11-ma'ruza: Katta hajmli ob’ektlarni tashkil qilish va boshqarish

Reja:

1. Katta hajmli multimedia ob’ektlar ularni boshqarish usul va vositalari

Katta hajmli multimedia ob’ektlar ularni boshqarish usul va vositalari

Zamonaviy axborot tizimlarida foydalaniladigan axborot (ayniqsa video, audio multimediali ma‘lumotlar) hajmini jadal o‘sishi oldimizga ushbu axborotni ma‘lumotlar bazasida saqlash va ularni boshqarishning yangi murakkab masalalarni qo‘yadi. Quyida katta hajmdagi multimedia ma‘lumotlarni tashkil qilishga, shuningdek ular bilan ishlash vositalari tahlil qilib chiqamiz.

Katta hajmli multimedia ma‘lumotlar to‘rtta asosiy xarakteristikaga ega bo‘ladi: hajm, xilma-xillik, tezlik, narhi.

1. Hajm. Insonlar va mashinalar tomonidan xosil qilinayotgan ma‘lumotlarning o‘sib borayotgan soni, axborot texnologiyalari infrastrukturalariga ma‘lumotlarni saqlash, ishlov berish va taqdim etishida yangi talablarni qo‘yadi.
2. Xilma-xillik. Turli strukturalarda taqdim etilgan ma‘lumotlar xilma-xil bo‘ladi. Bular kredit kartalar bo‘yicha operatsiyalar bo‘ladimi, ilmiy tadqiqot natijalari bo‘ladimi, fotografik suratlar bo‘ladimi, video va audio ma‘lumotlar bo‘ladimi barchasi o‘ziga xos ishlov va saqlash shart sharoitlarini talab etadi.
3. Tezlik. Tezlik deganda nafaqat ma‘lumotlarning ma‘lumotlar bazasiga kelib tushish tezligi, balki bu ma‘lumotlar bazasidan olinadigan ma‘lumotlarni chiqarib lish tezligi ham anglanadi.
4. Narhi. Katta hajmdagi ma‘lumotlar – qimmat resurs hisoblanadi. Ma‘lumot dolzarbligi, foydaliligi va ichidagi kontentiga qarab uning qanchalik qimat baholigi aniqlanadi.

Katta hajmli ma‘lumotlarni saqlash

Yangi axborot texnologiyalari, qurilmalari va kommunikatsiya vositalarining paydo bo‘lishi bilan, insonlar tomonidan ishlab chiqarilayotgan ma‘lumotlar soni ham geometrik progressiya bo‘yicha ortib bormoqda. Shu bilan birga ma‘lumotlarning foydaliligi koeffitsienti juda past bo‘lishi mumkin. Aksariyati befoyda, bekorchi ma‘lumotlardan iborat. Shundan kelib chiqadiki, barcha ishlab chiqarilgan ma‘lumotlardan oldin izchil ishlov berilgandan keyingina foydalanish mumkin bo‘ladi.

Katta hajmdagi ma‘lumotlar an’anaviy kompyuter tizimlari tomonidan amalga oshirib bo‘lmaydigan ishlov berishni talab qiladi.

Katta hajmdagi ma‘lumotlar turli tizim va dasturlar tomonidan yaratilgan axborotni o‘z ichiga oladi:

- Qora quti: vertolyot, samolyotlar, dengiz va kosmik kemalarning axborot ta’minoti qismiga kiradi. Uning vazifasiga boshqaruv ob’ekti xarakteristikalari to‘g‘risidagi axborotni, ekipaj ovozini yozib borish kiradi.
- Ijtimoiy multimedia ma‘lumotlar: ijtimoiy tarmoqlar orqali tarqatilatgan axborot.
- Fond birjalari: kompaniyalar orasidagi oldi-sotdi muomalalar to‘g‘risidagi ma‘lumotni saqlash.
- Energotizimlar: bunday ma‘lumotlar energetik tarmoqning uzellari va kuchlanishlarini o‘z ichiga oladi.
- Transport tizimi: modellar, xarateristkalar, masofalar, GPS media ma‘lumotlar – transport va yo‘l tarmog‘i haqidagi barcha ma‘lumotlarni qamraydi.
- Qidiruv tizimlari: turli ma‘lumotlar bazasidan izlash ishlari.

Natijada, katta hajmdagi ma‘lumotlar katta hajmga, katta tezlikka. ega bo‘lgan uch xil tipga ajratiladi:

- Strukturalangan ma‘lumotlar – relyatsion MB
- Yarim strukturalangan ma‘lumotlar - XML-fayllar

• Strukturalanmagan ma‘lumotlar – tasvir, video, audio, Word, PDF, Text formatdagi fayllar.

MapReduce taqsimlangan ma‘lumotlar modeli

MapReduce dasturiy tizim Google kompaniyasi tomonidan yaratigan bo‘lib, Google File System taqsimlangan fayl tizim asosida amalga oshirilgan. Bu maxsulot Google kompaniyasining xususiy mulki hisoblanadi va faqat shu kompaniyaning dasturiy maxsulotlarida ishlatiladi.
Mazkur modelning boshqa realizatsiyalari ham mavjud. Erkin tarqatuvda – Apache Hadoop loyihasida ishlab chiqilgan Hadoop MapReduce maxsulot ham bor. Mazkur texnologiya HDFS (Hadoop Distributed File System) taqsimlangan fayl tizimidan foydalanishga asoslanadi. MapReduce erkin va ochiq dasturi maxsulot hisoblanadi.
MapReduce modelida barcha hisoblash muolajalari "kalit-qiymat" deb atalgan kiruvchi juftlik to‘plami ustida amalga oshiriladi. Har bir hisob natijasida "kalit-qiymat" chiquvchi juftlik to‘plami xosil bo‘ladi.
MapReduce muhitida hisoblashlarni amalga oshirishda ikkita muhim funksiyalardan foydalaniladi: «Map» va «Reduce».
MapReduce amalga oshiriladigan loyihalar quyidagi xarakteristikali taqsimlangan klasterlar muhitida ishlashga yo‘naltirilgan:
• Dasturlarni bajarish muhiti uzellari odatda Linux operatsion tizimli umumiy kompyuterlardan iborat;
• Klaster yuzlab yoki minglab kompyuterlardan iborat;
• Ma‘lumotlarni saqlash uchun qimmat bo‘lmagan disk qurilmalaridan foydalaniladi;
• Mazkur disklarda joylashgan ma‘lumotlarni boshqarish uchun taqsimlangan fayl tizimidan foydalaniladi;
MapReduce texnologiyasi katta hajmdagi ma‘lumotlarga minglab kompyuterlar orqali ishlov berishga mo‘ljallangan. Shuning uchun, unda albatta alohida kompyuterlarning ishdan chiqishi holatlariga turg‘unlik bo‘lishi kerak.
Agar ma’lum uzel so‘rovlarga belgilangan vaqt oralig‘ida javob bermasa, tizim uni ishdan chiqqan deb hisoblaydi va boshqa uzelni jalb qiladi.
MapReduce texnologiyasi, minglab kompyuterlarni jalb qilish bilan, turg‘unlik va kengayish talablarini ta’minlay oladi. MapReduce avvaldan strukturalanmagan (matnli) ma‘lumotlar bilan ishlashga mo‘ljallanganiga qaramay, undan katta hajmdagi strukturalangan ma‘lumotlarga ishlov berishda foydalanish mumkin.

Hadoop texnologiyasi

Apache Hadoop maxsuloti oddiy qurilmalarda joylashtirilishi mumkin bo‘lgan erkin Java-platforma hisoblanadi. Loyiha Google File System materiallarini qayta ishlash natijasida tug‘ilgan bo‘lib, katta klasterlarda ishlaydigan taqsimlangan ilovalarni ishlashiga yordam beradi.
Mazkur texnologiya ilovalarga ma‘lumotlar bilan ishonchli va tez ishlashni ta’minlaydi. Maxsulotda MapReduce deb tanilgan hisoblash paradigmasi qo‘llangan. Shunga ko‘ra ilova ko‘p sonli kichik masalalarga ajratilgan bo‘lib, masalalarning har biri ixtiyoriy uzelda ishlashi mumkin. Qo‘shimcha, ma‘lumotlarni klasterning hisoblash uzellarida saqlashga mo‘ljallangan, taqsimlangan fayl tizimidan foydalaniladi. Bu klasterni juda yuqori agregatlashtirilgan o‘tkazish xususiyatiga erishitiradi.
Mazkur tizimlar ilovalalarni oson kengayishiga (minglab uzellarni petabayt ma‘lumotlarga ishlov berishiga) yo‘l qo‘yib beradi.
Hadoop texnologiya Facebook, Twitter, Rackspace i eBay kabi veb-loyihalarda foydalaniladi. Shuningdek IBM, EMC, Dell i Oracle kabi dasturiy maxsulotlarda ko‘llaniladi.
Hadoop ning asosiy texnik xarakteristikalariga quyidagilar kiradi:
• Kengayuvchanligi: platforma petabayt (1015) ma‘lumotlarni saqlash va ishlov bera olishi bilan chiziqli kengayishi mumkin;
• Ishdan chiqishga turg‘unligi: barcha saqlanayotgan ma‘lumotlar keragidan ortiq, barcha uzilib qolgan ishlov berish masalalari qaytadan boshlanadi;
• Krossplatformalik: Hadoop kutubxonalari asosan Java tilida yozilgan bo‘lib, Java mashinani qo‘llab quvvatlaydigan ixtiyoriy operatsion tizim ostida ishlashi mumkin.
• Masalalarni avtomatik tarzda parallellashtirish: Hadoop texnologiya dasturchilarga ko‘rinib turadigan "shaffof" abstraksiyalar xosil qiladi. Shu bilan ularni ma‘lumotlarni parallel ishlov berish natijalarini loyihalash, boshqarish va agregatsiya qilish ishlaridan forig‘ qiladi.
Hadoop dan foydalanishning afzalliklari quyidagilarda namoyon bo‘ladi:
• Qayshqoqlik: strukturalangan va strukturalanmagan ma‘lumotlar tipini saqlash va tahlil qilish;
• Samaraliylik: ko‘p hollarda terabayt ma‘lumotlarni saqlash va ularga ishlov berish boshqa mavjud texnologiyalarga

nisbatan arzon narhga tushadi.
• Klasterni arzon xosil qilish: Hadoop-klasterlarni xosil qilish uchun qimmat server apparat ta'minoti talab qilinmaydi.
• Nisbatan yengil moslashuvchanlik: Hadoop keng va aktiv rivojlanaytgan ekosistemaga ega;
• Minimal risk: platforma yadrosini noto'g'ri ishlashi bilan bog'liq risklarning minimalligi. Hozirgi kunda Hadoop platformadan petabayt ma'lumotlardan foydalanishda ishlatiladi;
• «Open Source» litsenziya: Hadoop platformani qo'llash va egalik qilishning arzon narhdaligi;
• «Open Source» litsenziya: Hadoop platformani qo'llash va egalik qilishning arzon narhdaligi;
• Platformadan foydalanadigan ishlab chiqaruvchilar sonining ko'pligi.

Forrester Research kompaniyasi analitiklarining fikricha, Apache Hadoop platforma barcha katta kompaniyalarning AT-infrastrukturalari uchun standart vazifasini o'taydi.

NoSQL yondashuv

NoSQL atama "nafaqat SQL " yoki "SQL emas" deganini anglatadi. Mazkur atama 2009 yildan boshlab, interet-texnologiyalar va ijtimoiy tarmoqlarning rivoji ma'lumotlarni saqlash va ularga ishlov berishga yangicha yondashuvlarni keltirib chiqarganda, mashhurlashdi. Bu paytga kelib, dasturchilar an'anaviy relyatsion ma'lumotlar bazasi o'ta qimmatga tushayotgani yoki yetarlicha tez ishlamayotganligi kabi masala va muammolarga ro'para kelgan edilar.
Shuni aytib o'tish kerakki, NoSQL-yechim relyatsion ma'lumotlar bazalaridan butunlay voz kechishni yoki ular almashtirishni ko'zda tutmaydi.
Afzalliklar sifatida quyidagilarni aytish mumkin:
• Kengayuvchanlik: mavjud an'anaviy MBBT lar uchun gorizontal kengayish masalasi odatda juda qiyin va qimmat hisoblanadi. Ko'p NoSQL-yechimlar shu sababga ko'ra loyihalashtirilgan.
• Tezlik: hisoblash samarasi – muhim omillardan hisoblanadi. Ko'p masalalar uchun an'anaviy MBBT relyatsion model, tranzaksiyalar, ishonchlilik va h.k. kabi xususiyatlarining hammasi bir payitda kerak bo'lavermaydi. Bu xususiyatlarning hammasi yoki ba'zilaridan voz kechish NoSQL katta tezlikka erishishiga olib keladi.
• Replikatsiyalar: serverni ishdan chiqishi yoki tarmoqqa ulanib bo'lmaslik ehtimolligi ixtiyoriy axborot tizimidan ishonchliylik xususiyatini talab qiladi. Barqaror ishlashning asosiy usuli – replikatsiya. Ma'lumotlar bazasinireplikatsiya rejimida ishlashga o'tishi NoSQL-yechimlarning imkoniyatlaridan biri.
• Yaratish va boshqarishning oddiyligi. O'rnatish va sozlash masalalari , yana qo'shimcha NoSQL-yechimlarni qo'llash, relyatsion MB ga ko'ra, oddiroq va kam harajat bilan amalga oshiriladi. Shuning uchun ishlab chiqish va tadbiq etish tezligi muhim omillardan sanalgan loyihalarda ko'pincha NoSQL-tizimlar tanlanadi.
Ba'zi tipdagi masalalar uchun ma'lumotlarni taqdim etishning relyatsion modeli har doim ham eng yaxshi usul hisoblanavermaydi.
Ilovalarni ishlab chiqishda relyatsion modelni ishlatilayotgan ma'lumotlar modeliga akslantiruvchi alohida oraliq ob'ektlardan foydalanish oddiy holga aylangan. Bunday holat loyiha tannarhini oshirib yuboradi va tizimni murakkablashtirib yuborishi mumkin.
NoSQL texnologiyasi ma'lumotlar modelining keng to'plamini taqdim etadi. Konkret masala uchun mos modelni tanlash kifoya qiladi: hujjat ko'rinishadgi ma'lumotlar modeli, maydonlardan tashkil topgan ma'lumotlar, "kalit-qiymat" yozuvlar, graflar va h.k.
Hujjatga yo'naltirilgan MBBT maydonlardan iborat hujjatlar kolleksiyasi ko'rinishidagi ma'lumotlarni saqlaydi. An'anaviy MB bunday ma'lumotlarni o'zaro bog'langan jadvallarda saqlaydi: asosiy ma'lumotlar yuitta jadvalda, qo'shimcha maydonlar bog'langan boshqa jadvallarda saqlanadi. Shu bilan birga hujjatga-yo'naltirilgan MB murakkab so'rovlar qilib bo'lmaydi. Bunday ma'lumotlarda hujjatlar bog'lanish bo'lmaydi.
Graflarga-yo'naltirilgan MB. Bunday MB graf ko'rinishida berilgan ma'lumotlarni samarali saqlaydi. Ular mohiyatlar to'plami va ularning o'zaro munosabati orasidagi bog'lanishlarni saqlashga ideal to'g'ri keladi.
Misol tariqasida ijtimoiy graflar, tizim ob'ektlari orasidagi bog'lanishlarni olish mumkin.

R dasturlash tili

R dasturlash tili universal til bo'lib quyidagi sohalarda foydalanish uchun ishlab chiqilgan: ma'lumotlarni tahlili, klassik statistik testlar, yuqori darajadagi grafika.
R tili katta hajmli ma'lumotlar sohasida foydali instrument hisoblanib, IBM SPSS, InfoSphere, Mathematica ga

qo‘shilgan.
Mazkur til ko‘proq statistikaga mo‘ljallangan. R tili kuchli skript tillar oilasiga kiradi. Unda matnga ishlov berishda doimiy ilovalardan foydalaniladi. Har xil ko‘rinishdagi va tartiblanmagan katta hajmdagi ma‘lumotlarga ishlov berishda R tili imkoniyatlaridan foydalanish mumkin.
Yana muhim xususiyatlaridan biri – tekin va erkin tarqatilishi mumkin. R tili ochiq kodga ega.
Kamchiliklarida shuni aytish mumkinki, R platforma ma‘lumotlarni saqlash joyi emas. Ma‘lumotlarni boshqa ilovada kiritib, keyin uni R muhitiga import qilinishi kerak.

11-mavzuga doir savollar:

1. Katta hajmli ob’ektlarga ta’rif bering.
2. MapReduce nima?
3. NoSQL texnologiyasini tushuntirib bering.
4. Katta hajmli ob’ektlarni boshqarish usul va vositalarini aytib bering.
5. Hadoop texnologiyasini tushuntirib bering
6. R dasturlash tili asosiy vazifalarini aytib bering.