

**PENGINTEGRASIAN DATA EXTENSIBLE MARKUP LANGUAGE (XML)
KE DALAM PANGKALAN DATA HUBUNGAN MENGGUNAKAN COMMON
WAREHOUSE METAMODEL (CWM)**

WAN MOHD HAFFIZ BIN MOHD NASIR

Tesis ini dikemukakan
sebagai memenuhi syarat penganugerahan
ijazah Sarjana Sains (Sains Komputer)

Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia

SEPTEMBER 2007

ABSTRAK

Permintaan kepada pengintegrasian data secara pantas menjadi semakin tinggi dengan semakin banyak sumber-sumber maklumat yang terdapat di dalam perusahaan moden. *Extensible Mark-up Language* (XML) telah menjadi satu piawaian baru bagi perwakilan dan pertukaran data dalam *World Wide Web* (WWW), contohnya di dalam aplikasi *Business to Business* (B2B) pada e-dagang. Ini memerlukan alatan analisis data untuk mengendalikan data XML di samping format data tradisional. Tujuan penyelidikan ini adalah bagi meningkatkan kaedah pengintegrasian data XML ke dalam data hubungan berikutan berlakunya beberapa masalah daripada hasil proses pengintegrasian seperti kehilangan data. Kaedah yang dicadangkan daripada penyelidikan ini adalah melalui pengaplikasian *Common Warehouse Metamodel* (CWM) sebagai piawaian bagi pertukaran dan pengurusan metadata yang menggabungkan satu perkongsian metamodel bagi sintak dan semantik metadata. Hasil daripada penyelidikan ini adalah satu cadangan peningkatan senibina dan pendekatan pengintegrasian melalui pengaplikasian CWM serta satu perisian antaramuka yang telah dibangunkan bagi memudahkan proses pengintegrasian sebagai penyelesaian kepada masalah kehilangan data.

ABSTRACT

The demand for rapid data integration is getting higher as more and more information sources appear in modern enterprises. Extensible Mark-up Language (XML) is fast becoming the new standard for data representation and exchange on the World Wide Web, e.g., in B2B e-commerce, making it necessary for data analysis tools to handle XML data as well as traditional data formats. The purpose of this research is to enhance the technique for XML data integration into relational data to solve integration problems such as missing data. The method that had been proposed is to apply Common Warehouse Metamodel (CWM) for metadata interchange and metadata management that incorporates a common shared metamodel for metadata syntax and semantics. The results of this research are the enhancement of the integration architecture and approach by applying CWM as well as the development of an interface software to simplify the integration process as a solution for the missing data problem.

KANDUNGAN

| BAB | PERKARA | MUKA SURAT |
|------------|--------------------------------|-------------------|
| | PENGESAHAN STATUS TESIS | |
| | PENGESAHAN PENYELIA | |
| | JUDUL | i |
| | PENGAKUAN PENULIS | ii |
| | DEDIKASI | iii |
| | PENGHARGAAN | iv |
| | ABSTRAK | v |
| | ABSTRACT | vi |
| | KANDUNGAN | vii |
| | SENARAI JADUAL | viii |
| | SENARAI RAJAH | ix |
| | SENARAI SINGKATAN | xiii |
| | SENARAI ISTILAH | xxi |
| | SENARAI SIMBOL | xxiii |
| | SENARAI LAMPIRAN | xxiv |
| 1 | Pengenalan | 1 |
| | 1.1 Pengenalan | 1 |
| | 1.2 Latarbelakang Masalah | 2 |
| | 1.3 Penyataan Masalah | 4 |

| | | |
|----------|---|-----------|
| 1.4 | Matlamat Utama | 5 |
| 1.5 | Objektif | 5 |
| 1.6 | Skop Penyelidikan | 6 |
| 1.7 | Kepentingan Penyelidikan dan Sumbangan Ilmiah | 6 |
| 1.8 | Struktur Thesis | 8 |
| 2 | KAJIAN LITERASI | 10 |
| 2.1 | Pengenalan | 10 |
| 2.2 | Penyelidikan Pengintegrasian Data | 11 |
| 2.3 | XML | 12 |
| 2.4 | Metadata | 12 |
| 2.5 | Perbincangan | 14 |
| 2.6 | Pengintegrasian Data | 16 |
| | 2.6.1 Sejarah dan Contoh Pengintegrasian Data | 16 |
| | 2.6.2 Contoh Pengintegrasian Data | 18 |
| 2.7 | Common Warehouse Metamodel (CWM) | 19 |
| | 2.7.1 Struktur CWM | 20 |
| 2.8 | Meta Integration | 23 |
| 2.9 | Pendekatan dan Senibina Pengintegrasian Sedia Ada | 24 |
| 2.10 | Kesimpulan | 27 |
| 3 | METODOLOGI PENYELIDIKAN | 28 |
| 3.1 | Pengenalan | 28 |
| 3.2 | Rangka Kerja Penyelidikan | 30 |
| | 3.2.1 Formulasi Masalah (Fasa 1) | 30 |
| | 3.2.1.1 Kajian Literasi | 32 |
| | 3.2.1.2 Analisis Sistem Semasa | 32 |
| | 3.2.1.3 Proposal Penyelidikan | 33 |
| | 3.2.2 Pembangunan Sistem (Fasa 2) | 33 |
| | 3.2.3 Implementasi dan Integrasi (Fasa 3) | 34 |
| | 3.2.4 Penulisan Laporan (Fasa 4) | 35 |

| | | |
|----------|---|-----------|
| 3.3 | Sumber Data dan Peralatan | 35 |
| 3.4 | Proses Pergerakan Data | 36 |
| 3.5 | Perolehan Maklumat | 37 |
| 3.6 | Langkah-Langkah Proses Pergerakan Data | 38 |
| | 3.6.1 Penakrifan Stor Sumber Data | 38 |
| | 3.6.1.1 Pengurangan Sumber Data | 39 |
| | 3.6.1.2 Penganalisaan Kandungan dan Struktur Data | 40 |
| | 3.6.2 Penakrifan Stor Destinasi | 41 |
| | 3.6.3 Aplikasi Transformasi Data | 42 |
| | 3.6.3.1 Kemungkinan-Kemungkinan Aplikasi Transformasi Data | 42 |
| | 3.6.3.2 Penentuan Transformasi-Transformasi Data | 43 |
| | 3.6.4 Penentuan Antaramuka | 44 |
| | 3.6.5 Penakrifan Keteguhan | 44 |
| | 3.6.6 Senibina Aplikasi | 45 |
| | 3.6.7 Penghasilan Senario Pengujian | 45 |
| | 3.6.8 Migrasi Akhir dan Kriteria Kejayaan | 46 |
| 3.7 | Aplikasi Kepada Maklumat Data Elektronik | 46 |
| 3.8 | Cadangan Senibina Pengintegrasian | 53 |
| 3.9 | Rumusan | 55 |
| 4 | PERLAKSANAAN | 57 |
| | 4.1 Pengenalan | 57 |
| | 4.2 Perbandingan Perlaksanaan Pengintegrasian | 58 |
| | 4.3 Antaramuka Pengintegrasian | 59 |
| | 4.4 Penyediaan Data dan Metadata | 64 |
| | 4.5 Pergerakan Data XML | 69 |
| | 4.6 Implementasi Pengintegrasian Data eBusiness | 70 |
| | 4.6.1 Langkah-Langkah Awal | 71 |
| | 4.6.2 Elemen-Elemen yang Digunakan | 72 |

| | | |
|----------|---|------------------|
| 4.6.3 | Proses Pemetaan | 74 |
| 4.6.4 | Penjanaan Migrasi Data | 82 |
| 4.7 | Kesimpulan | 87 |
| 5 | PENGUJIAN DAN ANALISIS | 88 |
| 5.1 | Pengenalan | 88 |
| 5.2 | Persekitaran Pengujian | 89 |
| 5.3 | Kriteria Perbandingan | 90 |
| 5.4 | Contoh Penyataan Pertanyaan | 91 |
| 5.5 | Matrik Pengujian Perbandingan | 93 |
| 5.6 | Pengujian Perbandingan Sebelum Peningkatan | 94 |
| 5.6.1 | Kajian Kes 1: e-Business | 94 |
| 5.6.2 | Kajian Kes 2: Sumber Manusia | 100 |
| 5.7 | Perbandingan Sebelum dan Selepas Peningkatan Pendekatan | 102 |
| 5.8 | Perbincangan | 109 |
| 6 | PERBINCANGAN DAN KESIMPULAN | 110 |
| 6.1 | Pengenalan | 110 |
| 6.2 | Hasil Penyelidikan | 111 |
| 6.3 | Pencapaian Objektif Kajian | 112 |
| 6.4 | Kebaikan dan Kelemahan Pendekatan Pengintegrasian | 113 |
| 6.5 | Cadangan Pembaikan | 115 |
| 6.6 | Penyelidikan Masa Hadapan | 115 |
| 6.7 | Kesimpulan | 116 |
| | RUJUKAN | 117 |
| | Lampiran A - C | 121 - 134 |

SENARAI JADUAL

| NO JADUAL | TAJUK | MUKA SURAT |
|------------------|---|-------------------|
| 3.1 | Skema Data dan Contoh Data bagi Pengujian | 46 |
| 4.1 | Perbandingan Pengintegrasian Sebelum dan Selepas Peningkatan | 58 |
| 5.1 | Contoh Penyataan Pertanyaan | 91 |
| 5.2 | Matrik Perbandingan Pengujian | 93 |
| 5.3 | Bilangan Kehilangan Data bagi P1 | 95 |
| 5.4 | Bilangan Kehilangan Data bagi P2 | 96 |
| 5.5 | Bilangan Kehilangan Data bagi P3 | 97 |
| 5.6 | Bilangan Kehilangan Data bagi P4 | 98 |
| 5.7 | Bilangan Kehilangan Data Kajian Kes 1 | 99 |
| 5.8 | Bilangan Kehilangan Data Kajian Kes 2 | 101 |

| | | |
|------|--|-----|
| 5.9 | Bilangan Kehilangan Data P1, P2, P3, dan P4 selepas Peningkatan | 103 |
| 5.10 | Peraturan Perbandingan Kehilangan Data bagi P1, P2, P3, dan P4 | 104 |
| 5.11 | Bilangan Kehilangan Data Kajian Kes 2 selepas Peningkatan | 106 |
| 5.12 | Peraturan Perbandingan Kehilangan Data bagi P5,P6, dan P7 | 107 |
| 5.13 | Peraturan Pengurangan Kehilangan Data Keseluruhan | 108 |

SENARAI RAJAH

| NO RAJAH | TAJUK | MUKA SURAT |
|-----------------|--|-------------------|
| 2.1 | Skema bagi Gudang Data | 17 |
| 2.2 | Common Warehouse Metamodel | 21 |
| 2.3 | Senibina Pengintegrasian Data Mikael R. Jensen | 24 |
| 3.1 | Rangka Kerja Operasi | 31 |
| 3.2 | Sumber Data dan Instrumentasi | 36 |
| 3.3 | DTD bagi Dokumen Jualan | 48 |
| 3.4 | Contoh Dokumen XML yang menuruti DTD | 49 |
| 3.5 | DTD bagi Dokumen Pemetaan | 50 |
| 3.6 | Dokumen XML bagi Pemetaan | 50 |
| 3.7 | DTD bagi Dokumen Komponen | 51 |
| 3.8 | Dokumen Komponen XML | 53 |

| | | |
|------|--|----|
| 3.9 | Senibina Pengintegrasian Data and Metadata XML | 54 |
| 4.1 | Skrin ‘Splash’ bagi Data Integration Interface | 59 |
| 4.2 | Menu Utama bagi Data Integration Interface | 60 |
| 4.3 | Antaramuka Data Integration Interface | 61 |
| 4.4 | Menu “About” bagi Data Integration Interface | 62 |
| 4.5 | Menu “Setting Manager” bagi Data Integration Interface | 63 |
| 4.6 | Antaramuka bagi Meta Integration Model Bridge (MIMB) | 64 |
| 4.7 | Dokumen XML bagi Tempahan Belian | 66 |
| 4.8 | DTD bagi Tempahan Belian | 67 |
| 4.9 | Model CWM bagi Pangkalan Data Jualan | 68 |
| 4.10 | Penggunaan MIMB dalam Penukaran Model Data | 69 |
| 4.11 | Pengintegrasian Data eBusiness | 72 |
| 4.12 | Operasi pada Pemetaan Atribut “ShipAddress” | 76 |
| 4.13 | Klaus ‘Where’ pada Pemetaan Kelas “Orders” | 78 |
| 4.14 | Pemetaan Kelas “ <i>Orders_Details</i> ” | 81 |

| | | |
|------|---|-----|
| 4.15 | Persediaan bagi Penjanaan Kod | 82 |
| 4.16 | Penentuan Destinasi | 83 |
| 4.17 | Set Paramater Masa Larian | 84 |
| 4.18 | Pemindahan Perpustakaan yang diperlukan bagi Penjanaan Kod | 85 |
| 4.19 | Hierarki Hasil Pemindahan Perpustakaan | 85 |
| 4.20 | Penjanaan Kod daripada <i>Command Line</i> | 86 |
| 5.1 | Graf Perbandingan Bagi Pertanyaan P1 | 96 |
| 5.2 | Graf Perbandingan Bagi Pertanyaan P2 | 97 |
| 5.3 | Graf Perbandingan Bagi Pertanyaan P2 | 98 |
| 5.4 | Graf Perbandingan Bagi Pertanyaan P4 | 99 |
| 5.5 | Graf Perbandingan Bagi Pertanyaan P1,P2,P3, dan P4 | 100 |
| 5.6 | Graf Perbandingan Bagi Pertanyaan P5, P6, dan P7 | 102 |
| 5.7 | Graf Perbandingan Sebelum dan Selepas Peningkatan bagi Kajian Kes 1 | 105 |
| 5.8 | Graf Perbandingan Sebelum dan Selepas Peningkatan bagi Kajian Kes 2 | 107 |

| | | |
|-----|--|-----|
| 5.9 | Graf Perbandingan Purata Kehilangan Data | 108 |
|-----|--|-----|

SENARAI SINGKATAN

| | | |
|------|---|--|
| CWM | - | Common Warehouse Metamodel |
| XML | - | Extensible Mark-up Language |
| DTD | - | Document Type Definition |
| MIW | - | Meta Integration Works |
| MIMB | - | Meta Integration Model Bridge |
| B2B | - | Business To Business |
| OLAP | - | Online Analytical Processing |
| ETL | - | Extract, Transform, & Load |
| BI | - | Business Intelligence |
| OMG | - | Object Management Group |
| UML | - | Unified Modeling Language |
| URL | - | Uniform Resource Locator |
| OIM | - | International Organization for Migration |
| WWW | - | World Wide Web |

SENARAI ISTILAH

| | | |
|----------------------------------|---|-----------------------------------|
| Sistem Pengurusan Pangkalan Data | - | <i>Database Management System</i> |
| Capaian semula maklumat | - | <i>Information retrieval</i> |
| Pertanyaan | - | <i>Query</i> |
| File Teks | - | <i>Text file</i> |
| Hubungan | - | <i>Relation</i> |
| Jadual | - | <i>Table</i> |
| Bahasa Pertanyaan Berstruktur | - | <i>Structured Query Language</i> |

SENARAI LAMPIRAN

| NO LAMPIRAN | TAJUK | MUKA SURAT |
|--------------------|--|-------------------|
| A | Aturcara kod bagi “Data Integration Interface” | 121 |
| B | Kod Aturcara Menu Utama Data Integration Interface | 126 |
| C | Contoh Sebahagian Kandungan Model Data CWM | 129 |

RUJUKAN

- Agosta, L. (2001). Reports of the demise of metadata are premature. *DM Review* 3
- Auth, G. dan Eitel, V.M. (2002). A Software Architecture for XML-based Metadata Interchange in Data Warehouse Systems.
- Bertino E. dan Ferrari E. (2001). XML and Data Integration. *IEEE Internet Computing*.
- Bremeau C. (2001). [XML Data Movement Components for Teradata](#).
www.metaintegration.com, July 2004.
- DAMA (2005). *"Metadata Based Impact and Lineage Analysis Across Heterogeneous Metadata Sources"*. Wilshire Meta-Data Conference. 24-25 May 2005.
- Do, H. H., Rahm, E. (2000). On metadata interoperability in data warehouses. Technical Report 1- 2000, Institute Informatics, University Leipzig.
- Haag, S., Cummings, M., dan McCubbery, D.J (2002). Alain Pinsonneault, Richard Donvan: *Managements Information System for the Information Age*, Third Canadian Edition, McGraw-Hill Ryerson.

Halevy, A.Y, Ashish N., Bitton, D., Carey, M.J., Draper, D., Pollock, J., Rosenthal, A., dan Sikka, V. (2005). "Enterprise information integration: successes, challenges and controversies". *SIGMOD 2005*, 778-787.

Holzner, S., (1997). XML Complete. McGraw-Hill

Inmon, W.H., dan Hackathorn R.D. (2001). *Using the Data Warehouse*, John Wiley & Sons.

Jensen, M. R., Møller, T.H., dan Pedersen, T.B. (2001a). Specifying OLAP Cubes On XML Data. *Tech Report R-01-5003*, Department Of Computer Science, Aalborg University.

Jensen, M. R., dan T. H. Møller (2001b). Constructing OLAP Cubes From XML Data. *Tech Report R-02-5003*, Department Of Computer Science, Aalborg University.

Kimball, R. dan Ross, M.(2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition)*, John Wiley & Sons.

Koshafian, S. dan Abnous, R. (1995). Object Orientation – Concepts, Analysis, and Design, Languages, Databases, Graphical User Interfaces, Standard. 2nd ed. Wiley.

Lahiri, T et. al. Ozone (1999). Integrating Structured and Semi structured Data, *Proceedings of the Seventh International Conference on Database Programming Languages*, Kinloch Rannoch, Scotland.

Lenz, H., (1997). Summarizability in OLAP and Statistical Databases, *Proceedings of the Ninth International Conference on Statistical and Scientific Database Management*, 39-48.

- Lenzerini, M. (2002). "Data Integration: A Theoretical Perspective".
PODS 2002, 243-246.
- Mimno, P. (2002). Successful Real-Time Business Analytics: A Data Warehousing Strategy. White Paper. Informatica Corporation.
- Pedersen, et. al.(2000). Extending OLAP Querying To External Object Databases,
Proceedings of the Ninth International Conference on Information and Knowledge Management, ms. 405-413.
- Pledge, K. dan McGarry, J. (2001). Data Warehousing for Actuaries. Versi (1.2): 3-12.
- Poole, J., Chang, D., Tolbert, D., dan Mellor, D. (2003). Common Warehouse Metamodel Developer's Guide. John Wiley & Sons Inc.
- Poole, J. (2000). The Common Warehouse Metamodel as a Foundation for Active Object Models in the Data Warehouse Environment. Position paper to ECOOP 2000 workshop on Metadata and Active Object-Model Pattern Mining – Cannes, France.
- Pyle dan Dorian (2003) *Business Modeling and Data Mining*. Morgan Kaufmann,
- Rafanelli, M. 1990). STORM: A Statistical Object Representation Model, *Proceedings of the Fifth Conference on Statistical and Scientific Database Management*, 14-29.
- Shanmugasundaram, et. al (1999). Relational Databases for Querying XML Documents: Limitations and Opportunities, *Proceedings of the Twenty-Fifth International Conference on Very Large Databases*, Edinburgh, Scotland.
- Shukla, A. (1996).Storage Estimation for Multidimensional Aggregates in the Presence

of Hierarchies, *Proceedings of Very Large Databases*, pp. 522-531.

Silicon Integration Initiative (2000). *The Electronic Component Information Exchange QuickData Architecture*.

Thomsen, E., (1997). *OLAP Solutions: Building Multidimensional Information Systems*, John Wiley & Sons, Inc.

W3C (2001). World Wide Web Consortium, *The XML Query Algebra*, W3C Working Draft, <http://www.w3.org/TR/query-algebra>, Dec. 4 2000.

Ziegler, P. dan Dittrich, K.R (2004). "Three Decades of Data Integration - All Problems Solved?". *WCC 2004*, 3-12.