UNIVERSITATEA DIN BUCUREȘTI

FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

TEZĂ DE DOCTORAT

Contribuții în teoria bazelor de date

(Rezumat)

CONDUCĂTOR ȘTIINȚIFIC,

PROF. UNIV. DR. IOAN TOMESCU

DOCTORAND,

VASILE SILVIU LAURENȚIU

BUCUREȘTI
2014

Apărute la granița dintre matematică și informatică, ca un răspuns la necesitățile din domeniul economic, bazele de date constituie în zilele noastre un domeniu de sine stătător, fundamentat matematic, cu aplicații în toate domeniile activității umane. Cu o istorie de peste o sută de ani, sistemele de gestiune a bazelor de date se dovedesc instrumentul atât de necesar pentru a răspunde la întrebări din ce în ce mai numeroase și din ce în ce mai complexe ale prezentului.

În abundență de ”on-line”, bazele de date au misiunea dificilă, dar și privilegiată, de a reprezenta instrumentul principal care permite stocarea și redistribuirea datelor. În plus, utilizatorul își dorește să primească informația cât mai rapid și într-un format cât mai aproape de modul de înțelegere/exprimare al acestuia. Ca lucrurile să devină și mai complexe, cei care produc informație ce se dorește a fi stocată, de cele mai multe ori, nu au niciun interes să ofere datele într-un format care să permită stocarea lor directă.

În acest decor, activitatea de cercetare din domeniul bazelor de date este foarte dinamică, dominată de o perpetuă schimbare, de dorința de a oferi un solid suport teoretic diferitelor aplicații ce folosesc bazele de date.

În lucrarea de față am prezentat cele mai importante direcții de dezvoltare a sistemelor de gestiune a bazelor de date, care fac obiectul cercetărilor actuale din domeniul abordat. Pornind de la sistemele tradiționale, am expus principalele probleme la care modelele relaționale nu oferă un răspuns. Am contribuit la dezvoltarea și implementarea celor mai cunoscute modele de baze de date care reușesc să modeleze și să manipuleze informația incertă: modelul aleatoriu și cel probabilistic.

În introducerea lucrării am prezentat o scurtă istorie a bazelor de date, pentru a sublinia, în primul rând, impactul pe care l-a avut, în general economia, asupra evoluției bazelor de date. Fiecare modernizare a sistemelor de gestiune a bazelor de date a avut la bază o problemă lansată de viața cotidiană, la care bazele de date au trebuit să răspundă. În prezent dimensiunea bazelor de date, natura informației ce trebuie prelucrată, timpul de răspuns sunt provocări la care

modelele relaționale răspund din ce în ce mai greu. Pe fondul acestor provocări o serie de sisteme de gestiune derivate din modelul relațional încep să câștige teren în fața modelelor relaționale.

În primul capitol este prezentat modelul relațional tradițional, regulile de integritate, precum și algebra relațională. Acest model este inclus deoarece în prezentarea modelelor de gestiune moderne se face o paralelă cu modelul relațional. De cele mai multe ori construcția sistemelor moderne se face pornind de la modelul relațional. În plus, principalele direcții de cercetare dezvoltate în lucrare – modelul aleator și cel probabilist – vin ca un răspuns la problemele întâmpinate în cadrul modelului relațional. Majoritatea cercetărilor din domeniu, care și-au propus realizarea unor aplicații care să nu mai folosească modelul relațional (*No-SQL*), au ajuns la concluzia că este utilă înlocuirea doar parțială a modelului tradițional (*Not-Only-SQL*).

În încheierea primului capitol sunt scoase în evidență principalele vulnerabilități ale celor mai cunoscute sisteme de gestiune a bazelor de date (*Oracle*, *SQL Server*, *MySQL*, *DB2*). Manipularea parțială a valorilor necunoscute reprezintă unul din motivele folosirii modelelor probabiliste.

În capitolul al doilea am abordat modelul bazelor de date aleatoare. Direcția de cercetare este actuală, a apărut ca un răspuns la prelucrarea unor volume mari de informații, date însoțite și de un anumit grad de incertitudine. În momentul în care manipulăm informație incertă, operațiile din cadrul modelului relațional nu mai sunt relevante. Pentru a prelucra astfel de informații, în cadrul acestui capitol am propus o generalizare a operatorilor tradiționali prin operații de aproximare. Am generalizat noțiunea de bază de date aleatoare, prin introducerea conceptului de bază de date aleatoare eterogenă, în care coloanele pot avea diferite repartiții de probabilitate. De asemenea, am studiat și repartiția numărului de linii rezultate în urma operațiilor de *ε-join.* În cadrul subcapitolului 2.2 prezint pașii urmați pentru a demonstra repartiția *Poisson* a numărului de linii al operațiilor de *ε-join*. Aproximarea cardinalității ε-operațiilor pentru bazele de date aleatoare eterogene este demonstrată prin intermediul unei teoreme. O problemă importantă în cadrul teoriei bazelor de date o reprezintă optimizarea cererilor. Un demers asemănător este considerat și în cadrul bazelor de date aleatoare. Se știe, din teoria modelului relațional, că cea mai costisitoare operație este *join*-ul, deoarece la bază este realizată o operație de produs cartezian. Pentru o optimizare a timpului de rulare în cererile ce folosesc mai multe tabele se

recomandă ca tuplurile care produc, cel mai probabil, dimensiuni reduse ale rezultatului operației de *join* să fie primele evaluate. Cunoașterea dimensiunilor operațiilor de *join* intermediare reprezintă o metodă importantă de optimizare a cererilor compuse. În urma testelor realizate, în cadrul bazelor de date aleatoare, am observat o anumită corelație între cardinalitatea tabelelor, valoarea lui ε, distribuția de probabilitatea a valorilor atributelor folosite în operația de *join* și numărul de linii din rezultat. În cadrul subcapitolului 2.3 furnizez un prim rezultat pentru aproximarea numărului de linii pentru operația de ε-*join*. Astfel, se poate aproxima cu ușurință numărul de linii obținute în urma operației de *join* pentru repartițiile de probabilitate normală, exponențială, uniformă etc. Rezultatele obținute au evidențiat o inegalitate între numărul de linii pentru distribuțiile de probabilitate normală, exponențială și uniformă. Aceste rezultate sunt demonstrate în subcapitolul 2.4.

Importanța practică a rezultatelor obținute este sintetizată în figura 2.6, unde arăt că pentru evaluarea operației de *join* dintre trei tabele cu un număr de 100 000 de linii, utilizarea rezultatelor obținute conduce la o reducere a timpului de execuție de aproape 3 ori. Pentru cereri în care numărul de tabele crește sau pentru tabele cu un număr mai mare de linii, rezultatele arată o îmbunătățire considerabilă a timpului de execuție. Rezultatele din acest capitol vor fi utilizate la îmbunătățirea unor algoritmi care folosesc, pentru optimizarea cererilor, diferite tehnici de măsurare a puterii de discriminare a tabelelor considerate în cereri. Pentru a măsura această proprietate a coloanelor în cercetările din prezent nu este luată în considerare și distribuția de probabilitate a valorilor din coloanele respective, dar în capitolul II am stabilit inegalități între puterea de discriminare a diferitelor repartiții de probabilitate.

În capitolul al treilea am extins modelul relațional pentru a putea lucra cu informație incompletă din punct de vedere al incertitudinii (o valoare are asociată o probabilitate). Am prezentat aplicații care generează, prelucrează și manipulează informație cu un anumit grad de incertitudine. Modelul probabilist de gestiune a bazelor de date a contribuit la dezvoltarea de tehnologii cu un impact deosebit în modul de reprezentare (*JSON*) și de regăsire a informației (*Knowledge Graph*). În subcapitolul 3.2 am prezentat aspectele teoretice care permit generalizarea modelului relațional pentru date probabiliste: bază de date probabilistă, răspuns posibil, urma unei cereri etc. Este formulată o teoremă prin care demonstrez că modelul probabilist propus este închis peste mulțimea operatorilor algebrei relaționale.

În continuare am introdus noțiunea de cheie primară cu ajutorul definiției tuplurilor echivalente. Folosind patru operatori tradiționali (*plus*, *minus*, *înmulțire*, *maximum*) am generalizat toți operatorii modelului relațional, implementarea în *PL/SQL* a acestora fiind inclusă în anexa acestei lucrări.

Actualitatea demersului din acest capitol este justificată de modificările la care asistăm în ultimii ani cu privire la conținutul paginilor web (documentele *JSON*, *Google Search* și *Google Now*). Impactul noilor tehnologii este covârșitor, mai ales pentru paginile ce oferă diferite produse spre comercializare. Pentru cei ce lucrează în domeniul optimizării paginilor web (SEO), traducerea paginilor, pentru a se alinia la noile tehnologii, devine o cerință obligatorie.

În perioada elaborării acestei lucrări, am folosit rezultatele obținute în cadrul a două aplicații: controlul extern de calitate a laboratoarelor de analize medicale și pentru simularea indicelui de secetă în România în perioadele 2021-2050 și 2071-2100 pe baza măsurătorilor din perioada 1961-1990.

În cadrul controlului extern de calitate, laboratoarele participă la testarea unor materiale de control cu o anumită frecvență, trimit rezultatele (de obicei numerice) coordonatorului schemei de control, datele sunt procesate și laboratorul primește un raport ce arată performanța rezultatelor laboratorului în comparație cu cea a altor laboratoare înscrise la același program. În cadrul manipulării datelor primite, coordonatorul trebuie să țină cont de incertitudinea aparatelor folosite de participant, în acest fel sunt realizate operații asupra unei baze de date aleatoare (două măsurători repetate furnizează rezultate diferite).

Pentru simularea indicelui de secetă au fost create diferite scenarii, prin generarea a diferite seturi de date, pornind de la măsurătorile existente. Fiecare interpretare a rezultatelor este însoțită de o anumită probabilitate, iar pentru stabilirea unor asemănări între diferite perioade de timp au fost folosite *ε-aproximări*.

Lucrarea de față se vrea a fi începutul unei cercetări mai ample, pe care doresc să o continui în anii ce vor urma. În viitorul apropiat am în vedere publicarea unor rezultate de optimizare a cererilor pentru baze de date probabiliste, precum și a rezultatelor studiilor în domeniul bazelor de date nerelaționale.

Consider că abordarea din această lucrare este una de actualitate, încadrându-se în tendințele moderne de dezvoltare a sistemelor de gestiune a bazelor de date, conținutul ei putând fi utilizat ca suport în studiul modelelor de baze de date nerelaționale.

Rezultatele au fost publicate în lucrările:

VELCESCU, L., VASILE, S. L., Relational operators in heterogeneous random databases, IEEE Proceedings of the 11th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, IEEE Computer Press, pp. 407-412, 2009.

VELCESCU, L., VASILE, S. L., Relations Between Approximate Join Cardinalities in Random Database Queries, Scientific Bulletin of "Politehnica" University of Timisoara, no. 4, vol. 57, 247-252, 2012.

VELCESCU, L., VASILE, S. L., Distribution of result sets cardinalities in heterogeneous random databases, MDIS (Modelling and development of intelligent systems), pp. 316-323, 2009.

BORONEANT, C., CAIAN, M., VASILE, S. L., COLL, J.R., Changes in drought characteristics for Romania projected by a regional climatic model, Proceedings of the Global Conference on Global Warming, pp. 228-238, 2011.

VASILE, S. L., Query optimization on random databases, Studies in Informatics and Control, vol. 23, pp. 257-265, 2014.

VELCESCU, L., VASILE, S. L., Inequalities between the relational result sets cardinalities in random databases, SACI (Symposium on Applied Computational Intelligence and Informatics), pp. 209 – 212, 2012.

VELCESCU, L., VASILE, S. L., Estimation of the selectivity factor for a set of queries, MDIS (Modelling and development of intelligent systems), pp. 220-225, 2011.

VASILE, S. L., Some aspects of clinical trials, A-16-a Conferință a Societății de Probabilități și Statistică din România, București, 2013.

VASILE, S. L., VELCESCU, L., Une application de la statistique dans l'optimisation des bases de données, 10ème Colloque Franco-Roumain de Mathématiques Appliquées, 2010, Poitiers, Franța.