

# **Data Mining in an Electronic Poll (e-Poll) System**

by

Mariani Bt Md Razale @ T. Mariani

Dissertation submitted in partial fulfillment of  
the requirements for the  
Bachelor of Technology (Hons)  
Information Technology

JUNE 2004

t  
QA  
769  
.D343  
M333  
2004

1. Data mining.

Universiti Teknologi PETRONAS  
Bandar Seri Iskandar  
31750 Tronoh  
Perak Darul Ridzuan

# CERTIFICATION OF APPROVAL

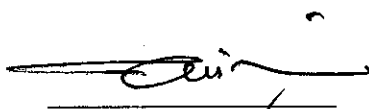
## Data Mining in an Electronic Poll (e-Poll) System

by

Mariani Bt Md Razale @ T. Mariani

A project dissertation submitted to the  
Information Technology Programme  
Universiti Teknologi PETRONAS  
in partial fulfillment of the requirements for the  
BACHELOR OF TECHNOLOGY (Hons)  
INFORMATION TECHNOLOGY

Approved by,



(Puan Aliza Sarlan)

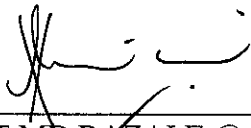
Aliza Binti Sarian  
Lecturer  
Information Technology & Information System Programme  
Universiti Teknologi PETRONAS  
31760 Tronoh  
Perak Darul Ridzuan, MALAYSIA

UNIVERSITI TEKNOLOGI PETRONAS  
TRONOH, PERAK

APRIL 2004

## CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgments, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.



---

MARIANI BT MD RAZALE @ T. MARIANI

## **ABSTRACT**

This paper introduces the Final Year Project entitled Data Mining in an Electronic Poll (e-Poll) System, with the problem being how the use of Data Mining in an Online Poll System can help managers to obtain statistics of customer feedback or opinions to help achieve their company objectives. The project's objectives are to conduct study on the use of Data Mining for an e-Poll system and how it affects the decision of system owners, to design a Data Mart for the Poll that will allow effortless management of the system's information, and lastly, to produce a working prototype of the system that allows capturing and retrieval of poll participation information from store members. The procedures identified to accomplish these tasks are by going through literature sources to better understand the proper tools and technique in the development of the system, by observing current online polling systems and online stores, and by creating a functional prototype of the e-Poll system to capture poll participation information in order for analysis to be performed onto them using Data Mining tools. Through the development of this system, the author finds that Data Mining offers managers to transform their raw data into useful data, save time in uncovering data trends and analyze vast amounts of data at a time. Proper design of the Data Mart using Dimensional Modeling or Star Schema allows optimal query performance and greater understandability without any loss of information. The e-Poll design allows correlation of the poll participant's feedback to their personal information, allowing proper analysis and a gateway for managers to gather potentially important information from all or a sample of their customers, regardless of geographical boundaries and/or time.

Keywords: Data mining, e-Poll system, online stores, Data Mart, segmentation, classification, analysis.

## ACKNOWLEDGEMENTS

*Bismillahirrahmanirrahim.* First and foremost, I thank The-Merciful Allah SWT for the guidance and wisdom to come this far in my academic life. With His Mercy alone was I able to successfully become a UTP Final Year student and thus now be able to present the findings of my Final Year Project (FYP) in this document.

I'd like to thank my FYP Supervisor, Puan Aliza Sarlan of the IT/IS Programme, for giving me the opportunity to work on the self-proposed title with only a few suggestions to make it FYP worthy. I thank her for her supervision, words of advice and encouragement throughout the semester, and for outlining my project's shortcomings in between it all. She has been truly supportive when it comes to entertaining my endless questions or concerns about the system or functions I had to develop. The value of your guidance, Puan Aliza, has been truly beneficial and priceless.

I'd like to extend my gratitude to the examiners who have been given the responsibility to judge my project's worth during the Internal and Final Presentations. Thank you to Mr. Ahmad Izudin Zainal Abidin of the IT/IS Programme and Mr. Askhoff Abd. Satar from the industry for giving me a chance to explain more about my findings during the first Q & A Session of the morning on 1<sup>st</sup> June 2004. And thanks to Mr. Khairul Shafee Kalid of the IT/IS Programme for his comments and suggestions during the pre-EDX IT/IS Exhibition on 21<sup>st</sup> April 2004. I appreciate the feedback and advice you've all shared, as alongside that of my Supervisor's, they have meant a lot for the writing of this report.

I also thank Mr. Justin Dinesh Devaraj for his suggestions on the conduct of my final presentation – including content and style of delivery. Thanks to my dearest FYP colleagues for their support, guidance and friendship – Syazwani Mohd. Yusop, Nashriq Ramly, Melissa Zainal Abidin, Wan Nur Asyikin Mohd. Yusof, Nurshazawati Abdul Hakim, Fadhlina Shoib, Amalia M. Kamarudin and others.

Lastly, without the financial and moral support from my husband Zulfadli Md Ariffin and my parents Md Razale Md Nor and Sharifah Maimunah Syed Mohammed, I would not be here today. Thank you and love goes to them, for the trust, encouragement and *doa* they've all given me during my brief stay here at UTP.

May Allah SWT bless you all. Amin.

## TABLE OF CONTENT

<b>CERTIFICATION OF APPROVAL</b> .....	i
<b>CERTIFICATION OF ORIGINALITY</b> .....	ii
<b>ABSTRACT</b> .....	iii
<b>ACKNOWLEDGEMENTS</b> .....	iv
<b>LIST OF FIGURES</b> .....	viii
<b>LIST OF TABLES</b> .....	ix
<b>ABBREVIATIONS</b> .....	x
<b>CHAPTER 1: INTRODUCTION</b> .....	<b>1</b>
1.1 Background of Study.....	1
1.2 Problem Statement.....	3
1.3 Objectives and Scope of Study.....	4
<b>CHAPTER 2: LITERATURE REVIEW OR THEORY</b> .....	<b>5</b>
2.1 Data Mining.....	5
2.2 Poll.....	7
2.3 Dimensional Modeling.....	10
<b>CHAPTER 3: METHODOLOGY</b> .....	<b>12</b>
3.1 Procedure Identification.....	12
3.2 Tools Required.....	15
<b>CHAPTER 4: RESULTS AND DISCUSSION</b> .....	<b>17</b>
4.1 Findings.....	17
4.2 Discussion.....	24

<b>CHAPTER 5:</b>	<b>CONCLUSION</b> .....	<b>40</b>
	5.1 Relevancy to the Objectives.....	40
	5.2 Suggested Future Work for Expansion and Continuation.....	41
<b>REFERENCES</b> .....		<b>43</b>
<b>APPENDICES</b> .....		<b>45</b>
	<b>Appendix A: Data Warehouse and Data Mining Terms</b> .....	<b>46</b>
	<b>Appendix B: Complete Listing of Database Schema (Dimensional Model)</b> .....	<b>49</b>
	<b>Appendix C: Star Schema Dimensions Values</b> .....	<b>53</b>



## LIST OF FIGURES

- Figure 2.1 A Dimensional Model Example
- Figure 3.1 The Evolutionary Prototyping Model
- Figure 4.1 Entity Relationship Diagram for the e-Poll System
- Figure 4.2 e-Poll Process Flow for Store Members
- Figure 4.3 e-Poll Additional Process Flow for Managers
- Figure 4.4 Star Schema for the e-Poll System
- Figure 4.5 New Store Member Registration
- Figure 4.6 Returning Store Member Login
- Figure 4.7 Poll for Store Members
- Figure 4.8 Updated Record in `fact_MemberPoll` Table
- Figure 4.9 Pop-up Window of Poll Results
- Figure 4.10 Adding Comments
- Figure 4.11 Confirmation of New Comment
- Figure 4.12 Graphical Representation of `fact_MemberPoll` (all records)
- Figure 4.13 Graphical Representation of `fact_MemberPoll` (married participants)
- Figure 4.14 Graphical Representation of `fact_MemberPoll` (Central region participants)
- Figure 4.15 Graphical Representation of `fact_MemberPoll` (Utara region participants)
- Figure 4.16 Results of *Figure 4.15* Analysis (Exported to Web Browser)

## **LIST OF TABLES**

- Table 4.1 e-Poll Database Schema
- Table 4.2 Online Store Personal Info Database Schema
- Table 4.3 Store Member-Poll Database Schema

## ABBREVIATIONS

CRM	Customer Relationship Management
DM	Dimensional Modeling
ERD	Entity Relationship Diagram
e-Poll	Electronic Poll
OLAP	On-Line Analytic Processing
SQL	Structured Query Language
WWW	World Wide Web

# CHAPTER 1

## INTRODUCTION

### 1. OVERVIEW OF PROBLEM

Perusing in an online store like amazon.com can be exhilarating to some of us, especially if their product offering catches our eye. Managers of these businesses spend quite an effort to ensure that quality products and service are promoted and delivered to their customers in a timely manner.

However, aside from making their business sustain the market expectations, there underlies a great need for the business owners to secure an advantage or ‘edge’ over their competition. This is hardly a trivial matter. These companies’ livelihood depends on their ability to retain existing customers in addition to securing potential ones. There is a need for the owners to ascertain that customer satisfaction is top-notch and that their online store offers bargains and features that are attractive and hard to resist. In a way, the owners must periodically reassess their business strategy to stay one step ahead of their competitors. And to achieve this is no easy feat.

In this document, the author introduces the use of a technique to help online business owners accomplish their company objectives. The technique utilizes data obtained from store customers themselves – data that upon thorough analysis will help owners plan their next strategic business move.

#### 1.1 Background of Study

Online stores like amazon.com and giant bookstore barnesandnoble.com are reaping serious business from visitors each and every day from all four corners of the world. Their product catalogue and tempting bargains are sometimes even personalized to

be catered to a specific customer profile upon the said customer's login state. In accomplishing this, statistical tools have been used to analyze the shopping habits of the customer and help system owners plan their marketing move.

However, owners of these online stores need to measure not only how much profit can be gained by catering to the styles of a specific customer group. But they must also determine whether their business on the whole is tempting enough to the group of visitors whom are just glancing through their site. Issues such as the website's content organization, or product offerings, to a product visitors most like to see featured are amongst important information that store owners should know to help them strategically plan or develop their business or marketing strategy.

In order to accomplish this, we must construct a platform for the visitors to provide feedback to the system owners. This could be accomplished in a number of ways, one of which is to have them answer simple questions in a Poll, updated monthly or periodically. This proposed project expands on the use of an electronic-Poll (henceforth e-Poll) system in an Online Store. Using specifically constructed poll questions, a system owner can gather useful data from their store customers to be used as guidelines for planning of new marketing strategies.

Data in its raw form is considered quite useless unless proper analysis of them is conducted. This can be accomplished by classifying the acquired data. A system owner may not really know what kind of data is useful for analysis – until they uncover a data pattern generated by a computer program through a simple query made by their programmer.

In order to have a useful analysis though, raw data acquired must be segmented or classified accordingly. Traditionally, it has been done either through tabulation of selected data in a spreadsheet form or others. But times have changed. The technique to have an accurate analysis of the collected data in the business world today is via use of Data Mining (henceforth DM).

With DM software or systems implementing DM in its functions, a system owner has the freedom and bonus of generating data analysis to help plan a strategic move.

Such a system will provide managers and owners to uncover links in the data gathered and allow them to obtain conclusions in the form of statistics, groupings or patterns.

The task of the project is to produce a good example of an e-Poll system in an Online Store that will use *Data Mining* to allow the owners to obtain statistics from the poll to formulate new ideas and plan their next strategic business move. To realize this, thorough research must be done to identify the proper techniques and procedures for implementing Data Mining onto an e-Poll system.

## **1.2 Problem Statement**

In this research project, the problem statement the author has identified is:

How the use of Data Mining in an Online Poll System can help managers to obtain statistics of customer feedback or opinions to help achieve their company objectives.

### **1.2.1 Significance of the Project**

The business world nowadays rely on the data they acquire to help them formulate future actions. Data is such a valuable asset, such that it could bring down the establishment due to data mismanagement or bring it to unprecedented heights of glory with its useful and valuable data analyses.

The latter is what companies are after. That is, to use and analyze the gathered data in such a way that it will help them strategically plan or develop their next business move. To do this, companies must not rely on traditional methods to acquire the statistics or patterns of the data, for example through tabulation of data in spreadsheets and graphs generated via Microsoft Excel.

Instead, companies should create platforms to automatically allow uncovering of patterns and trends based on classification of the data obtained. Such a system has been widely used in the industry – such as retailing, inventory control, and product

manufacturing. It has helped countless businesses improve its operations and with such, made it more profitable.

### **1.3 Objectives and Scope of Study**

#### **1.3.1 Objectives**

Among the objectives of this project are as follows:

- a. To study the use of Data Mining techniques for an e-Poll system, and how it affects decision of system owners;
- b. To design the Data Mart for the e-Poll system that will allow effortless management of the system's information;
- c. To develop a working prototype of the e-Poll system in an Online Store to capture poll participation of store members.

#### **1.3.2 Feasibility of the Project within the Scope and Time Frame**

The scope of study covers the following:

- a. Design of a well-structured database management system (DBMS) to store relevant information and to simplify and speed queries.
- b. Development of a dynamic Web-based Poll system to capture poll participation information.
- c. Use of Data Mining tools to provide system owners with accurate, useful and up-to-date statistics and classifications of data for their analysis needs.

The scope of the project has been tailored to accommodate the requirements of the Final Year Project. It is also designed in such a way that it will fulfil the objectives of the project within the time frame (or semester) given.

## CHAPTER 2

### LITERATURE REVIEW OR THEORY

#### 2. WHAT THE 'SOURCES' SAY

The main objective of this chapter is to discuss a few relevant and useful sources that itemizes and discovers the importance of Data Mining and Surveys or Polls in the business world. After completing this chapter, the author believes that a basic understanding of Data Mining and Polls will be grasped and will allow the readers to have little trouble to follow through with the rest of the discussion in the following chapters.

##### 2.1 Data Mining

###### 2.1.1 Theory

Data Mining (henceforth DM) “*consists of finding interesting trends or patterns in large datasets, in order to guide decisions about future activities*” (Ramakrishnan, R., Gehrke, J., 2000, p.707). It is understood that a connection between the DM system and a database is required for a user input or query to be processed. A finely structured data warehouse is normally required for DM to work properly.

DM has allowed businesses to flourish as it has allowed insights to system owners that upon careful investigation would provide valuable analysis. “*Data mining enables corporate managers to uncover trends, patterns, correlations, and relationships (possibly unexpected) of data by questioning the data warehouse*” (Siegel, J., Shim, J., 2003, p.111). It is largely agreed that upon implementation of a DM system, companies will have the power to potentially gather significant analysis



of their data collection and thus be able to use these information in a means that will bring unto them business success.

*“Successful data mining requires data feedback”* (Siegel, J., Shim, J., 2003, p.112). A DM software will allow automatic search for information in databases, which will then arrive at a certain conclusion based on its findings. For example, from a simple query, a system can identify the geographical area where a particular item or product sells better. From this, we see that the search item for the query would be ‘geographical area’ for ‘product A’, with the feedback being ‘location X’. This identifies a successful DM process.

DM may take the form of Knowledge Discovery process or KDD process in short. It is a *“bottom-up method that attempts to find new information about the data”* (Siegel, J., Shim, J., 2003, p.117) and *“deals with ‘knowledge discovery in databases’”* (Silberschatz, A., Korth, H. F., Sudarshan, S., 2002, p.830). Or it may come as a Hypothesis Testing that in some way will either prove or reject a view or idea that has been preconceived. Both forms of DM have its own application in the industry.

## **2.1.2 Data Mining Tools Currently in the Market**

This section of the chapter briefly describes a few of the DM tools that are in the market.

### **2.1.2.1. SAS Enterprise Miner**

Enterprise Miner by SAS addresses the entire data mining process through an intuitive point-and-click graphical user interface (GUI). Combined with SAS data warehousing and OLAP technologies, this tool creates an end-to-end solution that addresses the full range of knowledge discovery.

Enterprise Miner allows users to:

- Dramatically increase response rates from direct mail, telephone, e-mail, and Internet delivered campaigns.
- Identify your most profitable customers and the underlying reasons for their loyalty.
- Analyze clickstream data to improve e-commerce strategies.
- Determine what combination of products your customers are likely to purchase and when.
- Detect and deter fraudulent behaviour at your e-commerce site.

#### **2.1.2.2. Microsoft SQL Server 2000 - Analysis Services**

SQL Server 2000 Analysis Services provides integrated and Web-enabled analysis services. The OLAP component includes a middle-tier server that enables users to perform sophisticated analyses on large volumes of data. Data sources can include any OLE DB provider, such as SQL Server, Oracle, DB2, other relational databases, and text files. The data mining feature in SQL Server 2000 enables companies to discover patterns and trends and make predictions about future trends in their businesses.

Ventana Research (2004b) identifies that the software's "*Analysis Services extends the OLE DB standard with three specifications*":

- OLE DB for OLAP
- OLE DB for Data Mining
- XML for Analysis

## **2.2 Poll**

The word poll is synonym to a word we are all much familiar with. Survey. In this segment of the chapter the author will highlight the important aspects of a survey and its many uses in the industry.

### 2.2.1. Theory

Knowledge is the fuel that runs today's business. The success or failure of any company depends on knowing the attitudes, beliefs, and opinions of its people and also of the people it serves. The best way to determine these is by conducting a survey.

A survey may be called different things, depending on its purpose — a poll, a questionnaire, an *opinionnaire*, an evaluator, an assessment, an inventory, or a survey. Throughout this document, all of these various forms are referred to by the term “poll”.

Websurveyor Corporation (2003) identifies that “*a survey is a systematic, scientific, and impartial way of collecting information*”. For example, you can survey a group (or sample) of people about their feelings, motivations, plans, beliefs, and personal, educational, and financial background. This information is used to generalize conclusions or statements about the larger group (or population) from which the sample is drawn. The survey's intent is not to describe the particular individuals who take part in a sampling, but to obtain a statistical profile of the population.

Surveys used by many different organizations for many different purposes. Amongst them are to:

- Measure and improve their customers' satisfaction levels;
- Discover their workers' attitudes about issues affecting the work environment, quality, and productivity;
- Quickly evaluate opinions and attitudes;
- Provide data for long-range strategic planning and to enhance customer relations;
- Determine the effectiveness and attractiveness of their web page;
- Discover new market possibilities.

Resort managers survey their guests to gauge the level of their enjoyment of the facilities and discover opportunities for improvement.

Besides that, Dillman, Don A., Tortora, Robert D. and Bowker, Dennis (1999) tells us that “*a survey can accomplish many vital functions for an organization*”. They are to:

- Improve customer relations;
- Determine the quality of customer support;
- Evaluate your current and prospective customers;
- Indicate strengths and weaknesses;
- Pinpoint problems with productivity;
- Determine the quality of a web page.

Websurveyor Corporation (2002) tells us that in order to conduct an effective online survey, a company must first decide what specific topics you want to cover and what information you want to gather. To select the content of a survey, “*define your terms and clarify what you need to know*”. That is the most important criterion to plan and execute a successful survey. The author will illustrate this in relation to the current project.

Suppose that the Online Store managers would like to find out the level of customer satisfaction from the many new and returning store members that they have registered in their customer database. The *defined issue* at hand, i.e. Customer Satisfaction is pretty overwhelming and therefore must be ‘sliced’ up neatly before the poll is constructed.

In the example the author will illustrate later in this discussion, a poll to measure the level of promptness of delivery services is designed. Store members who have just registered may honestly answer that they have just made their first order whilst those who have already received their orders may have something to say. What the store members choose to say is what the store owners *need to know* in order to improve or preserve the value of their services. Without any means of identifying what the managers truly need to find out, the poll may be worthless for any sort of future analysis to be performed.

## 2.3 Dimensional Modeling

### 2.3.1 Theory

According to Dr. Joseph M. Firestone (2000), Dimensional Modeling is a favorite modeling technique in data warehousing, where the model containing tables and relations is “*constituted with the purpose of optimizing decision support query performance in relational databases, relative to a measurement of the outcomes of the business process being modeled*”. This model provides a highly efficient means to access the large data volumes for Decision Support System (DSS) analysis. It is a simple model that DSS users can easily understand.

He further states that conventional Entity-Relationship (E-R) models are in contrast constituted to remove redundancies in data model, facilitate retrieval of individual records having certain critical identifiers and therefore optimize Online Transaction Processing (OLTP) performance.

Nevertheless, Dr. Firestone proceeds to urge us that a *properly* constructed E-R model can be represented as a set of Dimensional or better known as Star Schema models without any loss of information. Which means that all information contained in an E-R model may not undergo any losses if or when converted to a Star Schema.

Dimensional Model is visually represented as a fact table surrounded by other smaller tables called dimensions, hence the name of “Star Schema”. A fact table contains the numeric measures of the business whereas the dimension tables contain the text descriptions of the subjects being measured. A relationship is defined where the fact table is an intersection for the dimension tables. A Dimensional Model is easy to modify, as it allows addition of new dimensions or fact tables.

The following figure shows an example of a Dimensional Model which summarizes the business information for product sales in relation to other information such as the product itself, the customer, time of sale and employee responsible for the transaction.

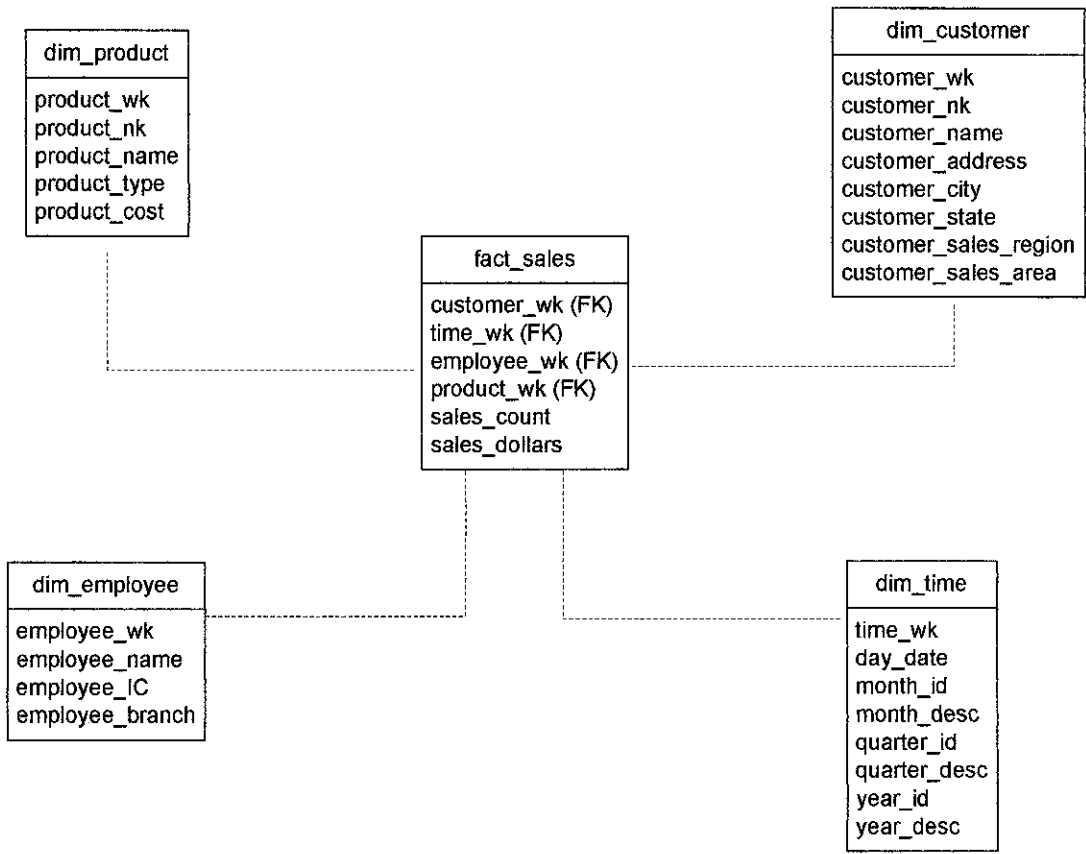


Figure 2.1: A Dimensional Model Example.

## **CHAPTER 3**

### **METHODOLOGY**

#### **3. PROJECT WORK OVERVIEW**

In this chapter, the author discusses the chosen method for the project work and the tools that had been used to make the system prototype development much easier and resulting in the success of project delivery.

##### **3.1 Procedure Identification**

The main purpose of this section is to list and outline the different methods that are used to meet the objectives of this project.

###### **3.1.1 Software Model**

The model depicted in the following figure is the Evolutionary Prototyping Model, where a version of a software prototype is developed by going through three separate phases, as will be discussed further later. They are the Planning, Analysis, and Design phases.

The two other remaining phases, namely Implementation and Support and Maintenance, however, will not be covered in this project development. Figure 3.1 illustrates these phases' relationship to the earlier four phases that will be undergone in the development of this project.

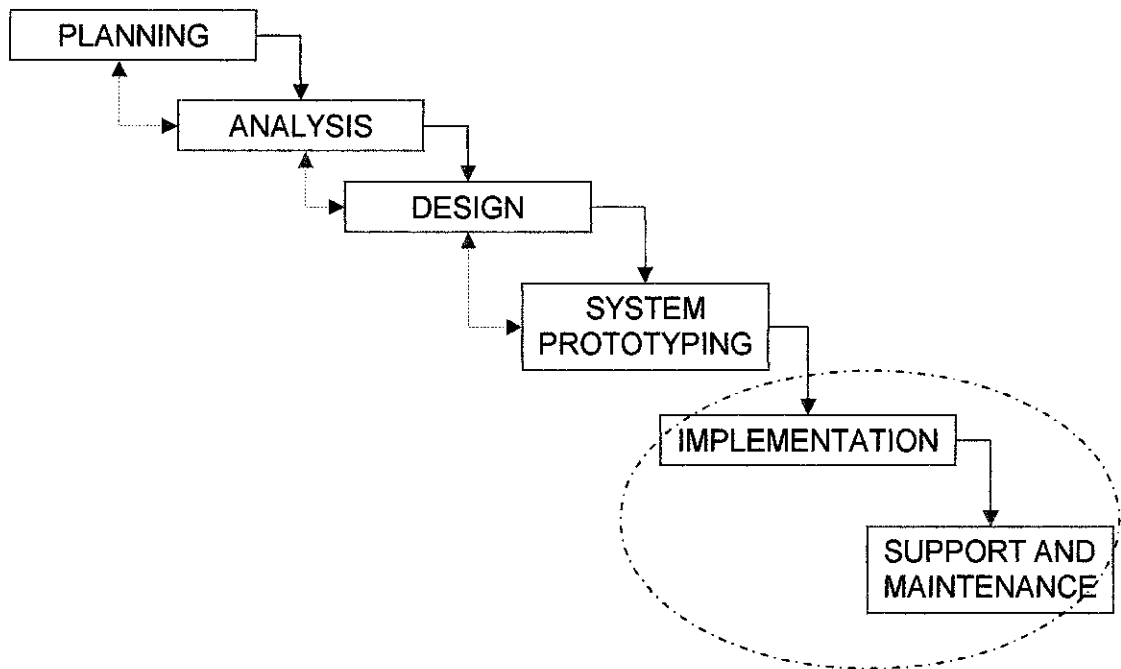


Figure 3.1: The Evolutionary Prototyping Model.

The procedures that the author has undergone during the specific phases are as listed:

#### ***3.1.1.1 Planning***

To identify the problems that can be solved by proceeding with the title. The problem statement, objectives and scope of study are acknowledged (please refer to the Introduction chapter for details) and approved before proceeding with the development of this project.

#### ***3.1.1.2 Analysis***

To understand the design requirements for the e-Poll Data Mart, to understand the different ways of processing an Online Poll and choosing one method that best corresponds to the desired results, and to understand the strengths of using Data Mining tools for analysis of the captured Poll data.

The author also determines the hardware and software requirements to completely and successfully execute the development of the project.



### ***3.1.1.3 Design***

To design the Data Mart for the e-Poll system and Online Store using a chosen database tool, to design the interface for the Online Store member registration, to design the interface for capturing and prompting results of Store member's poll participation and to establish the kind of analysis (analytical, graphical or otherwise) that can be useful for the Poll data captured.

### ***3.1.1.4 System Prototyping***

To use dynamic Web Programming tools with ODBC database connection to code and successfully show the working prototype of the e-Poll system, which includes seamless capturing of customer information for the Online Store and their vote in the prompted poll.

## **3.1.2 Detailed Literature Reviews**

Some details pertaining to the technical aspect of the project requires carefully gathering data and deciphering literature resources. This includes understanding the different forms and techniques of Data Mining and architecture of Online Stores and Polls from the WWW, magazines, books and/or journals.

## **3.1.3 Observation**

Through observations from the Web, the mechanism of an e-Poll system and Online Stores must be understood. Results from observing samples of these systems will allow the author to properly and correctly design the database structure for the system and the interface design for the prototype.

## **3.1.4 Prototyping**

A functional prototype of an e-Poll system that will demonstrate the concept of capturing Online Store member's poll participation must be developed, especially for

demonstration or presentation purposes. It can be developed simultaneously during the project development, as it allows for refinements of user requirements analysis and the system design.

Presentation of this prototype will be choreographed in such a way that the examiners will see the flow or steps to be taken before the e-Poll segment of the prototype is initiated. It will also allow the examiners to analyze the usefulness of the Data Mining analysis made for the system managers of the Online Store business.

### **3.2 Tools Required**

Among the tools needed to construct the system prototype are as follows:

- **Macromedia Dreamweaver MX** – a web programming tool used to create web pages.
- **XARA Webstyle 4** – a graphics design software that enables the creation of personalized web graphics, and complete website designs.
- **Microsoft Access** – suggested database to store the data collected throughout the project. Data will be filtered accordingly and the database design for them will be completed simultaneously.
- **Allaire ColdFusion** – an advanced web tool which will be useful to create dynamic pages (i.e. integration between the interface and ODBC).
- **Apache Web Server** – the chosen web server to run the system on a local, standalone PC without network connections.
- **Internet Connection** – for the process of data gathering and observation needed during the Analysis and Design phases of the chosen method.

- **Miner3D Excel** – a Data Mining software package that allows Microsoft Access data to be tabulated and represented in a graphical format. Further analysis can be performed such as refined query, and identifying values for min, max, count, average and sum.
  
- **Microsoft Visio** – for documenting and visually depicting the process flows and Data Mart structure of the e-Poll system.
  
- **Microsoft Project** – for keeping track of project development’s milestones and tasks.
  
- **Personal Computer Specifications**
  - Operating System : Microsoft Windows XP
  - Processor : AMD Athlon 1.8 GHz
  - Memory : 256 MB RAM
  - Hard drive : 40 GB
  - CD-ROM Drive : 56X speed
  - Sound : 16 bit CD quality
  - Peripherals : Mouse, Keyboard, Printer, Scanner, External CDRW Writer

## **CHAPTER 4**

### **RESULTS AND DISCUSSION**

#### **4. OVERVIEW OF RESULTS**

In this chapter, the author will elaborate in detail the project findings and following this, will discuss related results. This chapter contains in it the ‘heart’ of the project the author has worked for. At the end of this chapter, readers will understand the depth of the work completed which encompasses the scope of study listed in the Introduction chapter.

#### **4.1 Findings**

In this section, the author presents her findings that were recognized during the development of this project.

##### **4.1.1 General Observation**

###### ***4.1.1.1 e-Poll Database Requirement***

During the development phase of this project, the author had identified that there were more issues to be addressed concerning placement of an e-Poll system in a business website. An established data warehouse or data mart is required prior to implementation of the data mining technique (See *Appendix A* for Data Warehouse and Data Mining terminology). The author had also found that an e-Poll system on its own does not identify or maintain any customer or store member’s personal details or information – all of which is required to help find useful patterns in the data captured through the system.

We discuss an example. We want to find out from which countries or regions do customers over 40 years old think that the new interface design of the website is excellent. To pursue this, the e-Poll database has to be joined with an existing customer database in a query to find out what percentage of the group thought highly about the topic. If there were a 5-scale answering scheme denoting *Very Interesting* to *Very Ugly*, our first objective is to find the number of the over-40 year olds that answered *Very Interesting* to the question in relation to their sum.

Secondly, having the sum, we then segment it according to geographical location. The result: A table or chart showing the distribution of over-40 year old customers from different locations across the globe who judged that the site's interface design is excellent.

A good example of a system that maintains the said customer information is an Online Store such as amazon.com or barnesandnoble.com. These companies have hundreds of thousands of customers all over the world who buys all types of products from their websites. They keep records of their customers to enable the buyers to minimize their shopping time (if they ever return), by means of asking them to fill out a personal information form prior to making their initial purchases.

When the company has these customer data, they can distinguish the correlation between the buying habits of the customer with their geographical location, gender and age. In essence, they have enough information to help them classify their data in order to help them find meaningful analysis.

Therefore, in response to this observation, the author had extended the e-Poll system to be incorporated into an Online Store that will hold the customer information needed in order for correlation attempts to be successful. Aside from the customer or store member's personal details, the initial database schema for the e-Poll is listed as shown in *Table 4.1: e-Poll Database Schema*.

Table 4.1: e-Poll Database Schema

Table Name	Column Name	Data Type	Null
Poll_Question	Q_ID	VARCHAR2 (3)	N
	Question	VARCHAR2 (75)	N
	EntryDate	DATE	N
	ActiveDate	DATE	N
	ExpireDate	DATE	Y
	TotalVotes	NUMBER	Y
PollAnswer	A_ID	VARCHAR2 (4)	N
	Q_ID	VARCHAR2 (3)	N
	Answer	VARCHAR2 (25)	Y
	VoteCount	NUMBER	Y
	Percentage	NUMBER (6,2)	Y
PollArchiveQ	Q_ID	VARCHAR2 (3)	N
	Question	VARCHAR2 (75)	N
	EntryDate	DATE	N
	ActiveDate	DATE	N
	ExpireDate	DATE	Y
	TotalVotes	NUMBER	Y
PollArchiveA	A_ID	VARCHAR2 (4)	N
	Q_ID	VARCHAR2 (3)	N
	Answer	VARCHAR2 (25)	Y
	VoteCount	NUMBER	Y
	Percentage	NUMBER (6,2)	Y
PollComment	C_ID	VARCHAR2 (4)	N
	Q_ID	VARCHAR2 (3)	N
	Name	VARCHAR2 (25)	N
	Subject	VARCHAR2 (40)	Y
	Comment	VARCHAR2 (300)	N

On the other hand, *Table 4.2* in the following page lists the initial database schema for the Online Store member's information – including billing, shipping and personal details. However, even though a practical Online Store may have other tables such as for their items or products listings, transactions and others, it is determined to be not relevant to this project.

Table 4.2: Online Store Personal Info Database Schema

Table Name	Column Name	Data Type	Null
Member	MemberID	NUMBER	N
	MembershipDate	DATE	N
	Name	VARCHAR2 (75)	N
	Title	VARCHAR2 (5)	N
	Email	VARCHAR2 (25)	N
	Login	VARCHAR2 (25)	N
	Age	NUMBER	N
	Gender	VARCHAR2 (2)	N
	StatusID	NUMBER	N
	City	VARCHAR2 (30)	N
	State	VARCHAR2 (15)	N
MemberStatus	StatusID	NUMBER	N
	Status	VARCHAR2 (25)	N
MemberShipping	ShippingID	NUMBER	N
	MemberID	NUMBER	N
	ShippingAddress	VARCHAR2 (200)	N
	ShippingPostcode	NUMBER	N
	ShippingCity	VARCHAR2 (30)	N
	ShippingState	VARCHAR2 (15)	N
MemberCC	CCID	NUMBER	N
	MemberID	NUMBER	N
	Cctype	VARCHAR2 (10)	N
	CCnumber	NUMBER	N
	CCexp_mo	NUMBER	N
	CCexp_yr	NUMBER	N
	CCname	VARCHAR2 (75)	N
MemberLogin	Login	VARCHAR2 (25)	N
	Password	VARCHAR2 (25)	N
Group	Group	VARCHAR2 (10)	N
	Privilege	VARCHAR2 (40)	N
MemberGroup	Login	VARCHAR2 (25)	N
	Group	VARCHAR2 (10)	N
Item	ItemID	VARCHAR2 (10)	N
	Description	VARCHAR2 (25)	Y
	UnitPrice	NUMBER	N
Transaction	TransID	NUMBER	N
	MemberID	NUMBER	N
	Date	DATE	N
	ItemID	VARCHAR2 (10)	N
	Qty	NUMBER	N

As mentioned before, the only table of interest from the Online Store is the Store Member's personal details, namely tables Member and MemberStatus.

To find any correlation between the Store member's personal details and their poll participation information, a table must be created to hold the important and necessary values. The table created, MemberPoll, is as listed in Table 4.3.

Table 4.3: Store Member-Poll Database Schema

Table Name	Column Name	Data Type	Null
MemberPoll	MemberID	NUMBER	N
	Q_ID	VARCHAR2 (3)	N
	A_ID	VARCHAR2 (4)	N
	Date	DATE	N

The following is the resulting Entity-Relationship Diagram formed from the database schemas shown previously. The major tables needed for successful analysis of poll participation information are enclosed in the center square.

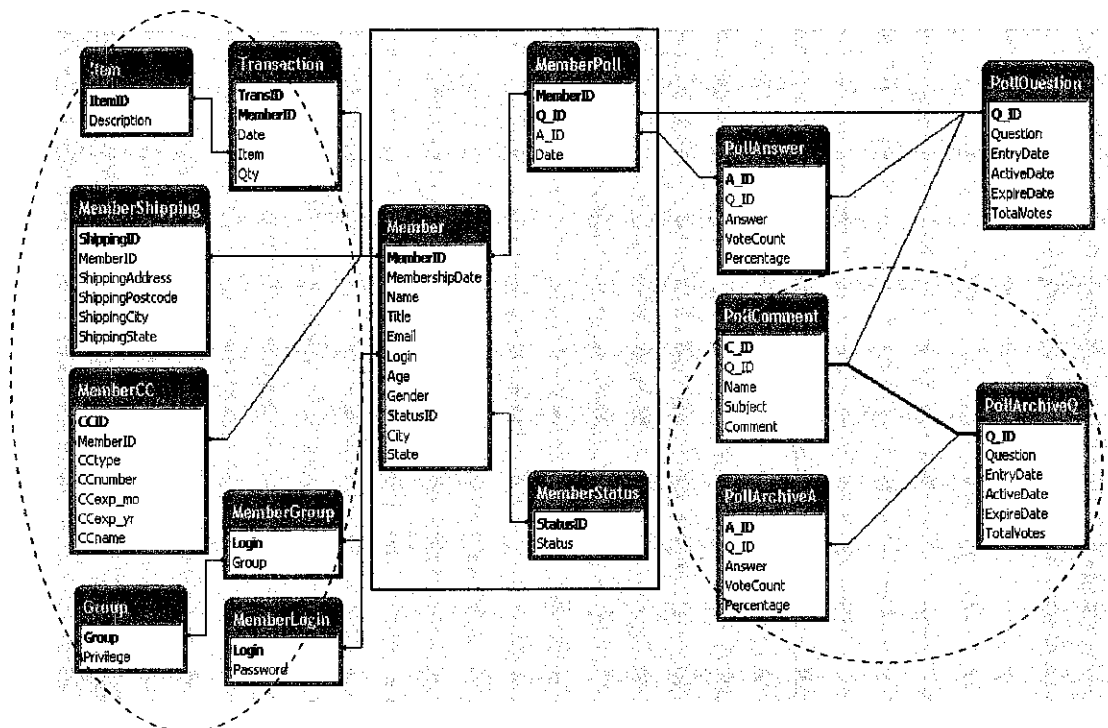


Figure 4.1: Entity Relationship Diagram for the e-Poll System



#### **4.1.1.2 e-Poll Placement**

On another note, it is imperative for us to understand the structure of an e-Poll system. A poll is often presented in a main page of a corporate website, intranet or internet, and is normally situated in the left- or right-frame of the window. At any given time, only one (1) set of question and answers are presented. The poll is left unchanged for a specified time frame, depending on the company's policy or Administrator scheduling. A noted disadvantage is the website's inability to guarantee that each visitor to the site would, in fact, answer the poll.

However, managing it inside an Online Store, and to have the rate of answers from customers to be 100%, the poll needs to be prompted upon a successful customer login or new registration. This means that the target users of the poll are the online store's registered customer (or member's) themselves. When the poll is answered, the customer will *only* then have total privilege to the members-only information, such as Member's-Only bulletin or bargains.

The author believes that the *placement* of the poll after a successful member login state should ascertain system owners that the poll results are obtained from their most valuable asset – their customers. The same set of questions could be asked or prompted again after a scheduled time frame, such as semi-annually or quarterly. This will enable the system owners to distinguish the difference in the captured data of the two or more time periods.

#### **4.1.2 Prototyping**

To accomplish the project objective, the prototype is developed to cater to the users of an Online Store. The store manager and members alike will share a common process, that is, the e-Poll system upon sign-in or login state is achieved.

For registered store members, they will be prompted with two (2) options to choose from *upon* answering the poll question. That is, when customers are seeing a window showing the current poll results and user comments. First option is to choose to enter any comments (if desired) or secondly, to exit the poll pop-up screen and e-Poll

system. The author believes that most store members would choose the latter option after being shown the horizontal bar chart of the current poll results.

The flow mentioned is featured in *Figure 4.2: e-Poll Process Flow for Store Members*.

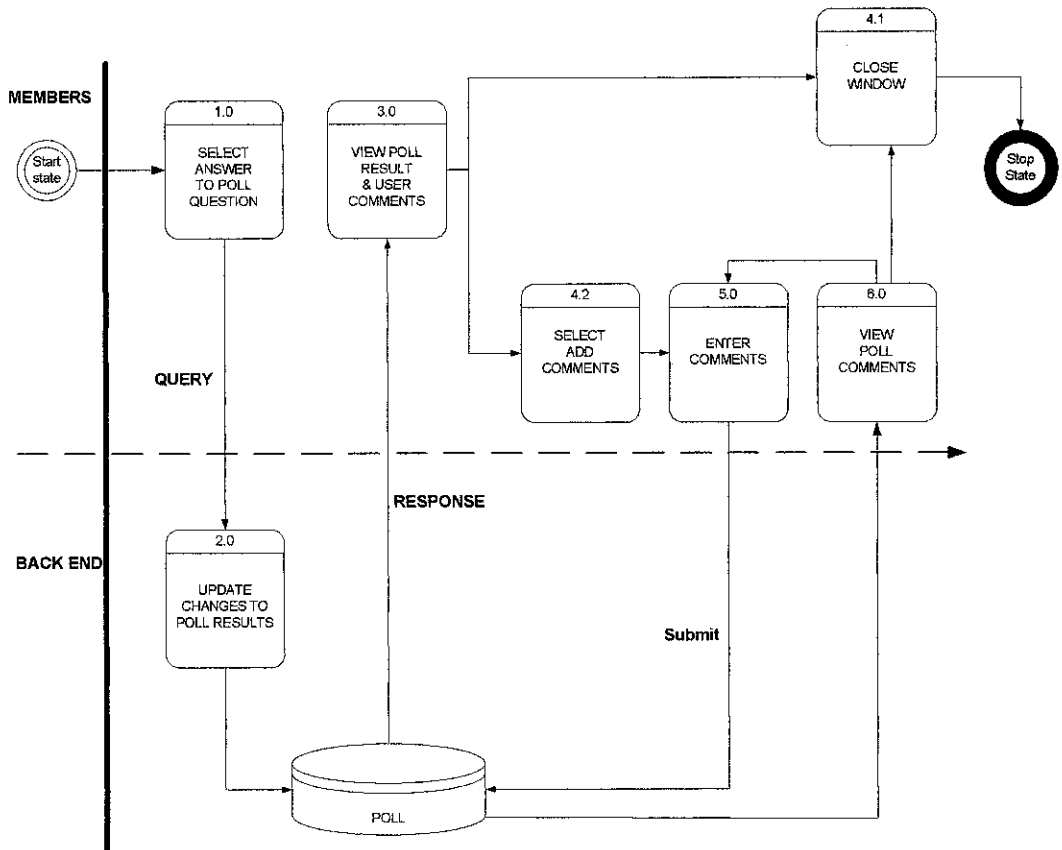


Figure 4.2: e-Poll Process Flow for Store Members

The system owners or managers will have the same privilege as the store members except that they also have access to analysis of the poll results. This tool will allow them to view the current status regarding their customer's opinions on the poll topic. Using a Data Mining software package, the graphical representation of the results can be shown. Depending on the functionalities or capabilities of the third-party software, the system managers may have entitlement to manipulate the analysis, by making further queries such as calculating the min, max, average and sum of certain values of interest.

The unique process flow for system owners mentioned is featured in *Figure 4.3: e-Poll Additional Process Flow for Managers*.

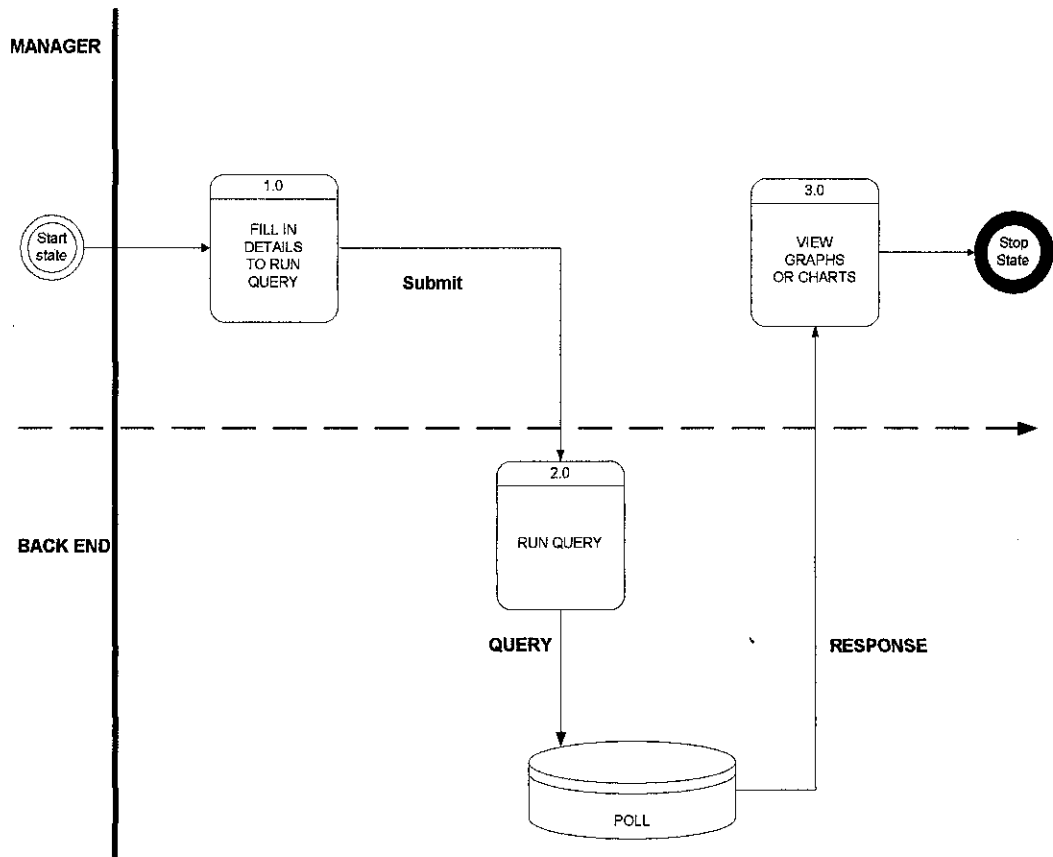


Figure 4.3: e-Poll Additional Process Flow for Managers

## 4.2 Discussion

In this section, the author discusses the results that were established after true completion of this project.

### 4.2.1 Data Mining

Data Mining allows system owners or managers to transform their raw data into a more useful and valuable data collection. They save time in uncovering data trends compared to more traditional means. Managers are able to analyze vast amount of

stored or captured data at a time, and the results of the analysis will allow and aid them to plan their next strategic move accordingly.

The use of Data Mining tools available in the market today has enabled many up-and-rising companies to flourish. Companies are now provided with a tangible Return of Investment (ROI), which is provided by analytics platform; where one obvious ROI is derived from the number of people-hours that are saved by automated analysis and report generation. These tools help companies continue to fulfill their mission and objectives, by means of interpreting analysis generated through the Data Mining means.

Aside from the tools available in the market to achieve this, Data Mining itself can be incorporated in a lot of ways pertaining to Customer Relationship Management (CRM). An online store may use Data Mining to acquire new customers by identifying prospects and converting them into customers. Also, the technique is useful to retain good existing customers, by predicting the group of customers that have a high probability of leaving and concentrating their effort in preserving the relationship with remaining ones.

The basic steps of data mining for an effective CRM are:

- To define business problem
- To build marketing database
- To explore data
- To prepare data for modeling
- To build model
- To evaluate model
- To deploy model and results

However, to make this discussion relevant to this project, Data Mining is important in determining patterns and classifications that Online Store managers can use to help them effectively plan their next strategic move. This is performed via a bottom-up method known as Knowledge Discovery, whereby Data Mining enables owners to explore and find new information about their customers' behaviors or opinions.

#### 4.2.1.1 *Limitations of Data Mining Tools*

In order to perform analysis of the captured poll data, the author has chosen to make use of a trial version of a Data Mining software called *Miner3D Excel*. This decision was based on the tool's graphical capabilities and manipulation of numerical data values. A discussion on a sample Poll analysis will follow later in this chapter.

However, the author has identified that if any Data Mining software is used, system owners or managers have to be ready to 'pay the price'. Not all visualization of a perfect data analysis could become a reality, especially when dealing with store-bought software packages. In contrast, if system owners hire a developer to come up with the type of analysis they need, perhaps what they envision could in fact take place.

Amongst the limitations the author has identified after using the analysis tool are:

- *Technical Knowledge of Software Developer –*

Software is as useful as it's programmed to be. When using such an analysis tool, we are faced with issues of less or insufficient functionalities or in some cases, undesired or useless functions. A less complex algorithm structure results in severe limitations, such as less functionality or usefulness of the software.

- *Deciphering Data Analysis –*

In *Miner3D Excel*, only numerical values from the database could be graphically represented and analyzed. Numerical functions allowed are to calculate the min, max, count, sum and average of the values. Although these analyses is very helpful for the project at hand, the person who 'reads' the results of the analysis must be one who understands what certain values for a field means.

For example, say we have a field for gender. Male is denoted by the value 1 while female is given the value 2. When *Miner3D Excel* represents the data, there may be a count value of 1's and 2's of the gender field, where there are 40 1's and 10 2's. If

a system owner does not understand what the value 1 and 2 represents, they may mistakenly think that there are 40 female and 10 male participants in the current poll results. This is where the limitation lies. A numerical number that represents a 'true' value may cause the wrong interpretation of the results.

- *ODBC and Interface Connection –*

Some software only allows integration with selected database sources. For example, *Miner3D Excel* works well with Microsoft Excel data, but fortunately for the author, it also allows analysis of copied Microsoft Access data. On the other hand, other tool such as Microsoft SQL Server 2000 with Data Mining extensions only allows analysis of data created using the software itself. This causes system owners to be apprehensive in selecting the 'right' tool for their analysis needs, as they should not invest in software that does not allow integration with their selected data source.

#### **4.2.2 Dimensional Modeling**

Dimensional Modeling (DM) or Star Schema is considered the best data modeling technique compared to Entity-Relationship Diagrams (ERD), as it possesses the following advantages:

- Superior (or optimal) decision support query performance in relational databases,
- Greater understandability,
- No loss of information (as any E-R model can be represented as a Star Schema).

In the earlier e-Poll Data Mart shown in *Figure 4.1*, query performance for the poll results may take a while. This is true whenever analysis of the marital status of the poll participant is to be determined when queried from the table `MemberPoll`. To extract the value, the query requires a double join of tables, from `MemberPoll` to `Member` and lastly to `MemberStatus`. This double join state will result in poor query performance.

To overcome this problem, a Dimensional Model or Star Schema of the ERD in *Figure 4.1* is developed. The resultant model is depicted in the following *Figure 4.4*.

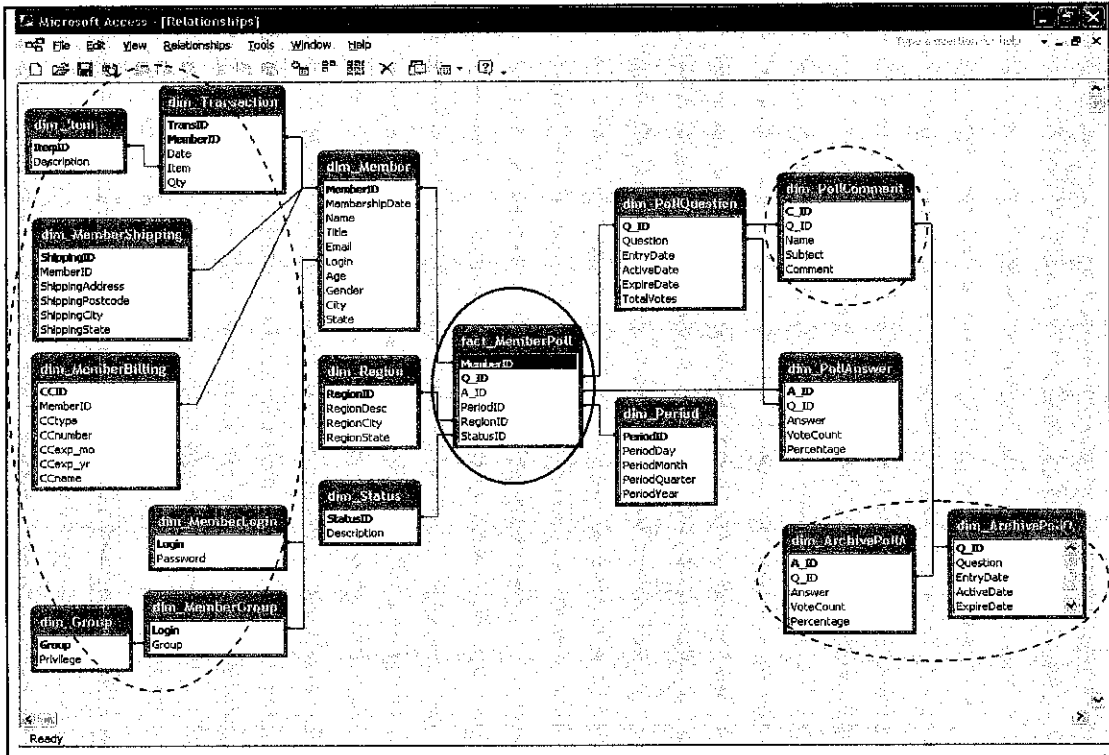


Figure 4.4: Star Schema for the e-Poll System

In the Star Schema illustrated, we see a central table named `fact_MemberPoll` become the intersection of other dimensions namely `dim_Member`, `dim_Region`, `dim_Status`, `dim_Period`, `dim_PollQuestion` and `dim_PollAnswer`. This fact table contains the foreign keys to the dimensions surrounding it and will eliminate the cause for double join queries that stunted the query performance in the ERD model.

Referring to *Figure 4.4*, the author stresses the relevancy of the relationship between the fact table and only the *directly attached* dimensions to this project, and sets aside all other dimensions that are enclosed in dotted blue circles from the model. These dimensions are necessary if Data Mining in Customer Relationship Management (CRM) is pursued and therefore is irrelevant to the discussion of this project work.

## 4.2.1 The e-Poll Process

Polls are widely used in marketing research to gather potentially important opinions, but must meet certain standards to ensure statistical accuracy. To improve its value, polls must be structured in a way that produces more statistically reliable results.

The process of participating in a poll begins after the login or new member registration state in the Online Store. Upon successful entry or registration, the e-Poll process begins. It takes into account the personal details in the registration form entered by a new customer (shown in *Figure 4.5*) or the extraction of personal data for a returning store member upon the person's successful login (shown in *Figure 4.6*).

Dinda Home Furnishings | Registration/Login - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: http://localhost/storeregistration.cfm

HOME About Us Products Shipping & Returns

Dinda Home Furnishings  
the stuff that dreams are made of

HOME  
About Us  
Products  
Shipping & Returns

Please login or register to continue...

Products > Shipping > Payment > Register

Please fill in all fields :

Choose a Login: Alia

Choose a Password: .....

Email address: alia@yahoo.com

Title: Miss

Name: Alia

Age: 22

Gender: Female

Marital Status: Single

City of Origin: Alor Setar

State: Kedah

Remember your chosen Login and Password; you will need them if you wish to log in again.

Reset OK

Dinda Home Furnishings  
the stuff that dreams are made of

Member Login

Login:

Password:

OK

Figure 4.5: New Store Member Registration



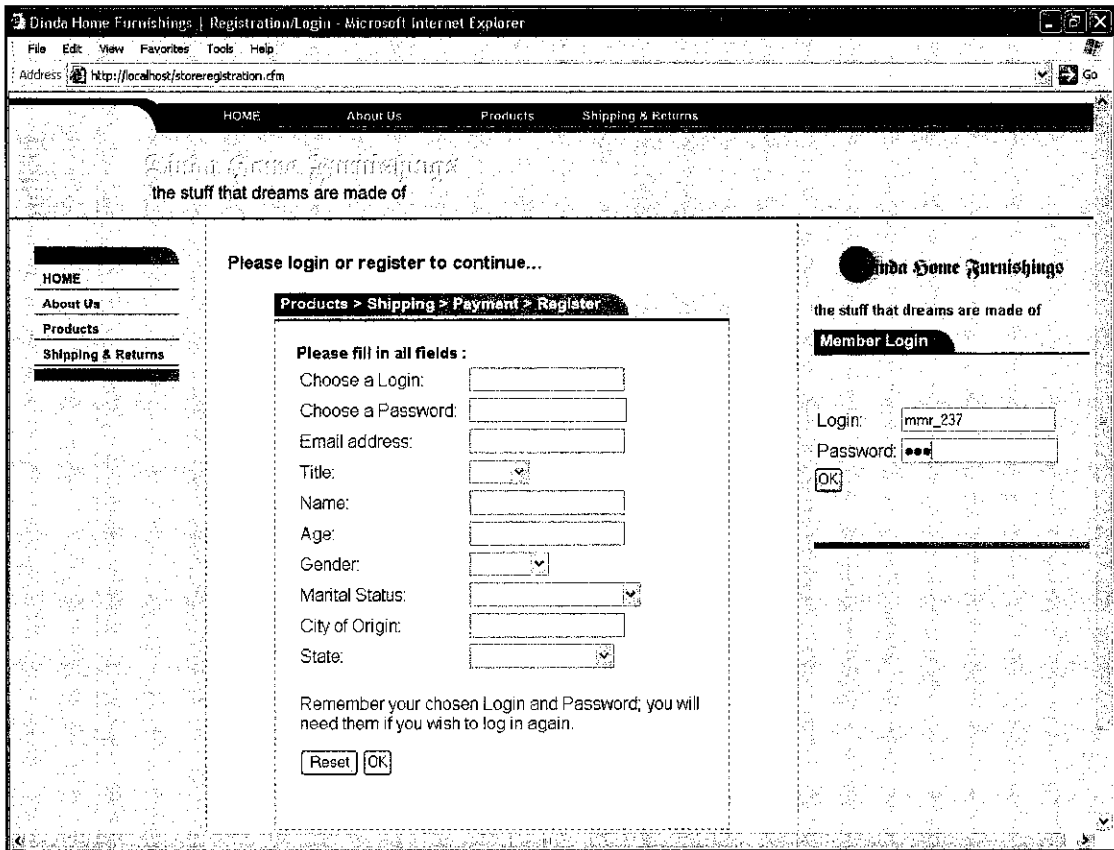


Figure 4.6: Returning Store Member Login

The details entered in *Figure 4.5* will be inserted into `dim_Member`, a very important table in the Data Mart that safe keeps the store member's personal information for future data retrieval and analysis.

Following this, the poll will be prompted as shown in *Figure 4.7*.

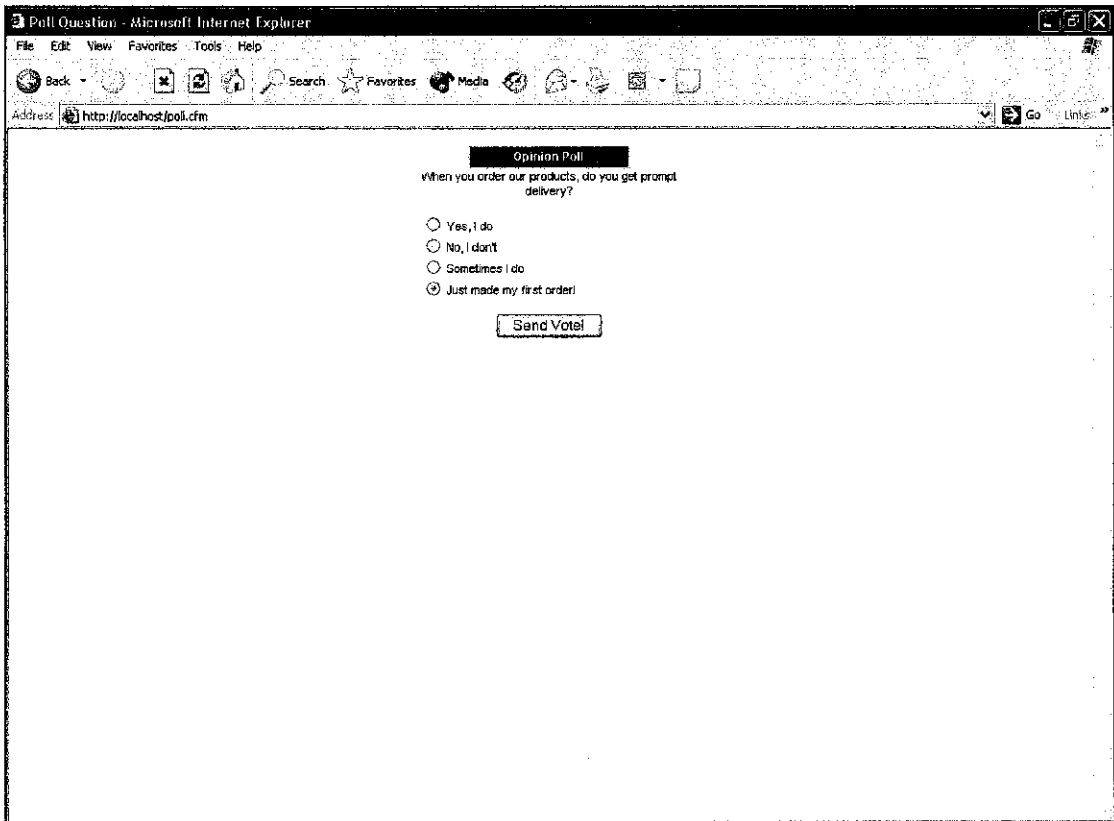


Figure 4.7: Poll for Store Members

After a store member chooses the appropriate answer to the question and clicks on 'Send Vote!', selected values from the participant's personal data and the answered poll (including poll question and answer ID's) will be inserted into the table `fact_MemberPoll`.

Following the flow from *Figure 4.5*, the resultant record inserted into table `fact_MemberPoll` is as enclosed in a box shown in *Figure 4.8*.

Microsoft Access - [fact\_MemberPoll : Table]

File Edit View Insert Format Records Tools Window Help

MemberID	Q_ID	A_ID	PeriodID	RegionID	StatusID
2		1004	2	32	2
3		1003	2	32	2
4		1001	2	3	1
5		1003	2	6	1
6		1001	2	5	1
7		1001	2	6	1
8		1004	2	5	3
9		1003	2	5	3
10		1004	2	5	1
11		1002	2	22	1
12		1002	2	27	1
13		1001	2	6	3
14		1003	2	11	3
15		1004	2	26	2
16		1001	2	28	3
17		1001	2	28	3
18		1002	2	22	2
19		1004	2	22	3
20		1003	2	1	2
21		1004	2	1	1

Figure 4.8: Updated Record in fact\_MemberPoll Table

However, on the store member's part, they will be shown a pop-up menu with the current poll results and user comments, as shown in *Figure 4.9*.

Following this, the poll participant may then choose to either close the pop-up window or to add new comments for the benefit of other users. This step is as shown in *Figure 4.10*.

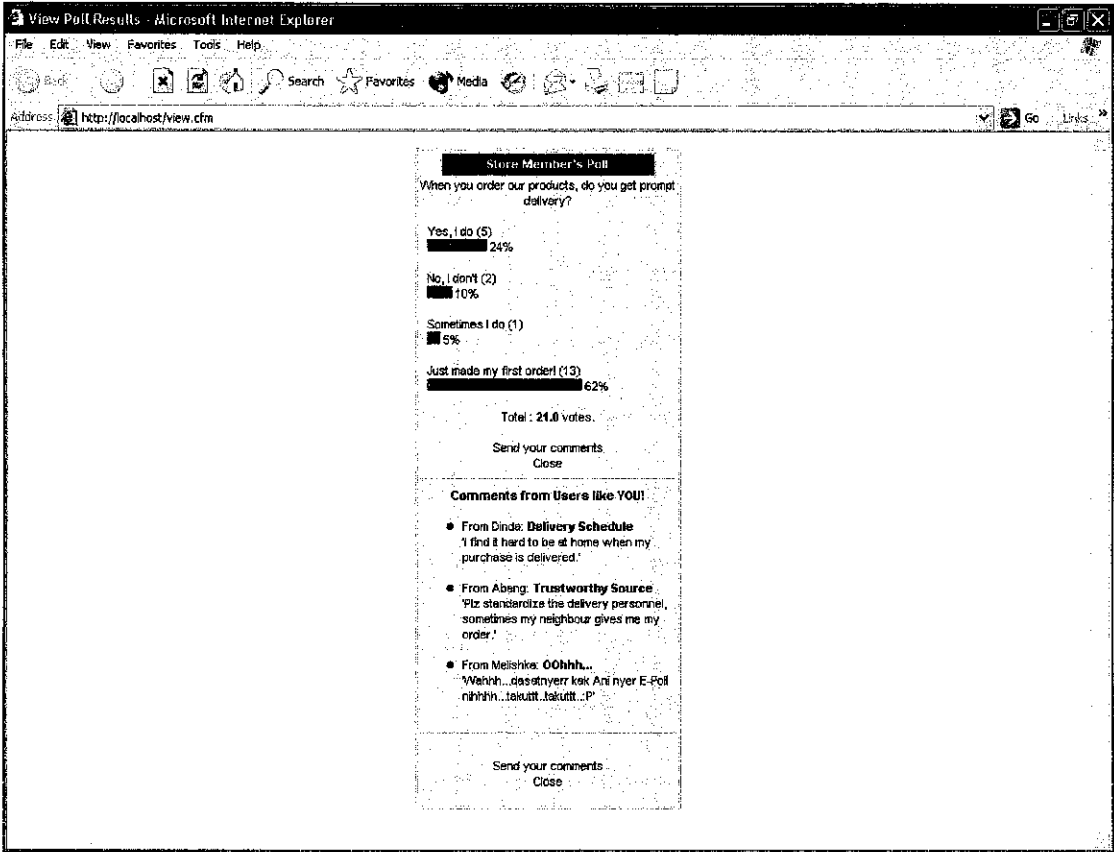


Figure 4.9: Pop-up Window of Poll Results

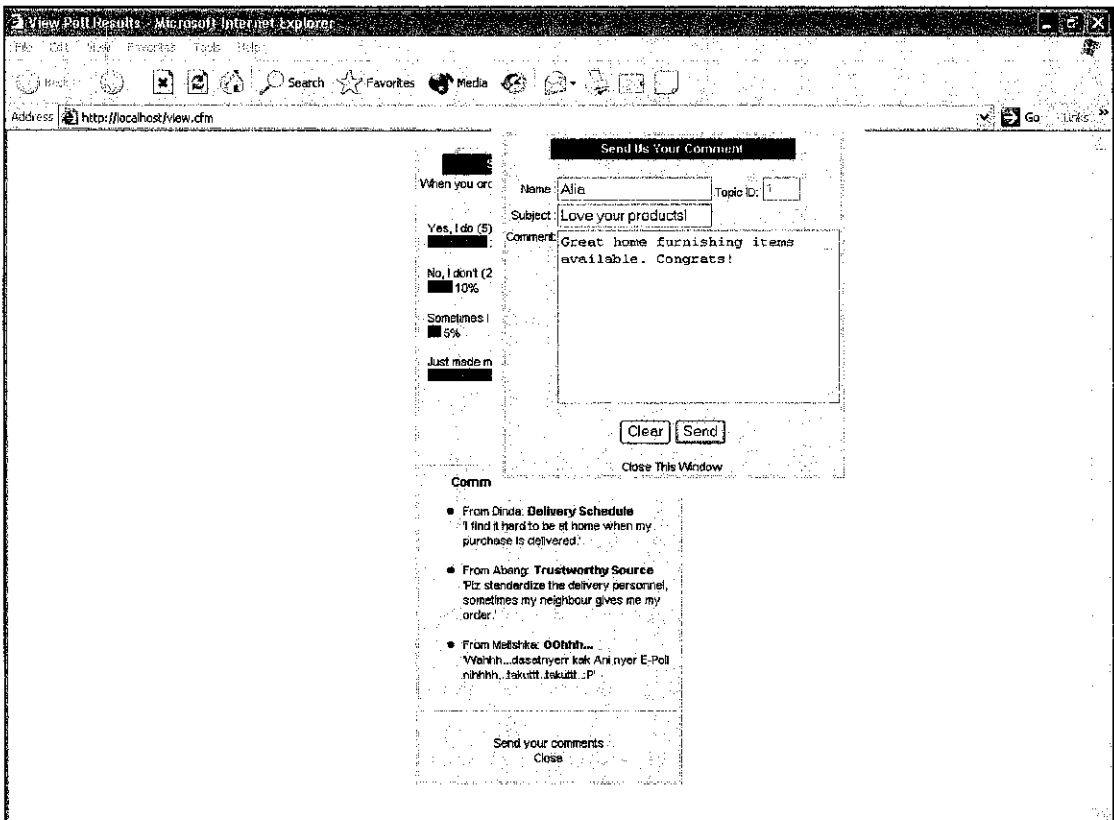


Figure 4.10: Adding Comments

After successful submission of the comment, a confirmation window will be shown as in *Figure 4.11*.

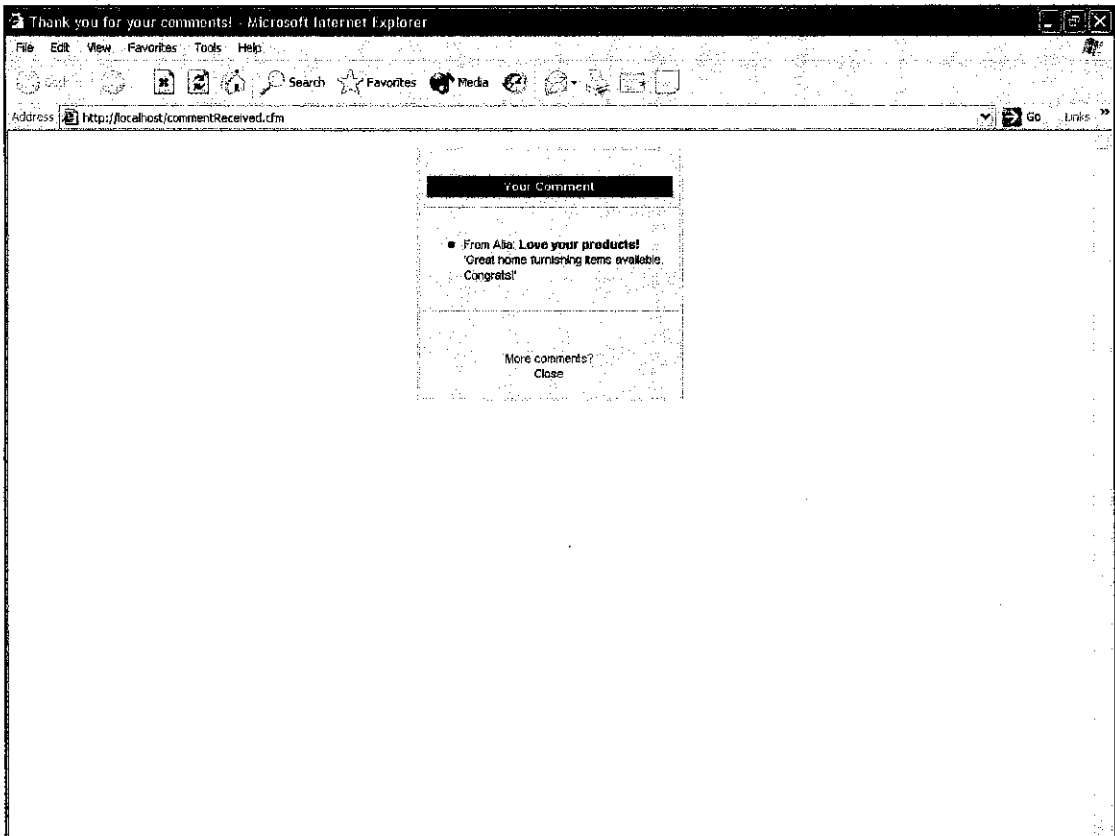


Figure 4.11: Confirmation of New Comment

The e-Poll process has now been completed for this particular store member. Upon clicking on 'Close', she will then be guided back to the Online Store for other tasks (if desired).

However, in relation to this project, the author would like to remind that the duty of these interfaces is to capture customer data and their poll participation information, which will then allow the next stages of the Data Mining process to be performed.

## 4.2.2 Poll Analysis for Managers

*Miner3D Excel Professional*, the Data Mining tool used in this project allows representation and analysis of numerical values from a database source. It allows the data to be tabulated in a scatter, bar or line graph format, and includes great options to make the data more ‘lively’ – such as changes in the data’s look, color, depth and perception (translate, rotate, reflect) in the XYZ-axes. This software also includes tools such as Selector (enclosed in a pink-colored square) and Calculator (enclosed in a lavender-colored square), as shown in *Figure 4.12*.

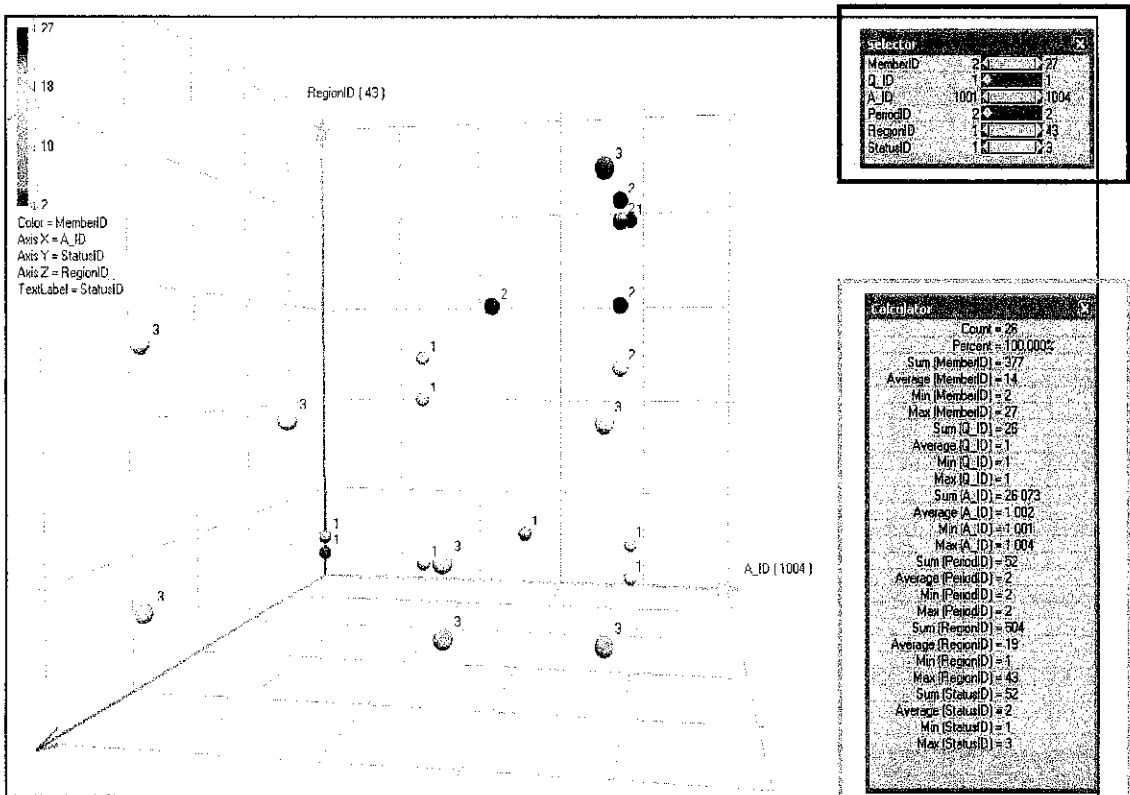


Figure 4.12: Graphical Representation of *fact\_MemberPoll* (all records)

The Selector tool allows the user to make more refined queries by dragging the tag for any data to either be a single value (item slider) or a range of values (range slider). In effect, this tool eliminates the need to write SQL statements ending with, as an example, ‘where StatusID = 1’.

On the other hand, the Calculator tool shows us the overall count and percentage of the records based on the queries made in the Selector tool (if any). For example, in *Figure 4.12*, the count shows the number of all the records (26 in all) with the percentage being 100% as we have not made any queries. This tool also prompts us with the sum, average, min and max values for *all* the fields in the database at all times, taking into effect any changes made in the Selector tool.

*Figure 4.12* illustrates the graphical representation of all the records copied from Microsoft Access's fact\_MemberPoll table. For complete database schema for all the tables used in the system, please refer to *Appendix B* of this document.

In *Figure 4.13*, the author demonstrates the power of the range slider in the Selector tool. Let's say the managers ask us to find out how many of their customers are married, with or without children. This refers to the StatusID field where the data are denoted by the values 2 and 3 (as shown by the blue arrow). The Calculator tool shows us the results – 62% or 16 customers are married.

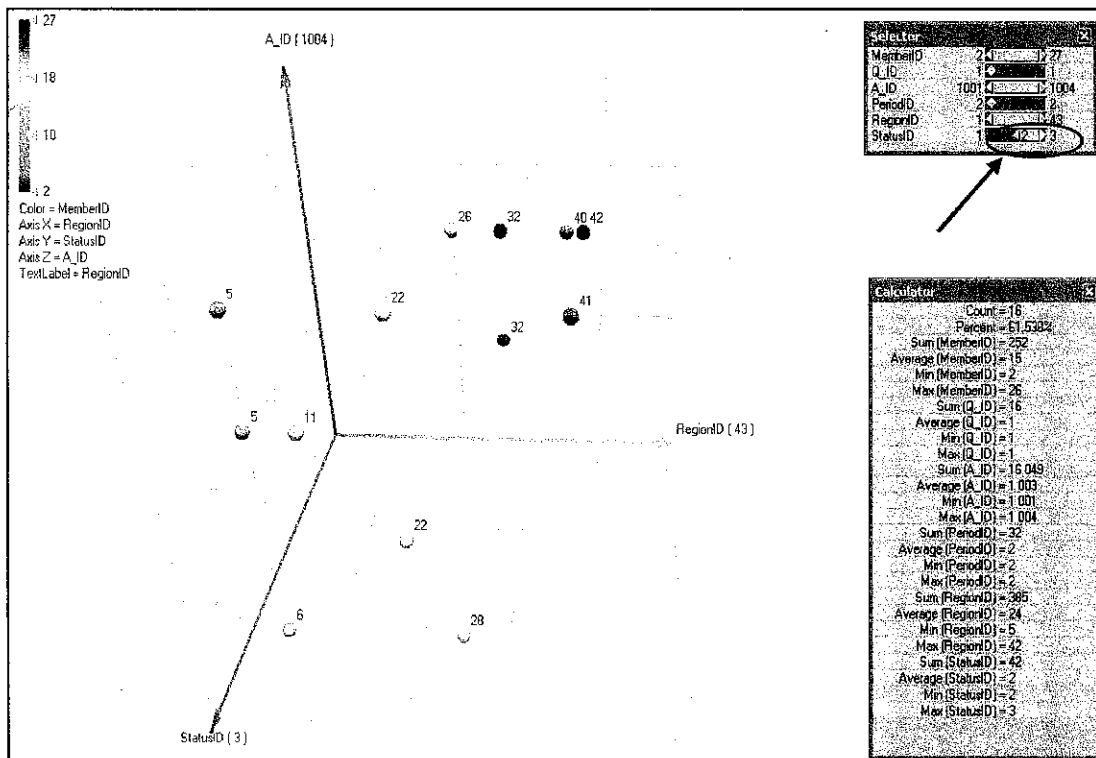


Figure 4.13: Graphical Representation of fact\_MemberPoll  
(married participants)

In *Figure 4.14*, the managers ask us to determine the number of customers living in the Central region (namely Selangor, Kuala Lumpur and Putrajaya) and what their marital status is. This refers to the `RegionID` field where the data are denoted by the values from 25 and 32 (as shown by the green arrow). The `Calculator` tool shows us the results – 23% or 6 customers are from the Central area. As can be viewed from the corresponding graph on the Z-axis (`StatusID`), only 1 customer is single while the others are married.

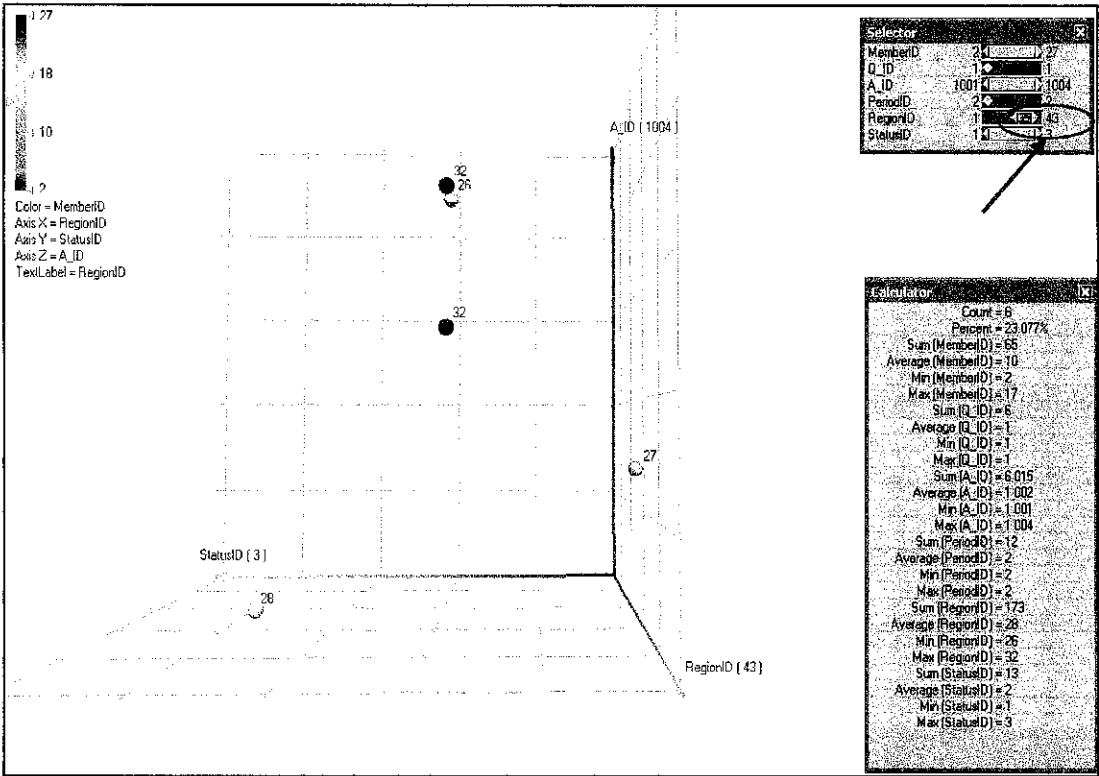


Figure 4.14: Graphical Representation of `fact_MemberPoll`  
(Central region participants)

We may analyze more of the data by rotating the graph in any direction to view the values for a given field in the XYZ-axes. From the above figure, we observe that we may determine what the customer's poll answers are (`A_ID`), which city they are from (`RegionID`) and what their marital status is (`StatusID`). The complete meaning of the numerical values for these fields is as listed in *Appendix C*.



In another example shown in *Figure 4.15*, we answer a manager’s query of the number of customers in the Utara region. This refers to the `RegionID` field where the data are denoted by the values from 1 to 8 (as shown by the red arrow). The Calculator tool shows us the results – 42% or 11 customers are from this region. As can be viewed from the corresponding graph on the X-axis (`A_ID`), the participants’ answer to the Poll is fairly distributed – where majority of them have just made their first order (`A_ID = 1004`).

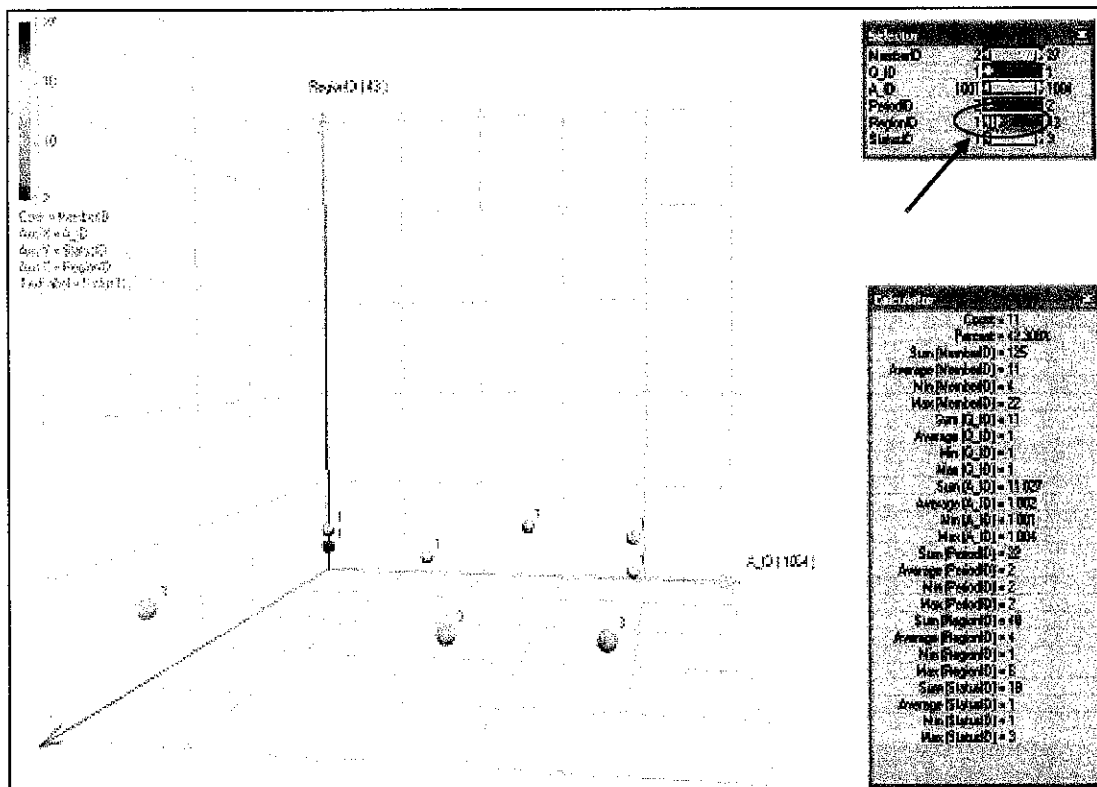


Figure 4.15: Graphical Representation of `fact_MemberPoll`  
(Utara region participants)

Lastly, another powerful tool of the *Miner3D Excel Professional* is that it allows the graph and details of the query to be imported to a Web Browser. We observe this feature in *Figure 4.16*, where the browser illustrates the analysis shown in *Figure 4.15*, complete with the corresponding records from the `fact_MemberPoll` table.

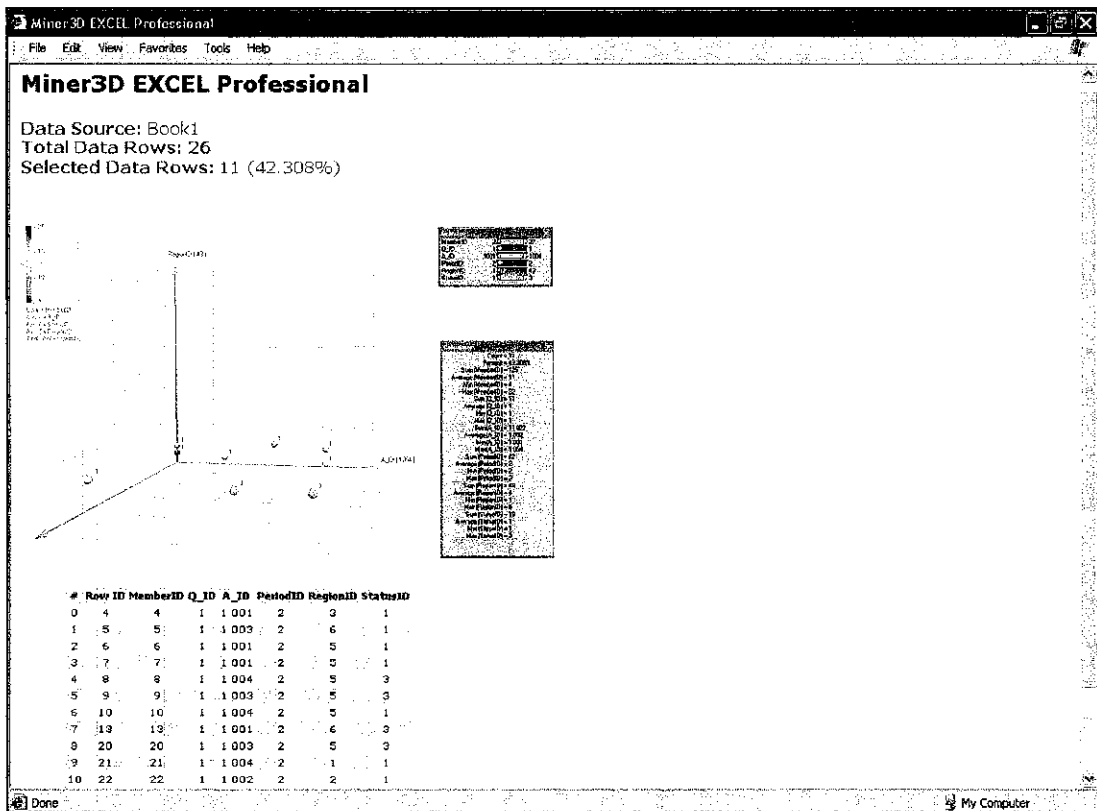


Figure 4.16: Results of Figure 4.15 Analysis (Exported to Web Browser)

## CHAPTER 5

### CONCLUSION

#### 5. OVERVIEW

Data nowadays, in its raw form, may be useless when initial analysis of it is being made. Only proper segregation and classification will allow the data to become useful for system owners and managers to perform analysis on them. This will thus result in better understanding of the collected data and will allow system owners to plan their next strategic business move accordingly.

The way to accomplish all this is via implementation of the *Data Mining* techniques onto the system. Such a technique has been used widely across many business platforms in the previous years and has allowed many organizations to enjoy benefits from its data warehousing enhancements. It saves time for uncovering data trends or patterns from the information gathered and such a technique involves analyzing vast amount of data at a time.

In this project, I examine the use of *Data Mining* on an e-Poll system stationed in an online store. The Knowledge Discovery form on Data Mining enables owners to explore and find new information about the store members and their behavior or opinions via a bottom-up method.

#### 5.1 Relevancy to the Objectives

In relation to the objectives stated earlier in this document, the author has identified the following:

- Data Mining allows system owners to (1) transform raw data into useful data, (2) save time in uncovering data trends, (3) analyze vast amounts of data at a time, and (4) plan their next strategic move accordingly, based on analysis that have been made. .
  
- Star Schema is the best data modeling technique compared to ERD as it allows for (1) superior or optimal decision support query performance in relational databases as it eliminates needs for double-joins of tables, (2) greater understandability of the relationship among the variables in the database, and (3) no loss of information, as an ERD can be represented as a Star Schema.
  
- e-Poll system in an Online Store allows the correlation between a store members' information and their feedback/answer to a poll, thus allowing system owners and managers a valuable 'gateway' to gather potentially important information from all or a sample of their customers, regardless of geographical boundaries, distance and/or time.

## **5.2 Suggested Future Work for Expansion and Continuation**

The title suggests many other improvements that could be covered if more time and resources could be given by any student in the future.

- *Data Mining in Customer Relationship Management (CRM) -*

To predict customer buying habits and/or patterns, Data Mining for CRM can be performed. In addition to findings patterns of customer feedback (from the e-Poll system), this new feature extension will allow owners to establish a returning and 'running' customers list, based on the product preference and buying habits of the customers.

This will result in a state of preservation of customer loyalty – whereby if the system were to provide listings of what customers might potentially be interested to buy, it

would thus cut short the amount of time a customer has to peruse to find what they need.

In addition, preservation of the company-customer relationship can be done by identifying reasons of customer satisfaction decline and rectifying the situation. The example here is when any 'favorite' items purchased by many of the customers are back-ordered by the company; many would not get their delivery in a timely manner. Perhaps upon noticing this slight, the business owners may avert any comparable situations in the future, thus upholding the trust their customers placed when they made their purchases.

## REFERENCES

### BOOKS

Labovitz, Mark L. (2003) *Business Driven Information Technology: Answers to 100 Critical Questions for Every Manager*, Question #21: What is Data Mining and What Are Its Uses?, Stanford University Press

Ramakrishnan, R. and Gehrke, J (2000) *Database Management Systems 2<sup>nd</sup> Ed.*, Singapore, McGraw Hill International Editions Computer Science Series

Riordan, D.G. and Pauley, S.E. (1999) *Technical Report Writing Today 7<sup>th</sup> Ed.*, Boston, Houghton Mifflin Company

Schwalbe, K. (2000) *Information Technology Project Management*, Thomson Learning, Course Technology

Sekaran, U. (2003) *Research Methods for Business – A Skill Building Approach 4<sup>th</sup> Ed.*, Singapore, John Wiley & Sons (Asia) Pte. Ltd.

Shelley, G.B., Cashman, T.J., and Rosenblatt, H.J. (2001) *Systems Analysis and Design 4<sup>th</sup> Ed.*, Cambridge, Course Technology

Siegel, J. and Shim, Jae (2003) *Database Management Systems - A Handbook for Managers and Their Advisors*, Mason OH, Thomson Learning

Silberschatz, A., Korth, H. F. and Sudarshan, S. (2002) *Database System Concepts 4<sup>th</sup> Ed.*, New York, McGraw Hill

Sommerville, I. (2001) *Software Engineering 6<sup>th</sup> Ed.*, Harlow, Addison-Wesley

Whitten, J. L., Bentley, L. D. and Dittman, K. D. (2000) *Systems Analysis and Design Methods 5<sup>th</sup> Ed.*, Indiana, McGraw-Hill

### INTERNET

Dholakia, Paul M. and Morwitz, Vicki G. (2002) *How Surveys Influence Customers* <[http://www.d.umn.edu/~rvaidyan/mba8211/How\\_Surveys\\_Influence\\_Customers.pdf](http://www.d.umn.edu/~rvaidyan/mba8211/How_Surveys_Influence_Customers.pdf)>

- Dillman, Don A., Tortora, Robert D. and Bowker, Dennis (5 March 1999) *Principles for Conducting Web Surveys*  
<<http://survey.sesrc.wsu.edu/dillman/papers/websurveyppr.pdf>>
- Firestone, Dr. Joseph M. (23 October 2000) *Dimensional Modeling and E-R Modeling in the DW* <<http://www.hpcwire.com/dsstar/00/1017/102300.html>>
- Gannon, Joseph (1 July 2001) *Creating Online Polls*  
<<http://www-106.ibm.com/developerworks/usability/library/us-polls/>>
- Gunderloy, Mike (10 March 2004) *Introduction to SQL Server 2000 Reporting Services* <<http://www.developer.com/db/article.php/3323401>>
- SPSS Inc. (2004) <<http://www.spss.com>>
- StatSoft Inc. (2003a) *Data Mining Techniques*  
<<http://www.statsoft.com/textbook/stathome.html>>
- StatSoft Inc. (2003b) *Data Mining with STATISTICA Data Miner in Marketing*  
<[http://www.statsoft.com/support/whitepapers/pdf/STATISTICA\\_DataMiner\\_marketing.pdf](http://www.statsoft.com/support/whitepapers/pdf/STATISTICA_DataMiner_marketing.pdf)>
- Ventana Research (2004a) *SAS Enterprise Miner*  
<<http://www.ventanaresearch.com/research>>
- Ventana Research (2004b) *Microsoft SQL Server 2000 – Analysis Services*  
<<http://www.ventanaresearch.com/research>>
- Websurveyor Corporation (2003) *Comparison of Traditional versus Online Survey Methods* <<http://www.websurveyor.com/pdf/webvsmail.pdf>>
- Websurveyor Corporation (2002) *How to Conduct Effective Online Surveys*  
<<http://www.websurveyor.com/pdf/howto.pdf>>

---

# APPENDICES

---



# **APPENDIX A**

---

## Data Warehouse and Data Mining Terms

---

## DATA WAREHOUSE & DATA MINING TERMS

### **Cookie**

Information that websites place on users' hard drives to identify users and record their usage patterns.

### **Data mart**

A repository of data that serves a particular community of knowledge workers. The data may come from an enterprise-wide database or a data warehouse.

### **Data mining**

The practice of extracting data from a data warehouse in order to analyze patterns, trends and relationships.

### **Data modeling**

The practice of analyzing an enterprise's data and identifying the relationships among the data.

### **Data scrubbing**

The practice of monitoring a data warehouse and removing data that is not trustworthy or timely.

### **Data warehouse**

A database that stores large amounts of historical business data.

### **Enterprise relationship management (ERM)**

The practice of analyzing customer data from sales, marketing, service, finance and manufacturing databases in order to relate efficiently to customers.

### **Intelligent agent**

A program that automatically performs a service, such as gathering specific information, or that personalizes information on a Web site based on a user's registration information and usage analysis.

### **Metadata**

Information (such as origin, collection criteria, formatting, categorization) about formatted information, or data.

### **Packet**

A chunk of data sent over a network.

### **Replication**

The process of making a copy of something. When using a groupware product, replication means copying a database from one server to another so that all users share the same information.

### **Search engine**

A program that delivers to users information and website addresses that relate to words they entered into the program's interface.

**Storage Area Network (SAN)**

A network built exclusively for storage devices, servers, backup systems and so forth. SANs can handle heavy bandwidth demands of storage data and segregate storage traffic to a network built specifically for storage needs.

**Structured Query Language (SQL)**

A programming language pronounced “sequel” that builds applications that move information in and out of databases.

**Third generation wireless (3G)**

A group of wireless technologies that move from circuit-switched communications to wireless broadband, high-speed, packet-based networks. These are preceded by first generation analog and second-generation digital communication technologies.

# **APPENDIX B**

---

Complete Listing of  
Database Schema  
(Dimensional Model)

---

## COMPLETE LISTING OF DATABASE SCHEMA (Dimensional Model)

*e-Poll* –

Table Name	Column Name	Data Type	Null
dim_PollQuestion	Q_ID	VARCHAR2 (3)	N
	Question	VARCHAR2 (75)	N
	EntryDate	DATE	N
	ActiveDate	DATE	N
	ExpireDate	DATE	Y
	TotalVotes	NUMBER	Y
dim_PollAnswer	A_ID	VARCHAR2 (4)	N
	Q_ID	VARCHAR2 (3)	N
	Answer	VARCHAR2 (25)	Y
	VoteCount	NUMBER	Y
	Percentage	NUMBER (6, 2)	Y
dim_PollArchiveQ	Q_ID	VARCHAR2 (3)	N
	Question	VARCHAR2 (75)	N
	EntryDate	DATE	N
	ActiveDate	DATE	N
	ExpireDate	DATE	Y
	TotalVotes	NUMBER	Y
dim_PollArchiveA	A_ID	VARCHAR2 (4)	N
	Q_ID	VARCHAR2 (3)	N
	Answer	VARCHAR2 (25)	Y
	VoteCount	NUMBER	Y
	Percentage	NUMBER (6, 2)	Y
dim_PollComment	C_ID	VARCHAR2 (4)	N
	Q_ID	VARCHAR2 (3)	N
	Name	VARCHAR2 (25)	N
	Subject	VARCHAR2 (40)	Y
	Comment	VARCHAR2 (300)	N

*Online Store Member's Personal Information –*

Table Name	Column Name	Data Type	Null
dim_Member	MemberID	NUMBER	N
	MembershipDate	DATE	N
	Name	VARCHAR2 (75)	N
	Title	VARCHAR2 (5)	N
	Email	VARCHAR2 (25)	N
	Login	VARCHAR2 (25)	N
	Age	NUMBER	N
	Gender	VARCHAR2 (2)	N
	City	VARCHAR2 (30)	N
State	VARCHAR2 (15)	N	
dim_Period	PeriodID	NUMBER	N
	PeriodDay	NUMBER	N
	PeriodMonth	NUMBER	N
	PeriodQuarter	NUMBER	N
	PeriodYear	NUMBER	N
dim_Region	RegionID	NUMBER	N
	RegionDesc	VARCHAR2 (10)	N
	RegionCity	VARCHAR2 (30)	N
	RegionState	VARCHAR2 (15)	N
dim_Status	StatusID	NUMBER	N
	Description	VARCHAR2 (25)	N
dim_MemberShipping	ShippingID	NUMBER	N
	MemberID	NUMBER	N
	ShippingAddress	VARCHAR2 (200)	N
	ShippingPostcode	NUMBER	N
	ShippingCity	VARCHAR2 (30)	N
	ShippingState	VARCHAR2 (15)	N
dim_MemberCC	CCID	NUMBER	N
	MemberID	NUMBER	N
	CCtype	VARCHAR2 (10)	N
	CCnumber	NUMBER	N
	CCexp_mo	NUMBER	N
	CCexp_yr	NUMBER	N
	CCname	VARCHAR2 (75)	N
dim_MemberLogin	Login	VARCHAR2 (25)	N
	Password	VARCHAR2 (25)	N
dim_Group	Group	VARCHAR2 (10)	N
	Privilege	VARCHAR2 (40)	N
dim_MemberGroup	Login	VARCHAR2 (25)	N
	Group	VARCHAR2 (10)	N
dim_Item	ItemID	VARCHAR2 (10)	N
	Description	VARCHAR2 (25)	Y
	UnitPrice	NUMBER	N
dim_Transaction	TransID	NUMBER	N
	MemberID	NUMBER	N
	Date	DATE	N
	ItemID	VARCHAR2 (10)	N
	Qty	NUMBER	N

*Store Member and Poll Information –*

Table Name	Column Name	Data Type	Null
fact_MemberPoll	MemberID	NUMBER	N
	Q_ID	VARCHAR2 (3)	N
	A_ID	VARCHAR2 (4)	N
	PeriodID	NUMBER	N
	RegionID	NUMBER	N
	StatusID	NUMBER	N

# APPENDIX C

---

Star Schema  
Dimensions Values

---



## STAR SCHEMA DIMENSIONS VALUES

*dim\_Region* –

RegionID	RegionDesc	RegionCity	RegionState
1	Utara	Arasu	Perlis
2	Utara	Alor Setar	Kedah
3	Utara	Sg Patani	Kedah
4	Utara	Grik	Kedah
5	Utara	Igoh	Perak
6	Utara	Kuala Kangsar	Perak
7	Utara	Tekuk Intan	Perak
8	Utara	Taiping	Perak
9	Timur	Pasir Mas	Kelantan
10	Timur	Kota Bharu	Kelantan
11	Timur	Kuala Terengganu	Terengganu
12	Timur	Kemaman	Terengganu
13	Timur	Kuantan	Pahang
14	Timur	Teluk Chempedak	Pahang
15	Selatan	Johor Bahru	Johor
16	Selatan	Segamat	Johor
17	Selatan	Muar	Johor
18	Selatan	Pasir Gudang	Johor
19	Selatan	Alor Gajah	Melaka
20	Selatan	Bdr Melaka	Melaka
21	Selatan	Jasin	Melaka
22	Selatan	Seremban	N. Sembilan
23	Selatan	Nilai	N. Sembilan
24	Selatan	Gemencheh	N. Sembilan
25	Central	Kuala Lumpur	WPKL
26	Central	Putrajaya	WPPJ
27	Central	Cyberjaya	Selangor
28	Central	Petaling Jaya	Selangor
29	Central	Subang Jaya	Selangor
30	Central	Shah Alam	Selangor
31	Central	Selayang	Selangor
32	Central	Rawang	Selangor
33	Sabah	Kota Kinabalu	Sabah
34	Sabah	Sandakan	Sabah
35	Sarawak	Miri	Sarawak

*dim\_Status* –

StatusID	Description
1	Single
2	Married
3	Married (with Children)
4	Divorced
5	Separated

*dim\_PollAnswer*\*\* –

A_ID	Q_ID	Answer	VoteCount	Percentage
1001	1	Yes, I do	6	21.00%
1002	1	No, I don't	4	14.00%
1003	1	Sometimes I do	5	18.00%
1004	1	Just made my first order!	12	46.00%

\*\* Note: Q\_ID refers to the Question entitled 'When you order our products, do you get prompt delivery?'.