eksperimen menunjukkan masa memproses pertanyaan dalam tiga kerumitan pertanyaan yang berbeza dikurangkan sebanyak 7%, 5% dan 8% untuk set data SigmodRecord dan 5%, 9% dan 4% untuk set data DBLP. A model Native JSON (N-JSON) baru telah dibangunkan untuk integrasi data berdasarkan Model NXD. Dalam model ini, algoritma baru telah dibangunkan untuk mengekstrak elemen dan atribut dari sumber data yang berbeza, kemudian simpan ke dalam RDBMS. Sementara itu, nilai setiap atribut dan elemen akan disimpan ke dalam JSON format. Tiga eksperimen telah dilakukan untuk perbandingan prestasi antara NXD dan N-JSON dari segi masa pemyimpanan data,masa memproses pertanyaan dan peratusan penggunaan CPU menggunakan set data SigmodRecord dan DBLP. Hasil untuk tindak balas masa penyimpanan data dikurangkan sebanyak 7% menggunakan SigmodRecord set data, dan 15% menggunakan set data DBLP. Hasilnya untuk tindak balas masa pemprosesan pertanyaan dalam tiga kerumitan pertanyaan yang berbeza dikurangkan masing-masing sebanyak 7%, 8% dan 5% menggunakan set data SigmodRecord dan masing-masing 10%, 3% dan 7% menggunakan set data DBLP. Sementara itu, hasil untuk peratusan penggunaan CPU dalam tiga kerumitan pertanyaan yang berbeza juga berkurang masing-masing sebanyak 12%, 8% dan 6% menggunakan set data SigmodRecord dan 11%, 3% dan 6% menggunakan DBLP. Seterusnya, algoritma N-JSOND baru dibangunkan berdasarkan pengubahsuaian algoritma N-JSON dengan menyingkirkan aksara ' " ' untuk setiap elemen dari sumber data. Berdasarkan algoritma baru dalam model N-JSOND, hasilnya menunjukkan prestasi yang lebih baik berbanding dengan NXD. Tiga eksperimen dari segi tindak balas masa penyimpanan data, masa pemprosesan pertanyaan dan penggunaan peratusan CPU untuk perbandingan prestasi antara N-JSOND dan NXD dilakukan. Hasil tindak balas masa penyimpanan data dikurangkan sebanyak 11% menggunakan set data SigmodRecord, dan 22% menggunakan DBLP set data. Seterusnya, hasil tindak balas masa pemprosesan pertanyaan dalam tiga kerumitan pertanyaan yang berbeza dikurangkan masing-masing sebanyak 11%, 11% dan 9% menggunakan SigmodRecord set data dan 15%, 7%, 13% menggunakan set data DBLP. Sementara itu, hasil penggunaan CPU dalam tiga kerumitan pertanyaan yang berbeza juga masing-masing berkurang sebanyak 15%, 12% dan 10% menggunakan set data SigmodRecord dan 13%, 11% dan 11% menggunakan set data DBLP. Berdasarkan kajian ini, dapat disimpulkan bahawa N-JSOND adalah salah satu penyelesaian yang lebih baik dalam integrasi data untuk pembangunan aplikasi perniagaan pintar.

Abstract of thesis presented to the Senate of Universiti Malaysia Terengganu in fulfilment of the requirement for degree of Doctor Philosophy

DATA INTEGRATION MODEL FOR FASTER DATA EXTRACTION AND RETRIEVAL FROM SEMI-STRUCTURED DATA FORMAT

MOHD KAMIR YUSOF

MAY, 2021

Main Supervisor	**:**	**Assoc. Prof. Ts. Mustafa Man, Ph.D.**
Co-Supervisor	**:**	**Rozniza Ali, Ph.D**
Faculty	**:**	**Faculty of Ocean Engineering Technology and Informatics**

Collections of data is crucial across a wide variety of field because of increasing data rapidly year by year. These collections are important for many organizations to make a correct decision using business intelligent applications. A business intelligent application must have capability to collect and integrate all data from different data sources. One of the challenges in development of business intelligent application is data integration. This challenge is happened because of data structure are different. This research is looking for suitable data integration model in order to allows data integration from different data sources. Native XML (NXD) is one the model has been used in data integration. In this model, elements and attributes for each data are extract and store into Relational Database Management System (RDBMS). Meanwhile, the value for each element and attribute are stored in XML format. Based on the experiments has been done by previous researchers, NXD can produce a better performance during data insertion response time and query processing response time using SigmodRecord and DBLP datasets. However, the efficiency of NXD still has room for improvement. In the initial experiment, list of special character can be removed to improve the performance of NXD has been idenfitied in the SigmodRecord and DBLP datasets. Cleansing Algorithm (CA) algorithm has been developed to remove all these special characters to improve the query processing response time. The experiments results show query processing response time in three different queries complexity are reduced by 7%, 5% and 8% respectively for SigmodRecord dataset and

5%, 9% and 4% for DBLP dataset. A new Native JSON (N-JSON) model has been developed for data integration based on NXD model. In this model, new algorithm has been developed to extract the elements and attributes from different data sources, then store it into RDBMS. Meanwhile, the value of each attributes and elements will stored into JSON format. Three experiments have been done for performance comparison between NXD and N-JSON in term of insertion data response time, query processing response time and percentage of CPU usage using SigmodRecord and DBLP dataset. The results for data insertion response time is reduced by 7% using SigmodRecord dataset, and 15% using DBLP dataset. The result for query processing response time in three different queries complexity are reduced by 7%, 8% and 5% respectively using SigmodRecord dataset and 10%, 3% and 7% respectively using DBLP dataset. Meanwhile, the result for CPU in three different queries complexity also reduced by 12%, 8% and 6% respectively using SigmodRecord dataset and 11%, 3% and 6% respectively using DBLP. Next, a new N-JSOND algorithm is developed based on modification of N-JSON algorithm by removing character of ‘ “ ’ for each elements from data sources. Based on the new algorithm in N-JSOND model, the result shows a better performance compared to NXD. Three experiments in terms of data insertion response time, query processing response time and CPU usage for performance comparison between N-JSOND and NXD was done. The result of data insertion response time is reduced by 11% using SigmodRecord dataset, and 22% using DBLP dataset. Next, the result of query processing response time in three different queries complexity are reduced by 11%, 11% and 9% respectively using SigmodRecord dataset and 15%, 7%, 13% respectively using DBLP dataset. Meanwhile, result of CPU usage in three different queries complexity also reduced by 15%, 12% and 10% respectively using SigmodRecord dataset and 13%, 11% and 11% respectively using DBLP dataset. Based on this research, it can be concluded that N-JSOND is one of the better solutions in data integration for development of business intelligent applications.

Abstrak tesis ini dikemukakan kepada Senat Universiti Malaysia Terengganu sebagai memenuhi keperluan Ijazah Doktor Falsafah.

MODEL INTEGRASI DATA UNTUK EKSTRAK DAN CAPAIAN DATA DENGAN PANTAS BAGI FORMAT DATA SEPARA BERSTRUKTUR

MOHD KAMIR YUSOF

MEI, 2021

Penyelia Utama	**:**	**Prof. Madya. Ts. Mustafa Man, Ph.D.**
Penyelia Bersama	**:**	**Rozniza Ali, Ph.D**
Fakulti	**:**	**Fakulti Teknologi Kejuruteraan Kelautan dan Informatik**

Pengumpulan data sangat penting di pelbagai bidang kerana peningkatan data yang cepat dari tahun ke tahun. Pengumpulan data ini penting untuk sesebuah organisasi dalam membuat keputusan yang betul dengan memggunakan aplikasi perniagaan pintar. Aplikasi perniagaan pintar mesti mempunyai kemampuan untuk mengumpulkan dan mengintegrasikan semua data dari sumber data yang berbeza. Salah satu cabaran dalam pembangunan aplikasi pintar perniagaan adalah integrasi data. Cabaran ini berlaku kerana struktur data yang berbeza. Kajian ini sedang mencari model integrasi data yang sesuai untuk membolehkan integrasi data dari sumber data yang berbeza. Native XML (NXD) adalah salah satu model yang pernah ada digunakan dalam integrasi data. Dalam model ini, elemen dan atribut untuk setiap data akan di ekstrak dan simpan ke dalam Sistem Pengurusan Pangkalan Data (RDBMS). Sementara itu, nilai untuk setiap elemen dan atribut disimpan dalam format XML. Berdasarkan eksperimen yang telah dilakukan oleh penyelidik sebelumnya, NXD dapat menghasilkan prestasi yang baik dari segi masa penyimpanan data dan pemprosesan pertanyaan menggunakan set data SigmodRecord dan DBLP. Walau bagaimanapun, kecekapan NXD masih mempunyai ruang untuk penambahbaikan. Dalam eksperimen awal, senarai aksara yang boleh singkirkan untuk membantu meningkatkan prestasi NXD telah dikenal pasti di dalam set data SigmodRecord dan DBLP. Algoritma Pembersihan (CA) telah dibangun untuk menyingkirkan semua aksara ini untuk menambahbaik prestasi masa memproses pertanyaan. Hasil