DATA WAREHOUSE DESIGN FOR EDUCATIONAL DATA WITH DATA MINING APPLICATION

College of Arts and Sciences in partial fulfillment of the requirements for the degree Master of Science (IS)
Universiti Utara Malaysia

By
NURA MUKHTAR
(Matric No: 804090)

© NURA MUKHTAR, 2010
All rights reserved
KOLEJ SASTERA DAN SAINS
(College of Arts and Sciences)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertanda tangan, memperakuan bahawa
(I, the undersigned, certify that)

NURA MUKTHAR
(804090)

calon untuk ijazah
(candidate for the degree of) MSc. (Intelligent System)

telah mengemukakan kertas projek yang bertajuk
(has presented his/her project paper of the following title)

DATA WAREHOUSE DESIGN FOR EDUCATIONAL DATA WITH DATA MINING
APLICATION

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan
dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory
knowledge of the field is covered by the project paper).

Nama Penyelii Utama
(Name of Main Supervisor):

Tandatangan
(Signature):

Tariikh
(Date):

ASSOC. PROF. FADZILAH SIRAJ

PROF. MADIYAH FADZILAH SIRAJ
Pensyarah
Bidang Sains Gunaan
Kolej Sastera & Sains
Universiti Utara Malaysia

16/05/2010
PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Master of Science in IS degree from University Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor or, in their absence by the Academic Dean College of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to University Utara Malaysia for any scholarly use which may be made of any material from thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to

Dean (Academic) College of Art and Sciences

University Utara Malaysia

06010 UUM Sintok

Kedah Darul Aman.
Abstract

Large data is stored in the data bases in Secondary Schools, which contain student's demographic and Examination Information. There is a need of these data to be integrated in one place. This study analyses the application of Data warehouse and Data mining application on data of student's previous performance on subject they have took. Where by the data from three Secondary schools were extracted, transformed and loaded into data warehouse. This study also builds student's performance multidimensional cube for each school, a Data mining model was build based on multidimensional cube designed using Microsoft Neural Network as a Data mining tool to predict the Student's performance in their SPM Exam based on their previous subjects performance. The result shows that subject BI from Sama Gagah Secondary School have the highest prediction.
ACKNOWLEDGEMENTS

First, I would like to express my appreciation to Allah, the most merciful and, the most compassionate, who has granted me the ability and willing to start and complete this study. I do pray to His Greatness to inspire and enable me to continue the work for the benefits of humanity. After that, my most profound thankfulness goes to my supervisor Assoc. Prof Fadzilah Siraj for her scientifically proven and creativity encouraging guidance and great support in this study. Also thank my evaluator Miss Thagirarani for her guidance and support.

Last, I wish to thank my parents and my lovely wife, who were always there for me by giving everything they have, my brothers and sisters for their love and support.

Thank you UUM.

Nura Mukhtar

May 13, 2010
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERMISSION TO USE</td>
<td>I</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>II</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>III</td>
</tr>
<tr>
<td>TABLE OF CONTENT</td>
<td>IV</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>VIII</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>X</td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>XI</td>
</tr>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>1.1 Background</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Problem Statement</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Research Questions</td>
<td>3</td>
</tr>
<tr>
<td>1.4 Research Objectives</td>
<td>3</td>
</tr>
<tr>
<td>1.5 Scope of The study</td>
<td>4</td>
</tr>
<tr>
<td>1.6 Significance of the Study</td>
<td>5</td>
</tr>
<tr>
<td>1.7 Organization of the Study</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER 2: LITERATURE REVIEW</td>
<td></td>
</tr>
<tr>
<td>2.1 Data Warehouse</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Multidimensional Data Model</td>
<td>8</td>
</tr>
<tr>
<td>2.2.1 Star Schema</td>
<td>9</td>
</tr>
<tr>
<td>2.2.3 Snowflakes Schema</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Data Mining Applications</td>
<td>12</td>
</tr>
<tr>
<td>2.4 Conclusion</td>
<td>15</td>
</tr>
</tbody>
</table>
CHAPTER 3: METHODOLOGY

3.1 Introduction .............................................................................. 16
3.2 System Development Research Methodology.......................... 17
3.3 Organization of Data of Interest .................................................. 19
3.4 Data Warehouse Design ............................................................. 20
   3.4.1 Multidimensional Model ..................................................... 20
   3.4.2 Conceptual Data Model ...................................................... 21
   3.4.3 Logical Data Model .......................................................... 22
   3.4.4 Physical Data Model ........................................................ 22
   3.4.5 Extract, Transformation and Loading ................................. 23
   3.4.6 Data Warehouse ............................................................... 24
   3.4.7 Data Marts .................................................................. 24
3.5 Neural Network ........................................................................ 28
   3.5.1 Model Creation ................................................................ 29
3.6 System Evaluation .................................................................... 31
3.7 Conclusion ............................................................................... 33

CHAPTER 4: RESULTS AND DISCUSSIONS

4.1 Introduction .............................................................................. 34
4.2 Dimensional Model Design ....................................................... 34
   4.2.1 Dimension Table ................................................................ 35
4.2.2 Fact Table.................................................................35
4.2.3 Star Schema............................................................36
4.3 Experiment and Results..................................................39
4.4 Student’s Performance of Sama Gagah...............................41
  4.4.1 BI Results............................................................41
  4.4.2 BM Result............................................................42
  4.4.3 MAT Result...........................................................43
  4.4.4 PENDO Result.......................................................44
  4.4.5 PI Result..............................................................44
  4.4.6 SCI Result............................................................45
  4.4.7 SEJ Result............................................................46
4.5 Student’s Performance of Kepala Batas.............................47
  4.5.1 BI Results............................................................47
  4.5.2 BM Result............................................................47
  4.5.3 MAT Result...........................................................48
  4.5.4 PENDO Result.......................................................49
  4.5.5 PI Result..............................................................50
  4.5.6 SCI Result............................................................50
  4.5.7 SEJ Result............................................................51
4.6 Student’s Performance of Kuala Ketil...............................52
  4.6.1 BI Results............................................................52
  4.6.2 BM Result............................................................53
  4.6.3 MAT Result............................................................53
CHAPTER 5: CONCLUSION

5.1 Summary ................................................. 60
5.2 Limitation ................................................ 60
5.3 Recommendation ....................................... 60

REFERENCES .................................................. 62
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Star Schema for pollution source</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Star Schema for Goal</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>A snowflakes schema for sales</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>Methodology</td>
<td>17</td>
</tr>
<tr>
<td>3.2</td>
<td>Data Warehouse Architecture</td>
<td>18</td>
</tr>
<tr>
<td>3.3</td>
<td>Sample Data</td>
<td>20</td>
</tr>
<tr>
<td>3.4</td>
<td>Logical Model Design</td>
<td>22</td>
</tr>
<tr>
<td>3.5</td>
<td>ETL Process</td>
<td>23</td>
</tr>
<tr>
<td>3.6</td>
<td>Data warehouse</td>
<td>24</td>
</tr>
<tr>
<td>3.7</td>
<td>Kepala Batas Data Mart</td>
<td>26</td>
</tr>
<tr>
<td>3.8</td>
<td>Sama Gagah Data Mart</td>
<td>27</td>
</tr>
<tr>
<td>3.9</td>
<td>Kuala Ketil Data Mart</td>
<td>27</td>
</tr>
<tr>
<td>3.10</td>
<td>Data mining wizard</td>
<td>29</td>
</tr>
<tr>
<td>3.11</td>
<td>Data mining Technique</td>
<td>30</td>
</tr>
<tr>
<td>3.12</td>
<td>Variable Selection</td>
<td>30</td>
</tr>
</tbody>
</table>
Figure 3.13: Deployment Process................................................................. 31

Figure 3.14: Deployment Process................................................................. 31

Figure 4.1: Dimension Table......................................................................... 35

Figure 4.2: Fact Table..................................................................................... 36

Figure 4.3: Data Source Wizard................................................................. 36

Figure 4.4: Data Source View Wizard......................................................... 37

Figure 4.5: Table selection ........................................................................... 37

Figure 4.6a: identifying Fact and Dimension Tables.................................... 38

Figure 4.6b: identifying Fact and Dimension Tables.................................... 38

Figure 4.7: Star Schema.................................................................................. 39

Figure 4.8: Variables Selection................................................................. 40

Figure 4.9: Data Types Selection................................................................. 41
List of Tables

Table 4.1: Classification Matrix for BI.................................................. 42
Table 4.2: Classification Matrix for BM.................................................. 42
Table 4.3: Classification Matrix for MAT.................................................. 43
Table 4.4: Classification Matrix for PENDO.............................................. 44
Table 4.5: Classification Matrix for PI..................................................... 45
Table 4.6: Classification Matrix for SCI.................................................. 45
Table 4.7: Classification Matrix for SEJ.................................................. 46
Table 4.8: Classification Matrix for BI..................................................... 47
Table 4.9: Classification Matrix for BM.................................................. 48
Table 4.10: Classification Matrix for MAT............................................... 48
Table 4.11: Classification Matrix for PENDO.......................................... 49
Table 4.12: Classification Matrix for PI................................................... 50
Table 4.13: Classification Matrix for SCI............................................... 51
Table 4.14: Classification Matrix for SEJ............................................... 51
Table 4.15: Classification Matrix for BI.................................................. 52
Table 4.16: Classification Matrix for BM............................................... 53
Table 4.17: Classification Matrix for MAT............................................... 53
Table 4.18: Classification Matrix for PENDO.......................................... 54
Table 4.19: Classification Matrix for PI.................................................. 55
Table 4.20: Classification Matrix for SCI............................................... 56
Table 4.21: Classification Matrix for SEJ............................................... 56
List of Appendices

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPENDIX A</td>
<td>Lift Chart and Mining Legend for student’s prediction in Sama Gagah</td>
<td>67</td>
</tr>
<tr>
<td>APPENDIX B</td>
<td>Lift Chart and Mining Legend for student’s prediction in Kuala Batas</td>
<td>71</td>
</tr>
<tr>
<td>APPENDIX C</td>
<td>Lift Chart and Mining Legend for student’s prediction in Kuala Ketil</td>
<td>76</td>
</tr>
</tbody>
</table>
CHAPTER ONE

INTRODUCTION

This chapter presents the background of the study, in which it explain the motivation behind the project and the domain on which the project is based on. This chapter also describes the problem statement, the objective to accomplished, significance and scope of the study. Finally it highlighted the way subsequent chapters will be organized.

1.1 Background

In the modern world, a lot of data is being received from various sources such as internet, barcode reading, remote sense and organisational database. Therefore huge amount of data is stored in a database system, which makes it difficult for organisations to properly organise and extract meaningful information. Although, traditional operational database store enormous data but only record and a real-time transactional data are managed. The operational database cannot answer a query such as what is the probable cause of profit decrease by 10% last year. This implies that operational database cannot help in making decision. This leads to the development of techniques called data warehouse and data mining which can handle this kind of situation.

As a result of increasing data in operational databases, these data needs to clean, transform and integrated into one centre called data warehouse, Online Analytical Processing (OLAP) is a good way to analyse data in a data warehouse and data mining is used to uncover useful information. Data warehouse is a subject oriented, integrated, time-varying and non-volatile database system for decision support (Vincent & Liu, 1998). The main purpose of data warehouse is to extract data from several sources i.e. external and internal sources, transform the data and make it in a
The contents of the thesis is for internal user only
REFERENCES:


