BAZE DE DATE NO-SQL: MongoDB

RUSU Florin, SARANCIUC Dorian

Universitatea Tehnică a Moldovei

Abstract: *Această lucrare descrie cadrul No-SQL, în cazul dat e vorba despre MongoDB. O bază de date NoSQL ignoră principiile RDBMS şi nu stochează date folosind tabele ci folosind chei de identificare. Datele pot fi regăsite în funcţie de cheile asignate. Acest tip de baze de date evadează din rigorile relaţionale prin lipsa unei scheme, lipsa necesităţii de normalizare a datelor şi de stocare a relaţiilor dintre tabele aducând astfel performanţe sporite aplicaţiilor care le folosesc. De asemenea, ele îmbunătăţesc şi răspunsul la schimbări de-a lungul timpului. Într-un sistem relaţional nu există flexibilitatea necesară pentru a asimila modificări în modelul de date. Faptul că bazele de date NoSQL nu au o schemă de date fixă face aceste baze de date să fie mult mai flexibile şi adaptabile la schimbări de model în cursul anilor.*

Cuvinte cheie: *MongoDB, NO-SQL, replication, sharding, OLTP*

1. Introducere

Deşi în prezent nu există competitori reali pentru SQL, problema se schimbă în cazul aplicaţiilor web. În acest caz, nu există o multitudine de asocieri *inner* şi *outer* pentru calcule complexe, ci mai degrabă se foloseşte o gândire obiect orientată, îndeosebi datorită MVC (Model-View-Controller). Pentru a transforma aceste modele obiect orientate în baze de date relaţionale au loc diverse procese de normalizare ce rezultă în ierarhii complexe de tabele şi îndepartează întregul model de principiile modelării orientate obiect. Faptul că limbajul SQL permite interogări dinamice asupra unor seturi de date complexe este inutil prin folosirea unei baze de date SQL doar pentru stocarea persistentă a datelor orientate obiect.

Aici intervin bazele de date NoSQL. Acestea permit dezvoltatorului de aplicaţii să stocheze date care nu au o schemă predefinită. Carlo Strozzi a folosit prima dată termenul de NoSQL în 1998, reprezentând numele bazei sale de date relaţionale open-source fără interfaţă SQL. Termenul a fost reintrodus în 2009 de către Eric Evans în cadrul unui eveniment cu tema „Baze de date distribuite open-source". De această dată termenul nu a fost folosit pentru a defini un întreg sistem, ci a fost folosit pentru a marca un pas în evoluţia de la baze de date relaţionale către baze de date cu performanţe sporite. Din acest moment s-au dezvoltat un număr vast de baze de date non-relaţionale. [1]

O definiţie a bazelor de date NoSQL este dată de site-ul nosql-database.org. Acesta caracterizează bazele de date NoSQL ca noua generaţie de baze de date ce îndeplinesc următoarele condiţii: nu sunt relaţionale, sunt distribuite, open-source şi se caracterizează prin scalabilitate orizontală. Alte caracteristici ce trebuiesc menţionate sunt lipsa unei scheme pentru a modela baza de date, prezintă suport pentru replicare, API simplu, nu respectă în întregime criteriile ACID (atomicitate, consistenta, izolare si durabilitate), stochează o cantitate mare de date. [2]

Conform definiţiei date de Rick Cattell, bazele de date NoSQL prezintă şase trăsături de bază:

- Abilitatea de a scala orizontal pe mai multe servere;
- Abilitatea de a replica şi distribui datele pe mai multe servere;
- CLI (call level interface) caracterizat prin simplitate (în contrast cu SQL binding);
- Un model concurenţial mai slab decat modelul relaţional (ACID);
- Utilizarea eficientă a indexării distribuite şi a RAM pentru o stocare eficientă;
- Abilitatea de a adăuga dinamic noi atribute la înregistrările existente.

O altă caracteristică importantă a sistemelor NoSQL este arhitectura "shared nothing" prin care fiecare nod/server este independent, nici unul din ele nu partajează memorie sau spaţiu. Datorită acestei caracteristici pot fi efectuate un număr mare de operaţii de citire/scriere pe secundă. O simplă operaţie poartă numele de OLTP (online transaction processing) şi este de asemenea comună şi în cadrul aplicaţiilor web moderne.

Conform unui studiu realizat de 451 Research Group intitulat „MySQL vs. NoSQL and NewSQL" între 2009 şi 2011 s-a înregistrat o scădere în utilizarea MySQL de la 82% la 73%. Studiul a fost efectuat asupra unui eşantion compus din 347 de utilizatori de baze de date open-source. 49% din respondenţi au abandonat soluţiile MySQL pentru a migra la soluţii NoSQL. Astfel se poate observa ameninţarea directă pe care o presupune NoSQL asupra MySQL.

Conform aceluiaşi studiu MySQL este direct ameninţat de apariţia NoSQL şi NewSQL, acesta din urmă reprezentând un nou set de baze de date relaţionale care încearcă să adauge la modelul relaţional performanţele şi funcţionalităţile bazelor de date NoSQL. (figura 1)

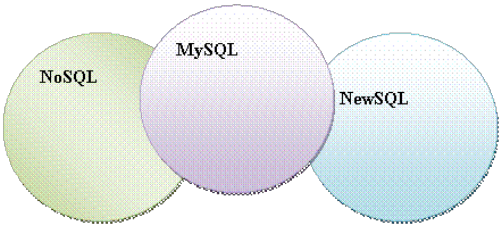

Fig. 1. Cucerirea pieții de NoSQL și NewSQL

2. Avantajele utilizării sistemelor NoSQL

În bazele de date relaționale optimizările se fac prin definițiile index. Modelul de date pune accent pe normalizarea, abstractizarea și integritatea datelor. Baza de date NoSQL a apărut ca noua generație de baze de date ce permite cereri la noi nivele și fac față nevoilor din ziua de azi, prin arhitecturi construite pe cloud și sisteme distribuite. Bazele de date non-relaționale sunt distribuite și memorează date care nu respectă garanțiile ACID.

Printre primele aplicații mari, ce au ridicat problema scalabilității, se numără și Facebook. Această rețea de socializare presupune un număr impresionant de utilizatori, la nivelul milioanelor. Prin adoptarea bazelor de date NoSQL scalarea masivă a adus un beneficiu considerabil.

Modelele de interogre în NoSQL se bazează pe căutarea unor chei primare sau a unui câmp ID și pe lipsa unei interogări pe alte domenii. Bazele de date MongoDB și CouchDB permit interogări mai avansate, cum ar fi cele static predefinite pe nodurile bazei de date. Proiectarea bazelor de date NoSQL a fost făcută cu caracteristici dinamice de interogare în favoarea obținerii performanței și a scalabilității.

Bazele de date NoSQL reprezintă o abordare în care atributele căutate sunt copiate într-un SQL sau text. Capacitățile de interogare ale acestei baze de date sunt utilizate pentru a prelua cheile primare de sortare în seturi de date, prin care baza de date NoSQL va fi mai târziu accesată. (figura 2,3)

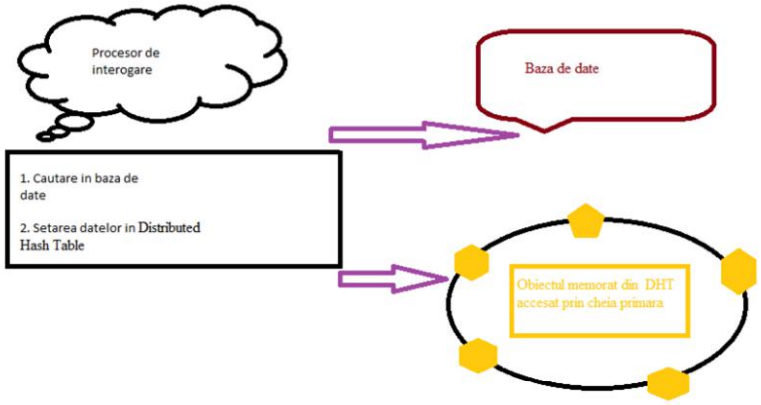

Fig. 2. Ilustrarea preluării cheilor

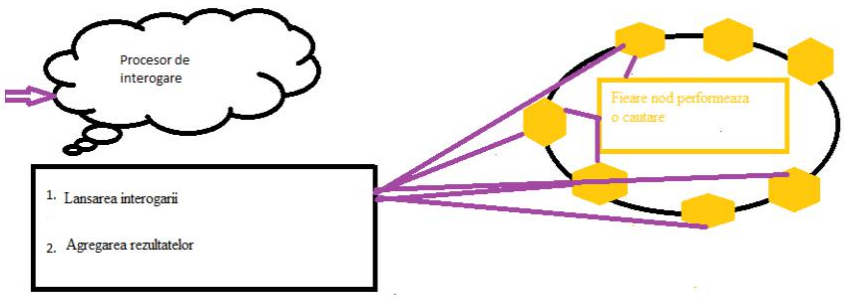

Fig. 3. O altă modalitate de căutare

3. Dezavantajele utilizarii sistemelor NoSQL

Este ştiut faptul că în ansamblul sistemelor, modelele alese joacă un rol vital. Datorită faptului că în sistemele NoSQL nu există o autoritate calificată pentru a defini un singur, bine-definit model, folosirea unui model non-normalizat poate conduce la suprapunerea obiectelor de date. La nivel de sistem este, de asemenea, important să se ţină cont de modelele de concurenţă şi de limitele serviciilor de date alese.

Din punctul de vedere al arhitecturii bazelor de date NoSQL există probleme legate de interfeţe şi interoperabilitate. Modelul DHT (Distributed Hash Table) conţine interfeţele sale încă standardizate, însă nu conţine o semantică intrinsecă pentru indicii. Interoperabilitatea este importantă în momentul în care datele trebuie accesate de mai multe servicii, moment în care se poate pierde din performanţă.

Spre deosebire de bazele de date relaţionale care s-au consacrat ca fiind stabile, sistemele NoSQL apar pe piaţă ca o alternativă ce poate fi pusă în aplicare doar cu precauţie, deoarece încă nu au ajuns la un nivel de maturitate cel puţin egal cu RDBMS-urile. De asemenea, în cazul apariţiei unor probleme, suportul sistemelor NoSQL este încă limitat, acestea fiind în mare proiecte open-source, iar companiile care oferă suport pentru bazele de date NoSQL nu oferă credibilitatea companiilor globale. [8]

Sistemele NoSQL se potrivesc foarte bine şi în tehnologia Cloud, care se bazează pe virtualizare. Există şi un punct slab al virtualizării ce ţine de performanţa I/O, limitările CPU-ului şi ale memoriei fiind de altfel în strânsă legătură. Bazele de date NoSQL folosesc proporţia cea mai mare pe memoria de disc, aceasta fiind locaţia principală de scriere, însă datorită scalării orizontale sunt capabile să memoreze datele eficient.

Bazele de date NoSQL prezintă un dezavantaj din punctul de vedere al administrării, pentru că necesită un anumit efort pentru a fi menţinute şi cunoştiinţe solide pentru instalare. Persoanele calificate pentru bazele de date NoSQL sunt mai puţine decât cele cu experienţă pentru RDBMS. De aceea preluarea lor de către firme trebuie să fie făcută în mod corespunzător, întrucât beneficiile reale aduse de acestea pot fi însoţite uneori şi de anumite probleme.

4. MongoDB

Pentru a putea exemplifica practic bazele de date NoSQL am ales să analizăm baza de date NoSQL MongoDB. Acesată bază de date este uşor de folosit pentru utilizatorii de RDBMS-uri. MongoDB lucrează cu date nestructurate şi organizează aceste date în format document. Implementarea acestei baze de date este mai uşoară decat un RDBMS deoarece ea urmăreşte modelul cheie valoare şi nu are nevoie de o schemă predefinită a datelor. Conceptele acesteia pornesc de la concepte tradiţionale, de aceea întelegerea filosofiei acestei baze de date este ceva uşor de realizat. Prezentul articol urmăreşte atât prezentarea generală a bazei de date cât şi instalarea şi utilizarea ei.

MongoDB este o bază de date open-source NoSQL scrisă în C++. Aceasta poate conţine mai multe baze de date, colecţii şi indecşi. În unele cazuri (baze de date şi colecţii) aceste obiecte pot fi create implicit. Odată create, ele se găsesc în catalogul sistemului db.systems.collection, db.system.indexes. Colecţiile conţin documente (BSON). Aceste documente conţin la rândul lor mai multe câmpuri. În MongoDB nu există câmpuri predefinite spre deosebire de bazele de date relaţionale, unde există coloanele care sunt definite în momentul în care tabelele sunt create. Nu există schemă pentru câmpurile dintr-un document, acestea precum şi tipurile lor pot varia. Astfel nu există operaţia de „alter table” pentru adăugare de coloane. În practică este obişnuit ca o colecţie să aibă o structură omogenă, deşi nu este o cerinţă, colecţiile putând avea structuri diferite. Această flexibilitate presupune uşurinţă în migrarea şi modificarea imaginii de ansamblu asupra datelor.

Dacă e să vorbim despre disponibilitate, atunci putem observa din figura de mai jos, că poate fi utilizată atât pe diferite platforme cât şi limbaje diferite. (figura 4)

Fig. 4. Diferenţierea limbajelor

Caracteristicile MongoDB sunt:

- Stocarea datelor sub formă de documente - Baza de date MongoDB stochează obiecte (documente). Aceste documente reduc nevoia de join;

- Prezintă support pentru indexare – Indexarea pe fiecare din atribute se face în modul tradiţional (RDBMS) asupra cheilor de regăsire ale documentelor;
- Disponibilitate - Disponibilitatea datelor este asigurată printr-un proces automat de failover;
- Auto-Sharding - Shardingul sau partiţionarea datelor pe orizontală se face automat. Citirile şi scrierile sunt distribuite pe partiţii. Lipsa joinurilor face ca interogările distribuite să fie rapide;
- Modificări rapide - MongoDB suportă operaţii de actualizare atomice cât şi pe cele tradiţionale. Operaţiile de mai jos demonstrează flexibilitatea limbajului NoSQL:
- De asemenea trebuie să menționăm că în MongoDB documentele sunt stocate în colecții. Astfel mai multe documente pot fi încadrate într-o singură colecție. (figura 5)

Fig. 5. Reprezentarea colecției

La fel, în cadrul acesteia datele sunt semi-structurate, având forma următoare. (figura 6)

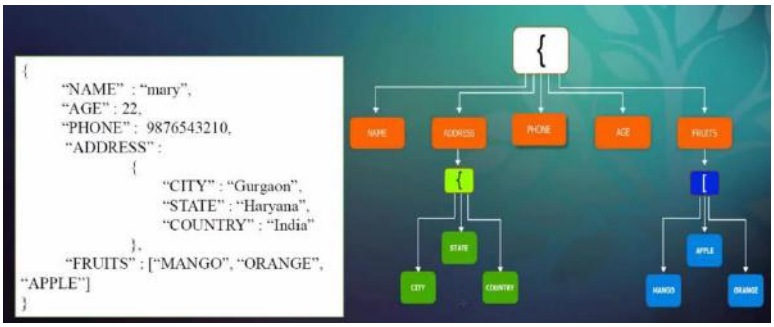

Fig. 6. Reprezentarea datelor semi-structurate.

5. Concluzii

Având la dispoziţie toate aceste informaţii despre conceptul NoSQL putem formula o definiţie proprie a ceea ce considerăm NoSQL. Astfel mişcarea NoSQL reprezintă o încercare de a depăşi limitările modelului relaţional şi un pas de trecere către NewSQL şi anume relaţional plus extra funcţionlităţi NoSQL.

Odată cu apariţia bazelor de date NoSQL, dezvoltatorii au oportunitatea de a beneficia de mai multă agilitate în modelul de date abordat. De asemenea aceste baze de date constituie modelul optim pentru aplicaţiile web. De aceea cunoaşterea caracteristicilor lor este foarte importantă, în special înainte de a migra la o astfel de soluţie.

Baza de date prezentată, MongoDB este o bază de date uşor de înţeles şi manipulat. Aceasta este ideală atât pentru proiecte mici, de test, cât şi pentru proiecte ce implică un volum mare de date. În opinia mea MongoDB este o bază de date ce va fi folosită din ce în ce mai mult pe viitor datorită tendinţei actuale înclinată către aplicaţii Web.

Bibliografie

1. SEEGER, M.: Key-Value stores: a practical overview, Computer Science and Media Ultra-Large-Sites SS09 Stuttgart, Germany, 21 Septembrie 2009.
2. The ultimate reference for NOSQL Databases. http://nosql-database.org/
3. WU, Suzanne: How Much Information IS There in the World. University of Southern California, 10 Februarie 2011.
4. CATTELL, R.: Scalable SQL and NoSQL Data Stores, SIGMON Record, Decembrie 2010, Vol 39, Nr 4.
5. Aslett, M.:" MySQL vs. NoSQL and NewSQL - survey results", 22 May 2012, 451 Research Group.
6. SHALOM, N.: The Common Principles Behind The NoSQL Alternatives, December 2009,- Blog post of 2009-12-15. http://natishalom.typepad.com/nati_shaloms_blog/2009/12/the-common-principlesbehind-the-nosql-alternatives.html.