

Guidelines for the Analysis of Student Web Usage in Support of Primary Educational Objectives

by

Dean John von Schoultz

2015

Guidelines for the Analysis of Student Web Usage in Support of Primary Educational Objectives

by

Dean John von Schoultz

Dissertation

Submitted in fulfilment of the requirements for the degree

Magister Technologiae

in

Information Technology

in the

Faculty of Engineering, the Built Environment and Information Technology

of the

Nelson Mandela Metropolitan University

Supervisor: Prof Kerry-Lynn Thomson

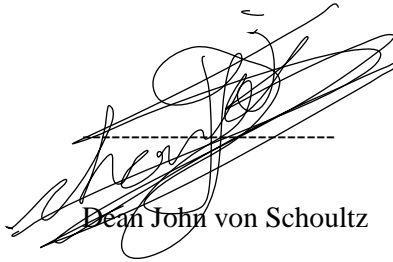
Co-Supervisors: Prof Johan van Niekerk and Dr Mariana Gerber

2015

Declaration

I, Dean John von Schoultz, hereby declare that

- the work in this dissertation is my own work
- all sources used or referred to have been documented and recognised
- this dissertation has not been submitted previously in full or partial fulfilment of the requirements for an equivalent or higher qualification at any other recognised educational institution.



Dean John von Schoultz

Abstract

The Internet and World Wide Web provides huge amounts of information to individuals with access to it. Information is an important driving factor of education and higher education has experienced massive adoption rates of information and communication technologies, and accessing the Web is not an uncommon practice within a higher educational institution. The Web provides numerous benefits and many students rely on the Web for information, communication and technical support. However, the immense amount of information available on the Web has brought about some negative side effects associated with abundant information. Whether the Web is a positive influence on students' academic well-being within higher education is a difficult question to answer. To understand how the Web is used by students within a higher education institution is not an easy task. However, there are ways to understand the Web usage behaviour of students. Using established methods for gathering useful information from data produced by an institution, Web usage behaviours of students within a higher education institution could be analysed and presented.

This dissertation presents guidance for analysing Web traffic within a higher educational institution in order to gain insight into the Web usage behaviours of students. This insight can provide educators with valuable information to bolster their decision-making capacity towards achieving their educational goals.

Acknowledgements

My sincerest gratitude goes to the following:

- My supervisor Prof Kerry-Lynn Thomson, for her impeccable guidance, consistently vibrant smile and warm presence which made the past 2 years of research a great privilege.
- My co-supervisor Prof Johan van Niekerk, whose enthusiasm for education, contagious zeal and distinct perspective has been truly encouraging and invaluable. In my experience, his tutelage is unrivalled.
- My co-supervisor Dr Mariana Gerber for her interest and assistance in my research. Her expertise and extensive research knowledge was greatly appreciated.
- My mother, who has been a nexus of inspiration and the essence of my persistence throughout my studies. Words cannot extend the gratitude I feel.
- My girlfriend, for her patience, encouragement and unwavering affection during this endeavour.
- My fellow student Jacque Coertze, for our regular discussions that provided crucial clarity, direction and ingenuity.
- The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the National Research Foundation.

Table of Contents

Declaration	i
Abstract	ii
Acknowledgements	iii
Chapter 1 - Introduction	1
1.1 Background.....	1
1.2 Problem Statement	2
1.3 Research Objectives.....	3
1.3.1 Primary Research Objective.....	3
1.3.2 Secondary Research Objectives	3
1.4 Research Design.....	3
1.4.1 Research Paradigm.....	4
1.4.2 Research Process.....	4
1.4.3 Research Methods	7
1.5 Scope and Delineation	9
1.6 Ethics and Permissions	10
1.7 Conclusion	10
Chapter 2 - Higher Education Institution Information Needs	11
2.1 Introduction.....	12
2.2 Information Access and Education	14
2.2.1 Information Access Paradigm Shift	14
2.2.2 Information Overload.....	15
2.3 - Primary Objectives of Education.....	18
2.3.1 Primary Objectives of Higher Education	18
2.3.2 Educators' Information Needs	22
2.4 Conclusion	25
Chapter 3 - Organisational Data and Web Usage Analysis	26
3.1 Introduction.....	27
3.2 Use and Purpose of Log Files	27
3.2.1 Log files in general	27
3.2.2 Networks and Web Logs.....	28
3.3 NMMU Web Log Collection and Comparison.....	33
3.4 Conclusion	40

Chapter 4 - Business Intelligence Dynamics	42
4.1 Introduction	43
4.2 Business Intelligence	43
4.2.1 Brief History of Business Intelligence	44
4.2.2 Business Intelligence Acquisition	45
4.3 BI System Development Method Argument and Discussion	59
4.3.1 Program/Project Planning	61
4.3.2 Program/Project Management	62
4.3.3 Business Requirements Definition	62
4.3.4 Technology Track — Technical Architecture Design and Product Selection and Installa....	63
4.3.5 Data Track – Dimensional Modelling, Physical Design and ETL Design and Developme.	64
4.3.6 BI Application Track – BI Application Design and BI Application Development	65
4.3.7 Deployment	66
4.3.8 Maintenance	66
4.3.9 Growth	66
4.4 Conclusion	66
Chapter 5 - SWAN Prototype Data Mart	68
5.1 Introduction	69
5.2 Program/Project Planning and Program/Project Management	70
5.2.1 Program/Project Planning	70
5.2.2 Program/Project Management	72
5.3 Business Requirements Definition	72
5.4 Kimball Lifecycle Tracks	74
5.4.1 Technology Track	74
5.4.2 Data Track	79
5.4.3 BI Application Track	83
5.5 Conclusion	86
Chapter 6 - SWAN Guidelines	87
6.1 Introduction	88
6.2 The SWAN Project Case and Context	89
6.3 Case Issue Isolation and Assertions	91
6.3.1 Data format and availability	92
6.3.2 Understanding the Business Process and Gathering Users' Information Needs	98
6.3.3 Designing a Dimensional Model	103
6.3.4 Manual Extract, Transformation and Load Process	112

6.3.5 Refining the Development Process and Maintaining Scope	124
6.4 Guideline verification	125
6.4.1 Expert Review Validation.....	126
6.5 Conclusion	131
Chapter 7 - Conclusion	133
7.1 Summary	134
7.2 Satisfying the Research Objectives.....	135
7.2.1 Primary Research Objective.....	135
7.2.2 Secondary Research Objectives	135
7.3 Revisiting the Research Problem	137
7.4 DSR Knowledge Contribution.....	138
7.5 Limitations	139
7.6 Publications Stemming from this Research	139
7.7 Future Research	139
7.8 Closing Remark	140
Bibliography	141
Appendices	148
Appendix A – Email liaison between associate professor and Director of ICT	151
Appendix B – Online survey form.....	153
Appendix C – System engineer interview.....	161
Appendix D – SWAN Project scope charter.....	164
Appendix E – Guideline validation review form	166
Appendix F – Guideline validation expert review responses.....	188
Appendix G – Web usage mining within a South African university infrastructure.....	193

List of Tables

Table 3.1 - Pre configuration error correction Fortigate firewall log entry example.....	35
Table 3.2 - Squid log example (Hossain, Rahman, & Kabir, 2012)	36
Table 3.3 - Post configuration error correction Fortigate firewall log entry example	37
Table 3.4 - Log files entry field comparison.....	39
Table 6.1 - Firewall log entry example (Von Schoultz et al., 2013).....	94

List of Figures

Figure 1.1 - Structure of dissertation	8
Figure 1.2 - Design science research process, research methods and chapter relationship	9
Figure 3.1 - MIMIC Architecture (Anand, Büchner, Mulvenna, & Hughes, 1999)	30
Figure 3.2 - A normal HTTP transaction (Luotonen & Altis, 1994)	31
Figure 3.3 - Overall setup of a proxy (Luotonen & Altis, 1994)	32
Figure 3.4 - A proxied HTTP transaction (Luotonen & Altis, 1994).....	33
Figure 3.5 - NMMU's logical network (Von Schoultz et al., 2013)	34
Figure 3.6 - Web log population	36
Figure 4.1 - A typical business intelligence system architecture adapted from Imhoff et al. (2003) and Watson & Wixom (2007).....	46
Figure 4.2 - Generic waterfall methodology lifecycle adapted from Ragunath (2010) and Rainardi (2008).....	50
Figure 4.3 - Iterative cycles (Rainardi, 2008)	51
Figure 4.4 - The structure of a top-down data warehouse (Inmon, 2002).....	55
Figure 4.5 - Data warehouse SDLC (top-down) (Inmon, 2002)	56
Figure 4.6 - The structure of a bottom-up data warehouse adapted from Golfarelli & Rizzi (2009) ..	57
Figure 4.7 - Data warehouse SDLC (bottom-up) adapted from Golfarelli & Rizzi (2009)	58
Figure 4.8 - The Kimball Lifecycle diagram (R Kimball et al., 2008)	61
Figure 4.9 - The Kimball Lifecycle tracks adapted from Kimball et al. (2008)	63
Figure 5.1 - Program/Project Planning and Program/Project Management milestones adapted from Kimball et al. (2008)	70
Figure 5.2 - Business Requirements Definition milestone adapted from Kimball et al. (2008)	73
Figure 5.3 - Technology Track Milestones adapted from Kimball et al. (2008)	75
Figure 5.4 - High level technical architecture model for SWAN adapted from Kimball et al. (2008) and Watson & Wixom (2007).....	76
Figure 5.5 - IT Infrastructure for SWAN.....	78
Figure 5.6 - Data Track milestones adapted from Kimball et al. (2008)	79
Figure 5.7 - SWAN Data Mart star schema	81
Figure 5.8 - SWAN Data Mart dimensional star schema diagram generated from SQL Server	82
Figure 5.9 - BI Application Track milestones adapted from Kimball et al. (2008)	83
Figure 5.10 - SQL Query used to produce a result set from the fact table.....	84
Figure 5.11 - SWAN result set.....	84
Figure 5.12 - Deployment, Growth and Maintenance adapted from Kimball et al. (2008)	85
Figure 6.1 - NMMU's logical network (Von Schoultz et al., 2013)	93
Figure 6.2 - Original sample log file.....	94

Figure 6.3 - Sample report extracted from the SWAN Data Mart	101
Figure 6.4 - Draft star schema.....	105
Figure 6.5 - Amended star schema.....	107
Figure 6.6 - Updated star schema.....	110
Figure 6.7 - Manual ETL process for the SWAN Data Mart.....	113
Figure 6.8 - Log file	115
Figure 6.9 - Entry with column divergence	116
Figure 6.10 - Entry split by “=”	116
Figure 6.11 - Entry split by a single space	117
Figure 6.12 - Column divergence	118
Figure 6.13 - SQL staging table import issue	119
Figure 6.14 - Data import in Excel	120
Figure 6.15 - Excel log entry Field=Value search formula.....	121
Figure 6.16 - Excel formula demonstrations.....	121
Figure 6.17 - Website dimensional load query	123
Figure 7.1 - DSRP Knowledge Contribution Framework (Grego & Hevner, 2013).....	138

List of Abbreviations

ARPA:	Advanced Research Project Agency
BI:	Business Intelligence
CIO:	Chief Information Officer
DS:	Design Science
DSS:	Decision Support System
DW/BI:	Data Warehouse/Business Intelligence
E-commerce:	Electronic Commerce
EIS:	Enterprise Information System
ETL:	Extract, Transformation, Load
ICT:	Information and Communication Technology
IDS:	Intrusion Detection System
IP:	Internet Protocol
IPS:	Intrusion Prevention System
IS:	Information Systems
IT:	Information Technology
KL:	Kimball Lifecycle
MIS:	Management Information System
NMMU:	Nelson Mandela Metropolitan University
SDLC:	Software Development Lifecycle
SQL:	Structured Query Language
SWAN:	Student Web Usage Analysis
TCP:	Transmission Control Protocol
WWW:	World Wide Web

Glossary of Educational Terms

The following terms have been defined according to the HIGHER EDUCATION ACT 101 OF 1997 (HEAA Council, 2001)

<i>academic employee:</i>	any person appointed to teach or to do research at a public higher education institution and any other employee designated as such by the council of that institution
<i>Department of Education:</i>	the Government department responsible for education at national level
<i>higher education:</i>	all learning programmes leading to qualifications higher than Grade 12 or its equivalent in terms of the National Qualifications Framework as contemplated in the South African Qualifications Authority Act, 1995 (Act 58 of 1995), and including tertiary education as contemplated in Schedule 4 of the Constitution
<i>higher education institution:</i>	<p>any institution that provides higher education on a full-time, part-time or distance basis and which is</p> <ul style="list-style-type: none">• established or deemed to be established as a public higher education institution under this Act• declared as a public higher education institution under this Act or• registered or conditionally registered as a private higher education institution under this Act
<i>private higher education institution:</i>	any institution registered or conditionally registered as a private higher education institution
<i>public higher education institution:</i>	any higher education institution that is established, deemed to be established or declared as a public higher education institution under this Act
<i>student:</i>	any person registered as a student at a higher education institution
<i>to provide higher education:</i>	(a) the registering of students for

(i) complete qualifications at or above level 5 of the National Qualification Framework as contemplated in the South African Qualifications Authority Act, 1995 (Act 58 of 1995), or

(ii) such part of a qualification which meets the requirements of a unit standard as recognised by the South African Qualifications Authority at or above the level referred to in subparagraph (i)

(b) the taking of responsibility for the provision and delivery of the curriculum

(c) the assessment of students' performance in their learning programmes

(d) the conferring of qualifications, in the name of the higher education institution concerned [definition of "to provide higher education" inserted by s. 1 (c) of Act 54 of 2000]

Chapter 1 - Introduction

“What is research but a blind date with knowledge.” – Will Harvey

This chapter introduces the research problem, research objectives and research design of this dissertation.

1.1 Background

Information and communication technology (ICT) is embedded in our daily lives. It is used in a personal and business capacity. Information exchange has become an effortless task and information is in abundance. Many organisations have experienced invaluable benefits from ICT. These organisations have ICT services that facilitate and automate certain internal processes. Whether an organisation is a Fortune 500 company, a primary school, a non-profit charity organisation or a public service, it is likely to have business practices requiring the processing of data and information. As a result, ICT has spread throughout the globe and into many areas of human activity (Torero & Braun, 2006).

The establishment of the World Wide Web was an information technology (IT) breakthrough and a catalyst for the accelerated rate of information exchange. The Web has bridged communication and information exchange gaps on a global scale. It has ushered in online business and electronic commerce (e-commerce), vast mediums of communication (email, social networking and video conferencing), efficient information gathering (search engines) and online education distribution (e-learning). The benefits of the Web are widely recognised and have resulted in increased rates of use and adoption. As a result, Web access is available in many organisations (Mehrtens, Cragg, & Mills, 2001). Educational institutions, especially those in higher education, have experienced high IT and Web adoption rates (Kukulska-Hulme, 2012; Njenga & Fourie, 2010).

Higher education institutions seek to provide tertiary education to students and information exchange is a central aspect of this provision. With the presence of the Web in higher education institutions, the ways in which information is obtained and exchanged within these institutions have grown significantly. For the facilitators of higher education such as lecturers, management and other role players, the information exchange behaviour of students can be an important factor in achieving their educational objectives. However, considering the nature of the Web, understanding how students manage information and identifying the effects of the Web within an educational environment is a huge task.

Various studies have been conducted to determine whether an abundance of information is detrimental to the academic environment or whether it provides the ideal means for educational achievement. For example, Wallace (2014) investigated how "Internet addiction" is affecting students' academic well-being. Hazelhurst, Johnson and Sanders (2011) analysed Web traffic within a university and correlated the results with the academic performance of students in order to examine relationships between certain Web usage behaviours and academic success. Rumbough (2001) uncovered a trend of unsavoury and controversial Web activity within a sample population of university students. In the present research study, the analysis and interpretation of Web usage data are the primary focus; this involves Business Intelligence (BI), a pervasive domain that is concerned with analysing organisational data in order to better serve the processes that produce it.

BI provides methods of converting organisational data into useful information for business users. Certain BI methods are used to understand and control Web usage data. For example, the Web usage behaviour of individuals is of interest to e-commerce organisations; online preferences and shopping trends of customers using the Web are invaluable to marketing agencies and e-commerce website owners. They can use these trends and preferences as leverage to drive sales by tailoring marketing material and predicting purchasing behaviours. This is achieved by analysing clickstream data, which is an account of each click made by a Web user on a website. It is produced by Web activity on browsers and Web servers (Cooley, 2000; Eirinaki & Vazirgiannis, 2003).

Analysing organisational data using BI methods to support decision-making is a common technique within many large companies (Chaudhuri, Dayal, & Narasayya, 2011). However, applying BI methods to students' Web usage data in higher education may require some calibration.

1.2 Problem Statement

From the background provided above it is evident that the implications of the accelerated change in information access within higher education institutions are far-reaching. However, within the Nelson Mandela Metropolitan University (NMMU) the means of determining the nature of the use of the most important source of information, the Web, and the effects thereof are not available.

The problem statement can thus be defined as:

There is currently no facility in place to analyse accurately student Web usage data within the NMMU in order to provide meaningful student Web usage information that will support educators' primary educational objectives.

The research objectives, which contribute collectively to the solution of this problem statement, are discussed in the following section.

1.3 Research Objectives

In order to address the research problem, the following research objectives were established to investigate and develop an appropriate solution.

1.3.1 Primary Research Objective

To provide a comprehensive set of guidelines to assist higher education institutions in the successful analysis of student Web usage data, in support of their decisions regarding primary educational objectives.

1.3.2 Secondary Research Objectives

- To determine what student Web usage data is currently available and its usability and format
- To identify educators' questions about student Web usage, in order to better inform their decisions regarding primary educational objectives
- To isolate best practices and heuristic aspects for applying data analysis methods to Web usage data
- To design, develop, implement and verify a prototype decision support system based on the above findings
- To consolidate a set of guidelines for the analysis of Web usage data within a higher educational context

These objectives will make a contribution to providing a solution to this organisational information system problem and/or reveal any additional requirements. The secondary objectives will collectively achieve the primary objective.

In order to achieve the objectives stated above, a research design needs to be applied.

1.4 Research Design

The problem statement driving this project is derived from an organisational information system issue, namely, the underutilisation of potentially important data within the system; more specifically, the lack of facilitation of these data. The objective of this research study is to design, develop and evaluate a solution to this issue. A sound research design is vital to the success of this objective and ensures that the elements required to solve the problem are determined and correctly presented. This

section discusses the research paradigm and process, the methods used, the scope and delineation, as well as ethics and permissions regarding the study.

1.4.1 Research Paradigm

The information systems (IS) domain exists where information technology (IT) converges with organizations (Hevner & Chatterjee, 2010). This research was conducted to solve an organizational problem, through the creation and analysis of a data mart which delivers useful student Web usage behaviour information. Therefore, this research resides in the IS domain. The nature of this study lends itself to the design science research paradigm (DSRP), which "seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artefacts" (Hevner et al., 2004, p. 75). The DSRP is highly relevant in addressing IS issues as it provides focus on organizational needs through the development of IT artefacts (Hevner & Chatterjee, 2010).

"It is incumbent upon researchers in the Information Systems (IS) discipline to further knowledge that aids in the productive application of information technology to human organizations and their management" (Hevner et al., 2004, p.76).

Design science research in information systems allows for the answering of questions towards the solving of organizational problems through the creation of novel artefacts (Vaishnavi & Kuechler, 2012). In doing so, the problem is further understood while providing a full or partial solution. Essentially, knowledge relevant to a problem domain and context is generated during the design and creation of artefacts, as they are designed to solve a specified problem (Hevner & Chatterjee, 2010).

1.4.2 Research Process

Peppers et al. (2006) define design as "the act of creating an explicitly applicable solution to a problem" (p. 84). The design science research process (DSRP) in IS has received some attention in the past. The need for design science research processes and heuristics has long been recognised by researchers in the field, resulting in a set of guidelines and a methodology process, amongst other contributions.

The research process for design science produced by Peppers et al. (2006) takes into account design approaches from various other relevant research paradigms and fields of study to provide a defined middle ground for IS research and design in general. Moreover, it includes other relevant contributions to design science in IS such as Hevner et al. (2004) and March and Smith (1995). These steps are applicable to the research at hand and can be implemented at any of the initial steps, at the

discretion of the researcher and determined by the nature of the study. The design science research process (DSRP) is used in this study.

The DSRP model (Peppers et al., 2006) includes six general steps:

1. Problem identification and motivation

Define problem and show importance.

2. Objectives of a solution

What would a better artefact accomplish?

3. Design and development

Develop artefact(s).

4. Demonstration

Find suitable context and use artefact to solve problem.

5. Evaluation

Observe how effective and efficient artefact(s) are, iterate back to design if necessary.

6. Communication

Generate scholarly publications and/or professional publications.

An overview of each step and how it is applied in this research follows below:

1. *Problem identification and motivation*

An investigation of literature dealing with the areas of BI, data warehousing and Web usage data analysis within academic institutions was conducted. This partially determined the current state of knowledge regarding the importance of the problem identified and the feasibility of a solution for the scope of this dissertation. Web use in higher education institutions, including the educational objectives within these institutions, and previous research undertaken will provide a broad context for this study.

2. *Objectives of a solution*

A survey targeting IT lecturers was conducted to investigate the need for information about the Web usage behaviour of their students. Furthermore, this survey allowed for the identification of a candidate for a face to face interview. This interview was conducted to refine the understanding of the needs of IT lecturers, thereby, providing exact requirements for the decision support system. Literature was consulted to determine what was required from the data if a decision support system was to support decision-making by IT lecturers. The focus was on determining the processes required for the creation, design and implementation of an organisation-specific decision support system in support of primary educational objectives. In addition, analysis techniques that were applicable to the available Web usage data were investigated. A BI system development method appropriate to the development of the prototype decision support system was selected by comparing established BI methods.

3. Design and development

The BI method selected in the previous step was used to develop a prototype data mart using NMMU Web usage data to produce useful Web behaviour information from a sample of students. An expert review was conducted during this development to refine a star schema, which is a central aspect of the BI method. A study of this prototype and its development was used to design and develop a set of guidelines that formed the delivered artefact.. This was achieved by consolidating assertions made from an analysis of the developer's account of the prototype development and associated documents, as well as research conducted prior to the prototype development. The prototype data mart serves as a platform for the design of the set of guidelines, which is the IT artefact delivered by this research. Therefore, the BI method used to develop the prototype is inherent in the IT artefact. However, it does not have any other bearing on the DSRP.

4. Demonstration

The context of the application of the guidelines was the NMMU. The guidelines were demonstrated during the development of the prototype from which these guidelines were derived. In other words, by developing the guidelines that provide guidance for a process, it was demonstrated that the process was effective and that the guidelines were a concise version of that successful process.

5. Evaluation

The set of guidelines, that is, the IT artefact created in this DSRP, were validated through expert review. The desired characteristics of an IT artefact, as well as the traits of ideal guidelines, were used to design the metrics for evaluation. The review resulted in some minor changes and consensus that they met the desired standard.

6. Communication

The guidelines, prototype data mart, and any relevant findings will be presented to the academic and appropriate organisational institutions.

In summary, this design process was selected because of its scope and relevance to this project. The research methods applicable to the above mentioned steps and their implementation in the project are discussed below.

1.4.3 Research Methods

The applicable method(s) applied in each research process step in the DSRP are listed below:

1. A *literature review* of Web usage analytics, business intelligence and data warehousing was conducted, and is discussed in Chapter 2. In addition, a *survey* was conducted by means of an online survey form, to identify potential users of information about Web usage behaviour, as discussed in Section 2.3.2.
2. The data required and its location was determined through the *literature review* and previous research and is discussed in Chapter 3. A *literature review* of the prerequisites for a decision support system and its development was conducted and is discussed in Chapter 4.
3. *Prototype development* was used to demonstrate how Web usage data could be analysed to provide meaningful information in an educational context (see Chapter 5). In the course of this development process, an *expert review* was undertaken to refine the data structuring technique required for the development, as discussed in Section 5.4.2. An *interview* was conducted to understand the information needs of an IT lecturer and this was used as input for some of the developmental steps of the prototype (see Section 5.3). The results from the survey mentioned above in step 1 were used to identify the interview candidate. A set of guidelines for the analysis of student Web usage in support of primary educational objectives is designed, and this is discussed in detail in Chapter 6.
4. The development of the *prototype* from which the guidelines were developed proved successful and therefore demonstrated the use of the guidelines, as discussed in Chapter 5 and Chapter 6.
5. An *expert review* was used to validate the guidelines (see Section 6.4.1).
6. This dissertation and scholarly publications will be used to present the findings of this research study.

Figure 1.1 illustrates the structure of this dissertation:

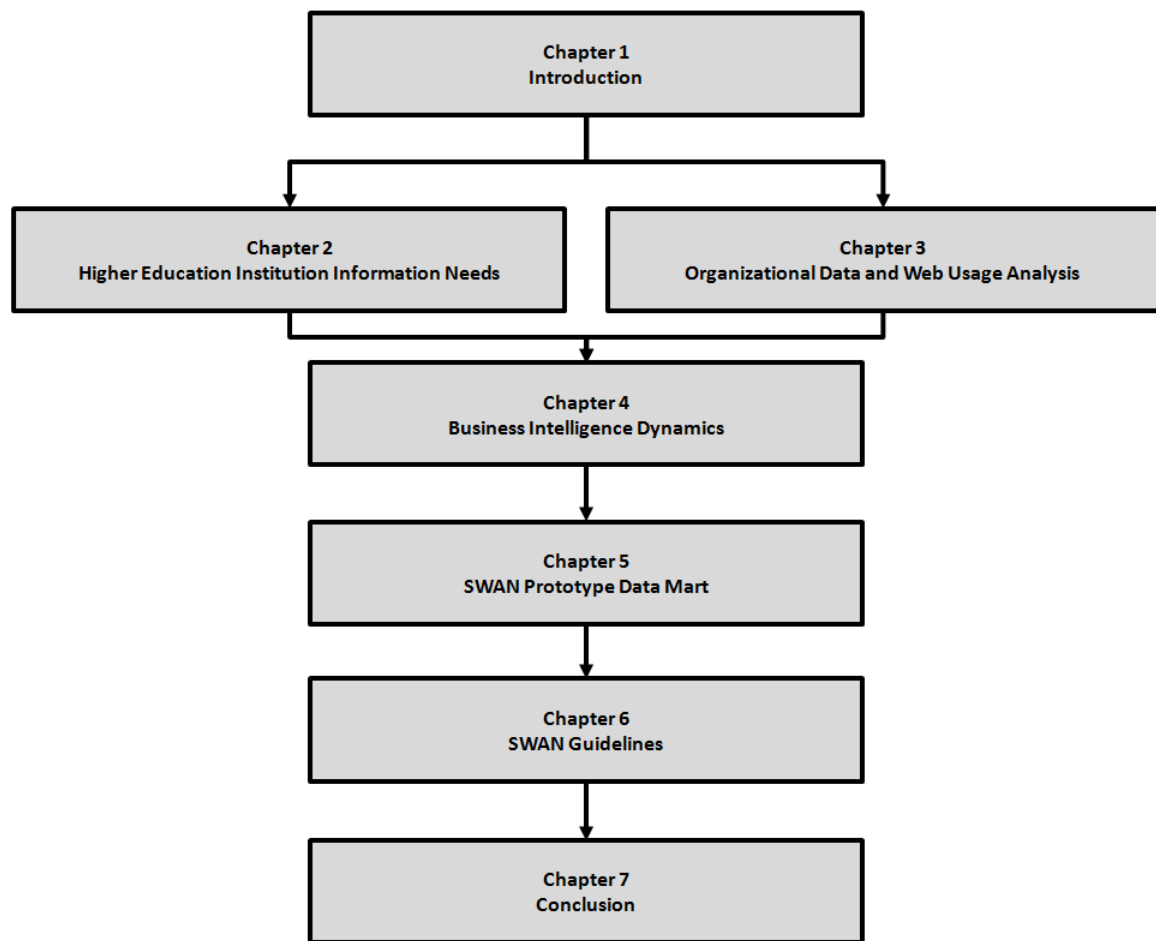


Figure 1.1 – Structure of dissertation

Figure 1.2 indicates the chapters in which the research process steps are discussed and the relationship of these steps to specific research methods:

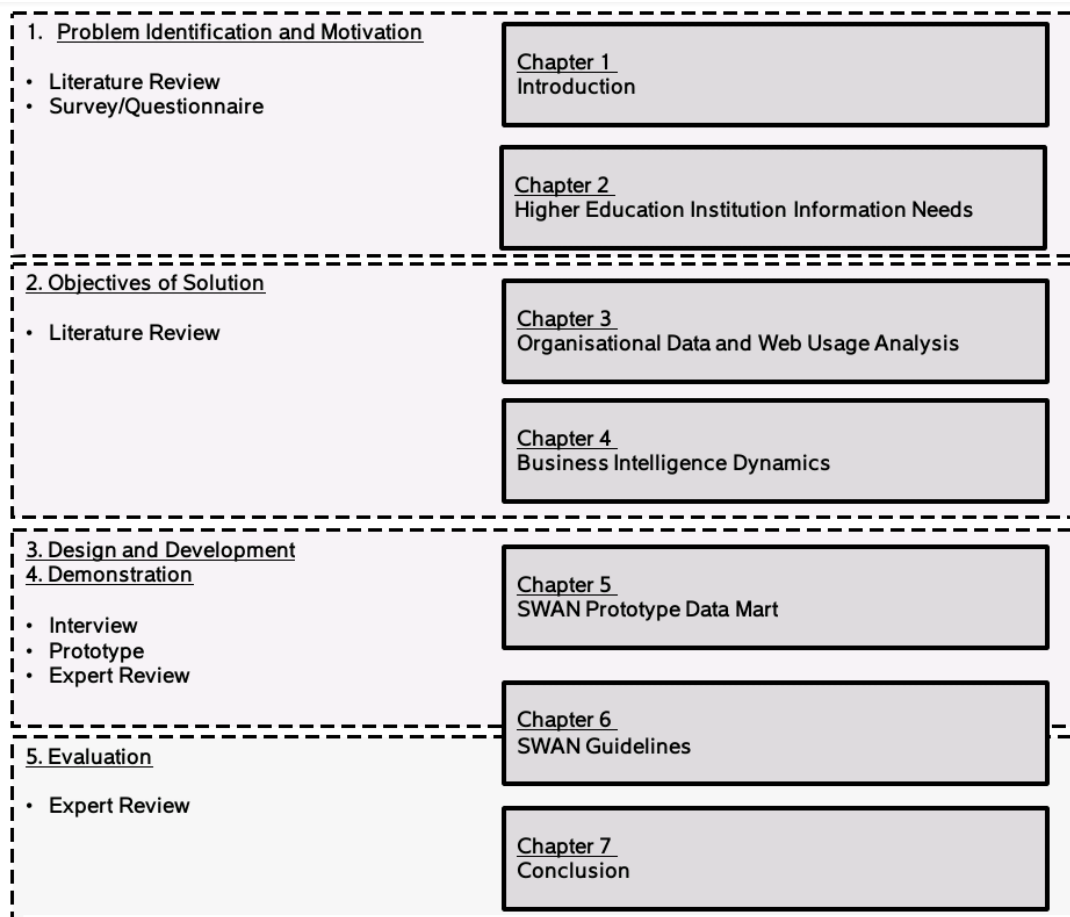


Figure 1.2 - Design science research process, research methods and chapter relationship

The process of this research study and the appropriate research methods has been established in the sections above. In summary, the literature review and survey allowed for the establishment of specifics for the development of a data mart which provides useful Web usage behaviour information. A data mart was then developed and, knowledge gained during its development, was used to design an IT artefact in the form of a set of guidelines. The scope and delineations of the study are discussed in the section which follows.

1.5 Scope and Delineation

The artefact produced by this research was designed specifically for the NMMU context and focuses on the School of Information and Communication Technology (ICT). The prototype used to derive and produce the artefact was limited to one business process and a single, select data set was used to answer selected questions based on the intended user of the information produced by this prototype.

1.6 Ethics and Permissions

The data collected for the prototype data mart comprised log files from the firewall that caches network traffic from all six campuses, i.e. it was a direct, raw reflection of each entry of a student's network Web activity. No individuals were identified, all details were kept confidential and users will remain anonymous in any findings presented in the communication step. Permission to use these logs was obtained from the Director of ICT; the email correspondence is attached as Appendix A. Furthermore, an interview with the system engineer who provided the logs confirmed that all logs were anonymous and that no individuals could be identified. In addition, the anonymising of the software was completed before the initial approach to the Director of ICT; the interview transcript is attached as Appendix C. The Director was informed of this anonymising software and the fact that this meant that the researcher had complied with the Directors request to ensure anonymity. Therefore, ethical clearance was not required for this dissertation.

1.7 Conclusion

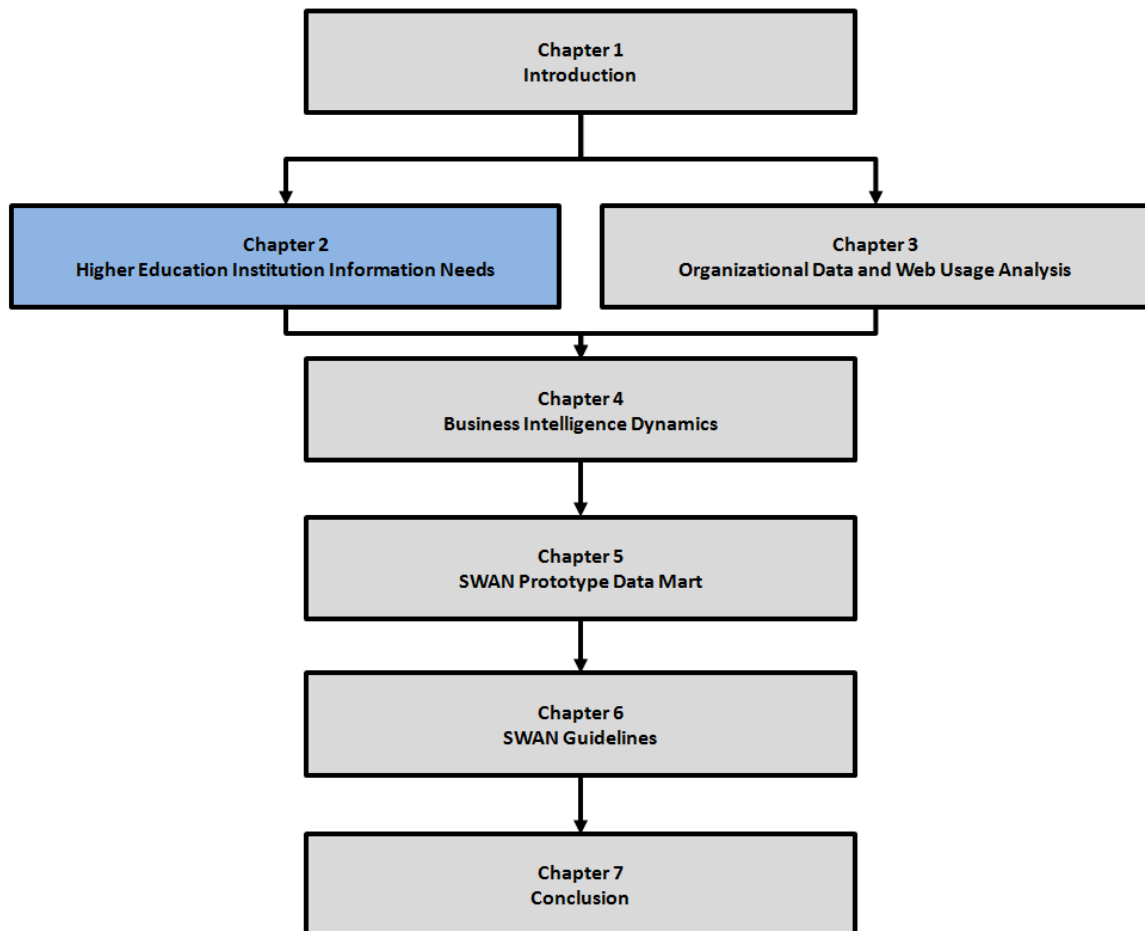
Questions regarding the effects of Web usage on university students are difficult to answer. The lack of useful information, despite the large number of available data sets generated from Web use within the NMMU has been highlighted. The potential of a data mart to facilitate the use of these data was presented.

The problem statement that refines the solution area upon which this research intends to focus was presented. In addition, the objectives to be met if an artefact with adequate utility was to be produced were established.

A research design approach and paradigm were established as a design science methodology approach appropriate to the nature and context of the study. Peffers et al.'s (2006) process for undertaking design science research was discussed and was adhered to in this study.

Chapter 2 - Higher Education Institution Information Needs

“Knowledge is power. Information is liberating. Education is the premise of progress, in every society, in every family.” – Kofi Annan



This chapter provides background information on the context of this research problem. The shift in information access mediums is discussed. Formal education is explored within a South African context, with a focus on higher education. The objectives of the various levels within higher education are explored.

2.1 Introduction

Education is a term familiar to most, and yet defining education holds many complexities. Education can be seen as something obtained or as a given process. A student can receive an education and education can refer to the process of attending classes and receiving information from an educator. Peters (2010) stipulates that the concept of education cannot be tied to one specific process; rather, it refers to certain criteria that could apply to a set of processes. For example, people could be educated by travelling, engaging in conversation, reading books or simply by observing and exploring their environments.

The definition and study of the concept of education resides in and is closely tied to sociology and psychology. However, the general consensus within literature is that the overall aim of education is to increase students' capacity to learn, and to provide students with analytical skills and to further their ability to process new information and derive independent conclusions (Gow & Kember, 1990).

In South Africa, education is provided by the government to the public through compulsory school attendance. According to the South African Schools Act, it is compulsory for a child to attend Grade 1 at a school from the first day of the year in which the learner reaches the age of 7 until the last school day of the year the learner reaches the age of 15, or the ninth grade, whichever occurs first. Schools provide for the enrolment of students from Grade R to Grade 12 (*South African Schools Act No. 84 of 1996*, 2011). Any official education or qualification past Grade 12 is obtained from private or public universities, which fall into the category of higher education.

According to Healey (2007), the purpose of a university is to create and disseminate knowledge through research and teaching. This notion is further solidified by Salmi (2003), who observes that higher education is vital for distributing, creating and applying knowledge and for building professional capacity. The application and advancement of knowledge results in social and economic progress (Salmi, 2003).

Knowledge is the suitable accumulation of information in such a way that it has use (Bellinger, Castro & Mills, 2004). Information can therefore be seen as raw material for knowledge creation, and as an essential part of this knowledge creation process. Dramatic changes have occurred over the past twenty years in the way that information is accessed. Advancement in information and communication technology (ICT) has accelerated the speed and increased the capacity of information distribution significantly. Many of these technologies have been widely adopted within education institutions. Higher education in particular is experiencing an increase in the use of technology, specifically e-learning (Kukulska-Hulme, 2012; Njenga & Fourie, 2010).

The Internet and World Wide Web in particular have revolutionised the way information is exchanged. Through the use of websites, email and other online services, communication and information exchange have become much easier than traditional means, to the point where heavy Internet users cut back on all forms of traditional media (Wellman & Haythornthwaite, 2008).

The integration of technology and education has resulted in the term “ubiquitous computing”. Ubiquitous computing can be defined in an educational context as teachers and students having access to computing devices and Internet access whenever it is required (Hooft & Swan, 2007). This rapid integration has resulted in a higher education environment that is rich in information retrieval and communication technologies (Amador & Amador, 2014).

Universities are responsible for teaching students how to interact with information (Breivik, 2005). Moreover, “higher education institutions must address changing expectations associated with the quality of the learning experience and the wave of technological innovations” (Garrison & Vaughan, 2008, p. ix). Lecturers and management staff at higher education institutions will probably have to conduct reform and implement changes to the systems in place in order to cater for the change in environment and student needs brought about by the integration of technology and education. If they are to be successful, they will need to have appropriate information about the effects of these technologies. As in many organisations, managing and using information and knowledge about business processes can be used to better serve the overall goals and objectives of that organization (Davenport & Prusak, 1998).

The purpose of this chapter is to establish what literature and research communities offer regarding information access and education, and to explain how information access has changed. The chapter also establishes major turning points in information access in education and in the understanding of the possible effects of these changes. The objectives of higher education in general are investigated. In addition, the objectives of higher education facilitators, from a government perspective (macro) to the educator’s or lecturer’s perspective (micro) are considered. In this study, educators’ perspectives were recorded in a survey, distributed via online survey forms. This was done in an attempt to understand what information would be of use to the institution and its educators when making decisions about teaching approaches and using ICT for access to information within their institutions. In this way, insights into how and where this information might be collected were gained.

The undertakings of this chapter contribute to the completion of step 1 of the DSRP. The problem is further defined and the importance of this problem is established.

2.2 Information Access and Education

“Over the past decades, one of the important issues for education reform and innovations has been the integration of technology into education” (Lee & Tsai, 2008, p. 2).

Laptops, computers and Wi-Fi access to the Web have become commodities in most universities and higher education institutions today and this is termed ubiquitous computing (Fried, 2008). Higher education has become responsible for teaching the necessary skills to identify the integrity of the information available to students (Breivik, 2005). This task is increasing in complexity. A contributing factor is that the amount of information shared and processed has been escalating exponentially (Heylighen, 2002). No longer are libraries and teachers or lecturers the only primary source of information within an education institution; the World Wide Web provides students with a vast amount of information.

2.2.1 Information Access Paradigm Shift

During pre-industrial times, communication was facilitated through letters carried by messengers on horseback. If it is assumed that an average letter equates to roughly 10 000 bytes and the journey took one month this can be estimated as about 0.03 bits per second. Modern fibre optics can transfer billions of bits per second. In a period of just 200 years the speed of information transfer has increased about 100 billion times (Heylighen, 2002). The most commonly recognised and pervasive contributors to this change in media and information dissemination are the Internet and the World Wide Web.

The Internet was an unforeseen end result of an attempt by the Advanced Research Project Agency (ARPA) to reroute communications around a disruption in the infrastructure, the result of collateral missile damage during the Cold War. ARPA was founded by the US government to spearhead technological advancement projects.

The result of ARPA's initial research was a data network named ARPANET, which allowed communications to be redirected where needed, should a vital infrastructure link be damaged. Subsequently, many similar networks were built across the globe. Robert Kahn and Vint Cerf designed Transmission Control Protocol (TCP) in 1973, to solve the anticipated problem of trying to connect or communicate with this multitude of heterogeneous networks (Mayo & Newcomb, 2008). TCP allowed for interconnecting networks; and the term “Internet” was coined. The Internet was initially used primarily by universities and select organisations to exchange and store files and for the most part remained relatively dormant.

This groundbreaking technology was a platform for the development of the World Wide Web, which went live in December 1990 (Berners-Lee, 2010). The Web provided a user-friendly interface to the Internet. Lawrence H. Landweber states that "the World Wide Web turns the Internet into a repository, the largest repository of information and knowledge that's ever existed" (Mayo & Newcomb, 2008, p. 4).

The benefit and vast utility of the Web was soon widely recognised. What resulted was the exponential and hyper acceleration of the use and growth of the Web (Mowery & Simcoe, 2002). Its commercial viability was also recognised and investment in the technology was substantial. Use and access to the Web has become a vital facet of life for huge numbers of people across the globe (Seo, Kang, & Yom, 2009).

Before the development of the Web, centralised broadcast media were used to distribute information. Everyone received the same information at the same time. In contrast, the contemporary information distribution paradigm is based on network media, through which an individual can contribute to the body of information and can decide what he or she would like explore (Boyd, 2010).

Current information gathering technologies provide vast amounts of unfiltered, auxiliary and superfluous information to users. One possible cause for this massive rate of progression and resulting overabundance could be ephemeralisation. Ephemeralisation is a progression dynamic through which constant emphasis is placed on trying to achieve or produce more with fewer materials or less effort (Fuller, 1973). The result is that that which is scarce becomes abundant.

2.2.2 Information Overload

The way information is accessed has clearly changed dramatically over the course of the 20th century. During this period, a problem emerged which has its roots in early civilisation. This problem has been termed "information overload" and despite being the most widely recognised problem of its kind, there is no single accepted definition for this phenomenon (Bawden & Robinson, 2009). This might be related to the term's having multiple meanings. On the one hand, it could be interpreted as having more relevant information than an individual is able to absorb; on the other hand, it could mean being overwhelmed by an immense amount of unrequested information, only some of which is relevant (Butcher, 1998). Feather (2000) observes that information overload occurs when there is such a large amount of information that it cannot be used effectively. Speier, Valacich and Vessey (1999) observe that "information overload occurs when the amount of input to a system exceeds its processing capacity" (p. 338).

Meadow and Yuan (1997) argue that information overload is actually an overload of possible sources of information vying for our attention, which creates the sense of being overwhelmed or overloaded.

They justify their argument by saying that in order to be information, messages need to be processed and evaluated, or they are merely data. Therefore, you can be overloaded by data but not by information. However, the term information is inherently ambiguous and it is unrealistic and bewildering to hold the term to one specific definition. The important uses for the term information refer to knowledge imparted and the process of informing. The term is also used to describe the things that carry out these uses, i.e. books or documents (Buckland, 1991).

The conditions under which information overload occurs are generally recognised and in this subsection the side effects of information overload are of interest. In this research study, information overload refers to having an excess of information available thereby affecting the information seeker's ability to gather useful information.

Information overload presents various related problems and side effects including information anxiety, infobesity, data smog and Internet addiction, amongst others.

Information anxiety is the culmination of five phenomena, as described by Wurman (1989):

1. Failure to understand information
2. Feeling overwhelmed by the amount of information to be understood
3. Not knowing whether certain information exists
4. Not knowing where to find information
5. Knowing exactly where to find the information, but not having the skills to access it.

Information anxiety shares many traits with information overload. There is an ongoing debate in the information overload research field on whether "information overload", "information anxiety" or "cognitive overload" should be used as the accepted definition (Girard, 2005). This is testament to the close association and various perspectives of the general problem of abundant information.

Our bodies are designed to store fats and sugar, that which is rare in nature. Our brains are programmed in a similar way. Information which is gross, sexual or violent in nature, or offensive, humiliating or embarrassing grabs our attention because it is stimulating. Based on this, the Web can produce a form of psychological obesity if a user does not guard against this (Boyd, 2010). This effect is also referred to as "infobesity". Essentially, infobesity is an overindulgence in large quantities of information, generally low in quality and integrity (Roberts, 2005). Infobesity is perhaps an issue resulting from "data smog", which is described as a large quantity of low quality information or data

(Shenk, 2009). Data smog is especially pervasive on the Internet, given the freedom the Web allows regarding information exchange and contribution.

Research into cyber psychology has shown that online users can become addicted to the Internet in the same way that people become addicted to drugs, gambling and alcohol (Young, 1998). “Problematic Internet addiction or excessive Internet use is characterized by excessive or poorly controlled preoccupations, urges, or behaviours regarding computer use and Internet access that lead to impairment or distress” (Weinstein & Lejoyeux, 2010, p. 277). As in any type of addiction, Internet addiction can lead to negative social and psychological issues. Tang, Yu, Du, Ma, Zhang, & Wang (2014, p. 744) define Internet addiction as “an inability to control one's use of the internet” leading to “negative consequences in daily life” and believe it poses a serious public health problem, particularly amongst adolescents. Many links exist between certain psychiatric disorders and Internet addiction (Ko, Yen, Yen, Chen, & Chen, 2012).

Various diagnosis metrics have been established for Internet addiction. However, many psychological, behavioural and demographic factors have to be considered when diagnosing an individual (Chou & Hsiao, 2000; Tao et al., 2010; Whang, Lee, & Chang, 2003). Stepanikova, Nie and He (2010) found that the way in which their surveys were constructed influenced the results in a study on the socio-psychological effects of increased Web use. However, when relationships did exist they were seemingly negative, not positive.

Despite the fact that “Internet addiction” is not recognised as an official disorder in the psychiatric community, partly as a result of the difficulty of isolating and diagnosing it, it is well documented as a growing problem (Chak & Leung, 2004; Ng & Wiemer-Hastings, 2005; Weinstein & Lejoyeux, 2010; Widyanto & McMurren, 2004). “Internet addiction is more widespread than just on university campuses where laptops and computer labs are within easy reach; it is also being seen in high school and middle school students” (Wallace, 2014, p. 12). Access to the Web and low cost computer equipment contributes to the trend that sees most students entering higher education having at least a moderate level of computer literacy (Gorgone, Davis, Valacich, Topi, & Feinstein, 2003), and thus many students are using and are familiar with the Web on entering higher education.

Access to vast amounts of information also brings with it the possible exposure to unsavoury content, which could result in undesirable behaviour and distorted ethics. A study of 985 university students undertaken by Rumbough (2001) attempted to shed light on just how much unethical behaviour occurs on the Web. Taking into account that unethical or unsavoury Web behaviour could include accessing websites containing content such as fake IDs, illegal drugs, illegal weapons, pornography, racism and gambling, he concluded that many of the users frequently accessed controversial material online.

A study using Web usage analysis techniques conducted at a South African higher education institution concluded that there was a significant correlation between higher Internet use and lower academic performance. Moreover, the usage patterns of good students were distinctly different from those of weaker student (Hazelhurst et al., 2011).

The current generation is undoubtedly dealing with much more information than their parents ever did (Breivik, 2005). The generation born in the 1980s and 1990s is considered to be immersed in entertainment, communication and any form of electronic information and media (Rosen, 2010). Kirschner and Karpinski (2010) observe that there has been significant media hype around the behaviour of this “NET generation”. An assumption that is sometimes made based on this generation’s tendency to overindulge in information is that they have exceptional multitasking skills and possess high technological prowess. However, empirical research shows that they do not necessarily possess such skills and that having multiple, constant information stimuli may hamper their information processing capacity. In a study of 102 undergraduate and 117 graduate students at a large public university, Kirschner and Karpinski’s (2010) main findings were that there was a clear and noteworthy negative relationship between the use of the widely used social network medium Facebook and academic performance.

Various factors can contribute to the way in which information abundance affects a student’s education. Kubey, Lavin, & Barrows, warned more than a century ago that academic practitioners should be aware of the growing problems associated with information overload, especially in higher education (2001). Investigating the objectives of education could provide a better understanding of information overload and the associated side effects in education.

2.3 Primary Objectives of Education

Despite consensus on the broad definition of education, its primary objectives vary depending on the type of education being provided. In this study the primary objectives of higher education are of interest and are explored in the following section.

2.3.1 Primary Objectives of Higher Education

Higher education is essential for the efficient formation, spread and utilisation of knowledge (Salmi, 2003). Knowledge has become a critical factor in economic progression. Fundamental long-term growth is dependent on the persistent growth of the national economic knowledge base (OECD, 1998). Therefore, higher education can play an important role in the well-being of a nation’s economy.

Higher education institutions are the intellectual training ground for students. The acquisition of professional skills is of major benefit to an individual. In addition, intellectual prowess and a proclivity for gathering and applying relevant knowledge can be developed. However, considering the entities involved in providing higher education, goals and benefits are likely to differ. Three major entities involved in education, namely the government, institutions and educators, are relevant to this study.

Government is responsible for vital decisions relating to the well-being of a country. Ideally, decisions made by the government should have very positive effects on the country and its citizens. One of the responsibilities of a government is to distribute the country's resources appropriately. The well-being of a country's citizens relies heavily on this allocation of resources. It is therefore imperative that resources are provided to areas that will benefit the overall positive growth of a country. Investment in education has been found to be a catalyst for economic growth. In many countries, education has provided a multitude of economic and non-economic gains (Blondal, Field, & Girouard, 2002).

In Africa, education and particularly higher education has been identified as an important driving factor in poverty alleviation. Added to this, improvements in national health, population growth reduction, technology development and the strengthening of governance are strongly linked to investment in higher education (Bloom, Canning, & Chan, 2006). Higher education should thus be regarded as an important area for investment for African countries desiring these benefits.

The Republic of South Africa is a democratic state. Therefore, the aims of the government are required to be documented in laws and policies and made transparent to citizens. According to the Higher Education Act of South Africa (HEAA Council, 2001), some of the broad goals of higher education from a South African government perspective are to:

- Establish a single coordinated higher education system that promotes cooperative governance and provides for programme-based higher education
- Restructure and transform programmes and institutions to respond better to the human resources, economic and developmental needs of the Republic
- Provide optimal opportunities for learning and the creation of knowledge
- Pursue excellence, promote the full realisation of the potential of every student and employee, tolerance of ideas and appreciation of diversity
- Respond to the needs of the Republic and of the communities served by its institutions

- Contribute to the advancement of all forms of knowledge and scholarship, in keeping with international standards of academic quality
- Allow higher education institutions to enjoy freedom and autonomy in their relationship with the State within the context of public accountability and the national need for advanced skills and scientific knowledge
- Provide for the establishment, governance and funding of public higher education institutions
- Provide for the registration of private higher education institutions.

The objectives of this act are to provide for the establishment of public or private higher education institutions that offer programme-based, international standard qualifications and facilitate an environment for creating scientific knowledge, thus allowing members of the population to gain higher education certifications.

These higher education institutions are run in a similar way to most organisations or enterprises in that they require management and effective administration to operate successfully. “Efficient management and administrative systems are of paramount significance to the productivity and effectiveness of any enterprise; academic institutions are no exception” (Teferra & Altbachl, 2004, p. 31). Relevant information about internal operations and related processes can act as a catalyst for more effective decision-making. Information and knowledge allows organisations to move forward (T. Davenport & Prusak, 1998). Therefore, it can be said that information and knowledge are vital to the overall success of an organisation and that they are fundamental to its well-being. In the context of an organisation, information and knowledge refer to what is known about the organisation itself, operational processes and anything relevant to or having an effect on the goals or aims of that organisation.

Each higher education institution, private or public, will have documented its objectives, goals, vision or ambitions at various levels. The NMMU has encapsulated its desired future in the Vision 2020 Strategic Plan, which mentions the following institutional educational objectives:

Vision

To be a dynamic African university, recognised for its leadership in generating cutting-edge knowledge for a sustainable future

Mission

To offer a diverse range of quality educational opportunities that will make a critical and constructive contribution to regional, national and global sustainability

To achieve our vision and mission, we will ensure that:

- Our values inform and define our institutional ethos and distinctive educational purpose and philosophy.*
- We are committed to promoting equity of access and opportunities so as to give students the best chance of success in their pursuit of lifelong learning and diverse educational goals.*
- We provide a vibrant, stimulating and richly diverse environment that enables staff and students to reach their full potential.*
- We develop graduates and diplomates to be responsible global citizens capable of critical reasoning, innovation, and adaptability.*
- We create and sustain an environment that encourages and supports a vibrant research, scholarship and innovation culture.*
- We engage in mutually beneficial partnerships locally, nationally and globally to enhance social, economic, and ecological sustainability.*

Educational purpose and philosophy

- We provide transformational leadership in the service of society through our teaching and learning, research and engagement activities.*
- To achieve this we are committed to developing the human potential of our staff and students in the full spectrum of its cognitive, economic, social, cultural, aesthetic and personal dimensions in the pursuit of democratic citizenship.*
- We adopt a humanising pedagogical approach that respects and acknowledges diverse knowledge traditions and engages them in critical dialogue in order to nurture a participative approach to problem-posing and -solving, and the ability to contribute to a multi-cultural society.*
- We inspire our stakeholders to be passionate about and respectful of an ecologically diverse and sustainable natural environment.*

- *We will be known for our people-centred, caring, values-driven organisational culture that will allow all members of the university community to contribute optimally to its life.*

Strategic priorities

1. *Formulate and implement an integrated strategic academic plan and distinctive knowledge paradigm.*
2. *Promote student success through excellence in teaching, learning and assessment.*
3. *Create and sustain an environment that encourages, supports and rewards a vibrant research, scholarship and innovation culture.*
4. *Position NMMU as an engaged institution that contributes to a sustainable future through critical scholarship.*
5. *Develop and sustain a transformative institutional culture that optimises the full potential of staff and students.*
6. *Formulate and implement a financial growth and development strategy to enhance long-term sustainability and competitiveness.*
7. *Improve institutional processes, systems and infrastructure to promote a vibrant staff and student life on all campuses.*
8. *Maximise human capital potential of staff.*

As represented by the NMMU Vision 2020 Strategic Plan (Council, 2010).

2.3.2 Educators' Information Needs

The primary educational goals of the South African government are clearly and concisely documented. In the same way, every institution will have set primary educational goals. However, educators who are at the chalk face may have a different understanding of their institution's primary objectives. Moreover, they may have their own, different primary educational goals. These goals require investigation to further understand their information requirements.

Method

A survey was conducted using an online survey form, attached as Appendix B. The purpose of this online survey form was to extract lecturers' primary educational objectives and to establish their thoughts on desirable Web usage behaviour among students. It was thus designed to enable an

understanding of how lecturers perceive the effect of Web use on their educational objectives, and to gain some insight into the type of information about students' Web usage that would be valuable to lecturers. In this way, the researcher could form a better understanding of the lecturers' educational information needs. Furthermore, the clarity regarding lecturers' educational information needs, provided by the online survey, allowed for the identification of an interview candidate used in the information requirements interview, as discussed in Section 6.3.2. This candidate indicated value in student Web usage information.

The target population for the survey was IT lecturers at the North Campus whose students were enrolled in an undergraduate National Diploma in Software Development and made use of North Campus facilities. This population was targeted because the Software Development course involves three years of study and many of the subjects are technical and programming-based, requiring practical classes in laboratories where students are able to access the Web for technical support and information retrieval. Moreover, given the scope constraints of this research, all IT subjects could not be considered. It is not assumed that all IT students share the same information needs, nor is it assumed that their needs are different based entirely on their subject domains. This online survey served to obtain sample data for a prototype.

It was assumed that lecturers would be interested in the Web usage behaviour of their students as it reflects the way they are using the Web during practical classes; such information could be valuable when making decisions regarding the planning and teaching of practical classes, amongst others.

A few examples of questions which were posed in the online survey are:

- Is the material provided to your students sufficient for the subject? In other words, they do require additional material for the subject, besides what you give them?
- Does your subject/s require your students to have up-to-date and recent information in that subject area?

In addition, lecturers were asked to indicate how strongly they agree or disagree with statements such as:

- Students' Web usage behaviour significantly affects your primary educational objectives
- Having information regarding the Web usage behaviour of your students would benefit your decision-making in achieving your primary educational objectives

Further, lecturers were asked to indicate to what extent example information regarding student Web usage behaviour would affect their decisions towards their primary educational goals. For example, lecturers were asked to indicate to what extent the following information could influence their primary educational goals:

- The length of time an individual student spends browsing the Web on campus per day. For example, student A spends an average X minutes on the Web on campus per day

A total of 14 lecturers were identified using the subject lists for the National Diploma in Software Development. Each lecturer who was asked to participate taught at least one of the subjects. 12 lecturer responses were received and covered 17 subjects in total. The online survey form is attached as Appendix B.

Most of the lecturers (9 out of 12) indicated that their primary educational objectives were to develop graduates who were able to think critically, or to facilitate knowledge dissemination (5 of the 12). These objectives were listed more than other possible objectives and lecturers did not indicate any other objectives. It was thus concluded that the lecturers were trying to develop their students' ability to process information, gather knowledge and to think with careful consideration about the information to which they were exposed. Their ability to process information was particularly relevant because information is the raw material for knowledge and often requires critical, rational and logical thought. Therefore, the dominant primary objectives in this sample were concerned with the information processing abilities of students. Given that the Web is frequently students' main source of information, knowing the way in which they obtain, gather and use their information from the Web could affect and contribute to these lecturers' primary objectives.

The majority of the lecturers (8 of the 12 who responded) indicated that their subject/s required up-to-date information and that the material they provided to their student was not entirely adequate for detailed knowledge beyond the scope of the course; thus students could be required to do further research in the subject, probably by using the Web. One participant mentioned that the study material provided was adequate for fundamental knowledge in the subject but not for detailed or in-depth knowledge. Three of the participants indicated that in the subjects they taught, their students were required to access online resources.

Half the respondents indicated that their students' Web usage behaviour significantly affected the primary educational objectives of the course. Moreover, just under half (5 out of 12) agreed that information about their students' Web usage behaviour would benefit their decision-making in the achievement of their primary educational objectives. This suggests that there is a need for information that reveals how students are using the Web and what they are using it for.

Almost two thirds of the lecturers (7 out of 12) indicated that knowing the length of time a student spends on the Web on campus each day would influence their teaching decisions. Four of these seven respondents indicated that knowing what students were browsing would be particularly important. Over three quarters of the respondents (9 out of 12) answered that knowing the length of time an

individual student spends browsing subject-relevant websites during a given practical class would influence their teaching decisions. Four of these lecturers considered that a very likely change to their decisions and planning would occur if they had this information. In addition, over half of the lecturers (7 out of 12) indicated that knowing the length of time an individual spends browsing websites unrelated or irrelevant to the subject during a given practical class would influence their teaching and planning decisions. All but one of the respondents indicated that knowing the website or Web service most used by a group of students would influence their teaching decisions. This further demonstrates the need for and value of information about students' Web usage behaviour. Such information would clearly benefit lecturers' decisions regarding their primary educational objectives. Moreover, known issues associated with Web use could be identified by analysing Web usage behaviour. All these respondents indicated that they were aware to some degree of the problem of information overload and all but one was aware of the growing threat of Internet addiction.

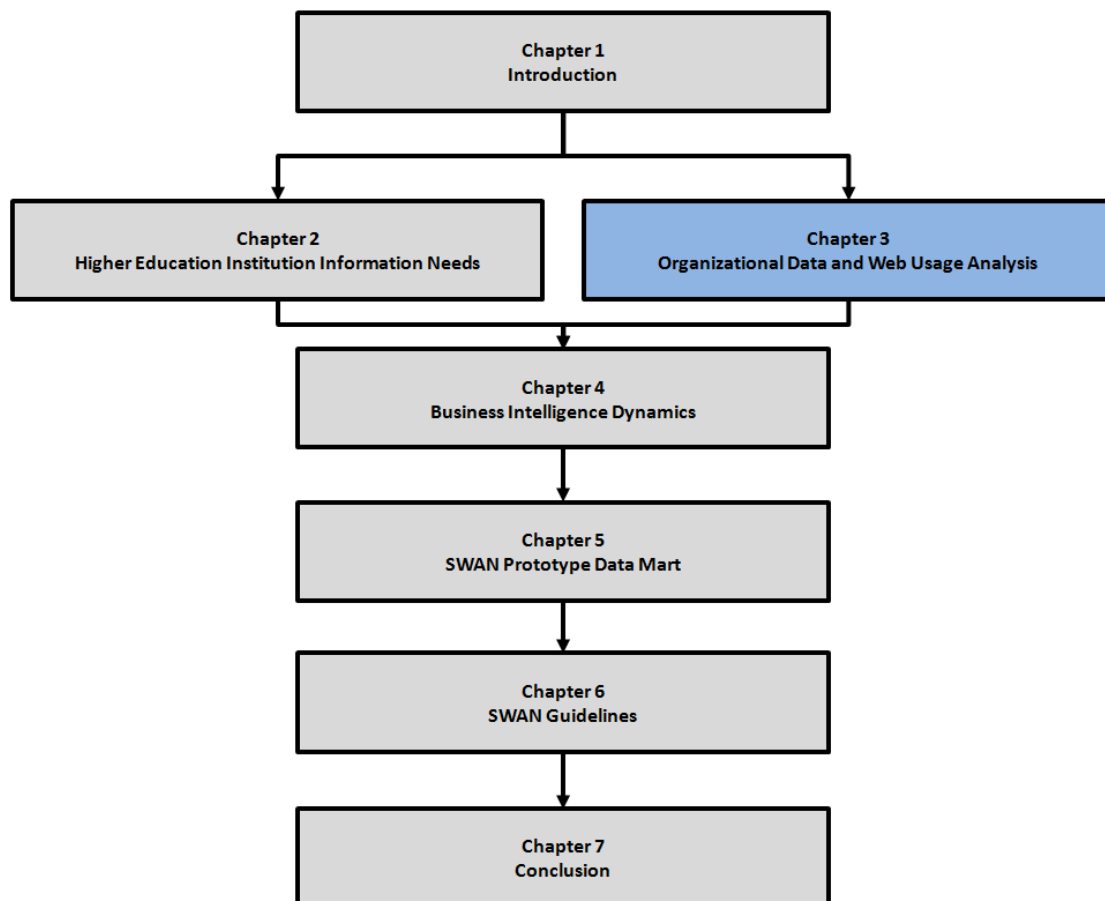
2.4 Conclusion

There has been a clear and widespread paradigm shift in information access over the last two to three decades. Information and information access are central to higher education and its related educational objectives. The importance of higher education to a country's well-being is clear. The differing nature of objectives in higher education is clear as well; there are overlapping commonalities but also clear differences in government's perspectives, institutions' perspectives and educators' perspectives of these primary objectives

By establishing the primary objectives of lecturers at NMMU, ways in which to provide for these information needs can be investigated. A need for information on students' Web usage was found to be a contributing factor in decision-making with regard to the respondents' primary educational objectives. How this information can be gathered and presented needs to be explored and established.

Chapter 3 - Organisational Data and Web Usage Analysis

“You can have data without information, but you cannot have information without data.” – Daniel Keys Moran



This chapter explores possible solutions to the problem of accessing Web usage information by investigating organisational data in general and its possible uses specifically. The context for this research is refined and previous research relevant to this study is introduced.

3.1 Introduction

Information systems and applications have become a prerequisite for most businesses and organisations. These systems are implemented to handle and control the information needed and created by an organisation. With this integration of information systems into business has come the need to monitor, trace and maintain them (Peng & Li, 2005). Most information systems store data on processes, access, errors and various other activities conducted by the system or its users. This data is commonly stored in a file known as a log or log file. Essentially, a log file is an account of all the activity on the system and is a primary source for information about this system (Sloan, 2001). By analysing these log files, very meaningful information can be gathered, which can be a great asset when assessing or managing the system's performance or other control mechanisms.

The undertakings of this chapter contribute to the completion of step 2 of the DSRP. Possible solutions are explored through investigation of a data source relevant to the problem.

3.2 Use and Purpose of Log Files

3.2.1 Log files in general

Log files can be used in many different contexts. The format and purpose of a specific log file depends on the system that created it and the type of information needed from it (Casey, 2008; Oliner, Ganapathi, & Xu, 2012). Different information systems create different types of log files. For example, a point of sales system would require log information about sales and access of the system by users. Such information could include details of the total amount per sale, payment method, the sales person who conducted the sale, the date and time the sale took place and how many items were purchased. A different type of system, such as a network intrusion detection system, might produce logs containing information about unauthorised access to or malicious activity on the network. The logs could detail the source IP address of the unauthorised attempt, the time and date of such activity or the threat level of a security breach, among others.

The use of a log file would depend on the person analysing it, based on the type of information needed. The analysis of a collection of log files can provide very meaningful and relevant information.

Examples of this include, but are not limited to:

- Jiang et al. (2009) used a set of log files in a customer support system called NetApp to study the characteristics of customer problem troubleshooting incidents.
- Asunka, Chae, Hughes, & Natriello (2009) analysed log files that contained user activity on an academic library website to extract general usage patterns on the site.

- Black, Dawson and Priem (2008) looked at log files generated by an online Moodle learning management system to establish whether students' perceptions of the course community were related to their actual activity on the online graduate level course.
- Hazelhurst, Johnson and Sanders (2011) analysed log files produced by a proxy server which contained Web usage data from 2153 undergraduate students in order to identify correlations between patterns of Web usage and students' academic performance.

3.2.2 Networks and Web Logs

One of the most pervasive forms of information systems is the computer network. Through various devices such as firewalls, routers and servers, computers are connected, allowing them to communicate and exchange information. The devices used to facilitate a computer network can be configured to create log files which can be extremely useful to technicians who maintain and optimise the network. Network administrators are responsible for ensuring that a network runs correctly and efficiently and that services on the network are available with the minimum of downtime. They rely on an array of resources to gather information about issues which cause downtime or other undesirable occurrences. These resources include log files that provide a direct account of the activity on the network the administrators oversee (George, Makanju, Zincir-Heywood, & Milios, 2008). This information allows them to isolate and correct problems and resolve persistent issues.

One facet of overseeing and maintaining a network is detecting and preventing attacks on it. Sabahi and Movaghar (2008) observe that Intrusion Detection Systems (IDSs) have come into use as a result of the proliferation of computer networks and the need for security. These IDSs are capable of detecting attacks on networks in many different environments (Sabahi & Movaghar, 2008). The inclusion of IDSs has become common in many organisations (Kazimierz et al., 2008). Most current IDS products detect network intrusions and write their details to a log file (Itoh & Takakura, 2006). By analysing the log files produced by IDS products, administrators can find relevant and useful information to use in isolating and preventing attacks on the network (Bin Hamid Ali, 2011; Kowalski & Beheshti, 2006; Kazimierz et al., 2008). The detail and usefulness of the log file depends on the information requirements of the analyst. If the administrator of a network implementing an intrusion detection system is interested in investigating a particular security breach that has caused a system failure then he or she would identify that event on a given day. However, if the administrator needs to know the total number of unsuccessful attempts made from one specific IP address to access the network over a three month period, for example, the approach and information gathered will be different. It is in this way that log files can be considered the raw data of many systems because the data has no meaning or context until it is analysed for specific information. Intrusion Prevention Systems (IPS) can be put in place to stop attacks that are flagged by an IDS (which could be included

as part of the IPS) and to prevent known threat signatures which have previously been identified (Abdelkarim & Nasereddin, 2011; Ierace, Urrutia, & Bassett, 2005; Patel, Qassim, & Wills, 2010). An IDS informs the administrator of attacks, whereas an IPS takes measures to halt them (Patel et al., 2010).

Administrators and developers of websites can use log files created by various devices when a user accesses their websites to gather information (Cooley, 2000). They can extract user activity which has taken place on their website by analysing the log files produced by the Web server and its content server.

Businesses also recognise non-technical uses for their log files. A great deal of information on customer behaviour can be found within certain log files. By analysing these files and establishing user behaviour patterns, many companies use this information to drive marketing strategy and to better understand customer trends in order to tailor their services to the needs of these users (Casey, 2008).

Organisations that conduct business through electronic commerce (e-commerce) allow customers to gather information about their products and to purchase these products or services via the organisation's website. These organisations can benefit enormously from gathering and analysing the behaviour of their online customers (Kohavi, 2001). Recommender systems are used by many large e-commerce websites. They use log file data to track a customer's preferences and browsing behaviour and to recommend products to that customer, based on these preferences (Cooley, 2000; Schafer, Konstan, & Riedl, 1999).

Figure 3.1 shows an example implementation (MIMIC Architecture), where log files have been used to drive marketing strategies. Log files generated by e-commerce Web browsing (online shopping) are fed into pattern recognition software known as MiDAS. Retail data and domain knowledge are then also fed into MiDAS, which processes this data and delivers navigational patterns. These patterns represent the behaviour of online customers on a given e-commerce website. The navigational patterns are then analysed by a marketing expert who adds meaningful input to the domain knowledge of the given business. Using this knowledge of the behaviour and trends of online customers together with retail data, allows the business to provide personalised offers of current products to online customers. These offers are likely to be more successful as they are tailored to a specific customer, based on his or her previous browsing and purchasing history and preferences.

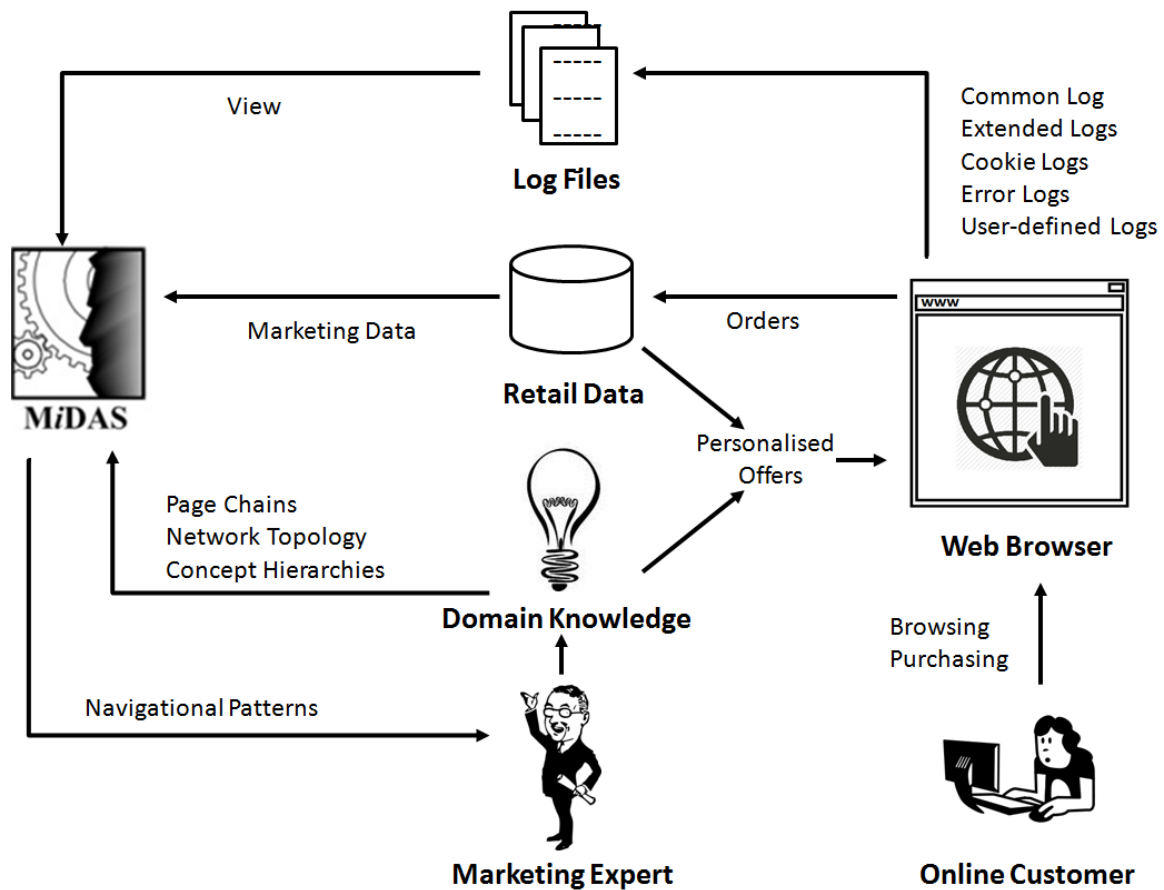


Figure 3.1 - MIMIC Architecture (Anand, Büchner, Mulvenna, & Hughes, 1999)

A similar approach has been adopted by Web services and websites such as search engines, social media sites and general information retrieval websites (for example, news, weather and sports), which provide customised content, based on browsing history and behaviour derived from log file data. Pariser (2011) notes that this has been termed “personalization”, and many major websites such as Google, Yahoo, Facebook, YouTube and Microsoft Live have adopted this as a core strategy in their efforts to remain relevant to an individual Web user in an increasingly competitive industry.

A typical user’s browsing a website from home usually results in numerous requests from the client (user’s computer) to the HTTP server (the server hosting the website in question). If the request is successful, the HTTP server will send the requested content to the client and the Web page will be displayed on the user’s screen. For example, if a user opens his or her browser and visits www.google.com, a request is sent to the Google server and it responds by sending the Google home page to the client. Any further activity triggered by the user on the website will result in similar requests. Figure 3.2 shows a typical example of such a request.

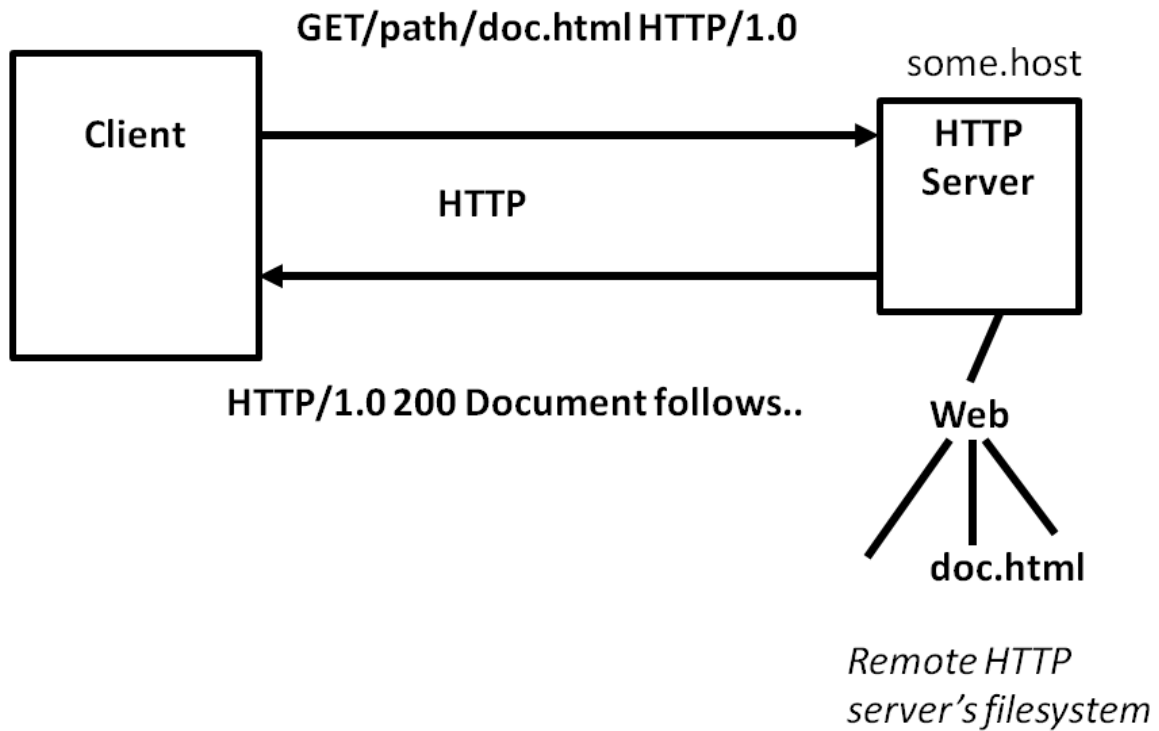


Figure 3.2 - A normal HTTP transaction (Luotonen & Altis, 1994)

In current business environments, most organisations have Web access (Kazimierz Kowalski & Beheshti, 2008; Thomason, 2012). The information streaming through their networks is a vital asset and the security of that information should therefore be an important concern if they wish to remain competitive (ISO/IEC 27002, 2005). Connecting a network to the Web opens it up to various threats (Thomason, 2012). An essential and widely adopted network security element is the firewall (Liu & Gouda, 2004, 2009; Mayer, Wool, & Ziskind, 2000). A firewall is a protective device that inspects all traffic entering a given network and accepts or rejects it according to a security policy (Al-Shaer & Hamed, 2003; Liu, Tornig, & Meiners, 2008; Zalenski, 2002). Firewalls can be configured to undertake a specific role or combination of roles, depending on the needs of the network it protects (Mayer et al., 2000; Zalenski, 2002). One of these roles is that of proxy server, put in place to allow access to the Web from within an organisation's network while remaining secure (Zalenski, 2002).

In a network where a proxy server is in place, all Web requests from within the firewall are intercepted by the proxy server and then forwarded to the requested HTTP server (Grace, 2011; Luotonen & Altis, 1994). Any undesirable requests can thus be denied. The network behind the proxy server can be protected in this way from external traffic (Zalenski, 2002). Figure 3.3 shows the general layout of a proxy server.

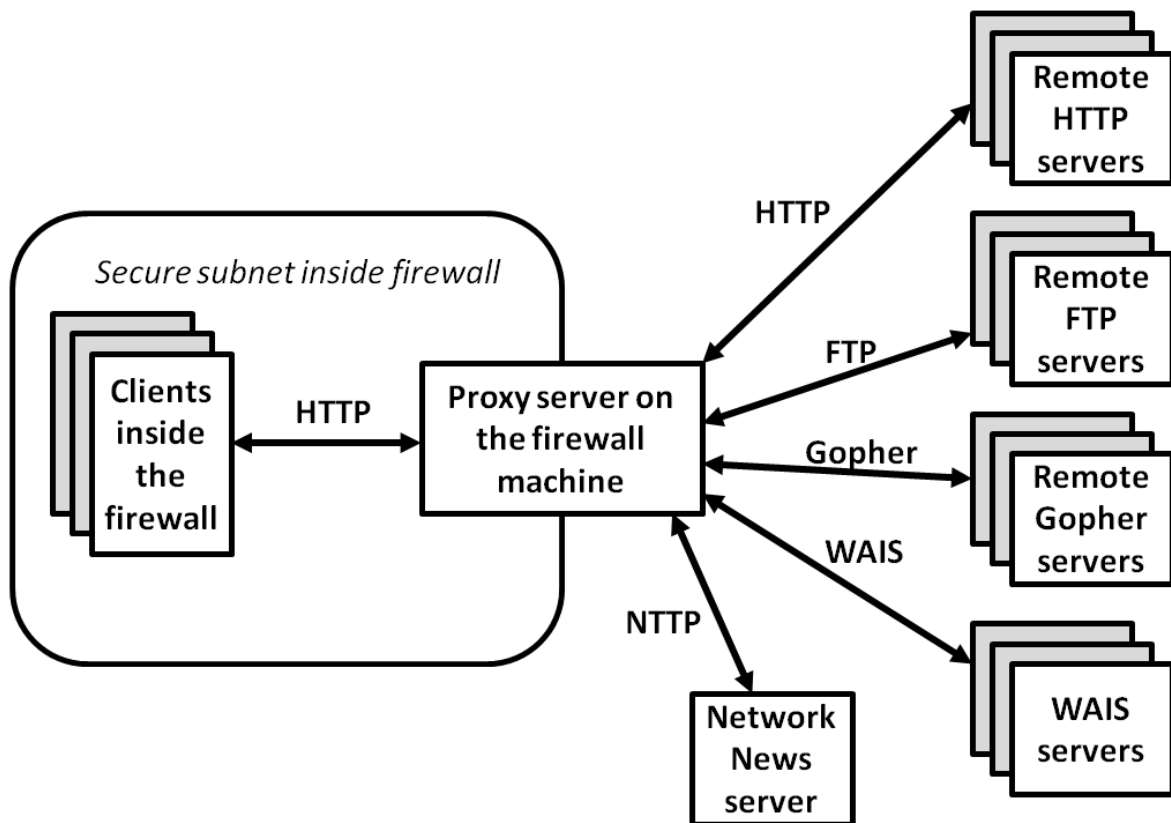


Figure 3.3 - Overall setup of a proxy (Luotonen & Altis, 1994)

Users browsing the Web within a proxied network will generate the same requests to HTTP servers hosting websites as a home user. Figure 3.4 demonstrates a typical proxied request.

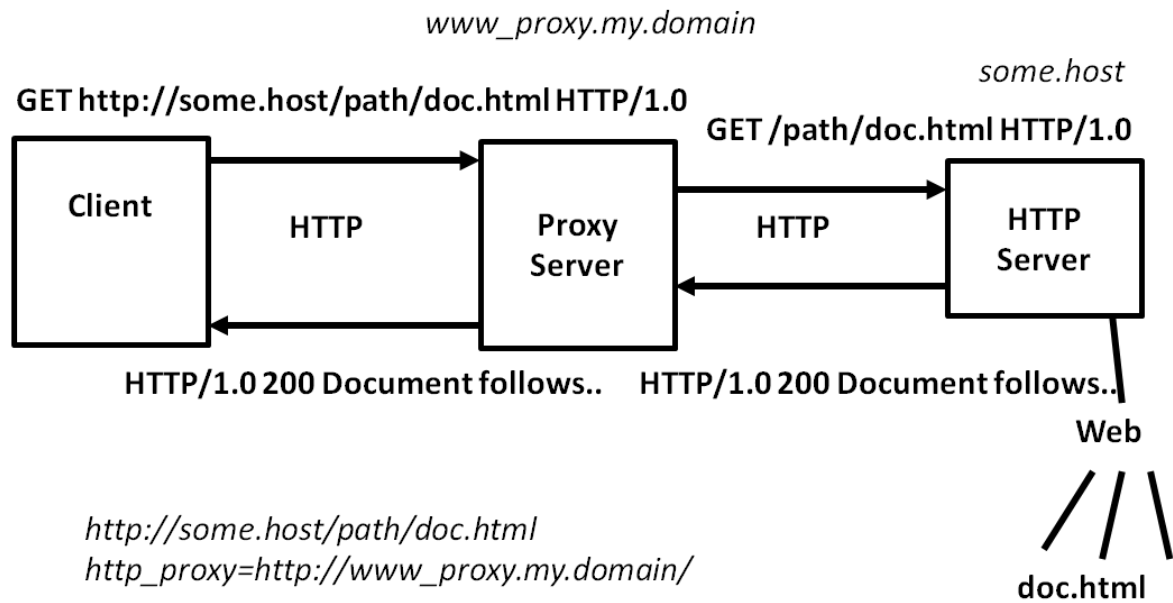


Figure 3.4 - A proxied HTTP transaction (Luotonen & Altis, 1994)

The proxy server acts as a gatekeeper or filter for the Web activity of the users within the proxied network. Given that these proxy servers intercept all the requests from the subnet, their log files will contain all Web requests from that subnet. The Web usage behaviour of users of the subnet can thus be extracted from these files.

3.3 NMMU Web Log Collection and Comparison

Following an interview conducted in prior research with two system engineers at North campus, it was established that the NMMU adopts a proxy model in which all Web traffic or outgoing requests from their sub networks pass through a single location (see Appendix G) (Von Schoultz, Van Niekerk, & Thomson, 2013) . In this case, a Fortigate firewall device serves as the gatekeeper for the proxy model and is similar to a proxy server. This Fortigate firewall device contains log analysing software called FortiAnalyser. However, it was established during a later interview with one of the same system engineers that similar log analysing software called Sawmill is used instead, the transcript of this interview can be seen attached as Appendix C. Sawmill allows for comprehensive network management information but does not provide any other decision support functionality. Any Web activity from a computer within the NMMU network will result in a request on the firewall machine to retrieve content from a Web server. Figure 3.5 shows how all sub networks must pass through the firewall to access any Web content on the Internet.

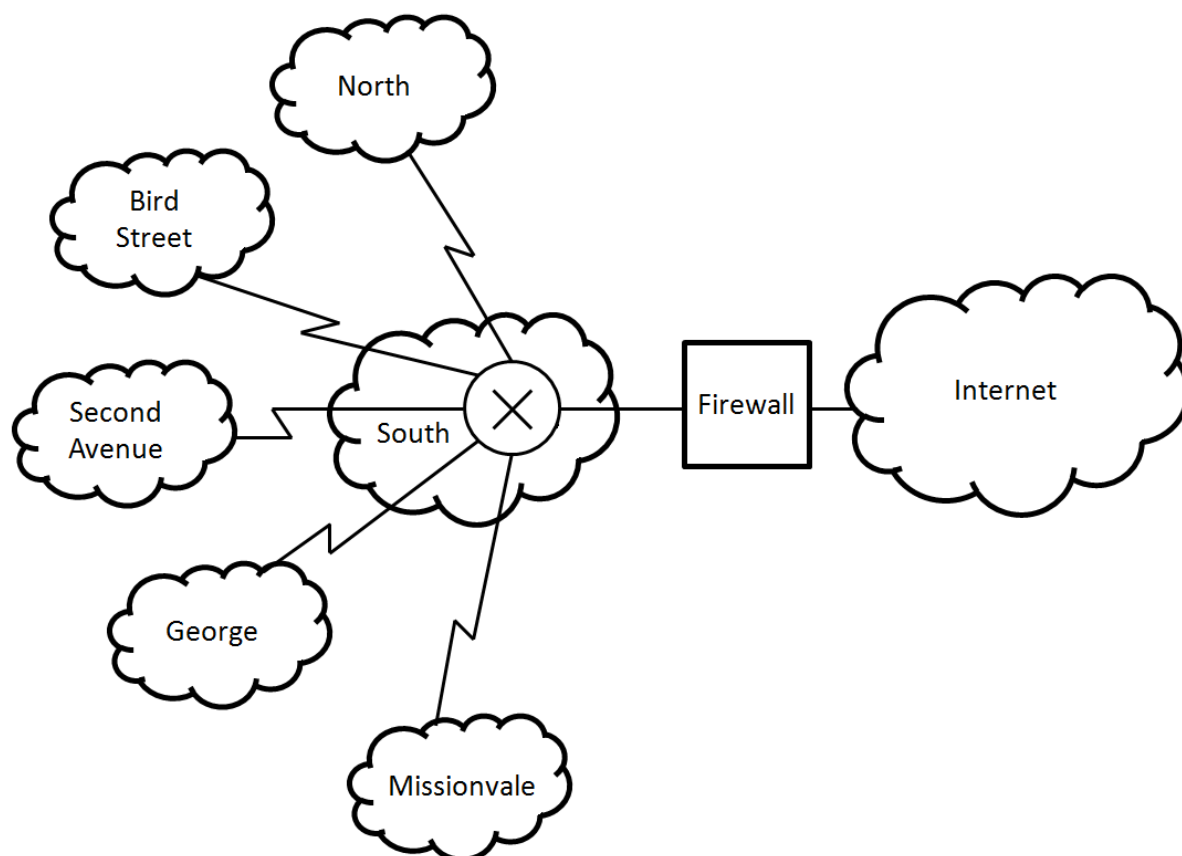


Figure 3.5 - NMMU's logical network (Von Schoultz et al., 2013)

The Fortigate firewall device stores log files containing the activity on the network in the form of network traffic. In the same prior study mentioned above, the two NMMU system engineers who were interviewed provided a sample of these network logs to determine whether there was enough detail within them from which to derive Web usage behaviour. It was concluded that the logs did contain raw network traffic accounts of Web usage by students on campus. However, the logs were not in an easily usable format as the log entries did not contain URL addresses or hostnames, as shown in Table 3.1. This was to the result of a logging configuration error on the firewall device, correction of which was pending (Von Schoultz et al., 2013).

Table 3.1 - Pre configuration error correction Fortigate firewall log entry example

Field	Value
Month	Apr
Day	12
Timestamp	11:52:15
unspecified	tyrael
date	2013-04-12
Time	11:52:15
devname	imperius
device_id	FGT1KC3912800514
log_id	0038000004
type	traffic
subtype	Other
pri	notice
vd	root
src	10.102.129.162
src_port	53149
src_int	"LAN_AGGR"
dst	173.236.49.82
dst_port	80
dst_int	"INTERNET"
SN	637641392
status	start
policyid	128
dst_country	"United States"
src_country	"Reserved"
tran_sip	192.96.15.20
tran_sport	61201
service	HTTP
proto	6
duration	0
sent	0
rcvd	0

The logging configuration error has since been resolved and logs retrieved from this firewall device now provide more detailed entries. The NMMU system engineers made a large number of detailed log files available for analysis. The Fortigate device is configured to write an entry into a log file whenever a Web request is made by a user on the campus subnet, as shown in Figure 3.6.

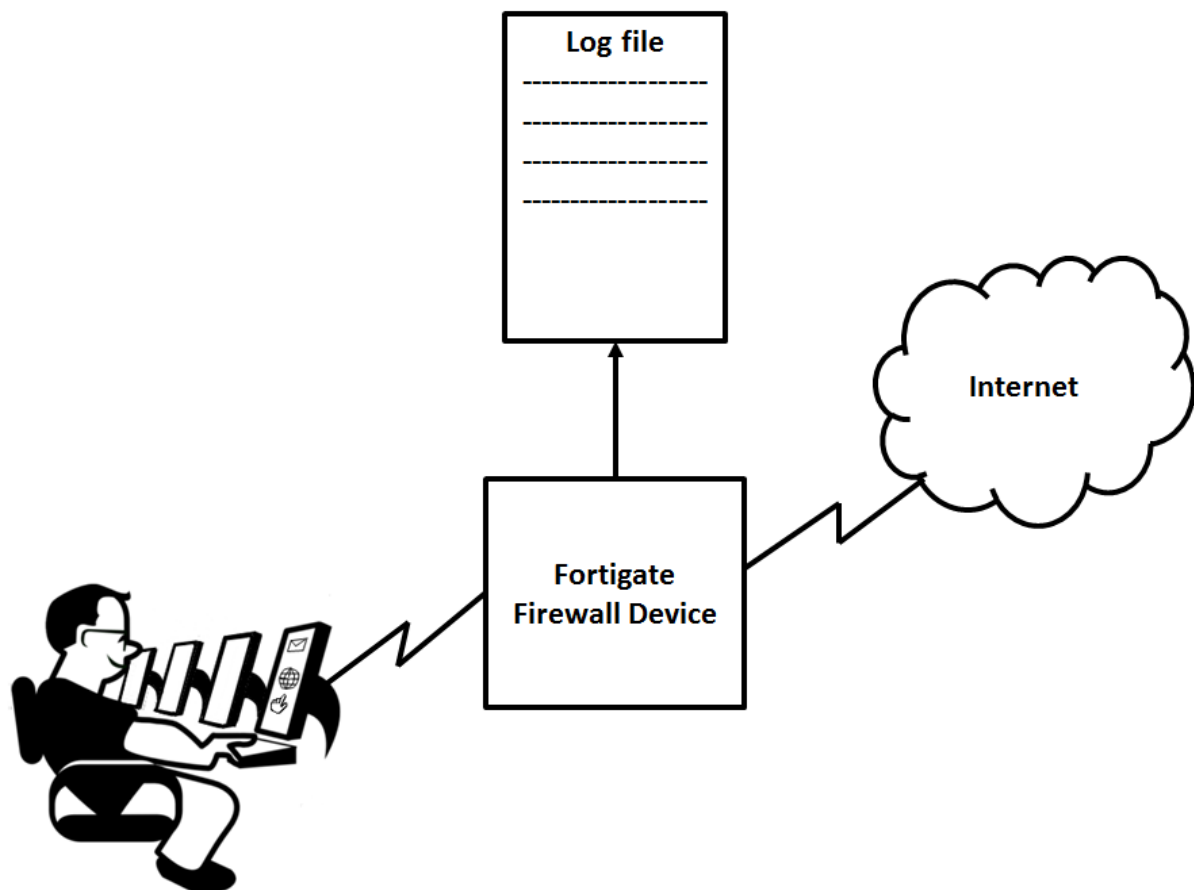


Figure 3.6 – Web log population

Hazelhurst et al. (2011) used proxy logs generated by a Squid proxy server to analyse student Web usage by uniquely identifying student sessions online and checking the URL of the websites they were browsing. A Squid proxy server is a type of proxy server (Wessels, 2001). Table 3.2 shows an entry in a log file generated by a Squid proxy server. It indicates the categories of information held by each entry.

Table 3.2 - Squid log example (Hossain, Rahman, & Kabir, 2012)

Field	Value
Timestamp	1322949739.102
Elapsed	77
Client	172.16.7.19

Action/Code	TCP_MISS/302
Size	350
URI	http://download.macromedia.com/pub/shockwave/cabs/flash/swflash.cab
Method	GET
Ident	-
Hierarchy/From	DIRECT/118.214.83.191
Content	-

In addition to other details, the logs generated by the Fortigate firewall device contained the same data as Squid proxy logs. Table 3.3 shows a single entry in a Fortigate log file, according to fields of data.

Table 3.3 - Post configuration error correction Fortigate firewall log entry example

Field	Value
month	Feb
day	5
time	07:03:02
unspecified	tyrael
date	2014-02-05
devname	imperius
devid	FGT1KC3912800514
logid	0317013312
type	utm
subtype	webfilter
eventtype	ftgd_allow
level	notice
vs	"root"
policyid	128
identidx	0
sessionid	2385876107
user	"3d89c96d3f"
srcip	10.102.13.1
srcport	58376
srcintf	"LAN_AGGR"

dstip	74.125.233.72
dstport	80
dstintf	"INTERNET"
service	"http"
hostname	"www.youtube.com"
profile	"NMMU BLOCK"
status	"passthrough"
reqtype	"direct"
url	"/share_ajax?action_get_share_box=1&video_id=1vuuhaSdlKM&list=PLCB014CC7A5726588&share_at=true&caption_text=Uploaded by pannellc"
sentbyte	923
rcvdbyte	0
msg	"URL belongs to an allowed category in policy"
method	domain
class	0
cat	25
catdsc	"Streaming Media and Download"

Table 3.4 reflects the field comparison of the log entries for the pre-configuration error and the post-configuration error logs once the error had been rectified. Certain fields were included in the post-configuration error log file entries, but not in the pre-configuration error log file entries, as listed below.

- eventype
- level
- vs
- identid
- sessionid
- user
- hostname
- profile
- reqtype
- url
- msg

- method
- class
- cat
- catdesc

With reference to table 3.4, cells highlighted in blue were unique to the post-correction logs, cells highlighted in pink were unique to pre-correction logs and cells highlighted in green were common to both log files.

Table 3.4 - Log files entry field comparison

Pre correction	Post correction
Month	month
Day	day
Time	time
unspecified	unspecified
date	date
devname	devname
device_id	devid
log_id	logid
type	type
subtype	subtype
src	srcip
src_port	srcport
src_int	srcintf
dst	dstip
dst_port	dstport
dst_int	dstintf
service	service
status	status
sent	sentbyte
rcvd	rcvdbyte
policyid	msg

pri	method
vd	class
Timestamp	cat
SN	catdsc
dst_country	eventtype
src_country	level
tran_sip	vs
tran_sport	identidx
proto	sessionid
duration	user
	hostname
	profile
	reqtype
	url

The only usable Web usage field in the pre-configuration error logs was the destination IP (dst) field, which is the IP address of the website or Web service used and is indicated by a sequence of numbers, e.g. 173.236.49.82. These dst field values still require a link to the hostname of the website to be meaningful. However, the post-configuration error logs contained crucial Web usage related fields. These included but were not limited to:

- The hostname field is the actual name of the website visited in that entry
- The url field is available to provide further information on the activity that took place within the website
- The user field uniquely identifies a single Web user amongst the entries.

The post-configuration logs were richer in Web usage related fields than the pre-configuration error logs.

3.4 Conclusion

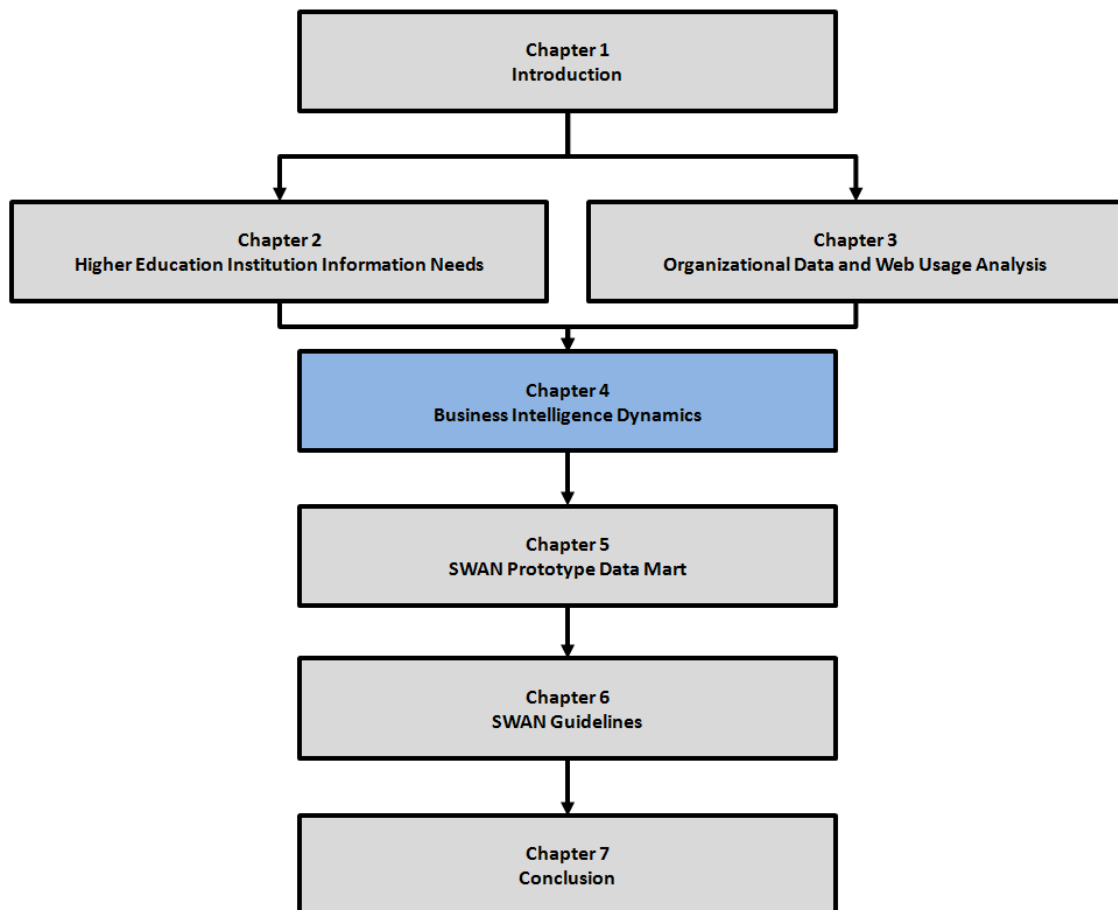
Log files are rich in valuable information. The NMMU system engineers made large numbers of detailed log files available for analysis. Analysis of these log files could provide a great deal of meaningful information about student Web usage to educators. Previously established information

needs and desired Web usage behaviour of students could be better addressed with this information. However, the log files needed to undergo a process of analysis to be meaningful and relevant.

The analysis of operational data with the intention of better serving the system that produced it lies within the domain of BI. The approaches, methods and heuristics of this well-established field needed to be investigated and applied to extract the true value from the available data. The logs available for analysis at the NMMU contained more detail than those used in a previous study at a South African higher education institution, which successfully derived useful Web usage information (Hazelhurst et al., 2011). The logs discussed in this study could thus be regarded as a viable data source of meaningful information about student Web usage behaviours at NMMU.

Chapter 4 - Business Intelligence Dynamics

“We’re not in an information age anymore. We’re in the information management age.” \rightarrow Chris Hardwick



This chapter examines the domain and dynamics of Business Intelligence. The components, development methods and application of a Business Intelligence system are discussed. Finally, an appropriate Business Intelligence system development method for this study is selected.

4.1 Introduction

A key aspect of effective decision-making is timely foundation and feedback information (Larson, 2006). Relevant and correct information about an organisation and its processes can be extremely valuable, especially to the organisation itself. Many business processes within an organisation are driven by decisions made by its employees. BI systems aim to provide these employees with concise factual information to facilitate better decision-making (Gangadharan & Swami, 2004; Nemati, Steiger, Iyer, & Herschel, 2002). It is clear that information is an influential factor in efficient decision making and any useful information could affect the quality of decisions.

Information is derived from data (Bellinger et al., 2004). Almost all business processes and activities will produce some data. Data, in some form of record or output, results from a business process and will provide an account of this process. For example, if a customer purchases items at a hardware store, the register or point of sales system will produce and store an invoice which captures data about the purchase. Another example would be if an insurance company processes a claim; the details of the claim will be captured and stored in a certain way, depending on the system used. If this data is collected and analysed, valuable information can be extracted and presented to decision makers to improve and optimise the processes. However, the task of transforming data into useful information for improved decision-making can be complex and differs from business to business. The domain that seeks to undertake these tasks is known as business intelligence (BI) (Al-Debei, 2011).

The undertakings of this chapter contribute to the completion of step 2 of the DSRP. Possible solutions are explored through investigation of a subject domain relevant to the problem.

4.2 Business Intelligence

As Scheeps (2013) observes, no single definition of BI exists. Vendors and consultants tend to define it in terms of their speciality. Academics, authors and consultants use a variety of definitions that are distinct and unique.

Moss and Atre (2003) describe BI as neither a product nor a system. "It is an architecture and a collection of integrated operational as well as decision-support applications and databases that provide the business community with easy access to business data" (p. 4). When defining BI, Negash (2004) refers to it as a system that can "combine data gathering, data storage, and knowledge management with analytical tools to present complex and competitive information to planners and decision-makers" (p. 178). Olszak and Ziemba (2003) consider BI to be a "set of concepts, methods and processes that aim at not only improving business decisions, but also at supporting realisation of an enterprise's strategy" (p. 856).

Despite the lack of consistency in the definitions of BI, its application generally results in the same elements, which are the technologies and tools for supporting decision-making (Schepps, 2013). BI is perhaps better understood when considering what it is used for or its end objectives. Negash (2004) believes that the objective of BI is to “improve the timeliness and quality of inputs to the decision process” (p. 178). Larson (2006) points to the objective of BI as providing "accurate, useful information to the appropriate decision-makers to serve as a foundation for the decision" (p.11), while Elbashir, Collier and Davern (2008) note that "BI systems provide the ability to analyse business information in order to support and improve management decision-making across a broad range of business activities" (p. 135). Finally, Imhoff et al. (2003) assert that BI seeks to "study past behaviours and actions in order to understand where the organization has been, determine its current situation, and predict or change what will happen in the future" (p. 3).

In this study, Schepps' (2013) definition of “any activity, tool, or process used to obtain the best information to support the process of making decisions” will be used (p. 11).

Despite the fact that the term “Business Intelligence” is fairly new, computer-based BI systems emerged, in various forms, around 40 years ago (Negash, 2004). BI, as an end goal or functional term, replaced management information systems, executive information systems and decision support and its origins are embedded in these early forms of decision support (Thomsen, 2003).

4.2.1 Brief History of Business Intelligence

The notion of decision support evolved from theoretical studies conducted in the late 1950s and early 1960s, as well as from technical work on interactive computer systems, undertaken predominantly at the Massachusetts Institute of Technology in the 1960s (Keen & Morton, 1978). Subsequently, decision support manifested itself in various types of information systems. The idea of decision support systems was first conceptualised in February 1964 by Michael Scott Morton (Power, 2007).

In the early 1960s, the emergence of mainframe computer systems afforded large companies the capacity to create Management Information Systems (MIS). These MISs made structured and periodic reports on business processes available to managers. A few years later, a type of new practical information system emerged in the form of Model Oriented Decision Support Systems or Management Decision Systems (Power, 2007).

In the 1970s, Decision Support Systems (DSS) emerged as applications designed specifically to support decision-making. A number of researchers and companies expanded the development of DSSs and it was recognised that these could be designed to support decision-makers at any level in an organisation (Power, 2007).

In the 1980s, academic researchers introduced software aimed at supporting group decision-making. These group DSSs were commercialised by a number of companies. Single user model-driven DSSs also emerged and gave rise to Enterprise Information Systems (EIS). EISs are software packages used to integrate and consolidate process based and transactional information and data within an organisation. EISs were developed in an attempt to provide business executives with vital performance information in an efficient and timely way (Davenport, 2011; Gosain, 2004).

In the 1990s, the term BI was used for the first time to describe a collection of methods to bolster business decision-making using fact based support systems. Data warehousing led to the expansion of the domain of EIS and data-driven DSSs, and data warehousing became one of the most important developments in the information systems field. During this period, Bill Inmon and Ralph Kimball encouraged an approach to the design and creation of DSSs that used relational database technologies. Constructing data-driven DSSs was Inmon and Kimball's focal point (Power, 2007; Watson & Wixom, 2007).

In 2007, a survey of 1400 chief information officers (CIOs) conducted by the Gartner company indicated that BI projects had become the priority for a number of them (Watson & Wixom, 2007). Currently, successful enterprises that do not use BI to improve their business results are rare (Chaudhuri et al., 2011). An estimated 95% of the Fortune 1000 companies have a data warehouse implemented or intend to develop one (Wixom & Watson, 2001). Moreover, data warehousing has been cited as the highest-priority post-millennium project by many information technology executives (Sen & Sinha, 2005). Many methodologies and tools have been created to assist in the development of BI systems, among them data warehouses (Sen & Sinha, 2005).

4.2.2 Business Intelligence Acquisition

Various methodologies, techniques and views exist on obtaining BI (List & Bruckner, 2002; Saroop & Kumar, 2011). However, BI systems share many common aims and desired outcomes (Scheps, 2013). Therefore, common core components are likely to become apparent when attempting to develop a BI system.

4.2.2.1 Components of a Business Intelligence System

Essentially, in order to facilitate BI, relevant data needs to be acquired, stored and made available to decision-makers in a format that is readable, relevant and understandable, as well as dynamic, to allow for improved decision-making that positively affects the well-being of the organisation. BI converts data into useful information and, through human analysis, into knowledge (Negash, 2004).

Most information systems are implemented as a computer system consisting of various software applications or information technology tools. BI systems exist predominantly as software applications on a computer system within some form of IT infrastructure. In order to operate correctly, BI systems usually rely on certain highly oriented IT tools and infrastructure. These include query and reporting tools as well as customised data bases such as data warehouses and data marts (Elbashir, Collier, & Davern, 2008). A data warehouse is a vital component and a central aspect of any BI system (Inmon, 2002; R Kimball, Ross, Thorthwaite, Becker, & Mundy, 2008; Scheps, 2013).

Figure 4.1 illustrates the three core facets of a BI system (Chaudhuri et al., 2011; Gangadharan & Swami, 2004; R Kimball et al., 2008; Rainardi, 2008):

- Data warehouse
- Extract Transformation Load
- Access/User Application

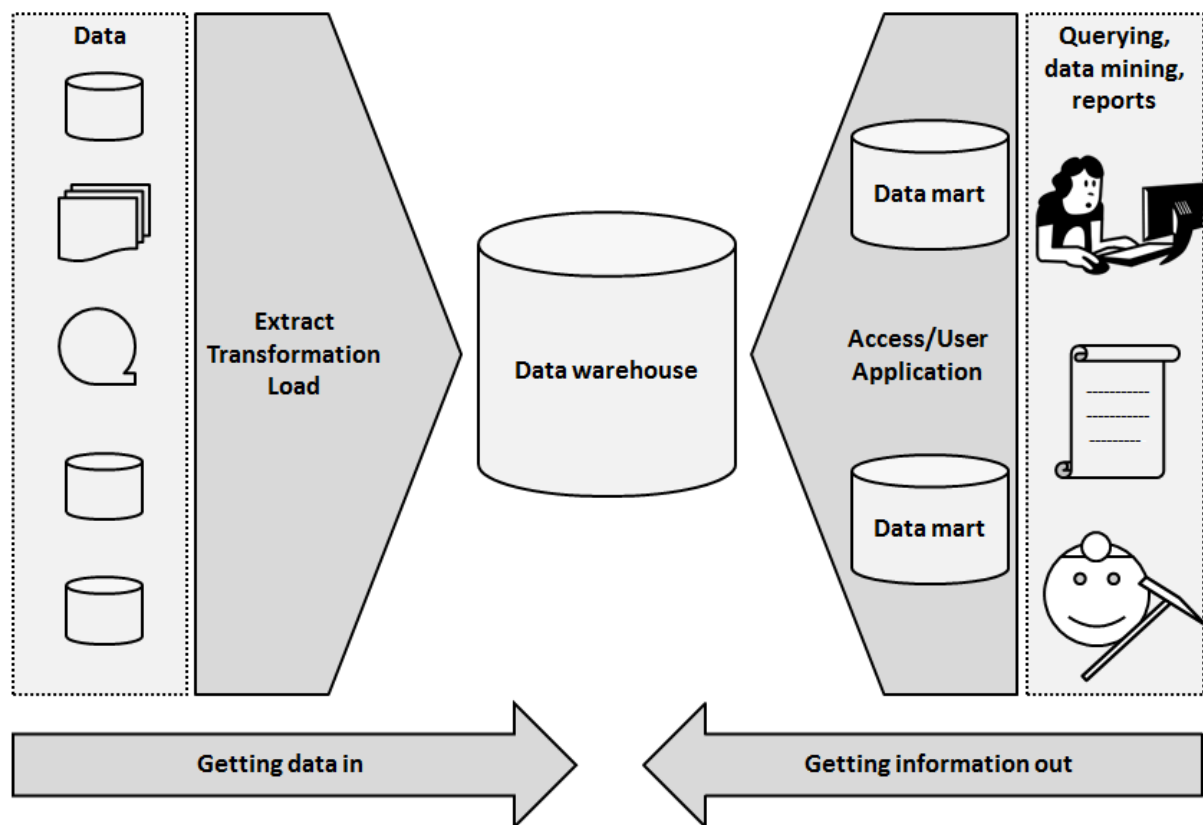


Figure 4.1 - A typical business intelligence system architecture adapted from Imhoff et al. (2003) and Watson & Wixom (2007)

As shown in Figure 4.1, the data warehouse stores raw data extracted from various sources. However, before the raw data is placed in the data warehouse it is subjected to a cleaning and

structuring process. This is commonly referred to as *Extract Transformation Load* (ETL) (Scheeps, 2013; Simitsis, Vassiliadis, & Sellis, 2005). Once the data is appropriately transformed and loaded into the data warehouse, it can be accessed by users and analysts in some form of BI application. This can be done using querying, reporting or data mining tools (R Kimball et al., 2008; Simon, 2009).

Data Warehouse

A data warehouse is a repository for data, structured to support decision-making (Bruckner, List, & Schiefer, 2002; M Golfarelli & Rizzi, 2009). It is an inherent part of a BI system, so much so that Kimball et al. (2008) refer to a complete end-to-end BI system as “data warehouse/business intelligence” (p. xxxi) (DW/BI), which demonstrates the strong relationship between data warehouses and business intelligence. Inmon (2002) believes that the data warehouse is the foundation of all DSS processing.

The World Wide Web has allowed global real time communication and has developed into a massive business hub. Data produced by Web activity is used in various ways in BI (Abraham, 2003). Data warehouses that are designed to store Web activity can be referred to as data webhouses (R Kimball & Merz, 2000). A form of Web data known as clickstream data comprises an account of each click made by a Web user while browsing (Hu & Zhong, 2005; R Kimball & Merz, 2000). By exporting clickstream data into a data webhouse, information about Web user activity in various contexts can be analysed (R Kimball & Merz, 2000).

Data warehouses that are developed specifically for a single department or business area are known as data marts (Larson, 2006). Data marts consolidate data and present this in a way that is tailored to a particular group of users, or to a single business process (R Kimball et al., 2008). Another important process is the loading of data into the warehouse, in addition to presenting it to analysts (Inmon, 2002).

Extract Transformation Load (ETL) — Getting Data In

The task of getting data into a data warehouse is commonly broken down into three sub tasks: extraction, transformation and loading (ETL). Extraction refers to the gathering of the data from its original source location/s. Transformation involves suitably altering the data. Loading simply means inserting the altered data into the data warehouse (R Kimball et al., 2008; Larson, 2006; Simitsis et al., 2005).

The ETL process, and the creation of the systems to facilitate this process, is estimated to make up the majority of BI system development (R Kimball et al., 2008; Watson & Wixom, 2007). An ETL

system is a set of routines that automatically performs the ETL process (Rainardi, 2008). The ETL processes ensure that the data is retrieved and cleansed in order to remove errors and to ensure that it is of a high quality (Simon, 2009).

Once the transformed data has been loaded into the data warehouse it must be made appropriately accessible to the relevant users.

Access and User Application — Getting Information Out

Users of a BI system will typically be knowledge workers, DSS analysts or anyone who requires specific information for decision-making (Inmon, 2002). In order for these users to access the dormant information within a data warehouse and to gain real value from it, some access or application must be made available to them (Watson & Wixom, 2007). Kimball et al. (2008) explain that the methods used to gather information from a data warehouse are known as BI applications. BI applications take the form of various information retrieval methods (R Kimball et al., 2008).

The most basic method of information retrieval is a *query*. A query is simply an electronic request for specific information from the data warehouse and is common in the database domain because it is so simple. Querying tools are available to assist users in translating their requests into queries through an intuitive interface (Scheps, 2013). *Reporting* is a second information retrieval method and is very similar to querying. However, reporting provides information in more intricate formats and layouts. Reporting tools allow the user more complex input and control options with regard to how the information will be presented (Simon, 2009). A common technique for calibrating queries and reports is Structured Query Language (SQL). SQL is a formalised set of keywords and parameter structures that allow for the retrieval of specific information, based on how the SQL query or report is constructed (Hand, Mannila, & Smyth, 2001; Scheps, 2013).

Data mining is a sophisticated method of information retrieval, typically applied to large amounts of data, and uses intelligent means to identify patterns within the data (Reddy & Srinivasu, 2010). Data mining uses complex mathematical algorithms to process detailed data in order to identify patterns, correlations and clustering within the data (Larson, 2006). This form of information retrieval is capable of predicting trends and forecasting future patterns by analysing historical data (Hirji, 2001). These trends and patterns can be extremely valuable and, through human interpretation, are able to provide a vital competitive edge in the business domain (Hofgesang, 2009).

Most BI applications are presented through the use of a *BI portal*. BI portals are a collection of BI applications working together to present information to a user. Essentially, BI portals are the interface and front end of a data warehouse. The design and requirements for a BI portal are defined by the business that intends using it (Kimball et al, 2008).

4.2.2.2 Methodologies for Developing Business Intelligence Systems

The use of BI systems has become very common in business and these days they are considered a priority (Watson & Wixom, 2007). Developing a BI system can be extremely complicated and the process requires guidance from domain experts and specialists. Despite some lack of consistency within the BI industry, many related guidelines and methodologies are available to organisations wishing to develop BI systems (Sen & Sinha, 2005). It is important that an appropriate methodology or set of guidelines is adopted (Moss & Atre, 2003).

Within software engineering, the subset which observes the process followed in creating information systems is called the system development life cycle (SDLC). Two divergent styles of SDLC exist: waterfall and iterative (Rainardi, 2008).

Waterfall Methodology

The waterfall methodology is sequential. Once a step is complete, the next step is initiated and so on until all steps have been completed; each step is thus reliant on the completion of the previous step (Inmon, 2002; Moss & Atre, 2003; Rainardi, 2008). Figure 4.2 illustrates how a sequence of generic steps is followed in the waterfall methodology:

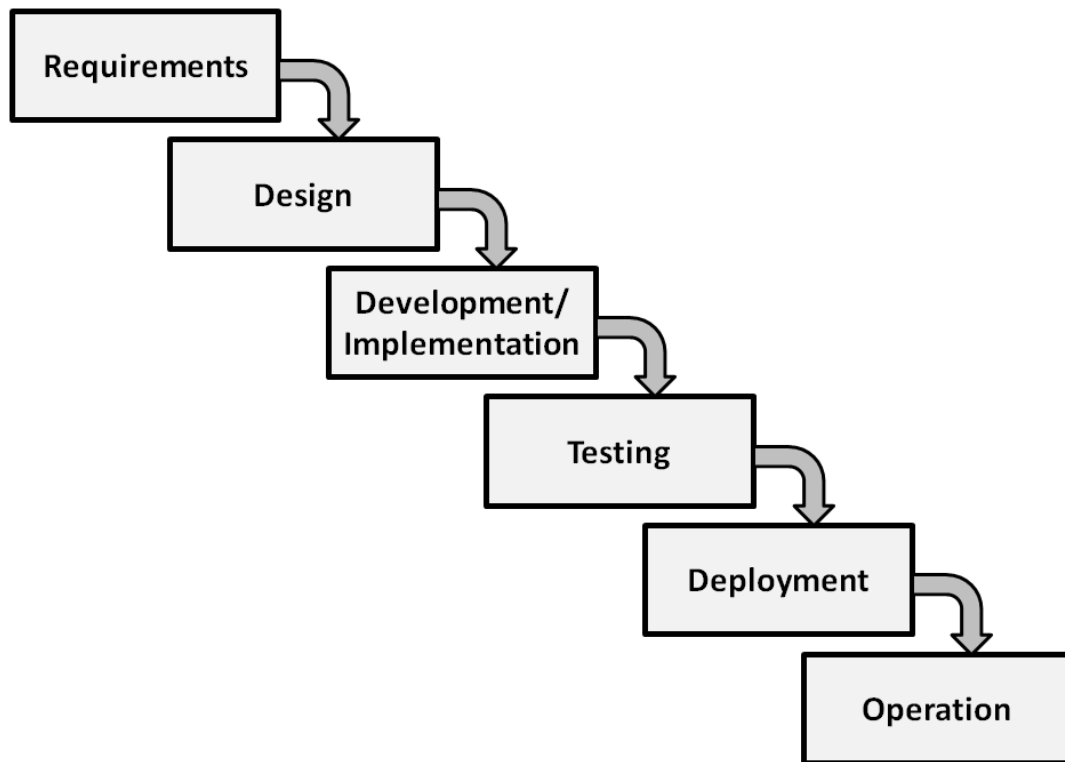


Figure 4.2 - Generic waterfall methodology lifecycle adapted from Ragunath (2010) and Rainardi (2008)

Iterative Methodology

The basic principle underlying iterative methodology is a focus on completing a part of the system. Each part of the system undergoes design, development, testing and deployment; each iteration deploys one part of the complete system. Once each part of the system has been deployed and is operational the complete system is in place (Rainardi, 2008).

The iterative methodology can be seen as a spiral, where each cycle represents an iteration (Rainardi, 2008). Figure 4.3 illustrates how the iterative cycles are completed, including the architecture design at the beginning of iteration 1:

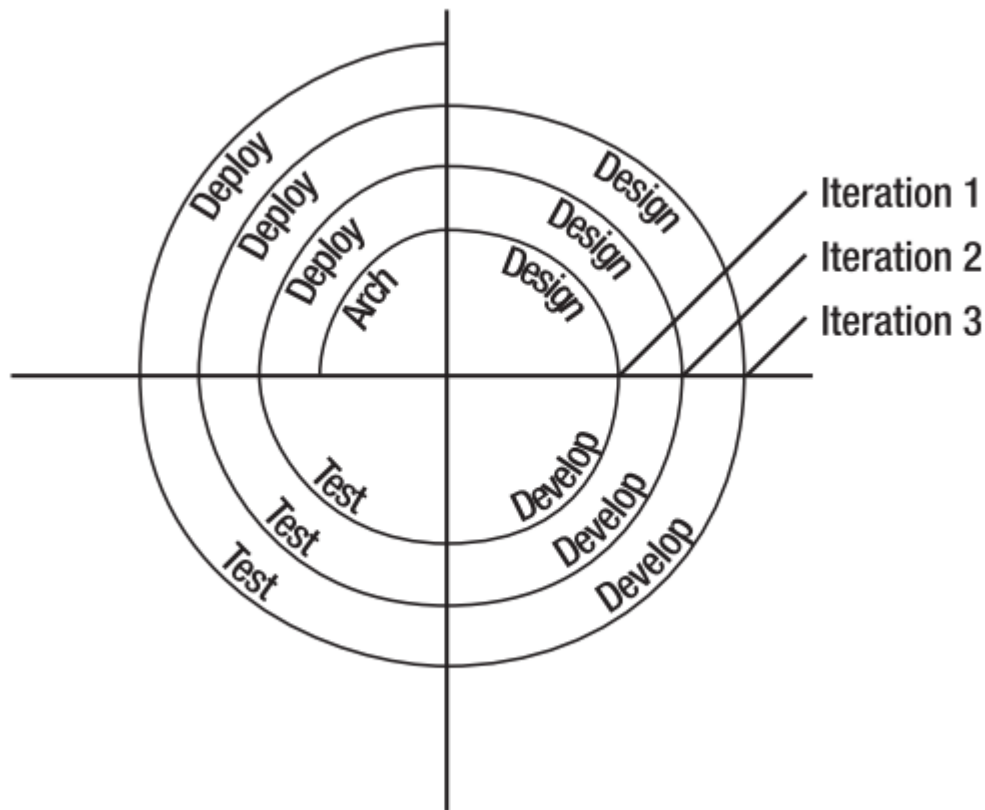


Figure 4.3 - Iterative cycles (Rainardi, 2008)

Waterfall versus Iterative Methodologies for Data Warehouse Projects

Waterfall and iterative methodologies are inherently different. The waterfall methodology is pervasive within many organisational system development projects, including data warehouse projects. Iterative methodologies are not as widely used but can be extremely beneficial if used correctly (Rainardi, 2008). In order to select an appropriate methodology, the advantages and disadvantages of each should be considered together with the recommendations of data warehousing industry experts. In addition, the requirements and scope of the data warehouse project at hand must be taken into consideration.

The advantages of the waterfall methodology include, but are not limited to:

- Good model for small organisations that have limited requirements (Obeidat & Nasereddin, 2013).
- Reinforces management at each step by checking deliverables of each completed step (Obeidat & Nasereddin, 2013).
- At each phase proper documentation is followed to ensure the quality of the development (Balaji & Murugaiyan, 2012).

- Requirements are clear before development starts (Balaji & Murugaiyan, 2012).
- Minimal resources are required to implement this model (Balaji & Murugaiyan, 2012).
- As it is a linear model, it is easy to implement (Balaji & Murugaiyan, 2012).
- Each phase is completed within a specified period; thereafter, the process moves on to next phase (Balaji & Murugaiyan, 2012).
- It is easy to determine the costs of the project, to establish a schedule and to allocate resources accordingly (Leau, Loo, Tham, & Tan, 2012).

The disadvantages of the waterfall methodology include, but are not limited to:

- Not all problems within a step are solved during the step itself because many can only be flagged after the step has been completed (Balaji & Murugaiyan, 2012).
- Should there be a change in a client's requirement, this cannot be implemented in the current development process (Balaji & Murugaiyan, 2012).
- May result in a slow development process as a result of the time spent on completing a single step before proceeding (Inmon, 2002).
- Not well suited to systems that require integration across departments within an organisation (Moss & Atre, 2003).
- The testing phase can carry high risks due to first time scenarios close to deployment, such as users seeing the data warehouse; all components being run together; test and production environments being used; running at maximum capacity (Rainardi, 2008).
- Does not support object oriented projects (Obeidat & Nasereddin, 2013).

The advantages of the iterative methodology include, but are not limited to:

- Users are exposed to the system during the development, which results in the detection and potential correction of errors (Rainardi, 2008).
- Eliminates first time risks associated with the waterfall methodology testing phase, because these are eliminated in the first iteration that is completed during the early stages of development (Rainardi, 2008).
- Results can be seen early in the development process (Inmon, 2002).
- Obviates lengthy development cycles, which may result in user requirements being overlooked (Simon, 2009).
- Feedback from users is integrated into the development of successive versions produced in subsequent iterations (Moss & Atre, 2003).

The disadvantages of the iterative methodology include, but are not limited to:

- May require the infrastructure to be delivered upfront (Rainardi, 2008).
- Using the spiral model, a type of iterative methodology, is costly and requires specific expertise for risk analysis (Kute & Thorat, 2014).

4.2.2.3 Data Warehouse Industry Recommendations and Justifications

Moss and Atre (2003) observe that a waterfall methodology is ill suited to BI decision support application release – instead, an agile development methodology tailored to BI decision-support applications should be used. Kimball et al. (2008) recommend that the data warehouse environment should be developed iteratively in manageable increments.

Inmon (2002) explains that a data warehouse is developed using a type of iterative methodology known as a spiral development method. This method warrants the creation of small sections of the data warehouse. Reasons for the importance of iterative development include, but are not limited to (Inmon, 2002):

- Its history of success in industry advocates its use.
- The end user cannot provide truly valuable requirements or foresee particular requirements until the first iteration has been completed.
- Providing quick, feasible and clear results will strengthen buy-in by management.
- Clear results can be seen early in development.

Sen and Sinha (2005) compared a range of different data warehousing methods for use in certain data warehousing projects. The majority (12 out of 15) of these methods were iterative. Sen and Sinha (2005) considered their collection to be a fair representation of the range of available methods at the time, which suggests that iterative methodologies were successful. Furthermore, a second study by Sen and Sinha (2007) found that of 30 commercial data warehousing methods analysed, iterative methodologies were by far the most popular. Simon (2009) suggests that in order to deliver a data mart, an iterative methodology should be followed because it allows for quick implementation which will maximise its business value.

There is a clear inclination towards using an iterative over a waterfall methodology in data warehouse development. Therefore, an iterative development methodology will be used for the purposes of this research study.

Iterative development methodology dictates the overall method of development that should be adopted for a data warehousing project. However, within these iterative development methods, various views and opinions exist as to how the data warehouse should actually be built, i.e. the detailed tasks within each iteration. The advantages, disadvantages, situational considerations and

industry experts' empirical views should all be considered in order to establish guidelines for this research.

Despite many divergent methods in the data warehouse industry, two core standards are widespread. All data warehouse development methods are generally based on either the top-down spiral method or the bottom-up dimensional life cycle method (Breslin, 2004; Sen & Sinha, 2007). One of the fundamental questions when building a data warehouse is which of these two methods to adopt (Ponniah, 2004).

Top-Down Method

Large organisations usually comprise departments that together make up the organisation. For example, there could be a human resource department, accounting department, marketing department, procurement department and so on. Each department has specific information needs depending on the nature of their operational activities (Inmon, 2002).

The top-down method seeks to build a large, enterprise-wide data warehouse tailored to providing for the entire organisation's needs and goals. This will be accessible by each department, thereby feeding all departments with business data (Breslin, 2004; Ponniah, 2004).

Figure 4.4 depicts the overall structure of a data warehouse using a top-down method. Operational data is fed into the data warehouse. Through transformative processing, the warehouse structures the data in such a way that it facilitates querying and reporting. This information is then detailed further into departmental specific data marts which serve each department by providing access in a way that suits its need for various levels of detail (Inmon, 2002).

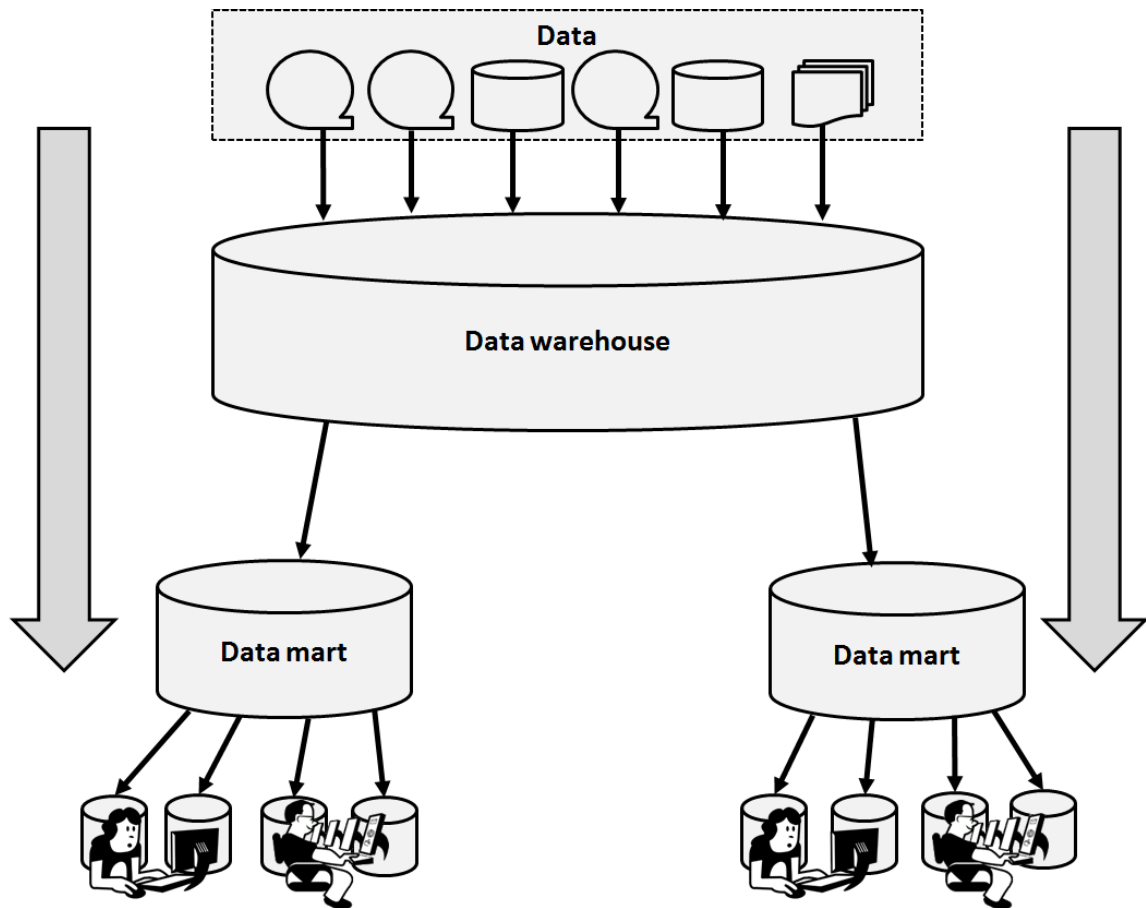


Figure 4.4 - The structure of a top-down data warehouse (Inmon, 2002)

Figure 4.5 illustrates the developmental lifecycle of the top-down method by which the data warehouse is developed and implemented, after which the programs to run a DSS are developed. These components will in turn both reveal and cater for the requirements of the user.

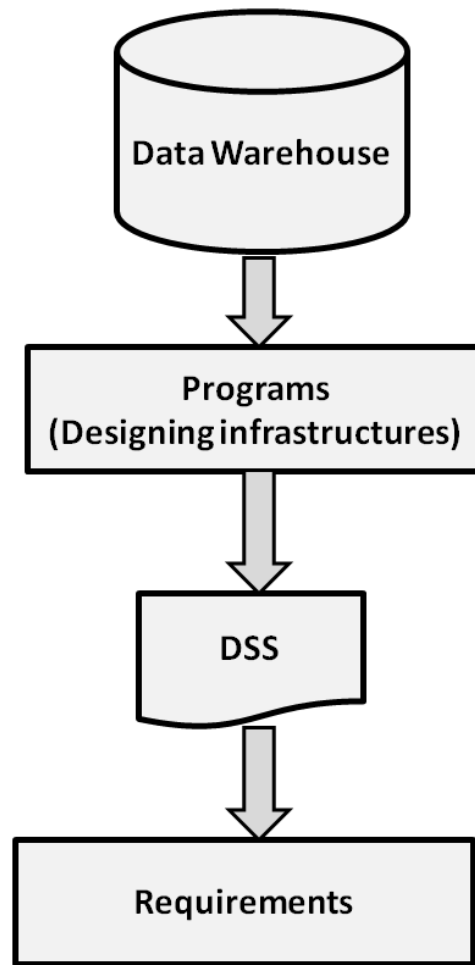


Figure 4.5 - Data warehouse SDLC (top-down) (Inmon, 2002)

The advantages of this method include (Ponniah, 2004):

- It is a collective and cooperative business effort that provides a high level, holistic perspective of the data with the entire enterprise in mind.
- It is structured to be universal, not a collection of department specific data marts.
- Metadata is stored centrally.
- Principles, management and authority are centralised.
- If carried out in iterations, results occur early.

The disadvantages are (Ponniah, 2004):

- Even with an iterative methodology, development is prolonged.
- It carries a higher risk of failure.
- Requires specialised, diverse and overlapping skill sets.
- Initial costs are high without proof of concept.

Bottom-Up Method

The bottom-up method is used when departmental specific data marts are developed one at a time in an iterative manor. Each data mart is based on the needs of one department. The collection of all the departmental data marts makes up the overall data warehouse system; essentially, each data mart is a component of the overall data warehouse system (M Golfarelli & Rizzi, 2009; R Kimball et al., 2008; Ponniah, 2004). Figure 4.6 depicts the overall structure of a bottom-up data warehouse as the collection of operational, department specific data marts.

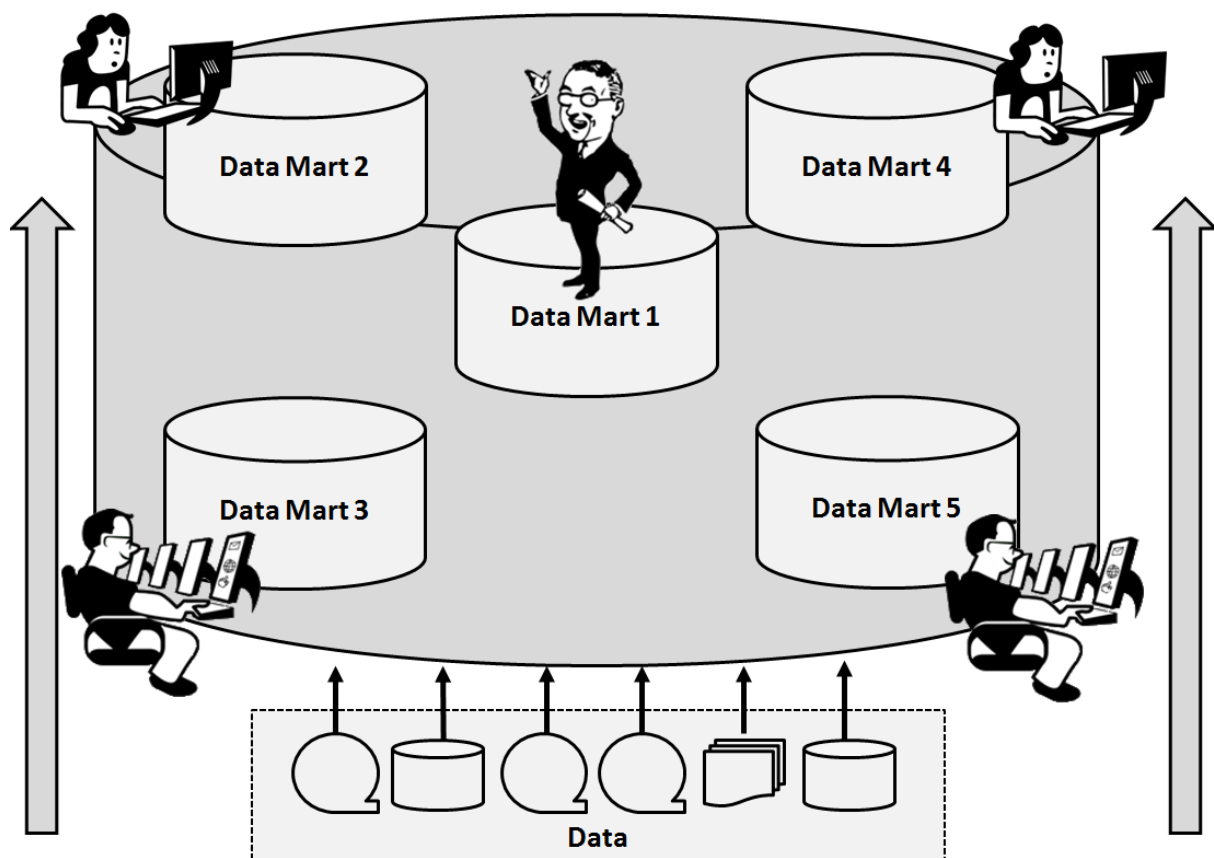


Figure 4.6 - The structure of a bottom-up data warehouse adapted from Golfarelli & Rizzi (2009)

Data is fed into the data marts from heterogeneous sources and handled according to the processes in place in the individual data marts. Figure 4.7 shows the developmental lifecycle of the bottom-up method, in which requirements are gathered and the programs needed to facilitate them are designed and developed.

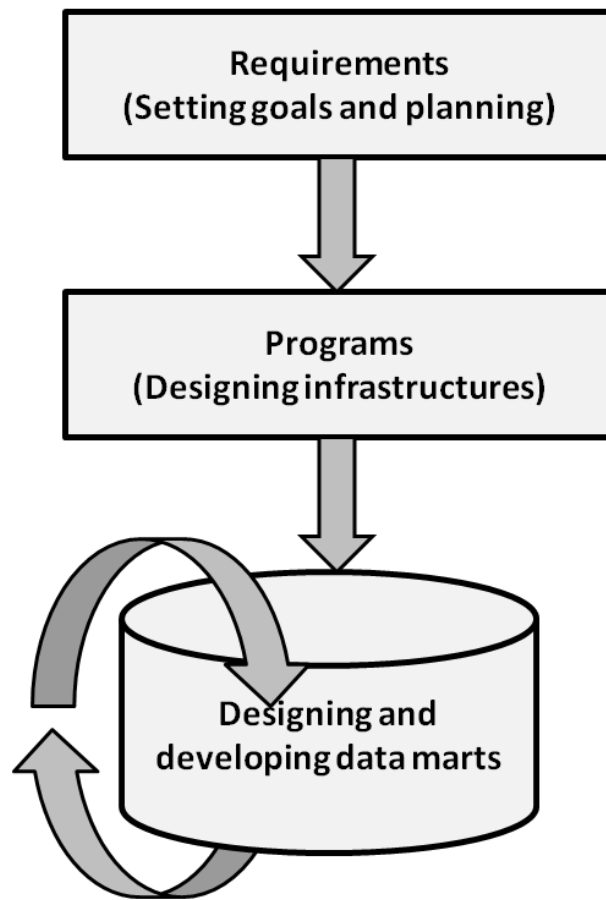


Figure 4.7 - Data warehouse SDLC (bottom-up) adapted from Golfarelli & Rizzi (2009)

Data marts are thus built in iterative cycles for each department. Once each department has a data mart in place, the overall data warehouse is operational, i.e. it is the collection of data marts working together through integration.

Advantages of this method are (Ponniah, 2004):

- Deploying smaller parts is quicker and less demanding
- This method results in desirable return on investment, proof of concept and apparent value
- Risk of failure is reduced
- The incremental nature of this method allows for the prioritisation of key data marts
- Project teams can gain invaluable experience.

Disadvantages include (Ponniah, 2004):

- Every data mart has a specific and limited perspective of the data
- Results in data redundancy throughout all data marts because data is duplicated regularly

- Sustains irregular and alienated data
- May cause troublesome interface management.

4.3 BI System Development Method Argument and Discussion

A number of relevant methods have been discussed and compared. This section will outline which of these will be used in this study, based on scope limitations and other considerations.

The data warehouse project in this study was limited in that it sought to provide BI for a single department. This was done without the intention of providing an entire enterprise or university scale data warehouse.

Using a top-down method involves a corporate effort that requires buy-in from the entire enterprise (Breslin, 2004; Inmon, 2002; Ponniah, 2004). It carries a higher risk of failure and more strenuous development efforts than a bottom-up method (Ponniah, 2004). This research study did not require an enterprise wide view of the data; it sought only to use department specific data. Furthermore, within the NMMU, there is a diversity of departments, subjects and lecturing approaches and subsequently, many distinct information needs. A top-down approach is inclined towards the creation of a one-size-fits-all data warehouse, from which all departments extract data as needed (Inmon, 2002; Ponniah, 2004). This means that an organisation-wide data warehouse must first be completed before the individual departments can access the data. A top-down approach would thus not initially allow for the provision of information according to the specific needs of lecturers at the School of ICT. A top-down approach focuses on the requirements of the entire organisation and not on those of a particular department, even though they may have strong ties (Ponniah, 2004). Therefore, a top-down method was not well suited to the proposed development of the prototype.

A bottom-up method seeks to create data marts for one department at a time and to build the data warehouse, which is the sum of all the developed data marts, incrementally (Breslin, 2004; R Kimball et al., 2008; Simon, 2009). This approach seeks to understand the requirements of the business user, in this case the lecturer, and to attempt to provide for those ad hoc requirements by building a data mart (Breslin, 2004). The proposed requirements for the prototype would only require a part of a DW/BI system. A bottom-up method is more suited to developing smaller parts than a top-down method (Ponniah, 2004). Moreover, using a bottom-up approach allows for integration with other data marts in future research (Matteo Golfarelli, 2010). Therefore, a bottom-up method was more appropriate for this project.

An industry thought leader in bottom-up data warehouse development and the author of numerous contributions to data warehousing is Dr Ralph Kimball. He is considered a pioneer in the data warehousing field and has initiated Web enabled data warehouses termed data Web houses

(Alsqour, Matouk, & Owoc, 2012; Simon, 2009; Ponniah, 2004). He advocates the use of a bottom-up method with dimensional modelling (Breslin, 2004). In this approach, the fundamental concept of assembling data dimensionally is realised through the use of a technique known as *dimensional modelling*. Dimensional modelling arranges data in a way that allows for easier querying of data by a specific end user. Kimball and Merz (2000) have refined the dimensional modelling of clickstream data specifically. Clickstream data is the primary data type used in this research. This method is clearly and comprehensively documented in the “Data Warehouse Life Cycle Toolkit” (R Kimball et al., 2008). Kimball is thus the author of both the data webhouse concept and the Kimball Lifecycle method, and as these are based on the same underlying concepts they can be seamlessly synchronised. For this reason, the bottom-up dimensional life cycle method as documented by Kimball et al. (2008) was applied in this study.

Kimball et al. (2008) observe that a successful instance of a DW/BI project encompasses the merging of various components and essential tasks. These components and tasks must be appropriately harmonised with sequential milestones, as illustrated in Figure 4.8.

Kimball et al. (2008) maintain that successful data warehousing using the bottom-up dimensional life cycle method relies on three fundamental principles:

- The business is the focal point
- Dimensionally assembled and arranged data are presented to the business through specialised queries and reports
- The data warehouse system is developed iteratively through viable lifecycle increments instead of in one large developmental endeavour.

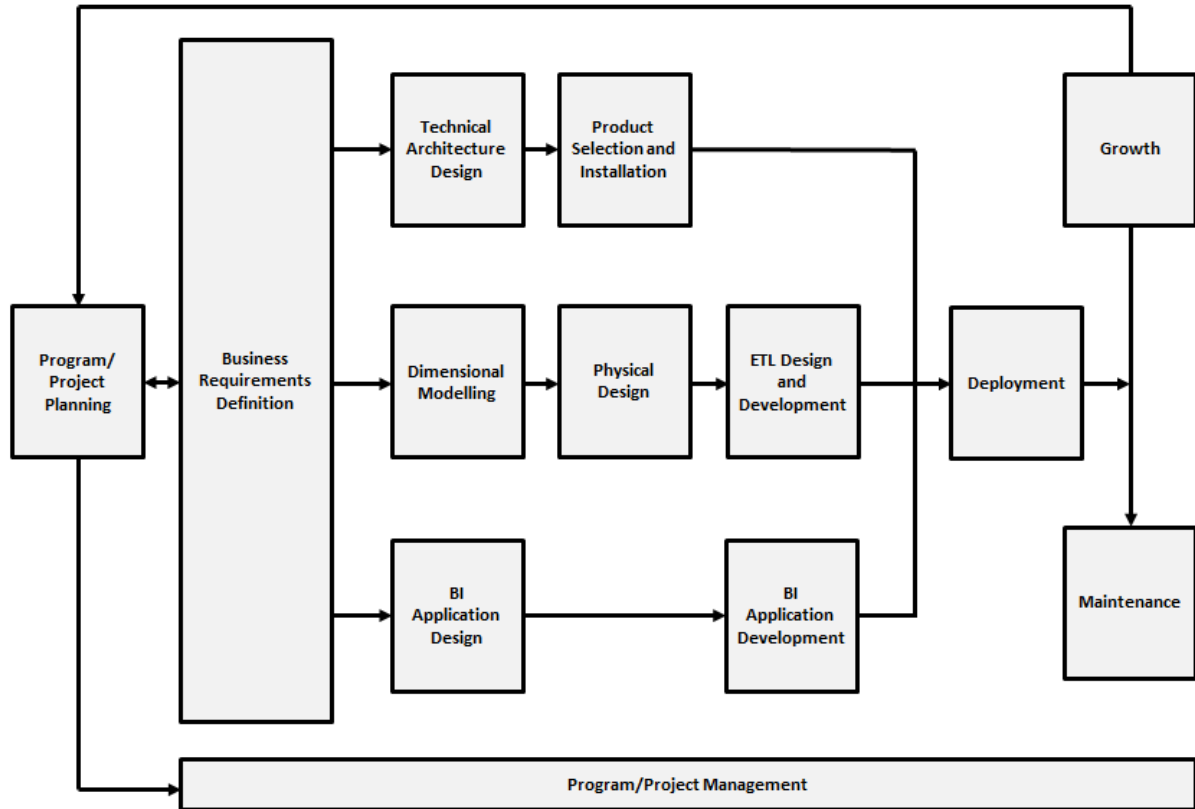


Figure 4.8 - The Kimball Lifecycle diagram (R Kimball et al., 2008)

The three fundamental principles mentioned above encompass the underlying virtues of the Kimball Lifecycle. Figure 4.8, the Kimball Lifecycle diagram, illustrates the associated tasks and milestones in the process of building a data mart that will result in a data warehouse system. Each milestone plays a role in this lifecycle and comprises various tasks and activities that must be completed. The milestones in Figure 4.8 will be discussed in the following sections (Kimball et al., 2008). The Data Warehouse Lifecycle Toolkit (Kimball et al., 2008) is used for reference in sections 4.3.1 to 4.3.9, as they directly explain and discuss the Kimball Lifecycle from this source.

4.3.1 Program/Project Planning

In the first task, Program/Project Planning, “project” refers to a single iteration of the lifecycle, i.e. the creation of a single data mart, whereas “program” concerns the overall development of the DW/BI system and all associated projects. In this study, the program received little focus as the intention was to stop once a data mart had been successfully built.

This milestone is reached when the scope of the project has been established, including resource allocation, task identification, assignment, duration and sequencing.

4.3.2 Program/Project Management

Program/Project Management is an ongoing facet of the lifecycle and spans all tasks following Program/Project Planning. Reaching this milestone involves ensuring that all tasks remain on track. In addition, it focuses on overseeing the project status, monitoring issues and maintaining scope and includes creating the means to mediate business and IT stakeholder communications.

4.3.3 Business Requirements Definition

A Business Requirements Definition involves gathering and understanding the needs of the business user whom the data mart intends to serve. A concrete understanding of the business user's requirements vastly improves the chances of a successful project.

The effective definition of business requirements is vital as this forms the foundation of all subsequent lifecycle tasks. It is in this way that Kimball's principle of focusing on the business is incorporated into the data warehouse system.

This milestone of the lifecycle includes communication with business users in order to understand their requirements and to establish the key factors driving their business. By understanding these key factors, requirements can be successfully translated into the project's design considerations.

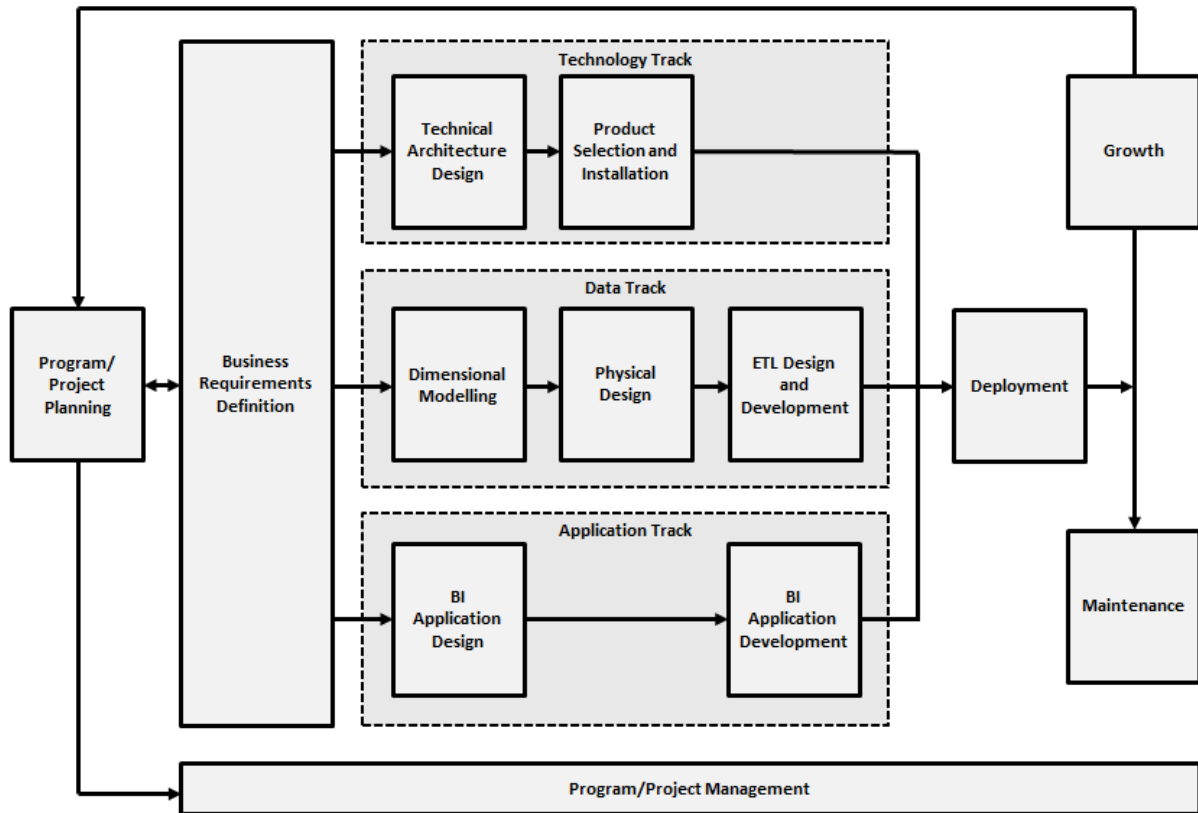


Figure 4.9 - The Kimball Lifecycle tracks adapted from Kimball et al. (2008)

Once the Business Requirements Definition milestone has been reached, the next phase of the lifecycle is entered, comprising three parallel sets of tasks. Each set is responsible for an aspect of the DW/BI project, defined as a track. The business requirements provide primary input for these three tracks. As indicated in Figure 4.9, these are the Technology, Data and BI Application tracks. The milestones for each track feed input to the succeeding milestone. There are implied dependencies between certain milestones, represented by their vertical alignment with milestones in other tracks. Parallel milestones may rely on each other for some input.

4.3.4 Technology Track — Technical Architecture Design and Product Selection and Installation

4.3.4.1 Technical Architecture Design

Technical Architecture Design is the establishment and documentation of the technological needs of the data warehouse project, based on the desired outcome and other factors. Essentially, it is a thorough investigation of any IT facilities that will be needed to facilitate the DW/BI system. The results of the investigation will be a form of framework which specifies the required technologies and integration considerations.

This milestone relates to business requirements, the current technical environment and the planned strategic technical directions required to establish a relevant technical architecture design.

4.3.4.1 Product Selection and Installation

Using the framework derived from Technical Architecture Design as a guide, the products required to facilitate the technical requirements of the data warehouse system are selected and installed. This framework features the two primary considerations for selecting products, that is, the functional and business requirements.

Reaching this milestone involves aspects such as the purchasing process, product evaluation, market research, short listing options and detailed evaluation, amongst others, which are to be taken into account when selecting the products that will cater to the business and functional needs of the data warehouse system. Once products have been selected and purchased they are then installed and made available for use.

4.3.5 Data Track – Dimensional Modelling, Physical Design and ETL Design and Development

4.3.5.1 Dimensional Modelling

Dimensional modelling is a data structuring technique that is used to deliver data in a comprehensible and timely manner. By using dimensional modelling, data can be structured to suit a specific information demand set, resulting in rapid querying. Essentially, if one knows what information is needed from the data, one can structure the data to cater for this by providing for an array of predictable queries. This will make the process more efficient because unnecessary data access is eliminated.

The Dimensional Modelling milestone involves investigating the current structure of available data and defining how it will be dimensionally modelled, based on the business requirements. It involves four steps, namely choosing the business process, declaring the grain (establishing exactly what information is being represented), identifying dimensions and identifying the facts.

4.3.5.2 Physical Design

Physical Design refers to the setting up of the product that is facilitating the data warehouse. The selected data base product must be calibrated in order to receive the data from the data source.

This milestone will include the replication of the dimensional model within a physical data base residing in the selected product/software.

4.3.5.3 ETL Design and Development

The ETL processes are responsible for delivering data to a data warehouse in a format that conforms to the dimensional model of that warehouse. This is done by retrieving data from heterogeneous sources, preparing them by configuring them so that they conform to their intended destination and then inserting them into the data warehouse. ETL is a massive undertaking and is done using automated processes carried out by software products and tools.

This milestone broadly involves the following tasks:

- Developing a plan for the ETL process with a source-to-target diagram at a high level, i.e. establishing where the source data is and where it needs to be
- Selecting a software tool for implementing the ETL process
- Developing strategies for error handling, dimension management and other processes
- Analysing the raw data to identify any complex restructuring needs or transforming and developing a preliminary sequence of jobs to fulfil these needs
- Building and testing historic dimension table loads
- Building and testing historic fact table loads, including surrogate key lookups and substitutions
- Building and testing dimension table incremental load processes
- Building and testing fact table incremental load processes
- Building and testing aggregate table loads
- Designing, building and testing the ETL system automation.

This milestone is regarded as one of the greatest challenges. However, the tasks and requirements for this milestone are driven by the scope of the DW/BI project at hand. Therefore, the complexity of this component is tied to the scope of this research.

4.3.6 BI Application Track – BI Application Design and BI Application Development

4.3.6.1 BI Application Design

BI Application Design refers to the design of the end user program or interface. The business user needs a way of accessing and interacting with the potential information residing in the data warehouse. This interface must be both relevant and usable. BI applications can deliver information by querying and analysing the dimensional models and presenting the results in the BI application in the form of reports, charts or other information delivering mediums.

This milestone comprises the presentation of relevant information to business users by considering their needs and technical capacity when designing the interface. Some users are capable of creating their own ad hoc queries while others are content to use predefined queries and reports from the data. These capabilities and needs are investigated at this point in the lifecycle.

4.3.6.2 BI Application Development

A BI portal or similar front end application is developed by taking into account the specifications gathered in the BI Application Design milestone. By integrating the business user's capabilities and preferences in the BI application the principle of focusing on the business is adhered to and queries and reports are made readily available to business users once the BI application is deployed.

4.3.7 Deployment

The Deployment milestone refers to the implementation of all preceding milestones. It is important that this implementation is well coordinated. It should be delayed if training, documentation and validated data are not primed for release. Extensive planning is required to ensure that all previous milestones have been fully tested and can be integrated.

4.3.8 Maintenance

Following the deployment and subsequent operation of the DW/BI system, various technical operational undertakings must occur to ensure that the system runs optimally. Amongst these are usage monitoring, performance tuning, index maintenance and system backups. In addition, the business users must be supported through education and communication.

4.3.9 Growth

Once the DW/BI system has been adopted and is running successfully, it is likely to expand. Using the previous milestones' tasks, some priority processes should be established to pass the growing business demands on to the DW/BI system.

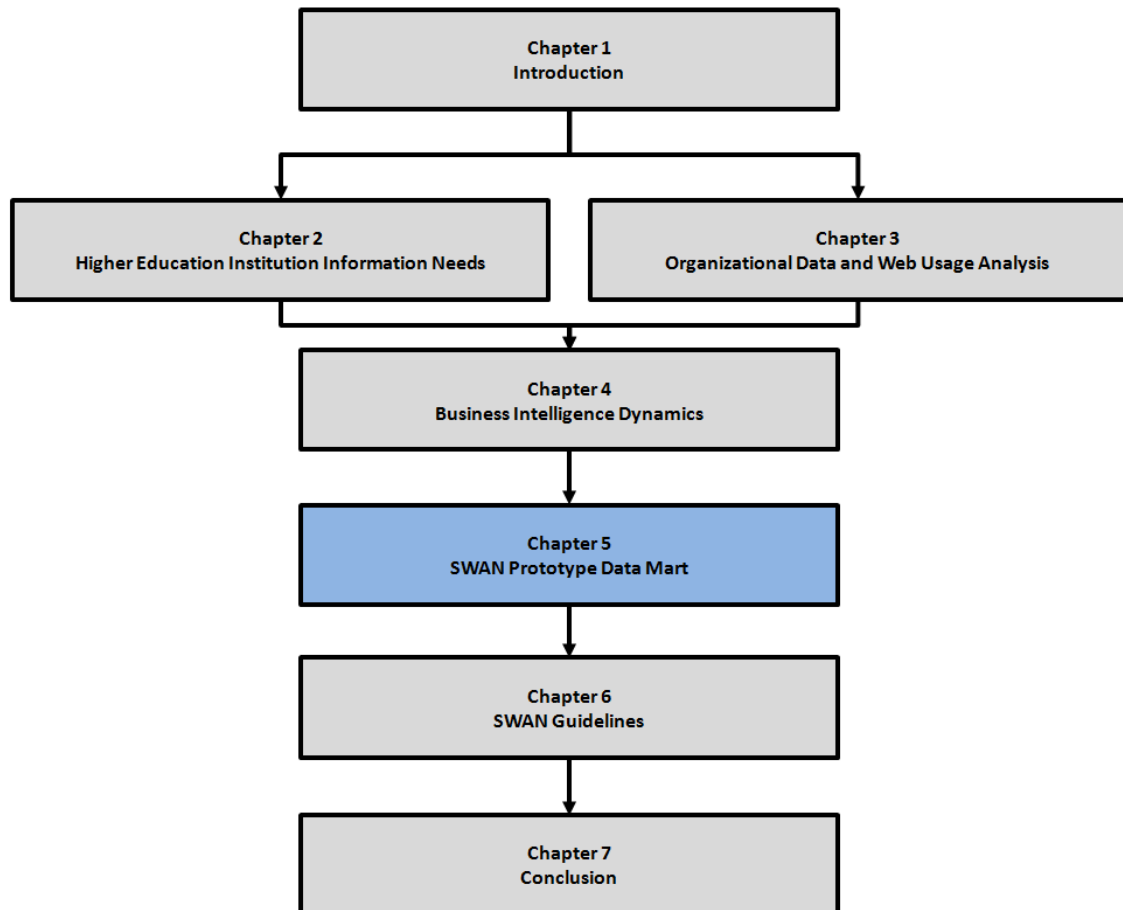
4.4 Conclusion

BI seeks to provide decision-makers with relevant information, thereby improving the quality of their decisions. There are many ways in which BI can be facilitated. Based on the comparison of a number of methods, it has been argued in this chapter that an iterative overall development methodology should be used for the development of a data warehouse, which is the central nexus of BI systems. Moreover, a bottom-up method was shown to be more desirable for such a data warehouse project. More specifically, the Kimball Lifecycle method, as documented by (R Kimball

et al., 2008) is particularly appropriate for the requirements and scope of this research study, given that it is an iterative bottom-up method, well documented, detailed and a very appropriate method.

Chapter 5 - SWAN Prototype Data Mart

“True genius resides in the capacity for evaluation of uncertain, hazardous, and conflicting information”. – Winston Churchill



This chapter presents the development of the SWAN data mart prototype that forms part of the proposed solution to the research problem. This development effort provides valuable input for the IT artefact design.

5.1 Introduction

The undertakings of this chapter contribute to the completion of steps 3 and 4 of the DSRP. The artefact designed in this research relies on knowledge gained during the development of the prototype data mart presented below. Therefore, a specific context is defined for use of the artefact, given that the development was limited to a specific scope and environment.

For the purpose of this research study, a single, self-serving, small scale prototype data mart was proposed. Prototyping is a process used in a number of disciplines to discover new possibilities and to solve complex problems (Kelley, 2001). A prototype can be referred to as a primitive form of an idea or design, and can be used for modelling, testing and evaluation (Wang, 2002). According to Kute and Thorat (2014), a prototype is a basic working version of something, is not a complete system, provides only overall functionality without any detailed functions and is based on current requirements. Prototyping is useful when developing systems that do not yet exist.

Developing a full DW/BI system for an organisation is a massive undertaking and exceeds the scope of this research. However, DW/BI systems provide the analytical capacity needed to demonstrate a solution to the research problem. Prototyping a small scale DW/BI system in the form of a prototype data mart can demonstrate a solution while remaining within the scope of the research. Moreover, it can demonstrate the feasibility of a possible extension of the project.

The prototype data mart is referred to as the *SWAN* (Student **W**eb usage **A**nalysis) data mart and the development effort is referred to as the *SWAN* project. The Kimball Lifecycle (KL) method was adopted to carry out the *SWAN* project, as discussed in Section 4.3. The Data Warehouse Lifecycle Toolkit (R Kimball et al., 2008) is heavily referenced in this chapter.

As noted above, the KL method is made up of 13 milestones that represent the stages and sequencing of a DW/BI system's high level tasks. Each milestone constitutes a number of major and minor tasks, as well as considerations and recommendations (Kimball et al., 2008). The completion of each milestone results in one iteration of the overall DW/BI system. One iteration results in one data mart serving a single business process. In this study, a single iteration of the Kimball Lifecycle was required for the development of the *SWAN* data mart.

The KL method integrates the scope and limitations of a project into the milestones. Therefore, tasks that were essential to the requirements of the *SWAN* project were undertaken, and unnecessary and irrelevant tasks were excluded. In other words, certain tasks are required for large DW/BI systems, but were not applicable to the *SWAN* data mart.

The iteration began with the Program/Project Planning and Program/Project Management milestones.

5.2 Program/Project Planning and Program/Project Management

Program refers to the overall DW/BI system, which is the complete collection of data marts as per the bottom-up approach. *Project* refers to a single iteration of the KL used to develop a single data mart (Kimball et al., 2008). Hence, the development effort for the SWAN data mart is referred to as the SWAN project.

In the following subsections, the project planning and project management and the associated milestone tasks, as applied to the SWAN project, are discussed. Figure 5.1 illustrates the Program/Project Planning and Program/Project Management milestones placement in the KL.

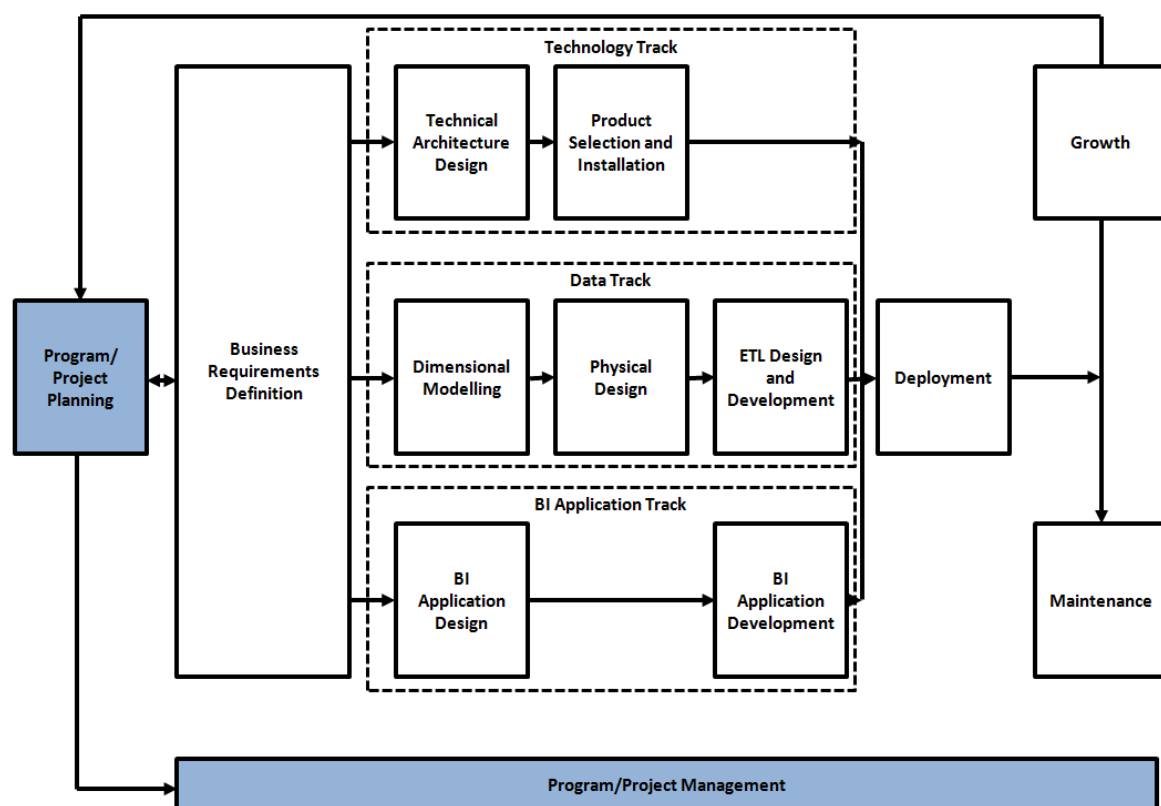


Figure 5.1 - Program/Project Planning and Program/Project Management milestones adapted from Kimball et al. (2008)

5.2.1 Program/Project Planning

The initiation of the SWAN project took this milestone into consideration by assessing the readiness and defining the scope of the project. This is discussed in the following paragraphs. Assessing readiness is achieved by considering certain initial factors such as senior sponsors, business motivation and technical feasibility (Kimball et al., 2008).

Senior sponsors refers to those individuals within the organisation – in this case, the NMMU – who bought into the DW/BI initiative and who have influence and demand respect. They recognised the need and value of the information the DW/BI could provide to the organisation. An ideal sponsor will provide useful input and demands within reason and be supportive of the DW/BI project. Having sufficient and suitable sponsors is an important factor when determining readiness for a DW/BI project.

Senior Sponsors or stakeholders for the SWAN project included various NMMU sponsors:

- Director of Information and Communication Technology at the NMMU. This sponsor indicated that value could be gained through the proposed analysis of the Web usage data.
- System engineer, NMMU. This sponsor corrected various issues that were awaiting attention as an indirect result of this research study. The correct logging of network traffic was deemed a concern by top management. The objectives of the SWAN project were in keeping with the objectives of the system engineer who thus supported the project.

These sponsors indicated either that they could gain some value from the SWAN project or certain tasks associated with it, or that they were inherent stakeholders in the project.

Business motivation has no real bearing on this research study, given the context; however, the research motivation was clear. The SWAN project was initiated to develop guidelines for the analysis of student Web usage data in support of primary educational objectives. Further motivation stemmed from investigative research coupled with certain needs of IT lecturers and system engineer management, discussed in Section 6.3.1. The research motivation was considered adequate for the SWAN project to proceed.

Technical feasibility of the DW/BI project readiness refers primarily to the required data. It is important to establish whether the data needed for the objectives or motivations of the BI/DW project is usable; in other words, the data must be accessible, interpretable and orderly. Without an appropriate source of data, the feasibility of the DW/BI project is threatened by issues that cannot be easily corrected (Kimball et al., 2008; Simon, 2009). Prior to the commencement of this study, it was established that the data required to drive the SWAN project was accessible, interpretable and orderly, as discussed in Section 3.4.

The level of importance of each of the above readiness factors is not equal as they all carry different weight when determining readiness (Kimball et al., 2008). However, given that none of the readiness factors had any apparent shortcomings, considering the scope of the SWAN project, and given the fact that all these readiness factors were in place and the project could proceed, the weight of each factor did not have to be considered.

Defining Scope

The scope of the SWAN project had to be documented. This ensured that the understanding of the relevant sponsors and developers involved would be consistent. The document that defines this scope is often referred to as a *scope charter*. This documents the project's focus, objectives, approach, anticipated data and target users, the parties involved and the stakeholders, the criteria for success, the assumptions and risks (Kimball et al., 2008). The scope charter for the SWAN project is discussed in detail in Section 6.2.

5.2.2 Program/Project Management

The Program/Project Management milestone comprises an ongoing process that runs throughout the KL, as shown in Figure 5.1. In this study it involved the coordination of tasks to ensure that the project adhered to the schedule and fulfilled the aims of the project. This was achieved through monitoring the status of the project, tracking problems and mitigating change to keep within the bounds of the scope (Kimball et al., 2008). In the SWAN Program/Project management, these tasks were the responsibility of the researcher and fell within the planning of the research itself.

5.3 Business Requirements Definition

Reaching the Business Requirements Definition milestone entails forming an understanding of the information requirements of the users of the DW/BI system. The business users and their requirements are the primary factors influencing the planning, design and implementation of a DW/BI system (Kimball et al., 2008). Thus the relationship between the Program/Project Planning and the Business Requirements Definition milestones is two-way, as shown in Figure 5.2.

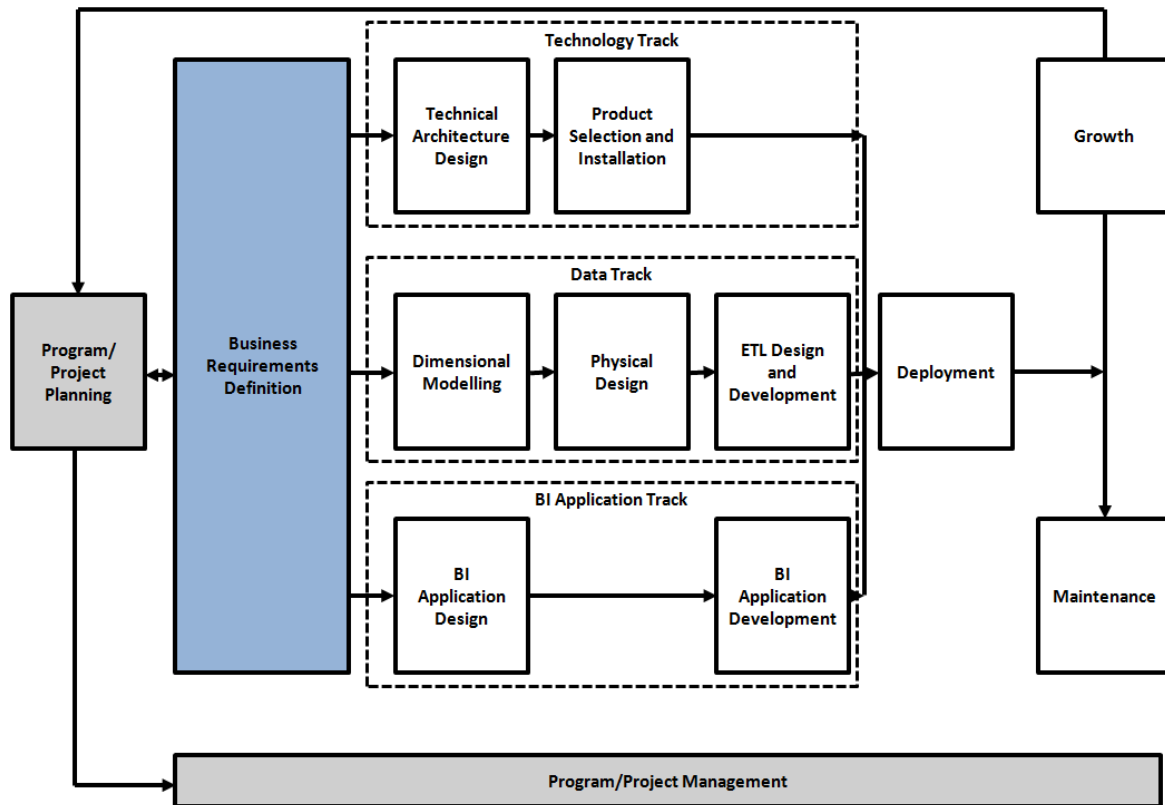


Figure 5.2 - Business Requirements Definition milestone adapted from Kimball et al. (2008)

In order to serve users' information needs, their requirements must be collected and understood. Face-to-face approaches provide a rich and valuable form of feedback (Inmon, 2002; Scheps, 2013; Simon, 2009). Interviews with business users are pivotal to understanding their requirements precisely (Kimball et al., 2008).

An interview with an NMMU IT lecturer was conducted to determine which student Web usage information would be most valuable to the lecturer when making decisions concerned with the achievement of her primary educational objectives. The questions posed in the interview, as well as the answers provided, are discussed in Section 6.3.2. The following was determined from the interview:

The primary educational objective of the IT lecturer was discussed and defined as:

To develop the students' ability to retrieve subject-relevant information from the Web.

The following information would be valuable to the IT lecturer:

The subject specific websites that students visit during practical classes in her subject.

These information needs helped to define the business requirements for the SWAN project. These requirements became the rationale for the design of the technical, data and BI application components of the SWAN data mart, ensuring that information of value would be provided, as these components were based on the user's direct requirements.

5.4 Kimball Lifecycle Tracks

Once the business requirements have been established, the KL splits into the three tracks, namely Technology, Data and BI Application, as shown in Figure 5.2 (Kimball et al., 2008). The Technology Track consists of milestones concerning the infrastructure and technological resource considerations of the project. The Data Track is made up of milestones dealing with the logistics of making the relevant data usable and extracting it from source location and transferring it to the data mart. The BI Application Track consists of milestones concerning the front end of the data mart, in other words presenting information to the business users.

The milestones of each track and their associated tasks as they applied to the SWAN project are discussed in the following subsections.

5.4.1 Technology Track

The milestones of the Technology Track, highlighted in Figure 5.3, are the Technical Architecture Design and Product Selection and Installation milestones.

It is implied that specific tasks on parallel tracks in milestones are carried out simultaneously where necessary. Working from left to right on the Technology Track as per Figure 5.3, the track begins with Technical Architecture Design.

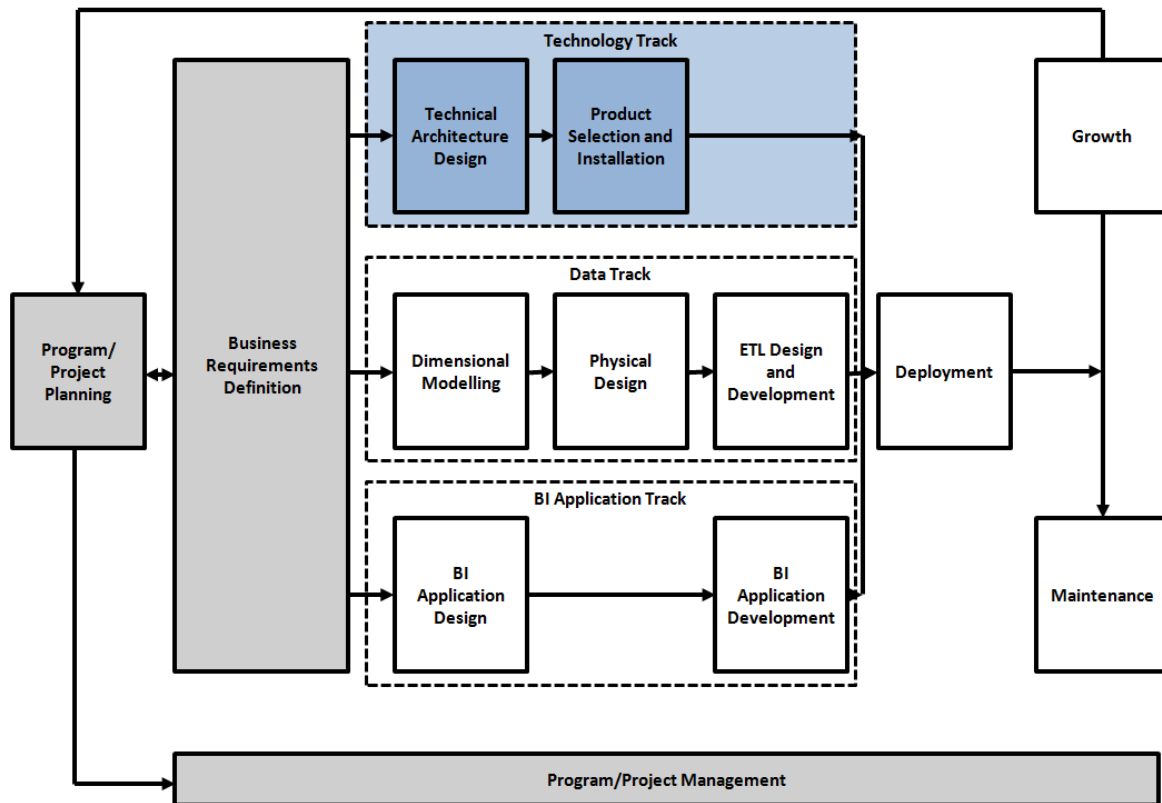


Figure 5.3 - Technology Track Milestones adapted from Kimball et al. (2008)

- **Technical Architecture Design**

This milestone outlines the technological requirements for the DW/BI system. Essentially, it involves considering the processing requirements of the DW/BI system in order to establish what IT hardware and software will be needed to facilitate the required processes. Subsequently, products can be selected and installed accordingly. “Where the business requirements answer the question ‘What do we need to do?’ the architecture answers the questions ‘How will we do it?’” (Kimball et al., 2008).

In order to understand the technical and business requirements of the project, the current technical environment and planned strategic technical directions are considered. A high level technical architecture model provides a holistic visual depiction of these considerations (Kimball et al., 2008).

Figure 5.4 reflects the high level technical architecture model for SWAN's technical needs by considering the associated processing and functional requirements that were necessary to move the data from its source to the presentation of relevant reports. The internal data was extracted from NMMU data, and included aspects such as subject, lecturer and timetable data as well as the Fortigate firewall traffic logs generated by Web browsing within the NMMU network.

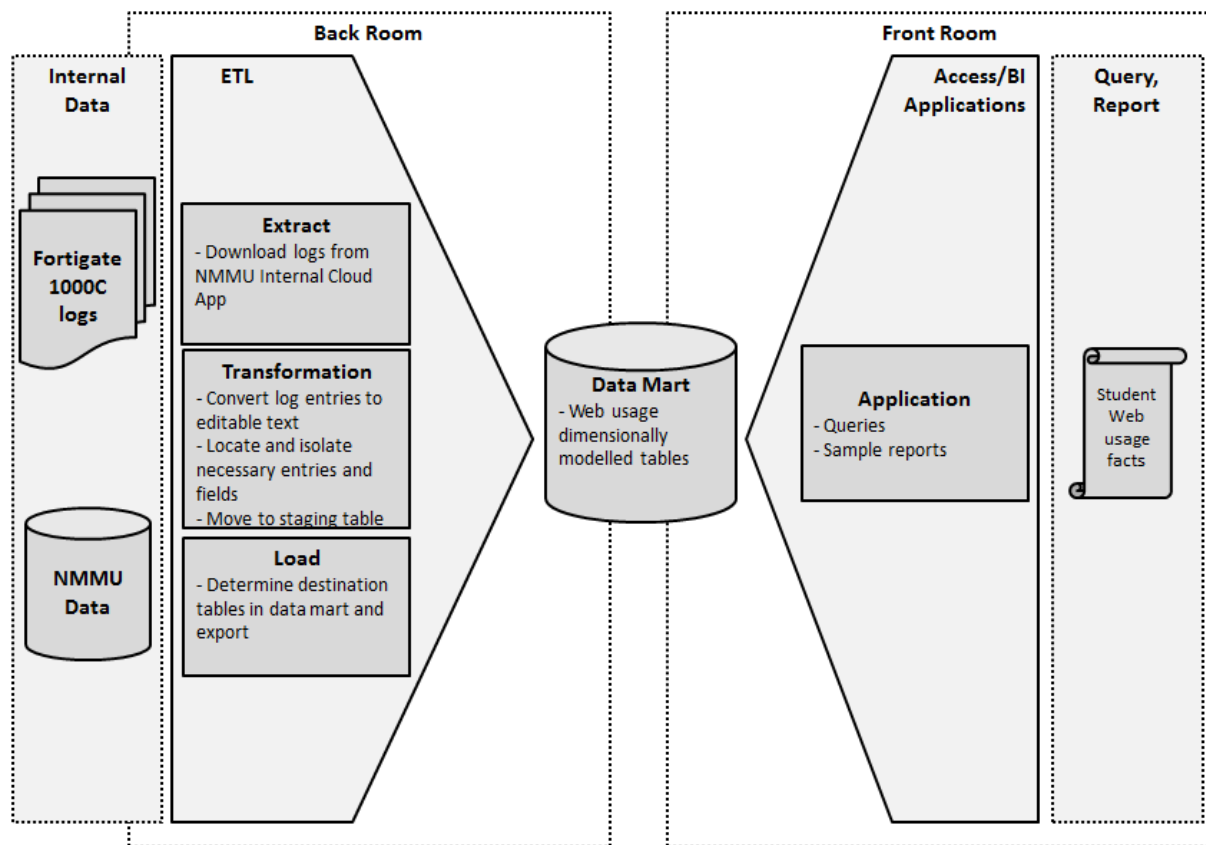


Figure 5.4 – High level technical architecture model for SWAN adapted from Kimball et al. (2008) and Watson & Wixom (2007)

A small set of subject and lecturer data from NMMU was needed for SWAN and this did not require any transformation; it simply needed to be extracted and loaded into the SWAN data mart according to the dimensional model established in the Dimensional Modelling Milestone, which will be discussed in Section 5.4.2. The Web usage data, however, comprised a large, inconsistent data set and required some data profiling in order to identify the ETL processing needed to make it usable in SWAN. Data profiling is the investigation of the content, structure and consistency of a data set (Kimball et al., 2008).

The values of each entry are described by a field. Sample log files were profiled by exporting the entries to a spreadsheet and sorting them by the *logid* field to determine consistency. Fortigate log file entries will have fields structured according to their *logid* (Fortinet, 2013). All entries with the same *logid* will have the same fields. The entries were found to have varying fields and inconsistent structure. Based on this basic data profiling, the following requirements were derived:

- The extract sub task of the ETL process requires that the logs be retrieved from the system engineer. Internal cloud based data exchange facilities, already in place, allowed for this exchange to occur easily.
- The transformation sub task involves converting the log entries into an editable format. Once this has been done, necessary entries and values can be isolated and the values can be concatenated or split according to the dimensional model specification.
- The load sub task is required to insert the clean log entries into the respective tables within the SWAN data mart. In order for this to occur, the SWAN data mart needed to contain tables and relevant fields to reflect the dimensional model, as discussed in Section 5.4.2. Query and report functionality was required to specify the student Web usage information to be extracted from the SWAN data mart.

With reference to Figure 5.4, the processing needs were split into two overall sets of tasks, the “back room” and the “front room”. If one considers the way a restaurant operates, the kitchen is regarded as the “back room”. The patrons do not see the raw ingredients being loaded into the store rooms from the wholesalers’ trucks, packed on shelves and in refrigerators, used as ingredients to prepare the meals, nor do they see the chef prepare the dishes. ETL processes are considered the “back room” of a DW/BI system as the users are not exposed to this processing. Extending this analogy, the restaurant seating area can be considered the “front room”. The patrons consume the meals prepared in the kitchen with no concern for the “back room” processes. In the same way, the users of a DW/BI system are only exposed to the information when it is presented to them via the reports, queries and other information retrieval techniques (Kimball et al., 2008; Kimball & Ross, 2002).

- **Product Selection and Installation**

As illustrated in Figure 5.4, the second milestone in the Technology Track is Product Selection and Installation. Once the technical needs have been established, products can be selected and installed to facilitate these needs.

Various software products are available for use by registered students at NMMU and are licensed as such. ICT services situated on campus provide this software via booking request systems. The following products were selected to facilitate the processing requirements established in the Technical Architecture Design milestone for each processing component in Figure 5.4.

- Extract – NMMU cloud application via Windows PC connected to NMMU network. The hardware and software that handles the logging of network traffic was already in place.

- Transformation –Microsoft Excel 2010. Excel spreadsheets were used to initially analyse, arrange and edit data. Excel spreadsheets are a recognised data source for data loading functionality in an SQL server (Larson, 2006; Rainardi, 2008).
- Load – Microsoft SQL server 2012. SQL server provides mass data loading functionality and is a widely used software which supports common information retrieval techniques such as queries, reports, data mining etc. (Rainardi, 2008; Simon, 2009).
- SWAN Data Mart – Microsoft SQL server 2012. Tables were created through SQL queries within a local database.
- Access/BI applications – Microsoft SQL server 2012. SQL queries were run on the present SWAN reports.

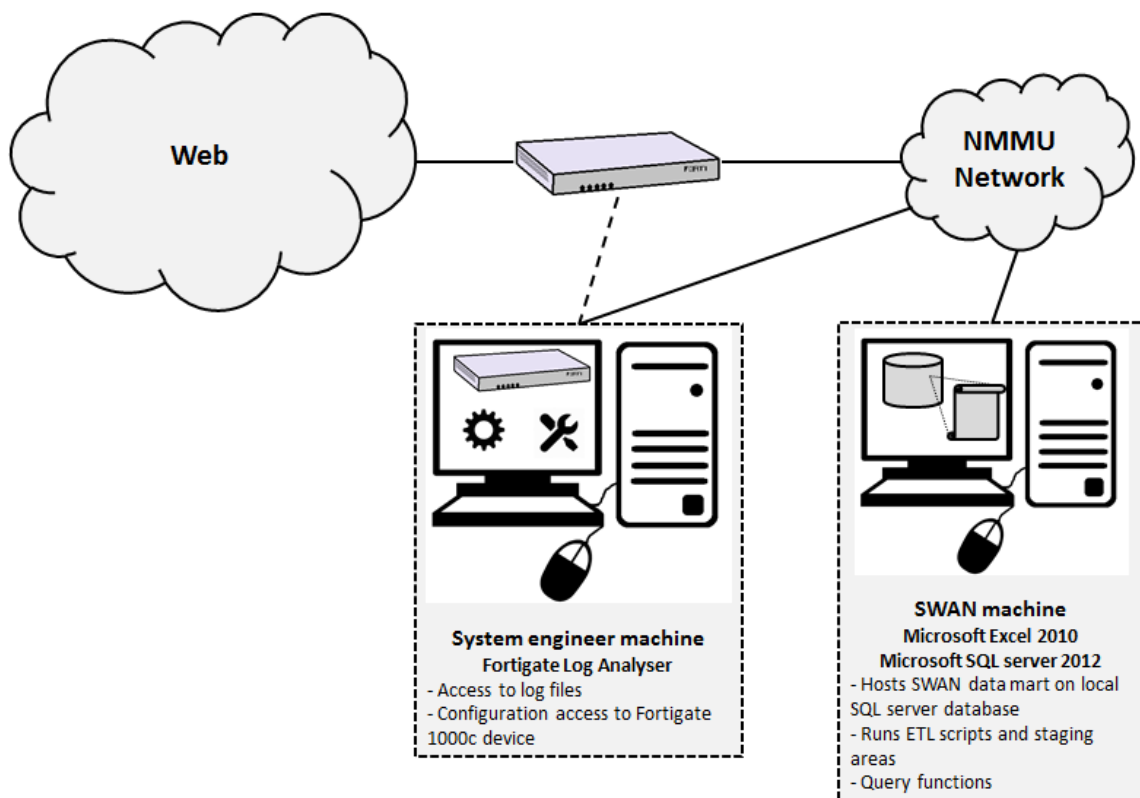


Figure 5.5 - IT Infrastructure for SWAN

The products were subsequently installed on the SWAN machine. Figure 5.5 shows the IT infrastructure that resulted from this installation. The SWAN machine was the PC used by the researcher to store and run the SWAN data mart. The system engineer machine was the PC used by the system engineer to configure and access logs from the Fortigate device and to upload logs to the cloud application for retrieval by the SWAN machine.

5.4.2 Data Track

The Data Track's milestones highlighted in Figure 5.6 are the Dimensional Modelling, Physical Design and ETL Design and Development milestones. The track begins with Dimensional Modelling.

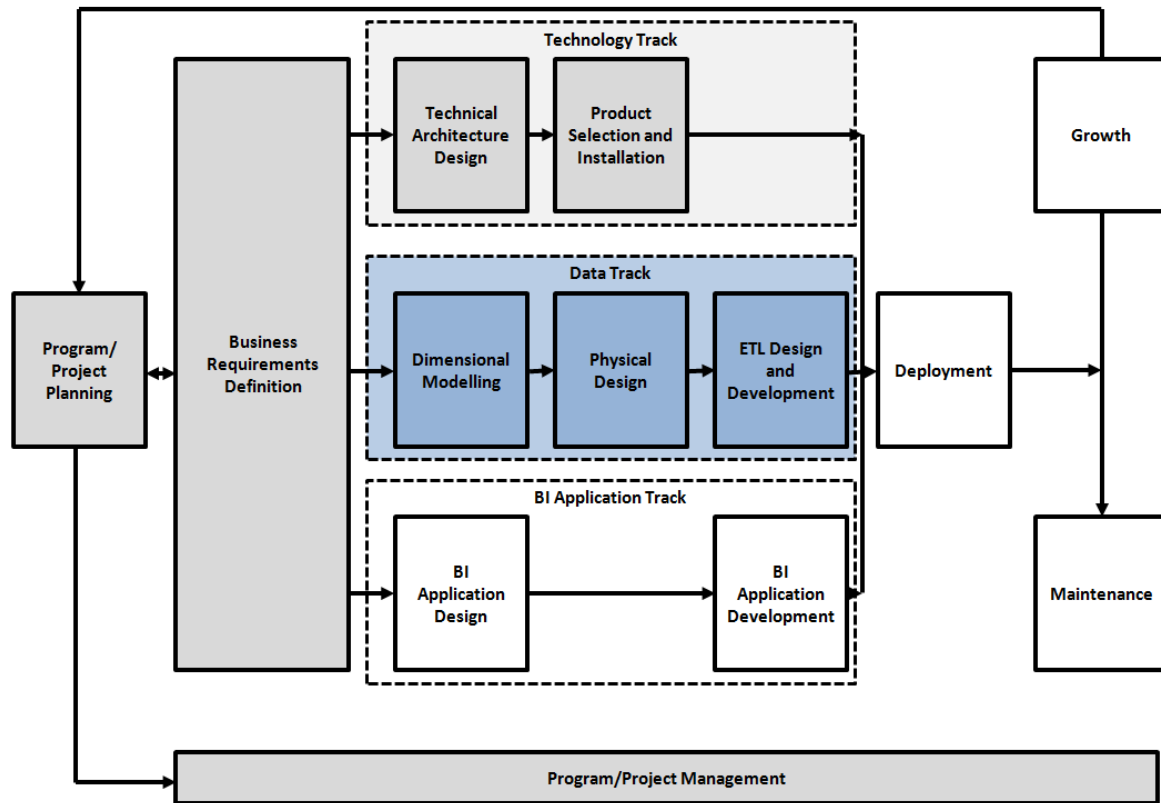


Figure 5.6 - Data Track milestones adapted from Kimball et al. (2008)

- **Dimensional Modelling**

Dimensional modelling structures data to align it with the information needs of the user, thereby making querying simpler and more efficient. This is achieved by dividing the data into *measurements* and *context*. Measurements result from organisational processes and systems and are usually numeric. Measurements are referred to as *fact*. Facts are enveloped by largely textual context that is true at the point at which the fact is documented. This context is intuitively divided into independent logical clumps called dimensions. Dimensions provide the "who, what, when, where, why, and how" context to the fact. Essentially, they give meaning to the data (Kimball et al., 2008; Simon, 2009).

Business processes in an organisation can be represented by a dimensional model that consists of a fact table containing the numeric measurements surrounded by a halo of dimension tables containing the textual context. A fact table is constructed, based on the business process being modelled. The

grain will be determined by the level of detail that is required. The design of dimensional data models is completed following a four-step process (Kimball et al., 2008), namely:

- Step 1 Choose the Business Process
- Step 2 Declare the Grain
- Step 3 Identify the Dimensions
- Step 4 Identify the Facts

This process is driven by the business requirements established in section 5.3. In this study, a draft dimensional model called a star schema was created using data webhouse star schema examples discussed by Kimball and Merz (2000) as well as input from one of the experts mentioned below. Constructing a suitable star schema requires some expertise in the data warehouse domain.

Consulting experts for research purposes can be done in various ways. An exclusive interview, a one-on-one interview with a specific individual who can provide information on a very specific event or context could be conducted (Hochschild, 2009). A communication platform between the experts could be set up to form a Delphi study (Okoli & Pawlowski, 2004). A focus group could be formed by meeting with the experts and allowing them to discuss a proposed issue (Morgan, 1997). In this study, three candidates were considered experts, based on their demographics as shown in Section 6.3.3. Their work commitments meant that these experts were unavailable for a focus group or a Delphi study. Moreover, given similar constraints, individual interviews were not possible and collating the interview results to allow for consensus would have proved difficult. For these reasons experts were sent a document containing the draft star schema and were asked to review this and indicate its suitability to this study, based on the business process and granularity for which it was intended. They were asked to suggest any changes they thought necessary. No strict research methodology or official process was found to fit the criteria and for the purposes of this study the process is referred to as an expert review.

The star schema model was refined and verified using the feedback from the expert review; details are discussed in Section 6.3.3. The resulting model is shown in Figure 5.7:

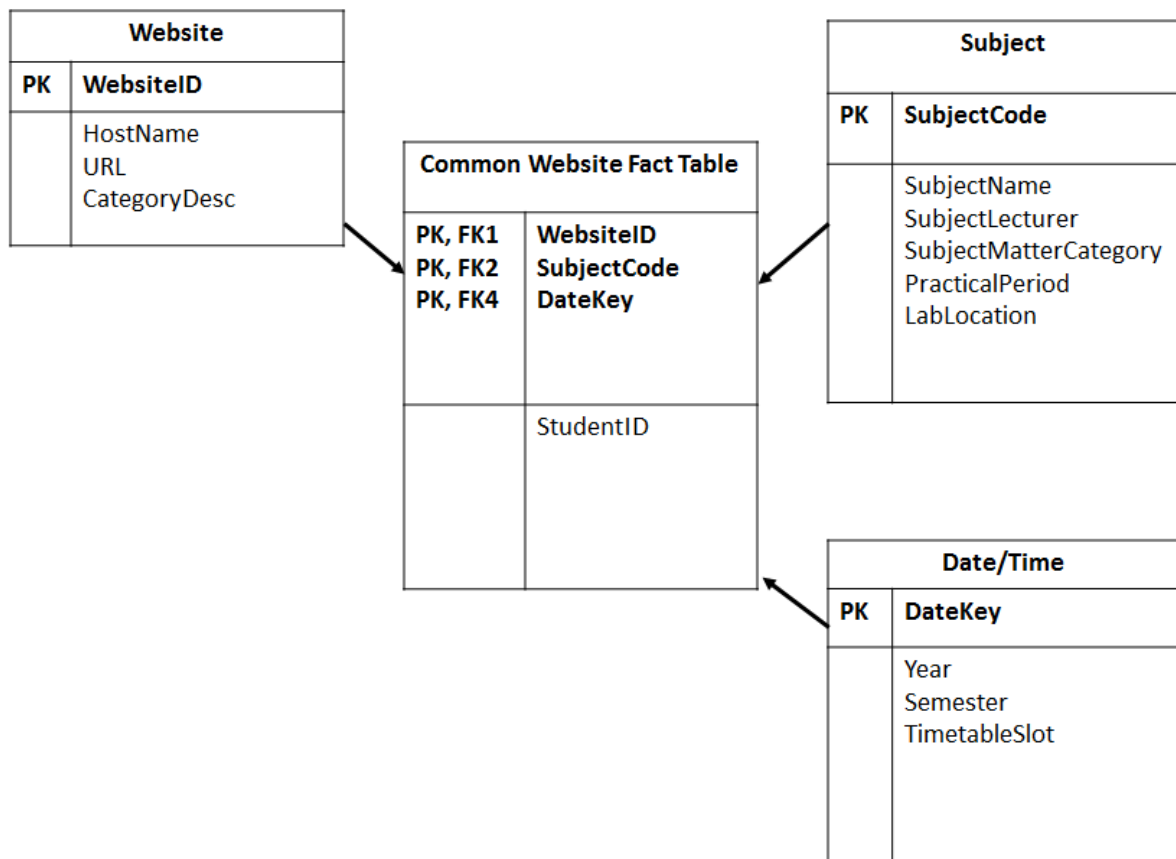


Figure 5.7 - SWAN Data Mart star schema

- **Physical Design**

The second milestone in the Data Track is Physical Design (see Figure 5.6). The physical design is the implementation of the dimensional model. This is achieved by creating tables in the SQL server to mimic the star schema in Figure 5.7. The tables in Figure 5.8 were created in SQL server in the SWAN database.

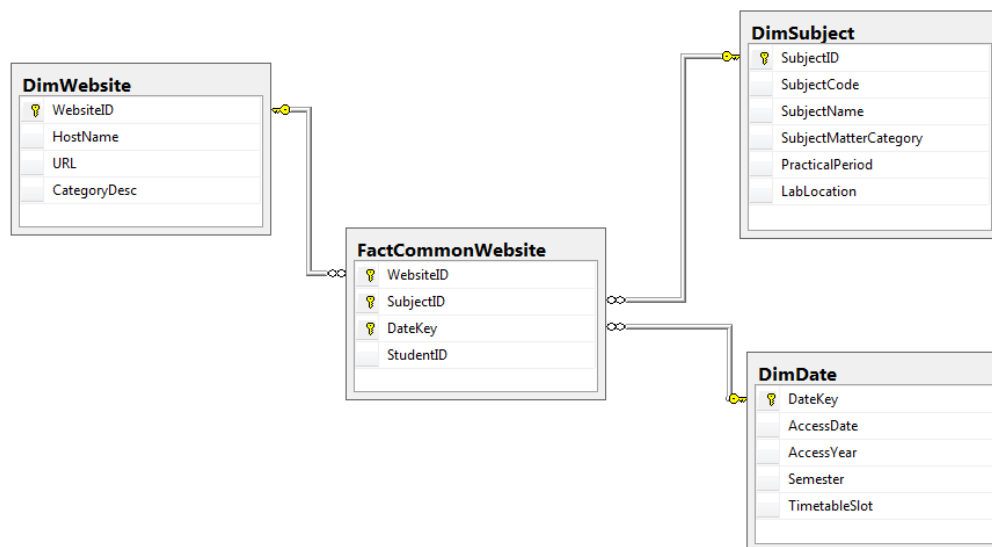


Figure 5.8 - SWAN Data Mart dimensional star schema diagram generated from SQL Server

The Star Schema design established in the Dimensional Modelling milestone was the framework for the SWAN data mart. With the star schema in place, the SWAN data mart required data to populate the tables.

- **ETL Design and Development**

As can be seen in Figure 5.6, the third milestone in the Data Track is ETL Design and Development. This milestone involves the extraction of the data from multiple sources and their transfer to the data mart in the required format, based on the dimensional model. It involves all the processes that make up a functional backroom of the DW/BI system. In a large scale data warehouse these processes take an immense amount of resources. Moreover, these processes will need to be automated to provide new data to the warehouse as it is generated by business processes (Golfarelli & Rizzi, 2009; Kimball et al., 2008). However, for the scope and requirements of the SWAN project, these processes were done manually and with a single data set (see Section 6.3.4).

Once the data has been loaded, the data mart can provide a report to satisfy the user requirements. This presentation of reports falls into the front room processing component of the data mart. The BI Application Track involves designing and developing front-end/presentation functionality which makes up the front room component.

5.4.3 BI Application Track

The BI Application Design and the BI Application Development milestones of the BI Application Track are highlighted in Figure 5.9. The track begins with BI Application Design.

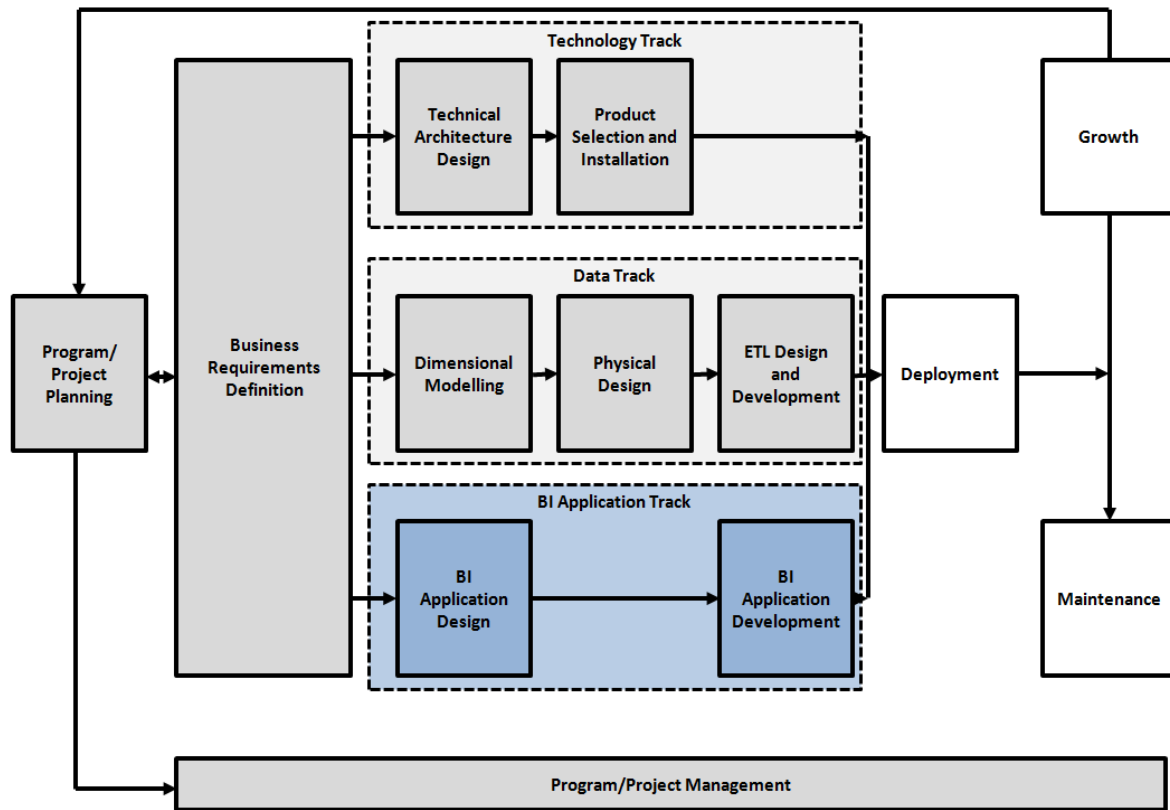


Figure 5.9 - BI Application Track milestones adapted from Kimball et al. (2008)

- **BI Application Design and BI Application Development**

BI Applications are used to present information to the business user. They can range from basic parameter-driven reports to complex analytical systems (Kimball et al., 2008). In the case of a large DW/BI system, careful consideration and effort will go into the design and development of the BI Application. In the case of the SWAN project, a simple report, given certain parameters, was needed to demonstrate the SWAN data mart's potential. Reports can be presented locally on the SQL Server; no BI Application design or development is required and the BI Application Track is mostly redundant.

A report was produced to extract information from the SWAN data mart, based on the IT lecturer's needs elicited by the interview conducted in the Business Requirements Definition milestone. This was achieved by querying the fact table. The query was constructed according to criteria based on the

IT lecturer's subject information and desired report. Figure 5.10 shows the query constructed to produce the result set shown as Figure 5.11. This result set was then copied to documentation software to be presented as a report. The fact table referenced all three dimension tables, allowing information to be viewed in different ways, given different criteria. The columns displayed in the result set can specify which data to include in the result set by applying filter parameters. Thus if the data mart is expanded to cover more subjects and practical classes, they can be specified in the WHERE clause. For example, this would allow reports to be generated for a given timeframe for a specific subject or based on a certain category such as "social media" or "video streaming". For the SWAN data mart prototype, the data set was purposefully limited. However, the star schema would be able to filter and present more subject data if it was further populated with more Web usage log data.

```

SELECT AccessYear,
AccessDate,
Semester,
SubjectCode,
HostName,
CategoryDesc,
LabLocation,
PracticalPeriod

FROM dbo.FactCommonWebsite
LEFT JOIN dbo.DimWebsite
ON dbo.FactCommonWebsite.WebsiteID = dbo.DimWebsite.WebsiteID
LEFT JOIN dbo.DimDate
ON dbo.FactCommonWebsite.DateKey = dbo.DimDate.DateKey
LEFT JOIN dbo.DimSubject
ON dbo.FactCommonWebsite.SubjectID = dbo.DimSubject.SubjectID

/*The filtering is specified i.e. drill across or drill up or drill down*/
WHERE Semester = '2' AND LabLocation = 'R128' AND SubjectCode = 'ONT3660' AND CategoryDesc = 'Information Technology' AND PracticalPeriod = '4'

```

Figure 5.10 - SQL Query used to produce a result set from the fact table

The result set was exported to Excel for viewing purposes. A follow-up interview on the Business Requirements Definition was conducted with the IT lecturer to determine whether the report contained the information that she had indicated would be valuable. The results of the interview confirmed that this type of information would indeed add value to the IT lecturer's decision-making. For more details on this, see the Information Requirements Validation Interview Transcript in Section 6.3.2.

	AccessYear	AccessDate	Semester	SubjectCode	HostName	CategoryDesc	LabLocation	PracticalPeriod
4	2014	2014-08-15	2	ONT3660	www9.addfreestats.com	Information Technology	R128	4
5	2014	2014-08-15	2	ONT3660	www3.addfreestats.com	Information Technology	R128	4
6	2014	2014-08-15	2	ONT3660	www.w3schools.com	Information Technology	R128	4
7	2014	2014-08-15	2	ONT3660	www.visual-paradigm.com	Information Technology	R128	4
8	2014	2014-08-15	2	ONT3660	www.updateyourbrowser.net	Information Technology	R128	4
9	2014	2014-08-15	2	ONT3660	www.trialpay.com	Information Technology	R128	4
10	2014	2014-08-15	2	ONT3660	www.techhive.com	Information Technology	R128	4
11	2014	2014-08-15	2	ONT3660	www.teamviewer.com	Information Technology	R128	4
12	2014	2014-08-15	2	ONT3660	www.subnettingquestions.com	Information Technology	R128	4
13	2014	2014-08-15	2	ONT3660	www.statcounter.com	Information Technology	R128	4
14	2014	2014-08-15	2	ONT3660	www.sourcecodeonline.com	Information Technology	R128	4
15	2014	2014-08-15	2	ONT3660	www.semsim.com	Information Technology	R128	4

Figure 5.11 - SWAN result set

- **Deployment, Maintenance and Growth**

The three tracks mentioned above all converge in the Deployment milestone, as indicated in Figure 5.12. In a large DW/BI system the Deployment milestone involves thorough planning to guarantee that the subcomponents created in previous milestones operate and interact and function correctly. Moreover, appropriate education and support are required (Kimball et al., 2008). However, given the short lifespan and non-deployable nature of SWAN, the Deployment milestone was not taken into account.

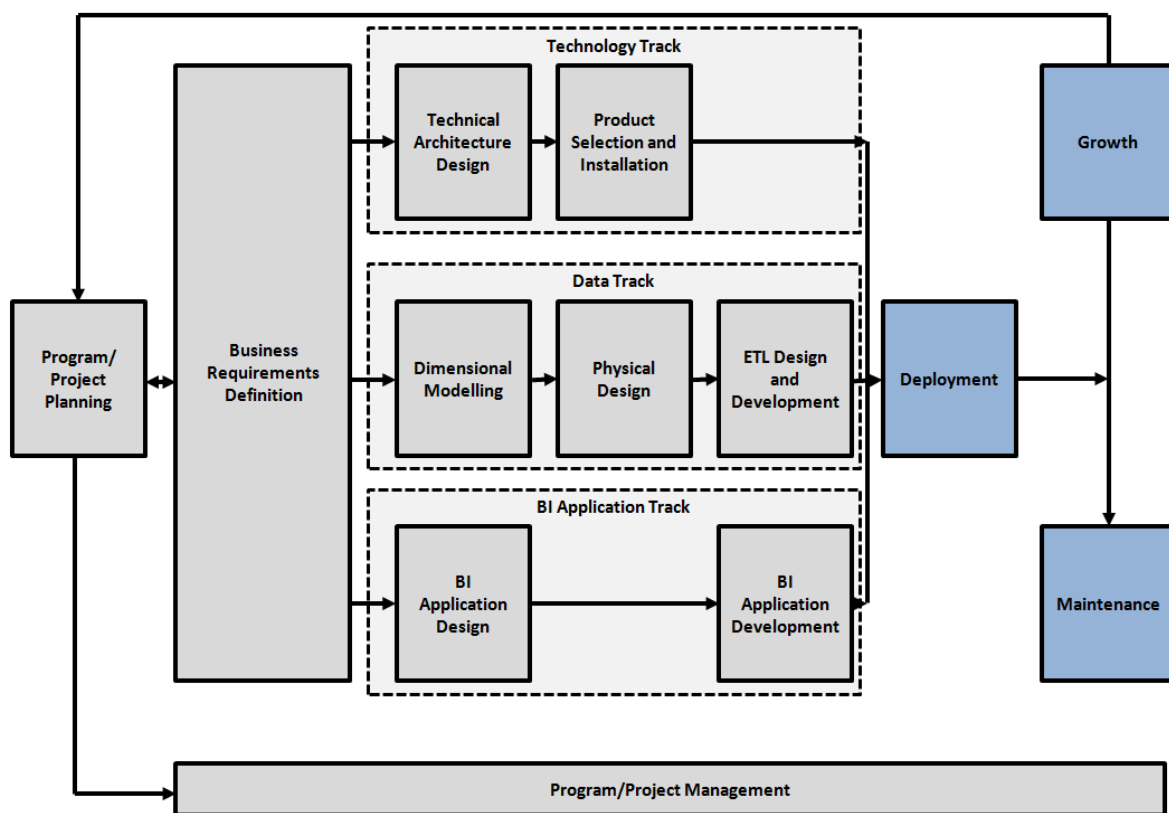


Figure 5.12 - Deployment, Growth and Maintenance adapted from Kimball et al. (2008)

Following Deployment the DW/BI system would require maintenance in the form of various technical operations such as usage monitoring, performance tuning, user support and back-up of the system (Kimball et al., 2008). Again, as the lifespan of SWAN was intended to be short and as it was non-deployable in nature, the Maintenance milestone was ignored.

The Growth milestone comprises the long term expansion of the DW/BI system; however, as was the case for the other two milestones discussed in this section, the Growth milestone was not taken into account in this project.

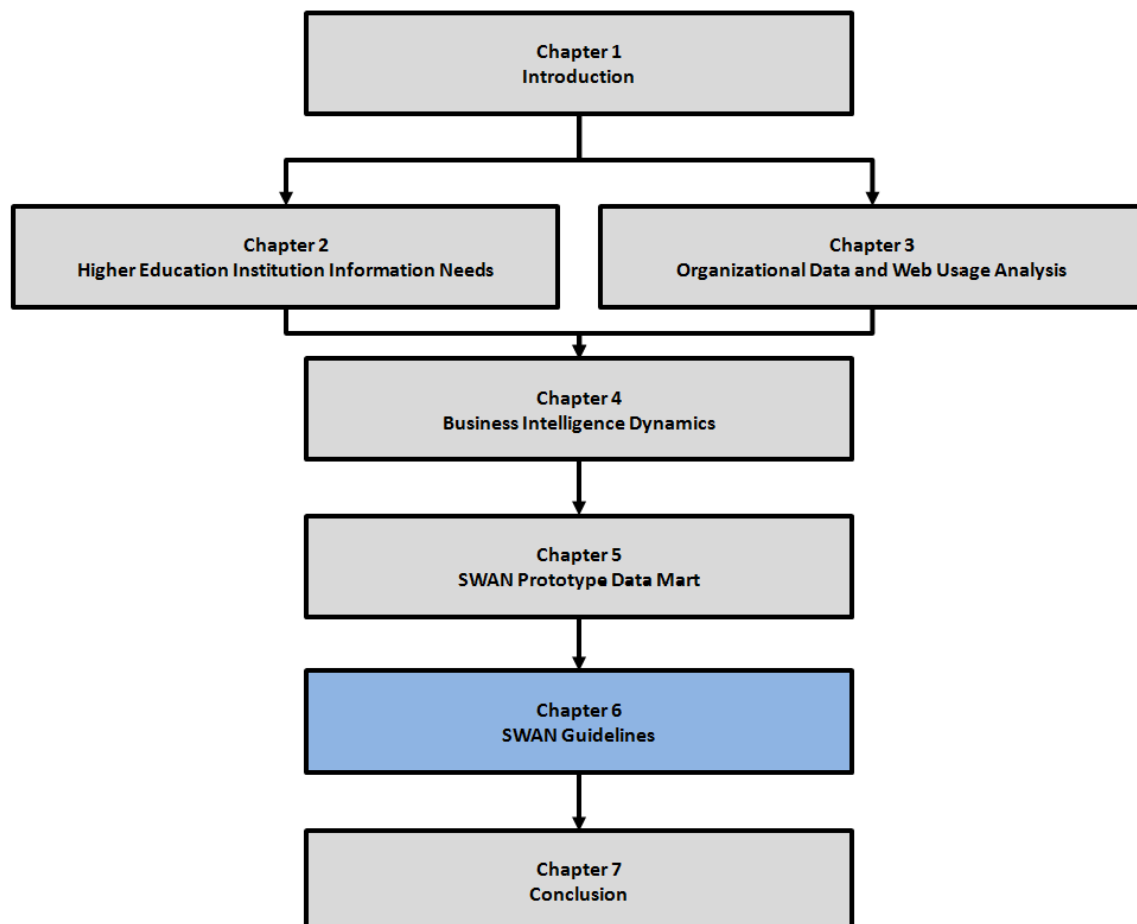
5.5 Conclusion

The SWAN project resulted in a single self-serving data mart (SWAN). This chapter has explained how the SWAN data mart was created using the KL bottom-up method and how it produced reports based on the business needs of a specific user (IT lecturer).

The criteria for success, documented in the SWAN project charter, were met, namely, a single question or a small selection of questions posed by IT lecturers regarding student Web usage behaviours were to be answered. The report produced by the SWAN data mart using NMMU log file data was confirmed as useful and valuable by an IT lecturer. This confirms the belief that the data can be analysed to provide information valuable to IT lecturers in assisting them in making decisions about their primary educational objectives. Furthermore, the SWAN data mart was considered successful based on the expert review of the development and implementation of a sound star schema, as discussed in Chapter 6. The SWAN data mart was thus shown to be a viable platform for addressing other questions regarding students' Web usage. The methods used to develop the SWAN data mart were also successful.

Chapter 6 - SWAN Guidelines

“A few observations and much reasoning lead to error; many observations and a little reasoning to truth” – Alexis Carrel



This chapter presents a set of guidelines that resulted from the consolidation of knowledge gained during and through the analysis of the SWAN project. These guidelines form the primary research output and IT artefact of this research study.

6.1 Introduction

The problem identified in this research study was that there was no way for IT lecturers in the School of ICT at NMMU to analyse Web usage data in a meaningful way. Given that the SWAN data mart was created to demonstrate how this could be achieved, it was hypothesised that knowledge gained from the SWAN project would make a valuable contribution to solving this research problem.

The undertakings of this chapter contribute to the completion of steps 3, 4 and 5 of the DSRP. The artefact is designed, presented and then evaluated through expert review. Furthermore, the evaluation demonstrates the artefacts use in the given context.

The question asked during this investigation was how was the SWAN data mart developed? In addition, how were certain steps in the SWAN project undertaken and why these steps were completed in that manner? By answering these questions some valuable assertions could be made. The boundaries and context of the SWAN project are very clear and specific. This project was conducted within a single department of the NMMU with organisationally unique data. A deeper understanding of the SWAN project, which is clearly defined, was sought in order to answer certain *how* and *why* questions.

Creswell's approach for presenting a case study is used. Whereby, the context of the phenomena is described and issues are isolated and discussed (2012). However, a case study strategy is not used. The SWAN project was examined to extract knowledge for the design of the IT artefact as part of the DSRP.

The author of this dissertation was the primary researcher, as well as the developer of the SWAN data mart. His accounts of the various issues resolved during the SWAN project were an important source of data. These reflections formed the accounts of the researcher during the development of the SWAN data mart. Issues that emerged during the SWAN project are described in the sections that follow. Essentially, the lessons learned and insights obtained during the SWAN project are presented. The SWAN project development process is studied by the same person who undertook it in order to solidify the lessons learned using an established and understood method to give it validity. It is in this way that it is considered a reflective process. Furthermore, it is noted that the SWAN project was initiated and completed as part this research and is not separate.

The lessons learned and assertions made, based on the case and the issues arising from it are consolidated into a set of guidelines. These guidelines are intended to be holistic, high-level and provide broad guidance through advice and suggestions. The lessons learned and assertions made are derived from the case through interpretative and evidential argumentation according to Mason (2002). Essentially, accounts of the developer of the SWAN data were considered alongside data resulting

from the development process and refined into concise and useful information in the form of a set of guidelines.

6.2 The SWAN Project Case and Context

The NMMU has six campuses. These are the Bird Street, South, North, Missionvale, Second Avenue and George campuses. All of these have extensive facilities to provide for a flourishing academic environment. The School of ICT resides in the North campus where it offers various IT qualifications, such as National Diplomas and Bachelor's Degrees in Support Services, Software Development and Computer Networks. These qualifications require students to attend lectures and to participate in practical classes in a number of computer laboratories situated around the campus. Postgraduate Master's and Doctoral Information Technology research qualifications are also offered by the School of ICT. The postgraduate research section contains offices in which these students can conduct their research. All the computers in the laboratories and in the postgraduate section are connected to the NMMU network which provides Web access. This allows students to use the Web as a valuable source of information and material to support their academic information needs and to provide them with technical support. A prototype data mart was developed in the postgraduate section to demonstrate a method of analysing the Web usage behaviour of students. This was undertaken as a component of this overall research project. The data mart that was developed was called the SWAN (Student Web usage ANalysis) data mart.

The development of the SWAN data mart and the results this produced was of interest. This case was studied in order to construct guidelines for the process of facilitating an analysis of student Web usage data, in a way that is useful and valuable, in the context of the NMMU and possibly other higher education institutions with a similar network infrastructure. Data was collected from the case in the form of documentation generated during the case, from transcripts of interviews with participants in the case and from the developer's reflective analysis of the case and its associated processes. Guidelines were proposed, based on conclusions generated and assertions made from the data through evidential and interpretative argumentation.

The SWAN project resulted from an initial inquiry to determine whether Web use was a positive or negative catalyst for students' academic well-being and success at the NMMU. This inquiry resulted in a study to determine whether online personalisation of search engine results was materialising in a computer laboratory in the School of ICT, and if so, how soon and to what extent the effects would be seen. The study concluded that personalisation was not a matter for serious concern in the information gathering ability of students at the campus (see appendix D) (Von Schoultz & Van Niekerk, 2012).

Following this study, further investigation into the effects of Web use in higher educational organisations resulted in the decision to consider analysing the network traffic of the organisation. In this way it was possible that students' Web usage behaviour could be determined and measured. However, the researcher was unclear as to how to go about finding and analysing the network traffic data. After a discussion with an associate professor it emerged that the system engineers responsible for the network infrastructure of the NMMU were based at the North campus and could be consulted. Following an exclusive interview with two of these system engineers it was determined that the network activity on all NMMU campuses, including Web based network activity, ran through a device for which they were responsible and therefore had access to. Furthermore, they provided a sample of log files extracted from the device that contained the network traffic of all six campuses, for further details see (Von Schoultz et al., 2013) attached as Appendix G. However, how the logs could be analysed to determine their usefulness, and the value of the information that could potentially be extracted from the log files, was not known.

The sample log files were profiled for usability. It was concluded that the data was an appropriate source of Web usage information from NMMU students. However, in its present state and format the data was not suitable for any process of analysis. The reason for this unsuitable format was identified as an issue that was awaiting attention from the system engineers at the time (Von Schoultz et al., 2013). It was agreed that this issue would be prioritised and that once the system engineers had resolved it, further data would be provided for profiling. This issue was later resolved and data samples were provided to the researchers. The new data samples were profiled and it was agreed that the new format and detail of the data was adequate to provide information on Web activity within the NMMU, as discussed in Section 3.4. Following an initial literature review, no heuristics were found regarding the analysis of Web usage data in the context of higher education. However, the domain in which Web usage data was analysed for information was extensive. Many formalised methods and uses for the analysis of Web data exist in literature, as discussed in Section 4.2. The information retrieved from the data would be valuable, especially to members of NMMU who could benefit from knowing more about their students' Web usage behaviour.

The IT lecturers were considered primary candidates for this information because they could better serve their educational objectives by making more informed decisions with this information. A survey was conducted to serve as an initial investigation into the value IT lecturers could see in having Web usage information about their students. The survey results indicated that many of the IT lecturers recognised the need for and perceived the value of such information, as demonstrated in Section 2.3. A possible method of providing for the analysis of Web usage data was found in the subject area of BI. Investigations of the literature in the BI subject area led the researchers to devising a plan to develop a prototype data mart. A data mart is a small-scale data warehouse that is a central component

of a BI system (Inmon, 2002; R Kimball et al., 2008; Simon, 2009). BI systems present operational and historical data in a format that makes it valuable to specific users. These users can gain insights based on factual, organisation-specific data that allows for more informed decision-making which in turn improves business processes. Developing a BI system requires certain important and complex tasks. There is no clear consensus within the data warehouse community about which method to follow when developing BI systems. In the present study this was resolved through method comparisons, taking into consideration the scope and limitations of the study. A bottom-up data warehouse development method was followed in developing the SWAN data mart, as explained in Section 5.2; specifically, the method used in the Kimball Life Cycle was applied. The initial plan and proposed criteria for the SWAN project were documented in the SWAN Project Scope Charter, attached as Appendix D. The Scope Charter mentions the stakeholder and developers involved in the project, the criteria for success, limitations to the scope, project focus and its objectives. Therefore, it contains direct accounts of the intentions of the case.

The SWAN data mart was developed and met the success criteria established in the planning stages as laid out in the SWAN Project Scope Charter. Meeting the success criteria required the resolving of some pressing issues which emerged during the project, all of which have been documented. The SWAN project, SWAN data mart, individuals involved in the SWAN project and associated documents will be used as data.

6.3 Case Issue Isolation and Assertions

As mentioned above, accessing relevant usable data that would provide information about the Web usage behaviour of students at North campus proved difficult. Initially, the data was unusable; after the data logging configuration issue was corrected by the system engineer so that the network traffic was logged in a more detailed format, the format of the data still required extensive analysis and processing in order to make it useful. Moreover, determining who the users of the information would be and whether they would consider the information valuable was an issue that required attention.

There was also the question of how the data would be analysed. Simply having the data in a readable format was not very useful. A method for gaining relevant information from the data was required. Following literature review, the DW/BI system development method known as the Kimball Lifecycle (KL) method was deemed to be the most appropriate. However, central to this method was the data structuring technique known as dimensional modelling. Dimensional modelling requires a design process that produces a star schema, as each organisation will have different data requirements and there is no one-size-fits-all dimensional model. The dimensional model requires an ad hoc design and must be developed from the ground up. Developing an ad hoc star schema is not a simple process.

The KL method requires an understanding of the business process that the system intends to serve by providing information from relevant data. Moreover, the development effort needs to have significant support from members of the organisation who should act as senior sponsors. In other words, the need for the information the system intends to provide must be recognised by respected members of the organisation as well as by the intended users of the system.

The most demanding aspect of the Kimball Lifecycle method in terms of time and resources is accessing the data, editing it to make it readable and usable and then exporting it to the required area of the DW/BI system. The complexity of this process proved to be dependent on the format of the data in this case. Essentially, there are some considerations when undertaking this process but there is no detailed generic method for doing this, owing to the myriad possible formats of data produced by different organisations.

In the following subsections particular issues that emerged in this case are discussed. Data on how these issues were resolved were collected and assertions drawn from each issue are presented as proposed guidelines.

6.3.1 Data format and availability

The issues associated with finding relevant and appropriate data for analysis of Web usage behaviour were explored to determine how they were resolved and why this was done in this way.

A study was undertaken prior to the case. This is attached as Appendix G (Von Schoultz et al., 2013). In this study, the NMMU system engineers were interviewed to ascertain the network infrastructure specifics. By analysing this infrastructure, the flow of Web traffic could be understood. In this way the device/s that logged or could log the network traffic could be isolated and investigated. It was found that the network infrastructure was set up as a proxy model. The system engineers confirmed that, in this model, all Web traffic from the six campuses passed through a Fortigate firewall device (as illustrated in Figure 6.1), configured to serve a similar role to that of a proxy server, that monitored incoming and outgoing network traffic. In other words, the device handles all Web requests and Web activity going out of the NMMU network and coming into the network from the Web. As all Web activity passes through the Fortigate device, this would be an ideal device to capture NMMU Web usage data.

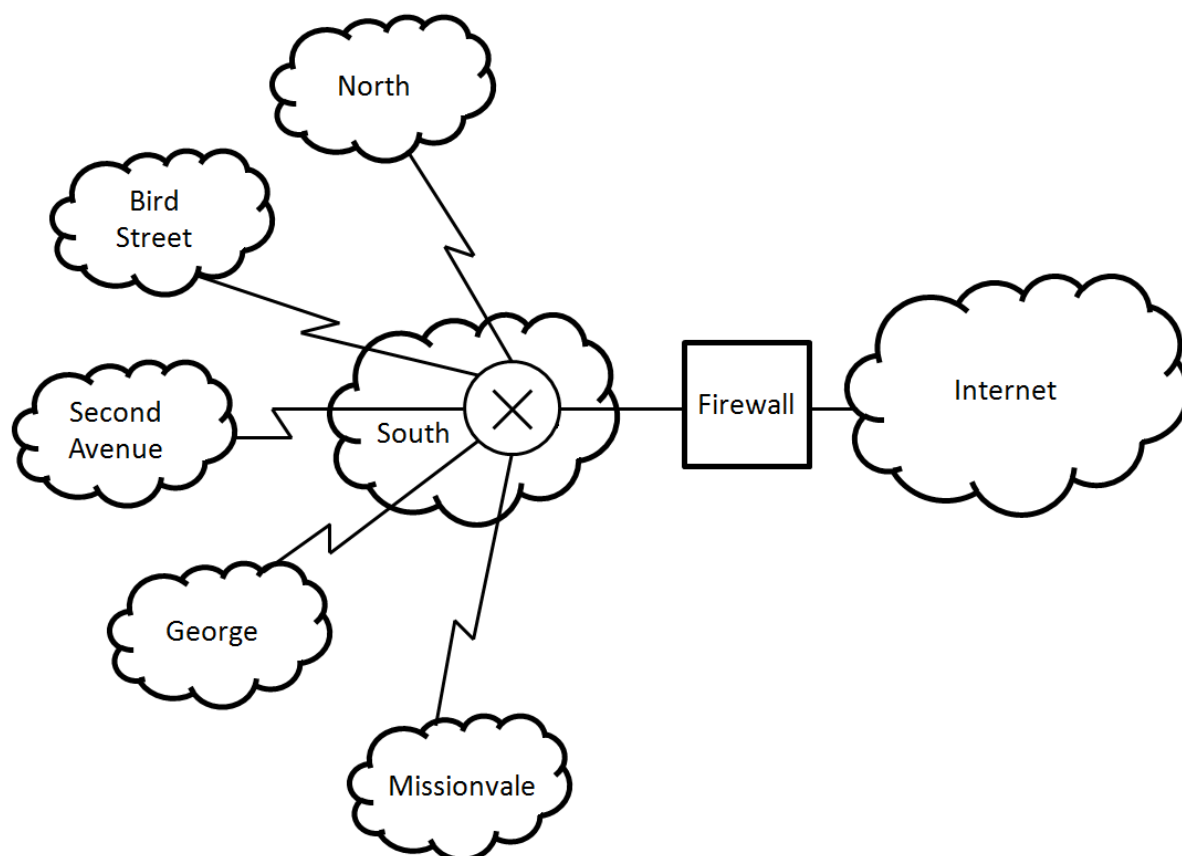


Figure 6.1 - NMMU's logical network (Von Schoultz et al., 2013)

One of the system engineers forwarded a sample log file from the Fortigate firewall device to the researcher for profiling. This was a large flat file (39.2GB) containing Web traffic from 10:00 on 5 April to 04:00 on 6 April. The file contained around 87 308 522 entries, each entry the account of a single Web access activity. In other words, one line in the sample log file showed one request to the Web, which was prompted by a user visiting a website by entering the hostname on the browser or clicking on links within a website, or other automated Web activity. Figure 6.2 shows the original sample log file, opened using a large flat file viewing application.

[illegible]

Figure 6.2 - Original sample log file

Each entry contained certain fields of data values that provided the details of the Web activity entry. Table 6.1 shows the fields in an entry and an example of values taken from a single random entry.

Table 6.1 - Firewall log entry example (Von Schoultz et al., 2013)

Field	Value
Month	Apr
Day	12
Timestamp	11:52:15
unspecified	tyrael
date	2013-04-12
Time	11:52:15
devname	imperius
device_id	FGT1KC3912800514
log_id	0038000004
type	traffic
subtype	Other
pri	notice
vd	root
src	10.102.129.162
src_port	53149

src_int	"LAN_AGGR"
dst	173.236.49.82
dst_port	80
dst_int	"INTERNET"
SN	637641392
status	start
policyid	128
dst_country	"United States"
src_country	"Reserved"
tran_sip	192.96.15.20
tran_sport	61201
service	HTTP
proto	6
duration	0
sent	0
rcvd	0

In order to derive Web usage behaviour, the actual URL and hostname information was required. This would show what websites had been visited. The only value in the sample log entries that could provide this information was the destination IP address (dst value), as highlighted in table 6.1. However, in order to find the hostname of that entry, for example, www.google.com, the destination IP (dst value) would have to undergo an NSLOOKUP to link the IP address to the hostname. An NSLOOKUP is a network command used to link an IP address to a website hostname. In other words, the destination IP address (dst value) has no real face value for analysis. Moreover, to force NSLOOKUP on each entry significantly increases the processing load of the analysis. The study concluded that, given that the destination IP address (dst) field was the only source of data to determine the actual websites visited, the sample log file in its current format was not viable for Web usage behaviour analysis. However, the firewall device was the appropriate device to log Web traffic. In addition, the reason the sample log file was in its present format was identified by the system engineers as a configuration error on the firewall device that was awaiting correction. The study made a recommendation that where a proxy modelled network setup was in place, and where data was needed for Web usage analysis, the availability and format of the data should be investigated as a starting point. The system engineers agreed to persevere in correcting the configuration error and to provide additional sample log files once this had been done. The correction of the configuration error would likely result in logs which are richer in Web related data.

The configuration error was resolved shortly after the study was conducted. One of the system engineers uploaded a new sample log file in the updated post configuration error format, to allow the researcher to profile it, via an internal cloud-based file exchange application. This updated log file was retrieved and profiled. A comparison was made with the original sample log file (pre-configuration error log file). The new file was found to contain more fields describing the entries as well as fields that had rich website values, such as URL, hostname and category descriptions. The analysis of the post configuration sample log files revealed that the new log file format was appropriate for use in Web usage behaviour analysis techniques, as discussed in Section 3.4.

Once it had been established that suitable data was available, an appropriate method by which valuable information could be elicited from this data was investigated. It was established that the domain of BI was the area where the solution might lie and that the development of a DW/BI system would facilitate information gathering. Various DW/BI system development methods were compared and the Kimball Lifecycle method (KL) was selected (see Section 4.3). The steps in the KL are referred to as milestones. The first milestone involves tasks for planning the project and emphasises the importance of assessing the readiness of the project (suitability of resources to move forward), based on certain considerations. A crucial factor before moving forward with the KL is determining the appropriateness of the available data through data profiling and analysis. This was found to be true prior even to the planning of the SWAN project, as the original sample log file was unusable as mentioned in the study discussed above. Any efforts to use the data in a meaningful way seemed futile. Moreover, during the planning of the SWAN project it became clear that if the data was not available or did not contain useful fields, as in the case of the original sample log file, the project would not be feasible. This was consistent with recommendations from the KL milestones used in the SWAN project.

Another important factor in assessing readiness was the availability of senior sponsors who were prepared to make a commitment to the project. These were senior members of NMMU who recognised the value in the SWAN project. These sponsors are stakeholders and parties involved in the SWAN Project Scope Charter (see Appendix D) who indicated their commitment to the SWAN project by expressing how they perceived its value. The Director of ICT was a major sponsor in that he is an authority in the area of system engineering. He indicated his buy-in following a discussion with an associate professor who was assisting in the SWAN project (see Appendix A for email correspondence). The Fortigate firewall configuration error, mentioned in the study discussed above, has been a pending priority and this was recognised by the Director of ICT. An interview with the system engineer who corrected the error, attached as Appendix C, confirmed that the researchers probing the data had flagged this configuration error and stressed its importance. It was established that correcting the configuration error was in the interests of the SWAN project, the Director of ICT

and the researcher. As a result, the configuration error was corrected, and the researcher was able to collect usable data. Had the Director of ICT not stressed the priority of the error, the senior sponsor, the system engineer, may not have considered this a priority as such a correction was not part of his usual workload. The support of this senior sponsor was instrumental in the success of the SWAN project. This supports the notion expressed in the KL that senior sponsors are a major factor in the feasibility of a DW/BI project.

Various approaches exist for developing a DW/BI system, as discussed in Section 4.2. The authors of many of these approaches have recognised the data, and the users of these data, as the truly important factor when developing successful DW/BI systems (M Golfarelli & Rizzi, 2009; Imhoff et al., 2003; Inmon, 2002; R Kimball et al., 2008; Moss & Atre, 2003; Scheps, 2013; Simon, 2009). Despite the common objectives of these approaches there is no standard method for DW/BI system development (Saroop & Kumar, 2011; Sen & Sinha, 2005; Wixom & Watson, 2001). A possible contributory factor to this situation is that successful DW/BI systems can be developed using one of many approaches (Sen & Sinha, 2005). Each instance of a DW/BI system development project will differ depending on the method used and the organisation in which it is implemented. An organisation's data and resources are tied to that specific organisation despite the overlap and similarity in types of data, for example, human resource data or accounting data. This is another contributory factor to the absence of standard approaches. For this reason, the steps in each DW/BI project will differ based on the organisation and the approach adopted.

The previous investigation of possible data sources in the SWAN project proved beneficial to the project and facilitated the identification of data owners. This allowed a working relationship to emerge. In this way, the senior sponsors were identified and value was gained. The initial study discussed above recommended early data source investigation, which was consistent with the KL approach used to undertake the SWAN project, and this was validated in a publication. Based on the present case, the assertions made about the analysis of the Web usage behaviour of students in the NMMU were consolidated in the following guideline:

Guideline 1

The data and the owner/s thereof should be investigated to gather and profile a sample of the required data before the DW/BI project should be considered feasible.

Recommendations

- Investigate which constituents administer the network infrastructure of the institution
- Establish which constituent/s of the institution is/are accountable for the control of the data

- Enquire about the network infrastructure to determine how Web access is provided to the institution
- Identify an appropriate device which logs, or could log Web usage data
- Acquire a sample of the data from the device
- Profile the sample to determine its current format

Considerations if a proxy model is in place

- The proxy server or device configured to fill the role of a proxy server should be investigated as a source of Web usage data
- Investigate the logging capacity of the proxy server or relevant device and consult the administrators to determine if it is plausible to configure the device to log Web usage data in an appropriate format. Logging Web usage data may have negative performance repercussions on the network

6.3.2 Understanding the Business Process and Gathering Users' Information Needs

The issues surrounding the delivery of valuable information to relevant business users were explored in order to address the questions of how these issues were resolved and why they were resolved in such a way.

The users of the SWAN data mart provided the input for the design of the data mart. The KL method mandates that focus on the business user is maintained throughout development. This ensures that the information produced by the data mart is relevant, has meaning and will result in a more efficient and effective business process as a result of the improved quality of decisions resulting from this information. Accordingly, a potential user was identified and interviewed. This user was selected based on the criterion that he or she taught at least one subject which was part of the curriculum of the National Diploma in Software Development. As part of the research for the SWAN project, a survey was conducted to determine IT lecturers' primary educational objectives in order to establish their views on desired Web usage behaviours among their students (see Section 2.3). From these results, lecturers who indicated that having information about their students' Web usage behaviour during practical classes or while on campus would assist them in making better decisions about their primary educational objectives were identified. These lecturers were considered suitable candidates for face-to-face-interviews to further refine what information they would like to have available to them. This information would serve as an important input for the SWAN project and subsequently the design of the SWAN data mart. The interview transcript was viewed for further detail on how the interview was conducted and the nature of the questions posed.

Information Requirements Interview Transcript

Interviewer: Mr Dean von Schoultz

Interviewee: North Campus IT Lecturer

Purpose of the interview

The purpose of this interview is to ascertain what information about students' Web usage behaviour would be useful to an IT lecturer when making decisions about the achievement of their primary educational objectives.

For example, early identification of students who are not participating in practical classes because of excessive Web browsing that is not subject-related; confirming whether students are browsing mandatory subject-specific websites during practical classes; identifying students who may be showing early signs of internet addiction.

The purpose of these questions is to inform the appropriate structuring of Web data so that it can be queried to provide relevant information for this research study.

Question 1:

What are your primary educational objectives?

Interviewee's response:

"For my third year project students to be able to go out there and find relevant information, find source codes, snippets of code that they can actually use within their projects. So they must learn to find the right material and obviously by that increase their educational learning."

Interviewer's comment

"Presumably using the Web as a primary source during your practical classes?"

Interviewee's response

"Yes, that will be their primary source."

Question 2:

If you could choose one report about the Web usage behaviour of your students, what information would you like to be in that report?

Interviewee's response:

"In that report I would like to know which of these kinds of education sites are they actually going to in terms of being able to find snippets of code, where they are actually getting their technical support and assistance from, that would be very helpful."

Question 3:

What information in the report would make the greatest difference in enhancing your decision-making in achieving your primary educational objectives?

Interviewee's response:

"The actual academic sites that they are visiting."

Interviewer's comment:

"So you would like to know the exact sites that they are using during your classes. That would then allow you to have a stronger knowledge base for making decisions about your primary educational objective."

Interviewee's response:

"That is correct, so if they are not using useful sites I would need to do more in class to make sure that they actually know which sites they should be going to and giving them maybe a little bit more support than I currently do. I mean students must actually learn to help themselves in the end."

Interviewer's comment:

"So you would find that information very valuable?"

Interviewee's response:

"Yes."

From the transcript it is apparent that emphasis is placed on encouraging the user to describe exactly what information he or she would find useful and how it would contribute to his or her decision-making in improving the business process. Furthermore, including open-ended questions allowed this interviewee to express her opinions free of any constraints. In addition, the interview approach allowed the interviewer to discuss and clarify the responses with the interviewee to ensure that he had understood them correctly.

This interviewee response information was used to design and drive the SWAN project, to fulfil a vital requirement and to formulate criteria for success, as indicated in the Scope Charter (see Appendix D). In order to verify that the correct information could be produced by the SWAN data mart, a follow-up interview was conducted with the lecturer. The query results from the SWAN data mart that were used to meet the success criteria of the SWAN project were exported into a report in an Excel spreadsheet, as shown in Figure 6.3. This was presented to the interviewee and she was asked whether she would find this information valuable.

	A	B	C	D	E	F	G	H	I	J	K
1	Year	Date	Semester	Subject	Website	Category	Laboratory	Practical Period			
2	2014	8/15/2014	2	ONT3660	zn_3sf17hmdsjbempn-techmedianetwork.siteint	Information Technology	R128	4 to 9			
3	2014	8/15/2014	2	ONT3660	za-cdn.effectivevemeasure.net	Information Technology	R128	4 to 9			
4	2014	8/15/2014	2	ONT3660	xa.xingcloud.com	Information Technology	R128	4 to 9			
5	2014	8/15/2014	2	ONT3660	www9.addfreestats.com	Information Technology	R128	4 to 9			
6	2014	8/15/2014	2	ONT3660	www3.addfreestats.com	Information Technology	R128	4 to 9			
7	2014	8/15/2014	2	ONT3660	www.w3schools.com	Information Technology	R128	4 to 9			
8	2014	8/15/2014	2	ONT3660	www.visual-paradigm.com	Information Technology	R128	4 to 9			
9	2014	8/15/2014	2	ONT3660	www.updateyourbrowser.net	Information Technology	R128	4 to 9			
10	2014	8/15/2014	2	ONT3660	www.trialpay.com	Information Technology	R128	4 to 9			
11	2014	8/15/2014	2	ONT3660	www.techhive.com	Information Technology	R128	4 to 9			
12	2014	8/15/2014	2	ONT3660	www.teamviewer.com	Information Technology	R128	4 to 9			
13	2014	8/15/2014	2	ONT3660	www.subnettingquestions.com	Information Technology	R128	4 to 9			
14	2014	8/15/2014	2	ONT3660	www.statcounter.com	Information Technology	R128	4 to 9			
15	2014	8/15/2014	2	ONT3660	www.sourcecodeonline.com	Information Technology	R128	4 to 9			
16	2014	8/15/2014	2	ONT3660	www.semsim.com	Information Technology	R128	4 to 9			
17	2014	8/15/2014	2	ONT3660	www.public-trust.com	Information Technology	R128	4 to 9			
18	2014	8/15/2014	2	ONT3660	www.nutripur.com	Information Technology	R128	4 to 9			
19	2014	8/15/2014	2	ONT3660	www.mozilla.org	Information Technology	R128	4 to 9			
20	2014	8/15/2014	2	ONT3660	www.microsoft.com	Information Technology	R128	4 to 9			
21	2014	8/15/2014	2	ONT3660	www.mediawiki.org	Information Technology	R128	4 to 9			
22	2014	8/15/2014	2	ONT3660	www.livefyre.com	Information Technology	R128	4 to 9			
23	2014	8/15/2014	2	ONT3660	www.linuxnix.com	Information Technology	R128	4 to 9			
24	2014	8/15/2014	2	ONT3660	www.jite.org	Information Technology	R128	4 to 9			
25	2014	8/15/2014	2	ONT3660	www.java2s.com	Information Technology	R128	4 to 9			
26	2014	8/15/2014	2	ONT3660	www.informationr.net	Information Technology	R128	4 to 9			
27	2014	8/15/2014	2	ONT3660	www.ikmnet.com	Information Technology	R128	4 to 9			
28	2014	8/15/2014	2	ONT3660	www.ihm.com	Information Technology	R128	4 to 9			

Figure 6.3 - Sample report extracted from the SWAN Data Mart

The transcript was as follows:

Information Requirements Validation Interview Transcript

Interviewer: Mr Dean von Schoultz

Interviewee: North Campus IT Lecturer interviewed in SWAN project

Purpose of the interview

The purpose of this interview is to determine whether the information produced by the SWAN data mart is the type of information you requested during the SWAN project information requirements interview.

Question 1:

Is this information adequate, based on what you originally requested during the information requirements interview?

Interviewee's response: "Yes, is this information only from my registered students, though?"

Interviewer's comments: "No, individual students cannot be identified in this SWAN data mart prototype for ethical reasons. However, as discussed, this report is on students in your practical lab during your practical class period on Friday 15 August and therefore contains all the activity of students who were present."

Interviewee's response: "I see, that makes sense. I could still see what websites they have been visiting from this report. I like that it is very useful."

Interviewer's comment: "The report shows unique information technology sites visited by students in your practical class, 469 in total. If the data mart is expanded in future research you could classify or rate the websites in the data mart itself as relevant to you and compare the ratio of relevant websites to others being browsed during the practical class. There are many avenues of report detail that could be explored further."

Interviewee's response: "I like it, and to answer your original kind of question, this report would be very useful to me."

The interviewee clearly sees value in the report. This demonstrates that using the requirements from the user to develop the SWAN data mart translated into valuable reports. Using a face-to-face interviewing technique allowed the user's needs to be clearly understood. This is consistent with authors in the data warehousing field who maintain that face-to-face approaches provide a rich and valuable form of feedback (Inmon, 2002; Scheps, 2013; Simon, 2009). In addition, Kimball et al. (2008) observe that interviews with business users are pivotal to understanding their requirements precisely. Moreover, by posing questions that address the exact type of information they would like to see resulted in clearly defined requirements that translated well in the SWAN project design. Therefore, based on the case in question, the assertions made about the analysis of NMMU students' Web usage behaviour were consolidated in the following guideline:

Guideline 2

Conduct face to face interviews with the intended users of the Web usage information and pose questions which directly focus on exactly what information they would consider valuable

Recommendations

- Identify potential users of the information which could be derived from the available Web usage data. This could be done by targeting a group which has perceived association with the information and gathering their information needs through a group meeting or survey

- Potential users would be identified from the meeting or survey results if they indicate strong influence to their decisions from the proposed information
- If no potential users are identified reconsider the perceived association mentioned above
- From these potential users, interview one or more and ask questions which directly focus on determining exactly what information they would consider valuable
- Clarify any misunderstandings with the interviewee
- Document the interview through recording and creating a transcript of the interview for future reference.

6.3.3 Designing a Dimensional Model

The issues surrounding the design and structuring of the data were explored to address questions of how the issues were resolved and why they were resolved in such a way.

DW/BI systems store current and all historical data produced by business processes and do not have the same data structuring foundations as traditional data base systems. In this way, they allow access to the history of the business, and this is manipulated to calculate trends and to predict future activity by recognising patterns that will allow for better decision-making. In order for this to occur in the DW/BI system, the data requires a structure and layout that goes against traditional data base design. Central to this structure and layout is dimensional modelling, as explained in Section 5.4.2.

Dimensional models are designed keeping in mind the business process for which the information derived from the data will be used to make more informed decisions. In so doing, a model is created to retrieve information that is valuable specifically to that business process. In order to design the dimensional model, as is the case with many specialised practical designs in a certain subject domain, experience and expertise in that particular subject domain is required. The dimensional model design for the SWAN project was no exception.

One example of a dimensional model is the star schema. The name derives from the way in which the tables are arranged: a single table surrounded by multiple connected tables, representing a star. Experts who had experience in designing and implementing star schemas were sought out. The School of ICT offers subjects that teach knowledge management, including dimensional modelling and star schemas. Three staff members were identified who either lectured in a knowledge management subject or in related subjects and/or who had previous industry experience in DW/BI systems. They were contacted and they all agreed to assist. Each expert's particular skill and experience in this field is listed below:

Expert 1)

- Six years' data warehouse and business intelligences experience in industry
- Five years' lecturing experience in data warehousing
- Cognos Business Intelligence Lifecycle Gold Level Certificate

Expert 2)

- Six years' lecturing experience in data warehousing at fourth year level
- Five years' data warehouse and business intelligences experience in industry in the area of data scrubbing and preparation

Expert 3)

- Three years' lecturing experience in data warehousing at fourth year level
- Data warehousing industry consultant

As a result of schedule constraints, regular meetings with all three experts were not possible. A draft schema was developed, using literature and input from one of the experts. This was then distributed to the other two experts with a Likert scale to indicate whether they considered the draft schema to be suitable for the business process for which it was intended. The business process and granularity was established in the information requirements interview, discussed in Section 6.3.2. The experts were asked to provide critique and suggestions. The first draft was in the following format:

Student Web analysis dimensional star schema iteration 1 document

Business Process: Planning for achievement of primary educational objectives

Specific Instance: Planning lecturing approaches to enhance students' ability to retrieve useful subject-related information from the Web during allocated practical project classes

Granularity: *Unique website visited by students during a practical class*

Dimensions:

- **Website** – WebsiteID, HostName, URL, CategoryDesc
- **Web User** – WebUserID, Type, EnrolledSubjects
- **Subject** – SubjectCode, SubjectName, SubjectLecturer, SubjectMatterCategory, PracticalPeriod
- **Practical** – PracID, SubjectCode, DateKey, NoOfPresentStudents, LabLocation
- **Date/Time** – DateKey, Year, Semester, TimetableSlot

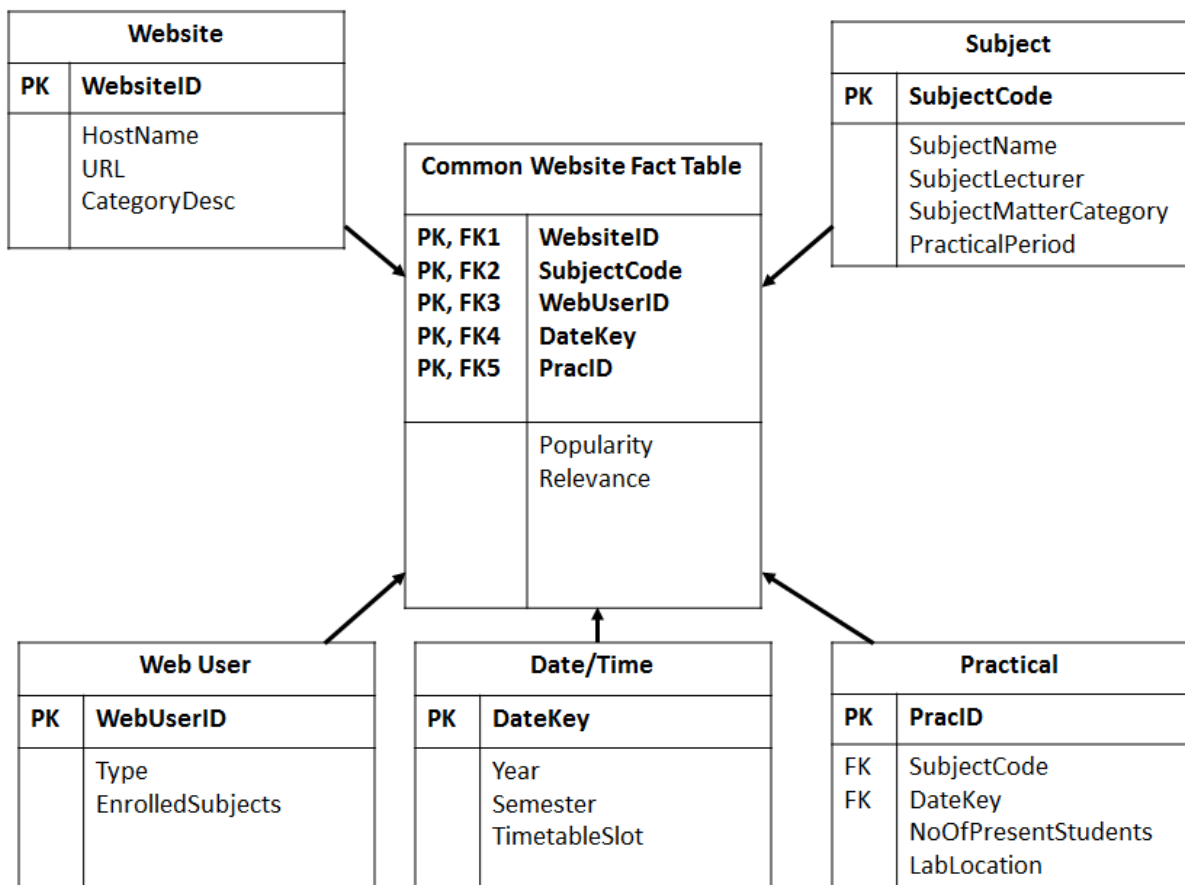


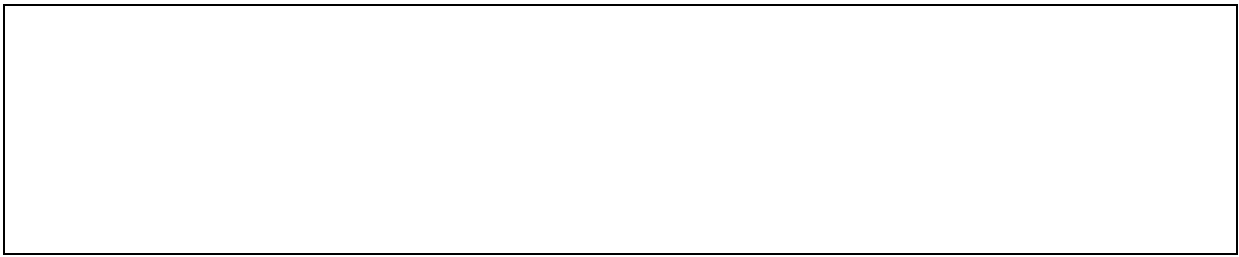
Figure 6.4 - Draft star schema

Please indicate how strongly you agree or disagree that the above dimensional star schema is suitable for the given granularity and business process, using the following scale:

- 1) **Strongly Disagree** — you believe the schema to be highly unsuitable for the given granularity and business process.
- 2) **Disagree** — you believe the schema to be unsuitable for the given granularity and business process.
- 3) **Neither Agree or Disagree** — you cannot agree or disagree with the schema's suitability.
- 4) **Agree** — you believe the schema to be suitable for the given granularity and business process.
- 5) **Strongly Agree** — you believe the schema to be highly suitable for the given granularity and business process.

Indicate level: _____

Comments and suggestions



The feedback from the expert review of this draft star schema was as follows:

Star schema expert review iteration 1 results document

Initial consensus among the expert review iteration 1 resulted in changes being made to the original star schema (Figure 6.4). Changes **A**, **B** and **C** are flagged in Figure 6.5, and are described as:

- **A** – Web User dimension changed to Student
- **B** – Common Website fact table facts “relevance” and “popularity” were removed to make it a factless fact table
- **C** – The Practical dimension was collapsed into the Subject dimension

Suggested changes are represented as numbers 1, 2 and 3 and are shown in Figure 6.5. Refer to paragraph 1.4 below for further details.

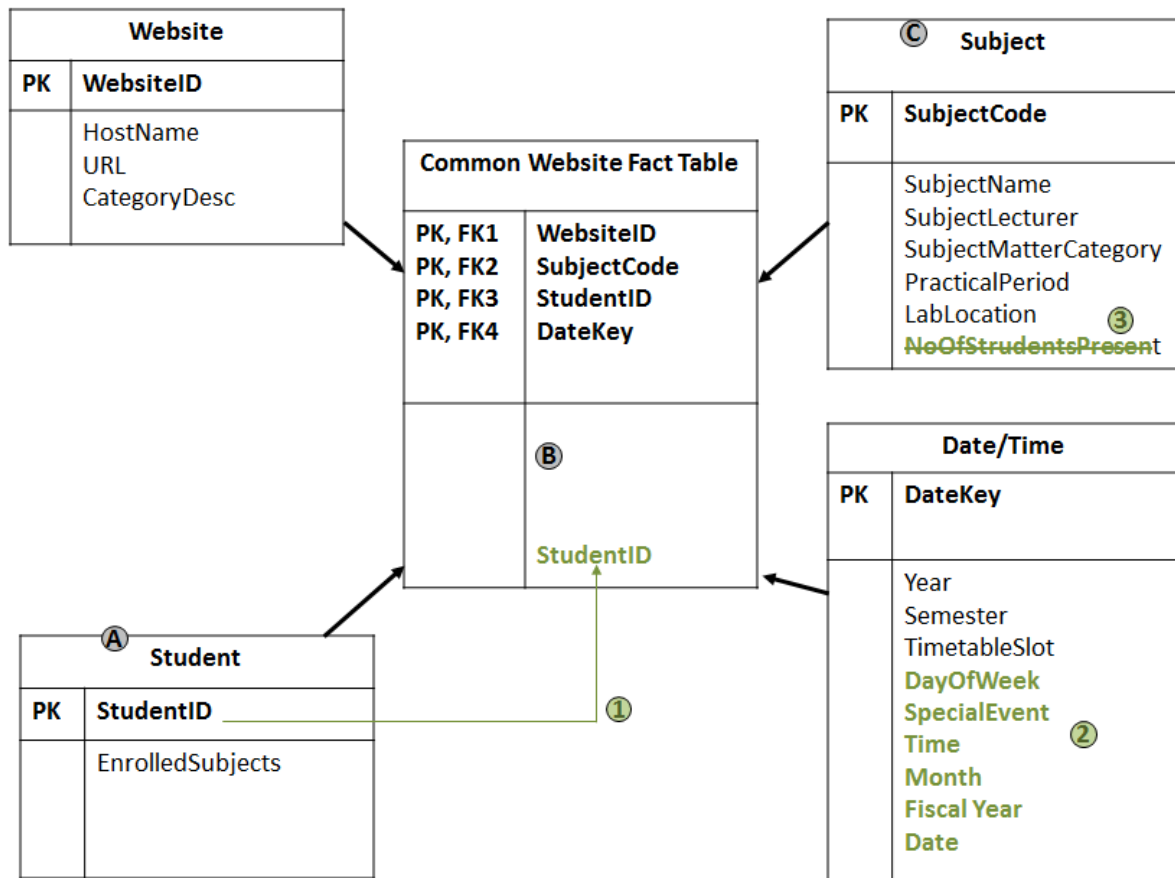


Figure 6.5 - Amended star schema

1.1) The comments on which the change A — *Webuser dimension changed to Student* was based are listed below:

Expert 3

- *Web User? I understand this to be student – odd naming convention.*

Expert 1

- *Why a Web user dimension? All students are web users? To replace Student Dimension?*

Expert 2

- *I assume Web User is a student table? One web user could have many subjects. The Enrolled Subject field does not make sense. One can either get the specific subject out of the fact table combinations OR one can create a stand-alone WebUserSubjectEnrolment Dimension for drilling across this one-to-many relationship.*

1.2)The comments on which the change **B** – *Common Website table facts “relevance” and “popularity” were removed to change to a factless fact table* was based are as follows:

Expert 3

- *I don’t understand what the facts ‘popularity’ and ‘relevance’ mean at this granularity. Seems to me it is an occurrence or not – factless fact table?*

Expert 1

- *Facts in fact table are not additive, suggest factless fact table design, except if you want to keep track of time duration, which can be averaged and totalled for all or for a particular student.*

Expert 2

- *Relevance and popularity in fact table are both metrics that could be constructed from SQL queries OR should be supplied by the specific lecturer. I see the rating of sites as relevant or not as one of the long term stumbling blocks for a project like this. It can easily become impossible to maintain, unless one has a separate stand-alone dimension that serves as a lookup table for sites.*

1.3)The comments on which the change **C** – *Practical dimension collapsed into subject dimension* was based are reflected below:

Expert 3

- *Subject and practical Dimension? Seems like it can be combined in one dimension without any apparent disadvantage.*

Expert 1

- *Schema is too normalised. Constraining across practical and subject will be difficult. Suggest a collapsed dimension with one row for every unique Student, Subject, Practical combination.*

Expert 2

- *The Practical Dimension in its current form has no data that cannot be collapsed into subject.*

1.4)Suggested changes are represented as numbers **1**, **2** and **3** in Figure 6.4. The corresponding suggestions are shown below:

Expert 2

- *1 – If web-user is only to capture the specific student involved (single value in dimension) this could be included in the fact table itself as an allowable non-numeric fact.*
- *2 – The NoOfPresentStudents field's value can be gained from a query against the fact table.*

Expert 3

- *2 – Date/Time can be attributed and filled in before the time; consider adding extra descriptors in there like 'day-of-week', 'special-event' (day before break for example) to allow lecturers to determine if there are other reasons for a trend. In the week when a major sports event takes place students may be side-tracked to sport sites, for example.*
- *3 – No_of_students present seems like a second fact table? But it's a dimension – confusing to me.*

Any change that was suggested by all three experts was regarded as indicating consensus and the change was duly made to the star schema. Any changes that were not suggested by all three experts were regarded as suggested changes and indicated as such on the star schema. None of the experts indicated that they neither agreed nor disagreed that the draft schema was suitable; the updated schema was therefore reviewed after the changes had been made to determine whether it was now suitable. The schema would be regarded as suitable if all three experts indicated this. The updated star schema, including suggested changes and all the comments made in the first review, was sent to all three experts as a second iteration of the star schema refinement. The results are shown below.

Student Web analysis dimensional star schema expert review iteration 2

Changes that were suggested by all three experts were applied to the schema.

Changes that were not suggested by all three experts were viewed as pending. These are indicated in Figure 6.6 by the numbers **1**, **2** and **3**.

Original specifications were:

Business Process: Planning for the achievement of primary educational objectives

Specific Instance: Planning lecturing approaches to enhance students' ability to retrieve useful subject-related information from the Web during allocated practical project classes

Granularity: *Unique website visited by a student during a practical class*

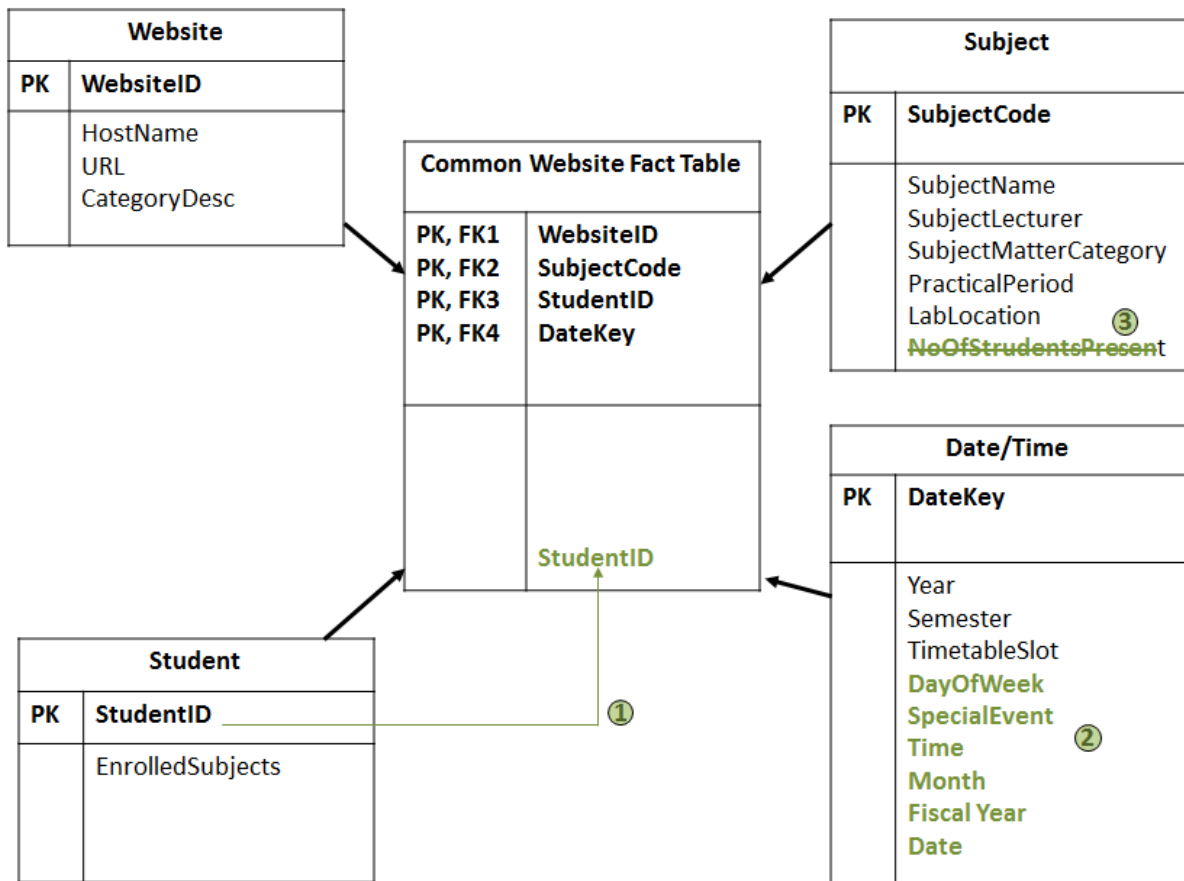


Figure 6.6 - Updated star schema

Question 1) Please indicate how strongly you agree or disagree that the above dimensional star schema is suitable for the given granularity and business process, using the following scale:

- 1) **Strongly Disagree** – you believe that the schema is highly unsuitable for the given granularity and business process.
- 2) **Disagree** – you believe the schema is unsuitable for the given granularity and business process
- 3) **Neither Agree or Disagree** – you cannot agree or disagree with the schema's suitability.
- 4) **Agree** – you believe that the schema is suitable for the given granularity and business process.
- 5) **Strongly Agree** – you believe strongly that the schema is suitable for the given granularity and business process.

Indicate level: ____

Comments and suggestions (Please indicate whether you agree that changes 1, 2 and/or 3 should be made.)

Question 2) **If you agree on certain changes** please indicate how strongly you agree or disagree that the above dimensional star schema is suitable for the given granularity and business process **if those changes were made**, using the following scale.

- 1) **Strongly Disagree** – you believe the schema to be highly unsuitable for the given granularity and business process, regardless of suggested changes.
- 2) **Disagree** – you believe that the schema is unsuitable for the given granularity and business process, regardless of suggested changes.
- 3) **Neither Agree or Disagree** – you cannot agree or disagree with the schema’s suitability.
- 4) **Agree** – you believe that the schema is suitable for the given granularity and business process if the changes you have agreed on are made.
- 5) **Strongly Agree** – you believe the schema to be highly suitable for the given granularity and business process if the changes you have agreed on are made.

Indicate level: ____

Student Web analysis dimensional star schema expert review iteration 2 results document

- 1) Expert 3 believed that the schema would be suitable for the given granularity and business process.
- 2) Expert 2 did not indicate whether he agreed or disagreed that the schema was suitable. However, he made the following suggestions, which he indicated would have no apparent advantage or disadvantage. They were therefore not considered as essential changes:
 - Not sure what PracticalPeriod in subject dimension is. Maybe a subject can have multiple practical periods – actually also true of lablocation.
 - Although we said practical can be folded in date/time, maybe this is a result of overloading or definition of terms. Maybe the “date/time”-type dimension should be described in a semantically richer manner – practical slots – which may again bring

up the argument of a separate dimension for “date” and “prac-opportunity” I’m a bit in two minds about this ... cannot clearly see the advantages/disadvantage.

- 3) Expert 1 believed that the schema was suitable for the given granularity and business process.

The first draft schema was clearly unsuitable. In the first review, none of the experts agreed that the draft schema was suitable for the business process. However, following the changes made based on the consensus reached by the experts and a second review, the star schema was deemed a suitable design. This schema was further validated when it was successfully implemented in the SWAN data mart and confirmed to be correctly implemented by expert 1. The expert review was invaluable to the design process of a suitable star schema, and without it the SWAN data mart would probably not have operated correctly. Based on these assertions and on the case in hand, the following guideline is proposed for the analysis of NMMU students’ Web usage behaviour:

Guideline 3

Consult individuals with data warehousing expertise when creating a star schema

Recommendations

- Identify individuals within the institution who could have data warehousing expertise
- Construct a draft schema and specify the business process for which it is intended and the required granularity
- Allow the individuals to rate the suitability of the draft schema and allow them make suggestions which may improve the design
- Make the appropriate changes and allow the individuals to review the draft schema once the suggested changes have been made if necessary, to consider the changes made and the input from other individuals. Repeat the review process until the schema is deemed suitable.

6.3.4 Manual Extract, Transformation and Load Process

Moving the data in its current format from its original data source to the SWAN data mart in an appropriate format was explored in order to address the question of how the data was moved and tailored and why it was done in this way.

It had been previously established that the sample log files were in the desired format. These were flat files containing huge amounts of Web activity data from the six NMMU campuses. In large DW/BI systems the process of extracting data from its source and presenting it to the users of the DW/BI system is a massive undertaking and requires the greatest resources and effort in the development project. The data has to undergo extensive editing, cleaning and profiling to make it readable and

meaningful to the users of the system. Furthermore, these systems store and handle data as it is produced and are therefore constantly updating their data tables. This allows for a live, almost real-time, account of the business processes. In order for data to be handled, this process is automated through an ETL system, (see Section 4.2 for further detail). In the case of the SWAN project, the ETL process of moving the data into the SWAN data mart to achieve the success criteria was done manually. An automated system would be difficult to achieve given the constraints of the scope of this project. Furthermore, there is no one-size-fits-all method for cleaning this type of data. The ETL process was thus approached in a trial and error fashion. A logical avenue was explored and, based on the results, abandoned or pursued. Figure 6.7 illustrates the process that proved successful, the details of which are discussed below:

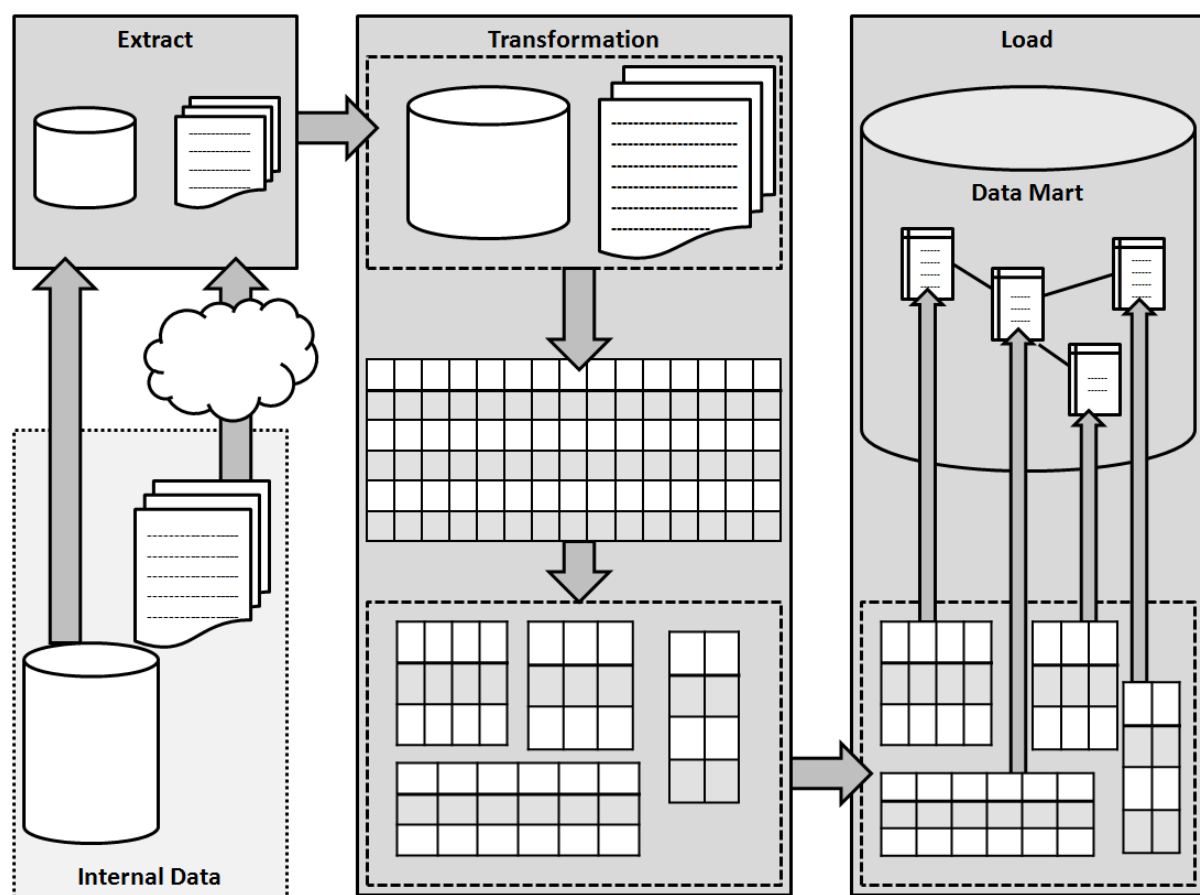


Figure 6.7 - Manual ETL process for the SWAN Data Mart

The first step was to obtain another log file from the system engineers who had supplied the previous sample, which was deemed usable. The software installed on the firewall device had a filtering capacity and ways to anonymise the data. Based on the type of report required, according to the business user interview as discussed above in Section 6.3.2, some criteria for the log file were established before it was requested. In this way, the system engineer could filter out unnecessary data on the firewall device and eliminate some irrelevant data. The criteria for the data were established as:

- Data from other campuses was not needed in this log file; the Web use activity of the students' of the business user (IT lecturer selected for information requirements interview) was needed, all of whom attend classes at North campus.
- Staff Web activity and non-Web based network traffic was of no use. The system engineer had already calibrated the device to exclude these in previous sample log files.
- Based on the practical class schedule of the business user, the log needed to contain entries for only one Friday in the academic calendar.
- Web activity was required from the R128 laboratory where the practical classes took place. The engineer could filter entries down to the building block in which the R128 laboratory was located but was unable to isolate the laboratory activity itself. This issue is discussed in the transformation section below.

Based on the above criteria, the system engineer uploaded the anonymised log file to a cloud-based application. All previous sample log files were anonymised in the same way; this meant that no identification whatsoever of individuals in the log file was possible.

A link to the cloud application was sent to the researcher along with a password to access the folder containing the requested log file. This was the first phase of the ETL process, that is extracting the log files from the cloud application. With reference to Figure 6.7, the internal data was the requested log file and other subject specific data such as practical time, subject codes, timetable information etc. The subject specific data is available to students on the shared network drives and students access it from the information section of the NMMU website. Hence, the data was extracted differently to the log file. No further extraction took place; the data needed to be explored and profiled to determine how it could be appropriately transformed to fit into the dimensional model residing in the SWAN data mart, as discussed in Section 5.4.2.

With reference to Figure 6.7, after the extraction phase, the data began to undergo transformation. The log file, which was in a flat file format (plain text file), was opened to profile the data, as shown in Figure 6.8:

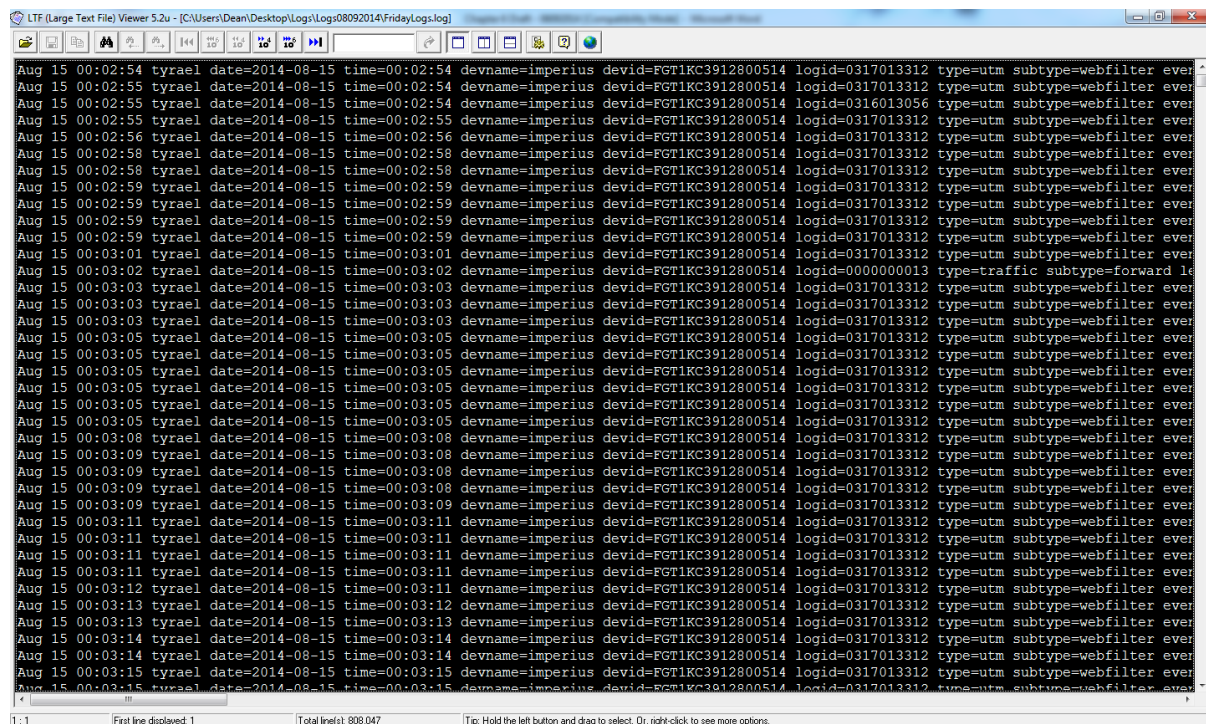


Figure 6.8 - Log file

At a glance it was determined that the log file contained one day’s worth of Web activity, specifically Friday 15 August from 00:00 to 00:00, in other words 24 hours. The format was consistent with the previous sample and contained the required fields such as hostname, URL, category descriptions, source IP etc. The next step was deciding how the data would be edited. The flat file in its current format was a read-only file so could not be edited. Even if it could, the number of entries was very large (808 047) and it would possibly have required some type of sorting functionality to handle them. The file needed to be moved to an application with the appropriate functionality for editing large data sets, which the Large Flat File Viewer application used to open the flat file (see Figure 6.8) did not support. Previous research had established that the SQL server and Microsoft Excel could provide this functionality. The SWAN data mart resides in the SQL server and initially this was used in an attempt to profile and edit the data. At the time this seemed logical because the data would then be in one place and could easily be moved into the SWAN data mart tables. However, moving the log file entries into the SQL server proved problematic; the SQL server imports data through a wizard-based process called “Import and Export data (64 bit)”. This requires the SQL server to have a destination table with the same number of columns as the source data (flat/log file) and it maps the data directly onto this. In this case, if the log file entries had had a consistent field format there would have been no difficulty in doing this. However, each log entry could contain various sets of fields. If the logs were to be imported to a data base table these fields would identify the column into which the value was to be inserted. Therefore, when importing into the SQL data base these inconsistent fields would result in inconsistent columns, and values would be incorrectly inserted into other destination columns.

Fields refer to the description of the data in each entry. For example, in Figure 6.9, the first entry contains logid, type, subtype etc. The values for those fields precede the '=' symbol, hence a field=value relationship exists. In data base terms these fields will become columns as they describe the value for the entry or line in a table.

An example is provided in Figure 6.9, where the two highlighted entries have different sets of fields. The top entry contains an eventtype field whereas the entry underneath does not. This is one of numerous instances of inconsistent fields in entries.

```
logid=0317013312 type=utm subtype=webfilter eventtype=ftgd_allow level=notice vd="root"
logid=0317013312 type=utm subtype=webfilter eventtype=ftgd_allow level=notice vd="root"
logid=0317013312 type=utm subtype=webfilter eventtype=ftgd_allow level=notice vd="root"
logid=0317013312 type=utm subtype=webfilter eventtype=ftgd_allow level=notice vd="root"
logid=0000000013 type=traffic subtype=forward level=notice vd=root srcip=255.39.192.175
logid=0317013312 type=utm subtype=webfilter eventtype=ftgd_allow level=notice vd="root"
logid=0317013312 type=utm subtype=webfilter eventtype=ftgd_allow level=notice vd="root"
logid=0317013312 type=utm subtype=webfilter eventtype=ftgd_allow level=notice vd="root"
```

Figure 6.9 - Entry with column divergence

The variance in fields is further expanded by the data importing wizard used by the SQL server. The wizard requires a column delineator in order to separate and identify the fields in the flat file entries. Flat file entries are primitive strings requiring explicit delineators to separate the required fields. Figure 6.10 shows a snippet entry in the log file as the entry is too long to place in one line and still remain readable for demonstration purposes. By specifying a column delineator, the importer will split the values between the delineator into columns. For example, if the "=" delineator is used, the entry will be split as shown in line 2 and 3 in Figure 6.10 below. In this case, this was not an appropriate delineator as it grouped multiple unrelated values together. The second value, underlined in line 3, grouped the devid value with the logid field, utterly compromising the integrity of the data.

```
Line 1) devid=FGT1KC3912800514 logid=0317013312 type=utm subtype=webfilter

      ↓               ↓               ↓               ↓
Line 2) devid=FGT1KC3912800514 logid=0317013312 type=utm subtype=webfilter

Line 3) [devid] [FGT1KC3912800514 logid] [0317013312 type] [utm subtype] [webfilter]
```

Figure 6.10 - Entry split by "="

The field=value relationship must be maintained. For example, the logid for the entry in Figure 6.10 is 0317013312 (line 1 logid=0317013312). These field=value grouped data values are separated by a single space in the entries.

Using the single space delineator as “ ” in the data importer resulted in the splitting of the entries, as illustrated in Figure 6.11.

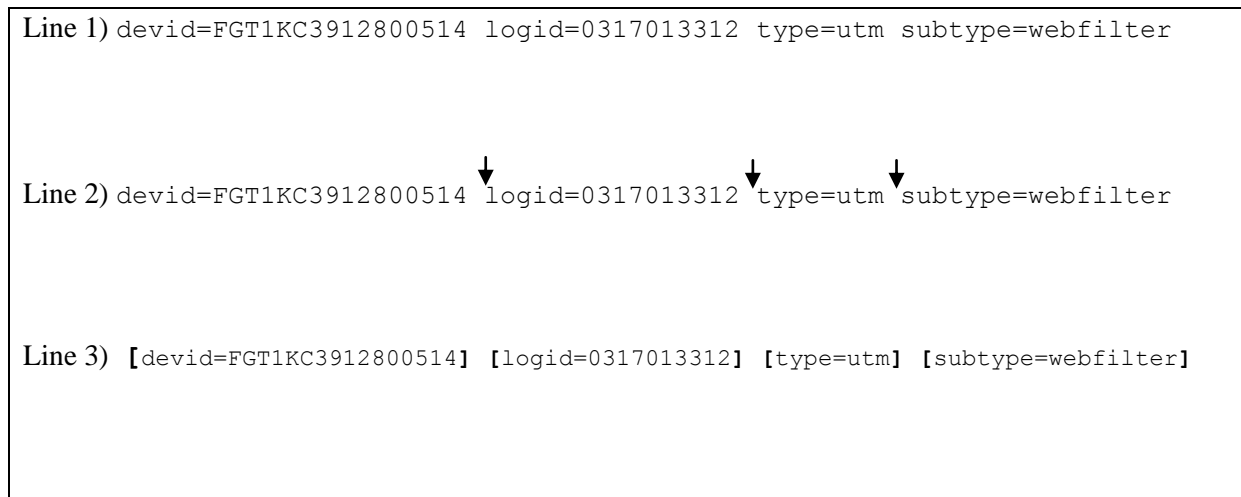


Figure 6.11 - Entry split by a single space

As indicated in Line 3 in Figure 6.11, the values are split while maintaining the field=value relationship. However, this approach creates a further increase in column divergence over and above the inconsistent fields in each entry. Figure 6.12, a snippet from another part of a log, demonstrates the reason for this. Line 1 is split by a single space, as in Figure 6.11. But in line 3 in Figure 6.12, it works as planned for srcip and srcport where the column is attached to the value as a field=value relationship. However, the value for srcintf has a single space within its value and is therefore separated into its own column and is detached from its field=value relationship. Where there should be three columns there are now four. It is in this way that this approach created further inconsistencies amongst entries and broke up the data in an undesirable way. Figure 6.12 illustrates just one of numerous instances of this problem within the log file.

```
Line 1) srcip=191.214.228.130 srcport=53340 srcintf="Internal 10G"

Line 2) srcip=191.214.228.130 ↓ srcport=53340 ↓ srcintf="Internal 10G"

Line 3) [srcip=191.214.228.130] [srcport=53340] [srcintf="Internal"] [10G"]
```

Figure 6.12 - Column divergence

The next step was to move the entries into the SQL server using this approach and then to attempt to correct the column deviance. However, the wizard-based data importer mentioned above did not allow the data to be inserted. It returned an error stating that the entries were too large for the destination table which contained columns automatically generated from the first line in the flat file. This was the result of the column divergence. A solution was found by creating a destination table with excess columns to cater for the largest set of columns in any possible entry. In this way all the entries could be inserted without the importer returning an error. The entries were successfully imported into a staging table in the SQL server. However, the number of entries in the flat file did not match the number indicated as successfully imported by the importer. There were fewer entries in the staging table than in the log file as some entries had somehow been lost during the importation. When this was investigated it became clear that certain entries with smaller column sets (shorter entries) had forced the importer to insert entire entries into the remaining columns in the lines in the table as these shorter entries did not completely populate a line. As illustrated in Figure 6.13 below, when a shorter log entry is inserted into the staging area it does not populate all the columns in the line on the staging table. Subsequently, the importer attempts to populate the excess space in that line by pulling the next log entry from the flat file into it. As a result, a number of entries may be lost during the importing of the log entries and the integrity of the data set is compromised. Retrieving the lost logs from the table incurs unnecessary costs. In this study it was decided that the approach of editing the log file using the SQL server would be too expensive to justify.

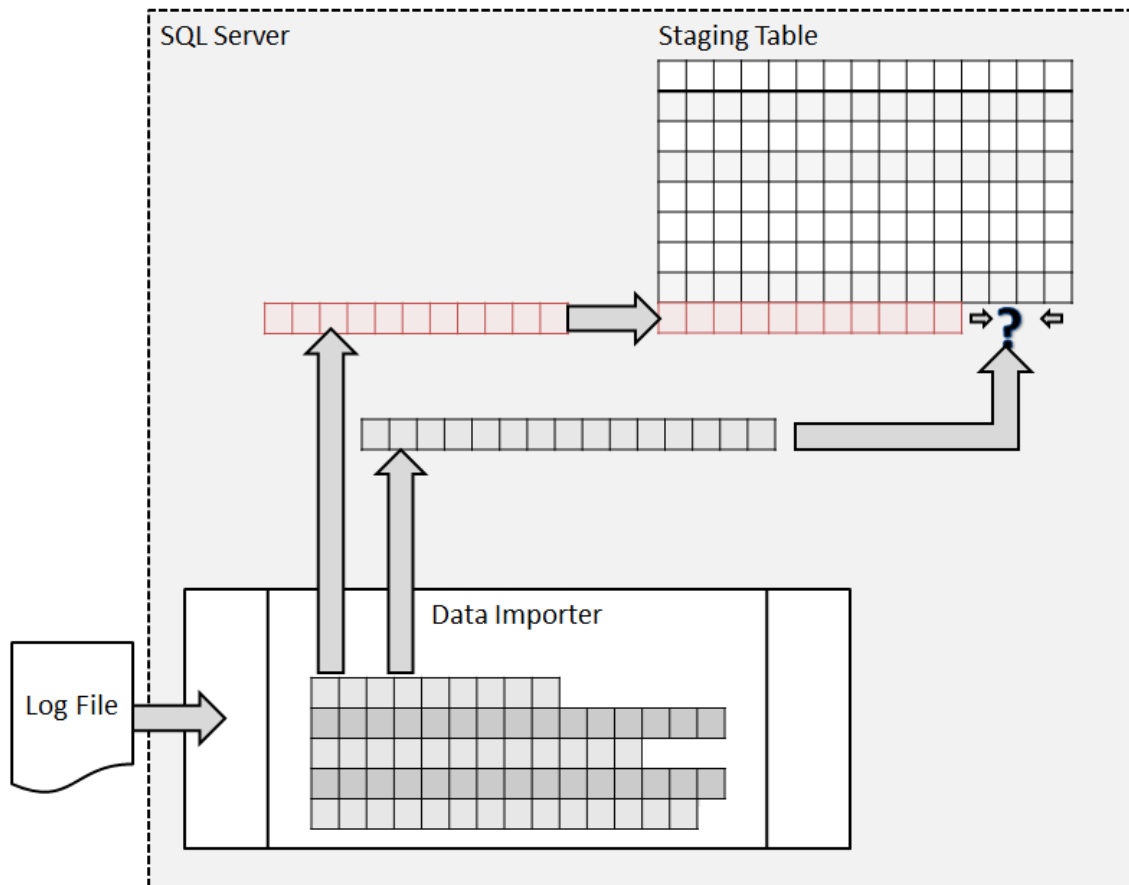


Figure 6.13 - SQL staging table import issue

Microsoft Excel is a software application that can perform the desired flat file editing. Excel has a data importing function similar to that of the SQL server. However, Excel has no table destination constraints and the logs were imported into an Excel spreadsheet using the single space delineator with no loss of entries. Excel spreadsheets have extensive column space and can hold around 1 000 000 lines in one sheet. As shown in Figure 6.14, if an entry was shorter it did not affect the next entry. Longer entries were simply extended into available column cells and shorter entries took up less space.

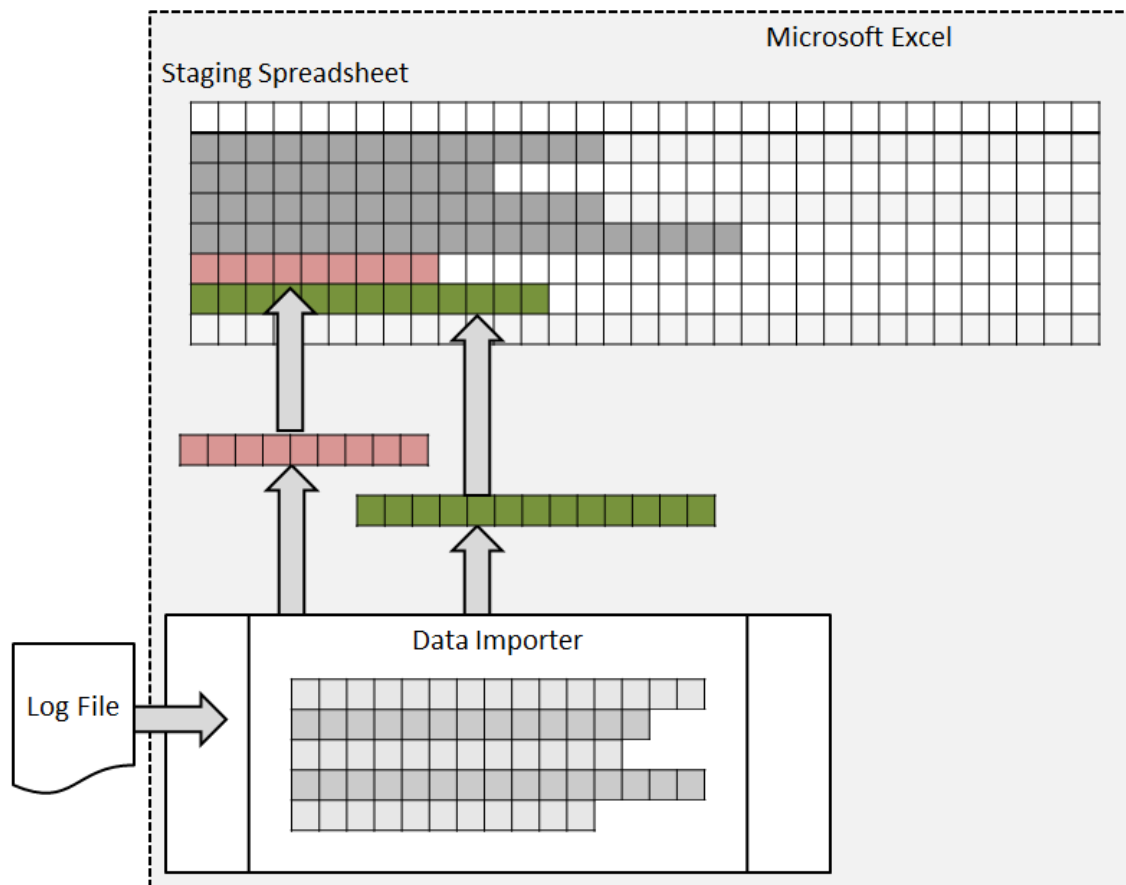


Figure 6.14 - Data import in Excel

With the log entries in the Excel spreadsheet, the data could be examined and edited. The extent of column divergence in the entries was determined by sorting the entries by columns. First they were sorted by logid. Logid is a numeric value given to an entry by the firewall device and indicating the type of log entry it is. Log entries of the same type have the same value types and therefore the same column size. By grouping the entries in this way and checking the various grouped entries it was determined that the majority of the logs shared the same logid. However, they still contained a high level of column deviance as URL and other values contained a number of spaces that caused a similar problem as the one described above in the SQL server staging table (see Figure 6.12). For this reason, it would be difficult to clean the logs by unifying or standardising all the columns. Instead, based on the fact that the values required from the entries were clear, another approach was considered. Given the column=value relationship, the values could be extracted to a separate column for each entry through a search formula supported by Excel. After various attempts, a formula was constructed. Figure 6.15 shows how the values were extracted.

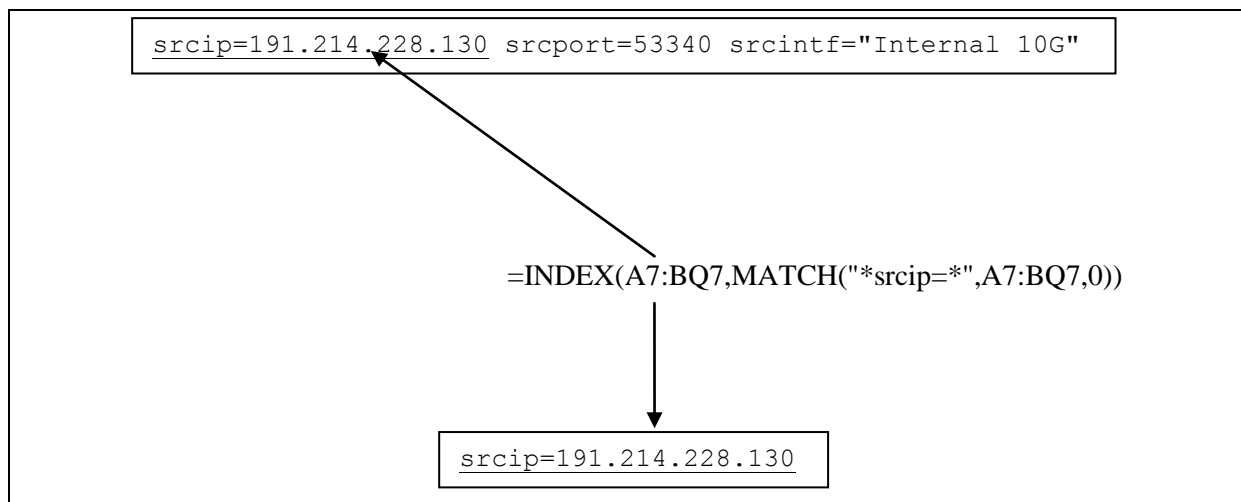


Figure 6.15 - Excel log entry Field=Value search formula

Using this formula, the field needed was specified, that is the cell that contained the field followed by the "=" symbol was extracted to another cell. The "=" symbol ensures that any other possible value that may contain the field name is not returned by the formula and that only the field name as logged by the firewall is recognised by the formula. This formula was then changed, based on the field and value needed, and then copied across all the entries, thereby extracting all the required fields and values specified by the parameter, while maintaining alignment with the other values of the entry, as shown in Figure 6.16.

	BV	BW	CA	CB	CC	CD	CE
1	Hostname	CleanHostname	Time	CleanTime	date	cleanDate	Userid
2	hostname="s.txtsrving.info"	s.txtsrving.info	time=14:44:57	14:44:57	date=2014-08-15	2014-08-15	user="6fe3ca459d"
3	hostname="www.google.co.za"	www.google.co.za	time=10:07:52	10:07:52	date=2014-08-15	2014-08-15	user="cecea379aa"
4	hostname="www.mp3olimp.net"	www.mp3olimp.net	time=09:59:01	09:59:01	date=2014-08-15	2014-08-15	user="323db5f592"
5	hostname="feed.yourfiledownloader.net"	feed.yourfiledownloader.net	time=10:38:58	10:38:58	date=2014-08-15	2014-08-15	user="23bb05e5e3"
6	hostname="feed.yourfiledownloader.net"	feed.yourfiledownloader.net	time=10:40:08	10:40:08	date=2014-08-15	2014-08-15	user="23bb05e5e3"
7	hostname="feed.yourfiledownloader.net"	feed.yourfiledownloader.net	time=10:40:16	10:40:16	date=2014-08-15	2014-08-15	user="23bb05e5e3"
8	hostname="feed.yourfiledownloader.net"	feed.yourfiledownloader.net	time=10:40:17	10:40:17	date=2014-08-15	2014-08-15	user="23bb05e5e3"
9	hostname="feed.yourfiledownloader.net"	feed.yourfiledownloader.net	time=10:40:17	10:40:17	date=2014-08-15	2014-08-15	user="23bb05e5e3"
10	hostname="feed.yourfiledownloader.net"	feed.yourfiledownloader.net	time=10:40:22	10:40:22	date=2014-08-15	2014-08-15	user="23bb05e5e3"
11	hostname="clkmon.com"	clkmon.com	time=09:46:45	09:46:45	date=2014-08-15	2014-08-15	user="e87355c673"
12	hostname="clkrev.com"	clkrev.com	time=11:13:49	11:13:49	date=2014-08-15	2014-08-15	user="5680910791"
13	hostname="clkrev.com"	clkrev.com	time=11:13:49	11:13:49	date=2014-08-15	2014-08-15	user="5680910791"
14	hostname="clkrev.com"	clkrev.com	time=11:13:49	11:13:49	date=2014-08-15	2014-08-15	user="5680910791"
15	hostname="clkrev.com"	clkrev.com	time=11:15:59	11:15:59	date=2014-08-15	2014-08-15	user="5680910791"
16	hostname="clkrev.com"	clkrev.com	time=11:15:59	11:15:59	date=2014-08-15	2014-08-15	user="5680910791"
17	hostname="clkrev.com"	clkrev.com	time=11:15:59	11:15:59	date=2014-08-15	2014-08-15	user="5680910791"
18	hostname="clkrev.com"	clkrev.com	time=11:16:53	11:16:53	date=2014-08-15	2014-08-15	user="5680910791"
19	hostname="clkrev.com"	clkrev.com	time=11:16:53	11:16:53	date=2014-08-15	2014-08-15	user="5680910791"
20	hostname="clkrev.com"	clkrev.com	time=11:16:53	11:16:53	date=2014-08-15	2014-08-15	user="5680910791"
21	hostname="zilliontoolkitusa.info"	zilliontoolkitusa.info	time=11:00:32	11:00:32	date=2014-08-15	2014-08-15	user="cecea379aa"

Figure 6.16 - Excel formula demonstrations

Once the field and values had been extracted they could be cleaned by removing field name from the cell value (“field=value”). The remaining data required for the dimensional model was extracted in this way. The following fields and values were extracted:

- Date
- Time
- SessionID
- UserID
- Hostname
- URL
- Category
- srcip

The URL raised difficulties as many of the values for the URL field contained single spaces and were therefore split across several adjacent cells, as demonstrated in Figure 6.12. Subsequently, the formula used to extract the URL values drew out only partial values. No solution to this was found. However, the URL data was not a critical attribute in the dimensional model and this was considered a minor issue.

Once all these values had been cleaned they were moved to a separate worksheet. This new worksheet contained only the desired columns and clean values. This worksheet was then imported into an SQL server staging table from which the relevant tables in the dimensional model tables could be populated. The transformation phase was considered successful once the data from the log files had been translated directly into the dimensional model. In other words, the attributes listed in each dimension could be loaded with the correct data, which reflected what was needed and was readable.

The Website dimension was populated in the SQL server by inserting the Hostname, URL and Category of the entries into the staging table, which now contained the data from the clean worksheet in Excel. These values were inserted into the Website dimension only if they met the criteria, as indicated in Figure 6.17. The entries had to have occurred between 10:00 and 15:10 on a normal academic Friday (no holidays) and had to have come from the R128 laboratory. In this way, the Web activity of the IT lecturer's students would be present in the Website dimension. The staging area contained the entire R-block's Web activity. The system engineer provided a list of IP addresses which the network had dynamically assigned to the laboratories during the day. Using this list, Web activity which was originating from the R128 laboratory could be isolated. This is done by retrieving entries whose source IP address corresponds with the R128 IP address list. The source IP address is the IP address of the computer which sent the request to the Web.


```

USE SWAN

INSERT INTO dbo.DimWebsite (HostName)

SELECT DISTINCT Hostname FROM dbo.LogStagingArea

/*This condition filters logs from the practical period*/
WHERE AccessTimestamp BETWEEN '10:45:00' and '15:10:00'

/*This condition isolates the R128 IT lab*/
AND (
    SourceIP = '130.6.186.60' OR SourceIP = '15.159.112.27' OR SourceIP = '16.4.109.39'
    OR SourceIP = '169.183.183.203' OR SourceIP = '170.98.142.218' OR SourceIP
    = '175.203.145.43' OR SourceIP = '194.216.232.41' OR SourceIP = '2.142.18.177' OR
    SourceIP = '210.202.25.65' OR SourceIP = '217.159.132.66' OR SourceIP
    = '246.247.113.47' OR SourceIP = '29.98.185.221' OR SourceIP = '46.185.216.58' OR
    SourceIP = '5.13.226.126' OR SourceIP = '74.233.141.32' OR SourceIP = '77.144.87.55'
    OR SourceIP = '93.74.121.130' OR SourceIP = '94.154.102.157' OR SourceIP
    = '94.172.55.67' OR SourceIP = '98.175.166.98'
)

AND CategoryNum = '52'

```

Figure 6.17 - Website dimensional load query

The remaining dimension tables were then populated manually. The fact table was populated with primary key values to link the dimensions together. The fact table was then queried to provide a report, discussed in Section 5.4.

Using Excel for the initial profiling and editing of the data proved a highly successful alternative to the SQL server. The importing of data from a flat file was allowed despite gross inconsistencies in the entries. This was desirable as flat files with log data are likely to be inconsistent and need to be edited. The sorting functionality in Excel proved effective in establishing the extent of column divergence. Moreover, the formulas available to enable searching the data and extracting values proved essential to this ETL process. Given the manual nature of this ETL process, the issues that arose when dealing with log files in this format and how these issues were resolved, the assertions made about Web usage behaviour analysis of students in the NMMU, based on the present case, were consolidated as the following guideline:

Guideline 4

Profile, sort, isolate and extract valuable values from the data in Excel or a similar spreadsheet base application first. Then load into the data marts dimensional tables, use discretion if the log formats are different.

Recommendations if the sample log file contains inconsistent fields

- Export the a sample log file into a spreadsheet using Excel or similar spreadsheet software
- Utilize sorting functionality to determine the level of inconsistency of the fields of the entries
- Use search formulas to pull required values from cells in each entry
- Move the values into a separate spreadsheet
- Export the spreadsheet into the appropriate dimensional tables or staging table in the destination database

6.3.5 Refining the Development Process and Maintaining Scope

The issues surrounding mapping and implementing the end to end development process while maintaining scope were explored to answer the questions of how the development of the data mart was approached and why it was done in this way.

DW/BI systems can be intricate, resource intensive, complex, time consuming and exceptionally large, requiring data warehousing expertise and multiple developers. During the early planning stages of the SWAN project, given the amount of data available and the scope limitations, it was decided that the development of a prototype data mart using a sub set of data would an appropriate approach. Considering what needed to be achieved from the development, the following steps were required:

- Determine whether the data could provide meaningful information using the Kimball Lifecycle (KL) method.
- Define a business process for which the Web usage information could be used.
- Present meaningful information to a potential user/lecturer.
- Implement and understand a single end to end iteration of the KL.
- Define a process for cleaning the Web usage data.
- Create and implement a table structure in the form of a suitable star schema.

Developing a DW/BI system within the limitations of time and scope posed by this research study would not allow the achievement of these steps. Moreover, the overall intention was to gather useful heuristics from the development process, not to deliver a complete system. Therefore, a prototype data mart was considered logical.

Developing a DW/BI system is a massive undertaking. By creating a prototype data mart instead of attempting to create a full system, some working relationships were developed and many issues, which may otherwise have been discovered only later on, became apparent early in development process. The logistics of moving the data from source to destination and methods to clean the data became clear. Information could be presented to the intended users within a reasonably short period, which confirmed the value to be gained from the data. Moreover, the prototype demonstrated that the

method used to develop it was appropriate. An understanding of how resource intensive a full DW/BI system development effort could be was formed. The use of a sub set of data allowed for detailed data profiling which meant that the process used to clean the data could be completed and thoroughly understood. Moreover, the method used to develop the prototype allows for expansion from a single data mart to a full DW/BI system by the development and connection of additional data marts. Based on these assertions, the following guideline was proposed for Web usage behaviour analysis of students at NMMU:

Guideline 5

Develop a prototype data mart which provides meaningful information for a single business process using only a sub set of data

Recommendations

- Develop a single, self-serving, small scale prototype data mart before considering a full DW/BI system
- Use a small set of sample data
- Focus on a single business process in one department

6.4 Guideline verification

In order to determine whether the proposed guidelines were correct, the guidelines as entity concept will be discussed, in order to better understand their purpose and use.

Guidelines can take various forms. Using a generic definition, guidelines are a set of recommendations and/or considerations for a well-defined and specified process and could include suggestions, models or prerequisites (Hevner et al., 2004; Mell & Grance, n.d.; Panel, 1998).

Some characteristics of guidelines are that they do not require strict adherence, they apply to a very specific setting and process and their validity is based on the evidence used to develop them. Guidelines as a research output in IS are used by some authors as a seemingly flexible research output (Gorgone et al., 2003; Hevner et al., 2004; Mendling, Reijers, & Aalst, 2010; Straub & Gefen, 2004). In clinical and healthcare research, guidelines are used extensively as research outputs. Medical practitioners commonly require clear assistance when undertaking their practice and making crucial, informed decisions, and guidelines can provide this assistance (Oxman, 2004). Hence, the extensive use of guidelines in this field is well supported.

Based on (Oxman, 2004), characteristics that guidelines in clinical and healthcare research should have are:

- Validity – based on sound evidence
- Clarity – presented in such a way that they cannot be misunderstood
- Flexibility – should be applicable to a variety of cases where the process applies
- Completeness – should cover the entire process

Further, given that the proposed guidelines form the IT artefact in this research study, they should have (Peffer, Tuunanen, Rothenberger, & Chatterjee, 2007a):

- Utility
- Efficacy
- Quality

Based on the desired traits of guidelines as an IT artefact and as used in clinical research, the guidelines were validated through expert review.

6.4.1 Expert Review Validation

The experts asked to review the guidelines had extensive knowledge of data warehousing and knowledge management in an educational context as well as in industry (see Section 6.3.3). They were given a review form, attached as Appendix E. Each guideline was rated for clarity and completeness, validity and flexibility. Furthermore, the guidelines were rated for utility as a deliverable set; thereby, encapsulating the desired traits of guidelines and desired IT artefact characteristics. The experts were asked to indicate how strongly they agreed or disagreed that the guideline met the desired characteristics. The responses from each expert are attached as Appendix F.

Guideline 1

The experts agreed that guideline 1 was clear and unambiguous. However, expert 1 required that the data should be referred to as Web usage data and would agree should this change be made. Subsequently, the guideline is updated to:

The Web usage data and the owner/s thereof should be investigated to gather and profile a sample of the required Web usage data before the DW/BI project should be considered feasible.

Recommendations

- Investigate which constituents administer the network infrastructure of the institution
- Establish which constituent/s of the institution is/are accountable for the control of the Web usage data

- Enquire about the network infrastructure to determine how Web access is provided to the institution
- Identify an appropriate device which logs, or could log Web usage data
- Acquire a sample of the Web usage data from the device
- Profile the sample to determine its current format

Considerations if a proxy model is in place

- The proxy server or device configured to fill the role of a proxy server should be investigated as a source of Web usage data
- Investigate the logging capacity of the proxy server or relevant device and consult the administrators to determine if it is plausible to configure the device to log Web usage data in an appropriate format. Logging Web usage data may have negative performance repercussions on the network

The underlined changes indicated above further describe the data to be clearer. Therefore, it was considered an improvement and would not negatively affect the other experts' reviews. The experts agreed that the guideline was correct and accurate. Furthermore, they agreed that the guideline would be appropriate in another university with a similar network infrastructure.

The experts agreed that guideline 2 was clear and unambiguous. Furthermore, they agreed that guideline 2 was correct, accurate and that the guideline would be appropriate in another university with a similar network infrastructure.

The experts agreed that guideline 3 was clear and unambiguous. However, expert 1 stated that the guideline should specify dimensional modelling expertise and not generalize to data warehouse techniques. Guideline 3 was subsequently updated to:

Consult individuals with data warehousing dimensional modelling expertise when creating a star schema

Recommendations

- Identify individuals within the institution who could have data warehousing dimensional modelling expertise

- Construct a draft schema and specify the business process for which it is intended and the required granularity
- Allow the individuals to rate the suitability of the draft schema and allow them make suggestions which may improve the design
- Make the appropriate changes and allow the individuals to review the draft schema once the suggested changes have been made if necessary, to consider the changes made and the input from other individuals. Repeat the review process until the schema is deemed suitable.

The underlined changes indicated above further describe the data to be clearer. Therefore, it is considered an improvement and would not negatively affect the other experts' reviews. Two of the 3 experts agreed that the guideline was correct and accurate, the third could did not agree or disagree and remained neutral. Two of the 3 experts agreed that the guideline would be appropriate in another university with a similar network infrastructure the third did not agree or disagree and remained neutral.

Two of the experts agreed that guideline 4 was clear and unambiguous; the third could did not agree or disagree and remained neutral. The experts agreed that the guideline was correct and accurate. Two of the experts could not agree or disagree; the third agreed that the guideline would be appropriate in another university with a similar network infrastructure.

The experts agreed that guideline 5 is clear, unambiguous, correct and accurate and would be appropriate in a university with a similar infrastructure.

Two of the experts consider the set of guidelines to be useful for their intended context. Expert 3 disagreed and stated that the set is very high level should specify a process. It should address how rather than what.

The guidelines as a set were intended to be a holistic, high level IT artefact and therefore the disagreement mentioned for the guideline set was not seen as a major consideration. Individually the guidelines are considered by the experts to be clear, unambiguous, correct and accurate and applicable in other universities with a similar network infrastructure. For some guidelines experts may not have an opinion or feel they cannot agree or disagree. In these cases the neutral response is not considered and the conclusion falls to the responses with a clear agree or disagree. However, no such case exists for any of the guidelines where a majority of the responses have disagreed. The responses all fell into the majority of either agreeing or strongly agreeing. Therefore, the guidelines proposed, as an IT artefact, are considered a plausible artefact based on the expert review. The guidelines are presented as follows:

SWAN Guidelines

Guidelines for the Analysis of Student Web Usage in Support of primary Educational Objectives

1) The Web usage data and the owner/s thereof should be investigated to gather and profile a sample of the required Web usage data before the DW/BI project should be considered feasible.

Recommendations

- Investigate which constituents administer the network infrastructure of the institution
- Establish which constituent/s of the institution is/are accountable for the control of the Web usage data
- Enquire about the network infrastructure to determine how Web access is provided to the institution
- Identify an appropriate device which logs, or could log Web usage data
- Acquire a sample of the Web usage data from the device
- Profile the sample to determine its current format

Considerations if a proxy model is in place

- The proxy server or device configured to fill the role of a proxy server should be investigated as a source of Web usage data
- Investigate the logging capacity of the proxy server or relevant device and consult the administrators to determine if it is plausible to configure the device to log Web usage data in an appropriate format. Logging Web usage data may have negative performance repercussions on the network

2) Conduct face to face interviews with the intended users of the Web usage information and pose questions which directly focus on exactly what information they would consider valuable

Recommendations

- Identify potential users of the information which could be derived from the available Web usage data. This could be done by targeting a group which has perceived association with the information and gathering their information needs through a group meeting or survey

- Potential users would be identified from the meeting or survey results if they indicate strong influence to their decisions from the proposed information
- If no potential users are identified reconsider the perceived association mentioned above
- From these potential users, interview one or more and ask questions which directly focus on determining exactly what information they would consider valuable
- Clarify any misunderstandings with the interviewee
- Document the interview through recording and creating a transcript of the interview for future reference.

3) Consult individuals with data warehousing dimensional modelling expertise when creating a star schema

Recommendations

- Identify individuals within the institution who could have data warehousing dimensional modelling expertise
- Construct a draft schema and specify the business process for which it is intended and the required granularity
- Allow the individuals to rate the suitability of the draft schema and allow them make suggestions which may improve the design
- Make the appropriate changes and allow the individuals to review the draft schema once the suggested changes have been made if necessary, to consider the changes made and the input from other individuals. Repeat the review process until the schema is deemed suitable.

4) Profile, sort, isolate and extract valuable values from the data in Excel or a similar spreadsheet base application first. Then load into the data marts dimensional tables, use discretion if the log formats are different.

Recommendations if the sample log file contains inconsistent fields

- Export the a sample log file into a spreadsheet using Excel or similar spreadsheet software
- Utilize sorting functionality to determine the level of inconsistency of the fields of the entries
- Use search formulas to pull required values from cells in each entry

- Move the values into a separate spreadsheet
- Export the spreadsheet into the appropriate dimensional tables or staging table in the destination database

5) Develop a prototype data mart which provides meaningful information for a single business process using only a sub set of data

Recommendations

- Develop a single, self- serving, small scale prototype data mart before considering a full DW/BI system
- Use a small set of sample data
- Focus on a single business process in one department

6.5 Conclusion

Web usage data can provide meaningful information about the Web usage behaviour of individuals. However, a seemingly simple process of converting data into information can become extremely complicated, requiring careful planning and vast resources. The SWAN project implemented well established techniques and documented logical solutions to problems, providing detailed accounts of the development process. The project has proved that there is a great deal to be learned from observing the logical processing efforts of others.

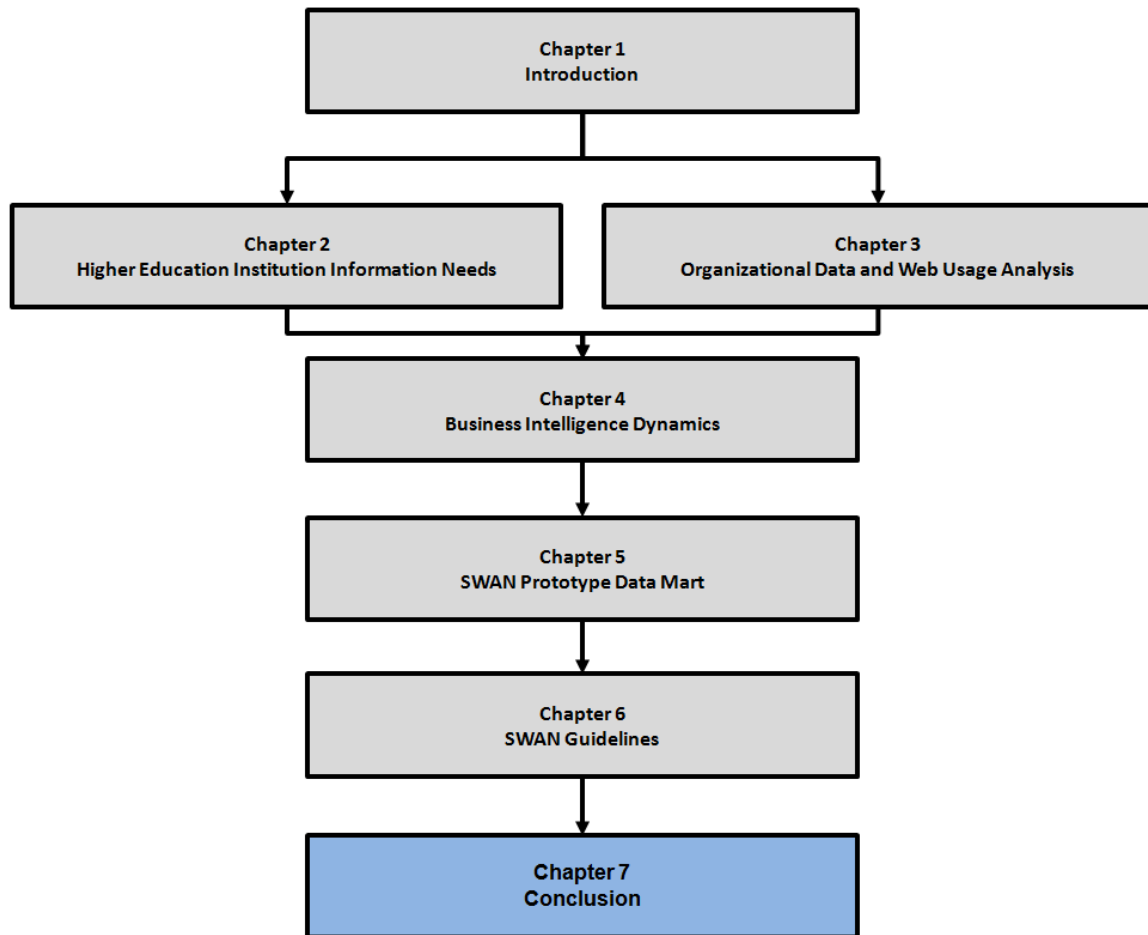
The findings from this case were translated into a set of proposed guidelines encapsulating the lessons learned and the logical assertions derived from the SWAN project. These guidelines could be used in a similar educational context to achieve similar outcomes. However, they are based on the account of observation of and exposure to a complex problem. Therefore, although it cannot be said that the case is a representation of all DW/BI projects in an educational context where Web usage is of interest, these guidelines could provide guidance in similar instances.

Hevner et al describe a category of accepted IT artefacts in DSR (method) which refers to the informing of a process (2004). Therefore, guidelines could be considered a form of method in this research as they provide guidance for a defined process. As a research output this type of contribution is considered "knowledge as operational principles" (Purao, 2002).

These guidelines as the delivered IT artefact provide assistance in the form of situational recommendations when developing a data warehouse or data mart for the analysis of student Web

usage data. The information produced by such a data warehouse or data mart could be of high value in informing decisions towards primary educational objectives and gaining insight into the information exposure and trends of students. This could provide useful information for decisions made in achieving the predominant objectives (see Section 2.3.2).

Chapter 7 - Conclusion



This chapter concludes this dissertation by summarising the findings and discussing the contributions made by this study.

7.1 Summary

Following the DSRP, this research involved the development of a prototype data mart designed to analyse student Web usage data using well established techniques. This prototype, and the development thereof, was studied to establish the principles of indicating Web usage behaviour of students. This resulted in the design of a set of guidelines as an IT artefact.

Chapter 1 introduced the background information and context to this research study. Higher education, Web usage analysis and the integration of IT in higher education institutions were introduced; this is the context within which an organisational issue was identified. The primary research objective was defined as a means to reach or contribute to a solution to this organisational issue. Secondary objectives were defined in order to examine the tasks required to satisfy the primary research objective. A design science research process as per Peffers et al (2007) was selected. The manner in which these steps were followed was discussed, together with the research methods.

Chapter 2 focused on information access and education. The way in which access to information has changed, with particular emphasis on the emergence of the Internet and the Web, was discussed. The implications of the paradigm shift in information access in higher education and the integration of IT were investigated. The primary objectives of higher education, viewed from the perspective of government, institutions and educators, were discussed. There are clear differences in the objectives expressed in these views. The educators' objectives were of particular interest and a survey was conducted using NMMU North Campus IT lecturers to understand these objectives better as well as the information needs of these lecturers. In establishing their needs, a solution domain was identified.

Chapter 3 investigated the ways in which organisations and institutions gather information about their business processes. In this regard, log files form a primary source of data and their general use was discussed. The collection of useful information from network traffic and Web usage logs was discussed. The findings of previously published research investigating Web usage logs at NMMU were presented. The systems administrators involved provided sample logs, which were profiled and deemed usable for the analytical techniques used in similar research.

Chapter 4 introduced the dynamics of BI and a brief history of the field. How BI is acquired was discussed and various methods were compared. A BI system development method was then selected, based on the comparisons and the scope and criteria for the desired output of the BI system. The Kimball Lifecycle method was selected and the milestones used to develop the system were outlined.

Chapter 5 presented the SWAN data mart prototype. Each milestone in the Kimball Lifecycle as it was applied to the SWAN project was explained and discussed. The requirements for the prototype were defined. An intended user of the Web usage behaviour information was consulted to provide

input for the design of the prototype. A data structure was then designed through dimensional modelling and refined using three data warehousing experts. This structure is physically represented in an SQL server as a set of tables and sample data is loaded into the server by means of manual extract transformation and loading processes. A report was then produced from this data and presented to the intended user. This individual confirmed that the information would have a positive impact on his/her decision-making about primary educational objectives. The prototype demonstrated that, by using BI system development techniques, valuable information could be extracted from the log files produced by NMMU Web activity.

Chapter 6 presented the analysis of the SWAN project. The development process of the SWAN project and the prototype itself were examined. Documentation, the developer's account of various issues that emerged and their resolution were used as data to design a set of guidelines. These guidelines were based on assertions made from the analysis of the data, using evidential and interpretive argumentation, as well as on findings from previous research that had led to the SWAN project. An expert review, using three individuals with experience in data warehousing and BI, was undertaken in order to validate these guidelines. They were found to satisfy the desired characteristics for a sound research output.

7.2 Satisfying the Research Objectives

7.2.1 Primary Research Objective

Provide a comprehensive set of guidelines to assist higher education providers in the successful analysis of student Web usage data, in support of their decisions about primary educational objectives.

This objective was met through the development and presentation of a set of guidelines to assist the successful analysis of student Web usage data, in support of decision-making about primary educational objectives within the School of ICT at NMMU. These guidelines encompass the fundamentals of a data mart designed for analysing student Web usage data. The details of how this was achieved are documented in this dissertation. This primary objective was achieved by satisfying secondary objectives that collectively created a confluence of contributions to satisfy the primary objective. The secondary objectives and the ways in which they were satisfied during this research are discussed in the following section.

7.2.2 Secondary Research Objectives

Determine what student Web usage data and analytical facilitation thereof is currently available.

This objective was set to ensure that appropriate data was available for prototype development. Without appropriate data and access to that data, an alternate avenue would have had to be investigated. Satisfying this objective was considered a vital aspect of the study and formed part of the early stages of the research. The possible location of the data and its owners were investigated. Once the data had been located it was profiled to determine whether it was appropriate for this study. Previous research, attached as Appendix G, was instrumental in satisfying this objective (Von Schoultz et al., 2013).

The owners of the data provided samples for profiling; based on the literature, this data was deemed appropriate and it was readily available from the system engineers. Section 3.3 and 3.4 explain how this was achieved. The system engineers confirmed that the analysis software installed on the device that produced the Web usage data was suitable for producing network management information, but did not provide other decision support, as discussed in Section 3.3.

Identify questions educators have about student Web usage, in order to better serve their primary educational objectives.

This objective was set to investigate what type of information regarding students' Web usage on campus educators would consider useful. In this context "educators" refers to IT lecturers at the NMMU North Campus. The rationale for this objective was that if the information needs of IT lecturers were understood, the investigation of ways to provide for these needs would be more informed. Moreover, the feasibility and justification of this research would become clearer. A survey was conducted using an online survey form for IT lecturers. The population selection and results are discussed in Section 2.3.2. From these results, one IT lecturer was identified to provide further details in a face-to-face interview regarding information she would find valuable in achieving her primary educational objectives. The results of this interview are discussed in Section 5.3. This objective was satisfied by identifying questions from IT lecturers in the survey results and addressing several specific questions established in the interview mentioned above. These questions were instrumental in achieving subsequent secondary objectives.

Isolate best practices and heuristic aspects for applying Web data analytics in this problem area.

Achieving this objective involved investigating how, where and why Web usage behaviour is determined. More specifically, the focus was on methods used generally and which of these could best be applied in this study. Several methods were found, all used in the field of BI. Section 4.2 discusses the BI field to demonstrate applicability of the selected method to this research. Finally, in Section 4.3, several methods were identified and compared in order to isolate an appropriate one for use in

designing a solution to the research problem. Thus the secondary objective was achieved through the process of selection of an appropriate method.

Design, develop, implement and verify a prototype decision support system based on the above findings.

This objective was set to create a basic working version of a proposed partial solution to the research problem. In achieving this objective, it was demonstrated that the methods used were appropriate and able to provide a solution to the problem. A prototype data mart called the SWAN data mart was developed using the method identified in this secondary objective, as explained in Sections 5.2, 5.3 and 5.4. This prototype, and its development, was used to design a set of guidelines, which satisfies the following secondary objective.

Consolidate a set of guidelines for the analysis of Web usage behaviour of students based on the lessons learned and conclusions derived from the case of the prototype development.

This objective involved the design of an artefact that contributed to the solution of the research problem. Analysing the development of a prototype, as discussed above, a set of guidelines was developed (see Section 6.3). Various data was used with reflective and evidential argumentation as well as assertions drawn from the case, to formulate the lessons learned from the case into guidelines. These guidelines provide recommendations for student Web usage analysis within NMMU and form the artefact for this research as mandated by the design science process.

7.3 Revisiting the Research Problem

There is currently no facility in place to analyse accurately student Web usage data within the NMMU in order to present meaningful student Web usage information to educators to support their primary educational objectives.

An IT artefact in the form of a set of specified guidelines based on research evidence was developed using the design science process. This artefact provides recommendations on how to provide for and conduct the analysis of Web usage data at the NMMU. This could ultimately lead to valuable information which can positively affect the decisions made by educators to achieve their primary educational objectives (see Section 2.3.2). The predominant objectives were concerned with the information processing capacity of students. The artefact contributes significantly to the gathering of information which is of high value to the decision making process. Furthermore, the artefact was designed using the research problem as input for the development. This artefact therefore contributes to the solution of this problem within the defined context and scope.

7.4 DSR Knowledge Contribution Summary

Various designs contributed to the creation of the IT artefact (SWAN guidelines). The design of the dimensional model (see Section 5.4.2) informed the development of the SWAN data mart. The design and development of the SWAN data mart provided various valuable lessons which were encapsulated in the design of the IT artefact. The SWAN data mart was developed according to a well established method for solving the problem of analysing historical data (Kimball Lifecycle). The knowledge contributed by this research is prescriptive in nature as it informs a process. Furthermore, this knowledge was derived from the application of an established solution from another field (BI) to a new problem (NMMU student Web usage data). This knowledge is inherent in the delivered IT artefact. Grego and Hevner's DSR knowledge contribution framework places this contribution into the Exaptation quadrant as described in the bottom right quadrant in figure 7.1 (2013).

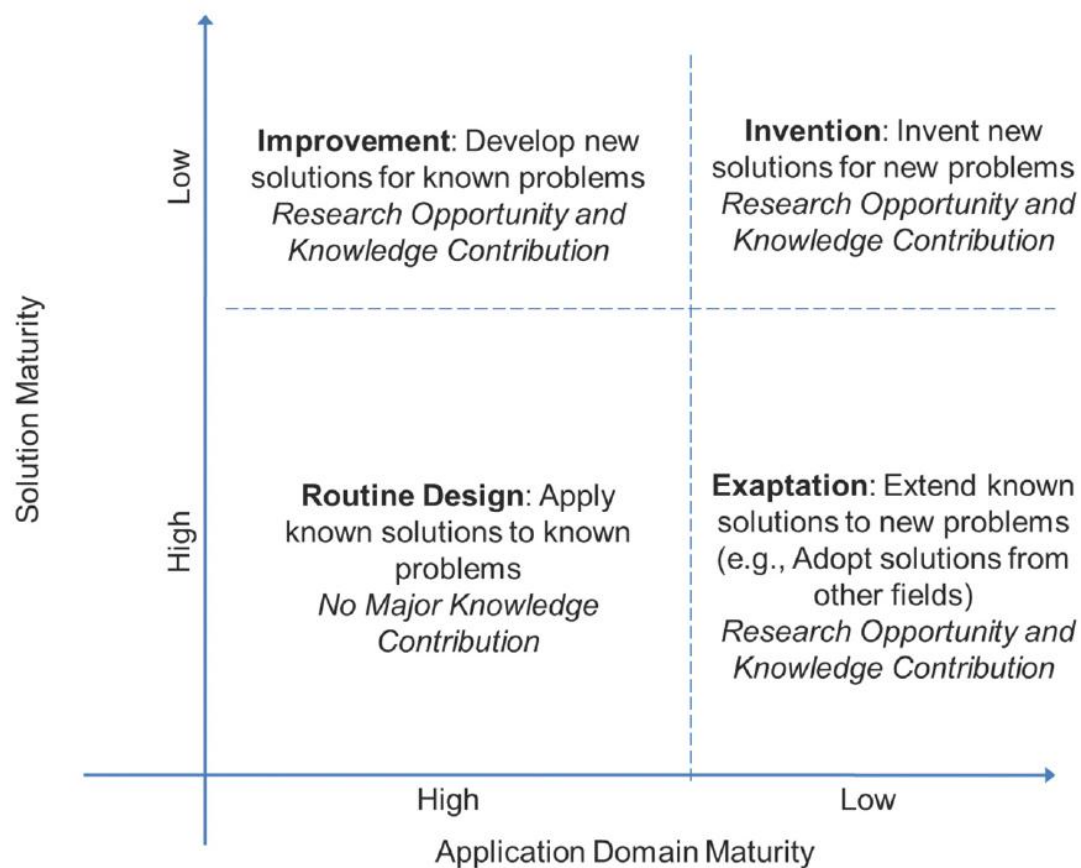


Figure 7.1 DSR knowledge contribution framework (Grego & Hevner, 2013)

Through the design of artefacts intended to solve a problem, knowledge and understanding was gathered and allowed for a significant contribution towards solving the problem defined in this research. It is in this way that DSR, as carried out by the DSRP, was valuable and relevant.

7.5 Limitations

This research presents a prototype data mart that uses Web usage data to present information on Web usage in various levels of detail. This data mart used a single IT lecturer, selected from a number who indicated interest in the information that the prototype could produce. A sub set of data was used to generate a sample report to satisfy this IT lecturer's particular information needs, demonstrating that the method was feasible and establishing a structured foundation for extension.

The primary output and IT artefact produced by this research was a set of guidelines derived from the process used to develop the prototype data mart. For this reason the guidelines are not entirely generalisable. However, they provide recommendations that might be applicable in similar circumstances. Higher education institutions could use these guidelines to analyse Web usage data and they would be particularly appropriate if an institution has a similar network structure and access to data in a similar format. While it is recognised that the contributions made by this study are limited, they could be extended. However, they are applicable to their intended context. The guidelines do not require strict adherence and merely contain recommendations for a specified process.

7.6 Publication Stemming from this Research

Von Schoultz, D., Van Niekerk, J. & Thomson, K.-L. (2013). Web usage mining within a South African university infrastructure, Towards useful information from student Web usage data. *Proceedings of the 14th Annual Conference on World Wide Web Applications Cape Town, 10-13 September 2013* (<http://www.zaw3.co.za>)

7.7 Future Research

A central aspect of this research study was the development of a prototype data mart that would be successful in using raw Web usage data to produce meaningful and valuable information about Web usage behaviour. The DW/BI systems that are made up of data marts are very large and the efforts required to develop these systems are time consuming. For this reason, a prototype was used. A DW/BI system for a single department requires a range of expertise and resources as well as adequate time. Expansion of the prototype is an area for further research. The prototype could be expanded to hold a larger, richer data set and to provide information for other business processes. In addition, the manual extraction, transformation and load processing could be automated to receive updated data from the Fortigate firewall device; this data produces the log files containing the raw Web activity of students at NMMU North Campus. The SWAN data mart could be used to provide Web usage data to all six NMMU campuses in the same way as Web activity at North Campus has been captured by the Fortigate firewall device.

The guidelines could be further refined in order to make them applicable to other contexts that experience the same problems.

7.8 Closing Remark

“Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information on it.” – Samuel Johnson

This research has provided guidance for educators who desire an understanding of how their students are experiencing information through the Web during their studies at a higher education institution.

It could be said that people have always been in an information age; information is all around us. What has clearly changed is how much access individuals have to this information and how easily it can be exchanged. In present day society where Web access is a part of many people’s daily routine, the choice of what information they want falls to them. No longer does everyone receive the same media at the same time through broadcast television and radio.

People are truly fortunate to have the power and the option to access seemingly limitless information on demand. With this comes the responsibility of choosing wisely: which information is valuable and which information is harmful? This is particularly the case when the world is in need of education and knowledge generated through education.

Bibliography

- Abdelkarim, A., & Nasereddin, H. (2011). Intrusion prevention system. *International Journal of Academic Research*, 3(1), 430–432.
- Abraham, A. (2003). Business intelligence from web usage mining. *Journal of Information & Knowledge Management*, 2(04), 375–390.
- Al-Debei, M. (2011). Data Warehouse as a backbone for business intelligence: Issues and challenges. *European Journal of Economics, Finance and Administrative Sciences*, 33, 153–166.
- Al-Shaer, E. S., & Hamed, H. H. (2003). Firewall Policy Advisor for anomaly discovery and rule editing. *IFIP/IEEE Eighth International Symposium on Integrated Network Management* (pp. 17–30). doi:10.1109/INM.2003.1194157
- Alsqour, M., Matouk, K., & Owoc, M. (2012). A survey of data warehouse architectures — Preliminary results. *Computer Science and Information Systems (FedCSIS)* (pp. 1121–1126). IEEE.
- Amador, P. & Amador, J. (2014). Academic advising via Facebook: Examining student help seeking. *The Internet and Higher Education*, 21, 9–16. doi:10.1016/j.iheduc.2013.10.003
- Anand, S., Büchner, A., Mulvenna, M., & Hughes, J. (1999). Discovering Internet marketing intelligence through web log mining. *ACM Sigmod Record*, 27(4), 54–61.
- Asunka, S., Chae, H. S., Hughes, B., & Natriello, G. (2009). Understanding academic information seeking habits through analysis of web server log files: The case of the Teachers College Library website. *The Journal of Academic Librarianship*, 35(1), 33–45. doi:10.1016/j.acalib.2008.10.019
- Balaji, S., & Murugaiyan, M. (2012). Waterfall vs v-model vs Agile: A comparative study on SDLC. *JITBM & ARF*, 2(1), 26–30.
- Bawden, D., & Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, 35(2), 180–191. doi:10.1177/0165551508095781
- Bellinger, G., Castro, D., & Mills, A. (2004). *Data, information, knowledge, and wisdom*. Retrieved from <http://www.systems-thinking.org/dikw/dikw.htm>
- Berners-Lee, T. (2010). Long live the web. *Scientific American*, 303(6), 80–85.
- Bin Hamid Ali, F. A. (2011). Development of host based intrusion detection system for log files. *2011 IEEE Symposium on Business, Engineering and Industrial Applications (ISBEIA)*, 281–285. doi:10.1109/ISBEIA.2011.6088821
- Black, E. W., Dawson, K., & Priem, J. (2008). Data for free: Using LMS activity logs to measure community in online courses. *The Internet and Higher Education*, 11(2), 65–70. doi:10.1016/j.iheduc.2008.03.002
- Blondal, S., Field, S., & Girouard, N. (2002). *Investment in human capital through upper-secondary and tertiary education*. OECD Economic Studies. Paris: OECD.

- Bloom, D., Canning, D., & Chan, K. (2006). *Higher education and economic development in Africa*. Washington, DC: World Bank.
- Boyd, D. (2010). Streams of content, limited attention: the flow of information through social media. *Educause Review*, 45(5), 26.
- Breivik, P. (2005). 21st century learning and information literacy. *Change: The Magazine of Higher Learning*, 37(2), 21–27.
- Breslin, M. (2004). Data warehousing battle of the giants. *Business Intelligence Journal*, 7.
- Bruckner, R., List, B., & Schiefer, J. (2002). *Striving towards near real-time data integration for data warehouses* (pp. 317–326). Springer Berlin Heidelberg.
- Buckland, M. (1991). Information as thing. *Journal of the American Society for Information Science*, 42(5), 351–360.
- Butcher, H. (1998). Meeting managers' information needs. *Aslib*.
- Casey, D. (2008). Turning log files into a security asset. *Network Security*, 2008(2), 4–7.
- Chak, K., & Leung, L. (2004). Shyness and locus of control as predictors of internet addiction and internet use. *CyberPsychology & Behavior*, 7(5), 559–570.
- Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM*, 54(8), 88–98. doi:10.1145/1978542.1978562
- Chou, C., & Hsiao, M. (2000). Internet addiction, usage, gratification, and pleasure experience: the Taiwan college students' case. *Computers & Education*, 35(1), 65–80.
- Cooley, R. (2000). *Web usage mining: discovery and application of interesting patterns from web data*. Minneapolis, MN: University of Minnesota.
- Council, N. Nelson Mandela Metropolitan University Vision 2020 (2010).
- Creswell, J. (2012). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage Publications.
- Davenport, T. H. (2011). Putting the enterprise into the enterprise system. *Harvard Business Review*, 76(4).
- Davenport, T., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston, MA: Harvard Business Press.
- Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology*, 3(1), 1–27. doi:10.1145/643477.643478
- Elbashir, M. Z., Collier, P. A., & Davern, M. J. (2008). Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *International Journal of Accounting Information Systems*, 9(3), 135–153. doi:10.1016/j.accinf.2008.03.001

- Feather, J. (2000). *Information society: A study of continuity and change*. London: Library Association.
- Fortinet. (2013). *FortiGate Log Message Reference v5.0 Patch Release 4*.
- Fried, C. B. (2008). In-class laptop use and its effects on student learning. *Computers & Education*, 50(3), 906–914. doi:10.1016/j.compedu.2006.09.006
- Fuller, R. (1973). *Utopia or oblivion*. New York: Overlook Press.
- Gangadharan, G., & Swami, S. (2004). Business intelligence systems: design and implementation strategies. In *Information Technology Interfaces, 2004. 26th International Conference* (pp. 139–144). IEEE.
- Garrison, D., & Vaughan, N. (2008). *Blended learning in higher education: Framework, principles, and guidelines*. Hoboken, NJ: John Wiley & Sons.
- George, A., Makanju, A., Zincir-Heywood, A. N., & Milios, E. E. (2008). Information retrieval in network administration. *6th Annual Communication Networks and Services Research Conference (CNSR 2008)* (pp. 561–568). IEEE. doi:10.1109/CNSR.2008.78
- Girard, J. (2005). Combating information anxiety: A management responsibility. *Organizacija Vadyba: sisteminiai tyrimai*, 35(1), 65–79.
- Golfarelli, M., & Rizzi, S. (2009). *Data warehouse design: Modern principles and methodologies*. New York: McGraw-Hill.
- Golfarelli, M. (2010). From user requirements to conceptual design in data warehouse design. *Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction, 1*. doi:10.4018/978-1-60566-756-0.ch001
- Gorgone, J., Davis, G., Valacich, J., Topi, H., & Feinstein, D. L. (2003). IS 2002 Model curriculum and guidelines for undergraduate degree programs in information systems. *Communications of the Association for Information Systems*, 11(1), 1.
- Gosain, S. (2004). Enterprise information systems as objects and carriers of institutional forces: the new iron cage? *Journal of the Association for Information Systems*, 5(4), 6.
- Gow, L., & Kember, D. (1990). Does higher education promote independent learning? *Higher Education*, 19(3), 307–322. doi:10.1007/BF00133895
- Grace, L. (2011). Analysis of web logs and web user in web mining. *International Journal of Network Security & Its Applications (IJNSA)*, 3(1), 99–110.
- Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining. Drug safety : an international journal of medical toxicology and drug experience* (Vol. 30). MIT Press. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/17604416>
- Hazelhurst, S., Johnson, Y., & Sanders, I. (2011). An empirical analysis of the relationship between web usage and academic performance in undergraduate students. *arXiv preprint arXiv:1110.6267*.
- HEAA Council Higher Education Act 101 of 1997 (2001).

- Healey, N. M. (2007). Is higher education in really “internationalising”? *Higher Education*, 55(3), 333–355. doi:10.1007/s10734-007-9058-4
- Hevner, A., & Chatterjee, S. (2010). *Design Research in Information Systems*. New York: Springer.
- Hevner, A., & Grego, S. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337-355.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). *Design Science in Information Systems Research*, 28(1), 75–105.
- Heylighen, F. (2002). Complexity and information overload in society: why increasing efficiency leads to decreasing control. *The Information Society*.
- Hirji, K. (2001). Exploring data mining implementation. *Communications of the ACM*, 44(7), 87–93.
- Hochschild, J. (2009). Conducting intensive interviews and elite interviews. *Workshop on Interdisciplinary Standards for Systematic Qualitative Research*. National Science Foundation.
- Hofgesang, P. I. (2009). *Modelling web usage in a changing environment*. Dutch Graduate School for Information and Knowledge Systems.
- Hooft, M. V., & Swan, K. (2007). *Ubiquitous computing in education: Invisible technology, visible impact*. Mahwah, NJ: Lawrence erlbaum associates.
- Hossain, S., Rahman, S., & Kabir, M. (2012). Network proxy log mining: association rule based security and performance enhancement for proxy server. *Computer Science and Engineering*.
- Hu, J., & Zhong, N. (2005). Clickstream log acquisition with web farming. *The 2005 IEEE/WIC/ACM International Conference* (pp. 257–263). IEEE.
- Ierace, N., Urrutia, C., & Bassett, R. (2005). Intrusion prevention systems. *Ubiquity*, 2005(June), 2–2. doi:10.1145/1071916.1071927
- Imhoff, C., Galembo, N., & Geiger, J. G. (2003). *Mastering Data Warehouse Design: relational and dimensional techniques*. Hoboken, NJ: John Wiley & Sons.
- Inmon, W. (2002). *Building the data warehouse* (3rd ed.). Hoboken, NJ: John Wiley & Sons.
- ISO/IEC 27002. (2005). *Information technology — Security techniques — Code of practice for information security management* (Vol. 2005). ISO.
- Itoh, T., & Takakura, H. (2006). Hierarchical visualization of network intrusion detection data. *Computer Graphics and Applications, IEEE*, 26(2), 40–47.
- Jiang, W., Hu, C., Pasupathy, S., Kanevsky, A., Li, Z., & Zhou, Y. (2009). Understanding customer problem troubleshooting from storage system logs. *FAST*, 9, 43–56.
- Keen, P., & Morton, M. (1978). *Decision support systems: An organizational perspective*. Reading, MA: Addison-Wesley.
- Kelley, T. (2001). Prototyping is the shorthand of innovation. *Design Management Journal*, 12(3), 35–42.

- Kimball, R., & Merz, R. (2000). *The data webhouse toolkit: Building the web-enabled data warehouse*. New York: John Wiley & Sons.
- Kimball, R., Ross, M., Thorthwaite, W., Becker, B., & Mundy, J. (2008). *The data warehouse lifecycle toolkit* (2nd ed.). Hoboken, NJ: John Wiley & Sons.
- Kimball, R., & Ross, M. (2002). *The data warehouse toolkit: the complete guide to dimensional modelling*. US: John Wiley & Sons.
- Kirschner, P., & Karpinski, A. (2010). Facebook® and academic performance. *Computers in human behavior*, 26(6), 1237–1245.
- Ko, C. H., Yen, J. Y., Yen, C. F., Chen, C. S., & Chen, C. C. (2012). The association between Internet addiction and psychiatric disorder: a review of the literature. *European Psychiatry : The Journal of the Association of European Psychiatrists*, 27(1), 1–8. doi:10.1016/j.eurpsy.2010.04.011
- Kohavi, R. (2001). Mining e-commerce data: the good, the bad, and the ugly. In ACM (Ed.), *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 8–13).
- Kowalski, K., & Beheshti, M. (2006). Analysis of log files intersections for security enhancement. *Third International Conference on Information Technology: New Generations (ITNG'06)* (pp. 452–457). IEEE. doi:10.1109/ITNG.2006.32
- Kowalski, Kazimierz, & Beheshti, M. (2008). Improving security through analysis of log files intersections. *IJ Network Security*, 7(1), 24–30.
- Kubey, R., Lavin, M., & Barrows, J. (2001). Internet use and collegiate academic performance decrements: Early findings. *Journal of Communication*, 51(2), 366–382.
- Kukulska-Hulme, A. (2012). How should the higher education workforce adapt to advancements in technology for teaching and learning? *The Internet and Higher Education*, 15(4), 247–254. doi:10.1016/j.iheduc.2011.12.002
- Kute, S. S., & Thorat, S. (2014). A review on various software development life cycle (SDLC) models. *IJRCCT*, 3(7), 776–781.
- Larson, B. (2006). *Delivering business intelligence with Microsoft SQL Server 2008*. New York: Osborne/McGraw-Hill.
- Leau, Y., Loo, W., Tham, W., & Tan, S. (2012). Software development life cycle AGILE vs traditional approaches. *International Conference on Information and Network Technology*, 37, 162–167).
- Lee, M.H., & Tsai, C.-C. (2008). Exploring teachers' perceived self efficacy and technological pedagogical content knowledge with respect to educational use of the World Wide Web. *Instructional Science*, 38(1), 1–21. doi:10.1007/s11251-008-9075-4
- List, B., & Bruckner, R. (2002). A comparison of data warehouse development methodologies case study of the process warehouse. *Database and Expert Systems Applications* (pp. 203–215). Springer Berlin Heidelberg.

- Liu, a. X., & Gouda, M. G. (2004). Diverse firewall design. *Parallel and Distributed Systems, IEEE Transactions*, 19(9), 1237–1251. doi:10.1109/DSN.2004.1311930
- Liu, a. X., & Gouda, M. G. (2009). Firewall policy queries. *Parallel and Distributed Systems, IEEE Transactions*, 20(6), 766–777. doi:10.1109/TPDS.2008.263
- Liu, a. X., Torng, E., & Meiners, C. R. (2008). Firewall compressor: An algorithm for minimizing firewall policies. *2008 IEEE INFOCOM – The 27th Conference on Computer Communications* (Vol. 1). IEEE. doi:10.1109/INFOCOM.2008.44
- Luotonen, A., & Altis, K. (1994). World-wide web proxies. *Computer Networks and ISDN Systems*, 27(2), 147–154.
- March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, 15(4), 251–266. doi:10.1016/0167-9236(94)00041-2
- Mason, J. (2002). Qualitative researching. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:qualitative+researching#0>
- Mayer, A., Wool, A., & Ziskind, E. (2000). Fang: A firewall analysis engine. *Security and Privacy, 2000. S&P ...*, 177–187. doi:10.1109/SECPRI.2000.848455
- Mayo, K., & Newcomb, P. (2008). How the Web was won. *Vanity Fair*. Retrieved from <http://www.vanityfair.com/culture/features/2008/07/internet200807>
- Meadow, C., & Yuan, W. (1997). Measuring the impact of information: Defining the concepts. *Information Processing & Management*, 33(6), 697–714.
- Mehrtens, J., Cragg, P. B., & Mills, A. M. (2001). A model of Internet adoption by SMEs, 39.
- Mell, P., & Grance, T. (n.d.). *The NIST definition of cloud computing recommendations of the National Institute of Standards and Technology*. Gaithersburg, MD: NIST.
- Mendling, J., Reijers, H., & Aalst, W. van der. (2010). Seven process modeling guidelines (7PMG). *Information and Software ...*, 2008. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0950584909001268>
- Morgan, D. L. (1997). *Focus groups as qualitative research*. Thousand Oaks, CA: Sage Publications. doi:10.4135/9781412984287
- Moss, L., & Atre, S. (2003). *Business intelligence roadmap: The complete project lifecycle for decision-support applications*. Reading: Addison-Wesley Professional.
- Mowery, D. C., & Simcoe, T. (2002). Is the Internet a US invention? An economic and technological history of computer networking. *Research Policy*, 31(8), 1369–1387. doi:10.1016/S0048-7333(02)00069-0
- Negash, S. (2004). Business intelligence. *The Communications of the Association for Information Systems*, 13(1), 54.
- Nemati, H., Steiger, D., Iyer, L., & Herschel, R. (2002). Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems*, 33(2), 143–161. doi:10.1016/S0167-9236(01)00141-5

- Ng, B. D., & Wiemer-Hastings, P. (2005). Addiction to the internet and online gaming. *Cyberpsychology & Behavior : the Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, 8(2), 110–113. doi:10.1089/cpb.2005.8.110
- Njenga, J. K., & Fourie, L. C. H. (2010). The myths about e-learning in higher education. *British Journal of Educational Technology*, 41(2), 199–212. doi:10.1111/j.1467-8535.2008.00910.x
- Obeidat, J., & Nasereddin, H. H. (2013). A new vision for information technology project management through selecting SDLC model. *American Academic & Scholarly Research Journal*, 5(4), 183–192.
- OECD. (1998). *Technology productivity and job creation best policy*. Paris: OECD.
- Okoli, C., & Pawlowski, S. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, 42(1), 15–29.
- Oliner, A., Ganapathi, A., & Xu, W. (2012). Advances and challenges in log analysis. *Communications of the ACM*, 55(2), 55–61.
- Olszak, C., & Ziemba, E. (2003). Business Intelligence as a key to management of an enterprise. *Informing Science Institute, Informing Science and Information Technology Education*. Pori, Finland.
- Oxman, A. (2004). Grading quality of evidence and strength of recommendations. *BMJ*, 328(June). Retrieved from <http://www.bmj.com/content/328/7454/1490.abridgement.pdf>
- Panel, N. (1998). *Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults*. Retrieved from <http://www.ncbi.nlm.nih.gov/books/NBK2003>
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London: Penguin.
- Patel, A., Qassim, Q., & Wills, C. (2010). A survey of intrusion detection and prevention systems. *Information Management & Computer Security*, 18(4), 277–290. doi:10.1108/09685221011079199
- Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Bragge, J., & Virtanen, V. (2006). The Design Science research process: A model for producing and presenting information system research. *Proceedings of the first international conference on design science research in information systems and technology (DESRIST 2006)*, 83–106.
- Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. doi:10.2753/MIS0742-1222240302
- Peng, W., & Li, T. (2005). Mining logs files for computing system management. *Second International Conference on Autonomic Computing (ICAC'05)* (pp. 309–310). IEEE. doi:10.1109/ICAC.2005.40
- Peters, R. (2010). *The concept of education*. New York, NY: Routledge.
- Ponniah, P. (2004). *Data warehousing fundamentals: a comprehensive guide for IT professionals* (Vol. 6). Hoboken, NJ: John Wiley & Sons.

- Power, D. (2007). A brief history of decision support systems. *COM, World Wide Web*, <http://DSSResources.COM/history/dsshitory.html>, version, 4.
- Purao, S. (2002). Design research in the technology of information systems: truth or dare. GSU department of CIS working paper. Atlanta: Georgia State University.
- Ragunath, P. (2010). Evolving a new model (SDLC Model-2010) for software development life cycle (SDLC). *International Journal of Computer Science and Network Security*, 10(1), 112–119.
- Rainardi, V. (2008). *Building a data warehouse: with examples in SQL Server*. Berkeley, CA: Apress.
- Reddy, G., & Srinivasu, R. (2010). Data warehousing, data mining, OLAP and OLTP Technologies are essential elements to support decision-making process in industries. *International Journal on Computer Science & Engineering*, 02(09), 2865–2873.
- Roberts, L. V. (2005). Information Overload. Capstone.
- Rosen, L. (2010). Welcome to the iGeneration! *Education Digest: Essential Readings Condensed for Quick Review*, 75(8), 8–12.
- Rumbough, T. (2001). Controversial uses of the Internet by college students. *Educause Quarterly*, 24(4), 70–71.
- Sabahi, F., & Movaghar, A. (2008). Intrusion detection: a survey. *Systems and Networks Communications, 2008. ICSNC'08. 3rd International Conference* (pp. 23–26). IEEE. doi:10.1109/ICSNC.2008.44
- Salmi, J. (2003). Constructing knowledge societies: new challenges for tertiary education. *Higher Education in Europe*.
- Saroop, S., & Kumar, M. (2011). Comparison of Data warehouse design approaches from user requirement to conceptual model: a survey. *2011 International Conference on Communication Systems and Network Technologies* (pp. 308–312). IEEE. doi:10.1109/CSNT.2011.161
- Schafer, J., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. *Proceedings of the 1st ACM conference on Electronic commerce*, 158–166.
- Scheps, S. (2013). *Business intelligence for dummies: Communications of the Association for Information* Hoboken, NJ: John Wiley & Sons.
- Sen, A., & Sinha, A. (2005). A comparison of data warehousing methodologies. *Communications of the ACM*, 48(3), 79–84.
- Sen, A., & Sinha, A. (2007). Toward developing data warehousing process standards: An ontology-based review of existing methodologies. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions*, 37(1), 17–31.
- Seo, M., Kang, H. S., & Yom, Y.-H. (2009). Internet addiction and interpersonal problems in Korean adolescents. *Computers, informatics, Nursing : CIN*, 27(4), 226–33. doi:10.1097/NCN.0b013e3181a91b3f
- Shenk, D. (2009, January). Data smog: Surviving the information glut. *Competitive Intelligence Review*. Harper Collins. doi:10.1002/(SICI)1520-6386(199724)8:4<89::AID-CIR19>3.3.CO;2-#

- Simitsis, A., Vassiliadis, P., & Sellis, T. (2005). Optimizing ETL processes in data warehouses. *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference* (pp. 564–575). IEEE.
- Simon, A. (2009). *Data warehousing for dummies* (2nd ed.). John Wiley & Sons.
- Sloan, J. (2001). *Network troubleshooting tools*. O'Reilly Media Inc.
- South African Schools Act No. 84 of 1996, *Government Gazette*, No. 34620 (2011).
- Speier, C., Valacich, J. S., & Vessey, I. (1999). The influence of task interruption on individual decision making: an information overload perspective. *Decision Sciences*, 30(2), 337–360. doi:10.1111/j.1540-5915.1999.tb01613.x
- Stake, R. (1995). The art of case study research. Thousand Oaks, CA: Sage Publications.
- Stepanikova, I., Nie, N. H., & He, X. (2010). Time on the Internet at home, loneliness, and life satisfaction: Evidence from panel time-diary data. *Computers in Human Behavior*, 26(3), 329–338. doi:10.1016/j.chb.2009.11.002
- Straub, D., & Gefen, D. (2004). Validation guidelines for IS Positivist, *The Communications of the Association for Information Systems* 13(1), 380–427.
- Tang, J., Yu, Y., Du, Y., Ma, Y., Zhang, D., & Wang, J. (2014). Prevalence of internet addiction and its association with stressful life events and psychological symptoms among adolescent internet users. *Addictive behaviors*, 39(3), 744–747. doi:10.1016/j.addbeh.2013.12.010
- Tao, R., Huang, X., Wang, J., Zhang, H., Zhang, Y., & Li, M. (2010). Proposed diagnostic criteria for internet addiction. *Addiction*, 105(3), 556–564. doi:10.1111/j.1360-0443.2009.02828.x
- Teferra, D., & Altbachl, P. (2004). African higher education: Challenges for the 21st century. *Higher Education*, 47(1), 21–50.
- Thomason, S. (2012). Improving network security: next generation firewalls and advanced packet inspection devices. *Global Journal of Computer Science and Technology*, 12(13).
- Thomsen, E. (2003). BI's promised land. *Intelligent Enterprise-San Mateo*, 6, 20–25.
- Torero, M., & Braun, J. Von. (2006). *Information and communication technologies for development and poverty reduction: The potential of telecommunications*. Intl Food Policy Res Inst. Retrieved from <http://books.google.com/books?hl=en&lr=&id=8b46AwAAQBAJ&oi=fnd&pg=PA1&dq=Information+and+communication+technologies+for+development+and+poverty+reduction:+The+potential+of+telecommunications&ots=sG0svz8yCQ&sig=A3CnBojarVqGKJqH-nnCF-TJPmg>
- Vaishnavi, V., & Kuechler, W. (2012). A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information Systems*, 13(6), 395–423.
- Von Schoultz, D., Van Niekerk, J. & Thomson, K.-L. (2013). Web usage mining within a South African university infrastructure, Towards useful information from student Web usage data.

- Wallace, P. (2014). Internet addiction disorder and youth. *EMBO reports*, 15(1), 12–16.
- Wang, G. (2002). Definition and review of virtual prototyping. *Journal of Computing and ...*, 1–14.
Retrieved from
<http://computingengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1399487>
- Watson, H., & Wixom, B. (2007). The current state of business intelligence. *Computer*, 40(9), 96–99.
- Weinstein, A., & Lejoyeux, M. (2010). Internet addiction or excessive internet use. *The American Journal of Drug and Alcohol Abuse*, 36(5), 277–283.
- Wellman, B., & Haythornthwaite, C. (2008). *The Internet in everyday life*. Hoboken, NJ: John Wiley & Sons.
- Wessels, D. (2001). Squid frequently asked questions. *Configuration Issues In:*
Retrieved from
<http://www.squid-cache.org/Doc/FAQ/FAQ-4.html>, Informationsabruf am, 20.
- Whang, L. S.-M., Lee, S., & Chang, G. (2003). Internet over-users' psychological profiles: a behavior sampling analysis on internet addiction. *CyberPsychology & Behavior*, 6(2), 143–50.
doi:10.1089/109493103321640338
- Widyanto, L., & McMurran, M. (2004). The psychometric properties of the internet addiction test. *CyberPsychology & Behavior*, 7(4), 443–50. doi:10.1089/cpb.2004.7.443
- Wixom, B., & Watson, H. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS quarterly*, 25(1), 17–41.
- Wurman, R. (1989). Information anxiety.
- Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.
- Yin, R. K. (2009). *Case study research: design and methods* (4th ed.). Thousand Oaks, CA: Sage Publications.
- Young, K. (1998). Internet addiction: The emergence of a new clinical disorder. *CyberPsychology & Behavior*, 1(3), 237–244.
- Zalenski, R. (2002). Firewall technologies. *Potentials, IEEE*, 21(1), 24–29.

Appendix A

Email liaison between associate professor and Director of ICT

To: Director of ICT

Cc: Associate Professor; Student Researcher

Subject: Permission to analyse clickstream data from firewall device

Dear Sir

As you are aware one of my students are working on identifying patterns in firewall data that might serve as early warning about academic problems for a student. **The data he'll be using will be anonymized before he analyses it.** However, in order to adhere to ethics guidelines we need some form of written permission from you to indicate that this has been cleared via the correct channels. Can you please provide this? I think an email reply to this message to include as an addendum to his research should suffice?

Kind regards

Associate Professor

School of Information and Communication Technology (ICT)

Summerstrand North Campus

Reply

To: Associate Professor

Cc: Associate professor; Student Researcher and System Engineers.

Subject: Permission to analyse clickstream data from firewall device

Dear Associate Professor,

The proposed study can provide valuable insights in usage patterns associated with academic performance of students, and we welcome this initiative. ICT Services will gladly provide access to the log data available. We do currently not have good anonymising tools available, and part of the work done by the researcher will have to include suitable tools which can be used by the ICT staff to anonymise info before providing the data. I think it will also be reasonable to request the researcher

to draft an undertaking that he will not use the information to attempt to identify individuals, or publish information that can be embarrassing to the institution – some form of review and approval that you can specify before findings are published.

I trust that this is in order?

Kind regards,

Director of ICT

Appendix B

Online survey form

Currently, research is being undertaken to analyse student Web usage data at the NMMU. This is being done in an attempt to gather meaningful information from this data to provide lecturers and educational decision-makers with relevant information about student Web use trends on campus. This could allow for more informed decision-making regarding student Web usage.

The purpose of this survey form is to extract lecturers' primary educational objectives and establish their views on desired Web usage behaviours of their students.

All responses will remain anonymous and will be represented as such. It is recognised that the desired Web usage behaviour of students may vary depending on the characteristics of their subject/s.

Therefore, if you lecture multiple subjects and the desired Web usage behaviour of students for those subjects differs, please complete one survey form for each subject. If, however, you feel your subjects are closely related and require the same Web usage behaviour from students, then please indicate this by including those subjects together in the subject code list below.

For example, if you teach three subjects and you feel that the demands of two of these are similar enough to warrant the same desired Web usage behaviour from students, then complete one survey form for both subjects and indicate as such below. Then for the third subject complete another survey form, as the desired Web usage behaviour differs. If all three subjects are closely related and warrant the same desired Web usage behaviour, then complete one survey form for all three and include all three subject codes below.

Please indicate the subject code/s for which you are a lecturer.

Question 1

Indicate your primary educational objective/s as a lecturer from the list below. Add more options in option "e" if necessary.

- a) Throughput (maximising student pass rates)
- b) Skill distribution
- c) Knowledge dissemination
- d) Developing critically thinking graduates
- e) Other, please specify _____

Additional comments

Question 2

a) Does your subject/s require your students to have up-to-date and recent information in that subject area?

- 1) Yes
- 2) No

Additional comments

b) Is the material provided to your students sufficient for the subject? In other words, they do require additional material for the subject, besides what you give them?

- 1) Yes

2) No

Additional comments

Question 3

Using the scale provided, indicate as instructed below.

The scale values have the following meanings:

Strongly Disagree means that you do not agree with the statement at all and know that the statement is entirely incorrect.

Somewhat Disagree means that you do not agree with the statement but believe it is not entirely untrue in certain aspects or scenarios.

Disagree means that you do not agree with the statement.

Neither Agree Nor Disagree means that you cannot agree or disagree with the statement.

Somewhat Agree means that you do agree with the statement but believe it is not entirely true in certain aspects or scenarios.

Agree means that you do agree with the statement.

Strongly Agree means that you completely agree with the statement and know that the statement is entirely correct.

For each of the following, indicate how much you agree or disagree with the statement. Please qualify your choice in the space provided by providing a reason or further explanation for your choice.

a) *Students' Web usage behaviour significantly affects your primary educational objectives.*

1) Strongly Disagree

2) Somewhat Disagree

- 3) Disagree
- 4) Neither Agree Nor Disagree
- 5) Somewhat Agree
- 6) Agree
- 7) Strongly agree

Choice Qualification

Additional comments

b) Having information regarding the Web usage behaviour of your students would benefit your decision-making in achieving your primary educational objectives.

- 1) Strongly Disagree
- 2) Somewhat Disagree
- 3) Disagree
- 4) Neither Agree Nor Disagree
- 5) Somewhat Agree
- 6) Agree
- 7) Strongly agree

Choice Qualification

Additional comments

Question 4

Using the scale provided, indicate as instructed below.

The scale values have the following meanings:

No Influence means that the information would not result in any change to your decision-making.

Minor Influence means that the information would result in a small change in your decision-making.

Some Influence means that the information would result in a clear change in your decision-making.

Strong Influence means that the information would result in a major change in your decision-making.

For each of the following assertions regarding student Web usage behaviour, indicate what level of influence the information would have on your decision-making as a lecturer.

- a) The length of time an individual student spends browsing the Web on campus per day. For example, student A spends an average X minutes on the Web on campus per day.

- 1) No Influence
- 2) Minor Influence
- 3) Some Influence
- 4) Strong Influence

Additional comments

- b) The length of time an individual student spends browsing subject relevant websites during a given practical class. For example, student A spends X minutes on the subject relevant website C during practical class B.

- 1) No Influence
- 2) Minor Influence
- 3) Some Influence
- 4) Strong Influence

Additional comments

- c) The length of time an individual student spends browsing websites unrelated and/or irrelevant to the subject during a given practical class. For example, student A spends X minutes on website C, which is unrelated and/or irrelevant to the subject during practical class B.

- 1) No Influence
- 2) Minor Influence
- 3) Some Influence
- 4) Strong Influence

Additional comments

d) The website or Web service used most frequently by a group of students. For example, students in subject B spend most of their online time on website C.

- 1) No Influence
- 2) Minor Influence
- 3) Some Influence
- 4) Strong Influence

Additional comments

Indicate any further question/s you would include in this question set.

Question 5

Using the scale provided, indicate as instructed below.

The scale values have the following meanings:

Not at all aware means that you have never heard of this problem before.

Slightly aware means that you have heard of the problem but have no understanding of it.

Somewhat aware means that you have heard of the problem and have some understanding of what it is.

Moderately aware means that you are aware of the problem and do understand it.

Extremely aware means that you are very aware of the problem and have a detailed understanding of it.

For each of the following problems associated with excessive Web use, indicate your level of awareness.

a) Internet addiction

- 1) Not at all aware
- 2) Slightly aware
- 3) Somewhat aware
- 4) Moderately aware
- 5) Extremely aware

b) Information overload

- 1) Not at all aware
- 2) Slightly aware
- 3) Somewhat aware
- 4) Moderately aware
- 5) Extremely aware

Additional comments

Appendix C

System Engineer Interview

Interviewer: Mr Dean von Schoultz

Interviewee: North Campus System Engineer

Purpose of the interview

The purpose of this interview is to formalise a casual liaison and to confirm uncertainties that emerged during the study.

Question 1:

What network log analysis software is installed on the Fortigate firewall? Sawmill?

Interviewee's answer:

"The Fortigate Device itself comes with FortiAnalyser software, which is a fairly expensive product. What we are using instead is we are using Sawmill. The sheer volume of logs takes a good couple of hours to actually process a day's worth of logs."

Interviewer's comment:

"Is that just for network statistics, you can't get that much information from it?"

Interviewee's response:

"It will give us like site names visited, who did what, that sort of stuff on the network so statistical stuff no it doesn't give us performance or session dwell and session times, in that sense."

Interviewer's comment:

"So, what do you use it for?"

Interviewee's response:

"It is comprehensive as far as network management is concerned"

Interviewer's comment:

"Not so for business intelligence?"

Interviewee's response:

"No it does not provide business intelligence functionality."

Question 2:

Did the initial request for the log samples from the Fortigate device result in the higher priority of the configuration error?

Interviewee's answer:

"Well, it was a pending issue. I would say that because you approached us the issue did become slightly more prioritised and stopped it from being pushed to the back burner because it was then not just an internal issue."

Question 3:

Does the Fortigate firewall device serve as a proxy server?

Interviewee's Answer:

"No, it is just a straight firewall, it can be configured as such but we don't use it as that."

Interviewer's comment:

"In terms of the proxy model, does it serve as a proxy server? Or is a proxy server very specialized?"

Interviewee's response:

"The firewall is more of a gatekeeper in the same way that a proxy server is in the proxy model. A proxy server will fetch Web material on your behalf. It breaks SSL POP point. A proxy server is more useful in a high bandwidth capacity. The firewall device is looking at and filtering the traffic in the same way as a proxy server, so in terms of the proxy model it serves the general role that a proxy server does."

Question 4:

The Director of ICT states in an email regarding permission to use the Web usage data that the logs must be anonymised before they can be analysed. To my knowledge, all the samples you sent me were anonymised. Was the director not aware of the anonymisation or is there more that needs to be done?

Interviewee answer:

“Oh, no, Steve had not yet been made aware at that time that the logs were completely anonymised. He has since been updated about this. We just had not yet pushed that information up to management back then.”

Appendix D

SWAN Project Scope Charter

Brief Background, Objectives and Approach

Nelson Mandela Metropolitan University (NMMU) North Campus IT Software Development lecturers would like to better understand the Web usage behaviours of their students. Certain patterns in students' Web usage may have a detrimental effect on the educational value to be gained by students from the institution. In addition, lecturers seeking to make more effective decisions about their primary educational objectives could benefit from information regarding the Web usage habits of their students. The SWAN project seeks to provide valuable input for research to assist in solving the problem of delivering Web usage data to IT lecturers in a usable format.

Project Focus

SWAN focuses on answering questions posed by NMMU IT lecturers about their students' Web usage on North Campus and on providing information that could be useful to lecturers' decision-making.

Anticipated Data

Partly filtered network traffic logs from the Fortigate firewall device, amounting to an estimated 900 000 entries (lines) in the log for a single day of North Campus network traffic. Log entries are likely to have inconsistent fields based on the entry type.

Target Users

The target users of reports generated by SWAN will be NMMU IT lecturers.

Involved Parties

Mr Bruce Smith – System Engineer. He will provide log files and will do initial filtering and anonymising during extraction of the log files from the Fortigate device.

Mr Dean von Schoultz – Student Researcher. Primary developer of SWAN

Stakeholders

- Professor Kerry-Lynn Thomson – Associate Professor in the School of ICT
- Professor Johan van Niekerk – Associate Professor in the School of ICT

- Director of Information and Communication Technology

Success Criteria

- A single question or a small selection of questions posed by IT lecturers regarding student Web usage behaviours will be addressed. This could enhance the lecturers' decision-making capacity in their primary educational objectives. This will be used as proof of the concept for a part of the research.

Assumptions and Risks

The following are considered possible risks:

- It may be difficult to structure certain important data fields such that they are usable.
- Availability constraints of relevant participants may result in problematic delays.
- Time constraints may cause further unforeseen limitations.

Exclusions from Scope

- No official BI application or front end application will be made available to business users.
- Data sources will not include staff Web usage data.
- Data sources may include data from other campuses. However, this will not be considered for analysis.
- Reports will be produced only to provide proof of concept for research purposes.

Appendix E

Guideline Validation Review Form

Preamble and brief background to previous research into Guidelines for the Analysis of Student Web Usage in Support of Primary Education Objectives

The paragraph below describes the process undertaken to develop a set of guidelines for the analysis of student Web usage data in support of primary educational objectives at the NMMU. This provides some context for the guidelines and allows for a better understanding of their origin and development.

An initial inquiry into how the Web was affecting students' academic well-being gave rise to the notion of analysing network traffic to derive the Web usage behaviour of students. This led to the collection and exploration of network traffic data from NMMU. It was found that the system engineers at the North campus administered part of the network for all six NMMU campuses. They revealed that the NMMU had adopted a proxy model to ensure secure Web access. The Fortigate firewall device, which handles incoming and outgoing Web traffic and activity, is configured and administered by these system engineers. They confirmed that all Web traffic to and from the NMMU computer laboratories must pass through the Fortigate firewall device that serves as a proxy server/gatekeeper in the proxy model. Moreover, the device logs Web traffic and is therefore a data source for Web usage within the NMMU. The system engineers provided a sample set of Web traffic log files that were profiled to determine the format. The format was compared to other research that had analysed Web traffic log files in an educational context and found to be ill suited as the URL and hostname data fields were not logged. This was the result of a logging configuration error on the firewall device. The system engineers later corrected this error and new samples were profiled and found to be in richer detail than those used in previously successful educational Web usage analysis research. Therefore, the new sample logs were deemed appropriate for analysis.

Methods for deriving useful information from these logs were examined to determine an appropriate domain and knowledge base. The data warehousing and BI field were found to be appropriate and various methods used in this field were compared. The Kimball Lifecycle method was judged to be most suitable for developing a prototype data mart that would satisfy the criteria of the study.

Potential users of Web usage information were identified from a survey. An interview was conducted with a potential user to determine the exact type of Web usage information about her students that would be valuable for her decision-making in achieving primary educational objectives. Using these information needs, a star schema was formalised and created on an SQL server database. A day's worth of student Web usage data was retrieved from the system engineers. The data was then

manually cleaned using Excel and exported to the star schema in the SQL server, as indicated in Figure 1. This formed the SWAN (Student Web usage ANalysis) prototype data mart.

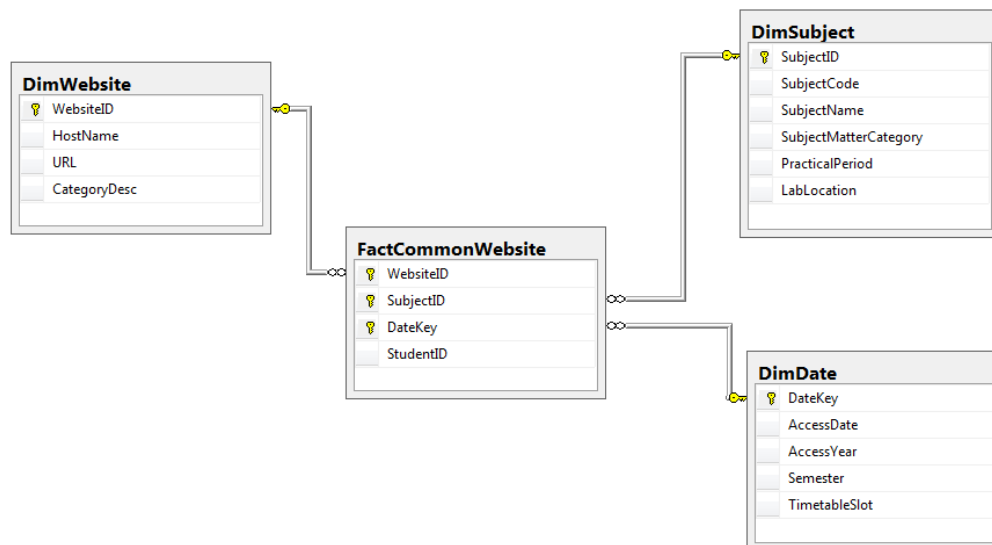


Figure 1 SWAN Data Mart Table Structure in SQL Server

Additional data was exported to the SWAN data mart, including subject, date and timetable data. Based on the user's information needs, a sample report was created to confirm the perceived value of the information and to ensure that the information produced fulfilled the user's expectations. The query used to construct the report used IP Addresses provided by the system engineers to isolate a given laboratory. Time and date values were used to isolate a practical slot, amongst other parameters, to refine the report. Figure 2 shows the sample report that was demonstrated to the user during an interview. The user confirmed that the information would be valuable.

AccessYear	AccessDate	Semester	SubjectCode	HostName	CategoryDesc	Lablocation	PracticalPeriod
4	2014	2014-08-15	2	ONT3660	www9.addfreestats.com	Information Technology	R128 4
5	2014	2014-08-15	2	ONT3660	www3.addfreestats.com	Information Technology	R128 4
6	2014	2014-08-15	2	ONT3660	www.w3schools.com	Information Technology	R128 4
7	2014	2014-08-15	2	ONT3660	www.visual-paradigm.com	Information Technology	R128 4
8	2014	2014-08-15	2	ONT3660	www.updateyourbrowser.net	Information Technology	R128 4
9	2014	2014-08-15	2	ONT3660	www.trialpay.com	Information Technology	R128 4
10	2014	2014-08-15	2	ONT3660	www.techhive.com	Information Technology	R128 4
11	2014	2014-08-15	2	ONT3660	www.teamviewer.com	Information Technology	R128 4
12	2014	2014-08-15	2	ONT3660	www.subnettingquestions.com	Information Technology	R128 4
13	2014	2014-08-15	2	ONT3660	www.statcounter.com	Information Technology	R128 4
14	2014	2014-08-15	2	ONT3660	www.sourcecodeonline.com	Information Technology	R128 4
15	2014	2014-08-15	2	ONT3660	www.semsim.com	Information Technology	R128 4

Figure 2 SWAN Data Mart Sample Report

A case study of the development of the SWAN data mart was undertaken to construct the guidelines. Refer to Appendix A for additional information regarding the construction of each guideline during the case study, as indicated in each of the guidelines.

The purpose of this document is to determine the perceived validity of guidelines produced from a research study that sought to contribute to the analysis of student Web usage data by providing lecturers with Web usage information to enhance their decisions in meeting their primary educational objectives.

These guidelines contain recommendations and considerations, but it is not required or mandated that they be followed. They are intended to be holistic, high-level and provide direction through advice and suggestions.

The evidence used to construct these guidelines was derived predominantly from a case of the development of the SWAN prototype data mart, used for the analysis of student Web usage data from NMMU's North campus. Each guideline used various data sources from the case as specified below.

Guidelines' context of use:

The proposed guidelines are designed for users within the NMMU who intend to gain useful information from Web usage data gathered from the institution.

Or

Users within a higher education institution with a similar network infrastructure who intend to gain useful information from Web usage data gathered from this institution.

Please rate each of the guidelines that follow based on the criteria and metrics provided. Include any suggestions and comments in the space provided.

Guideline 1

The data and the owner/s thereof should be investigated to gather and profile a sample of the required data before the DW/BI project should be considered feasible.

Recommendations

- Investigate which constituents administer the network infrastructure of the institution
- Establish which constituent/s of the institution is/are accountable for the control of the data
- Enquire about the network infrastructure to determine how Web access is provided to the institution
- Identify an appropriate device which logs, or could be configured to log, Web usage data.
- Acquire a sample of the data from the device
- Profile the sample to determine its current format

Considerations if a proxy model is in place

- The proxy server or device configured to fill the role of a proxy server should be investigated as a source of Web usage data
- Investigate the logging capacity of the proxy server or relevant device and consult the administrators to determine if it is plausible to configure the device to log Web usage data in an appropriate form, which contains URL and hostname fields. Logging Web usage data may have negative performance repercussions on the network

The case data used for this guideline included previously published research and documentation regarding the method used to develop the prototype, including literature findings, accounts of the developer's experience during development and resolving the issue of how to locate data for Web usage behaviour that was available, usable and rich with relevant information.

Further information on the evidence used to construct this guideline can be found in Addendum A.

Please rate the above guideline for each of the following characteristics

A) Clarity and Completeness

Please indicate how strongly you agree or disagree that the above guideline is clear and unambiguous.

5

Strongly Agree

Comments

--

B) Validity

Please indicate how strongly you agree or disagree that the above guideline is correct and accurate.

5

Strongly Agree

Comments

C) Flexibility

Please indicate how strongly you agree or disagree that the above guideline would be appropriate for use in another university with a similar network infrastructure.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Comments

--

Guideline 2

Conduct face to face interviews with the intended users of the Web usage information and pose questions which directly focus on exactly what information they would consider valuable

Recommendations

- Identify potential users of the information which could be derived from the available Web usage data. This could be done by targeting a group which has perceived association with the information and gathering their information needs through a group meeting or survey
- Potential users would be identified from the meeting or survey results if they indicate strong influence to their decisions from the proposed information
- If no potential users are identified reconsider the perceived association mentioned above
- From these potential users, interview one or more of the users and ask questions which directly focus on determining exactly what information they would consider valuable
- Clarify any misunderstandings with the interviewee
- Document the interview through recording and creating a transcript of the interview for future reference.

The case data used for this guideline included transcripts of interviews with specific users of the prototype, results of a survey conducted to identify possible users and documentation regarding the method used to develop the prototype, including literature findings, accounts of the developer's experience during development and resolving the issue of how to gather the information needs of intended users of the prototype.

Further information on the evidence used to construct this guideline can be found in Addendum A.

Please rate the above guideline for each of the following characteristics:

A) Clarity and Completeness

Please indicate how strongly you agree or disagree that the above guideline is clear and unambiguous.

5

Strongly Agree

Comments

--

B) Validity

Please indicate how strongly you agree or disagree that the above guideline is correct and accurate.

5

Strongly Agree

Comments

C) Flexibility

Please indicate how strongly you agree or disagree that the above guideline would be appropriate for use in another university with a similar network infrastructure

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Comments

Guideline 3

Consult individuals with data warehousing expertise when creating a star schema

Recommendations

- Identify individuals within the institution who could have data warehousing expertise
- Construct a draft schema and specify the business process for which it is intended as well as the required granularity
- Allow the individuals to rate the suitability of the draft schema and allow them make suggestions which may improve the design
- Make the appropriate changes and allow the individuals to review the draft schema once the suggested changes have been made, if necessary, to consider the changes made and the input from other individuals. Repeat this review process until the schema is deemed suitable.

The case data used for this guideline included documentation on three individuals with considerable educational and industrial expertise in the data warehousing field, who reviewed the star schema design for the prototype data mart. In addition, documentation regarding the method used to develop the prototype, including literature findings, accounts of the developer's experience during development and in resolving the issue of dimensional modelling requiring a sound star schema design was included.

Further information on the evidence used to construct this guideline can be found in Addendum A.

Please rate the above guideline for each of the following characteristics:

A) Clarity and Completeness

Please indicate how strongly you agree or disagree that the above guideline is clear and unambiguous.

5

Strongly Agree

Comments

--

B) Validity

Please indicate how strongly you agree or disagree that the above guideline is correct and accurate.

5

Strongly Agree

Comments

C) Flexibility

Please indicate how strongly you agree or disagree that the above guideline would be appropriate for use in another university with a similar network infrastructure.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Comments

Guideline 4

Profile, sort, isolate and extract required values from data entries in the flat file using Excel or a similar spreadsheet base application first. Then load the data into the data marts star schema tables, use discretion if the log formats are different.

Recommendations if the sample log file contains inconsistent fields

- Export a sample log file into a spreadsheet using Excel or similar spreadsheet software
- Utilize sorting functionality to determine the level of inconsistency of the fields of the entries
- Use search formulas to pull required values from cells in each entry
- Move the values into a separate spreadsheet
- Export the spreadsheet into the appropriate dimensional tables or staging table in the destination database

The case data used for this guideline included documentation regarding how data was moved from its source location to the data base software in which the prototype resided. In addition, documentation regarding the method used to develop the prototype, including literature findings, accounts of the developer's experience during development and in resolving the issue of how the data was to be moved and cleaned and why it was done in this way was included.

Further information on the evidence used to construct this guideline can be found in Addendum A.

Please rate the above guideline for each of the following characteristics:

A) Clarity and Completeness

Please indicate how strongly you agree or disagree that the above guideline is clear and unambiguous.

5

Strongly Agree

Comments

--

B) Validity

Please indicate how strongly you agree or disagree that the above guideline is correct and accurate.

5

Strongly Agree

Comments

C) Flexibility

Please indicate how strongly you agree or disagree that the above guideline would be appropriate for use in another university with a similar network infrastructure.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Comments

--

Guideline 5

Develop a prototype data mart which provides meaningful information for a single business process using only a sub set of data

Recommendations

- Develop a single, self-serving, small scale prototype data mart before considering a full DW/BI system
- Use a small set of sample data
- Focus on a single business process in one department

The case data used for this guideline included documentation of the steps in the development of the prototype. In addition, documentation regarding the method used to develop the prototype, including literature findings, accounts of the developer's experience during development and in resolving the issue of scaling the Web usage data analysis down into a project that could be undertaken in a short period by one developer, and providing a demonstration of the information presentation possibilities was included.

Further information on the evidence used to construct this guideline can be found in Addendum A.

Please rate the above guideline for each of the following characteristics:

A) Clarity and Completeness

Please indicate how strongly you agree or disagree that the above guideline is clear and unambiguous.

5

Strongly Agree

Comments

--

B) Validity

Please indicate how strongly you agree or disagree that the above guideline is correct and accurate.

5

Strongly Agree

Comments

C) Flexibility

Please indicate how strongly you agree or disagree that the above guideline would be appropriate for use in another university with a similar network infrastructure.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Comments

Guideline Set

D) Utility

Please indicate how strongly you agree or disagree that the above set of guidelines are useful for their intended context.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Comments

Addendum A

Guideline Evidence Briefing

Guideline 1

Findings from the case data revealed that various benefits could result from identifying the system engineers of the network infrastructure that provided Web access to the institution. These engineers provided detailed information about the network infrastructure and could pinpoint the devices used to handle Web traffic and their capabilities. Furthermore, the system engineers could provide sample data for profiling. They were also able to indicate which Web activity could be captured. Any limitations or possible options regarding the data and its collection became clear. Some limitations became apparent and a pending issue was identified as the cause of a logging error that made the logs unsuitable for the prototype project.

The senior management personnel who oversee the work load of these system engineers were identified through the system engineer liaison. They were made aware of the prototype development effort. It was recognised that mutual benefit could be gained from correcting the identified logging error and it was given some additional attention. Once the error had been corrected more detailed logs became available. These logs contained richer Web usage information that enhanced the development efforts and a pending issue was resolved by the system engineers. The previous investigation into possible data for the prototype proved beneficial to the project and allowed for the identification of data owners, which in turn resulted in an invaluable working relationship. Furthermore, senior sponsors were gained in this way and value was added. Moreover, previous research findings, which were validated through publication, recommended early data source investigation, which was consistent with the Kimball Lifecycle approach used to undertake the prototype project. Therefore, the above guideline can be considered a feasible recommendation.

Guideline 2

It became clear from the case data that the interviewee saw distinct value in the report, which demonstrates that using the user's information requirements to develop the prototype data mart translated into valuable reports. The face-to-face interviewing technique provided a clear indication of the user's needs. Moreover, posing questions that addressed the exact information she desired resulted in clearly defined requirements that translated well to the prototype project design. This is consistent with authors in the data warehousing field who maintain that face-to-face approaches provide a rich and valuable source of feedback (Inmon, 2002; Scheps, 2013; Simon, 2009). Interviews with business users are also pivotal to precisely understanding their requirements (Kimball et al., 2008). Therefore, given that in the context of the case and the recommendations in the literature it is considered

important to conduct face-to-face interviews to gather information needs from users of the data warehouse system, the above guideline can be regarded as a feasible recommendation.

Guideline 3

Findings from the case data revealed that the first draft star schema proposed to the experts was clearly unsuitable. None of the experts agreed in the first iteration of the review that it was suitable for the business process. However, following the changes made according to expert consensus and a second review, the star schema became a suitable design. This star schema was further validated by successful implementation in the prototype data mart; this implementation was confirmed as correct by one of the experts. In the context of the case, the expert review was invaluable to the design of a suitable star schema and without it the prototype data mart would probably not have operated correctly. Consulting individuals with data warehousing expertise to create a suitable star schema was vital to the success of the prototype. Therefore, the above guideline can be considered to be a feasible recommendation.

Guideline 4

From the case data it was determined that creating automated Extract Transformation Load (ETL) processes was an enormous task that went beyond the scope of the prototype development. Manual ETL processes were more suitable. Using Excel for the initial profiling and editing of the data proved to be a successful alternative to using the SQL server. The importing of data from a flat file was allowed in Excel despite gross inconsistencies in the entries, whereas the SQL server did not allow this import directly from the flat file nor any profiling or editing. The flat file import capacity of Excel was desirable as flat files with log data are likely to be inconsistent and need to be edited. The sorting functionality in Excel proved effective in establishing the extent of column inconsistency. Moreover, the formulas available in Excel to search through the data and extract values within an entry were essential to this ETL process. Without this facility for editing, profiling and ultimately extracting specific fields and values from the log files provided by Excel, the data would not have been suitable for insertion into the data mart. Therefore, the above guideline can be considered a feasible recommendation.

Guideline 5

Findings from the case data revealed that developing a DW/BI system is a massive undertaking. By creating a prototype data mart instead of attempting to create a full system, some working relationships were formed and many issues, which may have been discovered only later in the development process, became apparent early on. The logistics of moving the data from source to destination and methods for cleaning the data became clear. Information could be presented to the

intended users within a reasonably short time, confirming the value of the information gathered from the data. Moreover, the prototype demonstrated that the method used to develop it was appropriate. An understanding of how resource intensive a full DW/BI system development effort could be was gained. Using a sub set of data allowed for detailed data profiling, which meant that the process required to clean the data could be carried out completely and be fully understood. The method used to develop this prototype allows for expansion from a single data mart to a full DW/BI system by developing and connecting additional data marts. For these reasons, the above guideline can be considered a feasible recommendation.

Appendix F

Guideline Validation Expert Review Responses

Guideline 1 - *The data and the owner/s thereof should be investigated to gather and profile a sample of the required data before the DW/BI project can be considered feasible.*

Clarity and completeness

The experts indicated how strongly they agreed or disagreed that guideline 1 was clear and unambiguous.

Expert 1: Disagree: “I believe the guideline should be slightly modified to include the fact that this applies specifically to Web usage data in the primary guideline itself. Thereafter, I would agree.”

Expert 2: Agree: “data ownership? Multiple owners? Different data items vs data?”

Expert 3: Agree

Validity

The experts indicated how strongly they agreed or disagreed that guideline 1 was correct and accurate.

Expert 1: Agree

Expert 2: Strongly Agree

Expert 3: Agree

Flexibility

The experts indicated how strongly they agreed or disagreed that guideline 1 would be appropriate for use in another university with a similar network infrastructure.

Expert 1: Strongly Agree

Expert 2: Strongly Agree: “Again the concept of data ownership might need better definition”

Expert 3: Agree

Guideline 2 - Conduct face to face interviews with the intended users of the Web usage information and pose questions which directly focus on exactly what information they would consider valuable

Clarity and completeness

The experts indicated how strongly they agreed or disagreed that guideline 2 was clear and unambiguous.

Expert 1: Strongly Agree

Expert 2: Agree: “Unclear how I can perceive an association without users”

Expert 3: Agree

Validity

The experts indicated how strongly they agreed or disagreed that guideline 2 was correct and accurate.

Expert 1: Agree: “Other methods of identifying valuable data might also be appropriate but interviews are definitely the most feasible.”

Expert 2: Strongly agree: “the guideline statement is spot on, but the recommendations less so”

Expert 3: Agree

Flexibility

The experts indicated how strongly they agreed or disagreed that guideline 2 would be appropriate for use in another university with a similar network infrastructure.

Expert 1: Strongly Agree

Expert 2: Strongly Agree

Expert 3: Agree

Guideline 3 - Consult individuals with data warehousing expertise when creating a star schema.

Clarity and completeness

The experts indicated how strongly they agreed or disagreed that guideline 3 was clear and unambiguous.

Expert 1: Disagree: “I would be careful to state data warehousing in general. Maybe this should be specifically about dimensional modelling.”

Expert 2: Strongly agree: “It is really stating either the obvious or representing a ‘luxury’ as they may not be available”

Expert 3: Agree

Validity

The experts indicated how strongly they agreed or disagreed that guideline 3 was correct and accurate.

Expert 1: Agree

Expert 2: Neutral: “but then not a show-stopper. Assuming non-expertise of designer”

Expert 3: Agree

Flexibility

The experts indicated how strongly they agreed or disagreed that guideline 3 would be appropriate for use in another university with a similar network infrastructure.

Expert 1: Agree

Expert 2: Neutral: “as above”

Expert 3: Agree

Guideline 4 - Profile, sort, isolate and extract required values from data entries in the flat file using Excel or a similar spreadsheet based application first. Then load the data into the data marts star schema tables, use discretion if the log formats are different

Clarity and completeness

The experts indicated how strongly they agreed or disagreed that guideline 3 was clear and unambiguous.

Expert 1: Strongly agree

Expert 2: Neutral: “This seems to be suggesting a staging area similar to normal data warehousing exercises”

Expert 3: Agree

Validity

The experts indicated how strongly they agreed or disagreed that guideline 3 was correct and accurate.

Expert 1: Strongly agree

Expert 2: Agree: “Some staging necessary, but unsure about “technology-bond nature”

Expert 3: Strongly Agree

Flexibility

The experts indicated how strongly they agreed or disagreed that guideline 3 would be appropriate for use in another university with a similar network infrastructure.

Expert 1: Strongly agree

Expert 2: Neutral: “Probably, but other options to intermediate spreadsheets can work as well”

Expert 3: Neither agree nor disagree

Guideline 5 - Develop a prototype data mart which provides meaningful information for a single business process using only a sub set of data

Clarity and completeness

The experts indicated how strongly they agreed or disagreed that guideline 3 was clear and unambiguous.

Expert 1: Strongly agree

Expert 2: Strongly agree: “Picturing the low hanging fruit is always a good idea”

Expert 3: Strongly Agree

Validity

The experts indicated how strongly they agreed or disagreed that guideline 3 was correct and accurate.

Expert 1: Strongly agree

Expert 2: Strongly Agree

Expert 3: Strongly Agree

Flexibility

The experts indicated how strongly they agreed or disagreed that guideline 3 would be appropriate for use in another university with a similar network infrastructure.

Expert 1: Strongly agree

Expert 2: Strongly Agree

Expert 3: Agree

Guideline as a complete set

Utility

The experts indicated how strongly they agreed or disagreed that the set of guidelines 1—5 was useful for their intended context.

Expert 1: Agree

Expert 2: Disagree: “Very high level and specify process rather than ‘what’ only address some ‘how’

Expert 3: Agree

Appendix G

ZAWWW Conference Paper

Web Usage Mining within a South African University Infrastructure: Towards useful information
from student Web usage data

(Von Schoultz, Van Niekerk & Thomson, 2013)

Web Usage Mining within a South African University infrastructure: Towards useful information from student Web usage data

D.J. von Schoultz
Nelson Mandela Metropolitan University
Port Elizabeth
South Africa
s209063124@live.nmmu.ac.za

J. van Niekerk
Institute for ICT Advancement
Port Elizabeth
South Africa
Johan.vanniekerk@nmmu.ac.za

K. Thomson
Institute for ICT Advancement
Port Elizabeth
South Africa
Kerry-Lynn.Thomson@nmmu.ac.za

Abstract

It is widely known that the Internet is providing a deluge of information to anyone with access to it. Various studies have been conducted to determine if this abundance of information is detrimental to the academic environment, or if it is providing the ideal means for ultimate educational prowess.

Each university will have different infrastructures in place to handle varying amounts of network traffic data. Therefore, the 'gateway' within the network used to access the Web securely will likely differ according to specific institutional needs. Investigating the availability and format of the network traffic from this 'gateway' in addition to the collection thereof is a feasible start to gaining some useful information from this data.

This paper will present research which attempts to determine the state of raw network activity log data and the usability thereof within the Nelson Mandela Metropolitan University. This is done as an initial study towards using Data Mining techniques on this network data to provide knowledge regarding Internet usage by students.

Keywords: Web Usage Mining, Web usage behavior, Data Mining, Web Mining

1. Introduction

Web activity does not go unnoticed. Various marketing organizations pay top dollar for online preferences and interests. A person exposes his/her interests and preferences digitally every time the information goes from memory or thought to 1's and 0's online.

A multitude of signals are flagged and interpreted during a session of information indulgence via online browsing. From the source IP address used for connection, the type of browser being used, the amount of time taken to select a search result, to what is being clicked on (Pariser 2011).

'Click signal' is a term used to describe a bit of data which is created and stored when a user clicks on any link or interface element on a Web page. These signals are used extensively for Web personalization. If someone searches for "Pink Floyd" and clicks on the fourth link, it suggests this link is more relevant and could be seen as a tally. Furthermore, click signals form only a small part of the data collected by servers which are then processed using Data Mining techniques. These techniques are used to derive user behaviour patterns to provide powerful marketing tools such as personalization.

The first phase in Web personalization is the collection of Web data. This includes clickstreams (virtual trails left by a user's computer as they browse and surf the Web), click signals, past activities found in Web server access logs and/or via cookies or session tracking modules (Schafer, Konstan & Riedl 1999).

The ability to harvest and refine data to the detail of individual mouse clicks, allows for an astounding opportunity to personalize the Web experience of users (Mobasher, 2000).

Web personalization affects a user's experience with many of the websites used. It tailors search results, news feeds, advertising and content feeds to provide some relevance to a user's specific needs when sifting through the over abundant mound of information on the Web. This type of advertising is highly sought after by e-commerce companies. For every correctly tailored advertisement, there would have been a high amount of irrelevant ones which would essentially be wasted on that particular user. Therefore, personalized advertising is highly favourable in an online environment because wasted or unsuccessful advertising is a costly exercise in most circumstances.

If this data is so commercially valuable, how else could it be used to provide feedback for statistics, management alterations, comparative analyses or general knowledge discovery? Additionally, what techniques are used by e-commerce experts to correctly collect and analyse Web data and can they be applied to network data retrieved from an educational institution?

Identifying the data needed for the initial steps of the Data Mining approach would be a viable starting point. Additionally, it must be determined what data is available, if it is in a relevant or usable format and its level of availability.

Conclusions derived from this data may provide senior management with highly useful and relevant information. Subsequently, this may lead to the recognition of problem areas within the university Internet access policies which could be calibrated accordingly. This

could result in a more prolific academic environment for students by taking justifiable action based on hard facts provided by the analyses of Web data.

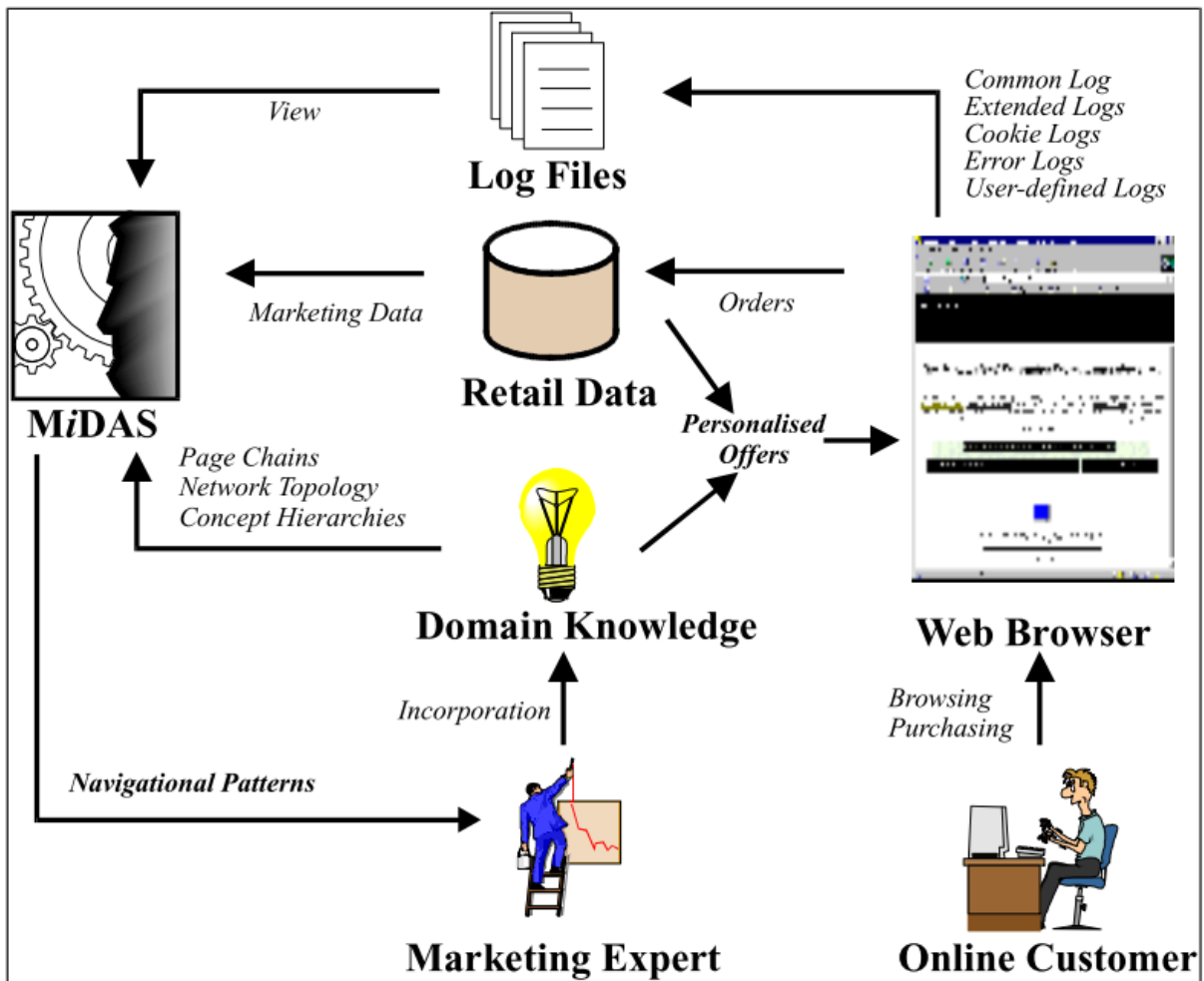
The initial aim of this study is determine if the appropriate network traffic data is available and if Data Mining is a feasible approach for deriving student Web usage information.

2. E-commerce Data Mining

The Internet has become a business mecca and e-commerce is a highly competitive environment. Therefore, organisations need to maintain and solidify relationships with high-value consumers (Büchner, Anand, Mulvenna & Hughes 1998).

Personalising online services and products is one way to improve customer relationships which is well recognised and widely used in online business. Jeff Bezos, CEO of Amazon.com, can be seen as being a pioneer of this type of innate online marketing. Amazon was launched in 1995 and it had the capability of providing unique recommendations to a customer based on their previous purchases. By monitoring the users Web activity using collaborative filtering methods developed at PARC, Amazon made ad hoc recommendations while the user browsed for books (Pariser 2011). The more Web traffic a user produced the higher the quality of the recommendation. This resulted in a highly successful and well recognised marketing approach. Amazon was one of the businesses on the forefront of e-commerce and proved that there is huge profit to be made online (Pariser 2011).

Figure 1: The MIMIC Architecture (Büchner et al. 1998).



Constant monitoring of customers to identify eventual market changes and calibrating accordingly is vital for increasing profit (Hofgesang & Kowalczyk 2005). Therefore, the collection of user Web usage data is highly justified by online businesses. This could account for or be a large contributing factor for the abundance of data available in Web servers and other data collection mechanisms in place online.

Figure 1 shows the MIMIC (Mining the Internet for Marketing Intelligence) architecture which was developed to provide marketing intelligence from stored log files and expert knowledge. This demonstrates how the constant collection of user activity is in place and is used to provide personalised advertising to the user based on their activity. The entity labelled as MiDAS (Mining Internet Data for Associative Sequences) is an algorithm which provides the marketing expert with navigation patterns (Büchner et al. 1998).

Data Mining is a method for translating data into information and with the correct approach, deriving highly relevant knowledge from that information. Moreover, Data Mining deals with data which is collected for a purpose separate to the purpose for which the data was originally collected, similar to the way in which MIMIC uses log files generated by an online customer. However, data in the incorrect format or a format which is at the incorrect granularity is not useful. Therefore, it can be said that the format of server logs and the

formatting of server logs is a vital area in analysing Web usage via Data Mining. Therefore, the approach to formatting data in Data Mining should be explored.

3. Data Mining

Hand defines Data Mining as the “analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner” (2007).

Data Mining encompasses many aspects of discovering or deriving information from datasets which are usually in high volume. The main difference between Data Mining and forms of statistical analysis is that Data Mining usually deals with data which was retrieved for another purpose other than data analyses (Hand, Mannila & Smyth 2001).

Essentially, Data Mining looks at data which is already collected rather than collecting data for analysis. Therefore, the goals of the Data Mining process will not be inherent in the original data collection method.

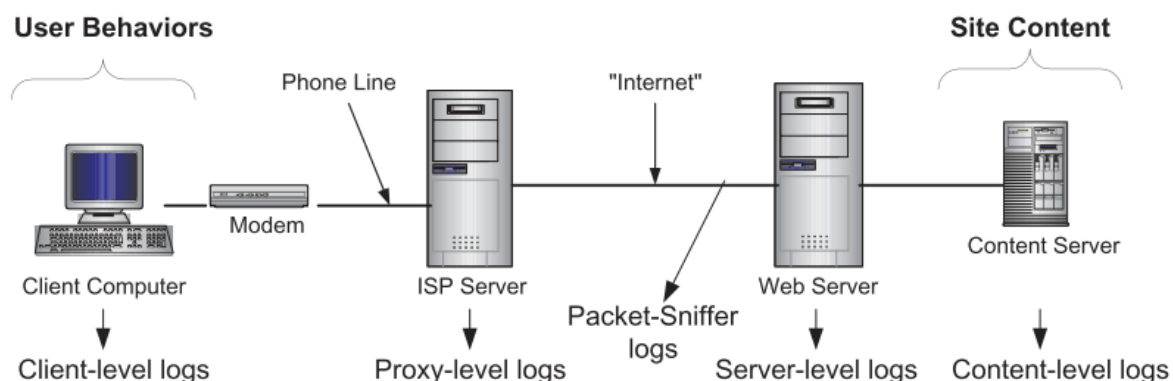
The World Wide Web is a ripe domain for Data Mining research, mainly due to the sheer volume of data available (Kosala & Blockeel 2000). The sub-category of Data Mining dealing with the mining of Web data is widely referred to as Web Mining

3.1 Web Mining

The Web provides users with seemingly limitless amounts of information. Furthermore, it can provide Web designers, system administrators and behaviour analysts, amongst others, with a massive amount of data about users. This is typically done via automatic caching (typically storing log files) at various levels of the client, proxy and Web server relationship. Log files can be found at three different places: Web server, Web proxy server and client browsers (Suneetha & Krishnamoorthi 2009).

Figure 2 shows a simple Web access diagram which graphically depicts the relationship mentioned.

Figure 2: Simplified Web Access Diagram (Cooley 2000).



A Web server stores and distributes data for one or more websites. A proxy server or Internet Service Provider (ISP) acts as a gatekeeper for subnets that require access to the Internet. Proxy servers are favourable because it provides a medium for Internet access which results in a higher level of security and single point of management of the network traffic (Luotonen & Altis 1994). Lastly, a client simply refers to the computer which the user is using to browse the Web. Clients are the source of the requests.

The servers which handle these requests will store data as the requests flow through them. This is done for various performance, analyses and administrative reasons. The storage of this data by the servers is commonly known as 'caching'.

Caching is more specifically used to store frequently used data, as well as Web pages, which are stored locally on a temporary basis for retrieval at a later stage to decrease commonly repeated requests to the Web server (Luotonen & Altis 1994). This is one of the ways that Web Mining can be considered a Data Mining process, because the original purpose of storing Web log data is clearly defined as set out for caching purposes. Therefore, it could be said that caching itself is not considered Web Mining or a form of Data Mining. Moreover, if the desired results for a Web mining project are close to or the same as the purpose of caching, it should not be considered strict Web Mining.

Caching occurs in proxy servers as well as Web servers. This caching output is usually found in the form of access log files in various formats which are briefly discussed in section 6. This is where the majority of the data used for Web Mining is retrieved from as well as the access log files from the Web server which is the main source.

Web Mining is commonly divided into the following three sub-areas (Hu & Zhong 2005 p.5):

- Web Content Mining: extracting knowledge from the unstructured or semi-structured content of Web pages;
- Web Structure Mining: studying the hyper-link structure of a website and the Web;

- Web Usage Mining: applying Data Mining techniques to Web usage data aiming at identifying interesting usage patterns, which is greatly employed for building a user profile (Adomavicius & Tuzhilin 2001).

Web content mining and Web structure mining is not considered in this study. The reasons for this are provided below.

Web Content Mining is not considered because the actual content viewed by a student can be deciphered once the data has been correctly formatted and arranged. Furthermore, the exact queries which would be posed have not been determined; this paper is considering how Data Mining can be used to initially gather information from logs.

Web Structure Mining is not considered at this stage for similar reasons because it refers to the organization of the Web page content.

Further, Web Usage Mining is the “application of Data Mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications” (Srivastava, Cooley, Deshpande & Tan 2000).

Web Usage Mining is utilized for personalization, system improvement, site modification, business intelligence and usage characterization.

Figure 3: Major Application Areas for Web Usage Mining (Cooley 2000)

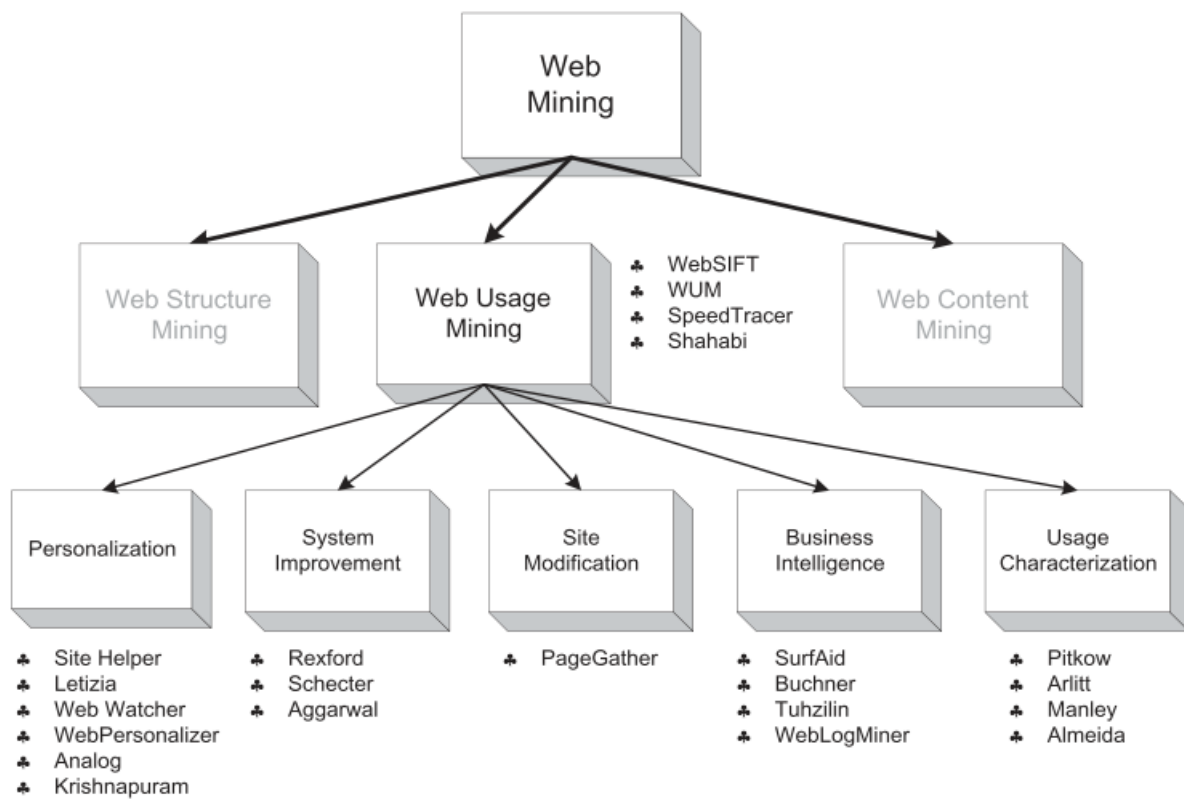


Figure 3 depicts examples of areas which utilize Web Usage Mining for ad hoc purposes. It includes the parallel types of Web Mining, namely; content and structure as mentioned earlier in this section.

The Web Usage Mining sub-area is in essence the discovery of usage patterns from Web Access logs i.e. raw usage data. It relies on the following steps (Suresh & Padmajavalli 2006).

1. Pre-processing
2. Pattern discovery
3. Pattern analysis

Figure 4: High Level Web Usage Mining Process (Cooley 2000)

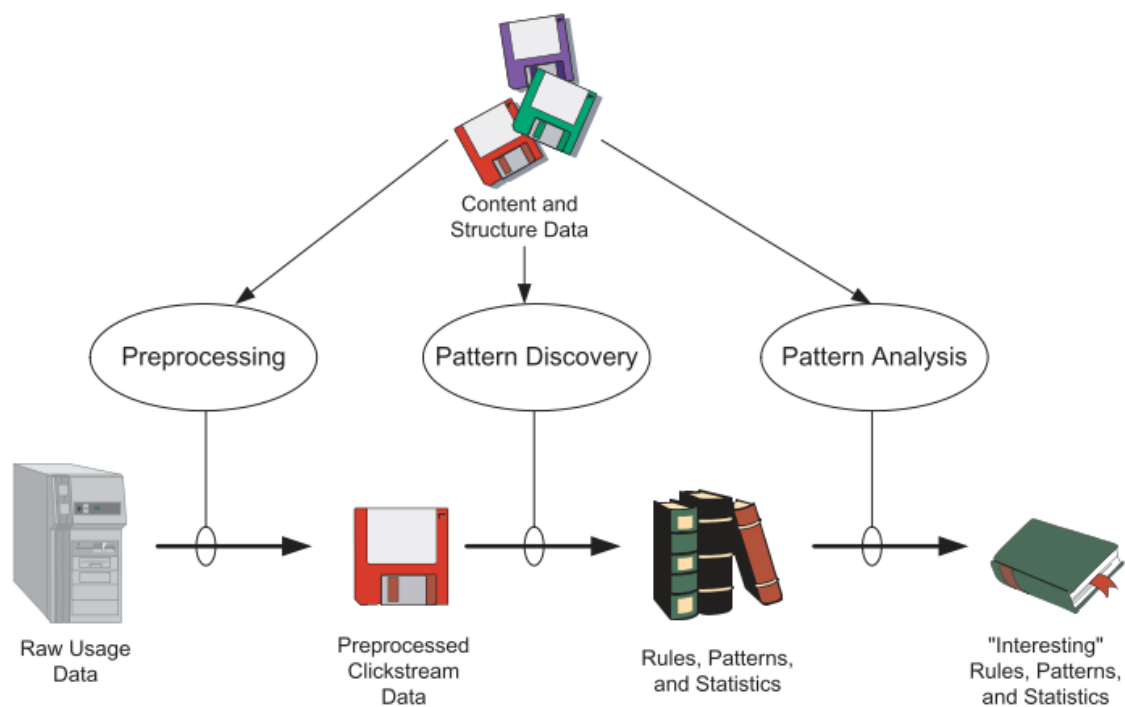
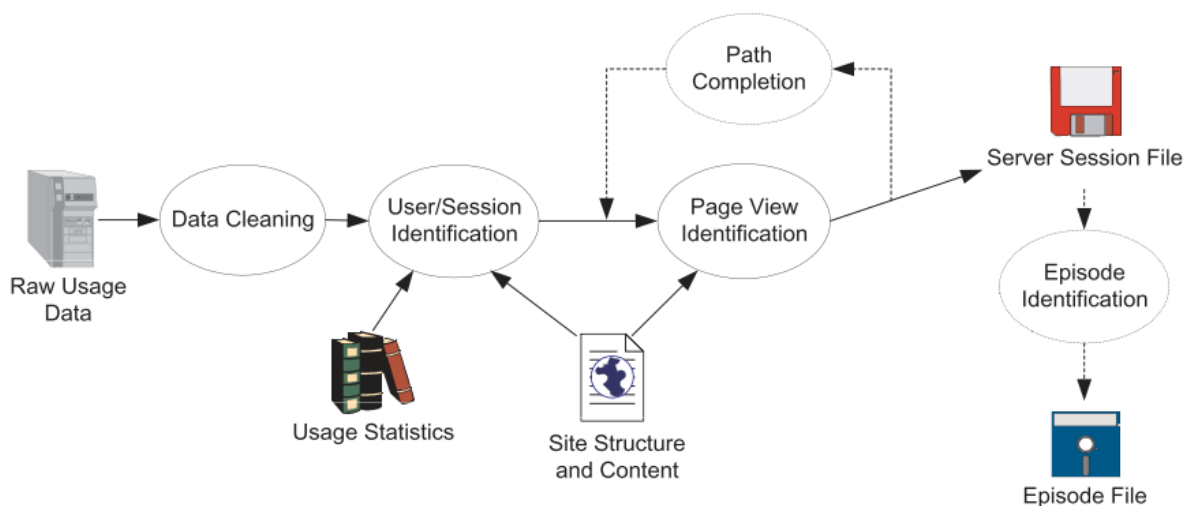


Figure 4 provides the overall Web Usage Mining process which shows the various inputs and outputs given and received by the Web Usage Mining steps.

Content and structure data is used in the pre-processing step of Web Usage Mining. This data acts as input for User/Session Identification and Page View Identification which falls beyond the scope of this paper as shown in Figure 5. Figure 5 confirms that site structure and content data is separate from the raw usage logs and that cleaning raw usage data is an initial phase of the pre-processing step.

Figure 5: Details of Usage Pre-processing (Cooley 2000).



Pre-processing is the overall higher level formatting of data from multiple sources such as raw usage data, site structure and content data into a usable format. This needs to occur in order to move into the pattern discovery step in Figure 4.

The scope of this paper will not move past the data cleaning phase of the pre-processing step of Web Usage Mining.

Inputs for the pre-processing step are server logs (raw usage data) and site files (structure and content data). Additional inputs could be usage statistics concluded from analysis undertaken beforehand (Cooley, Mobasher & Srivastava 1999).

These techniques are well studied due to the high level of adoption within online business and e-commerce. However, therein lies various complexities, such as, large amounts of data, user identification, proxy caching, user session separation and automated refreshing.

Data Mining is a well-established approach for deriving information from existing data. Many other approaches are available some of which are discussed in the following section.

4. Web usage analysis: previous qualitative approaches

The Internet is a highly distributed information resource. It has, therefore, been the interest and pertinent focus of many researchers since its exponential growth in the early 90s. Many studies have been conducted on the effects of this plethora of information.

There have been other approaches to deciphering Internet usage. Namely survey or interview based research. The studies which follow have taken a qualitative approach to infer some information about Web usage from a group of users.

Schumacher & Morahan-Martin concluded through the use of a survey that males tend to have a higher level of interest in computers, games and technology. Therefore, they may wield a higher level of competency with regards to Internet use (Schumacher & Morahan-Martin 2001).

Research into the extent of excessive Internet use among youth in Singapore was conducted via a questionnaire study involving 2735 adolescents (Mythily, Qiu & Winslow 2008). They found that excessive Internet use has various negative effects on many facets of the life of an adolescent including academic performance. It was concluded that males were twice as likely to be excessive Internet users, with excessive defined as more than 5 hours a day. They found that 17.1% of the participants were deemed to be excessive Web users.

Using 985 university students in a 55-item survey, Rumbough and Arts considered how taboo Web material was affecting college students (Rumbough 2001). They explored the extent to which students access contentious websites such as illegal drug sites, pornography sites, illegal weapon sites, racist sites, gambling sites and fake ID sites. They

attempted to answer or decipher the extent to which students partake in unethical behaviour such as academic cheating, fake e-mail, inappropriate e-mail and software piracy and if gender differences relate to these behaviours. They concluded that a significant amount of users do frequently indulge in unsavoury Web content.

Metzger, Flanagin and Zwarun used 356 undergraduate students in a questionnaire study to determine how reliant the students were on the Internet for academic material and general information (Metzger, Flanagin & Zwarun 2003). The conclusion they reached was that there is a problem in that students do not trust the integrity of the information they gather online, but they do not regularly verify this information.

There are abundant variables to consider when attempting to analyse the Internet use of an individual and that of a large group of individuals. Many of these issues have been identified during, prior and after research which has been conducted in previous studies.

5. Use in internal client analyses

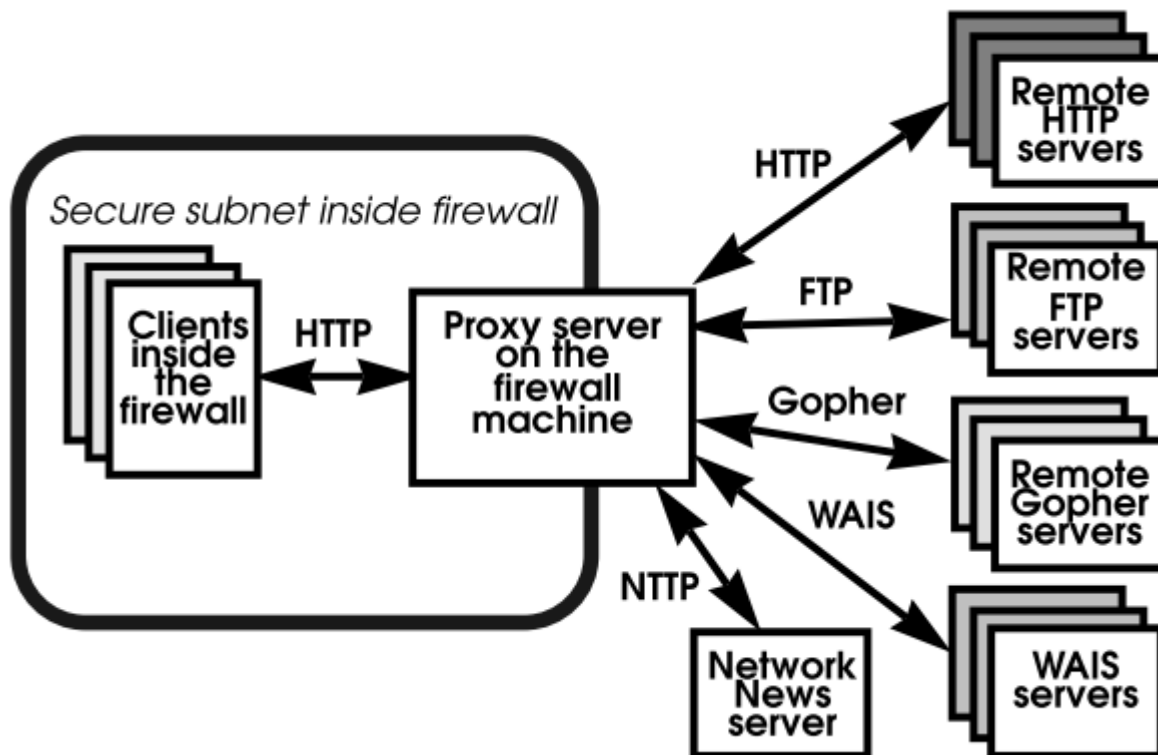
Various qualitative assessments have been conducted in this area of research, as demonstrated in section 4. This does, however, not provide a direct data representation or hard fact analyses of the Web usage behaviours of the candidates used in the various surveys. A more quantitative approach may provide a finer grain of information.

Many large organizations require secure access to the Web for all of its users who make use of proxy servers. Proxy servers are used primarily to allow Web access from within a firewall. A firewall inspects and filters sent and received network traffic from a secure network. Moreover, it enforces a security policy to this traffic by comparing incoming packets to a rule set (Fulp 2005). Thereby providing and maintaining a secure network. A secure network with Web access is highly desirable to organizations hence the wide adoption of proxy servers. In a usual case the same proxy is used by all clients within a subnet (Luotonen & Altis 1994). Therefore, all traffic requesting or receiving Web data has to go through the proxy server.

Figure 6 shows a general proxy server setup for a secure subnet requiring Web access. Figure 6 is included to provide a general common platform of comparison to see if NMMU subscribes to this proxy setup. A question is posed during the expert interview in section 8 to clarify this. By doing so it can be determined if there is indeed a single point of Web access which can provide web log data from all the students within the subnet. Furthermore, it could provide a common ground for other organizations that have carried similar research or are considering similar research.

Section 6 provides the approach and findings of a study conducted by another South African university who has derived findings using Web logs retrieved from a proxy server.

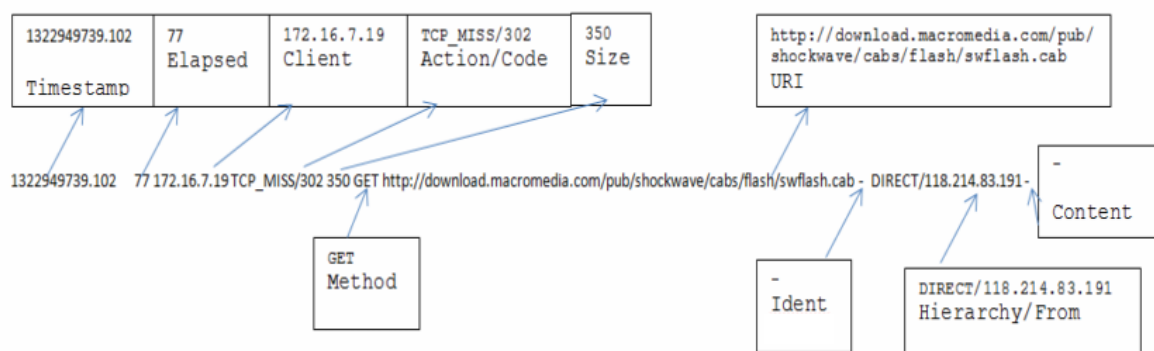
Figure 6: Overall setup of a proxy (Luotonen & Altis 1994).



6. Previous Data Mining approaches for University Web usage behaviour analyses

The University of Witwatersrand, Johannesburg (Wits) implemented a strict policy in 2007 to ensure that all Web browsing was done through an authenticated proxy. This provided a secure Web access medium similar to the model demonstrated in Figure 6. With this proxy in place any Web activity triggered by students is logged. Hazelhurst, Johnson and Sanders employed these proxy logs to analyse students Web usage (2011). The proxy server produced log files in a squid proxy access log format as shown in Figure 7.

Figure 7: Squid proxy access log format (Hossain, Rahman & Kabir 2012)



These squid access logs are available in two formats, the Common Log file Format (CLF) or the native log file format.

The Common Log file Format is adopted by many HTTP servers (Wessels 2001). It is composed of the following fields: remotehost rfc931 authuser [date] "method URL" status bytes.

The log files were extracted and inserted into a SQLite database. SQL provided the means for the bulk of the analysis. Additionally, Python scripts were used for identifying user sessions.

Another format considered in server log files is the Extended Common Log Format (ECLF) as shown in Figure 8.

Figure 8: Sample Extended Common Log Format (Cooley 2000)

IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
123.456.78.9	--	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
123.456.78.9	--	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
123.456.78.9	--	[25/Apr/1998:03:06:02 -0500]	"POST /cgi-bin/p1 HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
123.456.78.9	--	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla (IE4.2, WinNT)
123.456.78.9	--	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla (IE4.2, WinNT)
123.456.78.9	--	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla (IE4.2, WinNT)
123.456.78.9	--	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
123.456.78.9	--	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla (IE4.2, WinNT)
123.456.78.9	--	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
209.456.78.2	--	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
209.456.78.3	--	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1680	A.html	Mozilla/3.04 (Win95, I)

The purpose of the Wits study was to investigate the correlations between Web use behaviours and academic performance. Hazelhurst et al concluded that the Web is used mainly as a social medium, higher Internet use correlates to lower academic performance and the Internet usage profile for a prolific student is particularly dissimilar to the profile of a weak student (Hazelhurst et al. 2011).

The format of the log files within this study is clear and usable. Therefore, the format of the log files and availability thereof at the NMMU network department needs to be established. The main reason for this is to do a comparison to the above mentioned formats to determine how usable the log files are and if the same approach can be used.

7. Available data

The initial data retrieved in this study was readily available. It was requested from a system engineer stationed at the network department, who provided a .log file which contained firewall traffic data for the entire month of April 2013. However, whether it is in a suitable format for analysis using Data Mining techniques needs to be determined. It contains all packet transactions sent through the firewall device for the NMMU. The size of the .log file is 39.2 gigabytes. Table 1 depicts an example of an entry in the mentioned log file separated by each field and corresponding value.

Field	Value
Month	Apr
Day	12
Timestamp	11:52:15
unspecified	tyrael
date	2013-04-12
Time	11:52:15
devname	imperius
device_id	FGT1KC3912800514
log_id	0038000004
type	traffic
subtype	other
pri	notice
vd	root
src	10.102.129.162
src_port	53149
src_int	"LAN_AGGR"
dst	173.236.49.82
dst_port	80
dst_int	"INTERNET"
SN	637641392
status	start
policyid	128
dst_country	"United States"
src_country	"Reserved"
tran_sip	192.96.15.20
tran_sport	61201
service	HTTP
proto	6
duration	0
sent	0
rcvd	0

Table 1: Firewall log entry example

This data was retrieved from the Fortigate firewall device within the NMMU network. It provides a timestamp, source IP Address and destination IP Address but no URI (Uniform Resource Identifier) data. In order to decipher Web usage, actual HTTP addresses need to be documented. This is so that a URL (Uniform Resource Locator) can then be categorized for analysis. Given the above packet request the destination IP Address is 173.236.49.82 this holds no meaning until a nslookup request is made to retrieve the URL. Therefore, in order to process this data an nslookup needs to be run on each packet to determine the URL as per Figure 9.

The format should ideally subscribe to the Common Log file Format in order to be useful. The reason is that the Common Log file Format is widely used and it contains URL data and is very similar to the squid log files. This could provide more correlation to other proxy types later on.

To determine the details of what is used at NMMU and the exact problems experienced in retrieving the correct data format for Web Usage Mining, internal network experts were consulted.

8. Verification

The information required to confirm and clarify the problems mentioned in the previous section needs to be gathered. The system engineer who provided the initial snap shot data sample is directly involved in the configuration and maintenance of the Fortigate firewall device. Therefore, the information required can be retrieved from this engineer.

8.1. Methodology

Various qualitative approaches are available for extracting information from an individual. Considering that the individual in question is readily available for face to face meetings and is willing to participate in the project, an interview is an appropriate approach. Interviews are a systematic way of conversing with an individual in order to collect data from them (Kajornboon 2005).

The system engineer mentioned has highly specified knowledge in the context of this research. One of the main uses of elite interviews is to extract information and context that only a specific individual can give regarding a particular event or process (Hochschild 2009). Therefore, an elite interview with this system engineer is an appropriate research method for this scenario.

A semi-structured interview will be used so that any further detail or questions which may arise from the initial questions can be investigated and detailed immediately. Further, semi-structured interviews allow the interviewer to have more discretion over the conduct of the interview than unstructured interviews (Mikecz 2012). Therefore, a semi structured elite interview will be used to collect information from the system engineer. An additional system engineer within the same department was available and included in the interview to validate information and provide additional insight.

8.2. Elite Interview

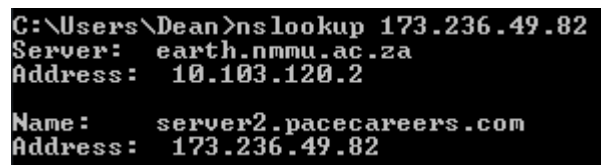
Two System Engineers from the NMMU North campus Network Team were used for the expert interview.

The purpose of this interview is to determine various factors specific to this university in order to clarify the context for further research towards the future aims of this project. Specifically the extent of responsibility of the network team, the type of gateway used for

the network, the similarities of this gateway compared to a basic proxy model and details regarding availability and format of network data. Based on this interview the following has been derived.

The network team in this department oversees the network needs for six campuses. These include the North, South, Bird Street, Missionvale, George and Second avenue campuses. This is favourable for further research because data is available for the entire university not only a specific campus. Furthermore, traffic from a distinct campus could be isolated for analysis.

Figure 9: Sample nslookup results



```
C:\Users\Dean>nslookup 173.236.49.82
Server:      earth.nmmu.ac.za
Address:     10.103.120.2

Name:       server2.pacecareers.com
Address:    173.236.49.82
```

Moreover, some of the entries in the log will not return a successful lookup as it may have formed part of an internal refresh or non-user activated request, which is not relevant to user Web behaviour.

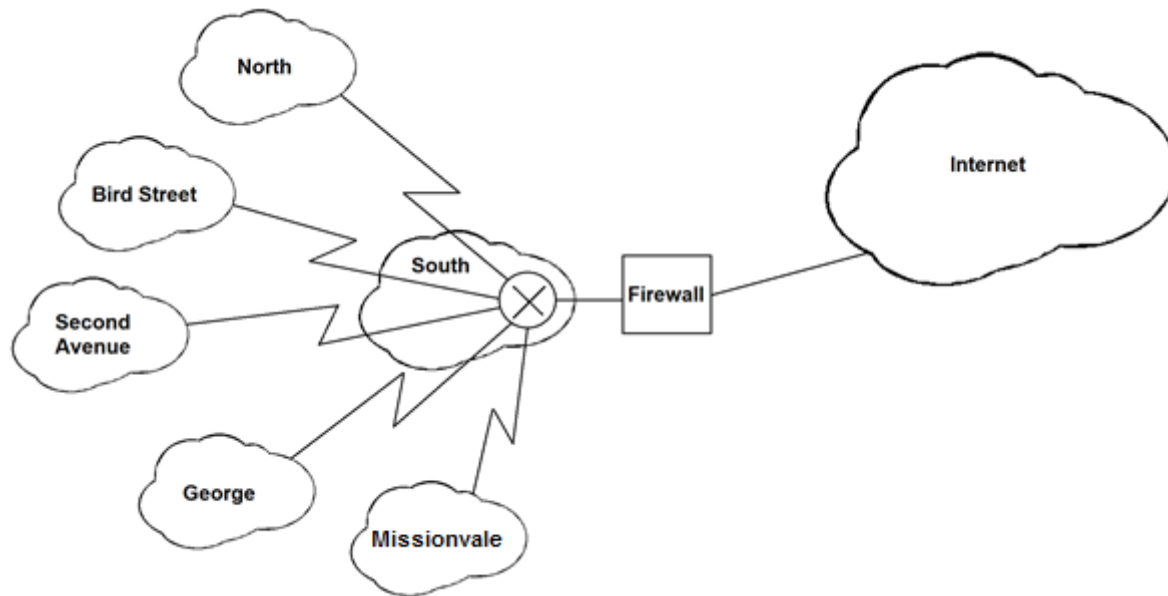
No proxy server is in place at NMMU. The reason for this is because the proxy server was in the early stages of being inadequate to handle the rapidly increasing network traffic volumes generated by the students and staff. This led to the adoption of a Fortigate firewall device which is more suited due to its larger capacity and filtering capability. A firewall does differ from a proxy server, however they accomplish a common task in blocking or limiting connections to and from a network. Therefore, the Fortigate firewall device will provide the same overall function as the authenticated proxy which provided the Squid proxy logs used in the study mentioned in section 6

Figure 9 demonstrates how the firewall is in a similar location to a traditional proxy server as labelled in Figure 6 as a 'Proxy server on the firewall machine'. Therefore, supporting the notion that the Fortigate firewall device is the appropriate source for Web traffic data collection to and from the university students when accessing the Web from one of the campuses.

Access.log files are available from the Fortigate firewall device. However it does not currently allow URL data to be written to these log files. Moreover, the log files are not available in a Common Log file Format.

The System Engineers have refined this limited format availability to a setting in the config file, the correct settings of which need to be identified and calibrated. The functionality is available; however the Fortigate needs to be calibrated correctly.

Figure 9: Nelson Mandela Metropolitan University logical network



9. Conclusion

The objective of this paper was to investigate and determine the state and format of Web log data within the NMMU network. Moreover, this is done to determine if Data Mining techniques could be applied to the data to derive information about the Web usage of the students within the university.

Data Mining is a well-studied field and widely adopted especially in e-commerce. Certain aspects of which could be used infer Web usage information from a university's network data. The area of Data Mining in the context of a subnet is not as thoroughly researched.

The data which is presently available from the NMMU network department does come from a firewall device which falls into the same model or structure as previous research and that which is stated in Web Mining. Therefore, it can be concluded that this data is a viable source to be used alongside Data Mining techniques.

However, at present the format of this data is not in a suitable state to begin the data cleaning phase of the pre-processing step of Web Mining. Furthermore, the reason for this has been identified in an expert interview as a technical configuration issue within the firewall device. Further research regarding ad hoc data cleaning can be done once the problems experienced by the System Engineers involved in the expert interview have been resolved.

The successful application of Data Mining in an organization using a proxy model is dependent on the type of gateway in place. This is due to the fact that not all gateway software provides the same format of log files. Therefore, it could be recommended that a Data Mining approach in determining Web usage information within a proxy modelled subnet should begin with the investigation of the available data from the gateway in place to access the Web.

Future research will be conducted in an attempt to provide a form of data warehouse which can be used to extract useful information from the data which is already being collected.

References

Adomavicius, G. and Tuzhilin, A. 2001. Using Data Mining Methods to Build Customer Profiles. *Computer* 34(2):74–82.

Büchner, A.G., Anand, S.S., Mulvenna, M.D. and Hughes, J.G. 1998. Discovering Internet Marketing Intelligence through Web Log Mining. *SIGMOD Record* 27:54–61.

Cooley, R. 2000. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. University of Minnesota.

Cooley, R., Mobasher, B. and Srivastava, J. 1999. Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and information systems* 1(1):5–32.

Fulp, E.W. 2005. Optimization of Network Firewall Policies Using Ordered Sets and Directed Acyclical Graphs. In: *IEEE Internet Management Conference*.

Hand, D.J., Mannila, H. and Smyth, P. 2001. *Principles of data mining*. MIT Press.

Hazelhurst, S., Johnson, Y. and Sanders, I. 2011. An empirical analysis of the relationship between web usage and academic performance in undergraduate students. In: *Proceedings of the Annual Conference of the South African Computer Lecturer's Association*. Ballito, South Africa, 29–37.

Hochschild, J. 2009. Conducting Intensive Interviews and Elite Interviews. In: *Workshop on Interdisciplinary Standards for Systematic Qualitative Research*. National Science Foundation.

Hofgesang, P.I. and Kowalczyk, W. 2005. Analysing Clickstream Data : From Anomaly Detection to Visitor Profiling. In: *ECML/PKDD Discovery Challenge*.

Hossain, S.S., Rahman, S.M.M. and Kabir, F. 2012. Network proxy log mining : association rule based security and performance enhancement for proxy server. 49:9852–9857.

Hu, J. and Zhong, N. 2005. Clickstream Log Acquisition with Web Farming. In: *The 2005 IEEE/WIC/ACM International Conference*. 257–263.

Kajornboon, A.B. 2005. Using interviews as research instruments. *E-Journal for Research Teachers* 2(1).

Kosala, R. and Blockeel, H. 2000. Web Mining Research : A Survey. *ACM Sigkdd Explorations Newsletter* 2(1):1–15.

Luotonen, A. and Altis, K. 1994. World-Wide Web Proxies. *Computer Networks and ISDN systems* 27(2):147–154.

- Metzger, M.J., Flanagin, A.J. and Zwarun, L. 2003. College student Web use, perceptions of information credibility, and verification behavior. *Computers & Education* 41(3):271–290.
- Mikecz, R. 2012. Interviewing Elites: Addressing Methodological Issues. *Qualitative Inquiry* 18(6):482–493.
- Mythily, S., Qiu, S. and Winslow, M. 2008. Prevalence and correlates of excessive Internet use among youth in Singapore. *Annals of the Academy of Medicine, Singapore* 37(1):9–14.
- Pariser, E. 2011. *The filter bubble: what the internet is hiding from you*. Penguin UK.
- Rumbough, T.B. 2001. *Controversial uses of the internet by college students*. Bloomsburg University of Pennsylvania.
- Schafer, J. Ben, Konstan, J. and Riedl, J. 1999. Recommender Systems in E-Commerce. In: *Proceedings of the 1st ACM conference on Electronic commerce*. 158–166.
- Schumacher, P. and Morahan-Martin, J. 2001. Gender , Internet and computer attitudes and experiences. *Computers in human behavior* 17(1):95–110.
- Srivastava, J., Cooley, R., Deshpande, M. and Tan, P. 2000. Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data. *ACM SIGKDD Explorations Newsletter* 1(2):12–23.
- Suneetha, K.R. and Krishnamoorthi, R. 2009. Identifying User Behavior by Analyzing Web Server Access Log File. *IJCSNS International Journal of Computer Science and Network Security* 9(4):327–332.
- Suresh, R.M. and Padmajavalli, R. 2006. An Overview of Data Preprocessing in Data and Web Usage Mining. *1st International Conference on Digital Information Management*:193–198.
- Wessels, D. 2001. SQUID Frequently Asked Questions [Online]. Available <http://www.squid-cache.org/Doc/FAQ/FAQ-4>.

Acknowledgements

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the National Research Foundation.