Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych

Warszawa, 27 września 2016 r.

D z i e k a n a t

Uprzejmie informuję, że na Wydziale Elektroniki i Technik Informacyjnych Politechniki Warszawskiej odbędzie się w dniu 11 października 2016 r. publiczna obrona rozprawy doktorskiej

mgr inż. Stanisława Adaszewskiego

temat: „Virtualization of neuroimaging data access and processing for multisite population brain studies"

promotor – dr hab. inż. Piotr Bogorodzki, prof. Politechniki Warszawskiej

recenzenci:
prof. dr hab. inż. Stanisław Kozielski z Politechniki Śląskiej
prof.dr hab. inż. Krzysztof Goczyła z Politechniki Gdańskiej

Obrona odbędzie się w dniu 11 października 2016 r. w sali 116 na Wydziale Elektroniki i Technik Informacyjnych – Gmach im. Janusza Groszkowskiego, Warszawa, ul. Nowowiejska 15/19; początek godz. 11.00 .

Po adresem: **www.elka.pw.edu.pl/Wydzial/Rada-Wydzialu/Harmonogram-obron-doktorskich-streszczenia-i-recenzje** zapewniony jest na stronie Wydziału dostęp do tekstów streszczenia rozprawy i recenzji, jak również do tekstu rozprawy umieszczonej w Bazie Wiedzy Politechniki Warszawskiej.

Dziekan

prof. dr hab. inż. Krzysztof Zaremba

Rodzaj pracy: rozprawa doktorska

Autor - mgr inż. Stanisław Adaszewski

Promotor – dr hab. inż. Piotr Bogorodzki, prof. Politechniki Warszawskiej

Tytuł rozprawy:"Virtualization of neuroimaging data access and processing for multisite population brain studies"

Streszczenie

Metody diagnostyki obrazowej stwarzają unikalną możliwość identyfikacji i badania wybranychprocesów neuronalnych w żywym mózgu człowieka. Zastosowanie wieloośrodkowych badań obrazowych w połączeniu z nowoczesnymi technologiami informatycznymi, pozwalającymi na gromadzenie i eksploracją rozproszonych zbiorów danych, otwierają drogę dla badań populacyjnych o charakterze podstawowym jak i poszukiwania biomarkerów ekspresji zmian chorobowych.
Niniejsza praca dotyczy zagadnienia wirtualizacji danych w kontekście wieloośrodkowych badań obrazowych. Zawiera ona tezę o cechach niezbędnych do realizacji analiz tego typów zbiorów danych z użyciem systemów bazodanowych oraz przedstawia nową modelową implementację części takiej architektury nazwaną przez autora WEIRDB (od ang. słów Wide, External, Imaging, Research, DataBase). WEIRDB realizuje mechanizm dostępu do danych o dowolnym rozmiarze w ich oryginalnej postaci bez potrzeby wczytywania i synchronizacji z systemem bazodanowym, z pominięciem komunikacji strumieniowej klient/serwer oraz formułowanie zapytań w języku SQL.
Porównanie wydajności WEIRDB z istniejących rozwiązaniami pod kątem obsługi danych zewnętrznych oraz prędkości wykonywania zapytań wykazało wzrost wydajności od 2 do 200 razy w stosunku do pozostałych baz danych. W dalszym toku pracy omówione zostały dotychczasowe zastosowania praktyczne proponowanego oprogramowania, a powyższe spostrzeżenia zostały poddane dyskusji pod kątem wcześniejszych przewidywań oraz celów przyszłych badań i rozwoju.

Abstract
Diagnostic imaging methods create a unique possibility for identification and in-vivo studies of selected neuronal processes in human brain. Multisite imaging studies in collaboration with modern information technologies allow to acquire and explore distributed datasets, creating the opportunity for population brain studies as well as research of expression biomarkers in disease-modulated changes. This work concerns uses of Data Virtualization techniques in multisite neuroimaging studies and presents a new model implementation of such architecture called WEIRDB (acronym for: Wide, External, Imaging, Research, DataBase). WEIRDB implements a mechanism of arbitrary-sized data access in their source form without the necessity for import procedure or synchronization with the database system. Furthermore, it demonstrates means of improving performance by eliminating streaming client/server communication and allows for formulating queries in SQL. A comparison of WEIRDB's efficiency against existing solutions is presented with respect to external data handling capabilities and query execution speed showing from 2 to 200 times improvement over prior art.
Subsequently this work lists current practical applications of the proposed architecture followed by discussion of the above observations with emphasis on expectations, goals and future research.

Prof. dr hab. inż. Stanisław Kozielski
Instytut Informatyki
Politechniki Śląskiej

Gliwice, 17.08.2016

KWESTIONARIUSZ - RECENZJA ROZPRAWY DOKTORSKIE J DLA RADY WYDZIAŁU ELEKTRONIKI I TECHNIK INFORMACYJNYCH POLITECHNIKI WARSZAWSKIEJ

Tytuł rozprawy: Virtualization of neuroimaging data access and processing for multisite population brain studies

Autor rozprawy: Stanisław Adaszewski

1. Jakie zagadnienie naukowe jest rozpatrzone w pracy /teza rozprawy/ i czy zostało ono dostatecznie jasno sformułowane przez autora? Jaki charakter ma rozprawa (teoretyczny, doświadczalny, inny)?

Praca dotyczy problemów przetwarzania dużych zbiorów danych powstających w projektach badań ludzkiego mózgu, w szczególności neuroobrazowania, realizowanych w kilku ośrodkach badawczych na świecie. Autor zwrócił uwagę na ten obszar badawczy ze względu na wyróżniające cechy powstających tam danych: zbiory o bardzo dużych objętościach, dane o bardzo zróżnicowanych cechach (typach), rozmieszczenie danych w wielu ośrodkach.

Od lat 60-tych ubiegłego wieku dane przetwarzane w systemach komputerowych najczęściej gromadzone były (i są nadal) w bazach danych. Systemy zarządzania bazami danych (SZBD) są złożonymi systemami programowymi udostępniającymi bogate zestawy narzędzi operujących na danych i gwarantujące wysoki poziom bezpieczeństwa w trakcie przechowywania i przetwarzania danych. Jednakże dane powstające w badaniach neurologicznych istotnie różnią się (m.in. w zakresie typów oraz organizacji i rozmiarów zbiorów) od prostych danych numerycznych i tekstowych, dla których stworzono klasyczne bazy danych. Dlatego naturalnym krokiem autora była ocena możliwości wykorzystania baz danych do gromadzenia, a w pewnym zakresie również przetwarzania danych powstających w procesach badań ludzkiego mózgu, w szczególności badań obrazowych. Autor przeprowadził taką analizę i wskazał całą listę rozszerzeń i mechanizmów, o które należałoby uzupełnić systemy zarządzania bazami danych, aby mogły spełnić wymagania niezbędne w badaniach związanych z neuroobrazowaniem.

Kluczową z tych funkcji jest wirtualizacja danych, zapewniająca dostęp do danych bez konieczności rozwiązywania technicznych szczegółów dotyczących postaci danych oraz fizycznej lokalizacji danych. W szczególności, co bardzo ważne, mechanizm wirtualizacji chroni przed koniecznością tworzenia i przechowywania kopii danych.

Teza pracy została sformułowana następująco:

Udane zaadoptowanie bazy danych jako podstawowego narzędzia badawczego dla projektów neuroobrazowania wymaga, aby rozważane oprogramowanie wspierało: wykonywanie algorytmów według modelu MapReduce przy wykorzystaniu klastrowego środowiska obliczeniowego, równoważenie obciążenia, integrację z narzędziami programowymi do obliczeń naukowych (np. MATLAB), wysokiego poziomu języki zapytań (np. SQL), struktury danych zoptymalizowane dla operacji w pamięci

operacyjnej (*in-memory*) oraz, co najważniejsze, wirtualizację danych (określaną także jako dostęp do danych *in-situ* lub baza danych *ad-hoc*) z wydajnym buforowaniem i mechanizmami wyszukiwania.

Jako cele pracy autor wymienił m.in.:

- implementację modelowego pakietu oprogramowania bazy danych wspierającego trzy spośród powyższych cech: język SQL, struktury danych zoptymalizowane dla operacji w pamięci operacyjnej i wirtualizacja danych z wydajnym buforowaniem i mechanizmami wyszukiwania;
- przeprowadzenie testów takiego oprogramowania z użyciem benchmarków syntetycznych danych, jak i w praktycznych scenariuszach badań związanych z neuroobrazowaniem.

Teza rozprawy została sformułowana jasno, chociaż zbyt rozwlekle.

Rozprawa ma charakter w części teoretyczny: autor opracował metody i algorytmy dla systemu zarządzania bazą danych; z przewagą części implementacyjnej i eksperymentalnej: autor wykonał programową implementację opracowanego systemu, eksperymentalnie sprawdził jego funkcjonalność i wydajność, a także przedstawił jego praktyczne zastosowania.

2. Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł / w tym literatury światowej, stanu wiedzy i zastosowań w przemyśle /świadczący o dostatecznej wiedzy autora. Czy wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonywujący?

Tematyka rozprawy mocno wiąże się z bazami danych, więc autor musiał się zmierzyć z niezwykle bogatą literaturą dotyczącą tej tematyki. Autor skupił się na tych problemach, które zasygnalizował w tezie pracy, a więc obecności (a raczej nieobecności) w systemach zarządzania bazami danych mechanizmów, które są niezbędne z uwagi na specyfikę rozważanych danych dotyczących neuroobrazowania. Stan wiedzy w podejmowanej dziedzinie badań i analiza źródeł przedstawione zostały w rozdziałach drugim i trzecim.

Na wstępie rozdziału drugiego autor przedstawił szersze spojrzenie na ten problem, a mianowicie, cytując literaturę, wskazał przeszkody utrudniające zastosowanie baz danych do przetwarzania danych pochodzących z badań naukowych. Są to: rozmiary danych, przekraczające możliwości istniejących SZBD; charakter dostępu do danych, odmienny znacząco od przetwarzania transakcyjnego; brak wygodnych interfejsów do środowisk programowania używanych w obliczeniach naukowych; ograniczenia w odwzorowaniu między typami danych wykorzystywanymi w obliczeniach naukowych a typami wspieranymi przez SZBD. Podano liczne przykłady badań naukowych, w których powstają dane o powyższych cechach. Wymienione kłopoty spowodowały powstanie pewnych rozwiązań (formaty plików, indeksy) umożliwiających zbieranie danych poza bazami danych (BIDS, OpenFMRI). Autor wskazał jednak całą listę zalet zastosowania baz danych do obróbki danych pochodzących z badań naukowych, w szczególności z neuroobrazowania. Należą do nich m.in.: możliwość zdalnego dostępu, filtrowanie i agregacja danych po stronie serwera, pojedynczy punkt aktualizacji, równoważenia obciążenia, współdzielenie przez wielu użytkowników programów do przetwarzania i analizy danych.

Zestawiając wady i zalety autor wskazał na mechanizm wirtualizacji danych jako rozwiązanie godzące rozbieżne podejścia. Zapewnia ono zachowanie plikowego dostępu do danych dla oprogramowania analitycznego wymagającego intensywnych operacji wejścia/wyjścia, przy wykorzystaniu wysokiego poziomu języków zapytań i innych zalet baz danych.

W przyjętym scenariuszu wirtualizacji dane są przechowywane w ich oryginalnym formacie i dostępne za pośrednictwem oprogramowania pośredniczącego odpowiedzialnego za przedstawienie ich organizacji w formie typowej tabeli bazy danych. Autor wskazał narzędzia programowe realizujące taką wizję wirtualizacji. Następnie przedstawił rozwój mechanizmów wirtualizacji danych w ramach standardu SQL o nazwie *Management of External Data* (SQL/MED). Omówione zostały dwie metody dostępu w SQL-owym systemie zarządzania bazą danych do danych zewnętrznych (plików spoza tej bazy). Jako jeden z popularnych formatów takich plików został wskazany format CSV (*comma separated values*), w którym plik składa się z linii danych oddzielonych od siebie przecinkami. Autor podkreślił jednak, że implementacja omówionych mechanizmów podlega wielu ograniczeniom dotyczącym typów danych oraz formatów i rozmiarów plików.

Następnie (w rozdziale trzecim) autor dokonał przeglądu istniejących systemów i narzędzi baz danych pod kątem obecności w nich mechanizmów przewidzianych standardem SQL/MED, zapewniających dostęp do zewnętrznych danych. Analiza ta objęła systemy baz danych typu *open source*: PostgreSQL, MySQL, Teiid (system wirtualizacji danych), HSQLDB, H2, SQLite, MonetDB oraz nowo powstające systemy NoSQL, w tym: CouchDB, Cassandra, Neo4j, Redis. Podsumowując autor stwierdził, że żaden z przedstawionych systemów NoSQL nie oferuje mechanizmów do zarządzania zewnętrznymi danymi, natomiast wszystkie badane systemy relacyjnych baz danych, udostępniając pewne mechanizmy tej klasy, wykazują liczne problemy z realizacją dostępu do zewnętrznych danych o specyfice charakterystycznej dla neuroobrazowania.

Jako główny powód tych problemów autor widzi m.in. preferowanie, już na etapie projektowania tych systemów baz danych, dedykowanych formatów plików, a jedną z przyczyn ograniczających wydajność badanych systemów relacyjnych jest ich słabe przygotowanie do wykorzystania coraz większych pamięci operacyjnych do przetwarzania całych plików danych.

Przeglądową część pracy pozwoliła na sformułowanie następującego punku widzenia autora. Dane dotyczące neuroobrazowania gromadzone są w plikach o specyficznej organizacji, tworzonych poza bazami danych. Specyfika tych danych utrudnia lub uniemożliwia ich przenoszenie, dla potrzeb analizy, do innych lokalizacji, czyli powinny być one przetwarzane z dostępem źródłowych plików danych. Zalety baz danych sugerują ich wykorzystanie jako ramy systemów przetwarzania rozważanych danych. Bazy danych posiadają wprawdzie pewne mechanizmy dostępu do „obcych" (zewnętrznych) plików danych, ale mechanizmy te nie są wystarczające dla danych dotyczących neuroobrazowania. Dlatego należy rozwinąć takie funkcje baz danych.

Takie ujęcie problemu było podstawą zaprojektowania w ramach pracy systemu umożliwiającego efektywny dostęp do danych i przetwarzanie danych dotyczących neuroobrazowania.

Podsumowując stwierdzam, że autor wykazał się dobrą znajomością literatury światowej i stanu wiedzy w zakresie przetwarzania danych dotyczących projektów neuroobrazowania oraz możliwości wykorzystania do tych celów systemów baz danych (relacyjnych i NoSQL).

3. Czy autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione?

Wychodząc z jednej strony z analizy specyfiki danych dotyczących neuroobrazowania, a z drugiej strony z oceny możliwości systemów baz danych, autor opracował koncepcję systemu programowego o cechach systemu zarządzania bazami danych, o funkcjach niezbędnych do dostępu do danych neuroobrazowania i ich przetwarzania. Jako dowód

realizowalności tej koncepcji autor zbudował system o nazwie WEIRDB (*Wide External Imaging Research DataBase*).

Implementacja architektury WEIRDB bazuje na języku C++ i kilku bibliotekach tego języka, m.in. Boost i Qt, wykorzystano także bazę danych SQLite. Wybór tej bazy był związany m.in. z jej bogatymi możliwościami indeksowania.

Jedną z najważniejszych możliwości systemu WEIRDB jest obsługa plików zewnętrznych (w stosunku do lokalnej bazy danych), stanowiąca główną funkcję procesu wirtualizacji danych. Dla operowania na takich plikach z wykorzystaniem języka SQL autor opracował silnik przepisywania zapytań, odpowiedzialny za budowę optymalnego planu zapytania. Zadaniem silnika i celem procesu przepisywania jest restrukturyzacja zapytania SELECT, prowadząca do wyodrębnienia elementów, które mogą być wykorzystywane do pobierania danych z plików zewnętrznych. Proces ten jest realizowany przez specjalny, opracowany przez autora, trzyetapowy algorytm przepisywania zapytań.

Dla przyspieszenia dostępu do plików zewnętrznych autor wykorzystał mechanizm odwzorowania plików w pamięci operacyjnej, zaimplementowany jako klasa biblioteki Qt o nazwie QAbstractItemModel. Rozwiązanie to pozwala wykorzystać zazwyczaj duże pamięci operacyjne do istotnego zwiększenia wydajności przetwarzania danych zewnętrznych.

Dla umożliwienia zdalnego korzystania z projektowanego systemu autor opracował dwa sieciowe protokoły dostępu do systemu WEIRDB: prosty, tekstowy protokół natywny oraz protokół kompatybilny z protokołem sieciowym PostgreSQL. Pierwszy z nich udostępnia format komunikatów, który może być stosowany również dla zbiorów danych o skali *big data* (m.in. dla tabel o nieograniczonej liczbie kolumn), natomiast drugi umożliwia dostęp do WEIRDB za pośrednictwem m.in.: ODBC, JDBC, a także API systemu PostgreSQL, co zapewnia kompatybilność z istniejącym oprogramowaniem bazodanowym.

Poprawność i użyteczność opracowanej koncepcji autor zweryfikował w eksperymentalnej części rozprawy. Do badań wykorzystano 7 zestawów danych pochodzących z projektów neurologicznych. Ocenie podlegał system WEIRDB oraz 7 systemów zarządzania bazami danych, opartych na języku SQL, omówionych w rozdziale trzecim. Badania skupiały się na dwóch aspektach: analizie możliwości wykonania zapytań przez poszczególne systemy oraz porównaniu wydajności systemu WEIRDB z rozważanymi bazami danych. Uzyskane wyniki pokazały, że parametry danych pochodzących z projektów neurologicznych przekraczają możliwości wielu dostępnych na rynku systemów baz danych, w przypadku dwóch zestawów danych jedynie system WEIRDB mógł zrealizować postawione zadania kontrolne. W ocenie wydajności system WEIRDB wyprzedzał inne bazy danych, tylko dwa razy WEIRDB był gorszy: w jednym przypadku od systemu MySQL, a w innym od PostgreSQL.

Bardzo ważne dla pełniejszej oceny możliwości systemu WEIRDB są przedstawione w rozdziale szóstym trzy przykłady praktycznego zastosowania tego systemu do analizy danych w badaniach neurologicznych dotyczących chorób Parkinsona, Alzheimera i innych. Celem wykorzystania WEIRDB w tych przykładach było udostępnienie narzędzia do analizy danych dotyczących neuroobrazowania, pomiarów laboratoryjnych i danych genetycznych. Typowym zadaniem WEIRDB było scalanie i wstępne przetwarzanie danych z różnych źródeł, charakterystyczne były przy tym specyfika formatów danych oraz bardzo duże objętości danych, co narzucało konieczność wirtualizacji danych i uniemożliwiało wykorzystanie klasycznych baz danych.

Przedstawione przykłady pokazały praktyczną użyteczność i zakres możliwości systemu WEIRDB.

Podsumowując stwierdzam, że autor rozwiązał postawione zadanie: zaprojektował system WEIRDB o unikalnych cechach, opracował metody i algorytmy dla wybranych funkcji tego systemu, przeprowadził programową implementację i eksperymentalną ocenę, a także przedstawił praktyczne zastosowania systemu WEIRDB.

4. Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy czy poziomu techniki reprezentowanych przez literaturę światową?

Oryginalnym rezultatem rozprawy jest opracowanie koncepcji środowiska programowego o funkcjach systemu zarządzania bazą danych, umożliwiającego efektywny dostęp i przetwarzanie zbiorów danych dotyczących neuroobrazowania. Zaproponowane rozwiązanie wypełnia lukę jaka powstała między możliwościami istniejących systemów baz danych (typu *open source*) a potrzebami dotyczącymi przetwarzania zbiorów danych powstających w projektach z zakresu neurologii, w tym neuroobrazowania. W szczególności do oryginalnych wyników uzyskanych przez autora zaliczam:

- Opracowanie silnika przepisywania zapytań umożliwiającego dostęp z poziomu języka SQL do zewnętrznych (w stosunku do bazy danych) plików danych.
- Opracowanie mechanizmów wykorzystania pamięci operacyjnej do zwiększenia wydajności dostępu do zewnętrznych plików danych.
- Opracowanie dwóch sieciowych protokołów: prostego ale bardzo wydajnego protokołu natywnego oraz protokołu zapewniającego kompatybilność z protokołem sieciowym PostgreSQL.

Bardzo ważną jest implementacyjna i eksperymentalna część rozprawy, która umożliwiła wykazanie realizowalności przedstawionej koncepcji, pozwoliła na ocenę proponowanych metod i wykazała przydatność systemu w udanych praktycznych zastosowaniach.

Uzyskanie powyższych wyników pozwala stwierdzić poprawność tez postawionych w pracy i osiągnięcie celu pracy.

5. Czy autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników /zwięzłość, jasność, poprawność redakcyjna rozprawy/?

Sposób prezentacji podejmowanych w pracy zagadnień jest w miarę jasny (niektóre wyjątki zaznaczyłem w p. 6 recenzji). Strona redakcyjna samego tekstu rozprawy nie budzi większych zastrzeżeń, drobne uwagi zamieściłem w punkcie 6 recenzji.

6. Jakie są słabe strony rozprawy i jej główne wady?

Lektura pracy nasunęła mi kilka uwag, częściowo o charakterze redakcyjnym:

1) We wprowadzeniu do pracy i w tezie pracy autor wymienia, jako jedno w wymagań dla projektowanego oprogramowania, możliwość realizacji algorytmów według modelu MapReduce – nie doszukałem się jednak szerszego omówienia tego problemu dla opracowanego systemu WEIRDB.
2) W rozdz. 5.3 analiza wyników eksperymentów jest przedstawiona mało przejrzyście – zdania są zbyt długie, przeładowane danymi liczbowymi i odwołaniami do tabel z wynikami.

3) Do której z baz danych odnosi się zestawienie struktur indeksów zawarte w tabeli 4.1.1 na str. 64?
4) Pewną wątpliwość budzi przyjęty sposób opisu treści części rysunków – cały tekst opisu zamieszczony został bowiem w podpisie pod rysunkiem. Z powodu małego rozmiaru czcionki opisy te są trudno czytelne.
5) Niektóre tabele a także rysunki są trudno czytelne, np. tabela 6.1 (str. 100), rys. 6.1 (str. 97).

Przedstawione uwagi nie obniżają mojej pozytywnej oceny pracy.

7. Jaka jest przydatność rozprawy dla nauk technicznych?

Wyniki rozprawy mogą mieć ważne zastosowania w systemach przetwarzania danych. Autor wykazał przydatność zbudowanego w ramach pracy systemu WEIRDB w trzech projektach przetwarzania zbiorów danych dotyczących neuroobrazowania. Te rezultaty dokumentują możliwości bezpośrednich zastosowań systemu WEIRDB. Natomiast metody i rozwiązania przedstawione w pracy mogą mieć szersze zastosowania w systemach przetwarzania danych, w których pojawiają się zbiory danych o formatach nietypowych dla klasycznych baz danych i rozmiarach przekraczających ograniczenia takich baz. Z takimi przypadkami często nie radzą sobie dostępne na rynku systemy baz danych, natomiast alternatywną propozycją mogą być opracowane w rozprawie metody wirtualizacji danych.

Podsumowując, wyniki rozprawy mają duży potencjał aplikacyjny.

8. Do której z następujących kategorii Recenzent zalicza rozprawę:

a/ nie spełniająca wymagań stawianych rozprawom doktorskim przez obowiązujące przepisy

b/ wymagająca wprowadzenia poprawek i ponownego recenzowania

c/ spełniająca wymagania

d/ spełniająca wymagania z wyraźnym nadmiarem

e/ wybitnie dobra, zasługująca na wyróżnienie

Zaliczam rozprawę do kategorii c/ - **rozprawa spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy.**

Podpis

Gdańsk, 2.02.2016

prof. dr hab. inż. Krzysztof Goczyła
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska
ul. G. Narutowicza 11/12
80-233 Gdańsk
kris@eti.pg.gda.pl

R E C E N Z J A

rozprawy doktorskiej mgr. inż. Stanisława Adaszewskiego pt.
„Virtualization of neuroimaging data access and processing for multisite population brain studies”
przygotowanej pod kierunkiem prof. nzw. dr. hab. inż. Piotra Bogorodzkiego
(Politechnika Warszawska, Wydział Elektroniki i Technik Informacyjnych)

1. Problem naukowy

Recenzowana rozprawa doktorska dotyczy ważnego z punktu widzenia współczesnych technologii informatycznych problemu, jakim jest przechowywanie i zarządzanie dużymi danymi medycznymi. Wraz z rozwojem metod diagnostycznych, technologie informatyczne wkraczają od pewnego czasu do medycyny, w tym do badań nad mózgiem człowieka, pomagając lekarzom w podejmowaniu właściwych diagnoz, a także wspierając naukowców w badaniach nad procesami zachodzącymi w mózgu. Do tego potrzebne są efektywne metody przechowywania, wydobywania, eksplorowania i prezentacji dużych zbiorów danych, zazwyczaj o charakterze obrazów 2- i 3-wymiarowych o bardzo dużej rozdzielczości.

Autor w swojej pracy podjął się rozwiązania tego problemu dla danych pochodzących z wieloośrodkowych badań obrazowych mózgu. Dane te charakteryzują się znaczną wielkością pojedynczych obrazów oraz bardzo dużą liczebnością tychże. Wymagane jest też, by systemy zarządzające takimi danymi umożliwiały łatwy dostęp i obróbkę danych. Tutaj z pomocą Autorowi przychodzi popularna ostatnio i często stosowana idea wirtualizacji. W ogólności, polega ona na ukrywaniu przez użytkownikiem (człowiekiem, ale też i systemem informatycznym) fizycznej struktury realizującej pewne usługi w sposób umożliwiający łatwe korzystanie z tych usług za pośrednictwem standardowych interfejsów. W kontekście paradygmatów przetwarzania, wirtualizacja stosowana jest w przetwarzaniu w chmurze (ang. *cloud computing*). W kontekście zarządzania danymi (a tego właśnie kontekstu dotyczy recenzowana rozprawa) wirtualizacja polega na ukrywaniu przez użytkownikiem lub aplikacją struktur danych i sposobów ich przechowywania oraz na udostępnieniu tych danych poprzez standardowe języki zapytań. W ten sposób można na przykład zarządzać plikami rezydującymi poza relacyjną bazą danych tak, jakby były zarządzane bezpośrednio przez RDBMS.

Reasumując — problem podjęty w rozprawie jest ważny i aktualny dla nowoczesnych zastosowań informatyki, w szczególności w medycynie, a podjęte zadanie badawcze uważam za bardzo interesujące i ambitne z punktu widzenia informatycznego.

2. Teza rozprawy

We wprowadzeniu do rozprawy Autor w sposób przekonujący i dogłębny motywuje podjęcie badań, przedstawiając ich cele na tle aktualnego stanu wiedzy i technologii. Następnie formułuje dość rozbudowaną tezę rozprawy. Streścić ją można następująco (w j. polskim):

> „Zastosowanie baz danych w badaniach dotyczących obrazowania procesów neuronalnych wymaga użycia takich technik i metod przetwarzania, jak wirtualizacja danych, algorytmy typu MapReduce, fragmentacja, równoważenie obciążenia, optymalizacja danych dla przetwarzania w pamięci operacyjnej, a także wymaga integracji ze standardowymi językami dostępu (np. SQL) i aplikacjami wspomagającymi obliczenia naukowe (np. MATLAB)".

Do tej tezy można mieć dwa zastrzeżenia. Po pierwsze, występującą w niej „wyliczankę" metod i technik można by skrócić, ograniczając ją do metod przetwarzania stosowanych w bazach typu NoSQL wspomaganych przez klasyczne techniki bazodanowe i aplikacje do obliczeń naukowych. Uczyniłoby to tezę rozprawy bardziej zwartą, co – jak wiadomo – sprzyja precyzji rozpraw naukowych. Po drugie, wydaje się, że Autor pomylił tu warunek konieczny z warunkiem wystarczającym. Przecież nie jest wykluczone, że korzystne rezultaty zastosowania baz danych w obrazowaniu procesów neuronalnych można też uzyskać innymi technikami niż te wymienione w tezie. W istocie, tezę Autora należałoby odwrócić: użycie wymienionych w tezie technik i metod przetwarzania **wystarcza** do tego, by technologie bazodanowe mogły być z powodzeniem zastosowane do badań związanych z obrazowaniem procesów neuronalnych.

Powyższe zarzuty nie mają charakteru deprecjonującego sformułowanie postawionego przez Autora, problemu, gdyż intencje Autora wydają się oczywiste, mimo że wyrażone nieprecyzyjnie.

3. Omówienie rozprawy i oryginalnego dorobku Autora

Rozprawa napisana jest w dobrym języku angielskim (w całej rozprawie natrafiłem na zaledwie kilka błędów o charakterze językowym). Składa się z wprowadzenia, 6 zasadniczych rozdziałów i podsumowania. Rozprawę uzupełniają: użyteczny słownik pojęć (choć umieszczenie w nim objaśnienia terminów GB i KB nie było potrzebne), obszerna bibliografia obejmująca 130 pozycji uporządkowanych według kolejności odwołań w tekście (co zresztą utrudniło mi odszukanie publikacji Autora związanych z tematyką rozprawy) oraz załącznik stanowiący skrótowy raport z wykonanej przez Autora, skądinąd bardzo wnikliwej i niewątpliwie pracochłonnej, analizy architektury popularnych systemów baz danych opisanych w rozdziale 3.

Układ rozprawy jest logiczny i przejrzysty. We **wprowadzeniu** do rozprawy (opatrzonym numerem 1) Autor uzasadnia podjęcie tematyki, formułuje cele i tezę rozprawy oraz prezentuje jej strukturę. Swoje zastrzeżenia do tezy rozprawy przedstawiłem powyżej, w punkcie 2 recenzji. Pozostała zawartość wprowadzenia nie budzi zastrzeżeń, a wręcz przekonuje czytelnika, że warto kontynuować studiowanie rozprawy.

W **rozdziale 2** Autor przedstawia podjęty problem badawczy bardziej szczegółowo, pokazując, na czym polega jego specyfika w stosunku do klasycznych problemów przechowywania i zarządzania zbiorami danych. Dokonuje też wyczerpującego przeglądu

istniejących tego typu rozwiązań w dziedzinie danych pochodzących z obrazowania procesów neuronalnych. Wyniki tego przeglądu zostały zestawione tabelarycznie. Już w tym miejscu Autor zaznacza, że istniejące rozwiązania są w praktyce niewystarczające zarówno pod względem funkcjonalnym, jak i wydajnościowym (w tym drugim aspekcie – odsyłając do rozdziału 5).

W tym rozdziale po raz pierwszy uwidacznia się pewna maniera redakcyjna, która w pewnym stopniu przeszkadzała mi w studiowaniu materiału. Otóż Autor przyjął zasadę szczegółowego opisywania rysunków w ich podpisach. Nie jest to dobra zasada. Po pierwsze, opisy stają się przez to mało czytelne z uwagi na to, że zastosowana w nich czcionka jest bardzo mała, szczególnie zważywszy na to, że praca pisana jest w formacie „książkowym". Po drugie, takie długie podpisy zabierają miejsce na stronie samemu rysunkowi. W efekcie teksty wewnątrz rysunków stają się mało czytelne. Dotyczy to również niektórych tabel (ale i rysunków), dla których układ poziomy, a nie pionowy byłby odpowiedniejszy ze względów na ich czytelność.

W **rozdziale 3** Autor dokonuje szczegółowego przeglądu istniejących systemów baz danych pod kątem możliwości obsługi danych obcych (zewnętrznych), tj. danych zapisanych w innych formatach niż naturalnych dla tych systemów. Jest to o tyle istotne, że dane z obrazowania neuronalnego mają, ze swojej natury, właśnie taki charakter. Autor ogranicza się do systemów typu *open source*, dostępnych na zasadach wolnych licencji, co jest sensowną decyzją, gdyż systemy komercyjne w aspekcie przetwarzania danych zewnętrznych nie wnoszą nic istotnego w stosunku do systemów niekomercyjnych. Analizie poddane zostały popularne systemy relacyjne i systemy typu NoSQL. Zaprezentowana analiza sprawia wrażenie systematycznej i dogłębnej, choć miejscami zbyt technicznej. Cenne jest to, że Autor dokonuje podsumowania dla każdego analizowanego systemu, w którym pokazuje jego ograniczenia w zastosowaniach objętych tematyką rozprawy.

Uwagi do tego rozdziału:

- W opisie systemu Teiid funkcja operująca na plikach CSV ma dwie różne nazwy: TEXTABLE i TEXTTABLE (str. 37).
- W paru miejscach Autor powołuje się na opublikowane testy wzorcowe (ang. *benchmarks*), nie podając stosownych odsyłaczy do źródeł (np. w analizie możliwości systemu HSQLDB na str. 40).
- W opisie popularnego systemu SQLite drugi akapit na str. 42 jest niedokończony lub zbędny.

Rozdziały 4, 5 i 6 składają się na zasadniczą, oryginalną część rozprawy. W **rozdziale 4** Autor prezentuje autorski system o nazwie WEIRDB (*Wide External Imaging Research DataBase*), który przeznaczony jest do efektywnego przechowywania i przetwarzania danych neuronalnych, także o charakterze strumieniowym. System ten integruje szereg gotowych rozwiązań bazodanowych z rozwiązaniami autorskimi, w tym z oryginalnymi algorytmami przeznaczonymi do przeformułowywania zapytań SQL (ang. *query rewriting*) dla obsługi danych zewnętrznych, z opracowanymi metodami wyrywkowego dostępu do danych i ze specjalnymi protokołami sieciowymi. W prezentowanym systemie nacisk położono na zwirtualizowaną obsługę danych zewnętrznych w formatach specyficznych dla wielośrodkowych badań obrazowych mózgu. Rozdział ten dowodzi, że Autor rozprawy bardzo dobrze potrafi integrować istniejące rozwiązania i narzędzia, wzbogacając je o własne rozwiązania algorytmiczne i architektoniczne. Jednak rozdział 4 miałby charakter wyłącznie akademicki, gdyby nie wyniki praktyczne przedstawione w dwóch następnych rozdziałach rozprawy.

W **rozdziale 5** Autor prezentuje wyniki eksperymentów przeprowadzonych na 7 rzeczywistych (a nie eksperymentalnych!) zestawach danych. Eksperymenty te zostały przeprowadzone w ten sposób, by ich wyniki można było porównać z istniejącymi, znanymi rozwiązaniami. Eksperymenty te dowodzą, że w większości przypadków zaprojektowana przez Autora architektura i jej realizacja pod postacią systemu WEIRDB znacząco przewyższa pod względem efektywnościowym rozwiązania oparte na dostępnych systemach. Pewne wątpliwości budzi jedynie postać zapytań nr 5 i 6, w których w warunku selekcji zastosowano operator OR (takie formułowanie warunków selekcji nie jest zalecane z uwagi na brak możliwości optymalizacji, np. z wykorzystaniem indeksów), i to w dodatku do wydobywania – nie wiadomo dlaczego – danych o identyfikatorach parzystych.

W **rozdziale 6** przedstawiono zastosowanie systemu WEIRDB w trzech projektach. Jest to szczególnie wartościowy element rozprawy, gdyż pokazuje on, że zaproponowane przez Autora rozwiązanie zostało gruntownie zwalidowane, co w pracach o charakterze praktycznym i doświadczalnym, a taki jest charakter recenzowanej rozprawy, jest szczególnie istotne. Niestety, i w tym rozdziale uwidoczniła się maniera Autora do prezentowania mało czytelnych rysunków z nazbyt obszernymi podpisami (patrz rys. 6.1 i 6.3.1).

W **rozdziale 7**, stanowiącym podsumowanie rozprawy, Autor rekapituluje swoje osiągnięcia, ale też – co jest bardzo istotne – wskazuje możliwości rozwoju i doskonalenia swojego podejścia do podjętego w rozprawie problemu naukowego.

Za **oryginalny dorobek naukowy** Autora, przedstawiony przede wszystkim w rozdziałach 4, 5. i 6 rozprawy, uważam:

- opracowanie i zaimplementowanie architektury systemu bazodanowego przeznaczonego do efektywnego przechowywania i przetwarzania dużych danych pochodzących z wieloośrodkowych badań obrazowych mózgu;
- poddanie weryfikacji i walidacji zaproponowanego rozwiązania drogą eksperymentów przeprowadzonych na rzeczywistych danych i w rzeczywistych projektach;
- dokonanie gruntownego, krytycznego przeglądu dostępnych systemów baz danych pod kątem możliwości ich bezpośredniego zastosowania w zarządzaniu dużymi danymi neuronalnymi.

Dorobek ten uważam za bardzo wartościowy dla rozwoju metod zarządzania danymi medycznymi, niekoniecznie tylko z zakresu badań mózgu. Uważam, że ten dorobek, po pewnych rozwinięciach, również tych nakreślonych przez Autora w podsumowaniu rozprawy, może znaleźć praktyczne zastosowanie w informatyce medycznej – dynamicznie rozwijającej się gałęzi informatyki stosowanej.

4. Uwagi krytyczne i polemiczne

Niektóre ważniejsze uwagi krytyczne, także natury redakcyjnej, sformułowałem już w punktach 2 i 3 recenzji. Poniżej zamieszczam pewne dodatkowe uwagi natury ogólnej, mające w istocie charakter polemiczny.

1) Praca ma charakter wyłącznie praktyczny i doświadczalny. Autor nie pokusił się o sformułowanie jakiegoś podłoża teoretycznego dla postawionego problemu. A szkoda, gdyż mogłoby to być ciekawym przyczynkiem do zagadnień integracji różnych modeli

danych, w tym modeli danych o charakterze strumieniowym (np. prostego modelu danych strumieniowych zaprezentowanego w „Mining of Massive Datasets", Chapt. 4 „Mining Data Streams"; J. Leskovec, A. Rajaraman, J.D. Ullman; Cambridge Univ. Press, 2014), które mają aktualnie bardzo duże osadzenie w środowiskach typu „Big Data". Ten aspekt jest słabością recenzowanej rozprawy.

2) W rozdziale 3 warto było odnieść się nie tylko do baz danych, ale także do hurtowni danych i znanych narzędzi typu *business intelligence*. Nie sądzę, aby wśród tej gałęzi systemów zarządzania danymi znalazły się jakieś rozwiązania gotowe do zastosowania w zarządzaniu dużymi danymi zewnętrznymi, ale taka analiza byłaby ciekawa.
3) Czy Autor rozważał zastosowanie w swojej architekturze środowiska Hadoop, np. z wykorzystaniem popularnej wolnej platformy Claudera? Z pewnością zagwarantowałoby to skalowalność poziomą proponowanego rozwiązania, co może być istotne w miarę postępu badań mózgu człowieka i związanego z tym wzrostu wielkości danych z obrazowania procesów neuronalnych.

5. Podsumowanie

W treści swojej rozprawy Autor wykazał się bardzo dobrym rozeznaniem w zagadnieniach przechowywania i przetwarzania dużych zbiorów danych medycznych. W oryginalny z punktu widzenia inżynierskiego sposób wykorzystał istniejący stan technologii informatycznej do opracowania własnej architektury adekwatnej do rozwiązania problemów z tym związanych. Zaproponowana architektura została z powodzeniem zaimplementowana i poddana walidacji w rzeczywistych sytuacjach na rzeczywistych danych. Ponadto Autor wykazał się dobrą znajomością współczesnej literatury z zagadnień związanych z tematyką rozprawy. Rozprawa jako całość świadczy o bardzo dobrym przygotowaniu Autora do dalszej pracy naukowej na polu informatyki w szeroko rozumianym zakresie baz danych i przetwarzania danych.

Konkludując, stwierdzam, że rozprawa spełnia wymagania stawiane rozprawom doktorskim przez stosowne przepisy i wnoszę o dopuszczenie jej do publicznej obrony.

Rozprawę zaliczam do kategorii: **spełniająca wymagania**.