A Data Warehouse Structure Design Methodology to Support the Efficient and Effective Analysis of Online Resource Usage Data

C. Ferreira

2012

Department of Computing Sciences

A Data Warehouse Structure Design Methodology to Support the Efficient and Effective Analysis of Online Resource Usage Data

By

Cornél Ferreira

Submitted in fulfilment of the requirements for the degree of Magister Scientae in the Faculty of Science at the Nelson Mandela Metropolitan University

Supervisors: Prof C Cilliers & Dr B Scholtz

Declaration

I, Cornél Ferreira, declare that the dissertation for the degree Magister Scientae is my own work and that it has not previously been submitted for assessment and completion of any postgraduate qualification to another university or for another qualification.

Cornél Ferreira

Acknowledgements

I would like to express my gratitude to my supervisors, Professor Charmain Cilliers and Doctor Brenda Scholtz, for their support throughout my work on this degree. Their invaluable advice, guidance and passion towards this research supported me significantly and kept me motivated. I am honoured by the opportunity that I was granted to be able to work with them and they helped me to grow as an individual in the field of computer science.

I would like to thank the NMMU Centre of Excellence and NMMU ICTS for funding this research. Finally, I would like to express my gratitude to the NMMU Department of Computing Sciences for the support I received throughout my degree.

I would like to acknowledge my parents for their endless support.

Abstract

The use of electronic services results in the generation of vast amounts of Online Resource Usage (ORU) data. ORU data typically consists of user login, printing and executed process information. The structure of this type of data restricts the ability of decision makers to effectively and efficiently analyse ORU data. A data warehouse (DW) structure is required which satisfies an organisation's information requirements. In order to design a DW structure a methodology is needed to provide a design template according to acknowledged practices.

The aim of this research was to primarily propose a methodology specifically for the design of a DW structure to support the efficient and effective analysis of ORU data. A variety of relevant DW structure design methodologies were investigated and a number of limitations were identified. These methodologies do not provide methodological support for metadata documentation, physical design and implementation. The most comprehensive methodology identified in the investigation was modified and the Adapted Triple-Driven DW Structure Design Methodology (ATDM) was proposed. The ATDM was successfully applied to the information and communication technology services (ICTS) department of the Nelson Mandela Metropolitan University as the case study for this research. The proposed ATDM consists of different phases which include a requirements analysis phase that was adapted from the identified comprehensive methodology. A physical design and an implementation phase were included in the ATDM.

The ATDM was successfully applied to the ICTS case study as a proof of concept. The application of the ATDM to ICTS resulted in the generation and documentation of semantic and technical metadata which describes the DW structure derived from the application of the ATDM at a logical and physical level respectively. The implementation phase was applied using the Microsoft SQL Server integrated tool to obtain an implemented DW structure for ICTS that is described by technical metadata at an implementation level.

This research has shown that the ATDM can be successfully applied to obtain an effective and efficient DW structure for analysing ORU data. The ATDM provides guidelines to develop a DW structure for ORU data and future research includes the generalisation of the ATDM to accommodate various domains and different data types.
Keywords: Data Warehouse Structure Design Methodology, Online Resource Usage Data

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AD	Active Directory
ATDM	Adapted Triple-Driven Data Warehouse Structure Design Methodology
BI	Business Intelligence
BIDS	Business Intelligence Development Studio
CIM	Computation Independent Model
DMB	Data Mart Bus
DSA	Data Staging Area
DW	Data Warehouse
EDW	Enterprise Data Warehouse
ETL	Extraction, Transformation and Loading
HMF	Hybrid Model Framework
ICT	Information and Communication Technology
ICTS	Information and Communication Technology Services
IDM	Independent Data Marts
IT	Information Technology
KPI	Key Performance Indicator
MSS	Microsoft SQL Server
NMMU	Nelson Mandela Metropolitan University
OLAP	Online Analytical Processing
ORU	Online Resource Usage Data
PE	Port Elizabeth
PIM	Platform Independent Model
PSM	Platform Specific Model
QVT	Query, View and Transformation
RDMS	Relational Database Management System
SQL	Structured Query Language
SSDE	SQL Server Database Engine
SSIS	SQL Server Integration Services
SSMS	SQL Server Management Studio
TDM	Triple-Driven Data Modelling
UML	Unified Modelling Language
XML	Extensible Markup Language

Chapter 1. Introduction

1.1 Background

Organisations provide an electronic infrastructure to employees to create an environment that allows employees to perform specific tasks. This infrastructure includes, amongst others, electronic communication, internet usage and devices to access electronic services. As a result of the usage of these services, vast amounts of data are generated on a daily basis through logging applications. This type of data is referred to as online resource usage (ORU) data and has the ability to provide an organisation's decision makers with information to support their decision making processes. In order to enable decision makers to effectively and efficiently make use of ORU data's ability to provide decision support, it is required that ORU data is efficiently analysed with the use of a business intelligence (BI) environment (Ariyachandra and Watson 2010).

A data warehouse (DW) is part of a collection of tools known as BI tools that together serve organisations as BI solutions to aid decision support initiatives (Sabherwal and Becerra-Fernandez 2009). DWs enable data processing and archival storage BI capabilities and the design and implementation of this technology are the initial steps towards an optimal analysis environment. Large companies have adopted DW technology as the standard practice to store integrated and centralised data extracted from different operational sources (Rob, Coronel and Crockett 2008). Thus data contained in a DW is housed separate from organisations' operational data stores because this enables analysts to execute complicated queries on the DW without conflicting with the transactional operations of operational systems (Inmon 2002). A DW supports organisations' analysis of data primarily by providing an infrastructure dedicated and optimised for this purpose (Golfarelli 2010). In contrast with traditional operational systems in which insertion and the updating of records are the primary objectives, DWs are subject-oriented to support the efficient analysis of information describing various subjects (Rob *et al.* 2008). The interest in DW technology has grown significantly since the 1980s given that with the use of this technology, trends and patterns in business processes can efficiently be elicited to aid planning towards establishing business strategies (Nemati *et al.* 2002; Jukic 2006).

A DW typically forms part of a larger DW architecture that describes the BI environment in which data is processed and maintained to support the analysis of processes of an organisation (Chaudhuri and Dayal 1997; Jukic 2006). The typical architecture of a DW (Figure 1-1) is represented by different components where data is passed from one component to the next after some critical operation is performed that prepares the data for the subsequent component (Simitsis and Theodoratos 2009; Ariyachandra and Watson 2010).

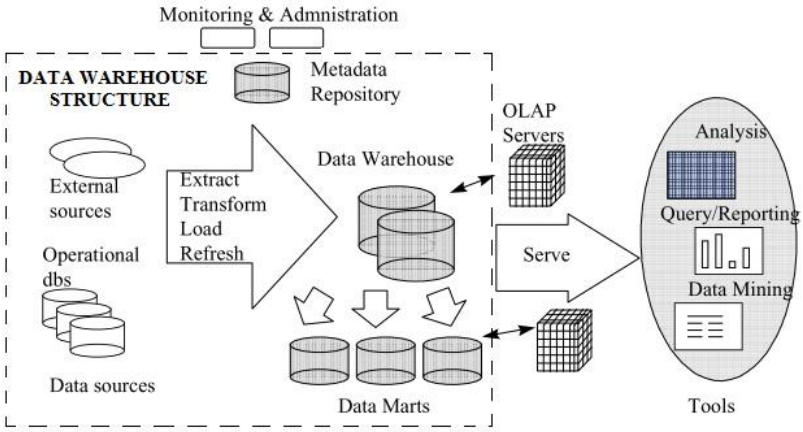


Figure 1-1: The Typical DW Architecture Adapted from (Chaudhuri and Dayal 1997).

The typical DW architecture (Figure 1-1) describes the following:

- Data sources;
- Tools for the extraction, transformation and loading (ETL) of data from data sources;
- DW and Data Marts that house integrated and cleansed copies of source data;
- Metadata repositories that store data describing the DW architecture and its processes;
- Online analytical processing (OLAP) servers; and
- Data access tools.

A DW structure is described in this research to consist of a subset of the components contained in a DW architecture, namely, data sources, DWs, data marts, ETL processes and metadata repositories (Figure 1-1). Due to varying requirements, several DW approaches have been proposed and are vigorously debated among experts (Sen and Sinha 2005). A DW approach refers to the manner in which data sources, ETL processes and DWs are structured in a DW structure (Jukic 2006). The selection of a particular DW approach influences the physical design of a DW structure.

Based on a selected DW approach, the design of a DW structure's DW may range from a single relational data structure, which represent an entire organisation's data, to a series of relational data marts, which are data models serving organisational departments (Kimball, Ross and Merz 2002). If a single relational DW is used to accommodate data for a DW structure, a single set of ETL processes is needed to populate the DW but, the use of a series of data marts need ETL processes to be designed for each data mart respectively.

One of the most important sets of processes used to ensure the successful functioning of a DW are sets of processes provided by ETL tools (Chaudhuri and Dayal 1997). ETL tools have the task of ensuring that relevant data is extracted from different data sources, cleansed, transformed and loaded to the allocated repository as part of the incremental updating of a DW to ensure that the data housed in a DW is consistent with the data contained in its respective data sources (Simitsis and Theodoratos 2009).

Maintaining and managing a DW may become a cumbersome duty for DW users if they are required to keep a detailed record of all processes executed on a DW (Vetterli, Vaduva and Staudt 2000). Metadata offers a solution to this problem and is a useful tool for both DW designers and users. Metadata is regularly referred to as data about data and forms an integral part of the management, administration and usage of a DW structure (Stöhr, Müller and Rahm 1999; Chaudhuri and Dayal 1997).

Metadata can be categorised into technical- and semantic metadata (Stöhr *et al.* 1999). Semantic metadata, also referred to as business metadata, describes the logical mapping between business models and data sources such as a logical DW design and the mapping of the concepts it represents to data sources. Semantic metadata is primarily used to allow users, who are not familiar with technical DW concepts, to understand the DW architecture and its contents. Technical metadata represents data that is typically used by designers which need a technical representation of DW structure processes such as the mapping between data sources and the DW at a physical and implementation level.

OLAP refers to a set of tools which allows users to perform more complex analysis of data which include data mining, operations research and advanced reporting functionalities (Rob *et al.* 2008). OLAP tools extract information from dimensionally modelled data repositories which are hosted on dedicated OLAP servers and in turn sources data from the DW or data marts (Chaudhuri and Dayal 1997). Although dimensional modelling is synonymous with OLAP data storage, dimensional modelling can also be implemented using traditional relational technology (Kimball *et al.* 2002). These relational dimensionally modelled data repositories are referred to as star- or snowflake schemas and are most commonly used as compared to dimensional OLAP repositories.

A lack of methodologies for the design of a DW structure to support the efficient and effective analysis of ORU data prompted this research. Although a lack of DW structure design methodologies for ORU data has been identified in literature, several DW structure methodologies exist (Golfarelli 2010). Existing DW structure design methodologies are used to elicit information requirements and data characteristics which are used to derive a logical design of a DW structure. Although a logical design of a DW structure can be derived using existing DW structure design methodologies, they fail to provide methodological support for the physical design and implementation of a system which are needed as part of a system development methodology (Satzinger, Jackson and Burd 2005).

The relevance of this project (Section 1.2) initiates a problem statement (Section 1.3) and a thesis statement (Section 1.4) in which the problem is identified and a hypothesis is stated respectively. The hypothesis proposes a methodology for the design of a DW structure as the solution to the problem identified. Research objectives (Section 1.5) are formulated which will be addressed by answering particular research questions (Section 1.6). The research questions will be addressed with an applicable research methodology (Section 1.7). This project is constrained to a particular scope (Section 1.8). This chapter concludes with a structure of the dissertation (Section 1.9).

1.2 Project Relevance

The Nelson Mandela Metropolitan University (NMMU) has a large network of devices that is used by its students and staff members on several university campuses. This network of devices is managed by NMMU's information and communications technology services (ICTS) and ORU data is recorded on dedicated servers.

This data is generated by individuals logging into a device and interacting with it. ORU data consists of application usage data, printing information, log-on and log-off information, each of which can be traced to a specific computer in a laboratory or office on a campus. ORU data serves as a foundation from which informed decisions are made regarding the usage of electronic resources in the university.

In any particular year, ICTS generates ad hoc queries issued on the logged ORU data that is extracted from different data sources and analysed to support the generation of usage reports. These queries can assist with decisions regarding computer resource needs at different facilities on different campuses. The current procedure for the analysis of ORU data consists of extracting data relevant to a particular report or question by issuing ad hoc queries on operational ORU data sources. This typical ICTS process leads towards queries that have to be executed on data sources which are typically experiencing large amounts of operational processes daily. Through issuing queries on the operational sources a contention for power emerges that results in longer query processing times and a decrease in overall productivity.

The operational ORU data sources are heterogeneous in location and design which complicates the integration of data from these operational sources. Reports that are frequently generated have resulted in the creation of a temporal data store that source the report generation processes. This temporal data store does not always apply to the need for ad hoc queries, thus resulting in users referring back to the operational sources to begin the design process for individual queries. ICTS need to constantly design infrastructure to obtain ad hoc query results which is a time intensive and redundant process.

A dedicated DW structure offers a solution to the restrictions experienced by ICTS in efficiently and effectively analysing ORU data. In order to develop a DW structure that will be able to accommodate the information requirements of the NMMU ICTS department, their information requirements need to be analysed to aid the design of an appropriate DW structure (Winter and Strauch 2003). In the generic representation of a DW architecture (Figure 1-1) data is extracted, transformed and loaded into a central data repository offering its users a cleansed and integrated copy of operational data and this can be achieved using an ETL tool (Chaudhuri and Dayal 1997). This strategy will allow information and communication technology (ICT) users to avoid the need to constantly refer back to operational sources when an ad hoc query needs to be addressed.

This research can be summarised as a need for a methodology for the design of a DW structure for ORU data exists based on the lack of methodologies identified in literature. Various DW structure design methodologies have been proposed, but no single DW structure design methodology that is appropriate for use with ORU data could be identified. This research proposes to investigate different DW structure design methodologies with the goal of proposing a consolidated DW structure design methodology specifically for ORU data.

1.3 Problem Statement

The problem statement of this research is as follows:

A DW structure design methodology to support the efficient and effective analysis of ORU data is required. Although several DW structure design methodologies have been proposed (Golfarelli 2010), no consolidated DW structure design methodology could be identified that supports the efficient and effective analysis of ORU data.

1.4 Thesis Statement

The thesis statement for this research is as follows:

> A DW structure design methodology can be proposed to support the efficient and effective analysis of ORU data.

To address the lack of DW structure design methodologies for ORU data that was identified in literature, existing DW structure design methodologies need to be investigated. Based on this investigation a comprehensive DW structure design methodology will be proposed specifically to support the effective and efficient analysis of ORU data.

1.5 Research Objectives

The main objective for this study is:

> To propose and evaluate a DW structure design methodology that supports the efficient and effective analysis of ORU data.

The main objective can be divided into the following secondary objectives:

1. To investigate DW structure design methodologies
2. To propose a DW structure design methodology for ORU data
3. To determine information requirements and data characteristics of the NMMU ICTS using the proposed DW structure design methodology
4. To design a DW structure for the NMMU ICTS using the proposed methodology
5. To implement the derived DW structure design for the NMMU ICTS using the Microsoft SQL Server (MSS) integrated tool
6. To evaluate the derived DW structure implementation for efficiency and effectiveness

1.6 Research Questions

The main research question for this study is:

What is an appropriate methodology to support the design of a DW structure for the efficient and effective analysis of ORU data?

The main research question for this study can be divided into the following sub-questions:

1. What DW structure design methodologies exist?
2. What is an appropriate DW structure design methodology for ORU data?
3. How can DW information requirements and data characteristics of the NMMU ICTS be determined using the proposed DW structure design methodology?
4. How can a DW structure be designed for the NMMU ICTS using the proposed DW structure design methodology?
5. How can a DW structure design be implemented using the MSS integrated tool?
6. How efficient and effective is the derived DW structure implementation?

1.7 Research Method

This research will use a combination of two research methods, namely, the interpretivism and positivism research philosophies. Interprevistic philosophies are based upon the notion that in order to understand a specific phenomenon, the motive and meaning of the phenomenon should be investigated (Abbott 2010). In contrast to the interpretivism philosophy, positivists argue that phenomenon can only be explained by means of natural sciences which typically utilise experiments to obtain quantitative measures.

The first phase of this research will incorporate the interpretivism research philosophy in which a DW structure design methodology to support the efficient and effective analysis of ORU data is proposed to address the need for a methodology which was identified in literature. A DW structure design methodology will be proposed based on an investigation of existing DW structure design methodologies. The second phase of the research will incorporate the positivism philosophy in which the proposed DW structure design methodology is applied to the NMMU ICTS to obtain a proof of concept. The DW structure that is derived from the application of the methodology will be subjected to a set of experiments to obtain empirical results which can be used to prove the creditability of specific facts, namely, the efficiency and effectiveness of the DW structure design (Saunders, Lewis and Thornhill 2009).

Table 1-1 represents the research questions (Section 1.6) which will be addressed by this research and how each question will be operationalised. Each question relates to a specific research objective (Section 1.5) and the process of addressing each question will be documented in a specific chapter that is indicated in Table 1-1.

Table 1-1: Research Methodology

Research Question	Method	Objective	Chapter
RQ 1) What DW structure design methodologies exist?	Literature Study	O 1	2
RQ 2) What is an appropriate DW structure design methodology for ORU data?	Literature Study	O2	2
RQ 3) How can the DW information requirements and data characteristics of the NMMU ICTS be elicited using the proposed DW structure design methodology?	Interviews Questionnaires ORU Data Analysis	O 3	3
RQ 4) How can a DW structure be designed for the NMMU ICTS using the proposed DW structure design methodology?	Literature Study Design	O 4	4
RQ 5) How can a DW structure design be implemented using the MSS integrated tool?	Implementation	O 5	5
RQ 6) How efficient and effective is the derived DW structure implementation?	Experimentation	O 6	6

Figure 1-2 summarises the process that will be undertaken to address the research questions presented in Table 1-1. A literature study will be conducted to investigate DW structure design methodologies from which a DW structure design methodology will be proposed. The DW structure design methodology will be applied to the NMMU ICTS to evaluate the methodology for a proof of concept. Information requirements for the NMMU ICTS are accumulated using interviews which will be formulated based on the proposed DW structure design methodology. Similarly the characteristics of the ORU source data will be analysed in order to design a DW structure as described by the proposed methodology. A prototype of the derived DW structure design will be implemented as a proof of concept in order to allow the evaluation of the derived DW structure design for effectiveness and efficiency. Criteria for evaluating the efficiency and effectiveness of a DW structure will be investigated in order to create a set of experiments from which efficiency and effectiveness results can be observed.

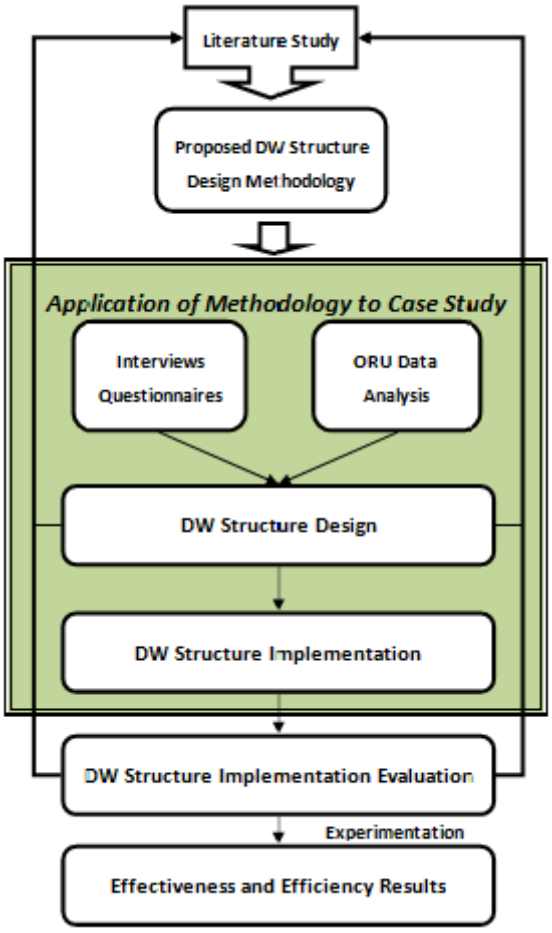


Figure 1-2: Process of Addressing Research Questions

1.8 Scope and Constraints

The scope of this research is limited to the proposal of a DW structure design methodology. The methodology is required to provide methodological support for the design of a DW structure that consists of data sources, DW, ETL processes and metadata (Figure 1-3). OLAP and related OLAP analysis tools will not be considered as part of the components for which the DW structure design methodology is proposed. The integrated tool that will be used for the implementation of the DW structure that is derived from the application of the proposed DW structure design methodology is Microsoft SQL Server (MSS), since this is the tool that is used at the NMMU ICTS.

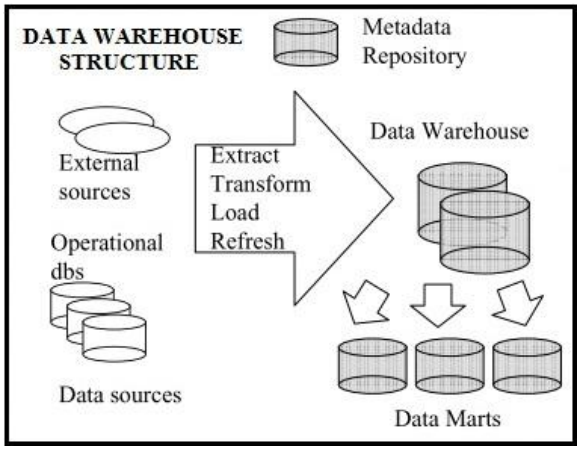


Figure 1-3: Research Scope and Constraints

1.9 Conclusion and Dissertation Structure

A limitation was identified in literature that no DW structure methodology could be identified that can be applied to derive a DW structure design that supports the effective and efficient analysis of ORU data. Figure 1-4 represents the structure of the dissertation. DW structure design methodologies will be investigated in order to derive a DW structure design methodology to support the effective and efficient analysis of ORU data (Chapter 2). The methodology will be evaluated as a proof of concept by applying the methodology to a case study. The application of the methodology consists of requirements analysis (Chapter 3), deriving a logical and physical DW structure design (Chapter 4) and implementing the derived physical DW structure design using an integrated tool (Chapter 5).

The implementation of the derived DW structure physical design will be subjected to a set of experiments in order to enable the observation of results from which the effectiveness and efficiency of the derived DW structure design can be determined (Chapter 6). Based on the conclusions drawn from this research recommendations and future research are discussed (Chapter 7).

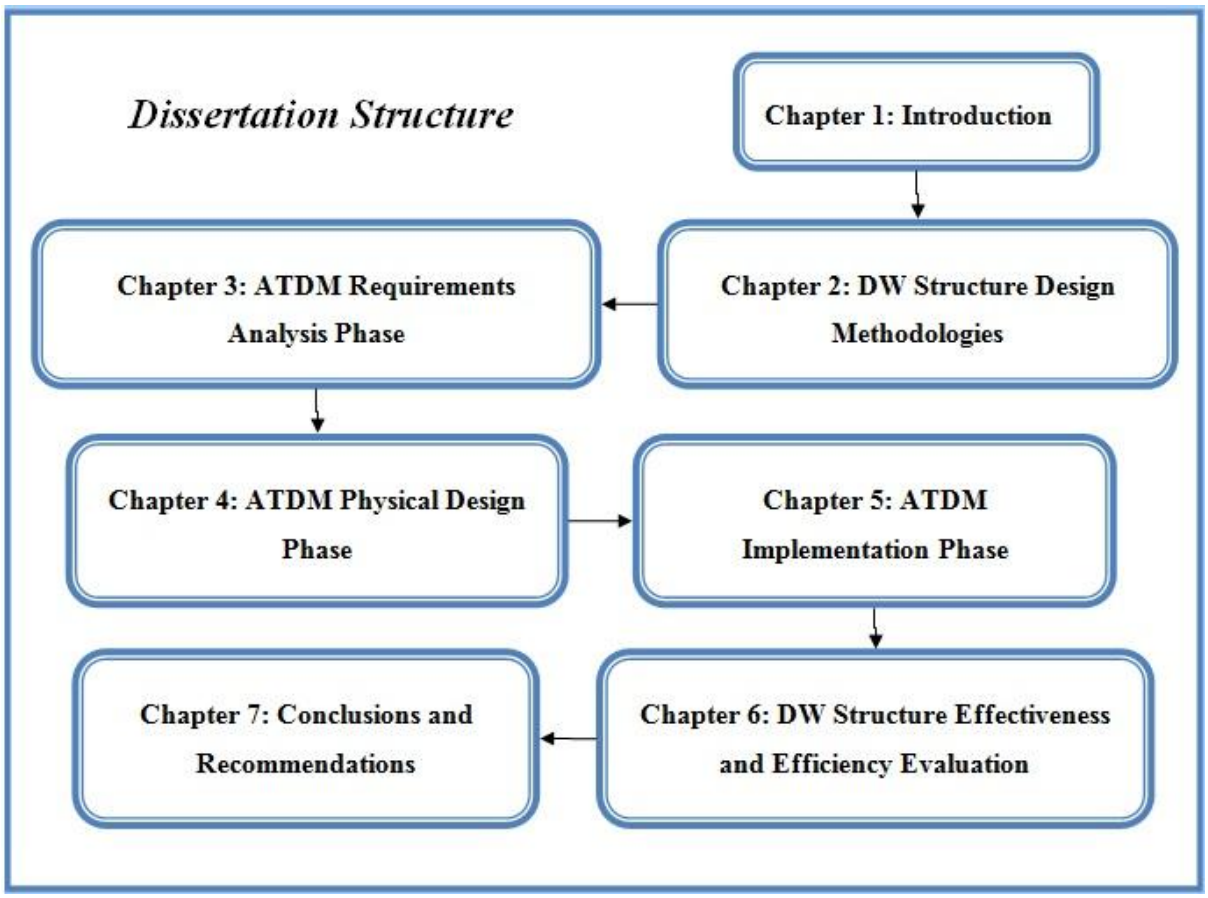


Figure 1-4: The Structure of the Dissertation for this Research.

Chapter 2. DW Structure Design Methodologies

2.1 Introduction

The management of organisations is typically interested in the flexible analysis of business processes in order to monitor if these processes contribute towards the organisation achieving predefined goals (Bruckner, List and Schiefer 2001). In order to ensure that this analysis of business processes is achieved effectively and efficiently, a dedicated data warehouse (DW) structure is required to provide organisations with analysis capabilities (Winter and Strauch 2004). The design of a DW structure is heavily dependent on the information requirements and the information supply of an organisation, and it is important that both the requirements and supply are well understood by designers (Bruckner *et al.* 2001; Golfarelli 2010; Winter and Strauch 2004). The designers should take into consideration that business processes are viewed differently by different individuals in an organisation. The management of an organisation would typically view the requirements of a DW structure in terms of vision statements and goals. In contrast, the users of a DW structure view DW structure requirements as various processes that have to be executed by this technology. It is thus important that different views of DW structure requirements are represented by a single DW structure design that describes all processes required to enable the analysis of operational source data.

The first research question that was presented for this research (Section 1.6) requires the investigation of different DW structure design methodologies. Since the information requirements of an organisation define the design of its DW structure, enterprise DW requirements are described, in terms of different levels of requirements abstraction (Section 2.2). To be able to gather and utilise these requirements to derive a DW structure design, different DW structure design methodologies (Section 2.3) are described. A comprehensive DW structure design methodology, the ATDM is proposed (Section 2.4) based on a comparison of the described DW structure design methodologies, in which limitations and differences are highlighted (Section 2.3). The DW structure design methodology is proposed in order to address the second research question for this research (Section 1.6).

2.2 Enterprise Data Warehouse Requirements

An organisation's stakeholders describe their goals, intentions and direction of their enterprise in terms of business views which can be accomplished, based on a set of business requirements to which a DW structure should conform (Bruckner *et al.* 2001; Golfarelli 2010). These business requirements are often ambiguous to DW designers who require a more detailed description of the DW structure they have to design. This ambiguity demands that business requirements have to be refined to lower levels of abstraction in their description to obtain an enterprise-wide description of the DW requirements of an organisation. Different levels of abstraction for enterprise DW requirements (Bruckner *et al.* 2001) should be considered during the requirements analysis phase and are represented by the hierarchy of enterprise DW requirements (Figure 2-1):

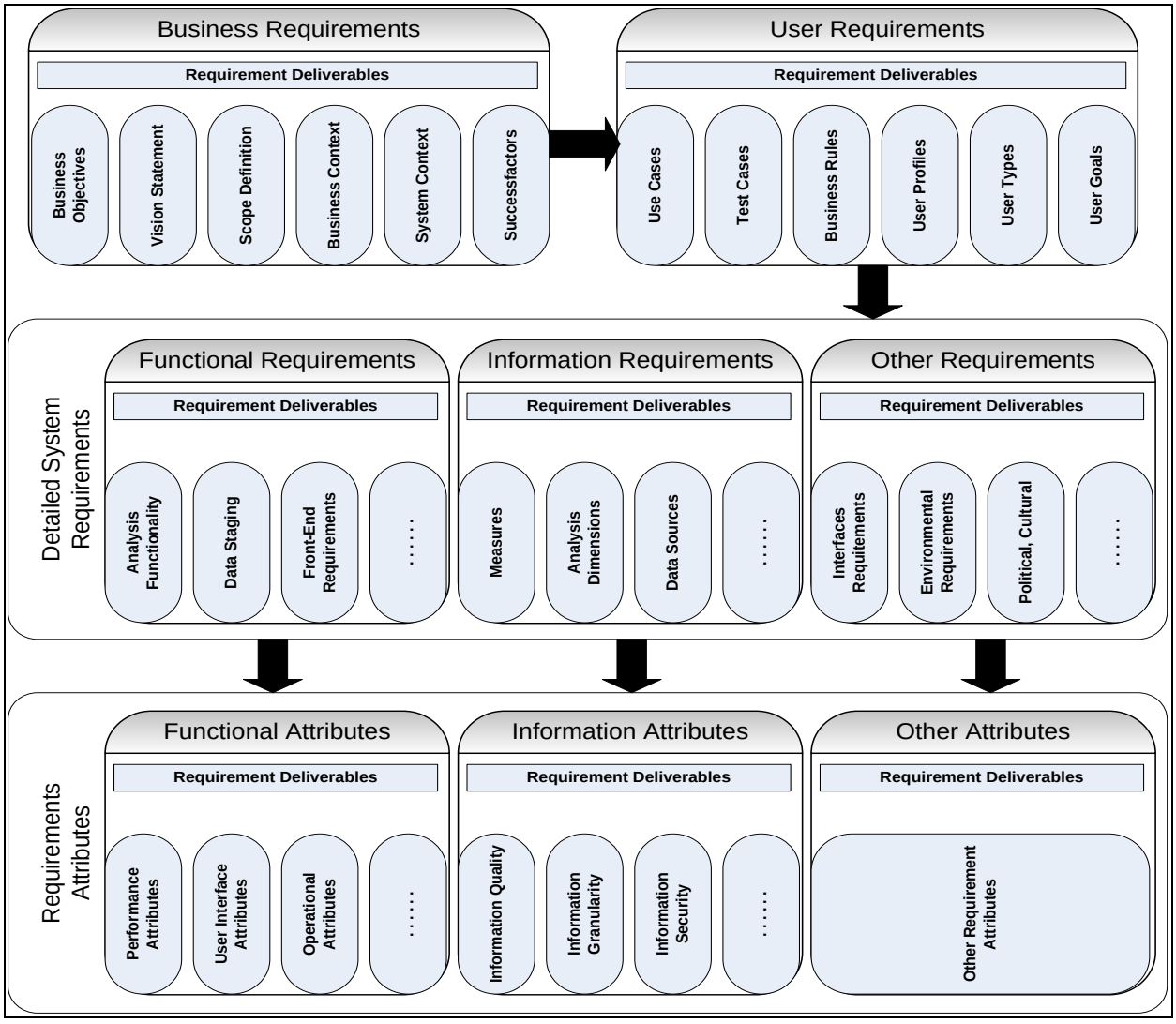


Figure 2-1: Enterprise DW Requirements Hierarchy (Bruckner *et al.* 2001)**.**

Business requirements are upper most in the hierarchy of enterprise DW requirements (Section 2.2.1). These requirements are described in more detail by the next set of enterprise DW requirements in the hierarchy, namely, user requirements (Section 2.2.2) which are in turn described by an increased level of detail by detailed system requirements (Section 2.2.3). Requirements attributes (Section 2.2.4) represents the most detailed enterprise DW requirements in the enterprise DW requirements hierarchy.

2.2.1 Business Requirements

Business requirements of an organisation are the highest level of abstraction of enterprise DW requirements (Bruckner *et al.* 2001). These requirements describe the expected benefits of using the system for decision making in terms of vision statements, objectives, success factors and business opportunities (Golfarelli 2010). The profiles of typical users and their potential benefit to the organisation when using the DW structure are identified in business requirements.

2.2.2 User Requirements

User requirements are more detailed than business requirements since these requirements describe the activities which the DW users should be able to perform on the DW structure (Bruckner *et al.* 2001; Golfarelli 2010). User requirements can be seen as the tasks that must be carried out to satisfy business requirements. User requirements are elicited from individuals who use the system more regularly than an organisation's management and provide a more in-depth description to the designers about the actual tasks, activities and non-functional requirements. Although user requirements are more low level than business requirements, they should be aligned with the business requirements obtained at a higher level of abstraction at the end of the data gathering process to avoid user disappointment in the DW structure design (Winter and Strauch 2004).

2.2.3 Detailed System Requirements

The detailed system requirements are derived from the user requirements (Bruckner *et al.* 2001). Detailed system requirements describe the detailed functional and information requirements. Functional requirements describe the activities and processes which should be supported by the system so that the users can complete a set of tasks to allow management to achieve a specific goal (Golfarelli 2010).

Functional requirements describe processes such as the functionality that supports analysis, ETL tasks and functionality to allow users to interact with the system (Bruckner *et al.* 2001). The information requirements of a DW structure describe the information that must be supported by the DW structure. This description of the information includes specification of the quality, data source, measures to be considered and how the data should be manipulated into a specified format suitable for the application of required processes. In addition to functional and information requirements, at this level of abstraction, additional requirements can also be considered which may include requirements for interface, environment, legal and cultural aspects depending on the problem domain.

2.2.4 Requirement Attributes

Users and designers of a DW structure require that tasks be performed by the DW structure, which conform to specific conditions considered as important to users and which are defined by requirement attributes (Bruckner *et al.* 2001). The detailed system requirements describe what the DW structure is required to do. Requirement attributes supplement these requirements by describing the characteristics, policies or standards to which the detailed system requirements should conform. These attributes describe, for example, how quickly processes have to be executed, the maximum or minimum capacity of data repositories, and the quality and granularity of data.

2.3 Methodologies for DW Structure Design

The design of a DW structure is largely dependent on the requirements analysis technique that is applied in which enterprise DW requirements (Section 2.2) are elicited and analysed to aid DW structure design (Golfarelli 2010). These techniques include supply-driven requirements analysis (Section 2.3.1) in which the information supply of an organisation is analysed in order to derive a DW structure design. Demand-driven requirements analysis techniques (Section 2.3.2) are techniques by which information requirements of stakeholders are analysed to derive a DW structure design. The supply- and demand-driven requirements analysis techniques both have disadvantages and advantages (Section 2.3.3). DW structure design methodologies are described (Section 2.3.4) which use a combination of supply- and demand-driven requirements analysis techniques in providing methodological support for DW structure design.

2.3.1 Supply-Driven Requirements Analysis

Supply-driven requirements analysis can be described as a bottom-up technique where requirements are accumulated starting with a detailed analysis of the data sources of DW structure (Giorgini, Rizzi and Garzetti 2005; Golfarelli 2010). The data sources for a DW structure typically consist of various operational systems which contain heterogeneous data with dissimilar structures and semantics. During the application of this technique, data sources are investigated to aid the development of a design for the DW structure, thus functional requirements are derived from the information supply (Guo *et al.* 2006). The only inputs from the users of the DW structure are the selection of the data sources, which will be considered in the design of the DW structure, and not functional requirements of users (Golfarelli 2010). This technique is considered to be stable and requires fewer resources for requirements gathering, but it provides little support to the DW designer to determine the facts and dimensions of the dimensional model which are significant for users of the DW structure (Song *et al.* 2001).

2.3.2 Demand-Driven Requirements Analysis

Demand-driven requirements analysis aids the design of the DW structure by firstly analysing the information requirements of the stakeholders of the project before the information supply is considered (Mazon and Trujillo 2009). Demand-driven requirements analysis can be divided into two different categories, namely, goal-driven and user-driven requirements analysis.

Goal-driven requirements analysis is based on organisational business goals set by the management of organisations that must be satisfied with support from a DW structure. Goal-driven requirements analysis involves conducting interviews with the top management of an organisation and is considered to be a top-down technique for data warehouse requirements analysis (Golfarelli 2010). This technique requires that different visions of the organisational processes be analysed and merged to derive a consistent business model.

When the standard business model is derived, it has to be translated into several relevant indicators that must be catered for in the design of a DW structure (Golfarelli 2010). These indicators are known as key performance indicators (KPIs) which are quantifiable measures of some business process in an organisation (Golfarelli 2010; Guo *et al.* 2006).

The derivation of KPIs is considered to be an important process used to ensure that the business processes for which a DW is being designed conforms to the business goals and visions of the organisation (Guo *et al.* 2006). The level of top management engagement in the design process will determine the applicability of goal-driven requirements analysis. Goal-driven requirements analysis is also expensive in terms of expenditure of time and money (Golfarelli 2010).

User-driven requirements analysis refers to the tasks that must be executed on a DW structure by its users. User-driven requirements analysis is a bottom-up requirements analysis technique in which the information requirements of the users of the DW structure are gathered (Winter and Strauch 2003). Once the information requirements of the users are accumulated, they are integrated and standardised to obtain a unique design for a DW structure (Golfarelli 2010). This requirements analysis technique is strongly driven by the participation of DW structure users. It is highly regarded by users because of the high level of user involvement (Guo *et al.* 2006). The user-driven technique requires significant guidance by the use of management skills because of the heterogeneous perspective of users that must be integrated to obtain an unique consolidated schema (Golfarelli 2010).

2.3.3 Comparison of Supply- and Demand-Driven Requirements Analysis

The different techniques for gathering requirements have their own set of strengths and risks (Table 2-1). The use of any particular technique is determined by the environment in which the particular DW structure is being designed (Golfarelli 2010). The supply- and user-driven requirements analysis techniques are bottom-up whilst the goal-driven requirements analysis technique is top-down. The supply-driven technique is the least resource expensive due to limited user involvement. This technique, however, may result in the incorporation of data into the DW structure's design which is not relevant to the organisation's goals and user tasks since the needs of users and management were not included into the DW structure design process (Winter and Strauch 2004). This unnecessary incorporation can lead to the waste of both time and money in managing data not relevant to the needs for which the DW structure is being designed. In contrast with the supply-driven technique, the two demand-driven requirements analysis techniques significantly incorporate the DW structure users and top management of an organisation in the process of accumulating information requirements respectively (Golfarelli 2010).

Table 2-1: Comparison of Supply- and Demand-driven Requirements Analysis Techniques (Golfarelli 2010)

		Demand-Driven	
	Supply-Driven	**User-Driven**	**Goal-Driven**
Basic approach	Bottom-up	Bottom-up	Top-Down
Users involvement	Low: DB administrators	High: Business users	High: Top Management
Constraints	Existence of a reconciled data level	Business users must have a good knowledge of the processes and organisation of the company	Willingness of top management to participate in the design process
Strengths	The availability of data is ensured	Raise the acceptance of the system	Maximise the probability of a correct identification of the relevant KPI's.
Risks	The multidimensional schemata do not fit business user requirements.	Quick obsolescence of the multidimensional schemata due to changes of the business users.	Difficulties in being supported by top management and in translating the business strategy into quantifiable KPI's
Targeting organisational level	Operational and Tactical	Depends on the level of the interviewed users, typically tactical	Strategic and tactical
Skills of project staff	DW designers	Moderators; DW designers	Moderators; Economist; DW designers
Risk of obsolescence	Low	High	Low
Number of source systems	Low	Moderate	High
Cost	Low	High	High

The high involvement of users and management in the user- and goal-driven techniques ensures that users' acceptance of the DW structure is increased and important KPIs are identified, whereas stakeholders are reluctant to participate in the supply-driven technique which in turn ensure the availability of data (Golfarelli 2010; Gardner 1998). The goal- and user-driven techniques are both considered to incur higher costs in the development of a DW structure (Golfarelli 2010). This increase in costs can be attributed to the larger number of skills and the number of sources that have to be consulted to aid the design of the DW structure.

User- and goal-driven requirements analysis techniques require more input from the DW structure users than the supply-driven technique and this can lead to greater acceptance of a DW structure (Giorgini *et al.* 2005; Golfarelli 2010). The user- and goal-driven techniques may, however, be constrained by users that may not have appropriate knowledge of the processes of an organisation nor of top management which is not willing to participate in the requirements analysis processes.

The supply-driven, goal-driven and user-driven requirements analysis techniques are usually applied in isolation from each other (Guo *et al.* 2006; Golfarelli 2010). This usually results in a design for a DW structure that does not satisfy all the requirements of a user's, management, operational and technical perspectives. In order to derive the best design for a DW, it is necessary that these isolated requirements analysis techniques be merged so that each can complement the short comings of the other (Guo *et al.* 2006). The requirements analysis technique which consists of a supply- and demand-driven requirements analysis technique is referred to as a mixed-driven requirements analysis technique. DW structure design methodologies, which apply a mixed-driven requirements analysis technique, are described (Section 2.3.4).

2.3.4 Mixed-Driven Requirements Analysis DW Structure Design Methodologies

Several DW structure design methodologies have been proposed which apply a mixed-driven requirements analysis technique. The DW structure design methodologies which have been proposed are:

- The Hybrid Model Framework (Section 2.3.4.1);
- The GRAnD Requirements Analysis Approach (Section 2.3.4.2); and
- The Triple-Driven Data Modelling Methodology in Data Warehousing (Section 2.3.4.3)

Although these DW structure design methodologies differ in the steps the use to elicit DW structure requirements, they agree on mapping the information requirements of users and management to the information supply of the DW structure being designed.

2.3.4.1 The Hybrid Model Framework

Mazon and Trujillo (2009) present the Hybrid Model Framework (HMF) (Figure 2-2) that is specifically designed to aid the development of dimensional schemata for a DW structure (Song *et al.* 2001). This framework begins with an investigation into the information requirements of an organisation to obtain an information requirements model as a computation independent model (CIM). The CIM model is used as the basis for the design of an initial conceptual model that is a platform independent model (PIM) representing the dimensional requirements of the organisation's decision makers for whom it is designed (Mazon and Trujillo 2009).

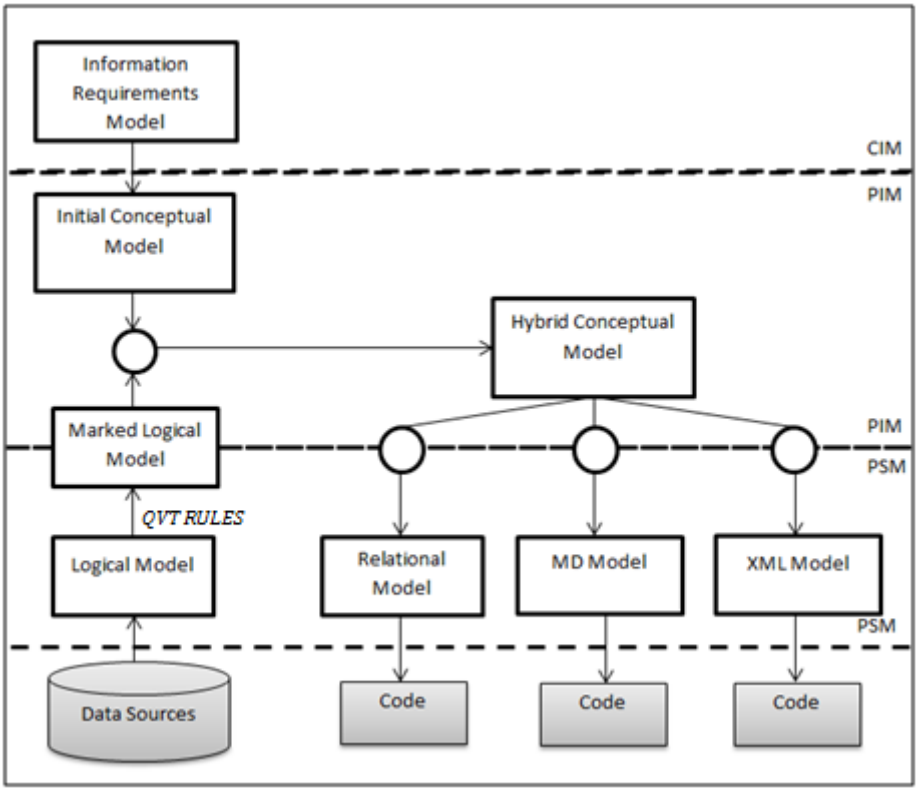


Figure 2-2: Hybrid Model Framework (HMF) for DW Structure Design (Mazon and Trujillo 2009).

The conceptual model at this stage only reflects the needs of decision makers dimensionally, without the model being justified by the organisation's available data sources. In order to ensure that the dimensional model is able to satisfy information requirements, it has to be mapped to the data sources of the organisation.

The data sources of the organisation are investigated in the application of the HMF to derive a logical model of the organisation that represents organisational data (Mazon and Trujillo 2009). This logical model of organisation data sources is subjected to query, view and transformation (QVT) rules to identify concepts in the logical model which can be considered as possible facts, measures and dimensions. The application of QVT rules to the logical model results in the marked logical model. The initial derived dimensional conceptual model is mapped to the marked logical model of the data sources using the QVT rules. This mapping between these models ensures that information requirements are mapped to information supply in to obtain the platform independent hybrid conceptual model.

The next step in the HMF requires that the derived hybrid conceptual model is translated to be platform specific by deriving a platform specific model (PSM) that represents the concepts of the PIM in terms of specific database technologies identified for implementation. These database technologies consist of relational, multidimensional or extensible markup language (XML) technologies.

2.3.4.2 The GRAnD Requirements Analysis Approach

The goal-oriented approach to requirements analysis in DWs (GRAnD) (Figure 2-3) proposed by Giorgini et al. (2008) can be applied using a demand-driven requirements analysis technique but when the schemata of the DW structure's sources are well defined and available, the information supply significantly aids the conceptual design of a DW structure.

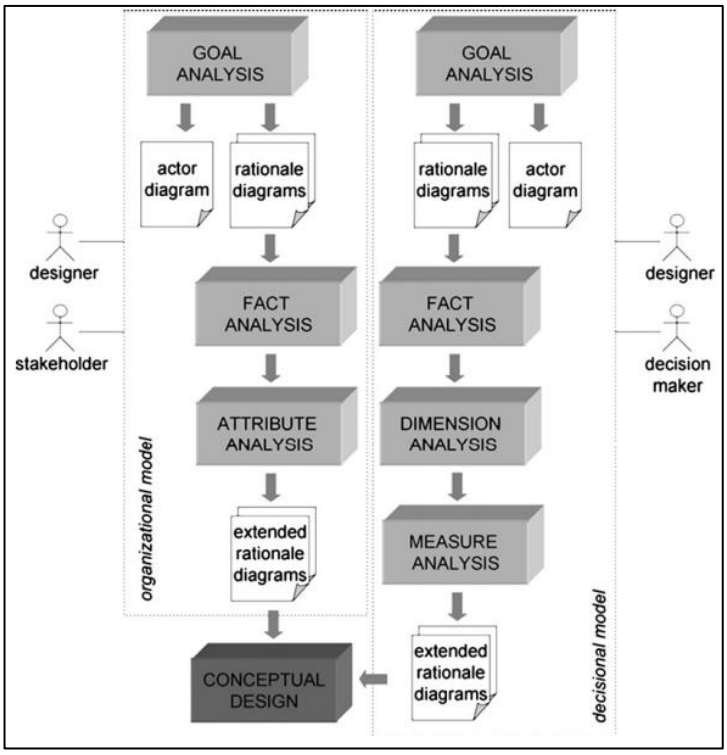


Figure 2-3: The GRAnD Requirements Analysis Approach (Giorgini et al., 2008).

A goal-driven requirements analysis technique for the design of a DW structure is used in the GRAnD approach in terms of an organisational and a decisional perspective. The goal of this approach is to model the organisational environment in which the DW structure will be used and also in a decisional environment which specifies the DW structure's functional and non-functional requirements (Giorgini *et al.* 2008). The organisational model of the GRAnD refers to the modelling of the information requirements of an organisation's management to the decisional modelling of the information requirements of organisation's decision makers.

In both modelling phases of this approach, various high-level actors and their dependence on each other to accomplish some set of goals are identified. These actors are then divided into sub-actors to obtain more detail, with the goal of identifying responsibilities attributed to a specific individual regarding a specific goal. Each individual goal is then analysed to obtain interdependencies between different goals. This process continues iteratively to obtain detailed rationale diagrams which describe all goals. These are further supplemented with the addition of descriptive facts, attributes, dimensions and measures. The next step in the process of using GRAnD is the mapping of the conceptual model representing the derived information requirements to the information supply (Giorgini *et al.* 2008). During this process, when the available source schemata are considered, the conceptual model is refined to better accommodate both the information supply and the user's requirements.

2.3.4.3 The Triple Driven Data Modelling Methodology

A DW structure design methodology using a combination of a supply- and demand-driven requirements analysis technique is proposed by Guo et al. (2006). This methodology is known as the Triple-Driven Data Modelling (TDM) methodology (Figure 2-4). The methodology begins with the goal-driven phase in which a goal-driven requirements analysis technique is applied. A supply-driven and a user-driven phase are followed in parallel with each other based on the organisational goals derived in the goal-driven phase. These phases apply a supply- and user-driven requirements analysis technique respectively. The goal-driven phase of this methodology starts with the investigation of organisational goals to develop a corporate strategy to elicit the organisation's goals and what measures the organisation will take to achieve these goals (Figure 2-4 Step 1.1) (Guo *et al.* 2006). This step is followed by the identification of the primary business fields (Figure 2-4 Step 1.2) which have to be supported by the DW structure to comply with the derived corporate strategy derived. The corporate strategy and business fields aid the identification of KPIs (Figure 2-4 Step 1.3) that serve as an indication of the performance of specific business fields relative to organisational goals. The KPIs that were identified and defined are the main deliverables of the goal-driven phase of this methodology (Figure 2-4 Step 1.6). Similarly to the GRAnD and HMF methodologies, users that will be using the DW structure are identified (Figure 2-4 Step 1.4) and are classified, based on their roles and responsibilities in terms of the corporate strategy (Guo *et al.* 2006; Mazon and Trujillo 2009; Giorgini *et al.* 2008).

Based on the primary business fields, subject areas are identified (Figure 2-4 Step 1.5) with the goal of arranging data, at a high level, around the primary business fields that are used to carry out various organisational processes (Guo *et al.* 2006). The identification of the subject areas and target users initiates the start of the supply- and user-driven phases of this methodology respectively (Guo *et al.* 2006).

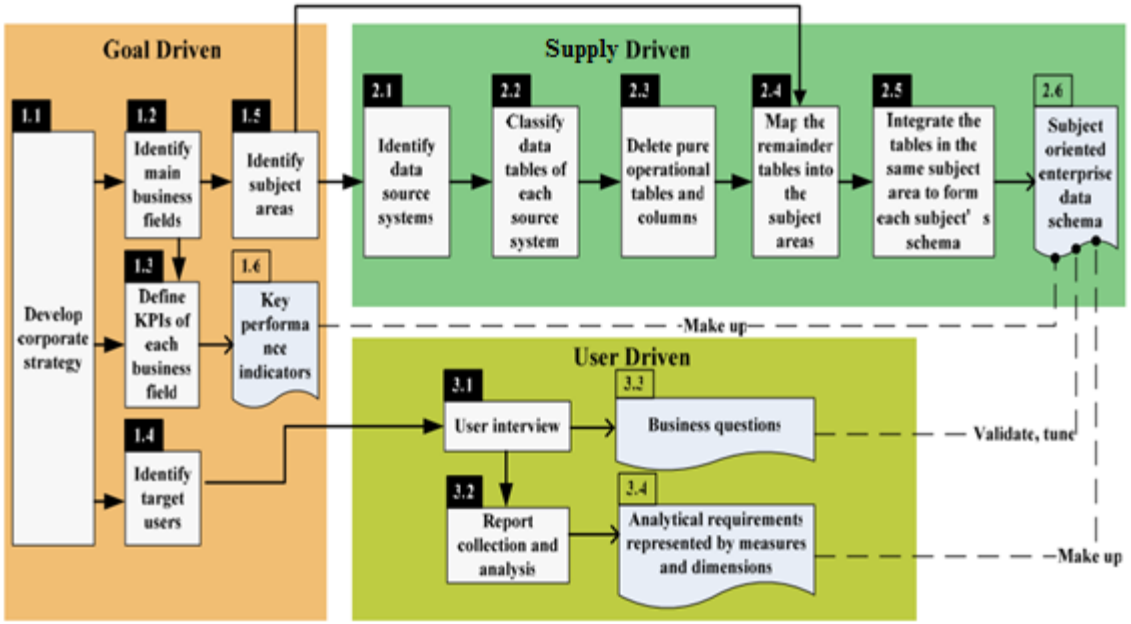


Figure 2-4: Triple-Driven Data Modelling (TDM) Methodology in Data Warehousing (Adapted from Guo et al. 2006).

The supply-driven phase continues with identifying DW source systems (Figure 2-4 Step 2.1) relevant to the subject areas identified, classification of relevant source system tables (Figure 2-4 Step 2.2) and the deletion of table attributes (Figure 2-4 Step 2.3) which are regarded as irrelevant for analysis. The remaining tables are mapped to the relevant subject areas (Figure 2-4 Step 2.4) based on the data they represent. Each table mapped to the different subject areas is finally integrated (Figure 2-4 Step 2.5) to obtain a data schema representative of each subject area identified.

Concurrently, the identified target users of the DW structure are interviewed (Figure 2-4 Step 3.1) to derive business questions (Figure 2-4 Step 3.3), identify different reports (Figure 2-4, Step 3.2) and derive analytical requirements for the proposed DW structure (Guo *et al.* 2006). Together with the business questions obtained (Figure 2-4 Step 3.3), measures and dimensions describing subject areas (Figure 2-4, Step 3.4) form the deliverables of the user-driven phase of this methodology.

The final step in the methodology is the utilisation of the KPIs, analytical requirements (Figure 2-4 Step 3.4) and business questions to validate and fine-tune the subject-oriented data schema to obtain the final logical data model of the DW structure's DW. The supply-driven approach for TDM is guided by its goal-driven approach (Mazon and Trujillo 2009; Guo *et al.* 2006).

The deliverable of the TDM is the subject oriented enterprise data schema (Figure 2-4 Step 2.6). The subject oriented data schema is derived from the subject oriented data schema of each subject (Figure 2-4, Step 2.5) which are adapted to accommodate the derived KPIs (Figure 2-4, Step 1.6), analytical requirements (Figure 2-4, Step 3.4) and business questions (Figure 2-4, Step 3.3) (Guo *et al.* 2006). The derived subject oriented data schema represents a logical design of a DW structure's DW.

2.3.4.4 Comparison of the DW Structure Design Methodologies

The GRAnD, HMF and the TDM DW structure design methodologies have several similarities and differences in the application of steps towards the design of a DW structure. Each of these methodologies consists of goal-, user- and supply-driven requirements analysis components. The GRAnD approach begins with the identification of relevant actors in the DW structure to derive the interdependencies between actors documented in rational diagrams. In contrast with the GRAnD approach, the HMF does not rely heavily on modelling the interdependencies between the various actors at first (Giorgini *et al.* 2008; Mazon and Trujillo 2009).

The HMF begins with identifying the strategic requirements followed by the lower level decisional requirements used to enable the satisfaction of the strategic requirements identified in the previous step (Mazon and Trujillo 2009). Based upon the strategic and decisional requirements identified, information requirements are derived to form the CIM. Only after the CIM has been derived are the roles and goals of the decision makers considered to form the PIM. At this stage of the process, the GRAnD and HMF are similar, in terms of the models derived from the goals of an organisation (Giorgini *et al.* 2008; Mazon and Trujillo 2009). The primary difference between the GRAnD and HMF is the manner in which the information supply is considered. The GRAnD approach significantly uses the information supply to guide the refinement of the conceptual design of the solution, but only if the information supply is well defined.

The HMF identifies logical source schemata without referencing the derived conceptual model through the application of the QVT rules. Thus the supply- and demand-driven requirements analysis technique in the HMF are mostly applied independently (Golfarelli 2010). This methodology however does map the information requirements with the information supply when the HMF hybrid conceptual model is derived (Figure 2-2) (Mazon and Trujillo 2009). The HMF hybrid conceptual model can be seen as an improved version of the platform independent model, since, during the process when demand is mapped with supply, adjustments are made to the conceptual model to accommodate source schemata. Although the conceptual model can be successfully restructured to accommodate information requirements and information supply, relevant information may be excluded in the derivation of the marked logical model.

The application of the TDM, in contrast with the GRAnD and HMF, does not result in initial conceptual models which are based only on organisational information requirements (Guo *et al.* 2006). This model defines an organisation's operational tasks in terms of subjects which guide a process that results in the design of a data schema that describes the available and refined data that represents each subject identified. In the HMF, the supply-driven component is independently investigated without reference to goals and user tasks, and the marked logical model may not contain data that is needed to satisfy information requirements (Mazon and Trujillo 2009). The GRAnD approach however only considers referring to the information supply for validating the conceptual model after the users' perspectives have been incorporated into the conceptual design (Giorgini *et al.* 2008). In addition, the TDM incorporates user analytical requirements to refine the design of a DW structure as well as to derive business questions from users which can be used to evaluate the effectiveness of the DW structure that is being designed.

2.3.4.5 Expert Data Warehouse Requirements

The described DW design methodologies heavily rely on the analysis of information requirements to aid the design of a DW structure. Expert DW requirements have been identified (Table 2-2), which provide methodological support for the critical phase of requirements analysis (Winter and Strauch 2004).

Table 2-2: Data Warehouse Expert Requirements (Winter and Strauch 2004).

Expert DW Requirement	TDM Supported	GRAnD Supported	HMF Supported
Multi-Level Hierarchies	■		
Information Map	■	■	■
"As Is" Information Provision	■	■	■
"To Be" Information State	■	■	■
Assigning Priorities			
Homogenisation of Concepts			
Development Phase Integration	■	■	■
Meta Data Documentation	■	■	■
Total Number of Supported Expert Requirements	6	5	5

Since mixed-driven requirements analysis techniques are preferred to the exclusive use of either a supply- or demand-driven requirements analysis technique (Section 2.3.3), only DW structure design methodologies which consist of a mixed-driven requirements analysis technique are compared based on the identified expert DW requirements. The multi-level hierarchy requirement specifies that information demand should be mapped to information supply at an aggregate level early in the application of a DW structure design methodology to ensure that information requirements could be derived from the information supply (Winter and Strauch 2004). The multi-level requirement is addressed only by the TDM (Section 2.3.4.4).

The mapping of the information supply with the information requirements of a particular organisation is documented in the information map which is created with the "As Is" information provision and "To Be" information state as input (Winter and Strauch 2004). The "As Is" information provision requirement refers to the specification of data sources. The "To Be" information state requirements refers to a planned state in which overlaps of information demand and information supply is analysed to meet information requirements.

In the GRAnD approach, the information supply is mapped to demand after a conceptual model is created that describes information demand at different levels of aggregation (Giorgini *et al.* 2008). During this process, the creation of the information map, in which the "As Is" information state and information demand are combined to form the "To Be" information state, is applicable.

Since both the TDM and HMF methodologies also require the information demand to be mapped to the information supply, the creation of the information map is applicable to both (Mazon and Trujillo 2009; Guo *et al.* 2006). Since all the described methodologies (Section 2.3) satisfy the requirement of the creation of an information map, these methodologies satisfy the "As Is" information provision and the "To Be" information state expert requirements.

The development of the phase integration expert requirement requires a methodology that supports the creation of DW structure models which are derived and directly in line with models created to represent information requirements (Winter and Strauch 2004). All of the methodologies support the creation of models, which are used to represent information requirements, and which are mapped to information supply. The ability of these methodologies to support the creation of these models ensures that the different phases of the respective methodologies are well integrated to guarantee consistency between the "To Be" information state, information supply and information demand. Based on the integration of the different phases of the respective methodologies, all of the methodologies satisfy the development integration expert requirement.

DW projects are considered to be significantly large projects which are developed over long periods of time and require a large number of resources for their development (Golfarelli 2010; Giorgini *et al.* 2008). In order to ensure that the most important requirements are addressed first, priorities should be assigned to the identified DW requirements (Winter and Strauch 2004). The assignment of priorities is an expert requirement that is not exhibited in any of the methodologies identified in this study and can be a valuable step to ensure that unsatisfied requirements are assigned the least amount of resources.

The homogenisation of concepts expert DW requirement needs to be addressed in a DW structure design methodology to ensure that the semantics of data at different levels of aggregation are homogenised to enable the easy integration of data from various sources (Winter and Strauch 2004). The homogenisation of concepts expert requirement is, in addition to the assignment of priorities, not exhibited in each of the described methodologies.

The metadata documentation expert requirement specifies that models, which are created in the application of a DW structure design methodology, are documented in such a way that metadata is easily transferable to the selected metadata management system (Winter and Strauch 2004). The application of the described methodologies results in the creation of models and the logical mapping between these models and data sources (Sections 2.3.4.1 – 3). These models and their respective mappings to their data sources represent semantic metadata (Section 1.1). If the described DW structure design methodologies are applied in such a way that the semantic metadata derived from their application is easily transferrable to a technical and implementation level, this expert requirement is applicable to any of the methodologies described.

Six of the eight expert DW requirements were found to be satisfied by the TDM methodology compared to the five expert DW requirements which are satisfied by the GRAnD and HMF (Table 2-2). Since the majority of the expert DW requirements are satisfied by the TDM methodology, it can be concluded that this DW structure design methodology can provide a more comprehensive requirements analysis phase for a DW structure design methodology. The only shortcomings which were identified were that the TDM methodology does not satisfy the assigning of priorities and homogenisation of concepts expert DW requirements.

A methodology for the design of a software product needs to consist of an analysis, design and implementation phase (Section 1.1). Although the TDM methodology consists of phases in which information supply and information requirements are analysed to derive a logical design of a DW structure, it fails to describe a physical design phase and implementation phase.

2.4 Adapted Triple-Driven DW Structure Design Methodology

The TDM methodology was identified to provide a comprehensive requirements analysis phase for a DW structure design methodology, but shortcomings were identified (Section 2.3.4.5). Based on these shortcomings a comprehensive DW structure design methodology is proposed based on the TDM methodology, namely, the adapted triple-driven DW structure design methodology (ATDM) (Figure 2-5).

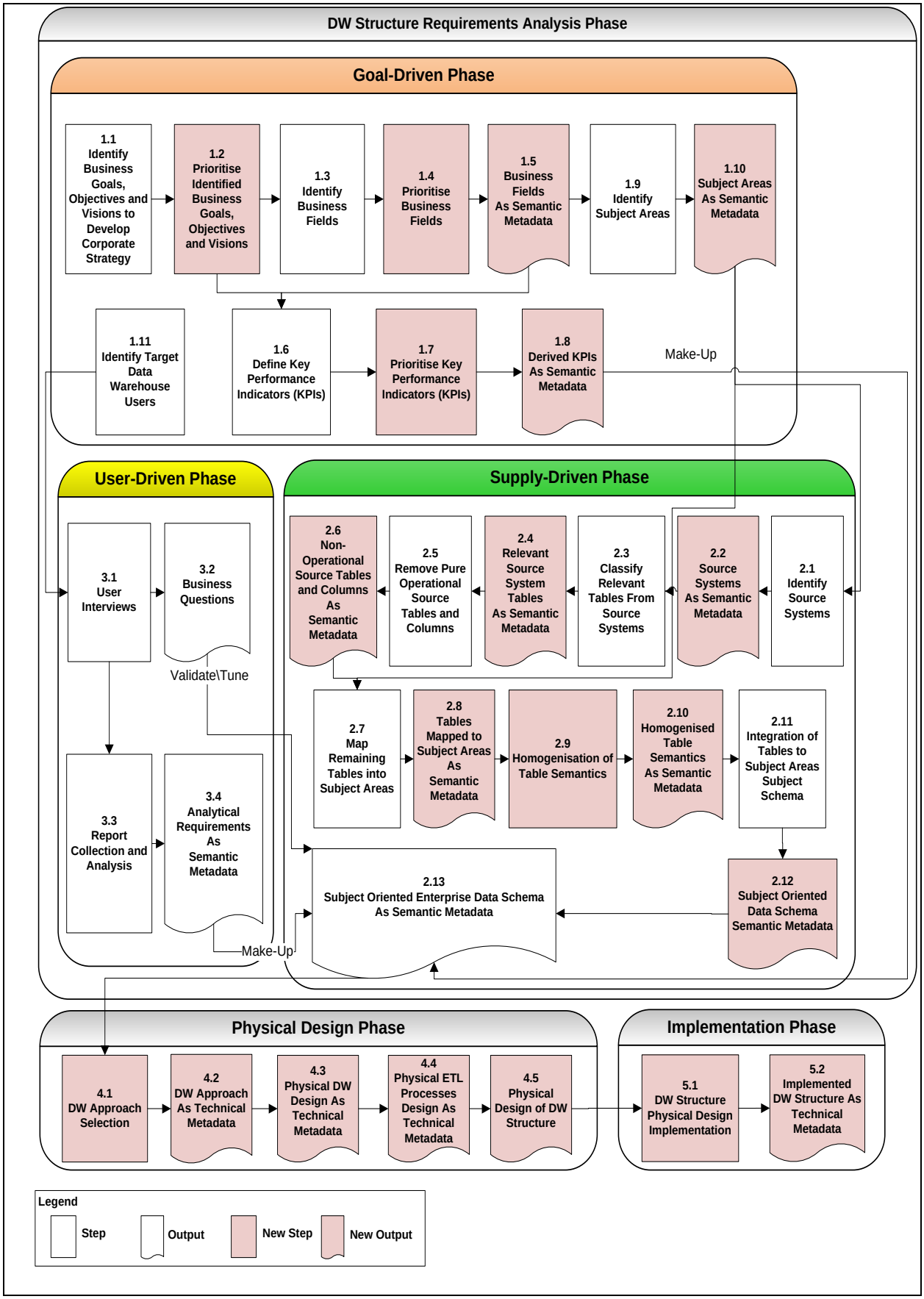


Figure 2-5: The Adapted Triple-Driven Data Warehouse Structure Design Methodology (ATDM)

The shortcomings of the TDM were addressed by introducing additional steps to the methodology as processes and deliverables, which are highlighted in red. The ATDM consists of three primary phases, namely, a requirements analysis, physical design and implementation phase. The requirements analysis phase consists of three secondary phases, namely, a goal-driven, user-driven and supply-driven phase which represent the existing phases of the TDM methodology (Section 2.3.4.3) which are used to derive a logical design of a DW structure. The physical design and implementation phase were included in the methodology in order to provide methodological support for the physical design of a DW structure and the implementation of the physical design using an integrated tool.

Three steps, which were addedto the goal-driven phase of TDM methodology in ATDM, are aimed towards prioritising the deliverables of a goal-driven process (Figure 2-5, Steps 1.2, 1.4 & 1.7). These steps were included to address the "assignment of priorities" expert DW requirement (Section 2.3.4.5). The business visions, objectives and goals which are identified (Figure 2-5, Step 1.1) need to be assigned each an appropriate priority value that are indicative to the DW designer of their importance compared to each other. Similarly, priority values are required to be assigned to each identified business field (Figure 2-5, Step 1.3) and KPI (Figure 2-5, Step 1.6). In addition to the prioritisation steps which were added to the ATDM's requirements analysis phase, three deliverables of goal-driven processes were added (Figure 2-5, Steps 1.5, 1.8 & 1.10) to explicitly indicate that a set of semantic metadata (Section 1.1) is required to be documented based on the metadata documentation expert DW requirement (Section 2.3.4.5). These deliverables are the business fields (Figure 2-5, Step 1.5), KPIs (Figure 2-5, Step 1.8) and subject areas (Figure 2-5, Step 1.10) which need to be documented as sets of semantic metadata and used in subsequent steps in the ATDM.

The supply-driven phase of the TDM methodology (Section 2.3.4.3) was adapted to explicitly indicate sets of semantic metadata which are required to be documented to describe the deliverables of the supply-driven phase's processes (Figure 2-5, Steps 2.2, 2.4, 2.6, 2.8, 2.10 & 2.12). To address the homogenisation of semantics expert DW requirement (Section 2.3.4.5), an additional process was added to the supply-driven phase of the TDM methodology (Figure 2-5, Step 2.9).

Once the subject oriented enterprise data schema is derived (Figure 2-5, Step 2.13), as the logical design of a DW structure, the physical design phase of ATDM begins with a process in which an appropriate DW approach (Section 1.1) needs to be selected (Figure 2-5, Step 4.1). The DW approach that is selected (Figure 2-5, Step 4.2) and the derived subject oriented enterprise data schema (Figure 2-5, Step 2.13) are used to derive the physical design of a DW structure (Figure 2-5, Step 4.3). Once the physical DW is derived, the physical design of the ETL processes are derived (Figure 2-5, Step 4.4). These processes are derived based on the physical design of the DW and the semantic metadata documented in the requirements analysis phase, since the semantic metadata describes the logical mapping between the data sources and the logical design of the DW. Once the physical design of the DW structure's ETL processes are derived, the completed physical design of the DW structure (Figure 2-5, Step 4.5) has been derived. The deliverables derived from the application of the physical design phase are documented as technical metadata (Section 1.1).

The final phase of the ATDM is the implementation phase in which the derived physical design of the DW structure is implemented (Figure 2-5, Step 5.1) to obtain the derived physical design of the DW structure at an implementation level (Figure 2-5, Step 5.2). The DW structure implementation needs to be represented in terms of technical metadata at an implementation level as allowed by the integrated tool used for implementing the physical design (Section 1.1).

2.5 Conclusion

This chapter described the investigation of DW structure design methodologies. The application of requirements analysis techniques were identified to be a critical part of the design of a DW structure (Section 2.3). Based on the importance of requirements analysis, different requirements analysis techniques were described (Sections 2.3.1 & 2.3.2) and compared (Section 2.3.3). Based on the comparison of requirements analysis techniques, it was found that using supply- and demand-driven requirements techniques mutually, the shortcomings of the respective requirements analysis techniques are eliminated. Based on this finding DW structure design methodologies which apply a mixed-driven requirements analysis technique were described and compared (Section 2.3.4).

In order to further distinguish the advantages and disadvantages of the respective mixed-driven DW structure design methodologies they were compared against a set of expert DW requirements (Section 2.3.4.5). Based on this comparison it was found that the TDM methodology complies with the most expert DW requirements and it was concluded that the TDM methodology provides a more comprehensive approach for DW structure requirements analysis. Although the TDM methodology satisfied the most expert DW requirements the TDM methodology fails to describe the physical design and implementation of a DW structure (Section 2.3.4.5). The metadata documentation expert DW requirement was found to be applicable to the TDM methodology (Section 2.3.4.5), but the TDM methodology fails to explicitly indicate the deliverables which are required to be documented as metadata.

Based on the limitations of the TDM methodology it was adapted to include steps to satisfy all expert DW requirements (Section 2.3.4.5) and that explicitly indicate the documentation of metadata as part of the proposed ATDM (Section 2.4). The ATDM also consists of a physical design phase and implementation phase in which the physical design and implementation thereof are addressed. The physical design and implementation phase of ATDM also explicitly indicates which deliverable of these respective phases are required to be documented as technical metadata at a physical and implementation level (Figure 2-5).

The NMMU ICTS requires a DW structure that allows them to effectively and efficiently address ad hoc questions (Section 1.2). The ATDM is applied to the NMMU ICTS case study to enable the evaluation of the proposed methodology. The evaluation of the methodology consists of two objectives namely:

1. To evaluate the ATDM for a proof of concept (Chapters 3, 4 and 5); and
2. To evaluate the effectiveness and efficiency of the DW structure derived from the application of the methodology (Chapter 6).

The first evaluation objective of the ATDM is addressed through the application of the requirements analysis phase (Chapter 3) on the NMMU ICTS case study to derive the logical design of the ICTS DW structure.

The first objective is further addressed through the application of the physical design phase (Chapter 4) and the implementation phase (Chapter 5) of the ATDM in which the physical design is derived and implemented respectively. The second evaluation objective for the ATDM is to evaluate the effectiveness and efficiency of the DW structure that is derived from the application of the methodology (Chapter 6).

Chapter 3. ATDM Requirements Analysis Phase

3.1 Introduction

The ATDM was proposed (Section 2.4) to address the limitations of existing methodologies in the design of DW structures which support the effective and efficient analysis of ORU data. The ATDM (Section 2.4) needs to be evaluated to determine its applicability to be used as a DW structure design methodology (Section 2.5). As part of the proof of concept evaluation of the ATDM, the requirements analysis phase of the methodology is applied to the NMMU ICTS case study to address the third research question (Section 1.6). This research question requires the information requirements and data characteristics of ICTS to be elicited. The ATDM's requirements analysis phase consists of three phases, namely, the goal-driven (Section 3.2), user-driven (Section 3.3) and supply-driven phase (Section 3.4). The aim of the application of these phases of the methodology is to elicit the information requirements and data characteristics to obtain a representative logical design of a DW structure for the NMMU ICTS.

3.2 Goal-Driven Phase

The application of the goal-driven phase of ATDM initiates with the identification of target DW structure users (Figure 3-1, Step 1.11) and the identification of business goals, objectives and visions (Figure 3-1, Step 1.1), to develop the ICTS corporate strategy, concurrently.

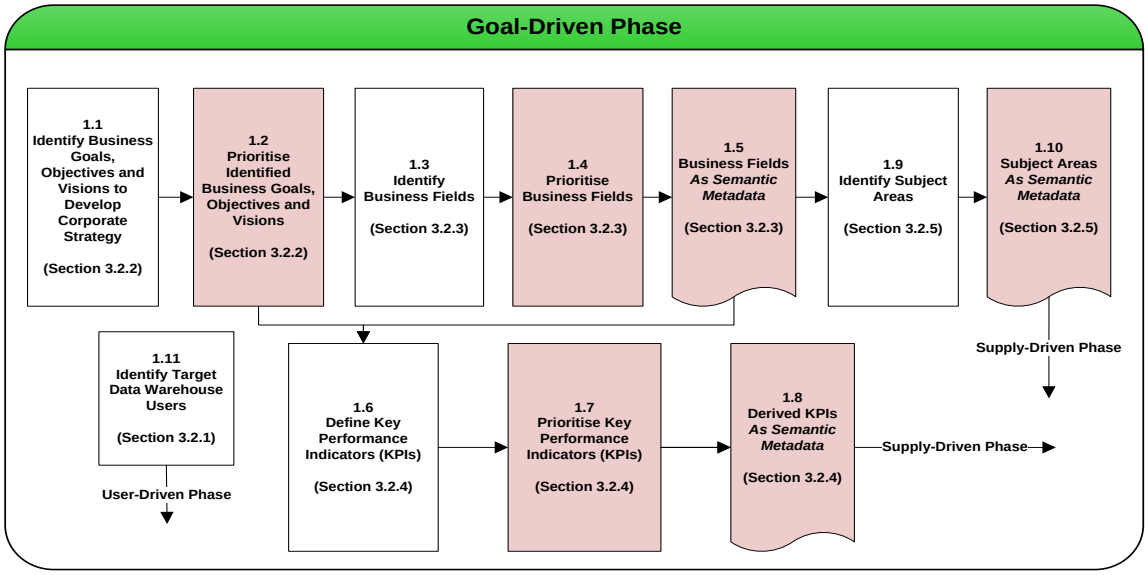


Figure 3-1: The Goal-Driven Phase of the ATDM.

3.2.1 Identification of Target Data Warehouse Users

Three individuals were identified as the target users of the DW structure in this step of the goal-driven phase (Figure 3-1, Step 1.11). The primary user that was identified is the facility planner of ICTS. The facility planner is involved in ICTS financial planning, allocation of resources and the management of the NMMU ICT infrastructure. Two secondary users of the DW structure were identified who are responsible for the generation of reports used to describe the state of ICTS resources.

3.2.2 Identification of Business Goals, Visions and Objectives

Three business objectives for the proposed DW structure of ICTS were identified (Figure 3-1, Step 1.1) by interviewing the ICTS facility planner, namely:

1. To ensure that ICT resources are used effectively and optimally;
2. To support the effective allocation of the annual budget to maintain the NMMU ICT infrastructure and to acquire additional resources to meet the needs of users which are derived from the ORU data; and
3. To identify scenarios in the current resource usage that can influence the future ICT resource allocation strategies of the university.

The identified business objectives were found to be similar, since all three business objectives require the analysis of resource usage to identify scenarios which could influence the ICT strategies of the NMMU. Thus a single objective is derived that encompasses these objectives namely:

> To identify trends in user online resource usage that can influence the short and long term ICT strategies of the university.

The subsequent step in the goal-driven phase requires the prioritisation of identified business goals, objectives and visions (Figure 3-1, Step 1.2). Since a single business objective has been identified for the NMMU ICTS case study the prioritisation step is not applicable.

3.2.3 Identification and Prioritisation of ICTS Business Fields

Three business fields (Table 3-1) were identified (Figure 3-1, Step 1.3) and prioritised (Figure 3-1, Step 1.4). These prioritised business fields represent the first set of semantic metadata that is documented (Figure 3-1, Step 1.5) in the application of the ATDM's requirements analysis phase.

Computer Management is the management of all the computer devices which enable users to connect to the NMMU ICTS infrastructure. The Computer Management business field was assigned a priority value of "1" that indicates that this business field has the highest priority compared to the other business fields which were identified.

Printing Management is the management of all printers which are allocated for use by an NMMU ICTS infrastructure user and was assigned a priority value of "2". The Computer Process Management business field is the management of all processes used to accommodate the needs of the NMMU users when using a computer and was assigned the lowest priority value of "3".

Table 3-1: ICTS Business Fields and Assigned Priorities (Figure 3-1, Step 1.5).

Business Field	Priority
Computer Management	1
Printing Management	2
Computer Process Management	3

3.2.4 Identification and Prioritisation of Key Performance Indicators

Two KPIs have been identified (Table 3-2) for each business field (Figure 3-1, Step 1.6) and these were then prioritised (Figure 3-1, Step 1.7) based on their importance as described by ICTS. The KPI with the highest priority was assigned a value of "1". The KPIs which were identified and prioritised constitute the main deliverable of the goal-driven phase of the ATDM's requirements analysis phase and represent a set of semantic metadata (Figure 3-1, Step 1.8) that is documented to be used in the application of the supply-driven phase.

Table 3-2: ICTS Key Performance Indicators (Figure 3-1, Step 1.8).

Business Field	KPI	Description	KPI Priority
1. Computer Management	1.1 Computer Usage Rate	The usage percentage of targeted computers allocated in the university.	1
	1.2 User Computer Usage Rate	The computer usage percentage of targeted users.	2
2. Printing Management	1.3 Printing Usage Rate	The usage percentage of targeted printers allocated at NMMU.	3
	1.4 User Printing Usage Rate	The printing usage percentage of targeted users.	4
3. Computer Process Management	1.5 Computer Process Usage Rate	The usage percentage of targeted processes on allocated computers.	5
	1.6 User Computer Process Usage Rate	The computer process usage percentage of targeted users.	6

3.2.5 Identification of ICTS Subject Areas

The identification of ICTS subject areas (Figure 3-1, Step 1.9) involves the identification of information that is required to support the analysis of specific subjects of interest within a business field (Section 3.2.3). The identified subject areas represent a set of semantic metadata (Table 3-3) that is derived and documented (Figure 3-1, Step 1.10) to aid the application of the supply-driven phase. Computer information and user information are required by each business field to support its analysis. These information sets are required to allow analysts to describe the computer and the user who was involved in some activity carried out on the NMMU ICT infrastructure. Activities include logging into a computer, printing a document with a printer and executing a computer process on a computer. In addition to computer and user information, other information sets are required to support the analysis of the respective business field subjects.

Table 3-3: ICTS Business Field Subject Areas (Figure 3-1, Step 1.10).

	Information Required to Describe Business Field Subjects				
Business Field	Computer Information	User Information	Login Information	Printing Information	Computer Process Information
Computer Management	■	■	■		
Printing Management	■	■		■	
Computer Process Management	■	■			■

The computer management business field requires user login information, such as login time, when users log into a computer. The printing management business field requires printing information which describes the use of the printer and details of any printing job executed on a particular printer. The details describing a printing job may include information such as the execution time, number of copies, number of pages of a particular colour and the paper size that were used in a particular printing job. The only additional information set that is required to support the analysis of a subject in the computer process management business field is computer process information that describes the processes executed, such as the name of the process and the time it was executed.

3.3. User-Driven Phase

The user-driven phase (Figure 3-2) of the requirements analysis phase of ATDM (Figure 2-5) as applied at the NMMU requires the investigation and accumulation of the information requirements of the target DW users of ICTS. The investigation involved interviewing the target DW users with the goal of identifying analytical requirements (Section 3.3.3) and the derivation of business questions (Section 3.3.1). In addition, the collection and analysis of reports are also required (Section 3.3.2) to derive analytical requirements which can supplement the analytical requirements identified in user interviews.

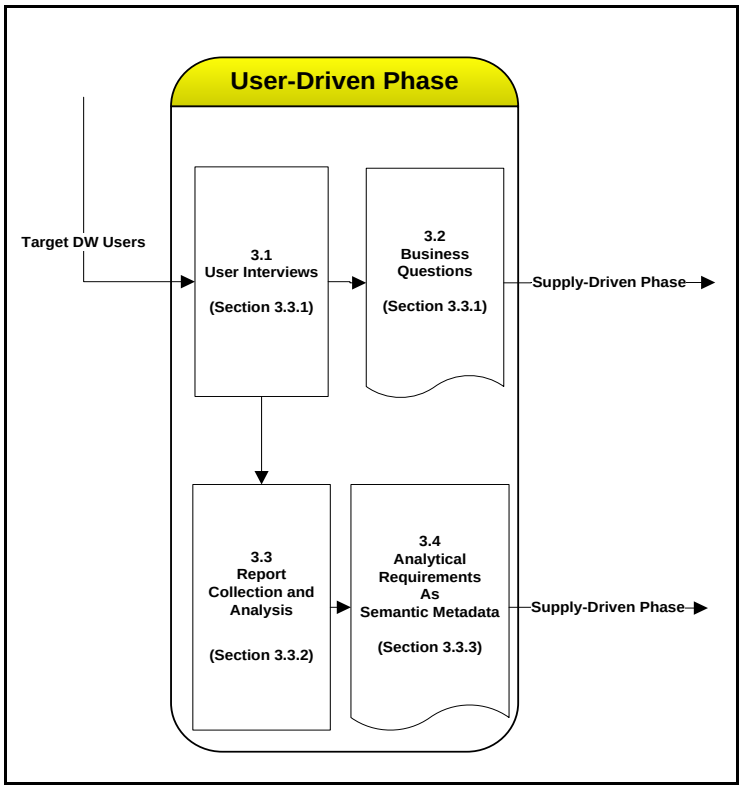


Figure 3-2: The User-Driven Phase of ATDM.

3.3.1 ICTS Business Questions

The interviews with the target DW users (Figure 3-2, Step 3.1) involved identifying business questions which are required to validate the logical design of the DW (Table 3-4). The business questions require the DW structure to support the extraction of resource usage measures in terms of the frequency of resource usage. In addition, the business questions require that the usage measures should be targeted by the specification of different levels of aggregation in terms of specified time periods, user and computer locations.

Table 3-4: ICTS Business Questions (Figure 3-2, Step 3.2).

Business Questions
How often do students use targeted university computers during a specified time period?
What set of computers is used most often by users during a specified time period?
What computer processes are used most often by users during a specified time period?
What processes are used most often on specified university computers?
How often do users use the university printing resources?

3.3.2. ICTS Report Collection and Analysis

The next step in the user-driven phase of the ATDM required the collection of ad hoc reports generated by ICTS (Figure 3-2, Step 3.3). ICTS was not able to provide the results of ad hoc reports, but the typical ad hoc report and query results were discussed and documented during user interviews. ICTS ad hoc report and query results are issued to the ORU data sources to report on the frequency of ICTS resource usage. The resources for which usage reports are generated can be mapped to particular business fields (Section 3.2.3). The measures used to indicate the usage of resources in ad hoc reports are typically the count and the average count of use of ICTS resources by a specific group of users, computers and printers within a specified time period.

3.3.3 Analytical Requirements

The analytical requirements for the NMMU ICTS (Figure 3-2, Step 3.4) were derived as a set of semantic metadata that was documented as part of the application of the user-driven phase (Tables 3-5, 3-6). The analytical requirements serve as an important set of semantic metadata (Section 1.1) that provides an initial indication for the designer to the logical design of a DW structure's DW, since analytical requirements are represented in terms of measures and dimensions (Guo *et al.* 2006).

Each dimension represents a unique combination of attributes descriptive of the associated measures and will be referred to as a dimension group (Table 3-5). The measures represent aggregated values with a granularity specified by unique combinations of the associated dimension groups (Table 3-6). The only dimension for which specific attributes were identified is the time dimension which includes the year, semester, quarter, month, week, day and hour in which a computer, printer or process was used.

Table 3-5: Analytical Requirements Dimensions.

Dimension	Dimension Attribute
Time	Year, Semester, Quarter, Month, Week, Day, Hour
Computer	Attributes which allow the analysis of specific groups of computers
User	Attributes which allow the analysis of specific groups of users
Printer	Attributes which allow the analysis of specific groups of printers
Process	Attributes which allow the analysis of specific groups of printers

Table 3-6: ICTS Analytical Requirements.

	Computer Usage	Printer Usage	Process Usage
Measures:	Count, Average	Count, Total Cost, Total Copies, Average, Average Page Count, Average Cost, Average Number of Copies	Count, Average
Dimensions:	Time, Computer, User	Time, User, Printer	Time, Computer, User, Process

The analytical requirements for the computers, printers and processes usage reports (Table 3-6) share the User, Time and Computer dimensions. These dimensions allow analysts to obtain usage measures of ICTS resources in terms of a level of aggregation based on a selection of attributes in the dimension identified. The analytical requirements identified, also represent the data required to obtain the information describing the identified KPIs (Section 3.2.4) since the KPIs also require the same measures which focus on the different ICTS, resources based on the same specified dimensions.

3.4 Supply-Driven Phase

The application of the supply-driven phase (Figure 3-3) to the case study begins with the identification of ICTS source systems (Section 3.4.1) followed by the classification of data tables from the identified source systems (Section 3.4.2) that are relevant to the required DW data. Once relevant tables have been identified, to be considered as potential source tables for the DW structure, pure operational tables and columns are removed (Section 3.4.3) since the data they contain will be of no significance to the analysis of resource usage.

The remaining tables and columns will then be mapped (Section 3.4.4) to the subject areas identified. The next step is to homogenise the semantics of the tables and columns (Section 3.4.5) to ensure consistency in the naming conventions. The last step of the supply-driven phase will involve the integration of the relevant tables (Section 3.4.6) to obtain a data schema for each subject area. During the application of the supply-driven phase a set of semantic metadata is generated at each step that describes the logical mapping between different deliverables in the supply-driven phase.

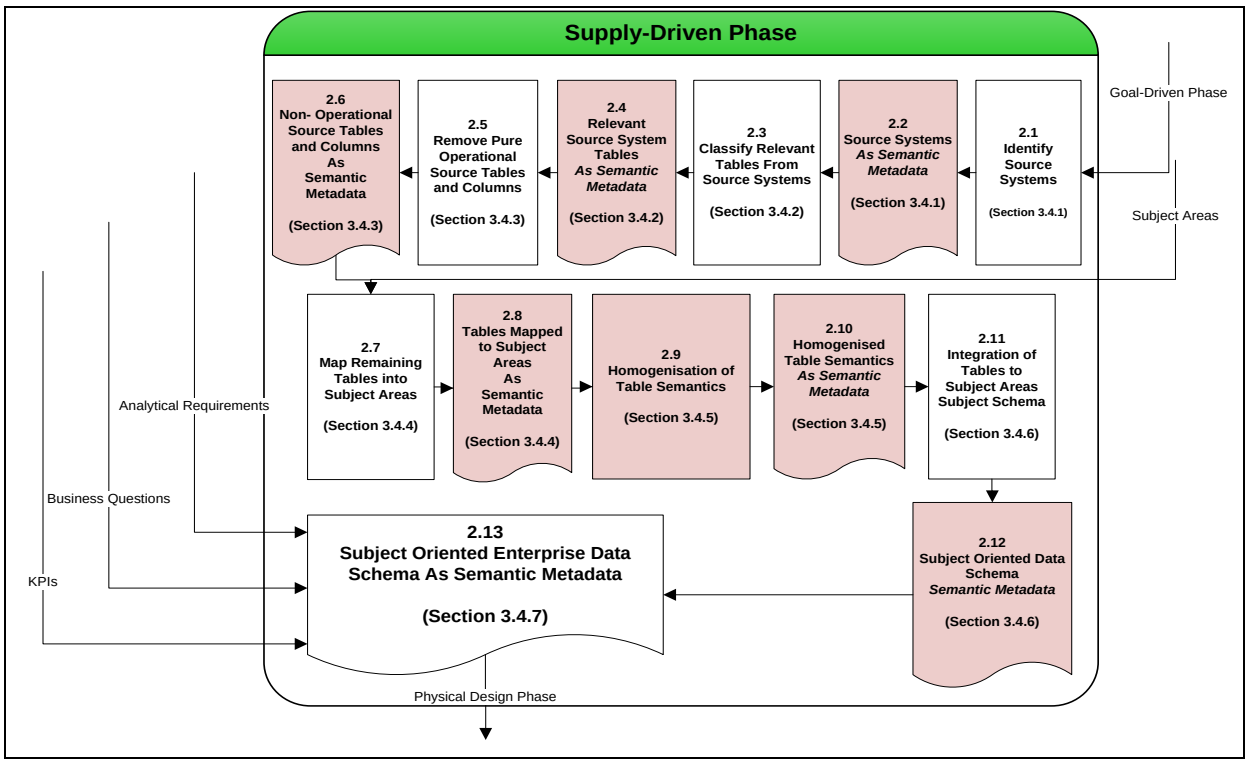


Figure 3-3: The Supply-Driven Phase of the ATDM.

3.4.1 Identification of ICTS Source Systems

The investigation of the ICTS case study resulted in the identification of five operational source systems (Figure 3-3, Step 2.1), namely:

1. Computer Logs;
2. User Computer Login logs;
3. Computer Process Logs;
4. Printing Logs; and
5. The NMMU active directory (AD).

The computer logs source system records information in a single log file, called the PC table, containing several rows describing the current operational status of computer workstations. The user computer login logs source system consists of the Logon table in which each user login to a university computer is recorded. The computer process log source system records the details of each process that is initiated in the Process table. The printing logs source system records the details of each printing job that is initiated in the Pcounter table.

The NMMU active directory (AD) is used in the management of ICT resource usage to describe the computers and users of the NMMU and consists of two tables in which information regarding users and computers is recorded. The NMMU_Computers table records information that describes the location of workstations in the NMMU. The AD also contains a table referred to as the NMMU_Users table in which information of users is recorded that describes their affiliation to the NMMU, and the campus and organisational unitto which the belong. The source systems which were identified represent the first set of semantic metadata documented in the application of the supply-driven phase (Figure 3-3, Step 2.2).

3.4.2 Classification of Relevant Tables from ICTS Source Systems

Table 3-7 represents the tables of the source systems which were identified as relevant to the proposed DW structure. These tables (Appendix A) contain information as described in Section 3.4.1 and were classified as described by Guo et al. (2006) (Figure 3-3, Step 2.3). The Logon, Process and Pcounter tables were classified as transaction tables, since they describe a logon, process and printing event respectively (Guo *et al.* 2006). The PC table was classified as a control table since it only describes the current operational state of computer workstations (Section 3.4.1).

The NMMU_Computers and NMMU_Users tables are classified (Figure 3-3, Step 2.4) as component tables, since they contain information that describes an event that occurred (Guo *et al.* 2006). The identified set of source system tables, which were classified, were documented as a set of semantic metadata (Section 1.1) that is used in the next step of the supply-driven phase in which operational tables and columns are removed (Section 3.4.3).

Table 3-7: Classified ICTS source systems' tables (Figure 3-3, Step 2.4).

Table Name	Source System	Category
PC	Computer Logs	Control Table
Logon	Computer Log-in Logs	Transaction Table
Process	Computer Process Logs	Transaction Table
Pcounter	Printing Logs	Transaction Table
NMMU_Computers	NMMU Active Directory	Component Table
NMMU_Users		Component Table

3.4.3 Removal of Pure Operational Tables and Columns

The six operational tables were classified (Table 3-7), as described by Guo *et al.* (2006), and pure operational tables and columns were removed (Figure 3-3, Step 2.5) to obtain a set of tables and columns which will be considered for subject analysis (Figure 3-3, Step 2.6) as a set of documented semantic metadata. The PC table was found to be a pure operational table (Guo *et al.* 2006) and does not contribute towards the analysis of ICTS resources since it only describes the current state of computer workstation usage, therefore this table was removed and this left five remaining tables. The remaining five tables are represented below in terms of their state after the operational columns have been removed. The NMMU_COMPUTERS and NMMU_USERS tables had no pure operational columns which required removal.

Logon (LogonId, PCNAME, USERNAME, StartDate)
Pcounter (PRINTUSERNAME, PRINTERNAME, PAPERSIZE, DUPLEX, COPIES, BWPAGES, COLORPAGES, NumberOfPages, DateTime)
Process (PROCESSID, PROCESSTITLE, LOGONID, StartDate)
NMMU_USERS (sAMAccountName, destinationOU)
NMMU_COMPUTERS (cn, destinationOU, operatingSystem)

3.4.4 Mapping of Remaining Tables to Identified Subject Areas

In order to ensure that source tables with the same or similar business semantics are easily integrated into the schemas, which will be created for the analysis for each business field's subject area, the remaining tables have to be mapped to the respective subject areas identified for the different business fields (Figure 3-3, Step 2.7). The ICTS computer management business field's subject area (Section 3.2.5) specifies that user login, user and computer information are required to analyse the computer management subject area.

In order to accommodate the data requirements, which are specified by the computer management business field's subject area, the Logon-, NMMU_COMPUTERS- and NMMU_USERS tables are respectively mapped to this subject area (Table 3-8). The information required for the subject area of the process management business field includes user, computer and process information. This information was identified to be provided by the Process, NMMU_COMPUTERS and NMMU_USERS tables. Similarly the tables mapped to printing management subject area include Pcounter, NMMU_COMPUTERS and NMMU_USERS. The mapping of the remaining source system tables to the respective business fields' subject areas is required to be documented as a set of semantic metadata (Figure 3-3, Step 2.8).

Table 3-8: Mapping of remainder tables to ICTS subject areas (Figure 3-3, Step 2.8).

			Remaining Tables			
Business Field	**Required Subject Information**	Logon	Process	Pcounter	NMMU_Computers	NMMU_Users
Computer Management	Computer, User, Login,	■			■	■
Printing Management	Computer, User, Printing			■	■	■
Process Management	Computer, User, Process		■		■	■

3.4.5 Homogenisation of Table Semantics

The homogenisation of table semantics step (Figure 3-3, Step 2.9) requires that the inconsistent naming conventions of tables and columns are removed to obtain a set of semantic metadata that represents the homogenised table semantics (Figure 3-3, Step 2.10). The table and column names will therefore be changed to ensure that entities such as users, computer, processes and printers are consistently represented in all tables in terms of more descriptive column names and the consistent use of cases. As a representative example, Table 3-9 shows the conversion of the Logon table in the process of the homogenisation of concepts. The names which were assigned to the Logon table and its columns by ICTS were found to be descriptive of the information contained in the table with the exception of the PCName column. The PCName field of the Logon table represents the unique names assigned to computer workstations to which computers are assigned.

Based on this observation the PCName column is renamed as WorkstationName. The conversions of the remaining tables in the process of the homogenisation of concepts (Figure 3-3, Step 2.10) are represented in Appendix B.

Table 3-9: Homogenisation step applied to the Logon table (Figure 3-3, Step 2.10).

	Current Semantics	New Homogenised Semantics
Table Name	Logon	Logon
Column Name	LogonId	LogonId
	PCName	WorkstationName
	UserName	UserName
	StartDate	StartDate

3.4.6 Integration of Tables to Form Subject Oriented Data Schema

The goal of this step (Figure 3-3, Step 2.11) is to integrate the remaining tables into a data schema for each subject area identified to form a subject oriented data schema (Figure 3-3, Step 2.12). Relationships have to be established between the tables for each subject by the identification of primary keys and foreign keys. This validates the mapping of the remaining tables to the business fields' subject areas and simplifies the design of the subject oriented data schema. Table 3-10 represents the identified tables' relationships in terms of primary key columns and tables which contain a foreign key reference column with the same semantics which forms part of the set of semantic metadata documented in this step (Figure 3-3, Step 2.12). This set of semantic metadata is used, together with the semantic metadata documented in previous steps of the supply-driven phase, to derive the ICTS subject oriented data schema that represents the primary set of semantic metadata documented in this step (Figure 3-4).

Table 3-10: Table relationships (Figure 3-3, Step 2.12).

Remaining Table	Primary Key Column	Foreign Key Reference Table
Logon	LogonId	Process
Process	ProcessId	
Printing	(*No Primary Key Identified*)	(*No Primary Key Identified*)
NMMU_Workstations	WorkstationName	Logon
NMMU_Users	UserName	Logon, Printing

The UserName column of NMMU_Users has been identified to uniquely represent the users and will be used as a primary key. The LogonId, WorkstationName and ProcessId columns of Logon, NMMU_WorkStations and Process were identified as primary keys respectively.

Although the Process table does not consist of a reference to the associated user and computer on which the process was executed, this information can be obtained through the LogonId column attributes. This column's attributes represent a foreign key for a record in the Logon table from which user and computer information can be derived. The tables that will be used to describe the information that is required to allow the analysis of the printer subject area includes the Printing_Process and NMMU_Users tables. The Printing_Process table does not contain information that links a printing transaction to a specific computer or login instance. Based on this observation the information identified to support the analysis of printer usage (Section 3.2.5) would not include computer information.

3.4.7 The ICTS Subject Oriented Enterprise Data Schema

The subject oriented data schema (Figure 3-4) for ICTS was derived based on the application of the supply-driven phase of the ATDM on the NMMU ICTS case study. In order to ensure that a logical data model is derived that is able to satisfy the information requirements of ICTS, the derived KPIs (Section 3.2.4) and analytical requirements (Section 3.3.3) have to be incorporated in the subject oriented data schema (Figure 3-4) to yield the ICTS enterprise subject oriented data schema (Figure 3-3, Step 2.13).

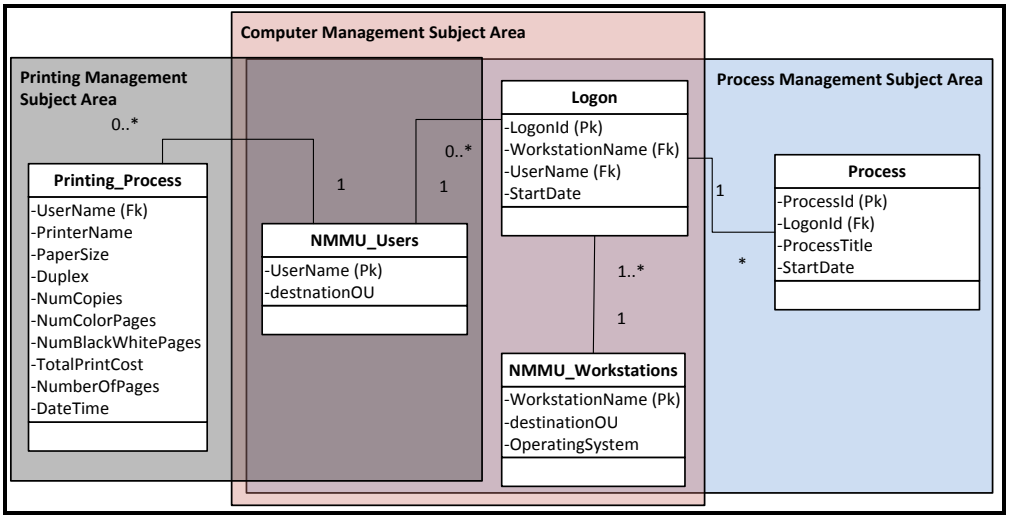


Figure 3-4: The ICTS Subject Oriented Data Schema (Figure 3-3, Step 2.12).

In the step by which tables are integrated to form a schema for each business fields subject area (Section 3.4.6), the printing management subject area could not be associated with the computer from which a printing job was executed as there were non-existent relationships between the required tables. This affects the analytical requirements (Section 3.3.3) identified in the user-driven phase for analysing printer usage.

Based on this information the analytical requirements for the analysis of user printing usage were adapted not to consider computer information. The integration of the KPI's, analytical requirements and subject oriented data schema to form the ICTS enterprise subject-oriented data schema, involved the logical integration of the measures and dimensions of KPIs and analytical requirements into the subject oriented data schema. The integration of these measures and dimensions resulted in the creation of a logical design (Figure 3-5) for the ICTS DW structure's DW which satisfies the identified analytical requirements and KPIs.

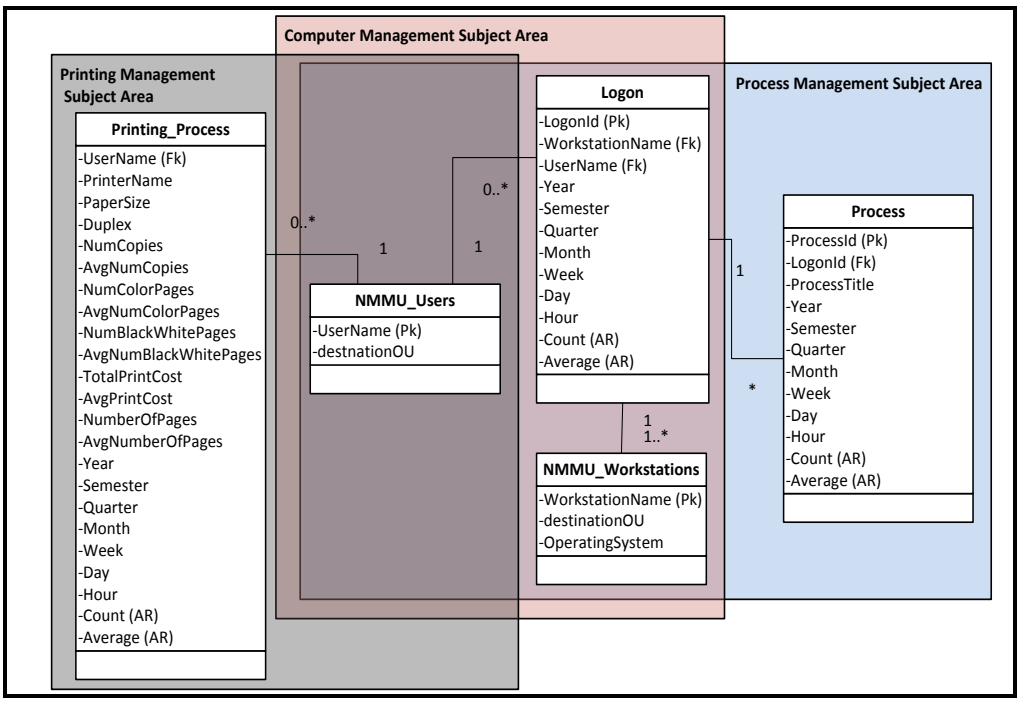


Figure 3-5: The ICTS Subject Oriented Enterprise Data Schema (Figure 3-3, Step 2.13)

The analytical requirements derived (Section 3.3.3), require that the respective subject areas are analysed in terms of a set of dimensions and measures. The measures Count and Average are required for the analysis of each subject identified of which Average is also the required measure for the analysis of the identified KPIs. Based on these requirements Count and Average were introduced into the Printing_Process-, Logon- and Process tables since these tables are considered to be transaction tables in which the details of a printer, computer and process usage are captured. In addition, these measures represent values which require an aggregation of the respective tables, based on a specified selection of dimension attributes. The integration of the subject oriented data schema with KPIs and analytical requirements resulted in the derivation of the specific attributes which are available for the analysis of the different ICTS subjects.

Based on the integration of KPIs, analytical requirements and the subject oriented data schema, the user dimension consists of the destinationOU attribute that represents a concatenation of attributes describing users. The computer dimension consists of the destinationOU attribute that describes computers as well as the OperatingSystem attribute which allows analysis in terms of the operating system installed on a computer workstation.

Measures, additional to these described in the derived analytical requirements, were identified from the subject oriented data schema (Figure 3-4). The additional measures were derived from the identification of attributes, in the Pcounter table, which enables the specification of additional measures than these specified by users (Section 3.3.3). The additional measures were identified since the user-driven and supply-driven phases are applied separately (Section 2.4).

The additional measures which were identified consist of NumBlackWhitePages, NumColorPages and NumberOfPages. In addition, an average value was assigned to each of these measures as additional measures for the printing management business field's subject area. The attributes Duplex and Papersize were introduced to the Printer dimension which further extends the capabilities of analysing printer usage.

Based on the consolidation of the analytical requirements, KPIs and the subject oriented data schema the derived analytical requirements were adapted to reflect the dimensions and measures depicted in the subject oriented enterprise data schema. This required the specification of the dimensions (Table 3-11) and the introduction of newly derived measures (Table 3-12).

Table 3-11: Specific analytical requirements dimensions

Dimension	Dimension Attribute Description
Time	Year, Semester, Quarter, Month, Week, Day, Hour
Computer	destinationOU, OperatingSystem
User	destinationOU
Printer	PrinterName, Duplex, PaperSize
Process	ProcessTitle

Table 3-12: Updated ICTS Analytical Requirements

	Analysis of Computer Usage	Analysis of Printer Usage	Analysis of Process Usage
Measures:	Count, Average	Count, NumberOfPages, TotalPrintCost, NumCopies, NumColorPages, NumBlackWhitePages Average, AvgNumberOfPages, AvgPrintCost, AvgNumCopies, AvgNumColorPages, AvgNumBlackWhitePages	Count, Average
Dimensions:	Time, Computer, User	Time, User	Time, Computer, User

3.5 Conclusion

The requirements analysis phase of the proposed ATDM (Section 2.4) was applied to the NMMU ICTS as the case study that was selected for this research. The application of the requirements analysis phase formed part of the evaluation (Section 2.5) of the ATDM in order to determine its applicability as a DW structure design methodology. The aim of the evaluation of the ATDM that was addressed in this chapter was to elicit the information requirements and source data characteristics of the NMMU ICTS case study. The information requirements and data characteristics were required to be combined to form a logical design of the ICTS DW that represents the ICTS information requirements and data characteristics. This was achieved through the application of the requirements analysis phase of the ATDM. The requirements analysis phase consists of three phases which were applied, namely, the goal-driven (Section 3.2), user-driven (Section 3.3) and supply-driven (Section 3.4) phase.

The application of the requirements analysis phase resulted in sets of semantic metadata which were derived and documented to represent the deliverables of the requirements analysis phase's steps which were applied to the case study. In addition, these sets of documented semantic metadata represent the logical design of the ICTS DW structure. The subject oriented enterprise data schema was derived (Section 3.4.7) that represents a logical data model for the ICTS DW. The derived subject oriented enterprise data schema is representative of the information requirements and data characteristics of the NMMU ICTS case study since the subject oriented data schema (Section 3.4.6), KPIs (Section 3.2.4) and analytical requirements (Section 3.3.3) were used to aid its design. The integration of the subject oriented data schema, KPIs and analytical requirements resulted in the identification of additional measures (Section 3.4.7) which were not specified in the derived analytical requirements.

This resulted in the need to update the derived analytical requirements to accommodate additional measures which were identified in the derivation of the subject oriented enterprise data schema, but is not accommodated by the proposed methodology (Section 2.4). Based on the need to update analytical requirements after the derivation of the subject oriented enterprise data schema, a limitation of the methodology was identified that was addressed (Section 3.4.7) to allow analytical requirements to be updated.

Chapter 4. ATDM Physical Design Phase

4.1 Introduction

The logical design of the ICTS DW structure was derived (Figure 3-5) that represents the information requirements and data characteristics which were analysed with the application of the ATDM's requirements analysis phase (Chapter 3). The application of the supply-driven phase (Section 3.4) of the methodology's requirements analysis phase resulted in the derivation and documentation of a set of semantic metadata that describes the logical mapping of the source data to the logical design of the ICTS DW. This chapter describes the application of the ATDM's physical design phase (Figure 4-1) to address the fourth research question for this research (Section 1.6) in which the physical design of the ICTS DW structure is required to be derived.

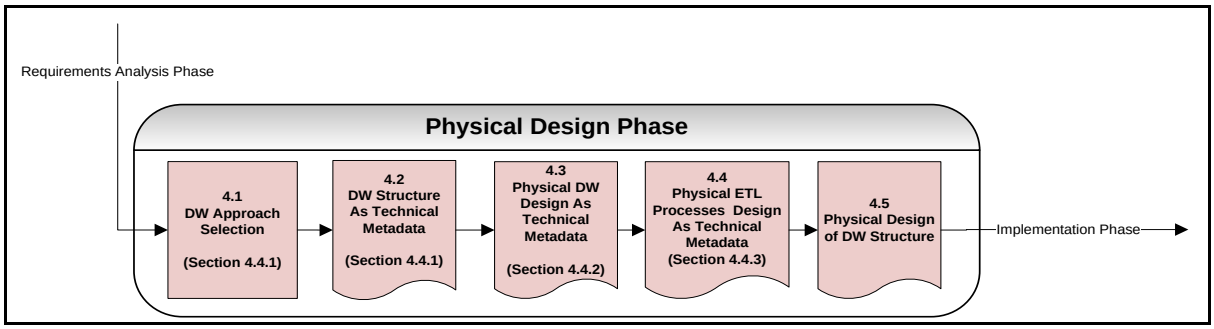


Figure 4-1: The ATDM's Physical Design Phase

The application of the physical design phase (Figure 4-1) involves selecting an appropriate DW approach, and the derivation of the physical design of the DW structure's DW and ETL processes to obtain the physical design of the DW structure. Different DW approaches are investigated and described (Section 4.2) as well as the process of selecting the most appropriate DW approach for a DW structure (Section 4.2.5). ETL processes are described as well as modelling techniques for ETL processes (Section 4.3). Based on DW approach selection factors and scenarios, the optimal DW approach for the ICTS DW structure is selected (Section 4.4.1) as part of the application of the physical design phase of ATDM (Section 4.4). After the DW approach has been selected, the physical design of the DW is described (Section 4.4.2) after which the physical design of the ETL processes is described (Section 4.4.3).

4.2 Data Warehouse Approaches

The selection of an appropriate DW approach is one of the most debated topics in the field of data warehousing (Vail III 2002; Jukic 2006; Ariyachandra and Watson 2010). Various DW approaches exist that are used in organisational DW structures (Jukic 2006). These DW approaches consist of: independent data marts (IDM) (Section 4.2.1), enterprise data warehouse (EDW) (Section 4.2.2), data mart bus (DMB) (Section 4.2.3) and federated DW approaches (Section 4.2.4).

4.2.1 Independent Data Marts (IDM) DW Approach

The IDM DW approach (Figure 4-2) consists of data marts which are independently created, each one usually for a different department in an organisation (Jukic 2006; Ariyachandra and Watson 2010). This DW approach results in DW data models which are created independently to suit the needs of the organisational departments for which they are designed and is considered to be one of the early attempts to supply a repository to aid decision support (Ariyachandra and Watson 2006).

In the IDM DW approach, each data mart, represented by a dimensional model, symbolises a specific subject area and aids the decisional processes for that subject area (Jukic 2006). Thus a dimensionally modelled DW is created for each business field in an organisation without sharing dimensions with other data marts (Jukic 2006).

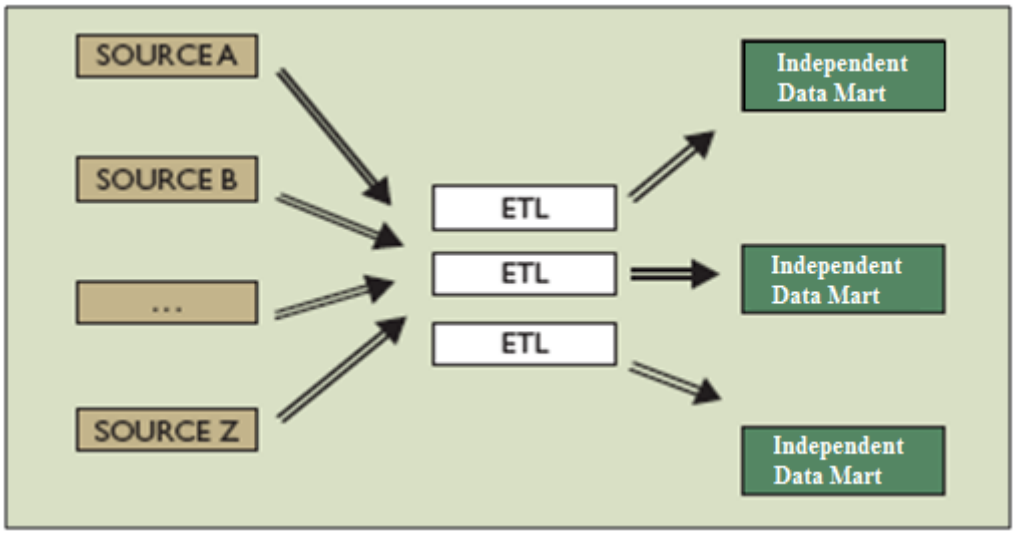


Figure 4-2: Independent Data Marts Data Warehouse approach (Jukic 2006).

The IDM DW approach consists of two major shortcomings (Ariyachandra and Watson 2010; Bontempo and Zagelow 1998; Bruckner *et al.* 2001). These shortcomings are:

- The ETL effort has to be duplicated for each data mart; and
- The restriction of cross-departmental analysis in an organisation.

ETL processes require vast amounts of processing power to allow the rapid integration of data from several heterogeneous data sources into a DW structure's DW (Bontempo and Zagelow 1998; Jukic 2006). This complex process must be designed for each data mart with the result that vast amounts of time and skill have to be invested by the organisation using the IDM DW approach.

Although each independently created data mart conforms to the requirements of the organisational department which it serves, the overall organisational view can be inconsistent if it is derived from multiple independent data marts (Bontempo and Zagelow 1998). The reason for this inconsistency is because each independent data mart has its own dimensions and does not share dimensions with other data marts in the organisation (Bontempo and Zagelow 1998; Hackney 2000). The independence of each data mart often results in different naming conventions and semantics used for data and this contributes to restrictions in cross-departmental analysis.

4.2.2 Enterprise Data Warehouse (EDW) Approach

The EDW approach (Figure 4-3), also known as the hub-and-spoke DW approach (Inmon 2002), involves integrating data from various sources into a relationally modelled, normalised DW for a DW structure containing atomic level data (Ariyachandra and Watson 2010; Jukic 2006). This normalised DW is intended to represent an enterprise-wide database that supports cross-departmental analysis of the data. Once created this enterprise-wide database, which is considered to be a single version of the truth, serves as a data source for dimensionally modelled data marts (Jukic 2006). The data marts can be either modelled dependently or independently from each other (Ariyachandra and Watson 2005). If the data marts are modelled to be independent of each other, the resulting DW approach would be known as a centralised DW approach.

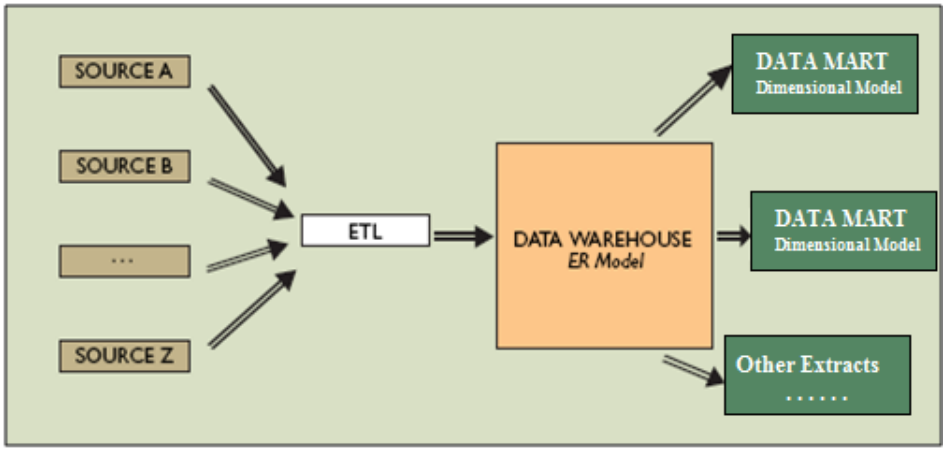


Figure 4-3: Enterprise Data Warehouse approach (Jukic 2006).

The EDW approach does not require ETL processing to be done for each of the data marts it contains. This approach was created with the purpose of accommodating a scalable data warehousing solution that is part of a concept that exceeds the traditional purpose of a DW structure's DW (Imhoff, Galemmo and Geiger 2008). In addition to only allowing dimensionally modelled data marts to source data from the relationally modelled DW, it was envisioned that additional non-dimensional extracts can be allowed to source data to further enhance decision support initiatives (Jukic 2006). Additional extracts from the relational model could include individual data tables, data sets used for data mining and flat files. Although this DW approach is one of the most effective, scalable and most used DW approaches, it is resource expensive due to the large amount of time spent on design, since these approaches often represent an entire organisation (Ariyachandra and Watson 2006; Jukic 2006).

4.2.3 Data Mart Bus Data Warehouse Approach

The data mart bus (DMB) DW approach (Figure 4-4) is defined as a collection of dimensionally modelled data marts in which a set of commonly used dimensions are first designed, based on the identification of business requirements for specific business processes (Jukic 2006; Ariyachandra and Watson 2010). The fact tables may be connected to several dimensions which are shared with other fact tables. Thus the result of applying this approach, which is proposed by Kimball *et al* (2002), is usually a collection of interconnected dimensional data marts (Ariyachandra and Watson 2010; Jukic 2006). This approach only makes use of one set of ETL tasks, which have to be performed periodically in order to populate a DW structure's DW with clean and integrated data. In contrast with the EDW approach the DMB approach does not provide support for additional extracts.

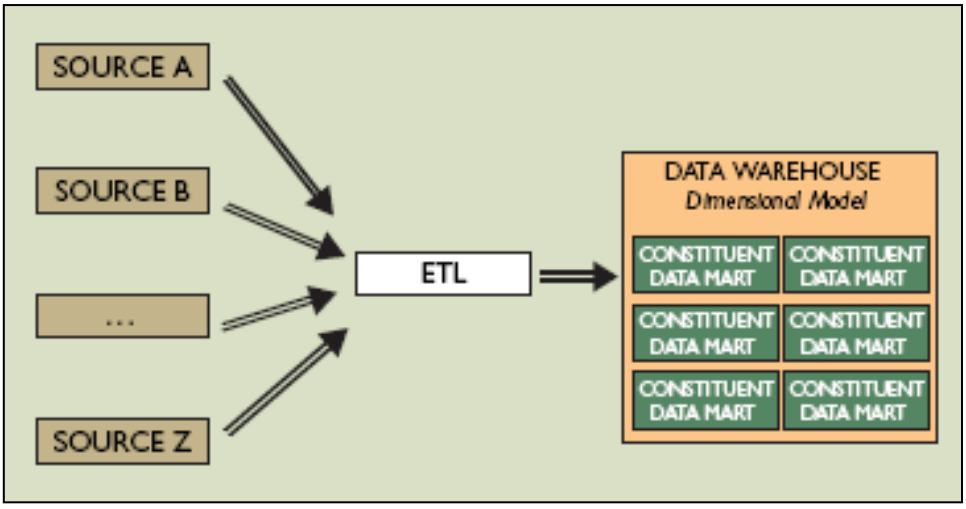


Figure 4-4: Data Mart Bus (DMB) Data Warehouse approach (Jukic 2006).

4.2.4 Federated Data Warehouse Approach

The federated DW approach (Figure 4-5) is suitable for organisations that do not wish to redesign their already complex and expensive, decision support infrastructure (Jindal 2004). The federated DW approach is also known as a virtual DW approach that consists of a set of views over operational sources (Han and Kamber 2006). The data is left intact at its sources because several challenges were faced during the implementation of different DW approaches (Hackney 1998). Thus this DW approach does not consist of an integrated repository containing data integrated from several heterogeneous data sources (Jindal 2004; Hackney 1998). In the scenario where data across several information systems is required to be integrated, to obtain a set of data for decision support, distributed queries, shared keys, global metadata and enterprise information integration are used (Ariyachandra and Watson 2010; Jindal 2004).

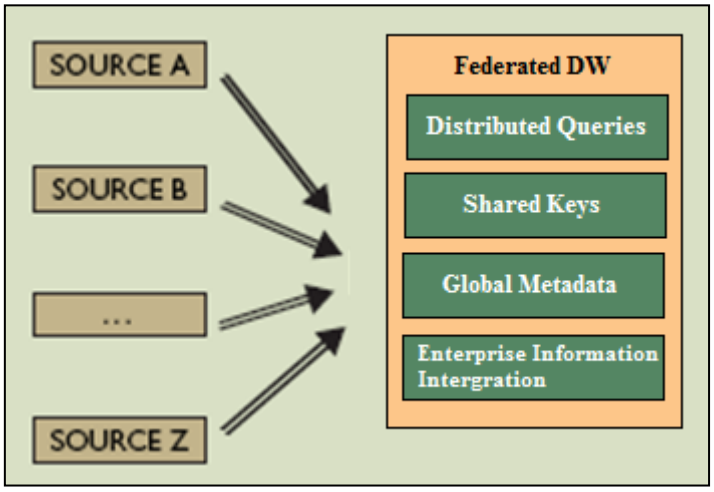


Figure 4-5: The Federated DW approach (Ariyachandra and Watson 2010; Jindal 2004).

4.2.5 Data Warehouse Approach Selection

The selection of a DW approach is one of the most important decisions that has to be made in the development of an DW structure, since the selection of the incorrect DW approach often results in shortcomings in performance and scalability (Ariyachandra and Watson 2010). Although the selection of a DW approach is considered to be important, limited research has been done to address DW approach selection. The process of selecting the most appropriate DW approach for an organisation depends on the environment of the organisation for which the DW approach is required (Ariyachandra and Watson 2005). A set of factors has been identified in a study by Ariyachandra and Watson (2005) which can be used to determine the most appropriate DW approach for a given organisational environment (Section 4.2.5.1). The factors are used to describe scenarios in which particular DW approaches are preferred (Section 4.2.5.2).

4.2.5.1 Data Warehouse Approach Selection Factors

Different organisations have different needs and six factors have been identified in the study by Ariyachandra and Watson (2010) that contribute to the selection of particular DW approaches for different organisational scenarios. The DW approach selection factors include:

- Resource constraints;
- Management's strategic view;
- Level of urgency;
- Perceived ability of IT staff;
- Organisational interdependence; and
- The level of sponsorship.

The constraints and availability of resources significantly influence the ability of organisations to develop information systems and a lack thereof significantly restricts development (Miller and Friesen 1982). The development of a DW structure, typically, requires large amounts of time, expert skills and funding, all of which are not always available.

The selected DW approach applied by an organisation determines the success of the DW structure in providing decision support. Since the selected approach is influenced by available resources, they directly influence decision support (Ariyachandra and Watson 2010).

The most appropriate selection of a DW approach requires that organisations explain their strategic view and what is required from the DW structure in the short- and long term (Griffin 2001). Organisations might require the functionality of a data warehousing solution to aid the decision support needs of a specific organisational sub-unit, organisational business unit or department or at an enterprise level (Kelly and Hadden 1997; Ariyachandra and Watson 2010).

The DW approach that is selected by an organisation, which has an urgent need for the processing capabilities of a DW structure, should allow the rapid implementation of the DW structure to support the urgent information needs of the organisation (Ariyachandra and Watson 2010). The perceived ability of staff to create a DW structure for decision support influences the DW approach selected (Ariyachandra and Watson 2010). The perceived ability of individuals to accomplish a task, such as developing a DW structure, is created through the accumulation of the required skills and subsequent experience (Ariyachandra and Watson 2010; Fichman 1997). The skills and experience obtained by individuals determines their knowledge barrier level which is used to select an appropriate DW approach. Individuals with high knowledge barriers are more likely to adopt more complex DW approaches and conversely individuals with a low knowledge barrier are more likely to select DW approaches with lower knowledge barriers (Breslin 2004).

Organisations need to share information internally to ensure superior organisational interdependence and it is sometimes critical to identify how organisational departments affect each other through their actions (Andres 2002; Ariyachandra and Watson 2010). A DW structure has the ability to enhance the capability of an organisation to process its information as an integrated set of information that represents the entire organisation instead of representing only specific departments (Devlin 1997; Ariyachandra and Watson 2010).

The integrated view of an organisation's information can further allow the development of information systems that support high interdependence within the organisation. The level of sponsorship for a specific project also influences the choices made in the creation of an entity (Swanson 1994; Watson *et al.* 1995; Bradshaw-Camball and Murray 1991; Jasperson *et al.* 2002). This is because sponsors have the ability to maintain power over individuals in an organisational environment. Sponsors usually favour the selection or creation of an entity from which they will directly benefit (Ariyachandra and Watson 2010).

4.2.5.2 DW Approach Selection Scenarios

The scenarios for which a specific DW approach is preferred are described to support the process of selecting a DW approach that is preferred for a given scenario. The study by Ariyachandra and Watson (2010) involved the analysis of the DW approaches of various DW structures of different organisations in terms of DW approach selection factors, by means of questionnaires which were presented to these organisations as well as to experts in the field of data warehousing.

Three DW approach selection factors were identified to significantly influence the selection of the IDM DW approach (Section 4.2.1) over other DW approaches. The IDM DW approach was identified as an approach that is used as a result of a series of independent decisions and is a preferred approach over the other approaches if there are serious constraints in resources (Ariyachandra and Watson 2010). This approach is usually implemented if the management's strategic view of a DW structure is limited to a specific organisational sub-unit and the perceived ability of an organisation's information technology (IT) staff is considered to be low.

Three DW approach selection factors have been identified which significantly influence the selection of the EDW approach (Section 4.2.2) above the other DW approaches (Ariyachandra and Watson 2005). These factors are organisational interdependence, management's strategic view and the perceived ability of IT staff. The EDW structure is most commonly used if the strategic view of DW structure consists of the entire enterprise (Ariyachandra and Watson 2006). This DW approach favours the scenario in which organisational interdependence and the perceived ability of IT staff are both high (Ariyachandra and Watson 2005). The EDW approach is considered to be a significantly

successful approach, but it is the most expensive in terms of development cost, maintenance cost and development time.

The DMB approach (Section 4.2.3) was identified as the preferred choice if constraints in resources are severe; the level of urgency is high; there must be high organisational interdependence and the level of sponsorship is high. Since only a limited number of organisations use the federated DW approach (Section 4.2.4), no DW approach selection factors could be identified which significantly favour the selection of this DW approach, however, it is most commonly preferred for smaller domains such as an organisational sub-unit (Ariyachandra and Watson 2005).

One of the most debated topics in the field of data warehousing is the selection of either the EDW or DMB DW approaches for the design of a DW structure (Ariyachandra and Watson 2005; Jukic 2006). These DW approaches were found to be similar in their respective success of supporting the decision support initiatives of different organisations which adopted these approaches as part of their DW structure design (Ariyachandra and Watson 2006). Three factors were identified which can be used to decide between the EDW and DMB DW approaches.

These factors are managements' strategic view, urgency and organisational interdependence (Ariyachandra and Watson 2005). Table 4-1 shows the scenarios in each DW approach is preferred over the other approach. The DMB and EDW DW approaches both support high organisational interdependence upon their deployment, therefore organisational interdependence is considered not to influence the decision between the DMB and EDW.

Table 4-1: The scenarios which influence the selection between the EDW and DBM DW approaches (Ariyachandra and Watson 2010).

Selection Between the EDW and DMB DW Approaches		
DW Approach	**Management's Strategic View**	**Urgency**
EDW Approach	Enterprise	Low
DMB DW Approach	Less strategic (departmental)	High

Organisations tend to select the EDW approach if their strategic view of the DW structure is enterprise-wide and the urgency for the DW structure is low (Ariyachandra and Watson

2010). If the organisational view of the DW structure is less strategic and the urgency high, the DMB DW approach is typically preferred.

4.3 Extraction, Transformation and Loading (ETL) Processes

To be able to perform consistent and reliable data analysis, it is expected that a DW gets populated with clean and transformed data which is extracted and integrated from different operational sources with ETL tools (Chaudhuri and Dayal 1997; Jukic 2006). The ETL component of a DW structure links the source data with the DW structure's DW. The ETL process is considered to be a back-end tool and is used as part of the refreshing of a DW which ensures that updates in the operational sources are reflected in a DW structure's DW (Chaudburi and Dayal 1996).

Issues regarding the refreshing of a DW structure's DW are, when and how to refresh it. DWs are typically refreshed during night times, usually if the operational sources systems are offline, or if the operational burden of operational source systems is relatively low. DWs should be refreshed according to the following set of tasks (Vassiliadis *et al.* 2001; Simitsis and Theodoratos 2009):

- The identification of data sources from which data gets extracted;
- Customisation and integration of the data into a common format;
- Cleaning the data in terms of database and business rules; and
- The loading of this data into the selected DW repository.

The ETL process (Figure 4-6) consists of the identification of relevant data sources, typically flat files and relationally modelled data bases, and the extraction of data to the data staging area (DSA) (Simitsis and Theodoratos 2009). Extraction routines used to extract the data from sources, provide snapshots of the source data which are sent to the DSA where it is transformed and cleaned. The results of the DSA are known as intermediate results which are propagated to the relevant DW structure's DW.

The transformation process in an ETL solution is regarded as the most important and complex part of the submission of data to an ETL cycle (Kimball *et al.* 2002). During this process, data is combined, cleansed and filtered. Surrogate keys are assigned and modified,

aggregates are computed and error handling is completed. The primary purpose of the transformation phase is to ensure the quality of the data that will be loaded into a DW (Simitsis and Theodoratos 2009).

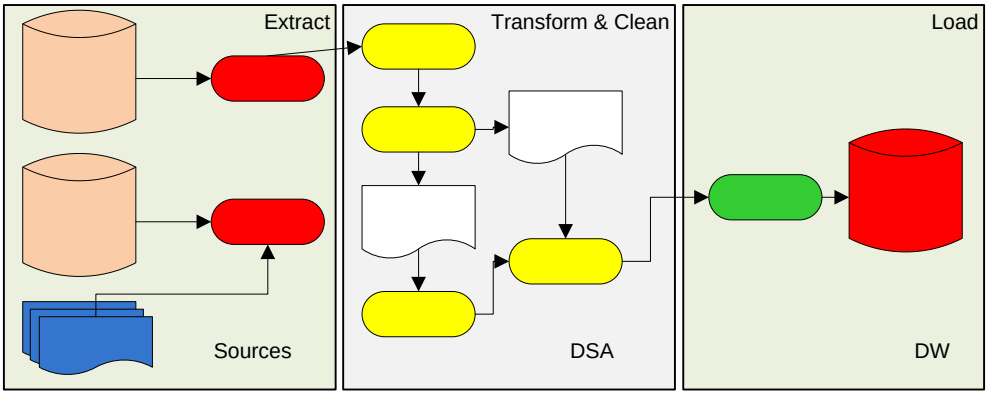


Figure 4-6: Extraction, Transformation and Loading Processes (Simitsis and Theodoratos 2009)

In some scenarios, the transformed and cleansed data is physically stored at the DSA in a relational format before it is transmitted to the presentation area for analysis (Kimball *et al.* 2002). The development of the infrastructure to support the ETL processes of a DW structure is usually the most time consuming and expensive task in the DW structure design effort (Moss and Atre 2003). The success of the ETL development phase can be supported by semantically- consistent, operational, source data models which require minimal transformation and cleaning before data is loaded to a DW structure's DW (Jukic 2006). This is not always possible, since an organisation may consist of several operational sources which differ significantly from each other (Vassiliadis *et al.* 2001).

A detailed model must be designed to address the complex task of extracting, transforming and the loading of source data onto a DW (Albrecht and Naumann 2008; Trujillo and Luj 2003; Munoz *et al.* 2008). The modelling of the ETL processes is aimed at providing clear definitions of the required processes to map source data to its destination (Vassiliadis and Simitsis 2002). Additionally ETL models are useful for communicating with individuals involved in the design process and the facilitation of easy revisions of the design task. Several ETL modelling techniques have been proposed (Munoz *et al.* 2008; Vassiliadis and Simitsis 2002; Albrecht and Naumann 2008; Trujillo and Luj 2003), all based on Unified Modelling Language (UML) (Satzinger *et al.* 2005).

The modelling technique that is proposed by Trujillo and Luj (2003) (Figure 4-7) uses UML class diagrams to represent source tables, destination tables as well as ETL mechanisms which are applied to the source data. The ETL mechanisms represent the various activities which are used to extract, transform and load data to destination tables (Trujillo and Luj 2003; Munoz *et al.* 2008). The destination table is denoted with a UML stereotype which represents the applied mechanism. The applied mechanism is described by a note which describes how each destination attribute is linked to a destination table attribute. This modelling technique represents source and destination table attributes in terms of their respective formats.

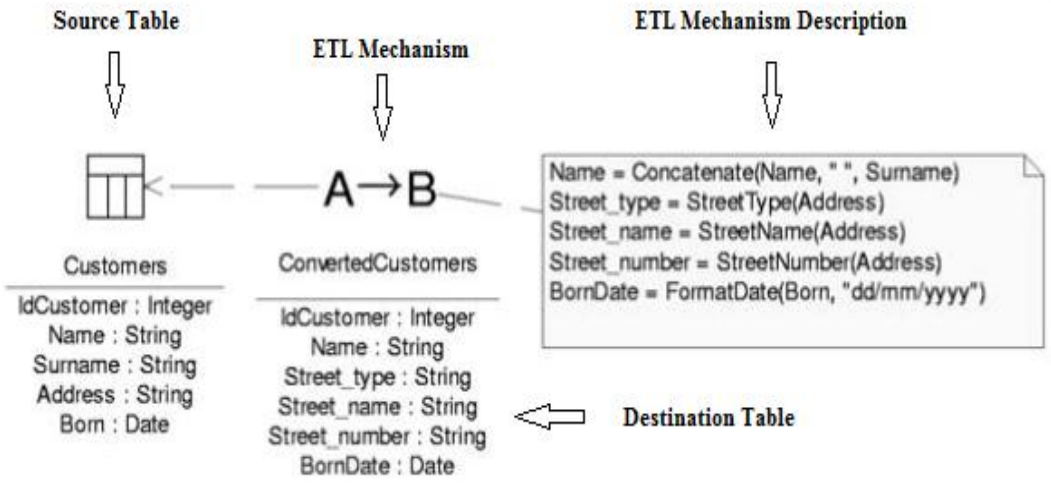


Figure 4-7: ETL modelling technique proposed by Trujillo and Luj (2003)**.**

Vassiliadis and Simitsis (2002) propose an ETL modelling technique which directly links source table attributes to destination table attributes (Figure 4-8) in a more representative manner compared to the technique proposed by Trujillo and Luj (2003). Each ETL mechanism is represented by a hexagon that is accommodated by a note that describes the operation needed to transform source attributes in order to derive a respective destination table attribute. In addition the primary key attributes, which are not specifically indicated by the modelling technique proposed by Trujillo and Luj (2003), are also represented by a hexagon.

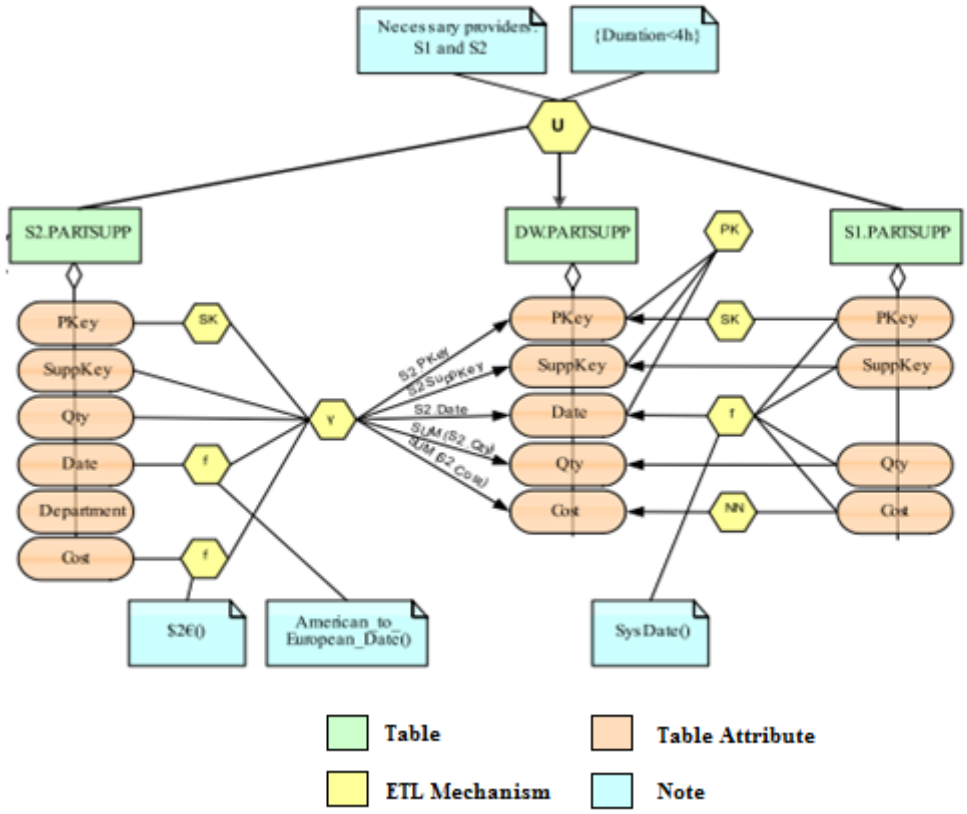


Figure 4-8: The ETL modelling technique proposed by Vassiliadis and Simitsis (2002)

By combining the different elements of the described ETL modelling techniques, a more descriptive ETL model can be derived (Figure 4-9) which complements the shortcomings of the described techniques. The ETL modelling technique that is proposed by Vassiliadis and Simitsis (2002) was adapted to represent the formats of data attributes of tables depicted in an ETL model based on the ETL modelling technique proposed by Trujillo and Luj (2003). ETL mechanisms are represented (Figure 4-9), similar to the technique proposed by Vassiliadis and Simitsis (2002) as hexagons which are described with an accommodated note. Hexagons which are connected only to table attributes and annotated with the abreviation “PK” are used to indicate the primary key of a table (Vassiliadis and Simitsis 2002).

In order to ensure that the link between corresponding destination and source attributes, and the difference between source, destination and temporary tables are clearly represented, additional ETL modelling concepts were introduced. Theses concepts respectively consists of related source and destintation attributes which are denoted with the same colour and notes which are used to describe and indicate whether a table is a source, temporary or destination table.

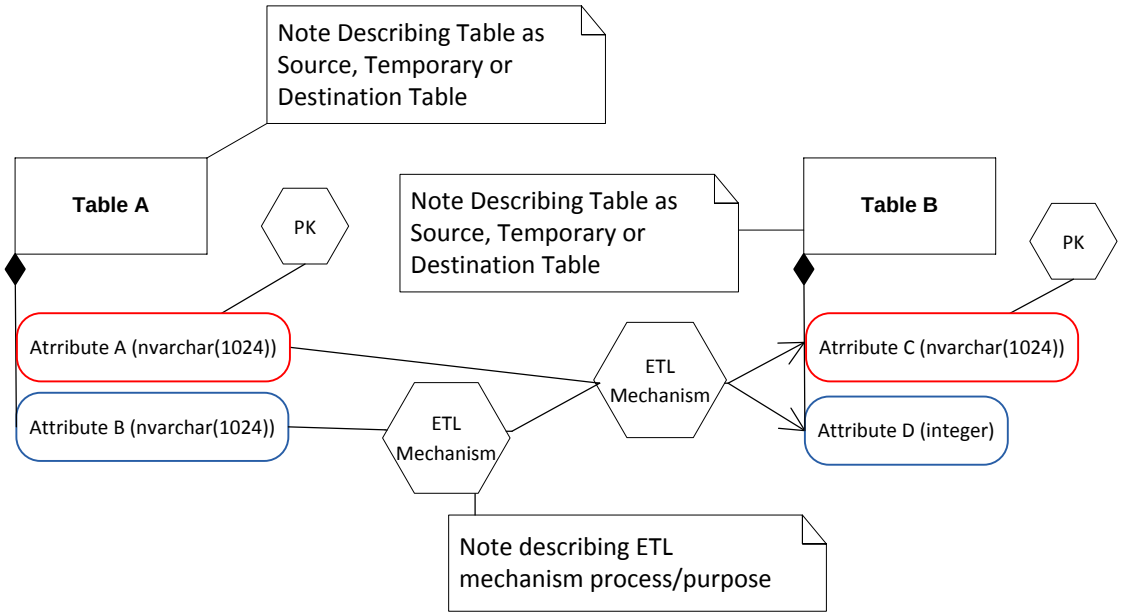


Figure 4-9: The Proposed ETL Modelling Notation.

4.4 The NMMU ICTS Case Study Physical DW Structure Design

In order to allow the NMMU ICTS to be able to effectively analyse their ORU data through the utilisation of a DW structure, an appropriate DW approach is required. The most appropriate DW approach for the NMMU ICTS case study is selected (Section 4.4.1) based on the described DW approach selection scenarios (Section 4.2.5.2). The physical design of the DW is derived (Section 4.4.2) and the physical design of the ETL processes which are required to populate the DW is designed (Section 4.4.3) using the proposed ETL modelling technique (Section 4.3).

4.4.1 ICTS Data Warehouse Approach Selection

An appropriate DW approach has to be selected as part of the application of the physical design phase of the ATDM (Figure 4-1, Step 4.1) to be documented and used as an important set of technical metadata (Figure 4-1, Step 4.2) that will be used to aid the physical design of the ICTS DW structure. The DW approach selection scenarios (Section 4.2.5.2), in which particular DW approaches are preferred, comprises of different combinations of six DW approach selection factors (Section 4.2.5.1) as significant factors influencing the selection of particular DW approaches (Section 4.2) (Ariyachandra and Watson 2005). Table 4-2 shows the DW approach selection factors for the NMMU ICTS case study. The ICTS selection factors (Table 4-2) were selected, under the assumption that they represent the ICTS case study, to demonstrate the process of selecting a DW approach.

The development of the DW structure for ICTS is considered to be constrained by time, funding and staff. A limited amount of funding has been made available for the development of the DW structure when the amount is compared to the typical cost of developing a DW structure (Ariyachandra and Watson 2006). The development of a DW to support ICTS in the analysis of ORU data is part of the NMMU ICTS strategy and does not represent the strategy of the entire NMMU.

Table 4-2: Summary of ICTS scenario in terms of DW structure selection factors.

Resource Constraints	Management Strategic View	Urgency	Organisational Interdependence	IT Staff Perceived Ability	Level of Sponsorship
High	Departmental	High	High interdependence between ICTS business fields	High	Departmental

Since a DW structure is required to support only ICTS in managing resources through the analysis of their ORU data, the strategic view of the management is constrained to a single organisational business unit or department. Since the need of the NMMU ICTS (Chapter 3) to be more capable in generating decision support the need for a DW structure is considered to be urgent.

In the case of the NMMU ICTS case study a DW approach is required that accommodates organisational interdependence between the different business fields (Section 3.2.3) of the ICTS department of the NMMU. The process of achieving interdependence between the different ICTS business fields involves the integration of their respective descriptive data into a homogenised representation which represents all business fields of ICTS. Based on the need to integrate the ORU data for the different ICTS business fields, the interdependence between the different business fields is considered to be high.

In terms of the development of the DW structure for the ICTS department of the NMMU, the perceived ability of the developer of the DW structure is considered to be high, based on skills and experience. The level of sponsorship for the development of the ICTS DW structure is the ICTS management who will benefit from the DW structure in terms of more effective and efficient analysis of ORU data to support decision making. Although the level of sponsorship for the development of the DW structure is not recognised at an organisational level, it is to be supported by the top management of the ICTS department of the NMMU.

The scenario for the NMMU case study has to be mapped to a specific scenario in which a particular DW approach is preferred (Table 4-3). Corresponding selection factors, as described in the ICTS scenario and the scenario in which a DW approach is typically selected, are highlighted to represent successful mapping between selection factors and between scenarios.

Table 4-3: Mapping of the ICTS scenario to the scenario in which the IDM DW approach is typically selected.

	Factors Influencing the Selection of the IDM DW Approach		
	Resource Constraints	**Management's Strategic View**	**Perceived Ability of IT Staff**
ICTS Scenario	**High**	Departmental	High
Favourable IDM Approach Selection Scenario	**High**	Organisational Sub-Unit	Low

The ICTS scenario only maps to one DW approach selection factor in the scenario in which the IDM DW approach is preferred, namely resource constraints, since the development of the proposed DW structure for ICTS is described as being significantly constrained by funds and time. The ICTS scenario is described as having staff with high ability, as perceived, to develop a data warehousing solution for the DW structure by the NMMU with a strategic departmental view. These factors for the ICTS scenario do not map to the corresponding factors in the scenario in which the IDM DW approach is favoured. The mapping between the ICTS scenario and the typical EDW approach selection scenario consists of two selection factors which correspond to each other, namely high organisational interdependence and high perceived ability of IT staff (Table 4-4). The ICTS scenario in contrast, consists of a strategic view that is departmental and does not correspond to the enterprise-wide strategic view for which an EDW approach is typically used.

Table 4-4: Mapping of the ICTS scenario to the scenario in which the EDW approach is typically selected.

	Factors Influencing the Selection of the EDW Approach		
	Organisational Interdependence	**Management's Strategic View**	**Perceived Ability of IT Staff**
ICTS Scenario	High interdependence between ICTS business fields	Departmental	High
Favourable EDW Approach Selection Scenario	High	Enterprise Wide	High

Three DW approach selection factors were identified which map the ICTS scenario to the scenario in which the DMB DW approach is preferred namely, high resource constraints, high urgency and high organisational interdependence (Table 4-5). If a data warehousing project is supported by high levels of organisational management, in addition to high resource constraints, urgency and organisational interdependence, the DMB DW approach is selected.

Table 4-5: Mapping of the ICTS scenario to the scenario in which the DMB DW approach is typically selected.

	Factors Influencing the Selection of the DMB DW Approach			
	Resource Constraints	**Urgency**	**Organisational Interdependence**	**Level of Sponsorship**
ICTS Scenario	High	High	High interdependence between ICTS business fields	Departmental
Favourable DMB DW Approach Selection Scenario	High	High	High	High

In the case of the ICTS scenario it is unclear if a level of sponsorship at a departmental level is considered to be high. The mapping between the typical DMB DW approach selection scenario and the ICTS scenario, however, resulted in the three successfully mapped DW approach selection factors, therefore, the selection of the DMB DW approach can further be justified. This is done by mapping the ICTS scenario to the scenarios (Table 4-6) which specify the selection between the EDW and DMB DW approaches. The EDW and DMB DW approaches have typical selection scenarios which were found to map the most to the ICTS scenario with two and three corresponding DW approach selection factors.

Table 4-6: Mapping between the ICTS scenario and the typical EDW approach selection scenario when selecting between the EDW and DMB DW approaches.

	Factors Influencing the Selection of the EDW Approach above the DMB DW Approach	
	Strategic View	**Urgency**
ICTS Scenario	Departmental	High
Favourable EDW DW Approach Selection Scenario	Enterprise Wide	Low

The ICTS scenario does not have a DW approach selection factor that maps to the scenario in which the EDW approach is typically selected if a decision has to be made between the EDW and DMB DW approaches. In contrast, however, the ICTS scenario successfully maps to the DMB DW approach selection factors which influence the selection of either the EDW or DMB DW approaches. The ICTS scenario corresponds to the DMB DW approach selection factors (Table 4-7) based on the departmental strategic view and a high urgency for a solution of the capabilities of a DW structure.

Table 4-7: Mapping between the ICTS scenario and the typical DMB DW approach selection scenario when selecting between the EDW and DMB DW approaches.

	Factors Influencing the Selection of the DMB DW Approach above the EDW Approach	
	Strategic View	**Urgency**
ICTS Scenario	Departmental	High
Favourable DMB DW Approach Selection Scenario	Departmental	High

Based on the mappings between the ICTS scenario and the different DW approach selection scenarios, the DMB DW approach (Figure 4-10) is selected to define the layout of the different layers of the ICTS DW structure. Based on this finding, a dimensionally modelled DW needs to be developed (Section 4.4.2) that consists of various constituent data marts (Jukic 2006) which each represent a business field subject area as described by the logical design of the DW (Figure 3-5). In addition, only one set of ETL processes has to be designed, when a DMB DW approach is used, to populate the DW structure's DW with data that is sourced from the identified ICTS source systems (Section 3.4.1).

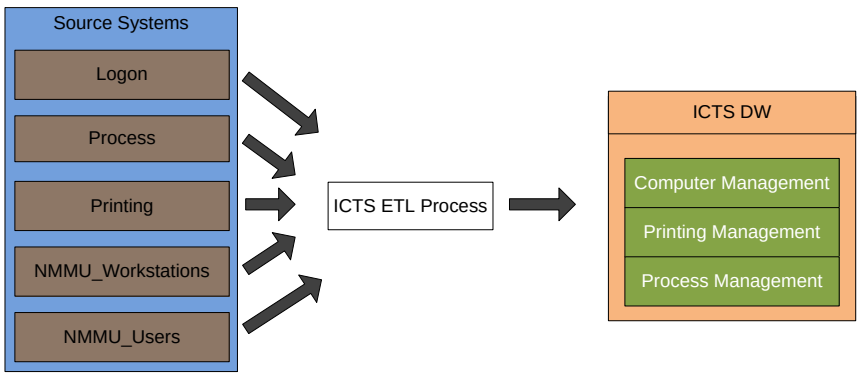


Figure 4-10: The DMB DW approach selected to be applied to the ICTS case study.

4.4.2 The Physical Design of the ICTS DW

The DMB DW approach was selected (Section 4.4.1) as the most appropriate DW approach based on the described DW approach selection scenarios (Section 4.2.5.2). This DW approach requires that a dimensional DW be implemented for the ICTS DW structure (Jukic 2006; Ariyachandra and Watson 2010). The subject oriented enterprise data schema (Figure 3-5) was derived based on the analysis of ICTS information requirements and source data characteristics (Chapter 3). This data schema represents the information requirements of ICTS which are consolidated with the relevant source data and represents the logical design of the ICTS DW structure's DW. Based on the selection of the DMB DW approach and the derived logical design of the ICTS DW, the physical design of the DW is derived (Figure 4-11) as part of the physical design phase of the application of the ATDM (Figure 4-1, Step 4.3) and is represented by an entity relationship diagram (Chaudhuri and Dayal 1997).

To ensure that the fact tables of the DW do not contain significant amounts of redundant data, dimension attributes for the time dimension (Table 3-11) are removed from their respective tables. This is done to create a physical dimension table in which the time dimension attribute values are stored with no significant redundancy in fact tables. The creation of the Time dimension table requires that a primary key is assigned, based on the unique combination of attributes for each record it contains. This primary key serves as a foreign key in the respective tables from which the time attributes were originally removed.

The destinationOU column in both the NMMU_Workstations and NMMU_Users tables in the logical design of the DW structure's DW (Figure 3-5) represents a concatenation of attributes which describe computer workstations and users respectively. These attributes must be split into the individual descriptive attributes to supplement the physical design of the DW to represent the most detailed description of the available NMMU ORU data sources. Based on the decomposition of the destinationOU column attributes of the NMMU_COMPUTERS and NMMU_USERS source tables, computer workstations and users were identified as being categorised into three different groups. This categorisation was done based on the number of similar descriptive attributes derived from the respective destinationOU column.

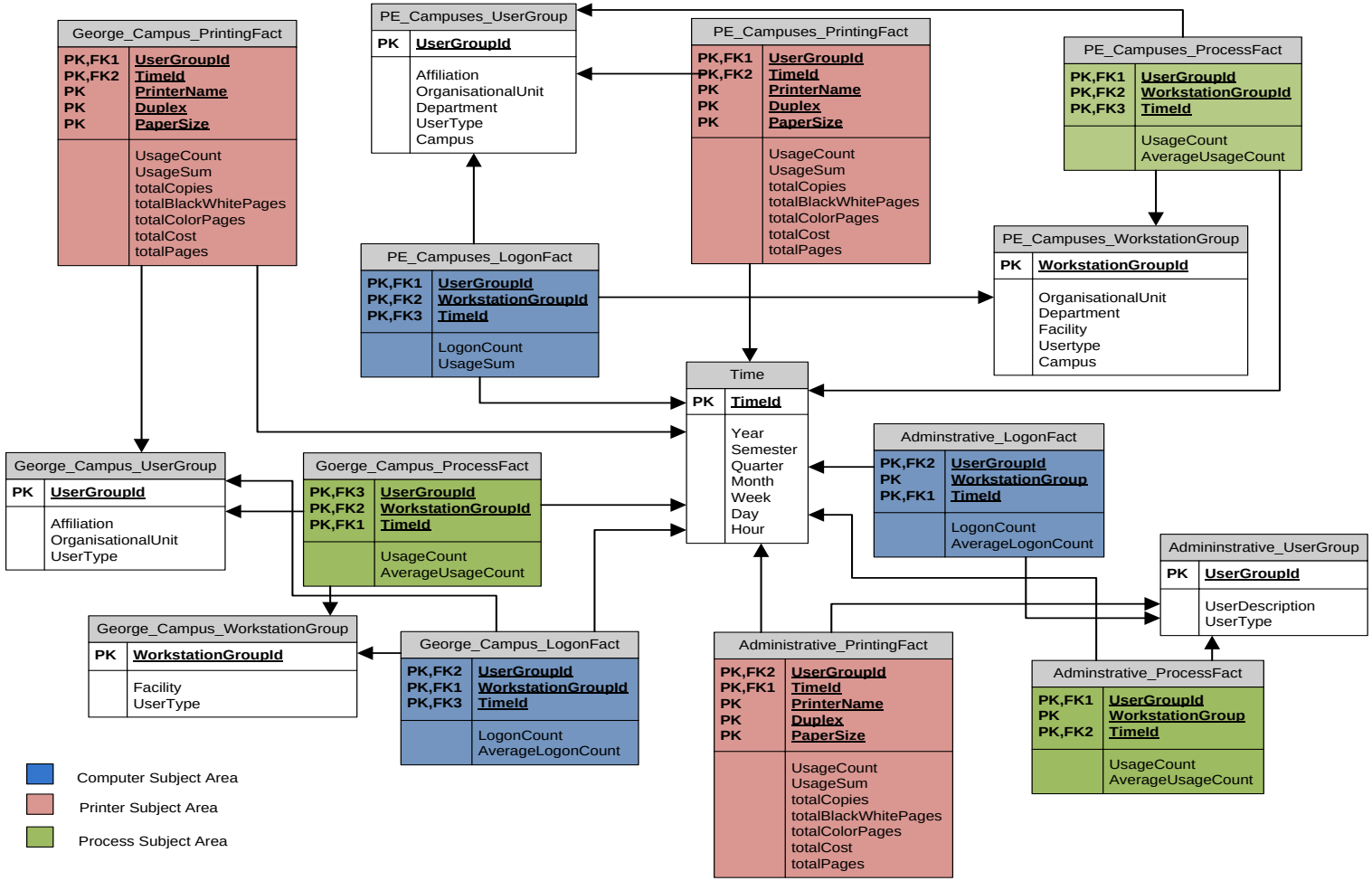


Figure 4-11: The physical design of the ICTS DW.

Users and computer workstations were categorised into the PE campuses-, George campus- and administrative groups for users and workstations respectively. This resulted in the creation of separate fact tables and dimensions for each group identified. Each fact table represents a data mart that corresponds to a subject area (Section 3.2.5) of the corresponding ICTS business fields (Section 3.2.3). The fact tables with their related dimensions, which are highlighted in blue, represent the computer management business field's subject area. The fact tables, which are highlighted in red, represent the printer subject area and these which are highlighted in green represent the process subject area. Each record in a fact table relates to only one record in the related dimension tables on the lowest level of aggregation. Similarly, one record in a dimension table may relate to one or to many records in related fact tables.

4.4.3 The Physical Design of the ICTS ETL Processes

In order to populate the ICTS DW with relevant source data a set of ETL processes must be designed. The application of the supply-driven phase of the ATDM (Section 3.4) resulted in the identification of relevant source tables and columns which are needed to populate the DW. In addition to the identified source tables- and columns semantics were homogenised (Section 3.4.5) which represents the logical design of the ETL processes required to populate the ICTS DW.

The DMB DW approach (Figure 4-10) illustrates that a single set of ETL processes are needed to populate the ICTS DW structure's DW, since the physical design of the DW (Figure 4.11) consists of a single dimensional data model with of a set of fact tables with shared dimensions. It is required that the physical design of the ETL processes for this case study is derived (Figure 4-1, Step 4.4) and documented as a set of technical metadata. The physical design of the ETL processes was derived based on the semantic metadata derived in the application of the supply-driven phase (Section 3.4) of ATDM, the selected DMB DW approach and the physical design of the ICTS DW structure's DW.

The physical design of the ICTS ETL processes is represented using the proposed ETL modelling approach (Section 4.3). A representative sample of the ICTS DW structure's ETL physical design is presented (Figure 4-12) to demonstrate the application of the proposed ETL modelling approach for the NMMU ICTS case study.

A temporary table, TempList C, is generated as intermediate results which will be used to populate destination tables represented in the physical design of the DW structure's DW. The StartDate column attribute must be transformed from a datetime2(3) format to a Date attribute format and is represented by a conversion ETL mechanism. This is done to ensure a consistent format in their representations in related dimension tables. Conversion attributes are annotated by using a note component in which the ETL conversion mechanism is discussed. The LogonId, PCNAME, and UserName columns do not require transformations to be applied thus their respective formats in the Logon table remains the same in the destination table. A filter mechanism is used to ensure that all data records in the Logon table, with a StartDate attribute value that represents a date prior to the last successful execution, are not loaded to TempList C as part of the incremental updating of the destination DW. Once the transformations have been applied to the selected Logon table columns, they are loaded to Templist C that represents intermediate ETL results.

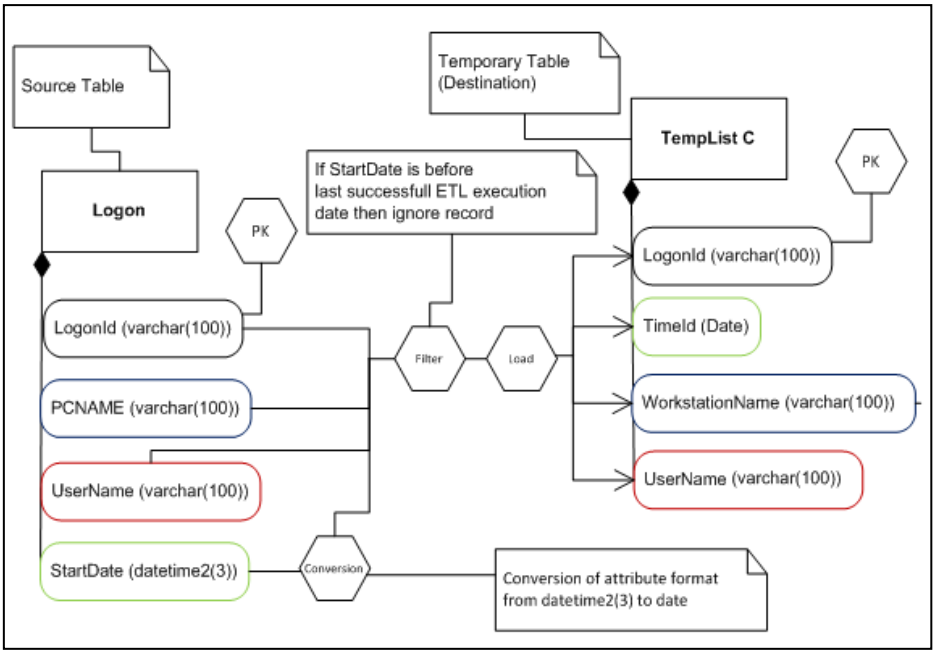


Figure 4-12: A Representative Sample of the DW Structure's ETL Processes Physical Design.

The loading process of the data sourced from the Logon table to the intermediate table, TempList C, is represented by a load ETL mechanism. The respective columns, which are loaded to TempList C, are represented in TempList C with corresponding colours (Section 4.3). The PCNAME column of the Logon Fact Table is represented in TempList C as WorkstationName that results from the homogenisation of the semantics step (Section 3.4.5) of the ATDM's supply-driven phase (Section 3.4).

Appendix C represents a subset of the ICTS ETL processes physical design which has to populate the DW with source data. This representation of the physical design of the ETL processes is representative of the ETL design for the NMMU ICTS case study and demonstrates the use of the proposed ETL modelling technique to populate different tables.

4.5 Conclusion

This chapter described the application of the physical design phase of the ATDM. This phase required a DW approach to be selected (Figure 4-1, Step 4.1 and 4.2), a physical design of the ICTS DW to be derived (Figure 4-1, Step 4.3) as well as the derivation of the physical design of the required D ETL processes (Figure 4-1, Step 4.4). In order to enable the application of the physical design phase on the case study DW approaches and methods of selecting an appropriate DW structure were investigated (Section 4.2). In addition DW ETL processes were investigated and a modelling approach for ETL processes was proposed based on a combination of existing ETL modelling techniques (Section 4.3).

The physical design of the ICTS DW structure's DW and ETL processes were derived as part of the evaluation of the methodology as a proof of concept (Section 2.5). The physical design of the DW (Figure 4-11) is directly related to the logical design (Figure 3-5) and the selected DMB DW approach (Figure 4-10). This proves that the ATDM successfully provided methodological support for the derivation of the physical design of the ICTS DW structure's DW. Similarly the physical design of the ICTS DW structure's ETL processes were derived (Figure 4-12) based on metadata from the supply-driven phase, DMB DW approach and the physical design of the DW. The following chapter describes the application of the final phase of ATDM in which the physical design of the ICTS DW structure will be implemented.

Chapter 5. ATDM Implementation Phase

5.1 Introduction

The output produced as a result of the application of the ATDM's physical design phase to the ICTS case study (Chapter 4) resulted in the derivation of the physical design of the ICTS DW structure. The design includes the physical design of the DW structure's DW (Section 4.4.2) and the ETL processes (Section 4.4.3) which together represent a set of technical metadata that was derived and documented, and describes the physical design of the ICTS DW structure. This chapter describes the implementation phase of the ATDM (Figure 5-1) and addresses the fifth research question for the research (Section 1.6) that requires that a DW structure's physical design is implemented using the MSS integrated tool.

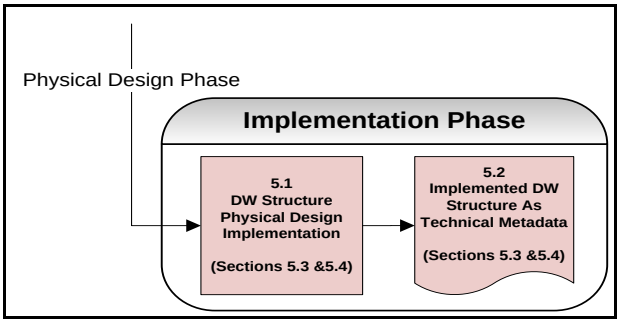


Figure 5-1: The implementation phase of the ATDM.

The physical design phase of the ATDM is required to be applied to the case study through implementing the physical design of the DW structure (Chapter 4) using Microsoft SQL Server (MSS) (Section 1.8). The application of the physical design phase forms part of the evaluation of the proposed ATDM as a proof of concept (Section 2.5). The MSS environment is required to support the implementation of the DW's physical design as well as the physical design of the ETL processes. In addition MSS should also allow the implementation of the DW structure to be described by technical metadata at an implementation level. The MSS environment that will be used to implement the physical design of the DW is described (Section 5.2). The implementation of the physical design of the DW is described (Section 5.3) as well as the implementation of the physical design of the DW structure's ETL processes (Section 5.4).

5.2 The Microsoft SQL Server (MSS) Environment

MSS was used to implement the physical design of the ICTS DW structure (Chapter 4) that was derived using the ATDM (Figure 2-5). MSS consists of a set of tools used to store, manage and maintain data (Microsoft 2012c). The set of tools contained in the MSS environment consists of a database engine, ETL tools, reporting tools, management interfaces and analysis tools. The implementation of the ICTS DW structure design required the utilisation of three components offered by the MSS environment, namely:

1. SQL Server Database Engine (SSDE);
2. SQL Server Management Studio (SSMS); and
3. SQL Server Integration Services (SSIS).

The SSDE is a relational, database management system (RDMS) that is used to host relational databases and allows the management and administration of the server on which databases are housed (Microsoft 2012c). In addition, the SSDE administers the rapid processing of transactions issued by applications or users. To allow users to manage and manipulate data, SSMS enables users to access and interact with the SSDE through the creation and management of databases housed in the SSDE (Microsoft 2012d). In addition, SSDE consists of the functionality that allows users to create and issue SQL queries on targeted databases.

A SSIS package is created using the MSS component, Business Intelligence Development Studio (BIDS) that allows the creation of ETL processes. SSIS packages consist of a control flow that represents a workflow of control items which is executed in a SSIS package (Opensource-Blogs 2011). The BIDS interface that is used to create and specify a control flow for a SSIS package is represented (Figure 5-2). Control flow items are categorised into three different types of items, namely (Microsoft 2012a).

- Control flow tasks;
- Containers; and
- Precedence constraints.

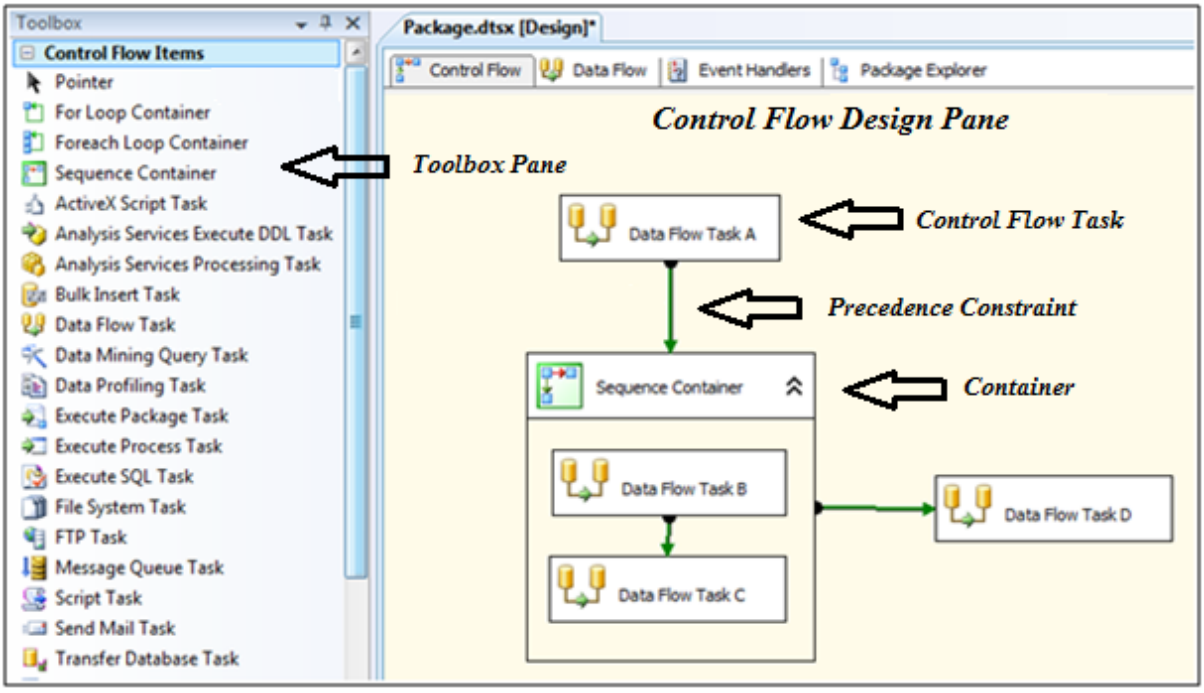


Figure 5-2:The BIDS control flow specification interface (Microsoft 2012a).

Control flow tasks represent the executable control flow items of a SSIS package which allows users to create a workflow of executable tasks dedicated to the management and maintenance of data (Mahadevan 2011). The BIDS interface allows users to drag a selected control flow task from the interface's toolbox pane to the control flow design pane in which a SSIS package's control flow is designed (Figure 5-2). The order in which tasks are executed in a SSIS package's control flow, is specified by precedence constraints represented by the green arrows in Figure 5-2 (Microsoft 2012a). A precedence constraint represents a link between two control flow tasks and constrains the execution of the destination control flow task based on the execution result of the preceding control flow task and the condition that is specified (Microsoft 2012b). Containers, which are also contained in the BIDS toolbox pane, are control flow items into which control flow tasks can be grouped and executed to provide structure to the control flow of a SSIS package (Microsoft 2012a).

To design an ETL processes for a DW structure, data flow tasks are used (Opensource-Blogs 2011). Data flow tasks are control flow tasks which are used to extract data from sources, apply necessary transformations and load data to specified data destinations. In order to use a data flow task as part of a control flow for a SSIS package, data flow task needs to be dragged to the package's control flow design pane (Figure 5-2). BIDS provides an interface in which a data flow task is specified (Figure 5-3).

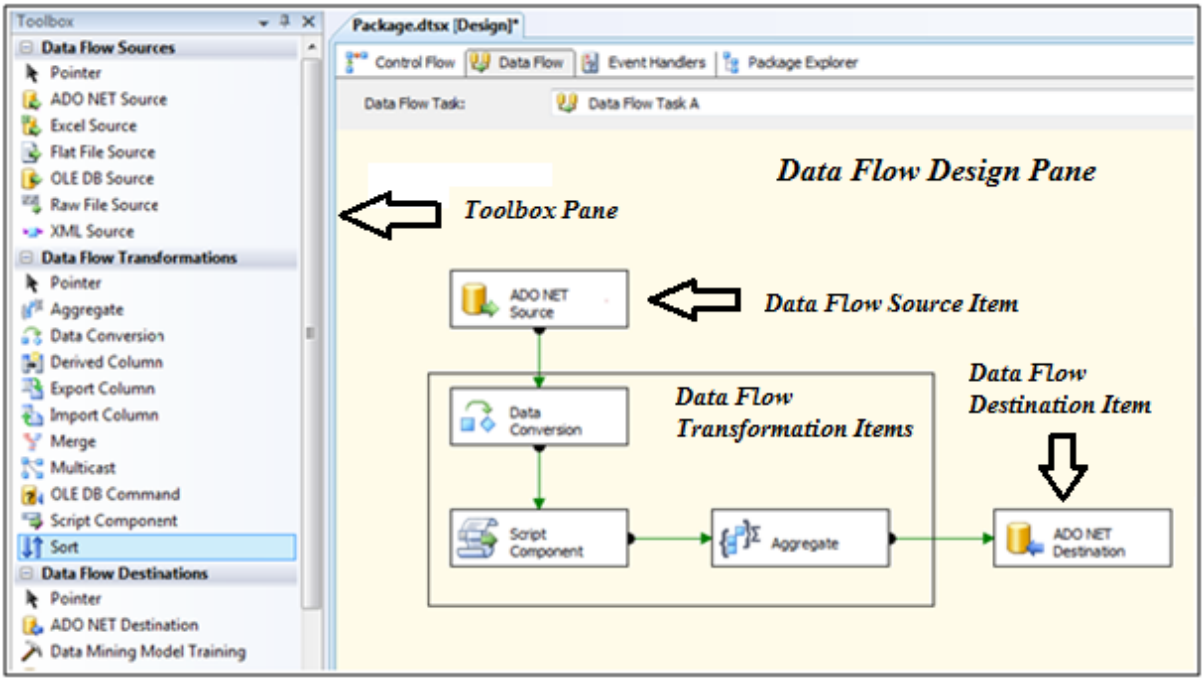


Figure 5-3: The BIDS data flow specification interface.

Figure 5-3 shows the ETL specifications of a data flow task created as part of a SSIS package's control flow (Figure 5-2). The data flow items are categorised into data flow source items data flow transformation items and data flow destination items (Opensource-Blogs 2011) (Figure 5-3). The data flow source items are used to extract data from different types of source data files which include databases, Microsoft Excel files, flat files or XML sources. Similarly, data flow destination items are used to load data propagated from data sources to a specified data file. Data flow transformation items are used to manipulate the data extracted from data sources into a format that is relevant to the destination to which the source data is propagated. SSIS offers a large variety of data flow transformation items which can be applied as part of a DW structure's ETL process.

5.3 Implementation of the DW

The physical design of the ICTS DW structure's DW (Figure 4-11) is implemented (Figure 5-1, Step 5.1) on the SSDE using SSMS. The physical design of the ICTS DW consists of sixteen different tables which include nine fact tables and seven dimension tables (Section 4.4.2). These tables were implemented as part of a database that is hosted on the SSDE. The implementation of the physical design of the DW required the specification of database tables. In addition, relationships between tables were specified by means of primary- and foreign key columns (Figure 4-11).

The implementation of the ICTS DW's physical design resulted in creation of a diagram in SSMS that represents the structure and relationships of the tables which were implemented. Figure 5-4 represents a representative sample of the implementation of the DW structure's DW (Figure 4-11). The representative sample of the DW's physical implementation consists of the structure of the PE_Campuses_LogonFact table as well as the structures and relationships with its respective dimension tables as represented on the diagram generated using SSMS (Section 5.2). This diagram represents a set of technical metadata (Section 1.1) at an implementation level (Figure 5-1, Step 5.2).

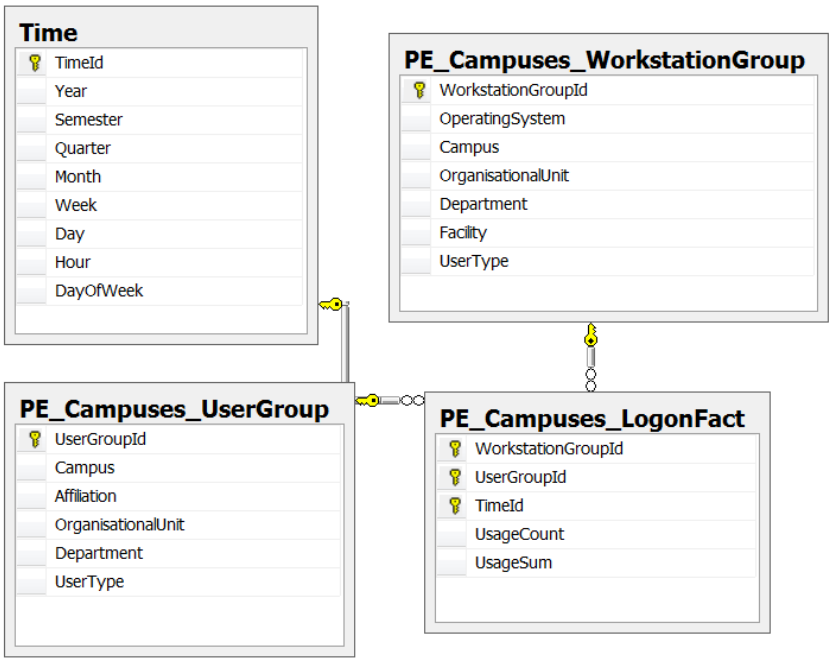


Figure 5-4: Implementation of the PE_Campuses_LogonFact table and related dimensions on the SSDE using SSMS.

5.4 Implementation of the ETL Processes

The physical design of the ETL processes (Section 4.4.3) was implemented (Figure 5-1, Step 5.1) using SSIS (Section 5.2). A SSIS package was created for which a control flow was developed (Figure 5-5). The control flow consists of three sequence containers which each contain a set of data flow tasks. Each sequence container in the control flow represents the population of the DW's tables (Figure 4-11) which relates to a specific category of users and workstations, namely, PE campuses, George campus and administrative (Section 4.4.2). The first sequence container represents a subset of the control flow in which the DW's tables, which represents PE campus users and workstations, are populated. Similarly, the second and

third sequence containers represent the control flow for George campus and administrative related DW table population.

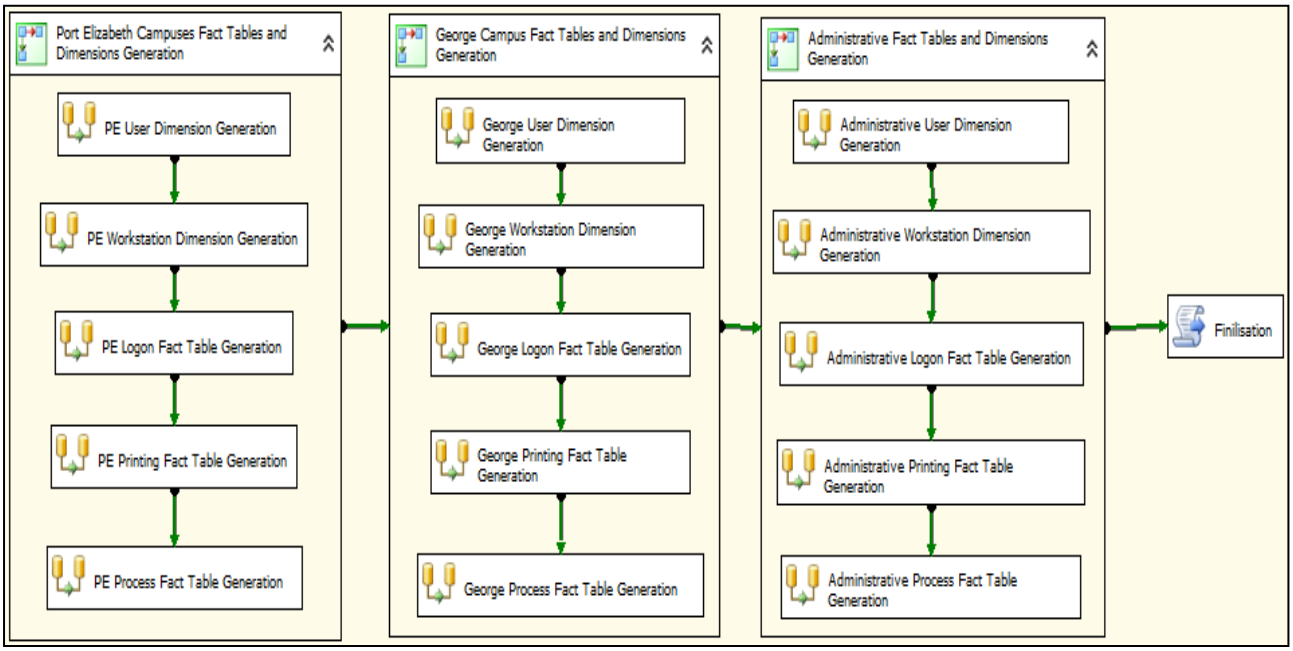


Figure 5-5: The control flow implemented for the ICTS DW ETL processes.

In addition to the sequence containers and their respective data flow tasks, a script component, annotated as "Finalisation", was included to be executed after all data flow tasks in all sequence containers were executed. This task is required to remove all temporary tables which were created and used by the ETL processes (Section 4.4.3).

The physical design of the ICTS ETL processes (Section 4.4.3) specifies the use of components which are required to extract data from sources, apply a set of conversions on subsets of data, aggregate data, filter data based on some criteria and load data to the respective tables of the DW (Figure 4-11) . The need for these components resulted in the identification of a set of relevant data flow items (Figure 5-6) which are required to develop each data flow task that is contained in the control flow (Figure 5-5).

The data flow items which are used to implement the respective data flow tasks include the aggregate-, data conversion-, script component-, ADO NET source and ADO NET destination data flow items. The ADO NET source and ADO NET destination items are used to extract and load data from data sources to data destinations respectively. The aggregate item is used to aggregate sets of data based on a set of input data. The data conversion item

enables designers to specify conversions from one data type to another. The script component data flow item allows designers to implement custom data transformation operations in a data flow task which cannot be supported by the existing data flow items of a data flow task.

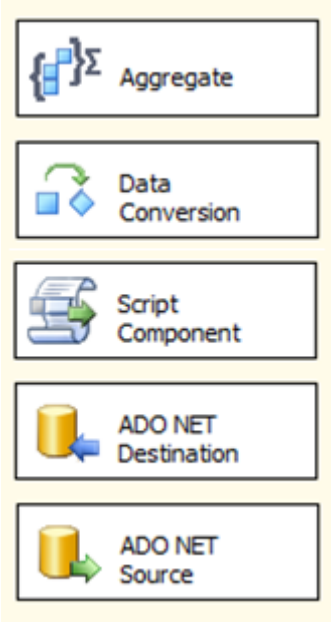


Figure 5-6: The SSIS data flow task's data flow items used for the implementation of the ICTS DW ETL processes.

The population of a DW table of the ICTS DW structure that was implemented using the SSDE (Section 5.3) is represented by a single data flow task in the control flow developed for the ICTS DW structure (Figure 5-5). The specification of the data flow tasks which are used for the population of the respective dimension- and fact tables are similar for PE campus-, George campus- and for administrative users and workstations. Therefore the data flow task specification for only PE campus users and workstations are described for illustrative purposes. In addition the data flow task specification of the respective workstation group dimensions are similar to the specification of user group dimensions. The data flow task specification to populate dimension tables are described (Section 5.4.1). The population of fact tables are demonstrated with the specification of the data flow task used to populate the PE_Campuses_LogonFact fact table (Section 5.4.2).

5.4.1 Dimension Table Population

The ETL processes for populating user group dimensions in the DW were implemented with using a script component and a data flow destination item in the specification of the respective data flow tasks. The ORU source data that is used to populate the PE_Campuses_UserGroup dimension table is obtained using a SQL query that is issued on the NMMU_USERS source table within the PE User Group Generation and Transformation

script component (Figure 5-7). This script component is also used to filter the data which is loaded from the NMMU_USERS source table as described in the physical design of the ETL processes (Section 4.4.3).

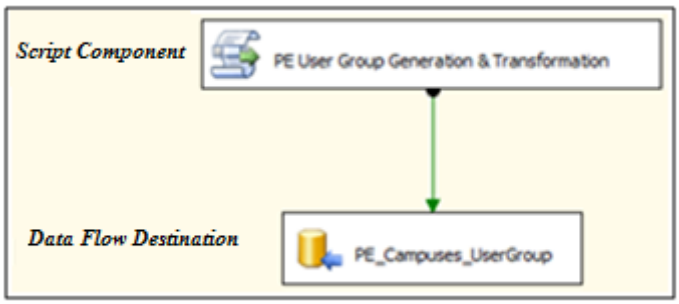


Figure 5-7: The data flow task items used to populate the PE user dimension.

The physical design of the DW structure's ETL processes (Section 4.4.3) specifies that a temporary list, namely, PE_Campuses_User_Group_List is created to be used for the population of PE related fact tables. Although data can be passed between the items of a data flow task, data cannot be passed directly from one data flow task to another in a control flow. In order to allow the data contained in the PE_Campuses_User_Group_List temporary list to be accessible in different data flow tasks in a control flow, a temporary list was created in the destination database as a table from which the data can be accessed by various data flow tasks.

A data flow destination item was not applicable for the loading of data into the temporary list PE_Campuses_User_Group_List. This component cannot be used since the specification of a data flow destination item in a data flow task requires a destination table to be created in the destination database before the SSIS package is executed. The PE_Campuses_User_Group_List does not form part of the physical design of the DW (Figure 4-11) and is only required for the execution of the ETL processes. This restriction was addressed through the execution of SQL queries from the script component in which the table was created and populated with relevant data.

5.4.2 Fact Table Population

The specification of the data flow task needed to populate the PE_Campuses_LogonFact table (Figure 5-8) consists of the use of: an ADO NET data flow source item; two ADO NET data flow destination items; a data conversion item; two data aggregation items; and two script components. The Logon ADO NET data flow source item is used to load each record

from the Logon source table individually to the ETL workflow created for the population of the PE_Campuses_Logon_Fact table.

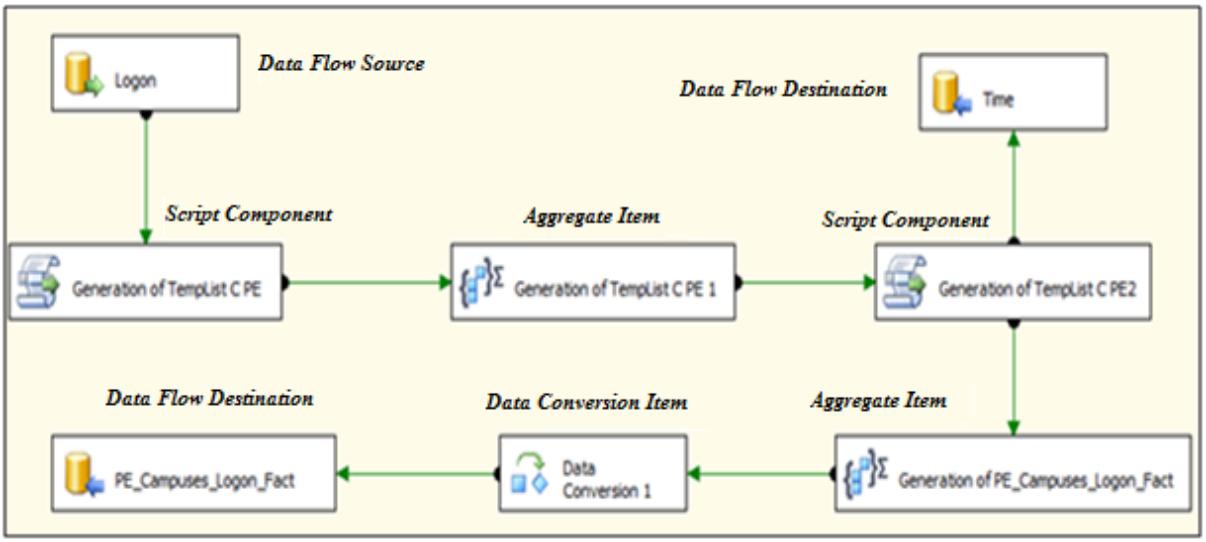


Figure 5-8: The data flow task items used to populate the PE_Campuses_Logon_Fact table.

The data records, which are loaded with the data flow source item, are propagated to the Generation of TempList C script component which applies a transformation on each record and immediately propagates the record to the following item in the sequence. Data records which are loaded to the data flow are filtered to only extract records from the Logon table which contains a value for the columns UserName and PCName that is contained in the respective temporary lists represented in the physical design of the ETL processes (Section 4.4.3). An additional filtering condition was implemented needed for the incremental the incremental updating of the fact table. This is required to ensure that data that is already loaded to the destination table is not repeatedly considered when the DW structure's DW is updated.

The filtered data is passed from the Generation of TempList C PE script component to the Generation of TempList C PE 1 aggregate item that provides the data required for the temporary list, TempList C PE 1, defined in the physical design of the ETL processes (Section 4.4.3). This item is required to perform an aggregation task on the data and an aggregation is defined through the selection of specified input columns as "group by" columns and the specification of measures in the aggregate item's specification wizard (Figure 5-9).

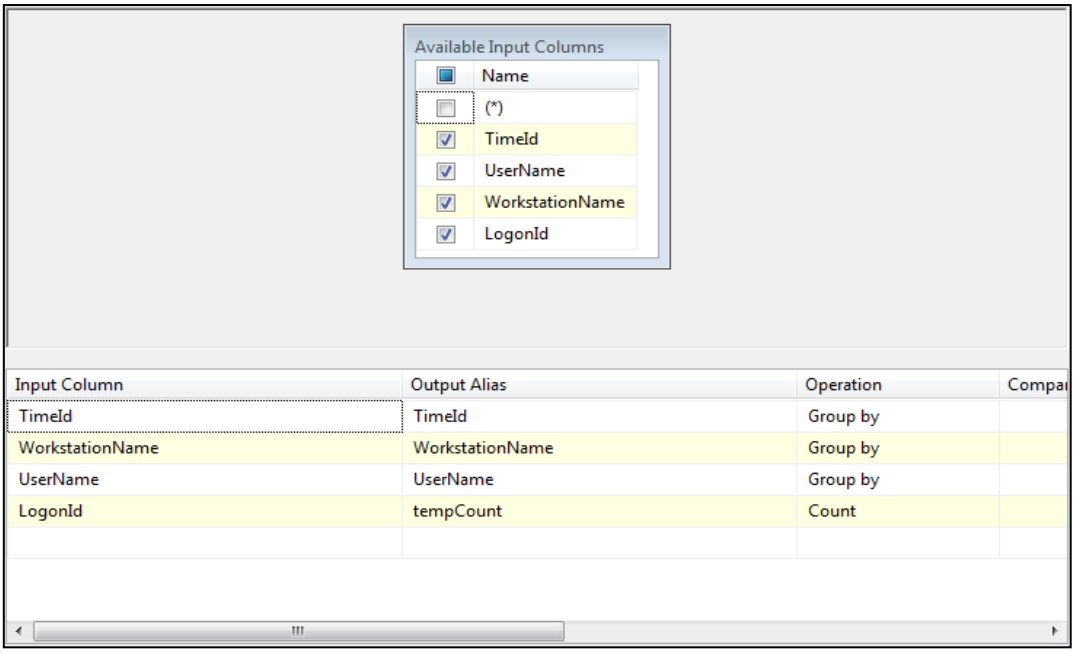


Figure 5-9: Aggregate data flow item specification wizard.

After the TempList C PE 1 temporary list has been populated, each of its records is propagated to the Generation of TempList C PE 2 script component in which the temporary list TempList C PE 2 is generated. Each record in this list is sent to the subsequent aggregation component which performs the required aggregation operation to populate the records for the PE_Campuses_LogonFact table which are then loaded to the destination table using the PE_Campuses_LogonFact data flow destination component.

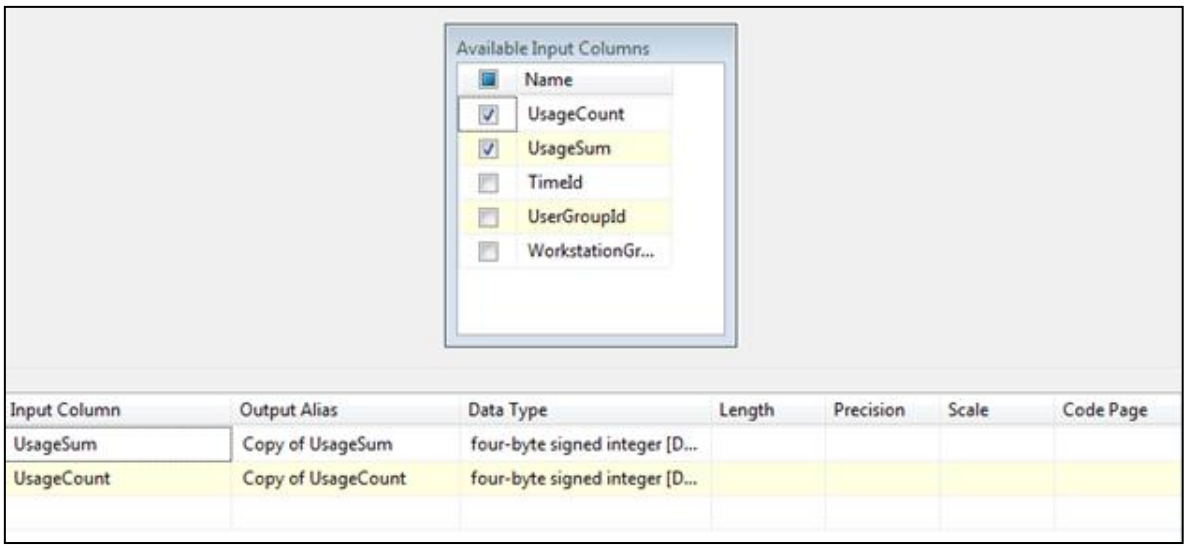


Figure 5-10: Specification of data conversion using SSIS data flow conversion component.

The data conversion component (Figure 5-8) is used to convert the numerical measures obtained from the previous aggregation component from an 8-byte unsigned integer format to a 4-byte signed integer format. The DW structure's DW implementation supports the 4-byte signed integer format and the conversion is required since the aggregation component's output is limited to the 8-byte unsigned integer format. Figure 5-10 represents the conversion component's wizard in which the conversion is specified. The columns which are required to be converted to a new format are selected from the list of available input columns. The selected columns are displayed in which the required conversion of data is specified.

5.5 Conclusion

The physical design of the ICTS DW (Figure 4-11) was implemented as a relational SQL server database that is hosted on a SSDE (Section 5.3). The physical design of the ICTS ETL processes (Section 4.4.3) was implemented using the MSS environment's SSIS and required the identification of relevant SSIS components to be used. Although various components were identified as applicable to the implementation of the ETL processes, a shortcoming was identified (Section 5.4.1). Data flow tasks have the ability of propagating data between their respective data flow components but are unable to propagate data to other data flow tasks. This shortcoming was addressed with the execution of SQL queries in script components in which a temporary list was created which contained information that was able to be accessed by all data flow tasks. The population of the respective fact tables (Section 5.4.2) required the introduction of a data conversion item in the specification of the respective data flow tasks. The conversion items were introduced to address the limitation of the data flow aggregation item's inability of supporting the 4-byte signed integer format of type integer (Section 5.4.2).

The physical design of the ICTS DW (Figure 4-11) was successfully implemented on the SSDE using SSMS (Section 5.3). Similarly the physical design of ICTS ETL processes (Section 4.4.3) were successfully implemented using SSIS (Section 5.4). The implementation of the physical DW structure design resulted in the creation of technical metadata at an implementation level which was created based on the specifications of the respective metadata describing the physical design of ICTS DW structure. The implementation of the ICTS DW structure needs to be evaluated for effectiveness and efficiency which are described in the next chapter.

Chapter 6. DW Structure Effectiveness and Efficiency Evaluation

6.1 Introduction

The physical design of the ICTS DW structure was derived (Chapter 4) and the implementation of the physical design was described (Chapter 5) as part of the application of the ATDM (Section 2.5) on the ICTS case study. Each phase of the ATDM was applied to the case study to evaluate the methodology as a proof of concept (Chapters 3, 4 and 5). This chapter addresses the sixth research question that was presented for this research (Section 1.6) which requires that the DW structure that is derived from the application of the ATDM to be evaluated for effectiveness and efficiency. The design for the effectiveness and efficiency evaluation is described (Section 6.2). The results of the evaluation are presented and described (Section 6.3).

6.2 ICTS DW Effectiveness and Efficiency Evaluation Design

The main goals of the evaluation of the derived ICTS DW structure are to measure the effectiveness and efficiency of the proposed DW in providing support for the analysis of ORU data. The DW structure is required to be effective in supporting ICTS to efficiently obtain answers for ad hoc queries (Section 1.2). The design for the effectiveness evaluation of the implemented DW structure (Chapter 5) is described (Section 6.2.1) as well as the design of the DW structure's efficiency evaluation (Section 6.2.2). The experimental design of the DW structure's effectiveness and efficiency evaluation is described (Section 6.2.3).

6.2.1 Effectiveness Evaluation

The DW structure that was derived for the ICTS case study, through the application of the ATDM, needs to support the effective analysis of ORU data. The derived DW structure should provide the functionality that ensures that the data contained in the DW is an accurate aggregated representation of the data contained in the respective data sources of the DW structure. The effectiveness of the derived DW structure is evaluated by measuring the quality of the functionality provided by the DW structure in providing accurate data for the analysis of ORU data. Since the effectiveness of the DW structure is determined by the quality of its functionality, the effectiveness of the derived DW structure is evaluated by measuring its accuracy.

The accuracy of the DW structure is determined by comparing the observed ad hoc query outputs of the DW structure to the expected output of ad hoc queries to determine the total number of DW structure ad hoc query output rows, for each ad hoc query, which are different to the expected ad hoc query output rows (ISO 2003). A difference between the observed query output rows and expected rows are determined by the total number of query output rows which are observed and are not in the set of expected output rows for a query.. The equations which will be used to calculate the accuracy of the derived DW structure are presented below.

$$X_i = \frac{A_i}{B_i} \qquad \text{Equation 6.1}$$

where

$X_i = difference\ ratio\ for\ ad\ hoc\ query\ i$

$A_i = \#\ DW\ structure\ ad\ hoc\ query\ output\ rows\ that\ differ\ from\ expected\ for\ query\ i; and$

$B_i = \#\ ad\ hoc\ query\ output\ rows\ expected\ for\ query\ i$

$$Y = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} Bi} \qquad \text{Equation 6.2}$$

where

$Y = total\ ad\ hoc\ query\ difference\ ratio$

$A_i = \#\ DW\ structure\ ad\ hoc\ query\ output\ rows\ that\ differ\ from\ expected\ for\ query\ i; and$

$B_i = \#\ ad\ hoc\ query\ output\ rows\ expected\ for\ query\ i$

$n = total\ \#\ ad\ hoc\ queries$

The output of a set of ad hoc queries are required to be obtained using the derived DW structure (Figure 6-1) and the typical process applied by ICTS (Figure 6-2) for obtaining output rows for ad hoc queries. Ad hoc queries' outputs are obtained by using the ICTS DW structure through executing the ETL processes once, to populate the DW structure's DW with source data (Figure 6-1). Once the DW has been populated, each ad hoc query is executed on the DW to obtain the output for each ad hoc query.

The output for ad hoc queries are obtained by using the typical ICTS process through executing a query on the DW structure's source data to populate a temporary data store on which an ad hoc query is executed to obtain its output rows (Figure 6-2). In contrast with the ICTS DW structure's process of obtaining ad hoc query output in which data is extracted from the sources only once (Figure 6-1), the typical ICTS process needs to extract data to a temporary data store for each ad hoc query that is processed.

The query output rows observed using the typical ICTS process (Figure 6-2) represent the expected output of the ad hoc queries considered for this evaluation. The query output rows obtained using the derived DW structure (Figure 6-1) are compared to the expected output rows for the respective queries in order to capture the number of output rows obtained from the derived DW structure which differ from the expected output rows. Based on the number of DW structure query output rows and the total number of expected rows the difference ratio for each query is calculated (Equation 6.1) as well as the total difference ratio (Equation 6.2). If a difference ratio is calculated to be zero for an ad hoc query it can be concluded that a DW structure is effective based on accurate ad hoc query results.

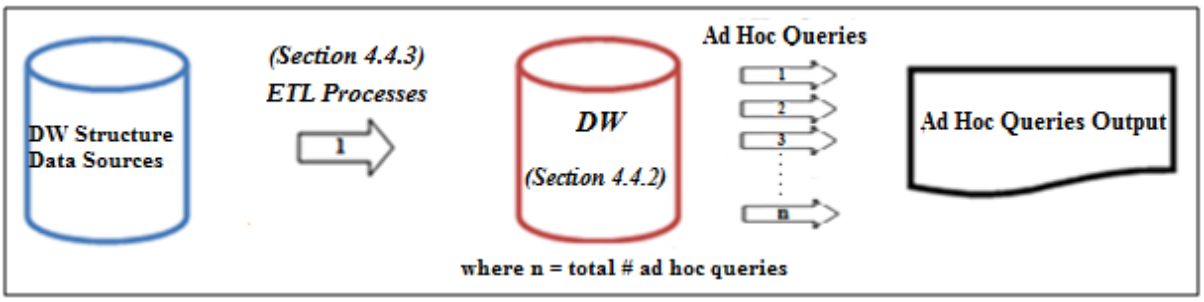


Figure 6-1: ICTS DW structure process of obtaining ad hoc query output.

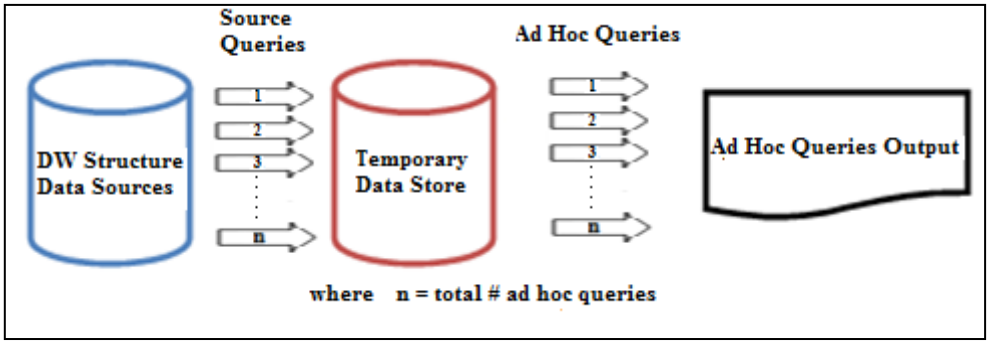


Figure 6-2: Typical ICTS process of obtaining ad hoc query output.

6.2.2 Efficiency Evaluation

The ICTS DW structure is required to support the efficient analysis of ORU data in terms of the amount of time required in which ad hoc query results are obtained. In order to enable the efficient execution of ad hoc queries using the DW structure, the DW structure's DW is required to be populated with source data using the described ETL processes (Section 5.4). The ETL processes of a DW structure typically run in strict time windows, thus it is required that ETL processes are executed as quickly as possible (Simitsis and Vassiliadis 2009). The efficiency of the derived DW structure is evaluated using the response time metric described in (ISO 2003).

The response time metric (ISO 2003) requires that the start- and end times of the execution of a process are recorded in order to obtain the time taken for a particular function to execute. The response time efficiency metric is used in two different experiments in which the efficiency of the derived DW structure is evaluated. The first experiment involves testing the efficiency of the ETL processes in populating the DW structure's DW with source data. The second efficiency evaluation experiment tests the efficiency of the derived DW structure in supporting ad hoc queries (Figure 6-1) compared to the typical ICTS process (Figure 6-2).

The aim of the first efficiency evaluation experiment is to determine whether the derived DW structure's ETL processes supports the efficient propagation of ORU source data from the various source systems to the DW structure's DW. A DW structure's DW is typically updated daily at night in order to ensure that data that was generated during a working day is propagated to the DW at a time which allows an acceptable recovery time window in case of an ETL failure (Vassiliadis and Simitsis 2009). Additionally the execution of a DW structure's ETL processes is typically preferred at night since at this time of day source systems are not experiencing high usage (Simitsis and Theodoratos 2009).

The derived ICTS DW structure's DW is required to be populated at midnight each day in order to ensure that the ETL processes are executed before source system usage commences at the start of the following working day that is eight hours after midnight. In order to determine whether the ICTS DW ETL process is able to propagate daily generated source data to the DW structure's DW during a eight hour time period, the ETL process were executed for source data for a specified number of days.

The equation that was used to calculate the efficiency of the derived DW structure's ETL processes is presented below (Equation 6.3).

$$W_i = |D_i - C_i| \quad \text{Equation 6.3}$$

where

$W_i = DW\ Structure\ ETL\ processes\ execution\ time\ for\ day\ i$

$C_i = DW\ Structure\ ETL\ processes\ execution\ start\ time\ for\ day\ i$

$D_i = DW\ Structure\ ETL\ processes\ execution\ end\ time\ for\ day\ i$

The execution response times of the ETL processes for a specified number of days are captured (Equation 6.3) in order to enable the derivation of descriptive statistics derived from the execution times of the ETL process for the specified number of days. Based on the ICTS case study, the DW structure's ETL processes is considered to be efficient for the population of the structure's DW with daily generated data, if the ETL processes' execution time is less than eight hours. In addition to the capturing of daily DW structure ETL processes execution times, the size of the set of data that was propagated to the DW structure's DW was captured to enable the analysis of the effect the size of data that is transmitted to the DW has on the ETL processes' execution times.

The second efficiency evaluation comprises of evaluating the efficiency of the derived DW structure in processing ad hoc queries (Figure 6-1) compared to the typical ICTS process of processing ad hoc queries (Figure 6-2). The efficiency of the ICTS DW structure and the typical ICTS process were compared in terms of the cumulative ad hoc query response times. The cumulative response times of processing ad hoc queries were used to evaluate the efficiency of the ICTS DW structure in processing a multiple ad hoc queries compared to the typical ICTS process. The equations which were used to calculate the efficiency of the DW structure and the typical ICTS process are presented below.

$$T = |E_{End} - E_{Start}| \quad \text{Equation 6.4}$$

where

$T = total\ DW\ structure\ ETL\ processes'execution\ time$

$E_{Start} = DW\ structure\ ETL\ processes'execution\ start\ time$

$E_{End} = DW\ structure\ ETL\ processes'execution\ end\ time$

$$S_z = (\sum_{i=1}^{z} |G_i - F_i|) + T \quad \text{Equation 6.5}$$

where

S_z = cumulative DW structure ad hoc query processing time for z # ad hoc queries

F_i = DW structure ad hoc query i execution start time

G_i = DW structure ad hoc query i execution end time

T = total DW structure ETL processes' execution time

$$R_z = \sum_{i=1}^{z} (|I_i - H_i| + |K_i - J_i|) \quad \text{Equation 6.6}$$

where

R_z = cumulative typical ICTS process ad hoc query processing time for z # ad hoc queries

H_i = typical ICTS process source query execution start time for ad hoc query i

I_i = typical ICTS process source query execution end time for ad hoc query i

J_i = typical ICTS process ad hoc query i execution start time

I_i = typical ICTS process ad hoc query i execution end time

6.2.3 Experimental Design

A set of data was made available by ICTS that was generated during a fourteen day period. The set of source data was housed on an instance of SQL server that was installed on the computer system on which the evaluation processes were executed. An application was developed which were used to capture the execution times of the different processes which were required to be developed to represent the typical ICTS process of obtaining results. In addition the execution times of the ETL processes were captured using the logging functionality provided by SSIS.

The execution times of queries issued on the data sets, which include the derived DW structure's DW and the data tables generated by the typical ICTS process, was captured using the query execution time functionality that is built into SQL Server Management Studio. A set of ICTS ad hoc queries was identified as a representative sample of ad hoc queries which were used in the effectiveness and efficiency evaluation of the ICTS DW structure (Sections 6.2.1 and 6.2.2). The identified set of ad hoc queries are representative of the ICTS business questions (Section 3.3.1) derived from the application of the ATDM.

The ad hoc queries which were identified are:

1. Which labs in the Computing Sciences department have the highest usage rate, at what time of day?
2. What are the top processes utilised during university office hours?
3. What are the top processes utilised over lunch period?
4. To what extent do users make use of double-sided page printing on a daily basis?
5. What groups of users uses university labs the most during lecture times?
6. What groups of users uses university labs the least during lecture times?
7. What functions do users perform most on their designated computers?
8. Which processes are used most often after hours by which groups of users?
9. How often do particular groups of users make printouts and when?

6.3 Effectiveness and Efficiency Evaluation Results

The effectiveness evaluation design (Section 6.2.1) and the efficiency evaluation design (Section 6.2.2) of the ICTS DW structure's implementation (Chapter 5) was done based on the described experimental design (Section 6.2.3). The results which were captured for the effectiveness evaluation (Section 6.3.1) and efficiency evaluation (Section 6.3.2) of the derived DW structure are described.

6.3.1 Effectiveness Evaluation Results

The output rows for each ad hoc query executed (Section 6.2.3) using the typical ICTS process (Figure 6-2) was captured and compared to the output rows obtained from issuing the respective queries on the ICTS DW structure's DW, after the DW was populated by the structure's ETL processes. The results of the effectiveness evaluation of the DW structure (Table 6-1) shows a difference ratio, X_i (Equation 6.1), for each ad hoc query that ranges between 0 and 0.081. It was observed that 55% (n = 5) of the ad hoc queries approximated a difference ratio of 0 and 22% (n = 2) of ad hoc queries approximated a difference ratio of 0.001. The only queries which have been identified with a difference ratio that is more than 0.01 are queries 1, 5 and 6. A total difference ratio (Equation 6.2) of 0.001 was observed. This indicates that only 0.1 percent of all observed ad hoc query output rows were found to be different from the expected ad hoc query outputs.

Table 6-1: The Effectiveness Evaluation Results.

Query Number (i)	Number of Expected Rows (B_i)	Number of Observed Rows	Number of Rows Different From Expected (A_i)	Difference Ratio (X_i)
1	37	37	3	0.081
2	4843	4844	2	0.000
3	2730	2730	0	0.000
4	14	14	0	0.000
5	392	397	8	0.020
6	392	397	8	0.020
7	41580	41615	69	0.001
8	24222	24222	2	0.000
9	3118	3118	0	0.000

6.3.2 Efficiency Results

The efficiency evaluation results for the ICTS DW structure are presented in terms of the observed response times of the ETL process for the daily population of the ICTS DW (Section 6.3.2.1). Additionally the comparison of the response times of the derived DW structure and the typical ICTS process of addressing the specified ad hoc queries respectively are presented (Section 6.3.2.2).

6.3.2.1 ICTS ETL Efficiency

The response times of the execution of the ICTS DW structure's ETL processes for the daily population of the DW structure's DW (Equation 6.3) for 14 subsequent days were captured and analysed (Table 6-2). In addition to the response times, the total amount of data that was transferred during the execution of the ETL processes for each subsequent day was captured.

The mean response time for the daily population of the DW structure's DW using the DW structure's ETL processes was calculated to be 00:02:05.770 and is within the 8 hour threshold (Section 6.2.2) during which the ETL processes are required to be executed. A linear positive relationship between daily ETL response times and the amount of data transferred was observed (Figure 6-3).

Table 6-2: DW Structure's ETL processing response times and amount of data transferred for 14 subsequent days.

Day (i)	Duration (W_i) (hh:mm:ss.000)	Data Transferred (KB)
1	00:01:53.531	1544
2	00:01:53.546	1512
3	00:01:53.984	1976
4	00:02:29.156	13600
5	00:02:13.047	11768
6	00:01:54.594	3784
7	00:01:59.219	3656
8	00:02:15.343	13976
9	00:02:22.437	13520
10	00:02:19.625	13272
11	00:02:07.328	12728
12	00:02:09.297	10424
13	00:01:53.813	3384
14	00:01:55.859	3432

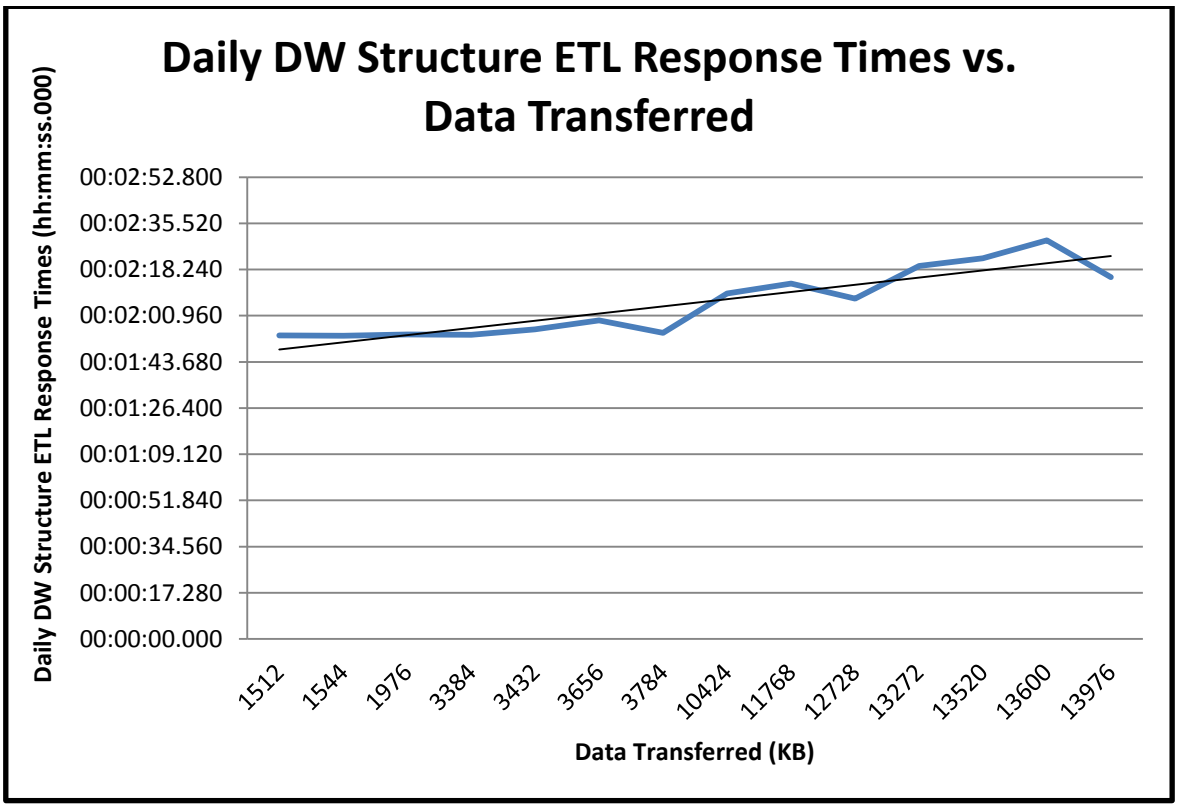


Figure 6-3: Relationship between daily

6.3.2.2 DW Structure Efficiency Compared to Typical ICTS Process Efficiency

The cumulative response times of the execution of the specified ad hoc queries using the typical ICTS process (Figure 6-2) and ICTS DW structure (Figure 6-1) respectively were captured and tabulated (Table 6-3). The response times are presented in terms of hours, minutes, seconds and milliseconds (hh:mm:ss.000).

Table 6-3: Cumulative ad hoc query execution times for the derived DW structure and the typical ICTS process.

Number of Queries (z)	Cumulative ICTS Response Time (R_z)	DW ETL Response Time (T)	Cumulative DW Structure Ad Hoc Query Response Time ($S_z - T$)	Cumulative DW Structure Response Time (S_z)	Difference ($\lvert R_z - S_z \rvert$)
1	00:00:08.567	00:14:03.875	00:00:00.129	00:14:04.004	00:13:55.437
2	00:21:36.250	00:14:03.875	00:00:02.805	00:14:06.680	00:07:29.570
3	00:28:32.772	00:14:03.875	00:00:02.997	00:14:06.872	00:14:25.900
4	00:29:29.023	00:14:03.875	00:00:03.485	00:14:07.360	00:15:21.663
5	00:30:21.486	00:14:03.875	00:00:04.098	00:14:07.973	00:16:13.513
6	00:31:12.100	00:14:03.875	00:00:04.415	00:14:08.290	00:17:03.810
7	01:05:32.652	00:14:03.875	00:00:38.891	00:14:42.766	00:50:49.886
8	01:13:51.336	00:14:03.875	00:00:41.337	00:14:45.212	00:59:06.124
9	01:14:54.417	00:14:03.875	00:00:41.634	00:14:45.509	01:00:08.908

It can be observed from Table 6-3 that as the number of queries (z) which are executed using the typical ICTS process (Figure 6-2) increases, the total response time of the ICTS process (Equation 6.6) increases to the point that it surpasses the cumulative response time of the same number of ad hoc queries using the derived DW structure (Equation 6.5). The main reason for this is due to the need for ICTS to extract data from the data sources each time an ad hoc query is executed (Figure 6-2). The differences between the cumulative response times ($|R_z - S_z|$) of processing ad hoc queries using the typical ICTS process compared to the DW structure indicates that the typical ICTS process of addressing ad hoc queries are more efficient if a small number of queries need to be addressed. The DW structure however proves to be more efficient than the typical ICTS process if a large number of ad hoc queries are required to be addressed.

6.4 Conclusion

The evaluation outlined and discussed in this chapter showed that the ICTS DW structure that was derived through the application of the proposed DW structure design methodology (Section 2.4) is effective in its task of providing support to ICTS for the analysis of ORU data. The effectiveness of the ICTS DW structure was evaluated through obtaining a measure of the DW's accuracy by comparing the observed ad hoc query output rows to the expected ad hoc query output rows (Section 6.2.1). It was observed that 5 ad hoc queries, which were executed using the DW structure, approximated a difference ratio of 0 (Section 6.3.1). Additionally a total difference ratio of 0.001 was observed. Based on these observations it was concluded that the implemented physical design of the ICTS DW structure (Chapter 5) is effective in providing an accurate representation of source data in the DW structure's DW. In addition it can be concluded that the application of the proposed DW structure design methodology on the ICTS case study (Chapters 3, 4 and 5) resulted in the design of an effective DW structure which can support the analysis of ORU data.

The efficiency of the ICTS DW structure was evaluated with the objectives of determining if the DW structure's ETL processes support the efficient population of the DW structure's DW with data generated on a daily basis and if the ICTS DW structure is more efficient than the typical ICTS process used to obtain results for ad hoc queries (Section 6.2.2). The daily population of the ICTS DW structure's DW with daily generated ORU data, using the DW structure's ETL processes (Section 6.2.2), was observed to have a mean response time of 00:02:05.770 (Section 6.3.2.1) that falls within the 8 hour period in the DW structure's ETL processes are needed to be executed on a daily basis. Additionally it was observed that a positive linear relationship exists between the ETL response time and the amount of data that is propagated from the DW structure's sources to its DW (Figure 6-3).

The ICTS DW structure was found to be more efficient than the typical ICTS process (Figure 6-2) used to obtain answers for ad hoc questions only if multiple ad hoc queries are required to be executed (Section 6.3.2.2). Based on this observation it is concluded that the ICTS DW structure is more efficient than the typical ICTS process that is currently being used, since ICTS requires that a large number of ad hoc queries are executed as fast as possible to support the efficient analysis of ORU data.

Chapter 7. Conclusion and Recommendations

7.1 Introduction

The main objective of this research was to propose a methodology that supports the design of a DW structure to support the effective and efficient analysis of online resource usage (ORU) data (Section 1.5). A DW structure consists of a subset of the components of a DW architecture (Figure 1-1), namely, data sources, ETL processes, DW and a metadata repository (Section 1.1). Existing DW structure design methodologies were investigated (Section 2.3) and limitations were identified (Section 2.3.4.5). Based on the limitations identified with DW requirements analysis methodologies and the lack of these methodologies in providing methodological support for the physical design and implementation of a DW structure (Section 1.1), the adapted triple-driven DW structure design methodology (ATDM) was proposed (Section 2.4).

The proposed ATDM was applied to the information and communication technology services (ICTS) department of the Nelson Mandela Metropolitan University (NMMU) in order to evaluate the methodology as a proof of concept (Chapters 3, 4 & 5). The application of the requirements analysis phase of the ATDM consisted of the elicitation of information requirements (Chapter 3). The main deliverable that was obtained from the application of the requirements analysis phase was the subject oriented enterprise subject schema (Figure 3-5) that represents a logical design of the DW structure's DW. Additionally, the application of the requirements analysis phase's supply-driven phase (Section 3.4) resulted in the logical mapping of source systems to the logical design of the DW. Each of the three phases in the requirements analysis phase of the ATDM (Figure 2-5) resulted in the generation and documentation of important sets of semantic metadata which were used to aid the design of the ICTS DW structure.

The application of the physical design phase (Chapter 4) commenced with the selection of an appropriate DW approach (Section 4.4.1) which was based on the investigation of DW approaches and associated selection criteria (Section 4.2). The physical design of the DW (Section 4.4.2) was derived based on the logical design (Figure 3-5) and the selected DW approach (Figure 4-10).

Based on the physical design of the DW and the logical mappings between the DW structure's sources and logical DW design, the physical design of the ETL processes was derived (Section 4.4.3). The DW structure's ETL processes were represented using an ETL modelling technique that was proposed based on the shortcomings which were identified with existing ETL modelling techniques (Section 4.3). The application of the physical design phase of the ATDM (Chapter 5) resulted in the generation and documentation of technical metadata (Section 1.1).

The physical design of the ICTS DW structure was implemented using the MSS environment (Section 5.2) as part of the implementation phase of the ATDM. The implementation of the physical design of the ICTS DW structure (Chapter 5) resulted in the generation of technical metadata (Section 1.1) on the Microsoft SQL Server (MSS) environment (Section 5.2) that describes the implementation of the DW structure at an implementation level. The derived DW structure was evaluated for effectiveness and efficiency in order to determine whether the proposed DW structure design methodology supports the design of a DW structure which in turn supports the effective and efficient analysis of ORU data. This chapter concludes the research by summarising the findings of the research and outlining the contributions of the work. Limitations of the research and problems encountered are described. The chapter closes by discussing recommendations and suggestions for future work.

7.2 Achievements of Research Objectives

This section revisits the research objectives defined for this research (Section 1.5) and discusses the achievements of the research in terms of these objectives (Sections 7.2.1 to 7.2.6). The thesis statement for this research was defined (Section 1.4) as follows:

> A DW structure design methodology can be proposed for the efficient and effective analysis of ORU data.

The main research objective for this research was to propose and evaluate a DW structure design methodology that supports the efficient and effective analysis of ORU data. The secondary objectives of this research used to address the main research objective were as follows:

1) To investigate DW structure design methodologies (Section 7.2.1)
2) To propose a DW structure design methodology for ORU data (Section 7.2.2)
3) To determine information requirements and data characteristics of the NMMU ICTS using the proposed DW structure design methodology (Section 7.2.3)
4) To design a DW structure for the NMMU ICTS using the proposed methodology (Section 7.2.4)
5) To implement the derived DW structure design a for the NMMU ICTS using the MSS integrated tool (Section 7.2.5)
6) To evaluate the derived DW structure implementation for efficiency and effectiveness (Section 7.2.6)

The extent to which these research objectives were met is discussed in the subsequent sections. This research has shown that the proposed ATDM can be successfully applied to derive a design for a DW structure that supports the efficient and effective analysis of ORU data.

7.2.1 Investigation of DW Structure Design Methodologies

Chapter 2 addressed the first research question of the research (Section 1.6). DW structure design methodologies were investigated and limitations were identified. Existing DW structure design methodologies (Section 2.3) only support the derivation for the logical design of a DW structure and do not provide methodological support for the physical design and implementation of a DW structure (Section 1.1). Expert DW requirements were identified (Section 2.3.4.5) which provide methodological support for the requirements analysis phase of DW structure design. The DW structure design methodologies were compared to the expert DW requirements and several limitations were identified (Section 2.3.4.5).

7.2.2 The ATDM DW Structure Design Methodology

Based on the limitations identified in existing DW design methodologies and the inability of these methodologies to provide methodological support for the physical design and implementation of a DW structure, the ATDM was proposed (Figure 7-1). The ATDM DW structure design methodology was proposed to address the second research question (Section 1.6) of this research.

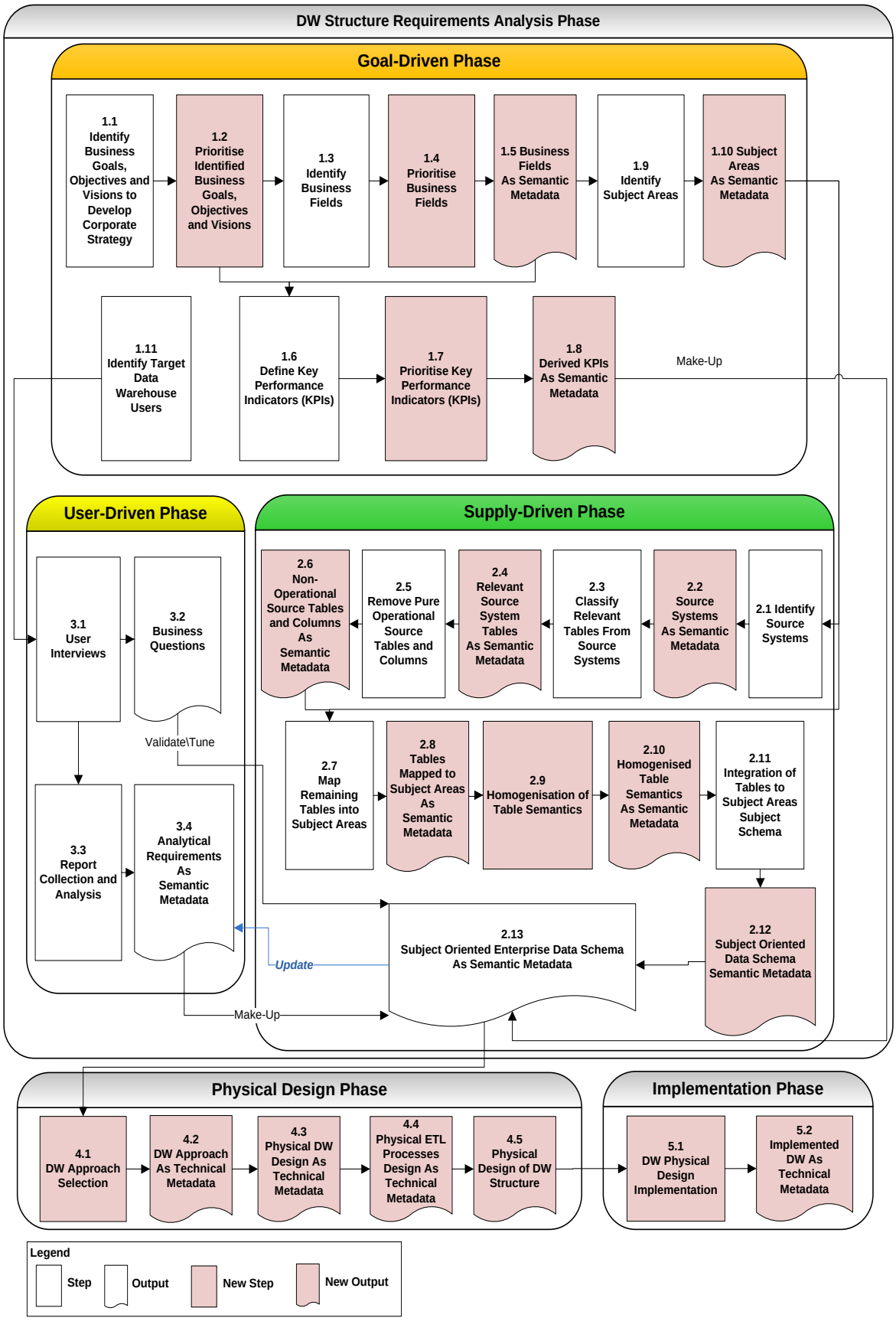


Figure 7-1: The Adapted Triple-Driven Data Warehouse Structure Design Methodology (ATDM).

7.2.3 The ICTS Information Requirements and Data Characteristics

The third research question for this research (Section 1.6) was addressed in Chapter 3 in which the extraction and analysis of the information requirements of the case study was described through the application of the ATDM's requirements analysis phase. The analysis of the extracted information requirements resulted in the derivation of a logical design of the ICTS DW structure that is represented by semantic metadata that was derived and documented. Based on the application of the requirements analysis phase, a limitation of the ATDM was identified that requires an iterative approach for updating analytical requirements based on the derived subject oriented enterprise data schema (Section 3.4.7). This limitation was addressed through the addition of an iteration (Figure 7.1), highlighted in blue, which allows analytical requirements to be updated once the subject oriented enterprise data schema is derived.

7.2.4 ICTS DW Structure Design

Chapter 4 addressed the fourth research question of this research (Section 1.6) in which the logical design of the DW structure were used to derive the physical design of the DW structure. The ATDM included a step in which a DW approach (Section 4.2) was selected (Section 4.4.1) that was also used to aid the physical design of the DW structure. ETL modelling techniques were investigated (Section 4.3) and based on the limitations of existing techniques a suitable ETL modelling technique was proposed which was used to model the physical design of the ICTS DW ETL processes (Section 4.4.3). The application of the physical design phase (Section 4.4) of the ATDM resulted in the derivation and documentation of technical metadata (Section 1.1).

7.2.5 ICTS DW Structure Implementation

The fifth research question of this research (Section 1.6) was addressed in Chapter 5 in which the physical design of the ICTS DW structure was successfully implemented using the MSS environment. The implementation of the physical DW structure design is represented in the MSS environment by a set of technical metadata at an implementation level that describes the DW structure's data sources, DW and ETL processes respectively.

7.2.6 ICTS DW Structure Effectiveness and Efficiency Evaluation

The last research question (Section 1.6) is addressed in Chapter 6 in which the DW structure that is derived from the application of the ATDM is evaluated for effectiveness and efficiency. The effectiveness evaluation (Section 6.2.1) involved testing the accuracy of the derived DW structure in providing results for ad hoc queries. The accuracy of the DW structure was calculated by comparing the observed DW structure ad hoc query results with the expected results for each ad hoc query. The results for the effectiveness evaluation showed (Section 6.3.1) that the derived DW structure is effective in supporting analysis since five ad hoc queries resulted in a difference ratio of zero and a total difference ratio that was close to zero, 0.001, was obtained (Section 6.2.1).

The evaluation for DW structure efficiency consisted of two experiments in which the response times of different tasks were captured and analysed (Section 6.2.2). In the first efficiency experiment the efficiency of the derived ICTS DW ETL processes for extracting daily generated source data was evaluated for a set of fourteen subsequent days. The ETL processes' execution times and the amount of data that was transmitted from the sources to the DW structure's DW for each day were captured and it was found that the ETL processes' response times were all below the specified 8 hour period (Section 6.3.2.1). A positive linear relationship was also observed between the DW structure's response times and the amount of data transferred.

The efficiency of the DW structure in providing ad hoc query results was compared to the typical ICTS process for obtaining results for ad hoc queries (Section 6.3.2.2). It was found that the derived DW structure is more efficient in processing ad hoc queries if several ad hoc queries are required to be processed. The reason for this is attributed to the need of the typical ICTS process to repeat the task of extracting data from the respective data sources, loading it to a temporary data repository and extracting data from the temporary list for each ad hoc query that is executed compared to the derived DW structure in which data is extracted only once using the ETL processes.

7.3 Summary of Contributions

The contribution of the research is discussed in terms of the theoretical and practical contributions. The theoretical contributions (Section 7.3.1) of this research relate to the methodological support for DW structure design. The practical contributions (Section 7.3.2) relate to the application of a DW structure design methodology to the NMMU ICTS case study which describes the logical design, physical design and implementation of a DW structure.

7.3.1 Theoretical Contribution

This research presents a comparison of DW structure design methodologies and their limitations in providing methodological support for DW structure design (Chapter 2). A DW structure design methodology, ATDM, was proposed (Section 7.2.2) which address the limitations identified with DW structure requirements methodologies for ORU data. The proposed DW structure design methodology can be used by DW designers to provide methodological support for the design and implementation of a DW structure's DW and ETL processes based on information requirements and the characteristics of data in data sources. Additionally the methodology supports the generation and documentation of semantic and technical metadata that describes a DW structure at a logical, physical and implementation level.

7.3.2 Practical Contribution

This research presented the application of the proposed DW structure design methodology to the NMMU ICTS case study that can be used by DW designers as an illustration of the outcome for the different steps in the respective phases of the methodology. The effectiveness and efficiency experimental design and associated results obtained from evaluating the derived DW structure provide a practical contribution. Researchers can use the experimental design for the evaluation of a DW structure's effectiveness and efficiency, and the results observed from evaluating the ICTS DW structure for efficiency and effectiveness for comparison purposes.

7.4 Limitation and Problems Encountered

The DW structure that was derived from the application of the proposed DW structure design methodology to the NMMU ICTS case study was implemented on a desktop computer, although it would be preferable to be implemented on a dedicated server for the DW. A dedicated server would typically consist of more processing-, storage- and memory resources which could have an effect on the efficiency of the DW structure for the execution of ad hoc queries. Based on this limitation a set of data generated over a period of fourteen days was used for evaluating the derived DW in order to ensure that the evaluation results are not affected based on the limitation of resources that was experienced.

Additionally, only nine ad hoc queries were used for the effectiveness and efficiency evaluation of the derived ICTS DW structure. A larger sample size of ad hoc queries would have been preferred to be used for the evaluation of the derived DW structure to obtain results which are more representative of the ICTS case study. The environment in which the derived DW structure was evaluated also presented a limitation, since it would have been preferred to evaluate the DW structure in a productive environment. The use of a single case study for this research also presented a limitation since the proposed ATDM cannot be generalised.

7.5 Future Research

Several opportunities for future research are possible from this research. The DW structure that is derived from the application of the methodology can be implemented on a dedicated server that represents the typical environment in which a DW structure is used. This will enable the evaluation of the derived DW structure in processing data of various sizes and the comparison of the results obtained. In addition the applicability of the methodology could also be evaluated with the use of different integrated tools and other case studies to determine the depth of generalisation. The proposed DW structure design methodology can be adapted in future research to include the methodological support for the design of OLAP technology which include multi-dimensional cubes, reporting, data mining and advanced analysis of source data.

References

Abbott, D., 2010. Sociology - Sociology Revision - Methodology, Positivism and Interpretivism. Available at: http://tutor2u.net/blog/index.php/sociology/comments/sociology-revision-methodology-positivism-and-interpretivism/ [Accessed October 17, 2011].

Albrecht, A. and Naumann, F., 2008. Managing ETL Processes. In *Proceedings of the VLDB International Workshop on New Trends in Information Integration (NTII)*. Auckland, pp. 12-15.

Andres, H., 2002. A Contingency Approach to Software Project Coordination. *Journal of Management Information*, 18(3), pp.41-70.

Ariyachandra, T. and Watson, H., 2005. *Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures*. University of Georgia.

Ariyachandra, T. and Watson, H., 2010. Key Organizational Actors in Data Warehouse Architecture Selection. *Decision Support Systems*, 49(2), pp.200-212.

Ariyachandra, T. and Watson, H., 2006. Which Data Warehouse Architecture Is Most Successful? *Business Intelligence Journal*, 11(1), pp.4-6.

Bontempo, C. and Zagelow, G., 1998. The IBM Data Warehouse Architecture. *Communications of the ACM*, 41(9), pp.38-48.

Bradshaw-Camball, P. and Murray, V., 1991. Illusions and other games: A Trifold View of Organisational Politics. *Organisational Science*, 2(4), pp.379-398.

Breslin, M., 2004. Data Warehousing Battle of the Giants. *Business Intelligence Journal*, 9(4).

Bruckner, R., List, B. and Schiefer, J., 2001. Developing Requirements for Data Warehouse Systems with Use Cases. In *Proceedings of 7th Americas Conference on Information Systems*. Boston: Citeseer, pp. 329–335.

Chaudhuri, S. and Dayal, U., 1997. An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26(1), pp.65-74.

Devlin, B., 1997. *Data warehouse: from Architecture to Implementation*, Reading, Massachussets: Addison-Wesley.

Fichman, R., 1997. The Assimilation of Software Process Innovations: An Arganizational Learning Perspective. *Management Science*, 43(10), pp.1345-1363.

Gardner, S., 1998. Buidling the Data Warehouse. *Communications of the ACM*, 41(9), pp.52-60.

Giorgini, P., Rizzi, S. and Garzetti, M., 2008. GRAnD: A Goal-Oriented Approach to Requirement Analysis in Data Warehouses. *Decision Support Systems*, 45(1), pp.4-21.

Giorgini, P., Rizzi, S. and Garzetti, M., 2005. Goal-Oriented Requirement Analysis for Data Warehouse Design. In *Proceedings of the 8th ACM international workshop on Data warehousing and OLAP - DOLAP*. New York, New York, USA: ACM Press, pp. 47-56.

Golfarelli, M., 2010. From User Requirements to Conceptual Design in Data Warehouse Design–a Survey. *Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction, IGI Global, Hershey*, pp.1–16.

Griffin, J., 2001. Information Strategy: Data Mart vs. Data Warehouse. *DM Review*, 12(3).

Guo, Y., Tang, S., Yunhai, T. and Dongqing, Y., 2006. Triple-Driven Data Modeling Methodology in Data Warehousing : A Case Study. In *Proceedings of the 9th ACM international workshop on Data warehousing and OLAP*. Virginia, USA, pp. 59-66.

Hackney, D., 2000. Architecture Anarchy and How to Survive it: God Save The Queen. *Enterprise Systems Journal*, 15(4), pp.24–31.

Hackney, D., 1998. Architecture of Architectures: Warehouse Delivery. *DM Review*, 12(4).

Han, J. and Kamber, M., 2006. *Data Mining: Concepts and Techniques* 2nd ed., San Francisco: Elsevier.

ISO, 2003. ISO/IEC TR 9126-2: Software Engineering - Product Quality - Part 2: External Metrics. Available at: www.iso.org/iso/catalogue_detail.htm?csnumber=22750 [Accessed September 12, 2012].

Imhoff, C., Galemmo, N. and Geiger, J.G., 2008. *Mastering data warehouse design*, Wiley-India.

Inmon, W., 2002. *Building the data warehouse*, New York: J. Wiley.

Jasperson, J., Carte, T.A., Saunders, C.S., Butler, B.S., Croes, H.J.P. and Zheng, W., 2002. Review: Power and Information Technology Research: A Metatriangulation Review. *MIS quarterly*, 26(4), pp.397–459.

Jindal, R., 2004. Federated Data Warehouse Architecture. *Wipro Technologies white paper*. Available at: http://www.bryongaskin.net/education/MBA TRACK/CURRENT/ISOM619/ASSIGNMENTS/Week3/FederatedDataWarehouseArchitecture_WP__5.pdf [Accessed April 23, 2012].

Jukic, N., 2006. Modeling Strategies and Alternatives for Data Warehousing Projects. *Communications of the ACM*, 49(4), pp.83–88.

Kelly, S. and Hadden, E., 1997. Data Marts: The Latest Silver Bullet. *DM REVIEW*, 7(2), pp.16–17.

Kimball, R., Ross, M. and Merz, R., 2002. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling*, New York: Wiley.

Mahadevan, 2011. SSIS Control Flow Items and Uses. Available at: http://mahadevanrv.blogspot.com/2011/07/ssis-control-flow-items-and-uses.html [Accessed October 31, 2012].

Mazon, J.-N. and Trujillo, J., 2009. A Hybrid Model Driven Development Framework for the Multidimensional Modeling of Data Warehouses! *ACM SIGMOD Record*, 38(2), pp.12-17.

Microsoft, 2012a. Control Flow. Available at: http://msdn.microsoft.com/en-us/library/ms137681.aspx [Accessed October 29, 2012].

Microsoft, 2012b. Precendence Constraints. Available at: http://msdn.microsoft.com/en-us/library/ms141261.aspx [Accessed October 31, 2012].

Microsoft, 2012c. SQL Server Overview. Available at: http://msdn.microsoft.com/en-us/library/ms166352(v=sql.90).aspx [Accessed October 30, 2012].

Microsoft, 2012d. Tutorial: SQL Server Management Studio. Available at: http://msdn.microsoft.com/en-us/library/bb934498(v=sql.105).aspx [Accessed October 31, 2012].

Miller, D. and Friesen, P.H., 1982. Innovation in Conservative and Entrepreneurial Firms: Two Models of Strategic Momentum. *Strategic management journal*, 3(1), pp.1–25.

Moss, L.T. and Atre, S., 2003. *Business intelligence roadmap: the complete project lifecycle for decision-support applications*, Boston: Addison-Wesley Professional.

Munoz, L., Mazon, J.-N., Pardillo, J. and Trujillo, J., 2008. Modelling ETL Processes of Data Warehouses with UML Activity Diagrams. *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*, pp.44-53.

Nemati, H.R., Steiger, D.M., Iyer, L.S. and Herschel, R.T., 2002. Knowledge Warehouse: An Architectural Integration of Knowledge Management, Decision Support, Artificial Intelligence and Data Warehousing. *Decision Support Systems*, 33(2), pp.143–161.

Opensource-Blogs, 2011. Control and Data Flow in SSIS. Available at: www.opensource-blogs.com/2011/10/313/ [Accessed October 31, 2012].

Rob, P., Coronel, C. and Crockett, K., 2008. *Database Systems: Design, Implementation & Management*, Cengage Learning EMEA.

Sabherwal, R. and Becerra-Fernandez, I., 2009. *Business Intelligence Practices, Technologies and Management* B. L. Golub, ed., Hoboken: John Wiley & Sons.

Satzinger, J., Jackson, R. and Burd, S., 2005. *Object-Oriented Analysis & Design with the Unified Process*, Thomson.

Saunders, M., Lewis, P. and Thornhill, A., 2009. *Research Methods for Business Students* 5th ed., Harlow: Pearson Education.

Sen, A. and Sinha, A.P., 2005. A Comparison of Data Warehousing Methodologies. *Communications of the ACM*, 48(3), pp.79-84.

Simitsis, A. and Theodoratos, D., 2009. Data Warehouse Back-End Tools. *Encyclopedia of Data Warehousing and Mining*, pp.572–579.

Simitsis, A. and Vassiliadis, P., 2009. Benchmarking ETL workflows. *Performance Evaluation and Benchmarking*, pp.199-220.

Song, I., Medsker, C., Rowen, W. and Ewen, E., 2001. An Analysis of Many-to-Many Relationships Between Fact and Dimension Tables in Dimensional Modeling. In *Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW)*. Interlaken.

Stöhr, T., Müller, R. and Rahm, E., 1999. An integrative and uniform model for metadata management in data warehousing environments. In *Proceedings of the International Workshop on Design and Management of Data Warehouses, Heidelberg, Germany*. Citeseer, pp. 1-16.

Swanson, E., 1994. Information Systems Innovation Among Organizations. *Management Science*, 40(9), pp.1069–1092.

Trujillo, J. and Luj, S., 2003. A UML Based Approach for Modeling ETL Processes in Data Warehouses. *Conceptual Modeling - ER 2003*, pp.307-320.

Vail III, E.F., 2002. Causal Architecture: Bringing the Zachman Framework to Life. *Information systems management*, 19(3), pp.8–19.

Vassiliadis, P. and Simitsis, A., 2002. Conceptual Modeling for ETL Processes. In *Proceedings of the ACM 5th International Workshop on Data Warehousing and Olap*. McLean, pp. 14-21.

Vassiliadis, P. and Simitsis, A., 2009. Extraction, Transformation, and Loading. Available at: http://www.cs.uoi.gr/~pvassil/downloads/ETL/SHORT_DESCR/08SpringerEncyclopedia_draft.pdf [Accessed April 10, 2012].

Vassiliadis, P., Vagena, Z., Skiadopoulos, S. and Karayannidis, N., 2001. Arktos : Towards the Modeling , Design , Control and Execution of ETL Processes. *Information Systems*, 26(8), pp.537-561.

Vetterli, T., Vaduva, A. and Staudt, M., 2000. Metadata standards for data warehousing: open information model vs. common warehouse metadata. *ACM Sigmod Record*, 29(3), pp.68–75.

Watson, H., Watson, R.T., Singh, S. and Holmes, D., 1995. Development practices for executive information systems: findings of a field study. *Decision Support Systems*, 14(2), pp.171–184.

Winter, R. and Strauch, B., 2003. A Method for Demand-Driven Requirements Analysis in Data Warehousing. In *Proceedings of Hawaii International Conference on System Science*. Big Island, pp. 1359-1365.

Winter, R. and Strauch, B., 2004. Information Requirements Engineering for Data Warehouse Systems. In *Proceedings of the 2004 ACM symposium on Applied computing*. Nicosia: ACM, pp. 1359–1365.

Appendix A: ICTS Source Data Tables

Pcounter

Column Name	Data Type	Allow Nulls
PRINTUSERDOMAIN	nvarchar(50)	☑
PRINTDOMAIN	nvarchar(50)	☑
PRINTUSERNAME	nvarchar(50)	☑
USERGROUP	nvarchar(50)	☑
PRINTERNAME	nvarchar(50)	☑
DOCUMENTNAME	nvarchar(50)	☑
PRINTTIME	nvarchar(50)	☑
PRINTDATE	nvarchar(50)	☑
CLIENTCODE	nvarchar(50)	☑
SUBCODE	nvarchar(50)	☑
PAPERSIZE	nvarchar(50)	☑
DUPLEX	nvarchar(50)	☑
COLOR	nvarchar(50)	☑
COPIES	bigint	☑
BWPAGES	bigint	☑
COLORPAGES	bigint	☑
TOTALPAGES	bigint	☑
COST	decimal(18, 2)	☑
BALANCE	decimal(18, 2)	☑
FQBALANCE	nvarchar(50)	☑
JOBTYPE	bigint	☑
JOBTYPEDESC	nvarchar(50)	☑
IsRefunded	nvarchar(50)	☑
NumberOfPages	bigint	☑
DateTime	datetime	☑
		☐

The Pcounter table.

Process

Column Name	Data Type	Allow Nulls
PROCESSID	varchar(100)	☐
PROCESSINTERNALN...	varchar(300)	☑
PROCESSNAME	varchar(300)	☑
PROCESSTITLE	varchar(300)	☑
PROCESSPATH	varchar(300)	☑
FILESIZE	varchar(50)	☑
OPENCLOSE	varchar(10)	☑
LOGONID	varchar(100)	☑
AnyAction	varchar(50)	☑
StartDate	datetime	☑
StopDate	datetime	☑
		☐

The Process table.

PC

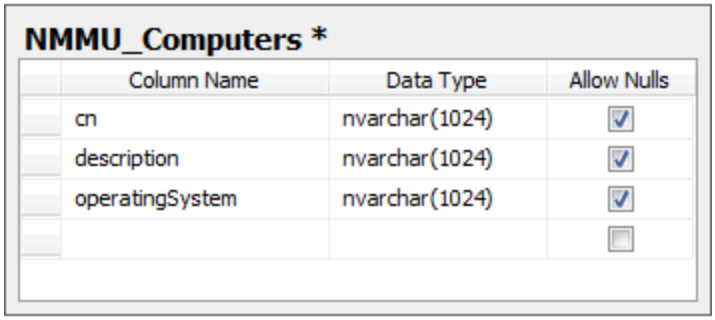
Column Name	Data Type	Allow Nulls
PCNAME	varchar(100)	☐
LoggedOnUser	varchar(200)	☑
FMSVersion	varchar(10)	☑
PCStatus	varchar(5)	☑
LastResponseDate	datetime	☑
LastSoftwareScan	datetime	☑
Status	varchar(50)	☑
		☐

The PC table.

Logon

Column Name	Data Type	Allow Nulls
LogonId	varchar(100)	☐
PCNAME	varchar(100)	☑
UserName	varchar(50)	☑
StartDate	datetime2(3)	☑
EndDate	datetime2(3)	☑
Active	varchar(50)	☑
		☐

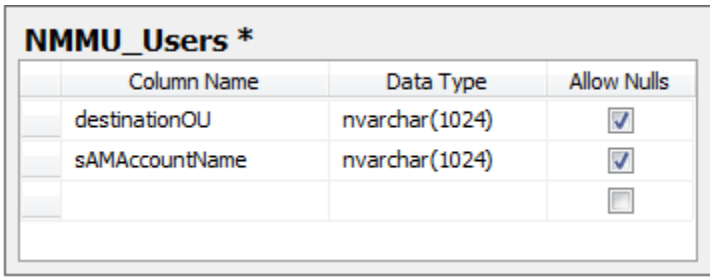
The Logon table.



NMMU_Computers *

Column Name	Data Type	Allow Nulls
cn	nvarchar(1024)	☑
description	nvarchar(1024)	☑
operatingSystem	nvarchar(1024)	☑
		☐

The NMMU_Computers table.



NMMU_Users *

Column Name	Data Type	Allow Nulls
destinationOU	nvarchar(1024)	☑
sAMAccountName	nvarchar(1024)	☑
		☐

The NMMU_Users table

Appendix B: Homogenisation of Table Semantics

	Current Semantics	Homogenised Semantics
Table Name	Logon	Logon
Column Name	LogonId	WorkstationName
	UserName	UserName
	StartDate	StartDate

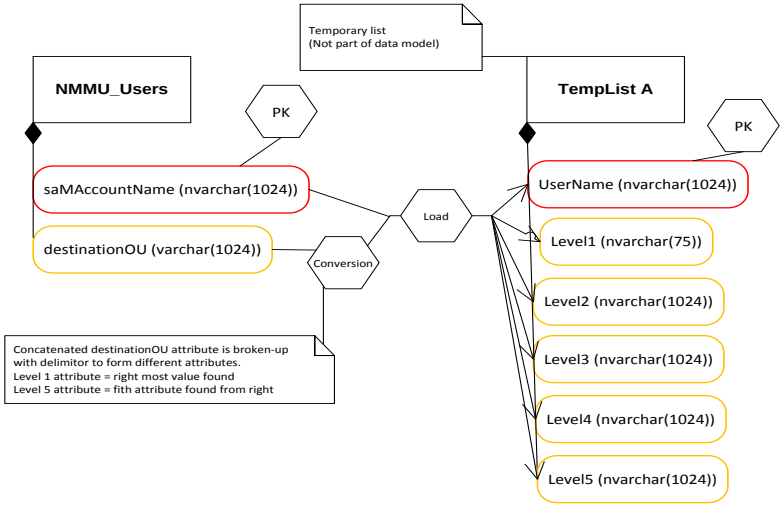
	Current Semantics	Homogenised Semantics
Table Name	Process	Process
Column Name	PROCESSID	ProcessId
	PROCESSTITLE	ProcessTitle
	LOGONID	LogonId
	StartDate	StartDate

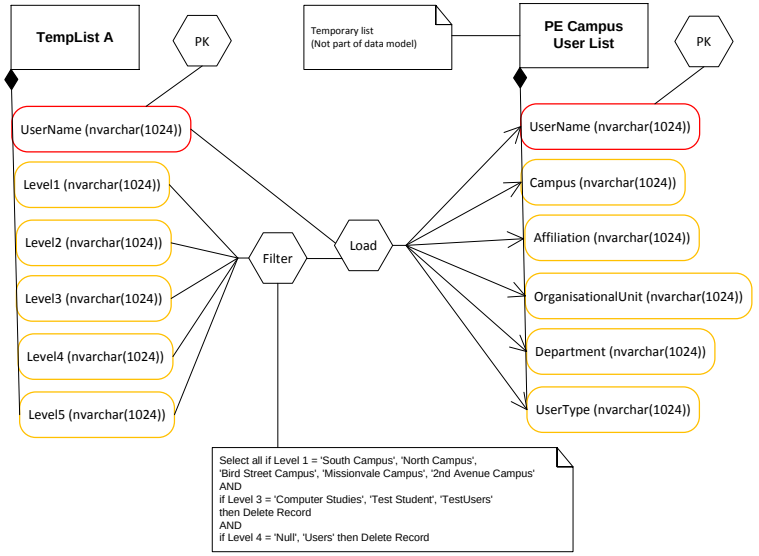
	Current Semantics	Homogenised Semantics
Table Name	NMMU_COMPUTERS	NMMU_Workstations
Column Name	cn	WorkstationName
	destinationOU	destinationOU
	operatingSystem	OperatingSystem

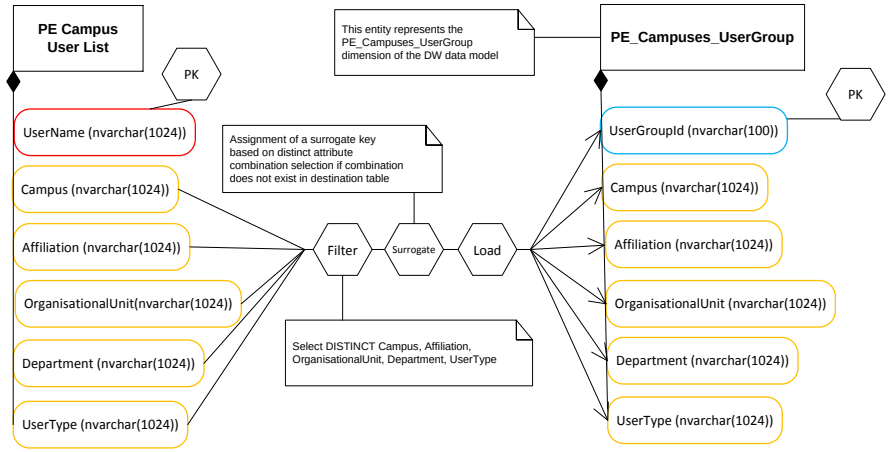
	Current Semantics	Homogenised Semantics
Table Name	Pcounter	
Column Name	PRINTUSERNAME	UserName
	PRINTERNAME	PrinterName
	PAPERSIZE	PaperSize
	DUPLEX	Duplex
	COPIES	NumCopies
	BWPAGES	NumBlackWhitePages
	COLORPAGES	NumColorPages
	COST	TotalPrintCost
	NumberOfPages	TotalNumPages
	DateTime	DateTime

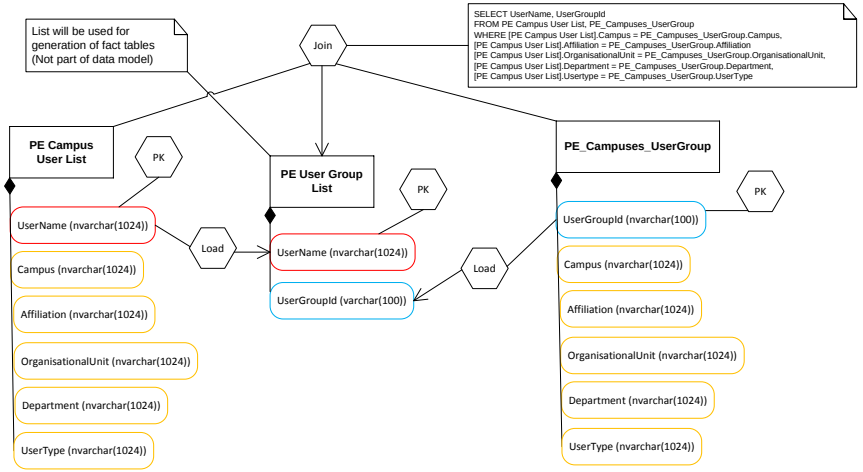
	Current Semantics	Homogenised Semantics
Table Name	NMMU_USERS	NMMU_Users
Column Name	sAMAccountName	UserName
	destinationOU	destinationOU

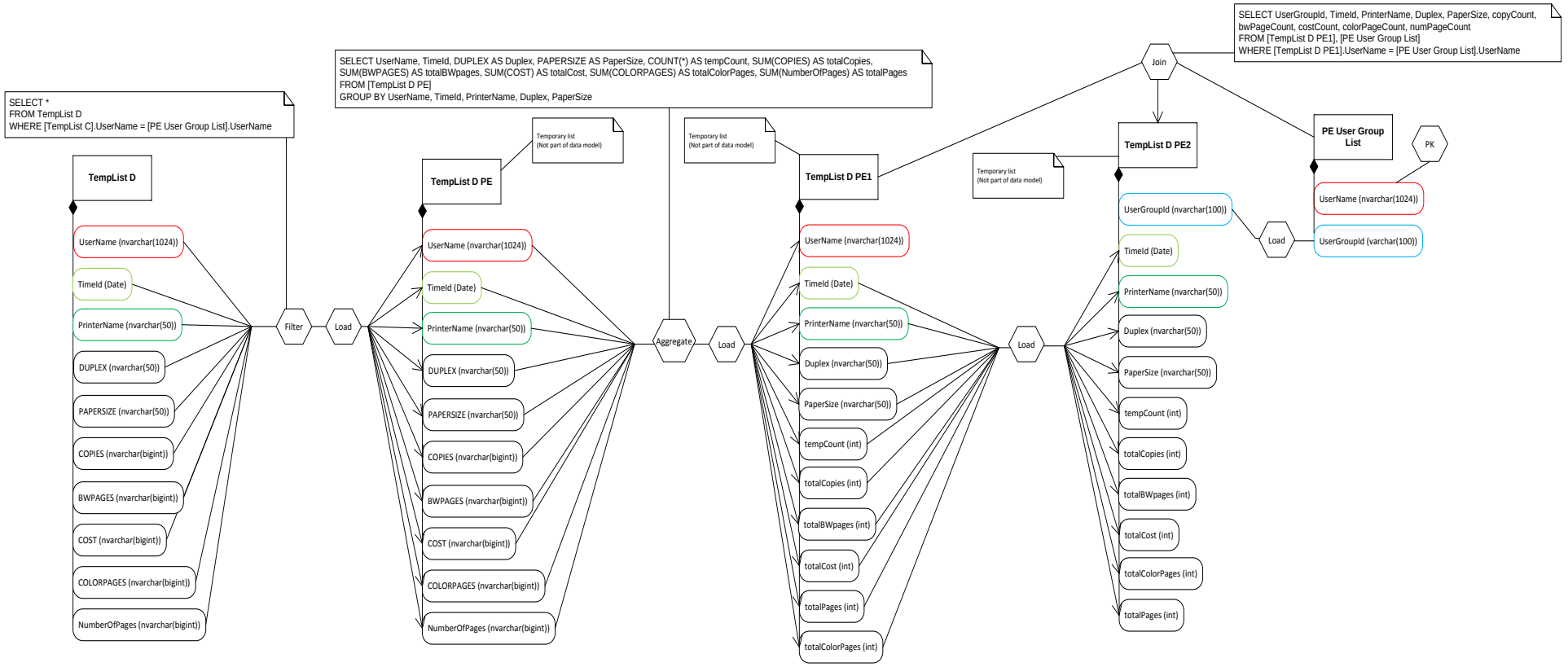
Appendix C: ICTS ETL Processes Physical Design










SELECT *
FROM TempList D
WHERE [TempList C].UserName = [PE User Group List].UserName
SELECT UserName, TimeId, DUPLEX AS Duplex, PAPERSIZE AS PaperSize, COUNT(*) AS tempCount, SUM(COPIES) AS totalCopies, SUM(BWPAGES) AS totalBWpages, SUM(COST) AS totalCost, SUM(COLORPAGES) AS totalColorPages, SUM(NumberOfPages) AS totalPages
FROM [TempList D PE]
GROUP BY UserName, TimeId, PrinterName, Duplex, PaperSize
SELECT UserGroupId, TimeId, PrinterName, Duplex, PaperSize, copyCount, bwPageCount, costCount, colorPageCount, numPageCount
FROM [TempList D PE1], [PE User Group List]
WHERE [TempList D PE1].UserName = [PE User Group List].UserName
TempList D
UserName (nvarchar(1024))
TimeId (Date)
PrinterName (nvarchar(50))
DUPLEX (nvarchar(50))
PAPERSIZE (nvarchar(50))
COPIES (nvarchar(bigint))
BWPAGES (nvarchar(bigint))
COST (nvarchar(bigint))
COLORPAGES (nvarchar(bigint))
NumberOfPages (nvarchar(bigint))
Filter
Load
Temporary list
(Not part of data model)
TempList D PE
UserName (nvarchar(1024))
TimeId (Date)
PrinterName (nvarchar(50))
DUPLEX (nvarchar(50))
PAPERSIZE (nvarchar(50))
COPIES (nvarchar(bigint))
BWPAGES (nvarchar(bigint))
COST (nvarchar(bigint))
COLORPAGES (nvarchar(bigint))
NumberOfPages (nvarchar(bigint))
Aggregate
Load
Temporary list
(Not part of data model)
TempList D PE1
UserName (nvarchar(1024))
TimeId (Date)
PrinterName (nvarchar(50))
Duplex (nvarchar(50))
PaperSize (nvarchar(50))
tempCount (int)
totalCopies (int)
totalBWpages (int)
totalCost (int)
totalPages (int)
totalColorPages (int)
Join
Temporary list
(Not part of data model)
TempList D PE2
UserGroupId (nvarchar(100))
TimeId (Date)
PrinterName (nvarchar(50))
Duplex (nvarchar(50))
PaperSize (nvarchar(50))
tempCount (int)
totalCopies (int)
totalBWpages (int)
totalCost (int)
totalColorPages (int)
totalPages (int)
Load
PE User Group List
PK
UserName (nvarchar(1024))
UserGroupId (varchar(100))