Институт по информационни и комуникационни технологии-БАН
Вх. № 167 / 28.02.2018 г.

РЕЦЕНЗИЯ

от проф. дмн Галя Младенова Ангелова, Секция ЛМОЗ, ИИКТ-БАН
за дисертацията на Йенс Колер

„*Optimizing Query Strategies in Fixed Vertical Partitioned and Distributed Databases and their Application in Semantic Web Databases*“

Оптимизиране на стратегиите на заявки в бази данни с фиксирано вертикално разделяне и в дистрибутирани база данни и тяхното приложение в бази данни на семантичния интернет)

представена за присъждане на образователната и научна степен „доктор“

Представената дисертация в професионално направление 4.6 „Информатика и компютърни науки“ е свързана с някои от най-актуалните проблеми в информатиката днес: постигане на сигурност и защита при достъпа до големи данни, които се съхраняват извън непосредствения контрол на потребителя (например в облак) и до които трябва да се осигури достъп за различни потребители чрез различни заявки, практически формиращи различни изгледи (views) към данните.

Дисертационният труд, написан на английски език, съдържа 212 страници. Основното тяло на труда е организирано в увод, 8 глави, заключение с изборяване на приносите, декларация за оригиналност, списък на използваната литература и списък на съкращенията. Текстът е богато илюстриран с таблици и фигури. Четирите приложения, разположени на последните 10 страници, съдържат списъци на таблиците, фигурите, както и на алгоритмите и фрагментите код в текста, а приложение 4 показва важни екрани на средата *SeDiCo*, използвана за тестване на предложените в дисертацията алгоритми.

Съгласно Правилника за специфичните условия за придобиване на научни степени и за заемане на академични длъжности в Института по информационни и комуникационни технологии (ИИКТ) при Българската академия на науките (БАН), кандидатът за получаване на образователната и научна степен "доктор" трябва да има "*поне 3 научни публикации, поне една от които да е в списание с импакт фактор или в специализирано международно издание*". Резултатите от дисертацията са представени в 16 публикации – една самостоятелна и другите в съавторство, като:

- 1 публикация е в международното списание „International Journal of Adaptive, Resilient and Autonomic Systems“,
- 10 са в Сборници трудове на престижни международни конференции (издадени от IEEE, Шпрингер и др.),
- 5 са в Сборници трудове на конференции, посветени на разработки по информатика в немски професионални училища.

Забелязани са 2 цитирания на трудовете на автора (в магистърска теза, представена в Университета на Окланд, Австралия и в една публикация от международна конференция).

Считам, че от формална гледна точка изискванията на ЗРАСРБ и ИИКТ за получаване на образователната и научна степен "доктор" са удовлетворени и надхвърлени.

Съдържание на труда

В Увода дистертантът формулира темата на изследването: да се покаже, че фиксираните вертикални разделяния и разпределяне на схемата осигуряват време за отговор на заявките, сравнимо с традиционната организация на релационните бази данни, с използване на облачни инфраструктури и съвременен хардуер. Изследванията на оптималността са фокусирани върху времето за отговор на заявки. Уточнява се, че не са известни предишни подходи за постигане на защита и сигурност на данните чрез прилагане на вертикални разделяния, и се въвежда средата *SeDiCo*, която е разработена с участието на автора преди докторантурата и при сегашните изследвания служи като платформата за измерване на времето за изпълнение на заявки за извличане на данни. Формулирани са и целите на труда: да се създадат и тестват практически стратегии за реализация на вертикални разделяния и разпределяне на базата данни, както и изходните хипотези на изследването: че фиксираните вертикални разделяния подобряват сигурността и защитата на данните и че преписването на заявките (query rewriting), както и кеширането и използването на твърди дискове осигуряват достатъчно бързи отговори на заявките.

Глава 1 (дефиниция на проблема) въвежда формално основните понятия с оглед темата на разработката, включително понятия от релационния модел и релационната алгебра. В така въведения формален контекст са дефинирани целите (оптимизация на времето за отговор на заявки) и хипотезите, които трябва да се потвърдят чрез провежданите изследвания, както и очакваните резултати. Въвежда се понятието „фиксирано вертикално разделяне“. Дадени са примери, които онагледяват обектите на изследване.

Глава 2 дефинира формално методологията за постигане на фиксирано вертикално разделяне и дистрибутивно представяне (FVPD) на релационни бази данни и показва теоретичната ѝ коректност. Показва се как методологията ще бъде реализирана на практика в оригиналната среда *SeDiCo* с цел проверка на хипотезите на дисертационния труд за ефективен отговор на въпросите. Фокусът е върху реализацията на операцията съединение (join) и нейната сложност, но са дискутирани също операциите създаване, четене, обновяване и изтриване на записи.

Глава 3 (свързани разработки, related work) описва архитектурни решения за реализацията на основните функционалности на средата *SeDiCo,* като ги обяснява от гледна точка на потребностите на дисертационния труд – а именно, наличие на тестваща платформа. Разгледани са решенията за осигуряване на сигурност и защита, облачната архитектура, вграждане на подхода за налагане на обекти и релации в заявките, реализация на кеширането. Еднпо по-компактно представяне на материала (без толкова голяма степен на детайлност) би откроило по-добре приносите на дисертанта.

Глава 4 (Концептуализация) представя три възможни подхода да се оптимизира времето за изпълнение на *FVPD*-заявки: преписване на въпроса (query rewriting), локално и отдалечено кеширане и *SSD*-подход с реализация върху твърди дискове (*SSD* - solid state drive). Първият подход е въведен с формален запис, а другите два са илюстрирани с много примери. Оценката на ефективността на предлаганите в дисертацията решения се извършва в рамките на тези три сценария за изпълнение на заявки.

Глава 5 (Реализация) описва практическата разработка в средата *SeDiCo* на трите подхода за оптимизация на времето за отговор на *FVPD*-заявки, които са въведени в Глава 4. Подробно е описана и имплементацията на заявки при различни видове кеширане: кеширане върху сървър, локално и отдалечено.

Глава 6 (Оценка на ефективността) обобщено представя наблюдения на експерименти, в които се изследва бързодействието на 3-те подхода за оптимизация на времето за изпълнение на заявки. Оценена е и ефективността на средата *SeDiCo* като платформа за тестване на *FVPD*-заявки и са набелязани нейни слабости и ограничения.

Глава 7 (Резюме на резултатите, свързани с постигане на ефективност при изпълнение на *FVPD*-заявки) дискутира как получените резултати отговарят на задачите и целите от Глава 1 и потвърждават хипотезите, поставени в основата на научно-приложното изследване. Заключението е, че подходите за преписване на въпроса (query rewriting) и кеширане са обещаващи направления за развитие на средата *SeDiCo* като сигурен разпределен облачен склад за данни. Следователно, основната идея на дисертационния труд – постигане на *защита-и-сигурност-чрез-разделяне* е разумна и трябва да бъде развивана като инструмент за достъп до големи данни в облачни платформи.

Глава 8 (Приложение в бази данни на семантичния интернет) представя един конкретен сценарий как предложеният подход за оптимизация на изпълнението на заявки може да се използва в бази от RDF-данни за SPARQL-заявки. Предложението е как да се подхожда към въпросите на сигурността и защитата в случая на свързани данни (linked data). Подобно на глава 2, понятията са въведени формално и е доказана коректността на подхода за разделяне. Дадена е оценка на сложността на подхода (тя е n^m, където n е броят на RDF-тройките, а m – броят на дяловете). Тъй като обикновено дяловете са няколко, сложността може да се разглежда като полиномиална, и това дава възможност да се мисли за практическо приложение на подхода. Тази глава убедително показва възможностите за широко приложение на идеята за вертикалното разделяне с цел осигуряване на сигурност и защита на данните. Намирам за подходящо, че главата съдържа кратко въведение в идеите на семантичния интернет и доста подробен обзор на съществуващи системи и решения; изнасянето на този обзор в предишни глави би затруднило четенето на труда. Главата представя едно завършено изследване, информативно и интересно за четене.

В Заключението се обсъждат получените резултати и тяхната важност. Поради актуалността и значителния приложен ефект на темата, бъдещата работа е очертана в няколко направления – развитие на средата *SeDiCo*, интеграцията ѝ в хетерогенни инфраструктури,

евентуално развитие към NoSQL-бази, разширяване към диагонални разделяния. Много добро впечатление прави списъкът с публикации, свързани с резултати по дисертацията, в който е дадено кратко резюме на съдържанието на всяка статия – така може да се проследи развитието на труда във времето. Представен е списък на изнесени лекции на конференции и семинари, посветени на проблемите на сигурността и защитата на данните в облачни архитектури, средата *SeDiCo* и оптимизацията на заявки. Изброени са 11 дипломни работи на студенти, работили под ръководството на автора при създаването на средата *SeDiCo*.

Авторефератът отразява коректно съдържанието и резултатите на труда, макар че би могъл да бъде по-кратък.

Приноси на дисертационния труд

Дисертацията е изградена върху една оригинална идея: да се използва вертикално разделяне на базата данни (vertical partitioning) с цел да се осигури сигурност и защита, като се предложи подход за оптимизация на времето за изпълнение на заявките и се постигне скорост на обработка, сравнима с времето на изпълнение на заявка в не-разделена база данни. Експериментите са извършени над средата *SeDiCo*, разработена с участие на автора преди започване на докторантурата, в която по време на докторантурата са вградени функции по оптимизация на изпълнението на заявките. Приемам приносите на труда, така както ги е формулирал авторът:

- дефиниране на принципа *сигурност-чрез-разделяне* за релационни бази данни;
- създаване на нови стратегии за обработка на заявки, зададени към бази с фиксирани вертикално разделени дистрибутирани (*FVPD*) данни. Въвеждане на формални дефиниции на основните понятия и доказване на коректността на подхода;
- интеграция на функционалност за обработка на *FVPD*-заявки в средата *SeDiCo* с цел експериментално тестване на предложения подход;
- оценка на ефективността за изпълнение на *FVPD*-заявки и сравнение с времето за отговор при бази данни с традиционна организация и
- предложение за адаптация на *FVPD*-подхода към бази данни от RDF-тройки и SPARQL-заявки в семантичния интернет.

Стремежът на автора към изчерпателно изследване на проблемите оставя отлично впечатление: разгледани са различни хардуерни среди с различна производителност, разнообразни методи за реализация (например при кеширане), както и няколко бази данни с традиционна организация за сравнение. Комплексният подход доказва техническата компетентност на автора като информатик и програмист в съвременния контекст на облачните платформи. Многобройните публикации и изнесените лекции по покана са свидетелство за интереса към резултатите на дисертацията. Интегрирането на дузина дипломни работи на студенти, които допринасят за постъпковото развитие на средата *SeDiCo*, показва уменията на дисертанта да ръководи разработката на сложни софтуерни платформи.

Лични впечатления и бележки

Според личните ми наблюдения от срещи и семинари, Йенс Колер е автор на представените иновативни резултати, което се потвърждава и от декларацията за оригиналност.

При запознаване с предишни версии на дисертацията съм отправяла редица бележки и коментари, с цел подобряване на качеството на труда. На предварителната защита дисертантът получи множество предложения за редакции от присъстващите колеги. В настоящата версия на труда се вижда, че всички бележки са старателно отразени и дисертацията се доближава до традиционното оформяне на докторска работа в България. Текстът е добре организиран, с ясно разделяне по глави и тематика, и с удобна за възприемане последователност на изложението (макар че на места има излишни детайли). Формалното изложение е значително подобрено.

Като критични бележки, освен споменатата вече излишна според мен детайлност, бих добавила необходимостта от последна техническа редакция на отделни параграфи, например на стр. 17 въведеният формат на елементите в клетките r_{kl} не съответства на формата на стойностите в релационния модел на Фигура 1.1. Дребните технически пропуски не променят научните качества на труда.

Заключение

Оценката ми за докторанта и за дисертацията е положителна. Получените резултати доказват професионалната компетентност и потенциала на кандидата за извършване на самостоятелни научни изследвания не само в ролята на докторант, но също и като ръководител на дипломанти и проекти. Личните ми впечатления от Йенс Колер са отлични поради неговото сериозно отношение към поставените задачи, отдаденост на работата и прецизност в качеството, професионализъм, непринуденост и дружелюбие.

Авторът е направил задълбочено изследване на поставения проблем и е предложил цялостно решение в нова и перспективна област. Изпълнени са всички изисквания на ЗРАСРБ, на Правилника за неговото приложение, както и специфичните изисквания за придобиване на академични степени в ИИКТ-БАН по отношение на обхват, обем и качество на дисертационния труд. На тези основания убедено предлагам на уважаемото научно Жури да присъди на Йенс Колер образователната и научна степен **"доктор".**

28 февруари 2018