UNIVERSITATEA BABEŞ-BOLYAI
FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ

Modele de date NoSQL: contribuții la reprezentarea structurii și evaluarea performanței

Rezumatul tezei de doctorat

Student-doctorand: Camelia-Florina ANDOR
Coordonator ştiinţific: prof. dr. Bazil PÂRV

2021

1. Introducere

În contextul actual, alegerea unui sistem de gestiune a bazelor de date (SGBD) nepotrivit specificului unei aplicaţii are un impact semnificativ asupra performanţei. Aplicaţiile moderne generează o cantitate impresionantă de date, care necesită SGBD-uri adaptabile şi performante, capabile să funcţioneze cu uşurinţă într-un mediu distribuit. Aceste date generate într-un ritm accelerat se află în diverse formate şi necesită procesări complexe. SGBD-urile relaţionale nu au fost proiectate pentru a funcţiona optim în asemenea condiţii. Acest lucru a dus la apariţia unei noi generaţii de SGBD-uri, adaptate nevoilor actuale de stocare şi de procesare a datelor. Aceste noi SGBD-uri fac parte din categoria NoSQL, sunt în general proiecte de tipul *open source* şi funcţionează cu succes într-un mediu distribuit. Există în prezent o multitudine de implementări NoSQL disponibile pe piaţă, fiecare având un anumit model de date, un limbaj de interogare propriu şi funcţionalităţi automatizate de replicare şi distribuţie a datelor. Interacţiunea cu aceste SGBD-uri este mai dificilă, acest fapt datorându-se particularităţilor fiecărei implementări dar şi naturii puternic specializate în rezolvarea anumitor probleme din sfera *big data*. Regulile de modelare a datelor diferă de la o implementare la alta, iar familiarizarea cu acestea necesită destul de mult timp şi efort. Unele implementări NoSQL nu beneficiază de o interfaţă grafică de tip client, iar interacţiunea cu acestea se face doar prin intermediul unei aplicaţii de tip consolă. Dezvoltatorii SGBD-urilor NoSQL sunt în general foarte preocupaţi de adăugarea unor noi funcţionalităţi şi de îmbunătăţirea celor existente, astfel că acestea au o evoluţie rapidă şi prezintă diferenţe semnificative între versiuni. Gradul înalt de performanţă şi flexibilitatea structurii datelor sunt două dintre cele mai mari avantaje ale SGBD-urilor NoSQL.

Alegerea SGBD-ului NoSQL potrivit pentru o aplicaţie dată este o sarcină destul de greu de realizat. Marea varietate de produse disponibile şi diferenţele foarte mari dintre ele îngreunează semnificativ realizarea unei comparaţii corecte. Documentaţiile tehnice ale produselor sunt de asemenea foarte complexe şi necesită un studiu îndelungat. Experimentele de evaluare a performanţei SGBD-urilor NoSQL oferă informaţii importante despre performanţa acestora în contexte specifice, fiind utile în procesul de alegere a SGBD-ului potrivit.

În această lucrare ne-am propus să simplificăm procesul de alegere a SGBD-ului NoSQL potrivit unei aplicaţii sau unui tip de aplicaţii date şi să facilităm interacţiunea utilizatorilor cu bazele de date NoSQL. Specialişti din toate domeniile lucrează în prezent cu cantităţi mari de date care necesită procesări complexe. Aplicaţiile care facilitează interacţiunea cu baza de date sunt de mare ajutor, în special în cazul utilizatorilor care nu sunt specializaţi în domeniul informaticii.

Contextul şi descrierea problemelor pe care ne-am propus să le soluţionăm în lucrarea de faţă sunt prezentate în capitolele 1, 2 şi 3.

Contribuţiile originale ale lucrării sunt prezentate în capitolele 4, 5 şi 6.

2. Structura tezei de doctorat

Teza de doctorat este compusă din şapte capitole.

În capitolul 1 sunt prezentate neajunsurile modelului de date relaţional, care au condus la apariţia SGBD-urilor NoSQL, dar şi cele patru modele de date NoSQL fundamentale. SGBD-urile NoSQL folosesc modele de date diferite de cel relaţional, sunt proiectate pentru a funcţiona fără probleme într-un mediu distribuit şi oferă un grad mai mare de flexibilitate a structurii datelor. Fiecare model de date NoSQL fundamental (*key-value*, *document*, *column-family*, *graph*) este descris într-o secţiune separată. De asemenea, sunt abordate şi alte teme importante din zona SGBD-urilor NoSQL, cum ar fi replicarea şi distribuţia datelor.

Capitolul 2 abordează problema dificultăţii alegerii SGBD-ului NoSQL potrivit pentru o aplicaţie dată. Diferenţele mari dintre SGBD-urile NoSQL, datorate modelelor de date diferite, limbajelor de interogare proprii şi altor caracteristici îngreunează realizarea unei comparaţii corecte între diferite implementări NoSQL. Experimentele de evaluare a performanţei SGBD-urilor NoSQL reprezintă o soluţie în acest caz, deoarece dezvăluie informaţii importante legate de performanţa fiecărei implementări NoSQL testate în condiţii specifice. Luând în considerare atât cazurile de utilizare comune, cât şi gradul de complexitate al modelului de date, sunt alese două modele de date NoSQL considerate mai versatile (*document* şi *column-family*) şi este descrisă câte o implementare NoSQL reprezentativă pentru fiecare (MongoDB, respectiv Apache Cassandra). În continuare, este abordat subiectul aplicaţiilor de evaluare a performanţei. Sunt descrise atât tipurile de aplicaţii de evaluare a performanţei cât şi rolul lor într-un experiment de evaluare a performanţei. Cea mai populară aplicaţie de evaluare a performanţei, *YCSB*, este de asemenea prezentată. Sunt descrise pe scurt şi studii experimentale de evaluare a performanţei din literatură în care a fost folosită aplicaţia *YCSB*.

În capitolul 3 este abordată problema interacţiunii dificile cu baze de date NoSQL, datorată în principal faptului că fiecare SGBD NoSQL are un limbaj de interogare propriu. O soluţie la această problemă ar putea fi aplicaţiile de tip *Visual Query System*, care folosesc o reprezentare conceptuală a structurii bazei de date şi a interogărilor şi permit utilizatorilor interacţiunea cu baza de date într-o manieră vizuală. Sunt prezentate diferite abordări din literatura de specialitate care folosec grafuri conceptuale pentru reprezentarea structurii bazei de date şi a interogărilor. Aceste abordări vizează bazele de date relaţionale şi datele stocate în format XML şi sunt implementate în aplicaţii de tip *Visual Query System*.

Capitolul 4 prezintă cinci studii de caz bazate pe experimente de evaluare a performanţei SGBD-urilor MongoDB şi Apache Cassandra. Metricile de performanţă analizate în aceste studii sunt *throughput*, *average latency* (latenţa medie) şi *total runtime* (timpul total de execuţie). În ultimele două studii de caz sunt folosite versiuni noi ale SGBD-urilor MongoDB şi Apache Cassandra. În urma analizei rezultatelor ultimelor două studii de caz, s-a dovedit că există diferenţe semnifica-

tive ale performanţei între versiuni diferite ale aceluiaşi SGBD. Aplicaţia de evaluare a performanţei folosită în acele cinci studii de caz este *YCSB*. În urma utilizării aplicaţiei *YCSB*, am remarcat anumite neajunsuri ale acesteia, precum lipsa unei interfeţe grafice bazate pe un model conceptual de reprezentare a structurii bazei de date şi a operaţiilor implicate în teste. Pentru realizarea unor experimente de evaluare a performanţei care necesită seturi de date şi seturi de operaţii personalizate cu *YCSB* sunt necesare modificări sau extinderi ale codului sursă.

În capitolul 5 propunem două metode de reprezentare a structurii bazei de date şi a interogărilor bazate pe grafuri conceptuale pentru MongoDB, respectiv Apache Cassandra. Aceste metode pot fi implementate atât în aplicaţii de tip *Visual Query System*, cât şi într-o aplicaţie de evaluare a performanţei SGBD-urilor.

Capitolul 6 prezintă două implementări ale metodei de reprezentare a structurii bazei de date şi a interogărilor MongoDB. Aplicaţia web *Conceptual Graphs for MongoDB* se încadrează în categoria *Visual Query System* şi oferă posibilitatea de a reprezenta structura bazei de date şi a interogărilor într-o manieră vizuală cu ajutorul grafurilor conceptuale. Aplicaţia transformă reprezentările conceptuale în instrucţiunile echivalente din MongoDB, care pot fi mai apoi executate. Aplicaţia desktop *DBMark* este o aplicaţie de evaluare a performanţei care oferă o interfaţă grafică bazată pe grafuri conceptuale. Momentan, aplicaţia *DBMark* poate fi folosită pentru evaluarea performanţei SGBD-ului MongoDB. Cu ajutorul acestei aplicaţii se pot realiza experimente de evaluare a performanţei în care se folosesc seturi de date şi seturi de operaţii personalizate. Interfaţa grafică bazată pe grafuri conceptuale facilitează personalizarea experimentelor, iar aplicaţia poate genera automat reprezentarea conceptuală a structurii unei baze de date deja existente. Bineînţeles, se pot crea în aplicaţie reprezentări conceptuale noi şi în acest caz se vor genera în mod automat instrucţiunile MongoDB echivalente.

În ultimul capitol sunt prezentate concluziile acestei lucrări şi direcţiile de cercetare viitoare. Ne propunem ca în viitor să dezvoltăm metode de reprezentare a structurii bazei de date şi a interogărilor bazate pe grafuri conceptuale pentru alte SGBD-uri NoSQL, dar şi să realizăm alte studii de caz bazate pe experimente de evaluare a performanţei SGBD-urilor NoSQL.

3. Contribuţii originale

Cele două probleme abordate în această lucrare sunt reprezentate de dificultatea şi complexitatea procesului de alegere a SGBD-ului NoSQL potrivit pentru o aplicaţie dată şi dificultatea interacţiunii cu SGBD-urile NoSQL, care se datorează în principal diferenţelor semnificative dintre limbajele de interogare. Contribuţiile originale vizează simplificarea procesului de alegere a SGBD-ului NoSQL potrivit unei aplicaţii date şi facilitarea interacţiunii cu SGBD-uri NoSQL în cazul utilizatorilor care nu sunt specializaţi în informatică.

Experimentele de evaluare a performanţei au un rol important în simplificarea procesului de alegere a SGBD-ului NoSQL potrivit pentru o aplicaţie dată. Acestea oferă informaţii importante despre performanţa SGBD-urilor evaluate în condiţiile specifice cazului de utilizare considerat. În capitolul 4 sunt prezentate cinci studii de caz realizate pe baza experimentelor de evaluare a performanţei. În aceste studii de caz a fost analizată performanţa a două dintre cele mai populare SGBD-uri NoSQL, MongoDB şi Apache Cassandra. Metricile de performanţă analizate sunt *throughput*, *average latency* (latenţa medie) şi *total runtime* (timpul total de execuţie). Primele trei studii de caz (SC1, SC2 şi SC3) analizează performanţa SGBD-urilor MongoDB şi Apache Cassandra din perspectiva *throughput*-ului, a latenţei medii şi a timpului total de execuţie. A fost realizată şi analiza statistică a rezultatelor obţinute. SGBD-urile NoSQL au o evoluţie rapidă, şi există diferenţe semnificative între versiuni diferite ale aceluiaşi SGBD care pot avea impact asupra performanţei. În ultimele două studii de caz (SC4 şi SC5) au fost refăcute experimentele din SC1 şi SC3 folosind versiuni noi ale SGBD-urilor, iar în urma analizei rezultatelor s-a dovedit că impactul asupra performanţei este semnificativ.

Experimentele de evaluare a performanţei SGBD-urilor NoSQL sunt realizate cu ajutorul unor aplicaţii de evaluare a performanţei, numite şi *benchmarking tools* sau *benchmarking frameworks*. În capitolul 2 au fost prezentate cele două tipuri de aplicaţii de evaluare a performanţei şi rolul pe care îl au acestea în procesul de evaluare a performanţei. În realizarea experimentelor care stau la baza celor cinci studii de caz prezentate în capitolul 4 a fost folosită cea mai populară aplicaţie de evaluare a performanţei SGBD-urilor NoSQL, *YCSB*. O descriere a acestei aplicaţii dar şi a SGBD-urilor MongoDB şi Apache Cassandra a fost realizată în capitolul 2. În urma utilizării aplicaţiei *YCSB* în toate cele cinci studii de caz, am identificat anumite neajunsuri ale acesteia cu privire la personalizarea experimentelor de evaluare a performanţei, cum ar fi lipsa unei interfeţe grafice care să permită o reprezentare la nivel conceptual a structurii bazei de date şi a operaţiilor executate într-un test de performanţă.

În capitolul 5 am propus două metode de reprezentare a structurii bazei de date şi a interogărilor cu ajutorul grafurilor conceptuale pentru SGBD-urile MongoDB şi Apache Cassandra. Acestea pot fi implementate într-o aplicaţie de evaluare a performanţei SGBD-urilor MongoDB şi Apache Cassandra care simplifică efectuarea experimentelor de evaluare a performanţei ce folosesc seturi de date

şi de operaţii personalizate. Metoda de reprezentare a structurii bazei de date şi a interogărilor MongoDB a fost implementată în aplicaţia *DBMark*, care este descrisă în capitolul 6. Aplicaţia *DBMark* oferă o interfaţă grafică care permite reprezentarea structurii bazei de date şi a interogărilor MongoDB cu ajutorul grafurilor conceptuale şi poate fi folosită în efectuarea experimentelor de evaluare a performanţei care implică seturi de date şi de operaţii personalizate.

Interacţiunea cu SGBD-urile NoSQL, care este îngreunată de marea varietate de limbaje de interogare, poate fi simplificată cu ajutorul aplicaţiilor de tip *Visual Query System*. Acestea facilitează interacţiunea cu baza de date, cu precădere în cazul utilizatorilor care nu sunt specializşti în domeniul informaticii. Într-o aplicaţie de tipul *Visual Query System*, utilizatorul poate reprezenta structura bazei de date şi a interogărilor într-o manieră vizuală, folosind un limbaj de modelare conceptual. Metodele de reprezentare a structurii bazei de date şi a interogărilor pentru SGBD-urile MongoDB şi Apache Cassandra propuse în capitolul 5 pot fi implementate şi în aplicaţii de tip *Visual Query System*. Metoda de reprezentare a structurii bazei de date şi a interogărilor MongoDB a fost implementată în aplicaţia *Conceptual Graphs for MongoDB*, descrisă în capitolul 6. Aplicaţia *Conceptual Graphs for MongoDB* simplifică interacţiunea cu baza de date în cazul SGBD-ului MongoDB, transformând reprezentările bazate pe grafuri conceptuale a structurii bazei de date şi a interogărilor în instrucţiuni ce pot fi executate în MongoDB.

În ceea ce priveşte direcţiile viitoare de cercetare, ne propunem să implementăm metoda de reprezentare a structurii bazei de date şi a interogărilor Apache Cassandra într-o aplicaţie de tipul *Visual Query System*, dar şi în aplicaţia de evaluare a performanţei *DBMark*, pentru a putea fi ulterior folosită şi în evaluarea performanţei SGBD-ului Apache Cassandra. În ceea ce priveşte experimentele de evaluare a performanţei SGBD-urilor NoSQL, ne propunem să efectuăm noi experimente care să includă şi alte SGBD-uri NoSQL şi să utilizăm o infrastructură *cloud-based*. Un alt obiectiv viitor este dezvoltarea unor metode de reprezentare a structurii bazei de date şi a interogărilor cu ajutorul grafurilor conceptuale pentru alte SGBD-uri NoSQL. Acestea ar putea fi ulterior folosite pentru a extinde funcţionalităţile aplicaţiei de evaluare a performanţei *DBMark* şi pentru a implementa noi aplicaţii din categoria *Visual Query System*.

Cuvinte-cheie: baze de date NoSQL; aplicaţii de evaluare a performanţei; MongoDB; Apache Cassandra; YCSB; grafuri conceptuale; modele de date; big data.

Cuprinsul tezei de doctorat