BAZE DE DATE DISTRIBUITE

MITRIUC Maria

Universitatea Tehnică a Moldovei

Abstract: *În articolul dat este descris conceptul principal al unei baze de date dsitribuite, mijloacele prin care se obține o performanță mai bună de trasmitere a informației, tipurile de baze de date distribuite, tipurile de proiectare şi arhitectura acestora. Este analizat conceptul de DBMS, modul de funcționare a mai multor calculatoare interconectate, principiul de autonomie locală la nivel de companie precum și metoda de procesare paralelă a datelor. Sunt analizate cele 3 nivele ale arhitecturii unei baze de date: intern, conceptual și extern, cît și distribuția spațiului de stocare a sistemului Hadoop pe mai multe servere.*

Cuvinte Cheie: *Baze de date dsitribuite, DBMS, autonomie, heterogenitate, proiectarea top-down, proiectarea bottom-up, nivelul intern, conceptual, extern, Hadoop, Facebook.*

1. Definiția, obiectivele și avantajele bazelor de date distribuite

O bază de date distribuită este o bază de date controlată de un sistem de gestiune a bazelor de date (Data Base Management System - DBMS), în care dispozitivele de stocare a datelor sunt atașate în mod distribuit la mai multe calculatoare. Aceste calculatoare pot ori să se afle fizic în aceeași locație (sală, clădire etc.), ori să fie dipersate într-o rețea de calculatoare interconectate. O bază de date distribuită trebuie să facă distribuirea transparentă (invizibilă) pentru utilizator. Obiectivul transparenței este de a face ca sistemul distribuit să apară ca un sistem centralizat.

Bazele de date distribuite au apărut ca necesitatea de a realiza autonomie locală – fiecare companie își poate controla mai apropiat baza de date. Performanța de a realiza schimbări – în cazul în care există o greșeală, aceasta va afecta doar o singură partiție. Partea economică este un factor primomdial de creare a bazelor de date distribuite: diminuarea costurilor de implementare şi întreţinere prin utilizarea de echipamente uzuale în nodurile de prelucrare. O altă necessitate de apariție a astfel baze de date este că subsistemele unei baze de date distribuite pot fi modificate, adăugate sau deconectate dinamic, fără să se afecteze alți clienți sau partiții.

2. Clasificarea bazelor de date. Sisteme omogene și heterogene. Proiectarea acestora

Clasificarea SGBD distribuite se poate face avînd în vedere trei caracteristici ale sistemelor aflate în nodurile locale ale sistemului: distribuție, autonomie şi heterogenitate.

Într-un sistem omogen (figura 1) toate site-urile utilizează acelaşi SGBD.

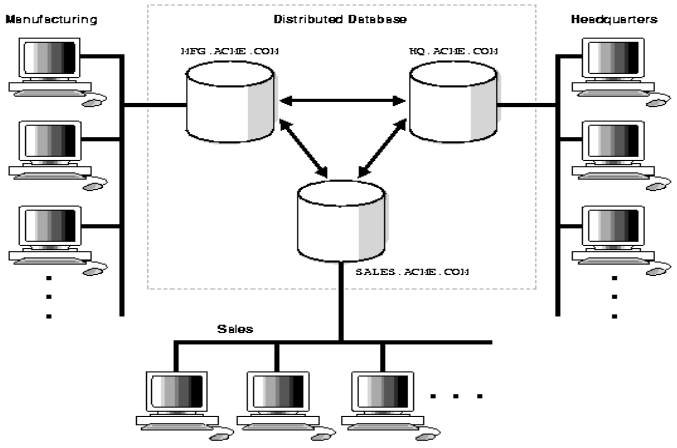

Fig. 1. Arhitectura unei baze de date omogene

Sistemele omogene sunt mai uşor de proiectat şi gestionat, adăugarea unui nou site în SGBDD este mai uşoară şi se obţin performanţe crescute prin exploatarea capacităţii de prelucrare paralelă a mai multor site-uri. Într-un sistem heterogen, site-urile pot utiliza SGBD diferite, bazate pe modele de date diferite: relaţional, în reţea, ierarhice, obiect orientate. De obicei, sistemele heterogene (figura 2) se crează atunci când site-urile individuale şi-au implementat propriile baze de date, iar integrarea este avută în vedere într-o etapă ulterioară. Într-un sistem heterogen sunt necesare traduceri pentru a permite comunicarea dintre diversele SGBD. Pentru

a realiza transparenţa sistemului, este necesar ca utilizatorii să poată efectua cererile în limbajul sistemului SGBD din site-ul local, apoi sistemul are sarcina de a localiza datele şi de a efectua orice traducere necesară.

Pentru a asigura respectarea obiectivelor specifice ale bazelor de date distribuite (BDD), şi anume: creşterea sigurantei sistemului şi a disponibilităţii datelor, descentralizarea resurselor sistemu-lui, o mai bună utilizare a acestora precum şi sporirea adaptabilităţii sistemului la modificările din structura organizatorică, în proiectarea BDD se urmareşte asigurarea respectării urmatoarelor principii:

- Maximizarea prelucrării locale a datelor, care presupune plasarea datelor cât mai aproape de aplicaţiile care le solicită. Se apreciază că într-o BDD corect proiectată aproximativ 90% din volumul de date trebuie sa fie accesate local si numai 10% sa fie accesate de la distantă.
- Asigurarea unui nivel sporit de siguranţă şi disponibilitate a datelor, care poate fi realizat prin replicarea datelor pe mai multe staţii. În aceste conditii, sistemul poate utiliza o copie alternativă atunci când cea care trebuia sa fie accesata în condiţii normale nu este disponibilă.
- Procesarea paralela a datelor. Distribuirea bazelor de date oferă posibilitatea de a utiliza eficient capacitatile procesoarelor din fiecare staţie pentru a maximiza gradul de paralelism în execuţia aplicaţiilor. Creşterea numărului de prelucrari realizate în paralel intra însa în contradicţie cu principiul de maximizare a prelucrărilor locale. În proiectarea BDD trebuie să se asigure un raport optim între cele doua tipuri de prelucrări.

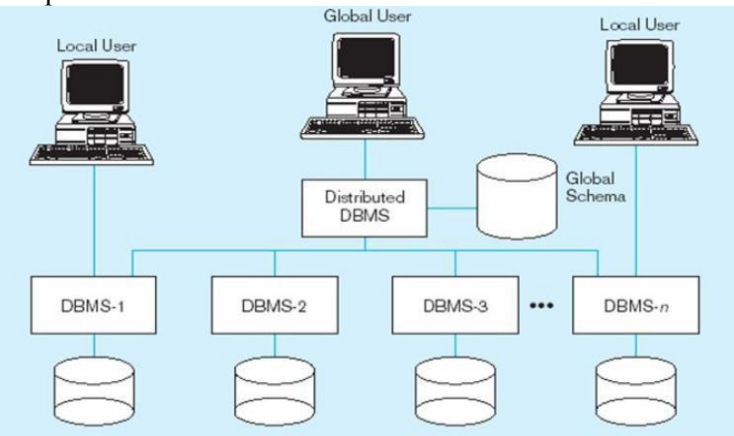

Fig. 2. Arhitectura unei baze de date distribuite heterogene

Tipurile de proiectare a BDD pot fi:

- Proiectarea BDD poate fi realizată, ca şi în cazul BD centralizate, în maniera top-down si/sau bottom-up.
- Proiectarea top-down urmăreste asigurarea unei distribuiri optime a datelor si este utilizată atunci când proiectarea BDD începe de la zero. În general, în urma abordarii top-down sistemele rezultate sunt omogene.
- Proiectarea bottom-up a BDD este recomandata în cazul în care BD locale există şi trebuie integrate într-un sistem unitar. În acest caz obiectivul principal urmarit în proiectarea BDD îl constituie asigurarea unei cooperari cât mai bune între BD existente care sunt, de regula, eterogene.

Proiectarea mixtă, partial top-down si respectiv parţial bottom-up este recomandată în majoritatea cazurilor.

3. Arhitectura bazelor de date distribuite

Arhitectura desemnează ceea ce percepe fiecare dintre utilizatorii finali, operatori, personal tehnic, echipă de dezvoltare, distribuitori. Arhitectura se referă odată la stadiul logic si apoi la cel fizic. Ca urmare există o arhitectura logică şi una fizică.

Arhitectura logică vizează organizarea funcţionalităţilor dorite în termeni de clase, relaţii şi interacţiuni, adică descrie modul în care se structurează şi se organizează sistemul la nivel conceptual, pentru a asigura funcţionalitaţile cerute de utilizator.

Arhitectura fizică descrie organizarea sistemului în structura de implementare şi exploatare, adică defineşte modul în care se implementează fizic structurile definite în cadrul arhitecturii logice. De exemplu în ACCESS softul se realizează pe proiecte din care, după compilare şi editare rezulta componente executabile instalate pe echipamente (calculatoare sau periferice) conectate cum ar fi server, staţie de lucru, etc. Asigurarea independenţei fizice şi logice a datelor impune adoptarea unei arhitecturi de baze de date organizată pe trei niveluri:

- nivelul intern (baza de date fizică);
- nivelul conceptual (modelul conceptual, schema conceptuală);
- nivelul extern (modelul extern, subschema, vizualizarea).

Nivelul central este **nivelul conceptual**. Acesta corespunde structurii canonice a datelor ce caracterizează procesul de modelat, adică structura semantică a datelor fără implementarea pe calculator. Schema conceptuală permite definirea tipurilor de date ce caracterizează proprietăţile elementare ale entităţilor, definirea tipurilor de date compuse care permit regruparea atributelor pentru a descrie entităţile modelului şi legăturile între aceste entităţi, definirea regulilor pe care trebuie să le respecte datele etc**.**

Nivelul intern corespunde structurii interne de stocare a datelor. Schema internă permite descrierea datelor unei baze sub forma în care sunt stocate în memoria calculatorului. Sunt definite fişierele care conţin aceste date, articolele din fişiere, drumurile de acces la aceste articole etc. La nivel conceptual sau intern, schemele descriu o bază de date. La nivel extern schemele descriu doar o parte din date care prezintă interes pentru un utilizator sau un grup de utilizatori. Schema externă reprezintă o descriere a unei părţi a bazei de date ce corespunde viziunii unui program sau unui utilizator.

Modelul extern folosit este dependent de limbajul utilizat pentru prelucrarea bazei de date. Schema externă permite asigurarea unei securităţi a datelor. Un grup de lucru va accesa doar datele descrise în schema sa externă, iar restul datelor sunt protejate împotriva accesului neautorizat sau rău intenţionat.

4. Bazele de date distribuite ale Facebook-ului

Baza de date a Facebook-ului este responsabilă pentru procesarea unor cantități mari de date, numite "Big Data", care variază de la raportarea simplă la cele de inteligență artificială la cele mai mari măsurări și rapoarte executate, informație amplă aflată în centre de date distribuite localizate în diferite locații geografice. Acestea sunt procesate cu ajutorul serverelor de înaltă tehnologie. Scalabilitatea și fiabilitatea sunt cerințe obligatorii în gestionarea bazelor de date ale Facebook-ului (figura 3), deservește miliarde de cereri și este responsabilă pentru răspunsul la solicitările utilizatorilor în doar câteva secunde. Facebook se bazează pe platforma Hadoop, care este foarte potrivită pentru a face față textului nestructurat, jurnalelor și evenimentelor.

Hadoop cunoscut HDFS (Hadoop Distributed File System) este conceput pentru stocarea în siguranță a seturilor de date foarte mari. Acesta a folosit într-un cluster mare mii de servere stocate direct și execută aplicația utilizator. Distribuind spațiul de stocare și calcul pe multe servere, care oferă abilități sistemului de a scala dinamic, resursa poate crește la cerere. Un HDFS poate fi format din sute sau mii de mașini server, fiecare dintre care stochează o parte din datele sistemului de fișiere. Accentul se pune pe o capacitate mare de acces la date. Pe plan intern, un fișier este împărțit în unul sau mai multe blocuri și aceste blocuri sunt stocate într-un set de noduri de date (figura 3).

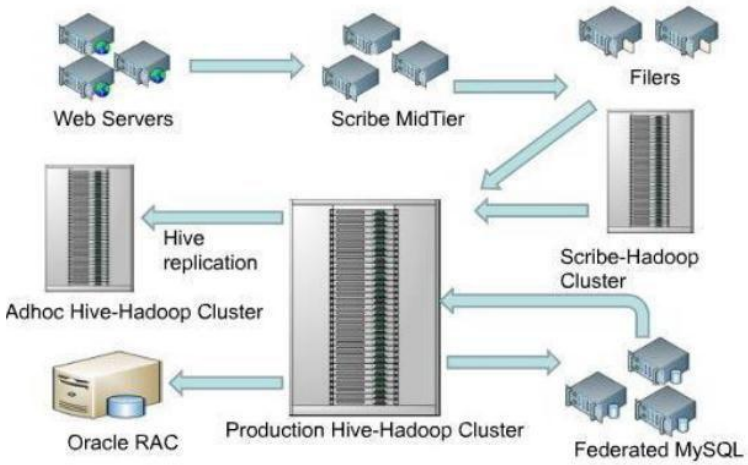

Fig. 3. Arhitectura de bază a sistemului Facebook

Tera Bytes și peta biți de date sunt procesate și analizate zilnic de centrele de date Facebook. Un mapper permite de a face o copie la toate datele ce sunt înregistrate în baza de date a sistemului. Toata informația este indexată și sortată după chei unice și trasmisă pentru a fi realizat un merge (conexiune) final a informației.

Concluzii

Bazele de date distribuite au fost create pentru a reflecta structura organizațională a companiei cât și pentru a permite fiecărei secții de servere de a-și controla mai aproape propriile distribuții de baze de date. În cazul folosirii bazelor de date distribuite partițiile care nu au fost folosite pot fi folosite pentru prelucrare pentru a asigura o performanță îmbunătățită. În dependență de necesitățile care apar, bazele de date distribuite pot fi împărțite în omogene și heterogene, fiecare cu caracteristicile și contribuțiile sale în sistemul de gestiune a bazelor de date ditribuite. Una dintre cele mai performante baze de date distribuite este sistemul Hadoop, creat și implementat de platforma Facebook. Acesta nu doar poate procesa cantități mari de informație dar oferă și o siguranță a datelor procesate.

Bibliografie

1. Studiu de caz: Facebook-Distributed-System-Case-Study-For-Distributed-System-Inside-Facebook-Datacenters, INTERNATIONAL JOURNAL OF TECHNOLOGY, 2014.

2. *Baze de date distribuite.* https://ro.wikipedia.org,. – 10 Noiembrie 2017.

3. Dorin Cârstoiu, *Sisteme de baze de date distribuite*, Bucuresti, 2013.

4. Adrian Runceanu, Baze de date. Universitatea "Constantin Brâncuşi", Târgu-Jiu, [Resursa electronica] – Regim de acces: http://www.runceanu.ro/adrian/wp-content/cursuri/bd2016/L10-BD_2016.pdf – 10 Noiembrie 2017.

5. Vitalie Cotelea, Unele aspecte de proiectare a bazelor de date distribuite [Resursa electronica] – Regim de acces: http://utm.md/meridian/2012/MI_1_2012/3_Art_Cotelea.pdf – 5 Noiembrie 2017.