacher schools had achieved the lowest scores average percent (68.80 %) compared with other types of admissions where they achieved 71.97 % and 69.14 % for regular and home respectively. Table 5.1 also shows that students who studied at high schools for Coed had achieved the highest scores average percent compared with students who studied at high schools for female and male respectively.

When we measured maximum scores average along location dimension, the Northern Province showed the lowest maximum scores average (95.43 %) compared with other provinces as shown in table 5.2. Table 5.2 shows that, again, the agricultural major had achieved the lowest maximum scores average (78.00 %) compared with other high school majors. It also shows that students who had graduated from private high schools had achieved the lowest maximum scores average compared with who had graduated of public high schools. In contrast, students who had graduated from coed high schools had achieved the highest maximum scores average compared with others who had graduated from female/male high schools.

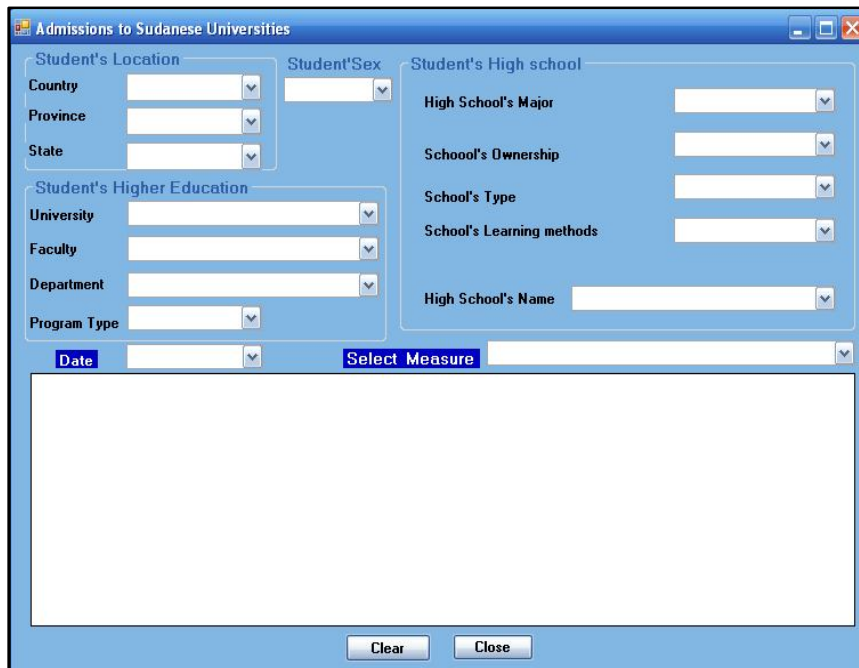
Nevertheless, all values on both tables 5.1 and 5.2 were summarized from reports that generated from our developed OLAP cube implementations.

Reports generated from OLAP cube often are not accessible by end users. To enable end users to access those reports, a smart graphical user interface (GUI) has been developed as a result of programming an OLAP system. GUI typically displays a representation to the end users in controllable form. The user could then specify a query by selecting certain dimensions. If we compared the developed OLAP system with traditional data analysis systems, we find that through OLAP system we could build a summarized base level of data; for instance, we could

query the system about scores average of students who applied to Sudanese universities at 2005, who belong to the Northern Province in Sudan, and are females. Also, we could benefit from the aggregation of data to navigate to various levels of aggregation with multidimensional views of the data. For example, using some OLAP operations as slicing, dicing, drilling down, and rolling up, we could measure admissions of students at different levels such as admissions per country, admissions per province, or admissions per state. Using our developed OLAP system, users can work with the databases and tables without need to know any knowledge of query language or SQL and do not need to type any query, but they only need to choose their desired dimensions according to their queries. As soon as users selected their dimensions to form their query, our developed OLAP will respond to them in seconds at the bottom area of the single user interface shown in figure 5.1.

Database programmers often need to write very complex codes in order to answer only one query about data; for example to answer a query such as: what is the scores average of students who live in Khartoum state, applied to Ahfad university for women, and got their high secondary certificate from public schools? To answer such queries, database programmers may need to develop multiple views among very complicated code; they also need more time to get their appropriate results. Which OLAP programmers do not? Our developed OLAP system in Figure 5.1 shows how users could easily answer such query by selecting their desired cases among different dimensions. Developing a client application based on traditional database will enforce end users to query the data about only their developed views and provides them with limited freedom of choosing multiple cases. In contrast, using OLAP system has beneficial effects to end users in allowing them to feel free in formulating their queries from different

dimensions. Those beneficial effects were confirmed due to the exceptional structure of OLAP.

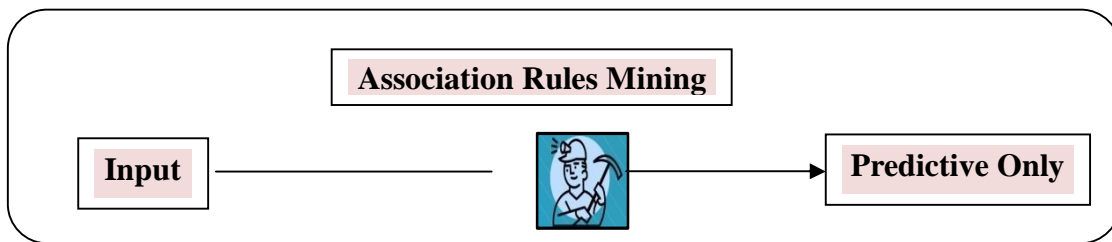


**Fig.5.1:** The developed OLAP system

By developing our OLAP system we could answer typical questions that related to the concept of data aggregation. To analyze data based on the concept of data correlation in order to discover useful knowledge, we developed several data mining models. All the collected data were partitioned randomly into two data sets: training and testing. The training data set was used to develop our developed models, whereas the testing set was used to validate the model developed by the training data set. All models applied only one type of data mining techniques, which was association rules mining technique. The only differences between all developed models were due to their different parameters or difference in ways of using the parameters. All models resulted in associations between several admissions' factors; for example, we could study associations between students who belong to a certain province and their preference college, we could study students' admissions based on the association between students who belong to a

certain province and their sex. In general, we could answer some of our research questions such as how the students based on their demographics (Sex, Country, State, etc.) affect the admissions process.

Luckily, however the association rules mining technique is categorized as one of descriptive data mining techniques; it could be also used as a predictive data mining technique. Through our research we could create several predictive models using the association rules mining technique. Those predictive models predicted outputs (Predictive Only) based on knowing their associated inputs. The term “Predictive only” indicates that output could not be used as input at the same time. Figure 5.2 illustrates the general idea of our developed predictive models’ task.



**Fig.5.2:** The predictive models’ task

Prediction results of our developed predictive models were evaluated in several ways. Since we used the association rules mining technique, our developed models’ prediction results were evaluated using reports of: itemsets found, generated rules, and dependency net views of the developed predictive models.

The itemsets reports produced the frequent itemsets associated with their support and size. Through the “Support” of generated itemset, we could measure the popularity of an itemset; whereas, through the “Size” of the found itemset, we could determine the number of iteration that the algorithm could scan in the data set and count the supports for each generated itemset. For example, the algorithm finds all frequent itemsets with size = 1 in the first iteration; the second iteration

found the frequent itemsets of size = 2 based on the result of first iteration (size = 1) and so on.

Reports of generated rules efficiently produced the popular generated rules; each rule was associated with its probability (confidence) and its importance. Through the “Probability” of the generated rule, we could measure the reliability of each rule; the higher the value of the probability (confidence), indicated the more often that set of items was associated together. Whereas, through the “importance” of the generated rule, we could measure the interesting score and the usefulness of each rule; i.e. the higher the importance score the better the quality of the rule is. Table 5.6.a shows the bases of applying the importance values for itemsets on two items A and B in order to specify the associations between them. Table 5.6.b shows the bases of applying the importance values for rules on two items A and B.

**Table 5.6.a:** Applying “importance” measure on two items A and B for itemsets

<b>Importance value</b>	<b>Description of associations</b>
= 1	A and B are independent items
< 1	A and B are negatively associated
> 1	A and B are positively associated

**Table 5.6.b:** Applying “importance” measure on two items A and B for rules

<b>Importance value</b>	<b>Description of associations</b>
= 0	There is no association between A and B
> 0	The probability of B goes up when A is true
< 0	The probability of B goes down when A is true

The study also evaluated the predictions from our developed models by using reports of dependency net views; the dependency net views graphically described the relationships between attribute values instead of model attributes.

Considering all pervious evaluation concepts, we applied those concepts on our developed models. By implementations of Student\_Admission\_Model, we noticed that most of the itemsets were discovered at size = 2; while the minimum support was set to 70. By taking a look at the itemsets reports generated from that model, as shown in table 5.3, we could deduce that, although the middle province usually contains students from all other provinces, the University of Omdurman Aleslamiah was very attractive for students who belong to the middle province. The University of Omdurman Aleslamiah has a lot of faculties of different specializations including most of the Islamic studies. It has many buildings that are distributed throughout all the country, enabling female students to study at separated buildings from male students. Female students are enforced to wear Islamic hiejab. Students' preference for such universities could indicate the high Islamic sprit of students who live in the middle province. The associations between other universities and provinces indicated that most of the students prefer to be admitted to universities that are located in the same province where they live. The minimum support values in table5.3 shows students who belong to South of Sudan have a lower preference for admission at the University of El Quraan Kareem, which indicates logically that these results are justified since most of people from the southern province are not Muslims.

To evaluate results in table 5.4, we applied concepts in table5.6. In regard to importance values provided in table 5.4, we deduced that students who belong to middle province may not be admitted to Elniel Alazraq university although they achieved the maximum probability, since their importance of rules have a negative correlation, as did the students who belong to west province. Rule sets generated



by Association rules mining technique showed that most of the students prefer to be admitted in their local province. Tables 5.3 and 5.4 summarized the results of itemsets and generated rules by implementations of Student\_Admission\_Model. Student\_Admission\_Model was also evaluated by using dependency net view as shown in figures 4.31.a and 4.31.b. Each edge of a dependency net view represents a pair wise association rule. The slider is associated with rules that scored the higher importance values. Figure 4.31.b shows that students who belong to the east province did not associate with any college. However, they achieved the highest importance as shown in table 5.4, where the east province was associated with Kassala Eltegneia. Referring to Figure 4.29.b, we deduced that rule had achieved a low probability (confidence) compared with other rules that associated other provinces with student admissions.

Student admissions were also evaluated against several high schools' factors. Two extra models were developed: High\_School\_Major\_Sex and Major\_Province model. Produced itemsets of both models were shown in table 5.5 in a summarized view. By evaluating results in table 5.5, we observed that the academic major of high school achieved the highest maximum support, whereas the industrial major of high school achieved the lowest minimum support of occurrence. In the achievement of the middle province, the highest maximum support indicated the high level of education in that province; the opposite occurred in the south province. Table 5.5 also shows female students had larger opportunities to get their education compared with male students, which may have happened due to the increasing number of female students over male students in recent years. Implementations of High\_School\_Major\_Sex model generated rules that described the associations between high school major and students' sex. Generated rules in figure 4.38 shows that the industrial major achieved the highest importance score for male students; it also shows female students were very

interested in studying commercial studies in contrast to male students. Implementations of Major\_Province model indicate that the middle province was attractive to a wider range of majors. Figure 4.38 shows how it enforced its appearance among all other provinces; whereas, dependency net views, in figure 4.39, shows how it implies other the high school majors.

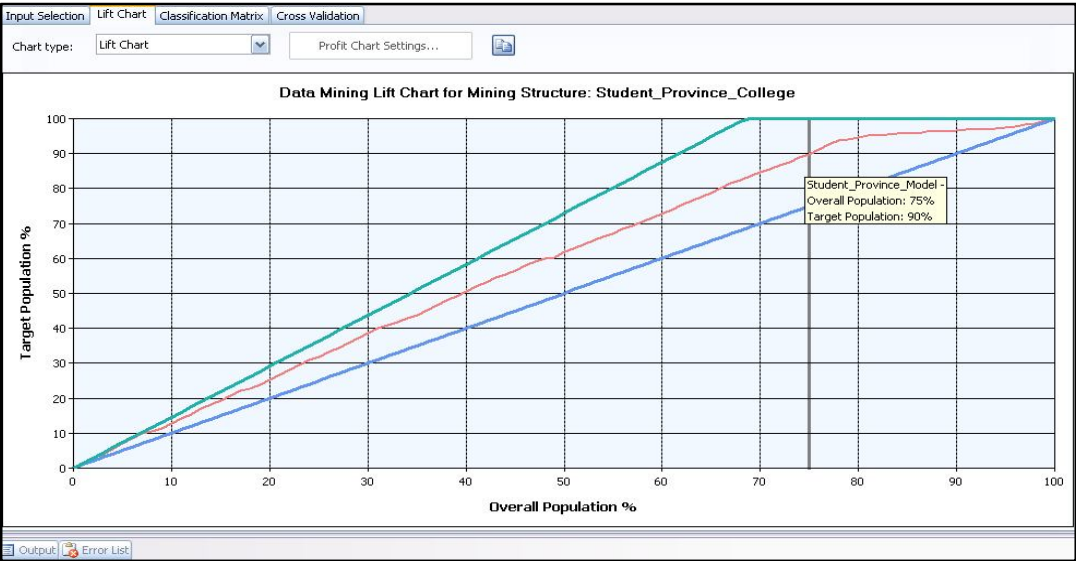
Throughout that study we developed several data mining models; all of which gave us logical results. However, do we need to determine which model is better? After we evaluated the results generated from our developed models implementations, we also evaluated our developed models for accuracy.

Fortunately, most of our developed models included an evaluation tools. In the contexts of exploring the generated results from data mining process, we argued the accuracy of two models and drew a comparison between them.

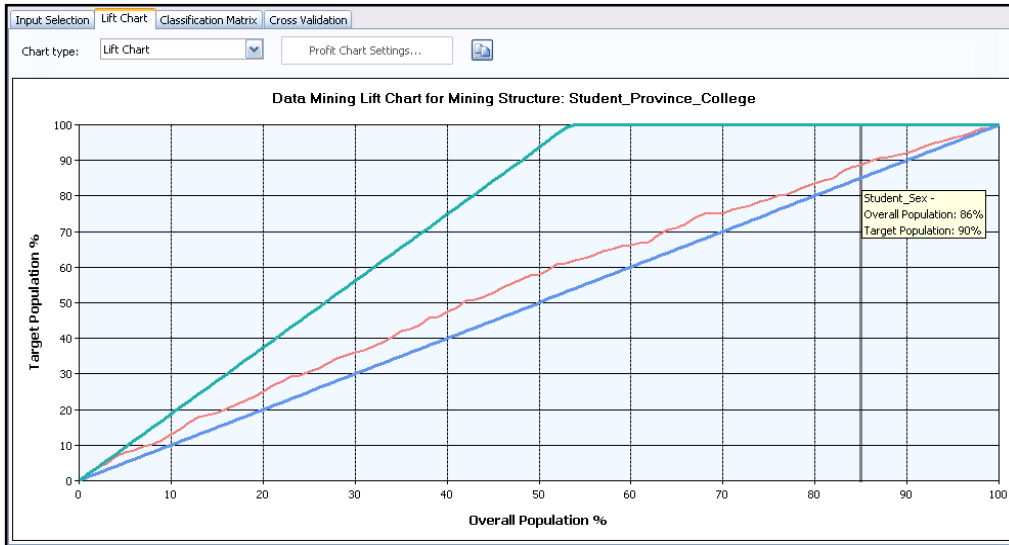
To gauge the quality and accuracy of the models, we used some of the Mining Accuracy Chart pane tools such as Profit chart, which performs predictions against our models, and compares the result to data for which we already know the answer. In that task, we used mining model test cases for two created models: Student\_province\_Model which predict what province student came from, and Student\_Sex\_Model which predict student's gender. For measuring the accuracy of the both models, we used a lift chart, shown in figure 5.3 that provided us with a discrete target value to predict for each model. A lift chart always contains a single line for each selected model; that line is often located between other two lines: an ideal line and a random line. As shown in figure 5.3, the ideal line often appears at the top of all other lines on the lift chart. It shows that an ideal line of Student\_province\_Model would capture 100 percent of the target data using 69 % of the data. This simply implies that 69 percent of the data indicates the desired target. Whereas, the random line appears at the bottom of the lift chart lines which is constantly a 45-degree line across the chart and indicates that if we were to

randomly guess the result for each case, we would capture 50 percent of the target using 50 percent of the data. Luckily, both of our model lines appear above the random line. Hovering model lines around the random guess line indicate that there was not adequate information in the training data to learn patterns about the target.

Using the lift chart instance in Figure 5.3, the Student\_province\_Model can get about 90 % of the target using only 75 % of the overall data; whereas the Student\_Sex\_Model, as shown in Figure 5.4, can get 90 % of the target only by using 86% of the data. Therefore, this instance indicates that the Student\_province\_Model performs better than the Student\_Sex\_Model.



**Fig.5.3:** The lift chart of the Student\_province\_Model, where 90 % of the target can be captured using 75 % of the data



**Fig.5.4:** The lift chart of the Student\_Sex\_Model, where 90 % of the target can be captured using 86 % of the data

Through this section, we could present some evaluation factors for most of our practical work in which we evaluated reports that were generated from our developed: OLAP cube, OLAP system, and data mining models created based on using association rules mining algorithm. We evaluated data mining models by three association rules mining techniques: itemsets found, generated rules, and dependency net view. We discussed some results based on the values and percentages provided by our developed data mining models. Through that discussion, we could extract different types of associations related to students' admissions to Sudanese universities i.e., the associations between students' college preference and the province where they live, the associations between students' sex and their high school major's preference, etc. Furthermore, we evaluated the accuracy of our developed models.

### 5.3 Conclusions

The objective of this study was to develop a methodology that uses data mining techniques to perform complex analysis tasks on higher education domain. Those analysis tasks were applied on the application and admission processes for students who were admitted in Sudanese universities within the period (2005 - 2009).

In this research, we tried to bring data warehouse environment to higher education domain using two types of its structures; star and snowflakes. The star structure was used to develop several data mining models; whereas the snowflakes structure was used to develop an OLAP system.

Before developing our OLAP system, we developed an OLAP cube through which we could generate several reports about students' admissions. The reports are often deployed by selecting different dimensions with different measures. Unfortunately, such reports cannot be deployed as stand-alone – application. The solution was to develop an OLAP system as one of our research objectives. Our developed OLAP system was built based on our OLAP cube; it provides a user friendly and single interface where end users can rapidly answer queries such as, : what is the maximum scores average of female students?, who live in Khartoum state? , graduated of commercial high school major? , and was admitted in Sudanese universities in 2007. Fortunately, an OLAP has amazing operations such as: slicing, dicing, drilling down, and rolling up, where end users can systematically explore the knowledge space to find out useful knowledge. Therefore, through our developed OLAP system, end users can answer queries containing such operations. They can drill down along the higher education dimension to answer queries such as; what are the scores averages of male students who were admitted in the University of Khartoum, Faculty of Science, and

Department of Physics. The end users will receive a system response in seconds. However, our developed OLAP system smartly enables end users to explore the data flexibly from different dimensions and at multiple abstraction levels; it requires further implementation to achieve full functionality as well as any developed user interface. Nevertheless, it meets most of the required functionality already identified; it also satisfied our objective of developing an OLAP system for end users with no experience of data mining or programming. One of the biggest challenges that higher education faces today is analyzing the huge amount of students' data. Higher education institutions would like to answer some questions that relate to the application and admission processes such as which students enroll in particular type programs. What types of universities attract more students, etc? By developing our OLAP system of admissions to Sudanese universities, we could provide the decision makers of the Ministry of Higher Education and Scientific Research with information necessary for determining how to enhance the application and admission processes.

Through our developed OLAP system, we could answer queries that have a statistical nature, (Max, Min, Average, etc.), to answer queries that related to studying correlations between items, we developed several data mining-based predictive models. Through these models, we could evaluate and analyze the dominant factors that affect a student's college preference. Using the association rules mining technique, we determined Predictions by interactions of factors such as province, high school major, student's sex, etc. We could answer most of our research questions such as: are there any relationships between students' demographic information such as their higher schools, states, etc. and their strategies for applying to Sudanese universities? Our developed data mining-based predictive models were a combination of OLAP structure with association rules

mining technique to create OLAM models. All our developed OLAM models (OLAP- Mining) were constructed using the SQL Server 2008 software.

Superficially, data-mining models may appear to be unconventional and non theoretical, but our developed models in this study proved that results were practical, reliable, actionable and, thus, highly desirable to admissions professionals. Due of its practical features, data-mining technology has a lot of potential in higher education in addition to the business world. By providing practical results, the Ministry of Higher Education and Scientific Research researchers can better inform and assist university administrators with data-supported decision making.

Through this work, we evaluated reports generated from our developed: OLAP cube, OLAP system, and data mining models. Furthermore, we also evaluated our developed data mining models in the perspective of accuracy. From the results analysis and data mining models evaluation, we can conclude that data mining techniques are applicable for the cooperative research; association rule mining can be effectively conducted to find the relationships between different factors related to students admissions. It provides a relatively easy way to rapidly identify previously unknown relationships among the attributes within a student cohort dataset. These relationships may provide the Ministry of Higher Education and Scientific Research policy analysts with the necessary information for supporting operational changes in order to enhance higher education institutions. They can build models that predict—with a high degree of accuracy—admissions to Sudanese universities. By acting on these predictive models, higher education institutions can effectively address issues that could limit some admissions problems such as transfers, dropping out, and retention. In general, such enhancements will affect the quality, effectiveness, and efficiency of higher education admissions.

In this study, we could explore a major difference between statistical models and data-mining predictive modeling, where classical statistical models do best drawing general conclusions about average and group means; whereas data mining–based predictive models make predictions for individual records using complex sets of rules.

In this study, the research questions are solved by developing an OLAP system and several mining models. We could answer typical questions such as: is there need for all these number of choices?, how many faculty choices qualify a student to be admitted to his/her preferred college?, what percent of choices do students usually fill in the application form in regard to the variation in residential regions? Moreover, we supported that data mining techniques have huge potential benefits in terms of multidimensional analysis and can help to solve Sudan’s education need for skilled analysts.

Those conclusions gave us hints to proffer recommendations and ideas for future research as discussed in the next section.

#### **5.4 Recommendations and future work**

Since the beginning of this PhD, in November 2011, the available systematic data were limited to years from 2005 up to 2009. This study included student data of south Sudan despite the separation event because the separation South of Sudan from North was at 2011. We recommend making a similar study that includes data after 2011. That will indicate any effects on education in Sudan resulting from the separation.

The greatest challenge that faces the researchers in Sudan is getting data due to two reasons: most of the available data has not been systemized yet and information security issues.



Results from our developed mining models seemed to suggest that admitted students did not enroll randomly at the Sudanese universities; their enrollment decision could be predicted. A reasonably large portion of student enrollment could be accurately predicted by several predictive models.

We live in the information age; people need access to information quickly and accurately. As a future endeavor, we recommend stronger improvements to this system. We recommend developing a web version of the system to be available on the internet for the students in Sudan and abroad.

Traditional data analysis methods became a source of difficulty for many researchers because they expended great efforts with slow process in finding information located in different files. In order to avoid wasting effort and time in such data analysis tasks, we recommend researchers be very keen to deal with data mining techniques. Cost, effort, and time are the most challenging issues facing the Ministry of Higher Education and Scientific Research officers in the processes of application and admissions. The Ministry of Higher Education and Scientific Research can be more cost effective in designing student application forms, directory books, employing extra officers to systemize information on the application forms, and dropping students with unqualified scores. Utilization of data mining will greatly enhance the enrollment processes at the Ministry of Higher Education and Scientific Research.

The development of OLAP systems could have broader application in other sectors in Sudan as well such as: health, law, business, economic, and in particular in the religion sector to rapidly solve some religious conflicts. Our challenges are extensive and using these research techniques only requires the willingness and dedication of researchers to explore each and every possibility.

## References:

**Alaa** Al Deen Mustafa Nofal, Sulieman Bani-Ahmed 2003, “*Classification based on Association Rule Mining techniques: a general survey and empirical comparative evaluation*”, Journal of Ubiquitous Computing and Communication, Vol. 5, No 3, pp.9-17. Jordan, Available from: <http://www.ubicc.org/> , [visited at 24 April 2013].

**Amit** Gupta 2010, Available from:  
<http://www.msbi-concepts.com/2010/09/analysis-services-storage-modes.html>,  
[visited at 28 April 2013].

**Antoaneta** Ivanova, Boris Rachev, 2004, “*Multidimensional models - Constructing Data Cube*” International Conference on Computer Systems and Technologies- CompSysTech, available from:  
<http://ecet.ecs.ru.acad.bg/cst04/Docs/sV/55.pdf>, [Downloaded at 24 April 2013].

**Ashutosh** Nandeshwar, Subodh Chaudhari, 2009, “*Enrollment Prediction Models Using Data mining*”, available from: [http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU\\_Project.pdf](http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf), [Downloaded on 22 March 2011].

**Bing** Lui, Alexander Tuzhilin, 2008, “*Managing Large Collections of Data mining Models*”, communications of the AMC, Volume 51, No. 2. 2008, USA, pp 85 -89, Available from: - -  
[http://www.cs.uic.edu/~liub/publications/papers\\_chron.html#Rule](http://www.cs.uic.edu/~liub/publications/papers_chron.html#Rule), [downloaded on 22-June-11].

**Bing** Lui, Kaidi Zhao, Jeffrey Benkler, Weimin Xiao, 2006, “*Rule Interestingness Analysis Using OLAP Operations.*” Philadelphia, Pennsylvania, USA. Available from: [http://www.cs.uic.edu/~liub/publications/papers\\_chron.html](http://www.cs.uic.edu/~liub/publications/papers_chron.html)  
[http://www.google.com/#sclient=psy&hl=en&source=hp&q=former+and+latter+definition+in+data+mining&aq=f&aqi=&aql=&oq=&pbx=1&bav=on.2,or.r\\_gc.r\\_pw.&fp=deb24ccd75812d3e&biw=1024&bih=431-](http://www.google.com/#sclient=psy&hl=en&source=hp&q=former+and+latter+definition+in+data+mining&aq=f&aqi=&aql=&oq=&pbx=1&bav=on.2,or.r_gc.r_pw.&fp=deb24ccd75812d3e&biw=1024&bih=431-) [downloaded at 22 June 2011].

**Cesar** Vialardi, Javier Bravo, Leila Shafti, Alvaro Ortigosa, 2007, “*Rcommendation in Higher Education Using Data mining Techniques*”, Available from: <http://arantxa.ii.uam.es/~jbravo/papers/vialardi-bravo-shafti-ortigosa-EDM09.pdf>, [Downloaded at 24April 2013].

**Dimitar** Hristovskia, Janez Starea, Borut Peterlinb, Saso Dzeroskic, 2001, “*Supporting Discovery in Medicine by Association Rule Mining in Medicine and UMLS*”. Available from: Databases  
<http://dml.cs.byu.edu/~cgc/docs/atdm/Hristovski.pdf>, [Downloaded on 10 Jul 2011].

**Fadi** Abdeljaher Thabtah, Peter Cowling, Yonghong Peng, 2006, “*Multiple Labels Associative Classifications*” *Journal of Knowledge and Information Systems*, Know Inf syst (2006) 9(1), Pages 109 – 1129, Bradford.

**Fadi** Thabtah, Peter Cowling, 2008, “*Mining the data from a hyperheuristic approach using associative Classification*”, expert systems with applications 34 (2008) pp.1093- 1101, Available from <http://www.sciencedirect.com>, [Downloaded at 26 May 2011].

**Galina** Bogdanova, Tsvetanka Georgieva, 2005, “*Discovering the Association Rules in OLAP Data Cube with Daily Downloads of Folklore Materials*”, International Conference on Computer Systems and Technologies - CompSysTech’, Available from: <http://ecet.ecs.ru.acad.bg/cst05/Docs/cp/SIII/IIIB.23.pdf>, [visited at 24 April 2013].

**Jamie** Maclennan, ZhaoHui Tang, Bogdan Crivat, 2009 “*Data mining with Microsoft SQL Server 2008*”, Wiley Publishing, Inc., Indianapolis, Indiana.

**Jiawei** Han, Micheline Kamber, 2006, “*Data mining: Concepts and Techniques*”, Morgan Kaufmann, San Francisco, CA 94111.

**Jigna** J. Jadav, Mahesh Panchal, 2012, “*Association Rule Mining Method On OLAP Cube*”, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 [www.ijera.com](http://www.ijera.com) Vol. 2, Issue 2, Mar-Apr 2012, pp.1147-1151. India. Available from: [http://www.ijera.com/papers/Vol2\\_issue2/GL2211471151.pdf](http://www.ijera.com/papers/Vol2_issue2/GL2211471151.pdf), [downloaded at 24 April 2013].

**Jing** Luan, 2002, “*Data mining and Its Applications in Higher Education*”, New Directions for Institutional Research, 2002: 17–36. doi: 10.1002/ir.35, No.113. Wiley Periodicals, Inc, California, Available from: <http://freepdfdb.com/pdf/data-mining-and-its-applications-in-higher-education-29039827.html>, [visited at 25March 2011]

**Joseph** Zalaket 2012 “*Enhancing the Search in MOLAP Sparse Data*”, information, www.mdpi.com/journal/information, ISSN 2078-2489, 14 November 2012, pp.661- 675, Available from: <http://www.mdpi.com/2078-2489/3/4/661/pdf>, [Downloaded at 24April 2013].

**Kifaya S.** Qaddoum, 2009 “*Mining Student Evolution Using Associative Classification and Clustering*”, Communications of the IBIMA Volume 11, 2009, ISSN: 1943-7765, Amman, Jordan, Available from: <http://eric.univ-lyon2.fr/~sabine/dolap06.pdf>, [Downloaded at 24April 2013].

**Kirk** Haselden, 2009, “*Microsoft® SQL Server™ 2008 Integration Services*”, Pearson Education, Inc., Library of Congress Cataloging, by Pearson Education, Inc- ISBN-13: 978-0-672-33032-2, United States of America.

**Lin** Chang, 2006, “*Applying Data mining to Predict College Admissions Yield: A Case Study*”, NEW DIRECTIONS FOR INSTITUTIONAL RESEARCH, no. 131, Fall 2006 © Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com) • DOI: 10.1002/ir.187, pp. 53 - 68 available from: [www.interscience.wiley.com](http://www.interscience.wiley.com), [Downloaded on 22 March 2011].

**Mike** Hotek, 2009 “*Microsoft SQL Server 2008 - Step by Step*”. Microsoft Press- a Division of Microsoft Corporation- One Microsoft way- Redmond, Washington.

**Mohammad** A. Rob, Michael E. Ellis, 2007, “*Case Projects in Data Warehousing and Data mining*”, University of Houston, Clear Lake, Volume VIII No.1. Available from: [http://iacis.org/iis/2007/Rob\\_Ellis.pdf](http://iacis.org/iis/2007/Rob_Ellis.pdf), [downloaded at 15 March 2011].

**Mohammad** Naderi Dehkordi, 2013, “*A novel Association Rule Hiding Approach in OLAP Data Cubes*”, Indian Journal of Science and Technology Vol: 6 Issue: 2 February 2013 ISSN: 0974-6846, India. pp. 89-101 available from: <http://www.indjst.org/index.php/indjst/article/viewFile/30587/26506>, [Downloaded on 27 April 2013].

M. N. **Quadri**, N.V. Kalyankar, 2010, “*Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques*”, Global Journal of Computer Science and Technology 2 Vol. 10 Issue 2 (Ver 1.0), April 2010, pp. 2 – 5, available from: <http://computerresearch.org/stpr/index.php/gjcst/article/viewArticle/128>, [Downloaded on 19 October 2010].

**Naeimeh** Delvarai, Somnuk Phon-Amnuaisuk, Mohammad Reza Beikzadeh, 2008 “*Data mining Application in Higher Learning Institutions.*” Informatics in Education-International Journal Vol7, No.1, Institute of Mathematics and Informatics Vilnius, pp 31 – 54, Available from:

[http://www.mii.lt/informatics\\_in\\_education/pdf/INFE111.pdf](http://www.mii.lt/informatics_in_education/pdf/INFE111.pdf), [Downloaded at 24April 2013]

**Nan Jiang**, Le Gruenwald, 2006 “*Research Issues in Data Stream Association Rule Mining*”, the University of Oklahoma, School of Computer Science, Norman, OK 73019, USA. SIGMOD Record, Vol. 36, No. 1. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.4587&rep=rep1&type=pdf>, [downloaded at 9 June 2011].

**Oladipupo O.O.**, Oyelade O.J., 2010, “*Knowledge Discovery from Students’ Result Repository: Association Rule Mining Approach*”, International Journal of Computer Science & Security (IJCSS), Volume (4): Issue (2), New Zealand, pp. 199 – 207, available from: <http://cscjournals.org/csc/manuscript/Journals/IJCSS/volume4/Issue2/IJCSS-249.pdf>, [Downloaded on 26 November 2011].

**Paulraj Ponniah**, 2010 “*Data Warehousing Fundamentals for IT Professionals*”. Published by John Wiley & Sons, Inc., Hoboken, New Jersey, Published simultaneously in Canada.

**Pierre Allard**, S´ebastien Ferr´e, Olivier Ridoux, 2011, “*Discovering Functional Dependencies and Association Rules by Navigating in a Lattice of OLAP Views*”, IRISA, Universit´e de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France, Pages 199 – 210, Available from: <http://cla.inf.upol.cz/papers/cla2010/paper18.pdf>, [Downloaded on 24 April 2013].

**Prachitee** B. Shekhawat, Prof. Sheetal S. Dhande, 2011, “A *Classification Technique using Associative Classification*”, International Journal of Computer Applications (0975 – 8887), Volume 20– No.5, April 2011, India, pp. 20-28, available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.3704&rep=rep1&type=pdf>, [Downloaded on 27 January 2013].

**Qasem** A. Al-Radaideh, Emad M. Al- Shawakfa, Mustafa I. Al-Najjar, 2006, “*Mining Student Data Using Decision Trees*”, the 2006 International Arab Conference on Information Technology (ACIT’2006), Jordan.

**Riadh** Ben Messaoud\_, Omar Boussaid, Sabine Loudcher Rabas´eda, 2006, “*Mining Association Rules in OLAP Cubes*” Available from: <http://eric.univ-lyon2.fr/~sabine/ITT%20dubai06.pdf> [visited at 24 April 2013].

**Riadh** Ben Messaoud, Sabine Loudcher Rabas´eda, Omar Boussaid, Omar Boussaid, Rokia Missaoui, 2006 “*Enhanced Mining of Association Rules from Data Cubes*”, DOLAP '06 Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, Pages 11 - 18 ACM New York, NY, USA ©, Available from: <http://eric.univ-lyon2.fr/~sabine/dolap06.pdf>, [Downloaded at 24April 2013].

**Sarah** N. Kohail, Alaa M. El-Hales, 2011, “*Implementation of Data mining Techniques for Meteorological Data Analysis*” (A Case Study for Gaza Strip), International Journal of Information and Communication Technology Research, Gaza. Volume 1 No. 3, Pages 96 – 100 Available from:



[http://esjournals.org/journaloftechnology/archive/vol1no3/vol1no3\\_2.pdf](http://esjournals.org/journaloftechnology/archive/vol1no3/vol1no3_2.pdf),  
[downloaded at 10 August 2011].

**Scott** Cameron, March 2009. “*Microsoft SQL Server 2008 Analysis Services - Step by Step*”, Hitachi Consulting, Microsoft Press- A Division of Microsoft Corporation- One Microsoft way- Redmond, Washington.

S. **Kotsiantis**, P. Pintelas, 2003, “*A Decision Support Prototype Tool for Predicting Student Performance in an ODL Environment*”, available from:  
<http://www.emeraldinsight.com/journals.htm?articleid=1612642&show=html>,  
[visited at 22 March 2011].

**Song** Lin & Donald E. Brown, 2002 “*Outlier-based Data Association: Combining OLAP and Data mining*”, Technical Report’, SIE 020011, Available from:  
[http://web.sys.virginia.edu/files/tech\\_papers/2002/sie-020011.pdf](http://web.sys.virginia.edu/files/tech_papers/2002/sie-020011.pdf), [visited at 24 April 2013].

**SQL** Server® Customer Advisory Team, SQL 2010, July 2010, “*Analysis Services ROLAP for SQL Server Data Warehouses*”, Microsoft Corporation. Available from:  
[http://www.google.com/#output=search&sclient=psy-ab&q=Analysis+Services+ROLAP+for+SQL+Server+Data+Warehouses&oq=Analysis+Services+ROLAP+for+SQL+Server+Data+Warehouses&gs\\_l=hp.12...490435.490435.0.494047.1.1.0.0.0.0.0.0..0.0...1c.2.14.psy-ab.4v8n4wN4eMM&pbx=1&bav=on.2.or.r\\_qf.&bvm=bv.46751780,d.d2k&fp=7d8dc92f0df78a3a&biw=1366&bih=625](http://www.google.com/#output=search&sclient=psy-ab&q=Analysis+Services+ROLAP+for+SQL+Server+Data+Warehouses&oq=Analysis+Services+ROLAP+for+SQL+Server+Data+Warehouses&gs_l=hp.12...490435.490435.0.494047.1.1.0.0.0.0.0.0..0.0...1c.2.14.psy-ab.4v8n4wN4eMM&pbx=1&bav=on.2.or.r_qf.&bvm=bv.46751780,d.d2k&fp=7d8dc92f0df78a3a&biw=1366&bih=625), [Accessed on 29 March 2013].

“*SQL Server 2000: OLAP Cubes and Queries Professional Skills Development*”, no date, Application Developers Training Company and AppDev Products Company pp. 5.1-5.6, Available from: <http://148.208.208.90/man/OLAP-AnalysisServices/AppDev%20-%20SQL%20Server%202000%20OLAP%20Cubes%20and%20Query%20Professional%20Skills%20Development.pdf>, [visited at 20 May 2013].

S S **Suresh**, Mugdha Mahale, 2011, “*Student Performance Analytics using Data Warehouse in E-Governance System*”, International Journal of Computer Applications (0975 – 8887), Volume 20– No.6, pp. 19 – 25, available from: <http://www.ijcaonline.org/volume20/number6/pxc3873284.pdf>, [Downloaded on 26 February 2013].

**Yo-Ping** Huang, Jung-Shian Jau, and Frode Erika Sandnes, 2007, “*Temporal-Spatial Association of Ocean Salinity and Temperature Variations*”, International Argo Project, Available from: [http://www.cc.ntut.edu.tw/~wwwoaa/oaa-nwww/oaa-bt/bt-data/98\\_phd/paper/41.pdf](http://www.cc.ntut.edu.tw/~wwwoaa/oaa-nwww/oaa-bt/bt-data/98_phd/paper/41.pdf), [downloaded at 16 September. 2011].

**Yuefeng** Li, Wanzhong Yang, Yue Xu, 2006, “*Multi-tier Granule Mining for Representations of Multidimensional Association Rules*” proceeding of the Six International Conference on Data mining (ICDM’06), 0-7695- 2701, IEEE, available from: <http://dl.acm.org/citation.cfm?id=1193270>, [downloaded on 22-June-11].

**Zlatko** J. Kovačić, 2010, “*Early Prediction of Student Success: Mining Students Enrolment Data*”, Proceedings of Informing Science & IT Education Conference

(InSITE) 2010 New Zealand, pp. 2 - 5 available from:

<http://repository.openpolytechnic.ac.nz/bitstream/handle/11072/646/Kovacic%20~%20Early%20prediction%20of%20student%20success.pdf?sequence=1>,

[Downloaded on 22 March 2011].