

**DATA WAREHOSUING/MINING OF FINANCIAL AND
ECONOMIC DATA: "A COMPREHENSIVE STUDY,
DESIGN, DEVELOP, IMPLEMENT AND MAINTAIN OF
THE FINANCIAL AND ECONOMIC MANAGEMENT
SYSTEM (FEMS)"**

By

TEE KIAM KHAI

Project Paper Submitted in Partial Fulfillment of the Requirement for the
Degree of Master of Information Technology

Digital Library OUM



0040283

OPEN UNIVERSITY MALAYSIA
(2007)

TABLE OF CONTENTS

| | |
|---|------|
| ACKNOWLEDGEMENTS..... | ii |
| TABLE OF CONTENTS..... | v |
| LIST OF FIGURES | viii |
| LIST OF TABLES x | |
| ABBREVIATION xi | |
| CHAPTER 1 INTRODUCTION..... | 1 |
| 1.1 Executive Summary | 1 |
| 1.2 Before the Project..... | 1 |
| 1.3 Problems Face by the Economists and Analysts and the Need of Having a Centralized Database System..... | 2 |
| 1.4 Project Objectives | 3 |
| 1.5 Financial and Economic Management System (FEMS) – An Business Intelligence (BI) / Online Analytical Processing (OLAP) Project..... | 4 |
| CHAPTER 2 LITERATURE REVIEW..... | 7 |
| 2.1 Overview..... | 7 |
| 2.2 Data Rich but Information Poor..... | 8 |
| 2.2.1 <i>Data</i> | 8 |
| 2.2.2 <i>Information</i> | 9 |
| 2.2.3 <i>Knowledge</i> | 9 |
| 2.3 Data Warehouses..... | 10 |
| 2.3.1 <i>Benefits of Warehousing for Data Mining</i> | 10 |
| 2.4 What is Data Mining? | 11 |
| 2.5 What can Data Mining do for the Economists/Analysts? | 11 |
| 2.5.1 <i>Automated Prediction of Trends and Behaviours</i> | 13 |
| 2.5.2 <i>Automated Discovery of Previously Unknown Patterns</i> | 14 |
| 2.5.3 <i>Steps Used for a Knowledge Discovery Process</i> | 14 |
| 2.6 The Evolution of Data Mining..... | 14 |
| 2.7 How Data Mining Works..... | 15 |
| 2.8 Data Mining Tasks | 17 |
| 2.9 Data Mining Techniques | 18 |
| 2.9.2 <i>Supervised Induction</i> | 18 |
| 2.9.3 <i>Association Discovery</i> | 19 |
| 2.9.4 <i>Sequence Discovery</i> | 19 |
| 2.10 Data Mining Tools | 19 |
| 2.11 Samples of the Case Studies | 20 |
| 2.11.1 <i>1st Cast Study - Predicting Stock Market Returns Using Neural Network</i> | 20 |
| 2.11.2 <i>2nd Case Study - Profiling the Tourism Market Using K-Means</i> <i>Clustering Analysis</i> | 23 |
| 2.12 Conclusion | 26 |
| CHAPTER 3 METHODOLOGY..... | 27 |
| 3.1 Selection and Evaluation of BI's Tools | 27 |
| 3.2 Functional Specification Documentation (FSD)..... | 29 |
| 3.2.1 <i>System Architecture and Operating Environment</i> | 29 |
| 3.2.2 <i>Cube Structure and Design</i> | 32 |

| | | |
|-----------|---|-----|
| 3.2.3 | <i>Extraction and Integration</i> | 34 |
| 3.3 | User Requirement Specifications (URS) | 37 |
| 3.3.1 | <i>Objectives of the User Requirement Specifications</i> | 37 |
| 3.3.2 | <i>Methodology Used in Determining the User Requirement Specifications</i> | 37 |
| 3.3.3 | <i>General Requirements and Project Overview</i> | 38 |
| 3.4 | System Design and Development | 40 |
| 3.4.1 | <i>Cube Design and Features</i> | 40 |
| 3.4.2 | <i>System Development - Process of Developing a Cube</i> | 42 |
| 3.4.3 | <i>Data Extraction</i> | 53 |
| 3.4.4 | <i>Security Administration and Controls/loggings</i> | 55 |
| 3.4.5 | <i>Economic Dashboard</i> | 56 |
| CHAPTER 4 | DATA ANALYSIS AND RESULTS | 57 |
| 4.1 | Data Analysis I - Virtual Cube and 255-Character Limitation | 57 |
| 4.2 | Data Analysis II - Single Sign-on | 60 |
| 4.3 | Data Analysis III - Performance Issue | 66 |
| 4.3.1 | <i>Server Hangs</i> | 66 |
| 4.3.2 | <i>Slow performance</i> | 66 |
| 4.4 | Data Analysis IV – Several Programming Bugs Detected and new request by users | 68 |
| 4.4.1 | <i>To Address Issues (a) and (b)</i> | 68 |
| 4.4.2 | <i>To address Issue “c” - The data structure designed for the Commodities’ dimension is not correct</i> | 72 |
| 4.4.3 | <i>To Address Issue “d” – The Historical or Old data in the cube must be deleted manually first before the revised data could be uploaded</i> | 75 |
| 4.4.4 | <i>To Address Issue “e” - New data structure was requested by the users to group the elements in the Commodities’ dimension</i> | 77 |
| 4.5 | Data Analysis V – Data Inconsistency Detected from the ANM’s Cube | 80 |
| 4.5.1 | <i>Unable to Use the “Suppress Zero” Function</i> | 81 |
| 4.5.2 | <i>Data Becomes Cumulative if the Process for the Same Dataset was Executed the Second Time</i> | 85 |
| 4.6 | Data Analysis 6 – Incompatibly between MS Excel and TM1 | 87 |
| 4.7 | Data Analysis 7 – Unable to Insert Data into the Cube via MS Excel Worksheet | 90 |
| CHAPTER 5 | QUALITY ASSURANCE EVALUATION ON THE IMPLEMENTATION | 93 |
| 5.1 | Term of Reference | 93 |
| 5.1.1 | <i>Data and Structures</i> | 93 |
| 5.1.2 | <i>Data Import Processes</i> | 93 |
| 5.1.3 | <i>Rules Implementation</i> | 93 |
| 5.1.4 | <i>Hardware</i> | 94 |
| 5.2 | Queries and Suggestions by the Independent Consultant | 95 |
| 5.2.1 | <i>Data and Structures</i> | 95 |
| 5.2.2 | <i>Data Import Processes</i> | 96 |
| 5.2.3 | <i>Rules Implementation</i> | 97 |
| 5.2.4 | <i>Hardware</i> | 100 |
| 5.3 | Quality Assurance - Summary | 101 |

| | | |
|------------------------|--|------------|
| CHAPTER 6 | RISK MANAGEMENT..... | 102 |
| 6.1 | Scope Risks | 102 |
| 6.2 | Schedule Risks | 102 |
| 6.3 | During organizational planning, resource planning, staff acquisition, cost estimating, procurement planning, and solicitation planning and document resource risks, including: | 102 |
| 6.4 | Backup Risks..... | 103 |
| 6.5 | Document Risks | 103 |
| CHAPTER 7 | CONCLUSION AND RECOMMENDATIONS | 104 |
| 7.1 | Current Scenario | 104 |
| 7.2 | Processes..... | 105 |
| 7.3 | Outcome..... | 106 |
| 7.4 | Findings and Recommendations | 107 |
| 7.5 | Moving Forward | 108 |
| APPENDIX I | Data Mining Tasks | 109 |
| APPENDIX II | Data Mining Tools..... | 111 |
| APPENDIX III | Estimated Size of all the Cubes at the Closing of FEMS Project and for the Next 2 Years..... | 113 |
| APPENDIX IV | Sample Metadata from the International Monetary Fund | 115 |
| APPENDIX V | Data Source: Properties and Extraction Details | 116 |
| APPENDIX VI | The Extraction Processes of the Respective Data Sources | 117 |
| APPENDIX VII | Tables and Charts to be included in the Economic Dashboard..... | 122 |
| APPENDIX VIII | Samples of The Rules Used for Fiss3 Cube Before and After Fine-Tuning..... | 123 |
| I) | Poor Performance is caused by Poorly Designed Rules | 123 |
| II) | Revised Rules Used which improved the performance significantly | 124 |
| III) | Conclusion | 124 |
| APPENDIX IX | Sample Of Metadata Rule Used For Grouping First 3-Character or 5-Character | 125 |
| I) | Coding to group first 3-character:..... | 125 |
| II) | Coding to group first 5-character | 125 |
| APPENDIX X | Sample of Instruction at The Prolog to Delete Old or Historical Data Before New Data is Uploaded for Monthly Data | 126 |
| REFERENCES..... | | 128 |

LIST OF FIGURES

| | |
|---|----|
| Figure 1: The process of knowledge discovery in databases (KDD)..... | 6 |
| Figure 2: Data Mining – Core of Knowledge Discovery Process in Database..... | 9 |
| Figure 3: Data Mining and Business Processes | 16 |
| Figure 4: Integrated Data Mining Architecture..... | 17 |
| Figure 5: Multiple TM1 Instances on a 32-bit TM1 Server..... | 30 |
| Figure 6: Proposed 64-bit TM1 Architecture on 64-bit Windows Server | 31 |
| Figure 7: Economic Dashboard..... | 33 |
| Figure 8: Overview of the FEMS Project | 38 |
| Figure 9: Create a New Dimension..... | 43 |
| Figure 10: Dimension Editor Dialogue Box | 43 |
| Figure 11: Dimension Element Insert Dialogue Box..... | 43 |
| Figure 12: Insert Element Name | 44 |
| Figure 13: Consolidated Element Added | 44 |
| Figure 14: Steps in Saving a Dimension | 44 |
| Figure 15: Insert the Dimension name | 45 |
| Figure 16: Update and Adding Additional Elements..... | 45 |
| Figure 17: Listing of Elements..... | 45 |
| Figure 18: Arrangement of Elements in the Dimension | 46 |
| Figure 19: Creation of a New Cube | 47 |
| Figure 20: Creating Cube Wizard | 47 |
| Figure 21: Create New Process | 49 |
| Figure 22: Turbo Integrator Wizard..... | 49 |
| Figure 23: Browsing the Source File | 49 |
| Figure 24: TM1 Warning Dialogue Box..... | 50 |
| Figure 25: Preview the Data..... | 50 |
| Figure 26: Selection of Variable Tab..... | 50 |
| Figure 27: Select Prolog Tab..... | 51 |
| Figure 28: Select Metadata Tab | 51 |
| Figure 29: Select the Data Tab..... | 51 |
| Figure 30: To Save the Process..... | 52 |
| Figure 31: Naming of the New Process | 52 |
| Figure 32: Snapshot of the Economic Dashboard..... | 56 |
| Figure 33: Integrated Logon..... | 60 |
| Figure 34: LDAP Login | 60 |
| Figure 35: TM1 logon | 61 |
| Figure 36: Tool Options..... | 61 |
| Figure 37: Search Result of ETLLDAP | 62 |
| Figure 38: TM1 Mapping..... | 63 |
| Figure 39: Session log..... | 63 |
| Figure 40: Cube Viewer – Client Properties..... | 64 |
| Figure 41: Clients/Groups on sdata..... | 64 |
| Figure 42: Unused elements appear in Commodities' Dimension | 68 |
| Figure 43: The same coding is used for Processing “Monthly” and “Revised” Data..... | 69 |
| Figure 44: Reading of Monthly Data | 70 |
| Figure 45: Reading of Revised Data | 70 |

| | |
|--|-----|
| Figure 46: The same metadata for “Monthly data” Process and the new metadata for “Revised data” Process | 71 |
| Figure 47: New Commodities’ Dimension Layout..... | 71 |
| Figure 48: The new elements were not uploaded correctly into their respective groups.... | 72 |
| Figure 49: The old code for Monthly Data | 73 |
| Figure 50: The new code for Monthly Data..... | 73 |
| Figure 51: Only group with 1 character e.g. HS-0, HS-1 etc resides directly under their parent “Total HS Commodities” | 74 |
| Figure 52: Parameters to be filled by the administrator..... | 76 |
| Figure 53: The code for clearing the cumulative SITC data at the prolog of the process .. | 76 |
| Figure 54: The 3-initial grouping of 9-character data source | 77 |
| Figure 55: The parent for the element was grouped by the first 1-, 2-, and 3-character..... | 78 |
| Figure 56: The parent for the elements was grouped by the first 1-, 2-, 3-, 4- and 5-character | 78 |
| Figure 57: The grouping for each element now was up to first 1-, 2-, 3-, 4-, and 5-character | 79 |
| Figure 58: Development of ANM’s Cube..... | 80 |
| Figure 59: When “suppress zero” icon was selected, all the data in the cube would not be displayed with the message “No values Available”..... | 81 |
| Figure 60: All the data inclusive the “Zero Values” would be displayed if the “Suppress Zero” icon was not selected | 82 |
| Figure 61: The old rule without feeder | 82 |
| Figure 62: The new rule | 83 |
| Figure 63: The feeders for the Perbelanjaan Mengurus’ s cube (oe-obj)..... | 83 |
| Figure 64: Only Data with Values would be displayed when “Suppress Zero” icon was selected..... | 84 |
| Figure 65: Sample of the wrong coding used..... | 85 |
| Figure 66: The amended code | 86 |
| Figure 67: To Set Calculation to Manual instead of Automatic Mode | 87 |
| Figure 68: Calculation was set to “Automatic” by default | 88 |
| Figure 69: Select the calculation to Manual..... | 88 |
| Figure 70: View of TM1 Cube..... | 90 |
| Figure 71: MS Excel that slides from TM1 Cube..... | 91 |
| Figure 72: Error Message while trying to add or modify any cell with formula | 91 |
| Figure 73: CPU Usage | 106 |
| Figure 74: The Cube Default View..... | 107 |
| Figure 75: Neural Networks..... | 111 |
| Figure 76: Decision Trees | 112 |
| Figure 77: DOSM, EPF, MIDA, ANM, Datastream & BIS Extraction Processes | 117 |
| Figure 78: IMF Cube Extraction Process..... | 117 |
| Figure 79: INTACTS, TERAS, CIS and FIRST Cubes Extraction Processes | 118 |
| Figure 80: PANTAS & TERAS (Historical Data) Cubes Extraction Processes..... | 118 |
| Figure 81: MEX Cube Extraction Process | 119 |
| Figure 82: Dbank Cube Extraction Process | 119 |
| Figure 83: CFIIS & RENTAS Cube Extraction Process | 120 |
| Figure 84: FISS Cube Extraction Process | 120 |
| Figure 85: ITEPS Cube Extraction Process | 121 |

CHAPTER 1 INTRODUCTION

1.1 Executive Summary

Every day, enormous amounts of information are generated from all sectors, whether it is economic, business, exchange rate, the scientific community, the World Wide Web (WWW), or more of many readily available off-line and online data sources. From all of this, which represents a sizable repository of data and information, it is possible to generate worthwhile and usable knowledge. As a result, the field of Data Mining (DM) has grown in leaps and has shown great potential for the future. The purpose of this project is to clean, centralize and warehouse all the raw data for the analysts or economists. The economists or analysts will identify many of the critical and future trends in the field of DM that helps the Bank to determine or predict the behaviors of the data. This assists in discovering information within the data that queries and reports can't effectively reveal. Raw data by itself however does not provide much information. Data mining, or knowledge discovery, is the computer assisted of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing the Economics Department of Bank Negara Malaysia (BNM) to make proactive knowledge-driven decisions. They scour databases for hidden patterns, finding predictive information that many economists /analysts may miss because it lies outside their expectations.

1.2 Before the Project

Data, data everywhere. Since the establishment of the Bank in 1959, a lot of data were collected by Economics Department (JEK) and stored in various format such as tables in

word documents, spreadsheets, Microsoft access databases or even text files. The data is stored either in the diskettes, hard drives or hard printed copies. Recent advances in data collection and data storage technology have enabled the Bank to accumulate large quantities of operational, economic and financial data from their daily activities.

Since the amount of raw data collected in the department is exploding, raw data by itself, however, does not provide much information. The dilemma faces by the department is data rich but information poor. The data collected by JEK is valuable for various purposes. For example, the data collected from the usage of credit cards issued by the financial institutions may be used to understand the pattern of purchasing behaviour and this invaluable information could be challenged to the right industries for further action. Traditional data analysis techniques, which are often based on statistical inference and hypothesis testing, encounter tremendous difficulties in handling large quantities of data. Motivated by the need to overcome the problems of analyzing large-scale data, analysts and economists have turned to data mining for assistance.

1.3 Problems Face by the Economists and Analysts and the Need of Having a Centralized Database System

The Department reveals that the present system of data collection and compilation has the following main weaknesses:

- Significant time spent in re keying and checking of data by different sections of the same department;
- Duplication of work in compiling and updating of data stored by different sections within the same department;
- No sharing of data compiled, except through exchange of diskettes or hard copy reports. Online access of data is impossible;

- Retrieval of information for detailed analysis is cumbersome, as there is no proper indexing of data files; and
- The raw or source data used by different sections may be different as the data is stored in different storage devices or with the owners/authors.

In order to resolve these problems, the Economics Department has recommended for the implementation of a financial and economic management system (FEMS), as part of the measures to reduce time spent on data compilation, to improve information accessibility and availability in order to improve the ability of the department to provide better quality reports, analysis, forecasting and recommending in a shorter time. The problems discussed can be summarized as follows:

- Data Explosion Problem
 - Automated data collection tools and mature database technology lead to tremendous amounts of data accumulated and/or to be analyzed in databases, data warehouses, and other information repositories.
- The Economics Department is drowning in data, but starving for knowledge!
- To resolve the problems, the department has suggested to implement data warehousing and data mining in stages:
 - Data warehousing and on-line analytical processing (OLAP) to filter and clean up existing databases in Phase I; and
 - Mine interesting knowledge (rules, regularities and patterns) from the existing databases in Phase II.

1.4 Project Objectives

The FEMS project is started in view of enhancing the capability of JEK and JASM's staff in performing their analytical activities. At present their capability is partly hindered due to the limitation of the current Time Series Management Tool, DBank, which acts mainly

as a data storage tool for the economic and financial data and unable to perform multi-dimensional analysis. The main objectives of the project are as follows:

- To facilitate multi-dimensional analysis and interactive reporting for all levels of users;
- To facilitate consolidation of data from several sources;
- To improve data analysis, estimation and forecasting by having the ability of performing scenario analysis;
- To segregate access to data based on security classification level;
- To keep track of data changes up to the lowest level;
- To generate pre-defined interactive and ad-hoc tables and charts for Annual Report, Quarterly Bulletin, Board Papers, Board Briefing, Travel Books and other management reports from the centralised database and timely submission of tables or reports to the management;
- To minimize the risk of losing important data since the data is centralised as opposed to current arrangement where spreadsheets are kept in individual PCs or with the authors;
- Data extracted from the centralised database will be more reliable and consistent;
- To reduce the searching and waiting time for the authors to generate the tables or reports; and
- To have the ability of searching any time series across the entire databases.

1.5 Financial and Economic Management System (FEMS) – An Business Intelligence (BI) / Online Analytical Processing (OLAP) Project

The FEMS – BI/OLAP is a step forward in using technology to facilitate a more efficient, effective and speedier way of generating information required for the management and publication. It will interface with the existing information system and non-system databases to produce database-linked outputs. An efficient and effective way of generating and disseminating of data or information is essential for the Economics Department; the

new system should be able to meet the needs and requirements for the entire department and the users. Data Warehousing follows by Data mining are chosen by the Economics Department for the following reasons:

- i. Clean and store all the raw or source data in a common place;
- ii. Filter, visualize and interpret the pattern of the data; and
- iii. Discover the useful knowledge from large data repositories that can be shared by all the economists and analysts.

For reason (i), it would be the responsibility of database administrator (DBA) while reasons (ii) and (iii) would be the responsibilities of the economists and analysts. The notion of usefulness has different meanings to different analysts or economists. Data mining or knowledge discovery is the computer-assisted process of digging through and analyzing enormous sets of historical economic and financial data, and then extracting the meaning of the data. This data mining tools predict behaviours and future trends, allowing the economists and analysts to be more proactive and make knowledge-driven decisions, as it can reveal previously unknown information about the trends or changes over time.

Data mining is often considered to be an integral part of knowledge discovery in databases (KDD). The process of KDD of the raw data into useful knowledge can be shown in Figure 1. It consists of a series of transformations, including data preprocessing and post processing. Data preprocessing transforms the raw data into a format suitable for subsequent analysis. It also helps to identify subsets of the data that are relevant for a particular data-mining task. As the raw data may be stored in different formats and in different databases, a large amount of time may be spent on data preprocessing which includes data cleaning, metadata repository, key families, naming convention, among others.

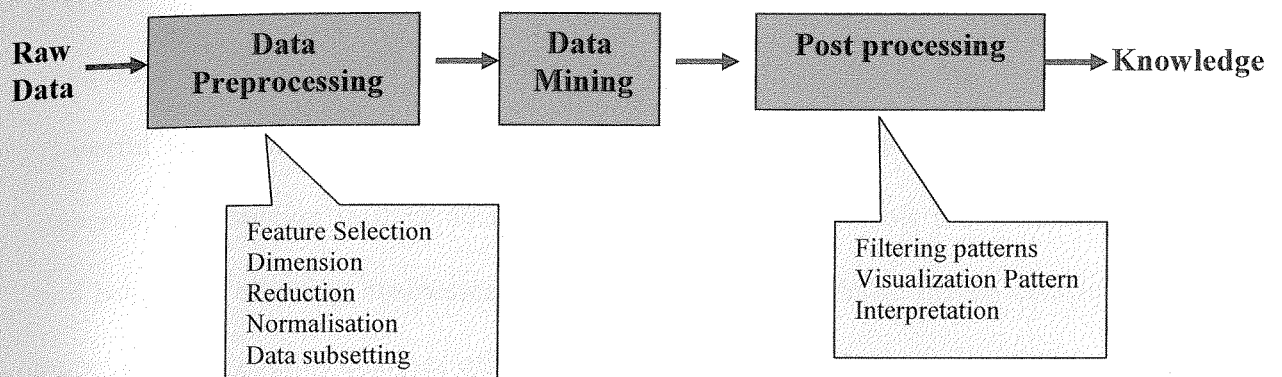


Figure 1: The process of knowledge discovery in databases (KDD)

Post processing encompasses all the operations that are performed to make the data mining results more accessible and easier to interpret. For example, spurious results can be filtered according to a variety of measures. Visualization techniques may also be employed to help analysts explore and understand the data mining results.

REFERENCES

1. Data Warehousing Concepts and Strategies *Stefan M. Neikes Sumit Sircar Bijoy Bordoloi*
2. Data mining by Doug Alexander at www.eco.utexas.edu/~norman/BUS.FOR/course.mat/Alex
3. H. Edelstein, "Technology How to: Mining Data Warehouses", Information Week (January 8, 1996) pg 48-47
4. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/data_mining.htm
5. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (Paperback) by Michael J. A. Berry, Gordon S. Linoff
6. Data Mining: Opportunities and Challenges, Wang, John (Editor). Hershey, PA, USA: Idea Group Inc., 2003
7. Data Mining in Times Series Databases, Mark Last, Abraham Kandel & Horst Bunke, World Scientific Publishing Company.
8. ERP & Data Warehousing in Organizations: Issues and Challenges, Gerald Grant Carleton University, Canada, Hershey, PA, USA: Idea Group Inc
9. The Research Project: How to Write It. Berry, Ralph. Florence, KY, USA: Routledge
10. How to write a thesis, Rowena Murray, Open University Press, McGraw Hill Education.
11. Discovering Knowledge in Data: An Introduction to Data Mining, Hoboken, NJ, USA: John Wiley & Sons, Incorporated.
12. <http://www.applic.com>
13. Data Sources: IMF's CDs which contain the BOP, GTS, DOT and IFS data
14. Guidelines provided by Financial Thomson in connecting to Datastream Server
15. Data Mining – Introductory and Advanced Topics by Margaret H Dunham, Pearson Education Inc