APPLICATION OF LEXICAL ONTOLOGY FOR SEMI-AUTOMATIC OF LOGICAL DATA DESIGN IN DATA WAREHOUSE

MIOR NASIR MIOR NAZRI

UNIVERSITI KEBANGSAAN MALAYSIA

APPLICATION OF LEXICAL ONTOLOGY FOR SEMI-AUTOMATIC OF LOGICAL DATA DESIGN IN DATA WAREHOUSE

MIOR NASIR MIOR NAZRI

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2012

APLIKASI ONTOLOGI LEKSIKAL DALAM REKA BENTUK LOGIKAL SEPARA AUTOMATIK GUDANG DATA

MIOR NASIR MIOR NAZRI

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2012

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

9 July 2012

MIOR NASIR MIOR NAZRI
P35238

ACKNOWLEDGEMENT

First of all, I would like to thank Allah S.W.T the most Merciful, and the most Magnificent for His grace and mercy that allowed me to complete my thesis. Undoubtedly, He increased my knowledge and gave me the strength to persevere.

I would like to express words of gratitude to my supervisor, Prof. Dr. Shahrul Azman Mohd Noah from Universiti Kebangsaan Malaysia (UKM), for his guidance in completing my study and also for his support, encouragement, and understanding throughout my research endeavour. His guidance chartered the direction of my research and writing of this thesis. I would like to express my appreciation to the International Islamic University Malaysia, the Dean of Kulliyyah of Information and Communication Technology (KICT), Prof. Dr. Mohd Adam Suhaimi and the Head of Department of Information Systems, Prof. Dr. Abu Osman Mad Tap for giving me the opportunity and support to continue my post graduate study.

I wish to express my warm and sincere thanks to my wife Dr. Zarinah Hamid for her boundless support and continuous source of inspiration, and my beloved children Nadzirah, Mior Nabil and Nusrah for their patience in dealing with their busy father. I am deeply indebted to my mother Halipah Ashari for raising me with lots of love and my late father Mior Nazri Mior Yusof for constantly reminding me about the importance of education. Lastly I would like to thank my friends at KICT and fellow PhD candidates at UKM for their support and encouragement throughout my study.

ABSTRACT

Designing a data warehouse based on the current operational database is a very complex and time consuming process. The most critical part in designing a data warehouse is to identify the suitable entities as fact tables. Many researches incorporate artificial intelligence algorithm in the form of knowledge-based systems in order to assist in designing process. Most of existing data warehouse design tools employ direct transformation of input into corresponding designs and rely on the users to identify suitable entities to be modelled as fact tables. As a result, the main task in knowledge intensive model design still relies heavily on the user input. Hence, the main objective of this research is to incorporate knowledge-based system for the task of designing multidimensional model for data warehouse. We propose a new method based on supply driven approach using lexical ontology as the knowledge domain. Our method is able to make intelligent decision in designing data warehouse model by extracting valuable information from the knowledge domain. Once fact table is identified, the following step is to generate data warehouse multidimensional model with minimal user intervention. The feasibility of the proposed method is demonstrated by a prototype called *ADW-tool* using WordNet as knowledge domain. The process starts with the conversion of an enterprise logical model into a specification language model as input. The input goes through a set of synthesis and diagnosis rules before it is accepted into the system internal tables. The next stage involves identifying fact table and dimensional table candidates with the help from WordNet database. In the final stage, a logical multidimensional model in the form of Star schema emerged around the selected fact tables. Two sets of tests are performed by using selected business enterprise logical model as input. The two schemas used as input are identical with the input schemas used by two previous researches. Hence the result from this experiment is comparable against the result from those mentioned research defined as the bench mark. The research has demonstrated the viability of exploiting ontology in assisting the process of semi-automated data warehouse design for selecting potential facts and dimensional tables. Ontology is not claimed to be a panacea, however exploiting lightweight ontologies such as WordNet is seen able to suggest the correct entities for potential fact tables which will remain unidentified for novice human designer.

ABSTRAK

Mereka bentuk gudang data dengan berlandaskan kepada pangkalan data yang beroperasi adalah proses yang kompleks dan memakan masa. Perkara yang paling penting semasa mereka bentuk sebuah gudang data adalah untuk mengenalpasti entiti yang paling sesuai digunakan sebagai 'jadual fakta'. Banyak kajian menggabungkan algoritma kecerdasan buatan dalam bentuk sistem berasaskan pengetahuan untuk membantu dalam proses reka bentuk. Sebahagian besar alat reka-bentuk gudang data yang sedia ada yang menggunakan proses transformasi langsung akan bergantung sepenuhnya kepada pengguna untuk mengenalpasti entiti berpadanan yang boleh dimodelkan sebagai 'jadual fakta'. Akibatnya, pengguna masih lagi dibebankan dengan tugasan utama dalam pemodelan yang bersifat intensif pengetahuan ini. Sehubungan dengan itu, tujuan utama penyelidikan adalah untuk mengaplikasikan pendekatan sistem berasaskan-pengetahuan dalam tugas proses mereka bentuk model multidimensi untuk gudang data. Kajian ini mencadangkan sebuah kaedah baru berdasarkan rangka 'supply driven' menggunakan ontologi leksikal sebagai sumber pengetahuan. Kaedah ini mampu untuk membuat keputusan sendiri tanpa campur tangan pengguna dalam merancang model gudang data berdasarkan maklumat penting yang diambil daripada sumber pengetahuan. Setelah 'jadual fakta' dikenal pasti, langkah seterusnya adalah untuk menjana model gudang data multidimensi tanpa banyak interaksi daripada pengguna. Sebuah prototaip yang dinamakan sebagai *ADW-tool* telah dibangunkan berlandaskan kaedah cadangan dengan menggunakan WordNet sebagai domain pengetahuan. Proses bermula dengan penterjemahan model logikal ke bentuk model spesifikasi sebagai input. Kemudian input tersebut akan melalui proses petua sintesis dan petua diagnosis sebelum diterima oleh sistem. Langkah seterusnya melibatkan pencarian calon 'jadual fakta' dan 'jadual dimensi' dengan bantuan WordNet. Pada tahap akhir, reka bentuk model multidimensi logik dikeluarkan berasaskan bentuk skema bintang berpusat sekitar 'jadual fakta' yang dipilih. Dua set ujian dilakukan dengan menggunakan model perniagaan syarikat yang dipilih sebagai input. Kedua-dua skema yang digunakan sebagai input untuk penyelidikan ini adalah sama dengan input yang digunakan oleh dua penyelidikan lain sebelum ini. Hasil daripada penyelidikan ini menunjukkan keputusan yang konsisten dengan hasil dari kajian tersebut yang telah ditakrifkan sebagai tanda aras. Kajian telah membuktikan bahawa ontologi leksikal dapat membantu proses reka bentuk gudang data separa automatik dalam memilih 'jadual fakta' dan 'jadual dimensi'. Ontologi tidak dianggap sebagai satu-satunya penyelesaian, namun dengan memanfaatkan ontologi leksikal seperti WordNet mampu menyarankan entiti yang betul untuk 'jadual fakta' tetapi tidak dapat dikenal pasti oleh penganalisis baru.

CONTENTS

		Page
DECLARATION		iii
ACKNOWLEDGEMENT		iv
ABSTRACT		v
ABSTRAK		vi
CONTENTS		vii
LIST OF FIGURES		xii
LIST OF TABLES		xv
LIST OF ABBREVIATIONS		xvi
CHAPTER I	**INTRODUCTION**	
1.1	Background	1
1.2	Statement of the Problem	3
1.3	Aim and Objectives of Research	4
1.4	Importance of the Research	5
1.5	Scope of the Research	5
1.6	Research Methodology	7
1.7	Organization of Thesis	8
CHAPTER II	**LITERATURE REVIEW**	
2.1	Introduction	11
2.2	Data Warehouse	11
	2.2.1 What is a Data Warehouse?	12
	2.2.2 History of Data Warehouse	13
	2.2.3 Why using a Data Warehouse?	14
	2.2.4 Data Warehouse Architecture	16
	2.2.5 Data Warehouse Model	18
2.3	Ontology	20
	2.3.1 Classification of Ontologies	21
	2.3.2 Type of Ontologies	22
	2.3.3 Application of Ontologies	24

	2.3.4 Lexical Ontology	25
2.4	Summary	27

CHAPTER III DATA WAREHOUSE DESIGN METHODS - REVIEW

3.1	Introduction	29
3.2	Data Warehouse Design Stages	30
	3.2.1 Requirement Analysis Stage	30
	3.2.2 Conceptual Design Stage	33
	3.2.3 Logical Design Stage	34
	3.2.4 ETL Process Design Stage	35
	3.2.5 Physical Design Stage	36
3.3	Multidimensional Model	36
	3.3.1 Star Schema	37
	3.3.2 Dimensional Fact Model (DFM)	38
	3.3.3 Multidimensional Data Model (MD)	39
	3.3.4 Multidimensional Entity Relationship Model (ME/R)	39
	3.3.5 StarER Model	40
	3.3.6 Structured Entity Relationship Model (SERM)	41
	3.3.7 Event-Entity-Relationship (EVER) Model	42
	3.3.8 Multidimensional Normal Form Model (MNF)	43
	3.3.9 UML Multidimensional Model	44
3.4	Multidimensional Modeling Techniques	45
	3.4.1 Guideline Modeling STAR Model – KRT98	45
	3.4.2 Semi-Automatic Modeling Dimensional Fact Model – GR98	46
	3.4.3 Guideline Modeling ME/R Model - SBHD98	48
	3.4.4 Guideline Modeling MD MODEL - CT98	48
	3.4.5 Guideline Modeling StarER Model – TBC99	50
	3.4.6 Guideline Modeling SERM – BvE99	51
	3.4.7 Guideline Modeling MNF Model – HLV00	52
	3.4.8 Guideline Modeling Star Model – MK00	53
	3.4.9 Semi-Automatic Hybrid Modeling Star Graph – BCC01	54
	3.4.10 Semi-Automatic Modeling ME/R Model – PD02	55
	3.4.11 Transformation Modeling Star Model – MR02	56
	3.4.12 Guideline Modeling Generic Model – WS03	57
	3.4.13 Semi-Automatic Modeling From XML Schemas – VBR03	58
	3.4.14 Semi-Automatic Goal-Oriented Modeling – GRG05	59
	3.4.15 Transformation-Oriented Modeling DFM – SN06	59
	3.4.16 Guideline Modeling UML-Based Model – PACW06	60
	3.4.17 Automatic Modeling By Example – RA06	61
	3.4.18 Semi-Automatic Modeling by Reconciling – MTL07	61

	3.4.19 Automatic Modeling SAMSTAR Model – SKD07	62
	3.4.20 Semi-Automatic Modeling from Ontologies – RA07	63
	3.4.21 Modeling Techniques Summary	64
3.5	Selection of Data Warehouse Design Method	69
3.6	Summary	70
CHAPTER IV	**RESEARCH METHOD**	
4.1	Introduction	72
4.2	Literature Review	73
4.3	The Development of Data Warehouse Prototype	73
	4.3.1 Analysis Phase	74
	4.3.2 Design Phase	76
	4.3.3 Implementation Phase	77
	4.3.4 Testing Phase	79
4.4	Evaluation Method	79
4.5	Summary	80
CHAPTER V	**A LEXICAL APPROACH TO AUTOMATE DATA WAREHOUSE DESIGN**	
5.1	Introduction	81
5.2	The Specification Language Model Formulation	82
	5.2.1 Translating Logical Model Into Specification Language Model	82
	5.2.2 Validation Process Of Input Record	85
5.3	Identify Fact And Dimensional Tables	91
	5.3.1 Overview - Identifying Fact Table	92
	5.3.2 Identify Entity Type	95
	5.3.3 Identify Facts	103
	5.3.4 Identify Potential Fact and Dimension Table	110
5.4	Creation Of Multidimensional Model	111
	5.4.1 Create Fact Table	114
	5.4.2 Create Temporal Dimensional Table	115
	5.4.3 Create Dimensional Table	115
	5.4.4 Add Dimensional Hierarchy Table	116
	5.4.5 De-Normalize Hierarchy Table	117
	5.4.6 Repeat the Process For Other Fact Table	118

5.5 Summary 119

CHAPTER VI TESTING AND EVALUATION

6.1 Introduction 122

6.2 Test Result from Selected Enterprise Logical Model 123

6.2.1 Test 1: Input Data from TPC-H Logical Model 123
6.2.2 Test 2: Input Data from Retail Logical Model 129

6.3 Evaluation of result 135

6.3.1 Evaluation of TPC-H result 135
6.3.2 Evaluation of Retail result 140

6.4 Summary 143

CHAPTER VII CONCLUSION AND FUTURE RESEARCH WORK

7.1 Introduction 145

7.2 Research Findings 145

7.3 Research Contributions 147

7.4 Future Research Work 149

7.5 Summary 150

REFERENCES 151

APPENDICES

A Overview of ADW-tool 159

B List of Lexicographer Files and Lexical Number 169

C Input Data - attributes and entity 181

D1 Dream House Logical Schema 187

D2 Specification Language for Dream House Table 188

D3 Detail Result of Dream House Table 190

E1 Specification Language for TPC-H Schema 192

E2 Detail Result of TPC-H Schema 194

E3 Fact Tables for TPC-H Schema 196

F1 Specification Language for Retail Schema 200

F2	Detail Result of Retail Schema	202
F3	Fact Tables for Retail Schema	205
G	Paper Presentation and Publication	208

LIST OF FIGURES

Figure No.		Page
2.1	Data Warehouse System Architecture	17
2.2	Data cube representing vehicle sale	19
2.3	Ontology level	22
2.4	An Ontology Spectrum	22
2.5	Meanings of the word “plant”	27
3.1	Example of Star Schema	37
3.2	Example of Dimensional Fact Model	38
3.3	Example of MD Model	39
3.4	Example of ME/R Schema	40
3.5	Example of StarER Model	41
3.6	Example of SERM	42
3.7	Example of EVER Model	43
3.8	Example of MNF Model	44
3.9	Example of UML Model	45
4.1	SDLC development stages	74
4.2	The three-stage supply-driven modeling approach	76
4.3	The Architecture of ADW-tool	78
4.4	Algorithm to transform logical schema into MD Model	78
5.1	Method Overview	81
5.2	Formulation Specification Language Process Overview	82
5.3	Syntax diagram for Specification Language Model structure	83
5.4	Syntax diagram for attribute name	83
5.5	Dream House Logical schema	84
5.6	Syntax for entity declaration	86
5.7	Syntax for attribute declaration	86
5.8	Valid data type for attribute	87
5.9	Syntax for primary key declaration	87
5.10	Syntax for foreign key declaration	88
5.11	Entities with invalid relationship	88
5.12	Algorithm for reading records	90

5.13	Algorithm to validate input records	91
5.14	Test for potential fact table	94
5.15	Overview process to identify fact table	95
5.16	Result of WordNet Browser for "customer"	97
5.17	Example 1 – "customer" entity	97
5.18	Result of WordNet Browser for word "bid"	98
5.19	Example 2 – "bid" entity	99
5.20	WordNet result for word "agent"	99
5.21	Result of "agent" entity	100
5.22	Algorithm to extract word meaning from WordNet database	104
5.23	Classification of entity LEASE	105
5.24	Algorithm to define and compute point for entity type	106
5.25	Algorithm to compute point for numeric fields	107
5.26	Result of numeric attributes "rent" and "deposit"	108
5.27	Algorithm to compute point for many-to-one relationship	109
5.28	Algorithm to identify candidates for fact table	111
5.29	Multidimensional Model Process Overview	112
5.30	Algorithm to generate multidimensional model	112
5.31	Confirmation Screen for Dream House Schema	113
5.32	Detail information on LEASE entity	114
5.33	F_LEASE Fact Table	115
5.34	Fact Table and Temporal Dimensional table	115
5.35	Fact Table and Dimensional Tables	116
5.36	F_LEASE Fact Schema	117
5.37	De-normalized F_LEASE Fact Schema	118
5.38	De-normalized F_VIEWING Fact Schema	119
5.39	Algorithm to Draw Schema Diagram	120
6.1	TPC-H Logical Model	124
6.2	Detail result for ORDER entity	126
6.3	Confirmation Screen for TPC-H schema	127
6.4	F_ORDER Fact Table	128
6.5	De-normalized F_ORDER Fact Table	129
6.6	RETAIL Logical Model	130

6.7	Detail result for SALE entity	131
6.8	Confirmation Screen for RETAIL schema	133
6.9	F_SALE Fact Table	134
6.10	F_SALE star schema	134
6.11	LINEITEM schema between Test 1 and Benchmark	137
6.12	PARTSUPP schema between Test 1 and Benchmark	138
6.13	TPC-H Constellation schema	140
6.14	SALE schema between Test 2 and Benchmark	142
A.1	ADW001 - ADW-tool main panel screen	160
A.2	CP001 - New project definition screen	160
A.3	Screen to select input file	161
A.4	ADW001 - Input table TPC-H.txt is selected and loaded	162
A.5	VT001 - Details of entities and attributes	163
A.6	VT001 - Detail definition of entity and attributes	163
A.7	DrawERD - TPC-H schema in graphical form	164
A.8	FT001 - List of Fact Table Candidates	165
A.9	FT010 - Detail information on 'ORDER' entity	166
A.10	Confirmation on selected fact tables	166
A.11	ADW001 - List of Fact Tables	167
A.12	DrawERD - List of Fact Tables	168
D.1	Dream House Logical Schema	187
E3.1	F_ORDER Fact Table	196
E3.2	De-normalized F_ORDER Fact Table	196
E3.3	F_LINEITEM Fact Table	197
E3.4	De-normalized F_LINEITEM Fact Table	198
E3.5	F_PARTSUPP Fact Table	199
E3.6	De-normalized F_PARTSUPP Fact Table	199
F3.1	F_SALE Fact Table	205
F3.2	De-normalized F_SALE Fact Table	205
F3.3	F_ITEM Fact Table	206
F3.4	De-normalized F_ITEM Fact Table	206
F3.5	F_FEE Fact Table	207
E3.6	De-normalized F_ FEE Fact Table	207

LIST OF TABLES

Table No.		Page
3.1	Comparison between approaches	33
3.2	Descriptions and Codes Summary	65
3.3	Modeling Techniques Summary	66
5.1	Validation Summaries	89
5.2	Summary of Lexical Number for Component Entities	101
5.3	Summary of Lexical Number for Transaction Entities	101
5.4	Summary of Dream House result	110
6.1	TPC-H Results Summary	127
6.2	RETAIL Results Summary	132
6.3	Comparison of result with Benchmark	136
6.4	Comparison of methods adopted for TPC-H Schema	158
6.5	Comparison of methods adopted for Retail Schema	143
B.1	Complete List of Lexicographer Files	169
B.2	Lexical Number of Known Component Entities	170
B.3	Lexical Number of Known Transaction Entities	171
B.4	List of words with type and tag count	172
B.5	Frequency of Transaction Type Matched Lexical Number	180
B.6	Frequency of Component Type Matched Lexical Number	180
D3.1	Summary of Dream House result	191
E2.1	TPC-H Results Summary	195
F2.1	RETAIL Results Summary	204

LIST OF ABBREVIATIONS

API	Application Programming Interface
CASE	Computer Aided Software Engineering
CLOS	Common LISP Object System
CTV	Connection Topology Value
DAG	Directed Acyclic Graph
DAML	DARPA Agent Markup Language
DARPA	Defense Advanced Research Projects Agency
DBMS	Database Management System
DFM	Dimensional Fact Model
DW	Data Warehouse
EIS	Executive Information Systems
ER	Entity-Relationship
ERD	Entity-Relationship Diagram
ERM	Entity Relationship Model
ERP	Enterprise Resource Planning
ETL	Extract-Transform-Load
EVER	Event Entity Relationship Model
GQM	Goal/Question/Metrics
KPI	Key Performance Indicator
KR	Knowledge Representation
LISP	Locator/ID Separation Protocol
MD	Multidimensional Data
MDBE	Multidimensional Design by Examples
ME/R	Multidimensional Entity Relationship
MIS	Management Information Systems
MNF	Multidimensional Normal Form
MOLAP	Multidimensional Online Analytic Processing
OIL	Ontology Inference Layer
OLAP	Online Analytic Processing
OLTP	Online Transaction Processing
OWL	Web Ontology Language

PC	Personal Computer
QVT	Query/View/Transformation
RDF	Resource Description Framework
ROLAP	Relational Online Analytic Processing
SDLC	System Development Life Cycle
SERM	Structured Entity Relationship Model
SQL	Structured Query Language
TPC/H	Transaction Processing Council/H
UML	Unified Modeling Language
UNSPSC	United Nations Standard Products and Services Codes
XML	Extensible Markup Language

CHAPTER I

INTRODUCTION

1.1 BACKGROUND

Organizations must be able to manage and use their information resources efficiently in order to survive in the competitive business world (Burstein et al. 2008). Operational data base system is designed to support the day-to-day running of the business process. On the other hand, a different type of decision support system is needed for strategic information. Many organizations are sitting on vast amount of data already accumulated in their operational database. The huge data repository is a gold mine waiting to be explored for potential strategic business information to support corporate executives making important decisions. As a result, the organization turned to the creation of a data warehouse, where the data are integrated, transformed, cleaned and loaded into multidimensional schemas which in turn provide strategic information to top management (Ponniah 2001).

The basic structure of a star schema multidimensional model consists of fact table as the "center" of the star with several dimensional tables forming the "points" of the star around it (Kimball & Ross 2002). Fact table captures the core data of the business transaction that can be used for analysis such as *customer*, *sale amount*, *product*, and *transaction date*. Measure is the numeric attribute for fact which can be aggregated. Examples for measures are *quantity sold* and *sale amount* (Moody & Kortink 2000). Dimension provides business perspective on the transaction such as who is the *customer*, which *branch* handles the sale, which *product* has been purchased and the *date* when the transaction happened (Cabibbo & Torlone 1998). Dimension can be broken down into hierarchies such as *Country*, *State* and *City*.

Many organizations that already owned matured operational information systems started to take advantage of their existing data by applying several tools for information analysis, decision support, and strategic planning. Organizations are moving towards the knowledge-based technologies which are also known as "Business Intelligence" (Watson & Wixom 2007). The new requirement leads to the creation of data warehouse systems that are based on analysis of data done on multidimensional databases (Kimball & Ross 2002). Currently there are many available products off-the-shelf for data warehouse that support building physical data models and On-Line Analytical Processing (OLAP). On the contrary, the tasks of designing conceptual and logical data warehouse remain largely on the skill of system analyst or data warehouse analyst of respective vendor or organization that build the data warehouse (Dori et al. 2005).

Designing a data warehouse based on the existing operational database is a very complex and time consuming process (Romero & Abelló 2009). Most of the existing Data Warehouse design tool employs direct transformation of input into corresponding designs (Golfarelli et al. 1998; Boehnlein & Ulbrich-vom Ende 1999; Hüsemann et al. 2000). Many literatures have indicated that identifying fact is the most important step in data warehouse design process (Dori et al. 2005; Romero & Abelló 2009), and most of the time pointing fact is usually done manually (Moody & Kortink 2000; Romero & Abelló 2007). As a result, many of the existing tools do not have complete cycle of the design process, hence, an algorithm need to be formalized in order to recognise the potential fact table in designing data warehouse. Noah and Williams (2004) indicate that design automation is more intelligent with the ability to tap business knowledge about semantics of the application domain. Romero and Abelló (2007) propose a semi-automatic method to generate data warehouse multidimensional concept using domain ontology. The ontology is used to discover business multidimensional concept buried in the business domain. Sugumaran and Storey (2006) use domain ontology to represent the domain knowledge in order to assist database designer.

1.2 PROBLEM STATEMENT

Designing a Data Warehouse based on the current Operational Database is a very complex and time consuming process. Many organizations have invested substantial amount of resources and time in developing data warehouse (Chenoweth et al. 2003). Computer industries and researchers have developed several methodologies and approaches to design data warehouse model (Romero & Abelló 2009). Most of the techniques required hands-on involvement of data warehouse expert in painstaking exercise to determine the relevant multidimensional structure buried in the source operational data (Ponniah 2001).

In order to improve productivity, researchers build programming tool to support user with data warehouse design process. Despite a great number of Computer Aided Software Engineering (CASE) tools developed for data warehouse design automation, the tools are more successful towards capturing data for documentation purposes, designing physical model and setting up foundation for On-Line Analytical Processing (Dori et al. 2005). The CASE tools are lagging with real knowledge that deprived them of diagnostic capability or understanding the basic semantic of words that represent entities or attributes (Noah & Williams 2003). On the contrary, real human designers are able to analyze problems, provide solutions and solve ambiguities because of their experience and knowledge of the real world (Noah & Williams 2000).

Many researches (Sugumaran & Storey 2006; Romero & Abelló 2007) incorporate artificial intelligence algorithm in the form of knowledge-based systems in order to assist in design process. Nevertheless, most of existing Data Warehouse design tools employ direct transformation of input into corresponding designs and rely on the users to identify suitable entities to be modeled as fact tables (Phipps & Davis 2002). Currently existing tools are unable to detect such entities and leave them to users for decision. This represents a major gap in data warehouse tools lacking the most important function for the automation process of data warehouse design. Therefore, it is the intention of this research to close the functional gap by providing semantic intelligence for the task of designing Data Warehouse model.

Entities that represent concept related to business transaction hold the best criteria needed to be selected as potential candidates for fact table. Moody and Kortink (2000) suggest that the most suitable candidates for fact table are the transaction type candidates such as *booking* and *purchase* while component type candidates such as *customer* and *region* are the best candidates for dimensional table. In addition, Baekgaard (1999) proposes for entities that signify events in business transactions such as *sales* or *order* are the best candidates for fact tables. An experience data warehouse analyst can easily identify the transaction or event entity by the semantic meaning of the entity name. To solve the above-mentioned problem, machines must have semantic intelligence to understand the meaning of entity name that belongs to transaction type or indicates business events in order to select the best candidate for fact tables.

According to Lassila and McGuinness (2001) ontologies are categorized as either light weight or heavy weight and both categories can provide an acceptable specification for term names and term meanings. Light weight ontology such as lexical ontology is sufficient enough to provide the semantic intelligence for identifying suitable entities for fact tables. Therefore, a lexical ontology WordNet is proposed as the source of knowledge domain for configuring the semantic meaning of entity name.

1.3 AIM AND OBJECTIVES OF RESEARCH

The main aim of this research is to investigate the possibility of providing semantic intelligence for the task of designing multidimensional model for Data Warehouse automatically using Lexical Ontology. To accomplish the stated aim, below are the lists of objectives to be fulfilled.

i. To propose a semi-automated method for the task of data warehouse design in accordance with the knowledge-based system engineering technique
ii. To construct algorithms for transforming logical model into data warehouse multidimensional model utilizing semantic knowledge stored in lexical ontology for decision making

iii. To compose a set of diagnosis guidelines for validating input data for the semi-automated process
iv. To develop an intelligent data warehouse case tool employing the above objectives for assisting data warehouse designer
v. To evaluate the validity of the proposed method and guidelines

1.4 IMPORTANCE OF RESEARCH

This research explores the feasibility of developing a CASE tool that assist user analyst to design data warehouse logical model automatically from existing operational database system. The CASE tool is equipped with artificial intelligence capability in the form of knowledge-based system that allows the tool making smart decision in choosing candidate fact tables. The decision making ability contribute in closing the gap of the existing data warehouse CASE tool. In addition, the tool is capable to generate corresponding data warehouse model in star schema. The output of the CASE tool setup the basic working logical model for the analyst to further refine and move to the next stage of transforming the model into physical model.

1.5 SCOPE OF RESEARCH

According to List et al. (2002) data warehouse development methods can be divided into three different approaches such as supply-driven, goal-driven and demand-driven. Supply-driven approaches begin with the study of the existing operational database which in turn reengineered to produce a multidimensional model. Several researchers manage to semi-automate the data warehouse design process within the supply driven framework (Phipps & Davis 2002; Sitompul & Noah 2006). On the other hand, goal-driven approaches integrate corporate strategy and business objective as the requirement for data warehouse design. Giorgini et al. (2005) adopt organizational and decisional modelling while Guo et al. (2006) use Key Performance Indicators as the guideline towards the data warehouse design. Lastly, demand-driven approaches compel users to play very critical role in shaping and contributing to data warehouse design (Winter & Strauch 2003).

In a traditional database design, the construction process starts with requirement study, conceptual design, logical design and follow by physical design (Connolly & Begg 2005). However, there is no standard comparable methodology to the database design, although there are many literatures discussing similar method based on "first principles" approach in designing data warehouse (Hüsemann et al. 2000). Similarly with database design, data warehouse design can be divided into three levels; the conceptual design, logical design and physical design. Construction of a data warehouse design starts with formation of conceptual model which is independent of implementation details such as software requirement or hardware platform. In the logical design stage, logical model is structured according to specific data model but independent of any particular Data Base Management System (DBMS) or hardware platform. Finally at the implementation level, the logical model is transformed into physical model directly related to a specific DBMS and optimized according to the hardware recommended performance.

For the purpose of this research, the research scopes are limited to solving problem at the logical design stage, constructing data warehouse model based on data-driven development approach, and focusing data warehouse solution restricted to business application. Our decision for the aforementioned scope is based on the following reasons. All major corporations already running full blown operational database system, and almost as many are using Management Information System (MIS) for data analysis. Hence, naturally the supply-driven development approach is the best method to design their new data warehouse system based on the existing operational database (Inmon 2005). The logical design stage is selected for the research because the blue print of the existing operational database is captured in the logical schema. At this level, the logical schema is very stable as compared to conceptual schema and flexible enough for modification as compared to physical schema. In another word, analysts are free from conceptual issues such as normalization of entities, many-to-many relationship, the superclass and subclass entities, and relationship with attributes (Codd 1970; Chen 1976). Finally, our research is restricted to the business domain since majority of the organizations are focusing on the financial aspect of the business. Furthermore, the specification of rules and algorithms can be customized relevant to the business industry.

1.6 RESEARCH METHODOLOGY

The research methodology for this work consists of three major phases. The first phase is the understanding of related works done in the same area of research and the formulation of proposed method for solving the research problem. The second phase is the implementation of method derived from the first phase with a development of a prototype tool. And the third phase is to derive an evaluation method in order to validate the result from the prototype.

The first phase starts with the purpose of comprehensive understanding about data warehouse and work related to data warehouse design. The process is done through literature review of various books, journals, conference papers and technical reports in the following related areas:

i. Research works on the basic function of data warehouse and designing data warehouse model. The works provide necessary understanding of data warehouse and how it supports organization with strategic decision making process. Furthermore, the study also provides the basic foundation on multidimensional model and important principles in data warehouse design.
ii. Research works on various multidimensional modeling methods. Review on the approaches employed by research works, the models used to describe data warehouse concept, type of data sources used as input and the deployment of automation process.
iii. Research works on ontologies and the application of ontologies in automation design.

After reviewing the research works, we document all the relevant methods and the results produced from each method. We study the reports and identified any gap existed in the current research. Based on the review, we formulate the research questions that need to be answered by this research.

From the review of all the research works, we identify that many of the works presented are unable to recognize fact tables automatically (Moody & Kortink 2000; Romero & Abelló 2007). The process of identification is done manually by user

analyst or is verified automatically against user queries. To improve on the current method, we proposed a transformation-oriented approach with artificial intelligence capability in decision making process. The knowledge provided to the intelligent process originates from lexical ontology acting as the knowledge domain. To support the transformation-oriented algorithm, we need to formulate the following:

i. A specification language model. The function of this model is to store semantic contents of the enterprise logical schema used as input.
ii. A set of syntax guidelines. These guidelines filter any syntax errors found in the input records and suggest diagnosis solution to the problem.
iii. Fact table algorithm. This algorithm is used to identify potential fact tables among the entities presented in input relational schema.

In the second phase, a prototype is developed to demonstrate the feasibility of the proposed transformation-oriented algorithm. The prototype is developed based on an incremental system development life approach. The development process is divided into four phases as the following:

i. Analysis Phase - The objective of this stage is to understand the requirements, to analyze previous works in the same area, to select the right hardware or software.
ii. Design Phase - The objective of this stage is to design the solution based on the problem defined during analysis phase.
iii. Implementation Phase –The objective of this stage is to implement the system design into actual programming coding represented by a prototype.
iv. Testing Phase - The objective of this stage is to verify the correctness and consistency of the prototype.

The third phase is focus on formulating evaluation method for validating the result of the prototype. There is no standard testing available for validating the output model of data warehouse design.

1.7 ORGANIZATION OF THESIS

This thesis consists of seven chapters. Chapter I is the introduction of the thesis that explains the background of data warehouse and the problem related to data warehouse

design. It also presents the aim and objectives of the research together with the importance of research and the scope of the research. This chapter also describes the research methodology.

Chapter II covers two main subjects, general information about data warehouse and description of ontology. Discussion on the data warehouse consists of the following area; data warehouse in general, history of data warehouse, the importance of data warehouse, data warehouse architecture, different types of data warehouse and data warehouse models. Second part of the chapter describes about ontology, types of ontology.

Chapter III analyzes several aspects of data warehouse development process such as data warehouse design stages, multidimensional model and multidimensional modeling techniques. The five design stages are requirement analysis, conceptual design, logical design, extract-transform-load (ETL) process design and physical design. The chapter also discussed about nine different types of multidimensional models. Some of the models enhanced on the existing entity-relationship model while few use graphical presentation to present facts and dimensions. The biggest portion of chapter III compares twenty previous modeling techniques works especially on the design approach employed, data sources used, input and output abstraction level, and the automation utilized.

Chapter IV explains about the research method implemented in this research. The research method consists of conducting extensive literature review, developing a running prototype implementing the proposed method and designing evaluation process to validate the result.

Chapter V explains in great detail the approach chosen and algorithm employed for the data warehouse modeling method. The three main stages for the modeling process are; i) translation of logical data model into specification language model, ii) identification of fact and dimensional tables, and iii) generation of multidimensional model. Each of the major steps is broken further into several smaller tasks.

Chapter VI describes the result obtained from testing the ADW-tool developed based on the objective of the research. The results are obtained by running two sets of data using the data warehouse tool. Output from existing researches done using the same input data for the same objective are set as the benchmark for the test. Results from the test are compared to the output from the benchmark in order to verify the accuracy and validity of the test.

Chapter VII or the final chapter presents the conclusion of this thesis on the research contribution in the automation of data warehouse design. The chapter also elaborates future work that can be explored which will improve the process of data warehouse design.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter discusses primarily on the main topics namely general information about data warehouse and description of ontology. Discussion on the data warehouse consists of the following area; data warehouse in general, history of data warehouse, the importance of data warehouse, data warehouse architecture, different types of data warehouse and data warehouse models. Second part of the chapter describes about ontologies, classification of ontologies, types of ontologies, application of ontologies and lexical ontology.

2.2 DATA WAREHOUSE

In today's global business environment, every giant corporation is using computer system to run businesses. The system is used to process ledger, order, inventory, billing, training and many other applications. As the production database become more mature, comprehensive, and easily access throughout the globe, the focus of organizations requirement has also shifted from merely automation to also include "Business Intelligence" (Watson & Wixom 2007). Organizations have stored large volume of business data in their operational databases. Many of the data have been kept for many years. As business become very competitive, corporate executives are looking at ways on how to tap valuable information from these massive data in order for them to stay competitive and improve on the profit margin. Executive need answers to their strategic decision making questions, set up enterprise goals, identified objectives and monitor results. Although they have the data themselves, but accessing

the production data for analytical processing in not easy. Database from the traditional OnLine Transaction Processing (OLTP) system are designed to provide information for day-to-day operations. The need for different kinds of information suitable for OnLine Analytical Processing (OLAP) geared towards strategic information triggered the design of Data Warehouse (Ponniah 2001).

Information has becomes one of the most important assets of any organization. Business communities as well as government agencies are looking into data warehousing as one of the solutions in transforming their vast amounts of data into something meaningful for their strategic decision making process(Watson et al. 2004). Data Warehouse also provides the raw data for powerful data analysis technique such as data mining and multidimensional analysis as well as the fundamental queries and reporting. As data warehousing become one of the most important tools for business, by the year 2005, organizations were estimated spending about US$150 billion for hardware and software in the data warehousing market (Chenoweth et al. 2003). Top management improved on decision making performance with the implementation of data warehouse (Park 2006). The benefit generated from data warehouse can only be acquired if the data warehouse structure has been designed correctly to meet the user queries and strategic demand for decision making.

2.2.1 WHAT IS A DATA WAREHOUSE?

Bill Inmon (2005), considered to be the father of data warehouse provides the following definition: "A Data Warehouse is a subject-oriented, integrated, non-volatile and time–variant data supporting management's decisions". Connolly and Begg (2005) elaborate Inmon definition as follows:

- *Subject-oriented*: In data warehouse, data is structured based on important subject of the organization such as customer or product. This data provides top management with strategic information for decision making process. In operational systems, data is organized by function to support specific business process on day to day activities. For example, an order entry system is to capture data about customer ordering some products from the company website.

- *Integrated*: Data from across the organization residing on multiple applications, platforms and architectures is integrated producing a single enterprise-wide view. The integrated data goes through transformation process such as validation, correction, cleaning, standardizing, streamlining and summarizing.
- *Time-variant*: The data does not reflect the actual operation data at real time. Data represents several snapshots of value at specific points in time, or a comprehensive history of the data. The data also has historical value since it is kept for longer period of time. Attribute time will always be part of data warehouse data structure.
- *Non-volatile*: The data are not updated real time and new data are loaded from operational system at specific time frame or frequency. It contains the copy of transaction data which are not subjected to update or change once written into the data warehouse.

One of the fundamental philosophies of building a data warehouse is to structure data as the central pillar and design application around it (Ballard et al. 1998). Data mart is a subset of a data warehouse which focuses on particular needs of the department or business unit of the organization. According to Marakas (2003) relevant data from the data warehouse is loaded to the data mart at certain interval. Those data are more stable since they have been cleaned and verified at the data warehouse level.

2.2.2 HISTORY OF DATA WAREHOUSE

According to Ballard (1998) the concept of data warehouse started in the early 1980s with the widespread usage of relational database management system. Many organizations have already developed comprehensive operational database systems. Over the years, companies have accumulated mountains of data through their operation system. These data are more suitable to answer the day-to-day operation inquiries but not appropriate for questions related to strategic decision making. Senior business managers require strategic information for decision making. In order to fulfill the senior management request, Information Technology personnel writes program to

extract data from online database. In most cases, the data originate from multiple systems reside on multiple platforms which employ diverse file structures. Several different programs are written to handle each different situation. These data are extracted at particular time interval and stored in special database design for decision making process. The data represent snap short of operational data at particular time interval.

As a result of extraction process, system such as Management Information System (MIS) or Executive Information Systems (EIS) mushroom throughout the industries to support ad-hoc queries and standard reporting function. The decision support system data bases are separated from the operational system in order to improve system performances. Data from the operational system are extracted and loaded into the decision support system at regular interval. The snapshots of the operational data are accumulated in the database for end-users to access using standard queries and commercial reporting tools. Data model used are subset of the operational data model since the records are loaded directly from the operational system database. Top management utilized MIS or EIS as tools for them to analyze business trend, historical data, customers' preference, suppliers' price and much more valuable information.

2.2.3 WHY USING A DATA WAREHOUSE?

One of the reasons why an organization should migrate from EIS to Data Warehouse based system is because the EIS data is not reliable and not consistent. EIS database is loaded by using extraction data from time to time. Inmon (2005) stated that many operations started as a simple extract, and then there are more extracts; and extracts of extracts. As time goes by, the extraction process has become out of control and the source of data is not credible. Reports or queries resulting from two different departments regarding the same sales projection are more likely not to be the same. This is because of the following factors such as timing for extraction, level of extraction, incompatible external data and multiple sources of data. Big corporation usually has multiple EIS systems; one for each different division. This creates another

set of problems such as code standardization. One code might stand for different meaning as compared to different division.

Senior managers evaluate their business in term of business dimensions. Requirement for operational managers are different than the senior managers of a corporation. Operational managers are interested at micro level of the operation as compared to senior managers who are looking at macro level of the corporation. As end-users become more matured and the organizations grow bigger and global, the executive strategic demands from MIS or EIS systems also expanded. The systems could no longer depend on local homogeneous databases. The executives are now interested with national business reports. The operational system or the Information Technology department could not support the overwhelming demands of business strategic users. With the expanded scope, the requests for the reports are very tedious and time consuming through reading process from multiple platforms and various incompatible database systems. Hence executives' requests are bound by the limitation imposed by the system. Data Warehouse integrates data from across and outside the organization in a complete and consistent manner (Marakas 2003). The data warehouse will provide the complete puzzle consists of pieces coming from the entire organization for strategic management users to use as tools for their decision making process.

Business analysts are interested to study business trend over a long period of time. Operational database usually contains current data or only active data which are designed to support operational or clerical function. Completed transaction or closed account will be pushed to the historical file stored in secondary medium of storage. Since data warehouse consists of historical data across several years, business analysts are able to study several analyses such as business trend, market competitiveness, consumer behavior, and supplier reliability. Data warehouse can support middle and top management to discover new information which is vital for the organization future direction and business survival.

Business analysts are trained to operate the data warehouse analysis tools themselves. Hence, this newly acquired skill free them from relying on the IT

department. They can have access to the data warehouse anytime for their business analysis process. Together with the standard report presented to them, business users like to mix and match the data with multiple combinations with a process commonly defined as slicing and dicing (Kimball & Ross 2002). This process can be done through several analysis tools available for data warehouse.

2.2.4 DATA WAREHOUSE ARCHITECTURE

In this section we present the general overview of data warehouse architecture and the major component of a data warehouse. Data warehouse architecture can be divided into four main components; data source, data staging area, data warehouse system, and information delivery system as illustrated in Figure 2.1 (Ponniah 2001).

Data warehouse does not have any significant value without the data content (List et al. 2002). The first component of data warehouse architecture is the data source that can be grouped into two main categories; i.e. the internal data and external data. Internal data consist of production data and archived data. The operational system is the legacy systems that capture the day-to-day business transactions record throughout the organization and store the production data. In contrast, the archived data are old data which are no longer needed by the operational system and are kept in secondary storage for future reference. On the other hand, the external data are produced by external agencies such as the Statistical Department, Chamber of Commerce Organization or The World Bank. These data can be used to measure the performance of the business organization against the outside world or the competitors. Selection from both types of data is the potential candidate for data warehouse.

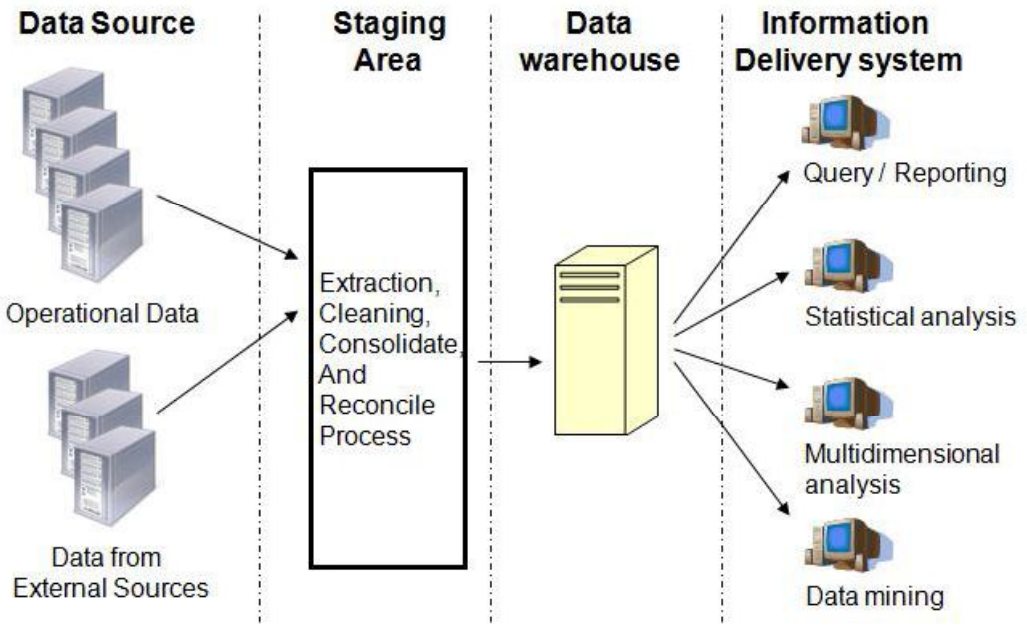

FIGURE 2.1 Data Warehouse System Architecture

Data extracted from the data source must process at the staging area to resolve data conflicts before being loaded into the data warehouse. This process is commonly known as extract-transformation-load (ETL) process (Simitsis & Vassiliadis 2003). In the staging area, the data will go through transformation process such as cleaning of data from misspelling, locating missing data, standardizing code coming from multiple business units, combining data from multiple sources, deleting duplicate data, assigning warehouse keys and foreign keys. The final step of ETL process is to load the transformed data into the data warehouse. The initial ETL process involves massive work of reading production and archived data. Subsequently, regular ETL is scheduled for the production data and external data at specified time interval.

Data warehouse is central repository systems for the enterprise business data designed to support the decision making within the organization. Large volume of data is kept over the span of several years as to represent the historical data for the organization. Metadata describe detailed information about the data similar to the data dictionary or a library card catalog. This information is very useful for user finding and navigating data using the end-user application tools. Data warehouse structure model is based on subject orientation such as customer or product in contrast to the

operational system which is structured by applications. Data mart is a subset of data warehouse and usually tailored to the need of specific department or user group.

The last segment is the information delivery system where the end-users are able to interact with the data warehouse system using several end-user access tools such as reporting and query tools, EIS tools, OLAP tools and data mining tools (Connolly & Begg 2005). Reporting tools generate the regular operational reports required by the organization during nightly batch job. Query tools allow end-users to submit predefined standard query statements or execute customized Structured Query Language (SQL) statements through a user friendly application that supports 'point-and-click' creation of SQL. Senior managements are able to utilize the EIS to support their strategic decision making process. Sophisticated users utilize OLAP tools to analyze data using complex multi-dimensional view navigating through massive amount of data or answering difficult and complicated questions such as identifying business trend or projecting future forecast (Lakshmanan et al. 2008). The last tool under information delivery system is the powerful data mining that helps users to discover meaningful patterns or trends hidden among the huge data repository using several techniques such as statistical, mathematical and artificial intelligent algorithm (Han & Kamber 2006). Empowering the users with data warehouse access tools contribute tremendously to the success of data warehouse system investment.

2.2.5 DATA WAREHOUSE MODEL

A model is the representation of an abstract idea and implementation independent for the designer to visualize before the construction of the physical data warehouse (Ballard et al. 1998). With a data model, it is much easier for developers and users to understand the structure and relationship of data in the data warehouse. Data model provides a means for designer to construct representations of reality. The developers of operational database system have been using Entity-Relationship (ER) Diagram to represent their conceptual model. The ER model which is introduced by Chen (1976) has become the standard model for conceptual database design. However, ER modeling with normalized tables is not suitable for data warehouse system (Kimball et al. 1998). Multidimensional model is a new modeling technique that has emerged to

support data warehouse performance especially on data analysis function. Data warehouse model can be classified into three major types: cube models, multidimensional models, and statistical models (Pedersen 2000).

a. Cube Models

Cube models is a collection of data cells arranged data in the form of n-dimensional cubes (Gyssens & Lakshmanan 1997). The number of dimensions for cube is not limited to 3 dimensions; a cube can have any number of dimensions and a cube that has more than 3 dimensions is called hypercube. Each cell holds measures or also called quantifying data which can be used to describe fact. The axes of the cube are called dimensions or qualifying data that provide another view for analyzing the fact.

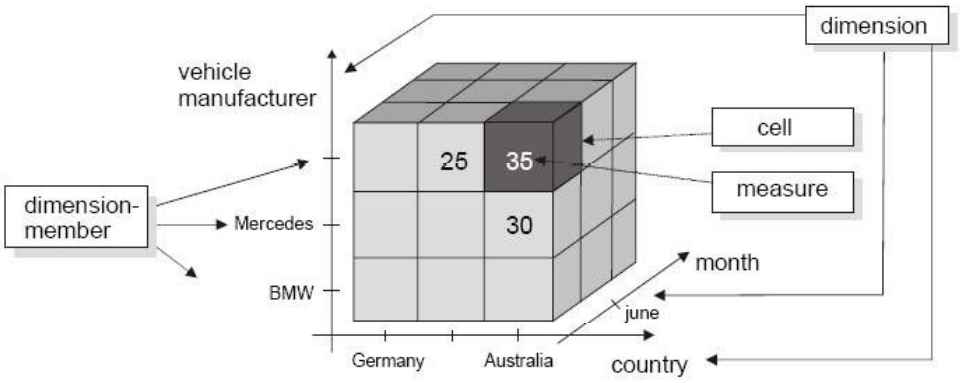

FIGURE 2.2 Data cube representing vehicle sale (Sapia et al. 1998)

Figure 2.2 depicts a cube that represents vehicle sale in several countries, by month and manufacturer. Using the cube as an example, user knows that only 30 Mercedes cars were sold in Australia for the whole month of June. The month dimension can be aggregated into coarser granularities such as quarter or half-year. The only drawback is that only 3 dimensions cube can be shown graphically.

b. Multidimensional Models

The basic constructs of multidimensional model is facts, dimensions, and dimension hierarchies. Fact table is the primary table that contains primary key, foreign key and

numerical attributes. Measure is the numerical attribute for fact table which can be aggregated and can be used to measure the performance of the business organization. An example of fact table is the vital data for the organization business transaction that can be used for analysis such as *Sales*. Examples for measures are *Product sold* and *Price*. Dimensional table provides business perspective on how to analyze fact such as *Sales* based on *Geography* (Cabibbo & Torlone 1998). Granularity level represents the level of detail in the fact table. The higher the granularity level the more summarized the data. On the other hand, the lower the granularity level the more detail the data (Ponniah 2001).

Dimension attributes provide the query restriction, groupings and report labels for user analysis requirement. Dimension table often defines into hierarchies representing the business relationship. An example of hierarchy is when products can be rolled up into brand and then into categories.

c. Statistical Models

Statistical data model is constructed based on the notions of summary table, summary attribute and category attribute. The model consists of a structured classification hierarchy with an explicit aggregation function that allows a researcher to work only on a single measure; hence enable to answer one specific set of questions. Even though this approach is not flexible, it provides some protection against incorrect query. Some examples of statistical models are STORM (Rafanelli & Shoshani 1990) and Mefisto (Rafanelli & Ricci 1993).

2.3 ONTOLOGY

The question about the universe has been asked since the time when human started to ponder about their own existence. In Philosophy, the word ontology means a systematic explanation of being. Hence, ontology is the study of being whether it is a physical being such as the sun or the abstract being like the angel (Sowa 2000). Gruber (1993) defines ontology as "an explicit specification of a conceptualization". The KACTUS project defines ontology as means for describing conceptualization in a

modular component with the ability to redesign and reuse of knowledge-intensive system components (Schreiber et al. 1995).

Merriam Webster registered two definitions for ontology since 1721 (McGuinness 2003). While ontologies are known to human for a long time, they remained the topic of discussion among philosophers, linguists and librarians until recently caught the interest of researchers from Artificial Intelligence, Computational Linguistics and Database Theory areas (Guarino 1998). Ontology covers wide variety of subjects that comprise everything about knowledge. In order to make ontology more manageable, researchers have defined domain ontology as semantics terms and relationships about some domain of interest (Sugumaran & Storey 2006). A fast and cost effective strategy in building ontology is to extract several domain specific ontologies and knowledge bases from large ontologies such as SENSUS (Swartout et al. 1996). Design automation will be more intelligent with the ability to tap business knowledge about semantics of the application domain (Noah & Williams 2002).

2.3.1 CLASSIFICATION OF ONTOLOGIES

There is no clear classification of ontologies since its represent diverse spectrum of concepts, build with various structures, and use for multiple applications. Some significant works on the classification of ontologies are conducted by several researchers (Mizoguchi et al. 1995; van Heijst et al. 1997; Guarino 1998; Lassila & McGuinness 2001).

Guarino (1998) classifies ontologies based on their level of generality into four types; top-level, domain, task and application ontologies as shown in Figure 2.3. Top-level ontologies represent general concepts like space, time and matter which are independent of any particular domain. While domain and task ontologies are related to a generic domain such as medical or generic task such as selling. On the other hand, application ontologies express concepts depending on certain domain or task.

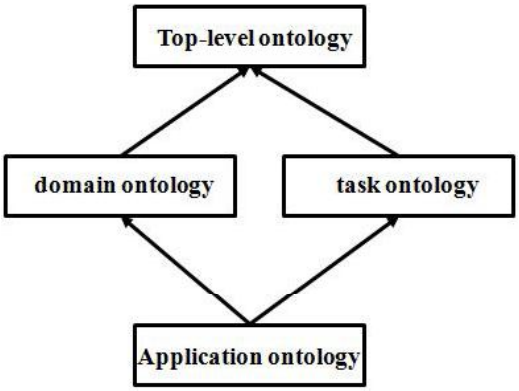

FIGURE 2.3 Ontology level (Guarino 1998)

Lassila and McGuinness (2001) categorize ontologies based on the information that the ontology represents and the complexity of the internal structure. Ontology is used to present specification of term names and term meaning from the simplest form such as catalog to a more complex representation with logic constraints between terms. Lightweight ontologies are mainly concept taxonomies, relationship between concept, and properties that describe concepts. On the other hand, heavyweight ontologies include axioms and constraints on top of the lightweight ontologies. Both lightweight and heavyweight ontologies can be modeled using multiple knowledge modeling techniques into various type of languages (Uschold & Gruinger 1996). Figure 2.4 depicts ontologies in a continuous line starting from lightweight and progressively becoming heavyweight ontologies.

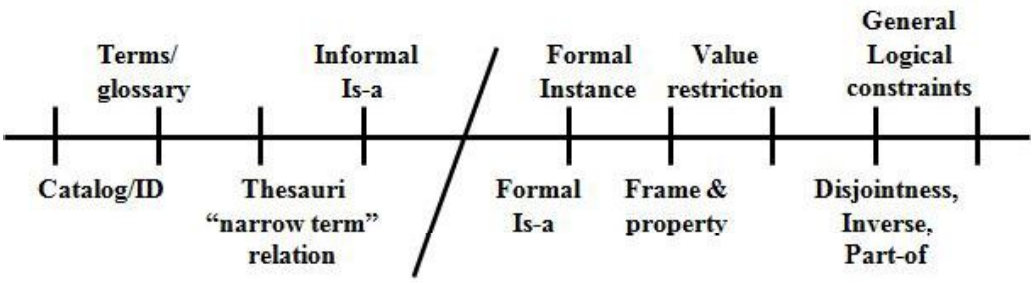

FIGURE 2.4 An Ontology Spectrum (Lassila & McGuinness 2001)

2.3.2 TYPE OF ONTOLOGIES

This section covers the most outstanding ontologies based on their usage in major projects and theoretical contributions. The selected four types of ontologies are

knowledge representation ontologies, top-level ontologies, linguistic ontologies and domain ontologies.

a. Knowledge Representation Ontologies

Knowledge representation (KR) ontologies are the basic modeling attempt to formalize knowledge in a KR paradigm with primitive items such as classes, relations and attributes (van Heijst et al. 1997). The most popular KR ontology is the Frame Ontology developed by the Knowledge Systems Laboratory at Stanford University (Gruber 1993). While Resource Description Framework (RDF) is designed specifically for describing Web resources with metadata (Lassila & Swick 1999). Ontology Inference Layer (OIL) is developed as an extension of RDF using a layered approach (Fensel et al. 2000). DARPA Agent Markup Language (DAML) is project funded by United States Defense Advanced Research Projects Agency (DARPA) started in 1999. DAML+OIL is generated as an extension of RDF but not in different layer while Web Ontology Language (OWL) is derived from DAML-OIL language (Antoniou & van Harmelen 2003).

b. Top-Level Ontologies

Top-level ontologies describe general concepts that cross all domains which are the umbrella for all existing ontologies. Guarino and Welty (2000)define concepts which instances are universals as the top-level ontology of universals that has four attributes; rigidity, supplies identity, carries identity and dependency. On the other hand, Cyc's Upper Ontology which is part of Cyc Knowledge Base consists of large amount of common sense knowledge (Guha & Lenat 1990) while Sowa's top-level ontology has 27 concepts derived from logic, linguistic, philosophy and artificial intelligence (Sowa 2000).

c. Linguistic Ontologies

The purpose of linguistic ontologies is to describe the semantic constructs bound by the grammatical units rather than to model a specific domain. They offer large volume