18-§. Katta hajmdagi ma’lumotlarni qayta ishlash

Reja:

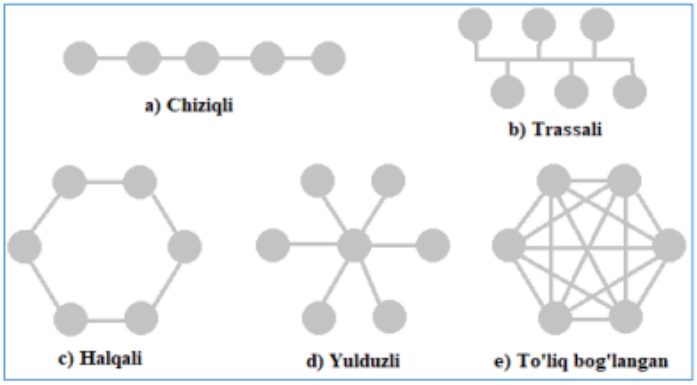
1. *Katta hajmdagi ma’lumotlarni yaratish va boshqarish usullari.*
2. *Katta ma’lumotlarni boshqarish arxitekturasini yaratish.*

Bugungi kunda zamonaviy ilm - fan taraqqiy etishi, dunyo bo’ylab ma’lumotlarni raqamlashtirish jarayonining tezlashishi katta va kichik serverlarda hamda shaxsiy qurilmalarda hozirgi zamonaviy texnologiyalar yordamida to’plash va qayta ishlash murakkab bo’lgan katta hajimli turli xil ma’lumotlar oqimi (Big Data) ning hosil bo’lishiga sabab bo’lmoqda. Bu katta hajimli turli xil ma’lumotlar oqimi ularni saqlash va qayta ishlash bilan bog’liq bir qancha muammolarni hosil qilishi bilan birgalikda, ulardan samarali foydalanish ya’ni uni to’liq va to’g’ri tahlil qilish malumotlar ishonchligini oshirib, insonlarga barcha sohalarda to’g’ri qarorlar qabul qilishda katta imkoniyatlarni ochib berish imkoniyatiga ega.

Big Dataning asosiy xususiyati ya’ni katta hajimli ma’lumotlarni tezkorlik bilan qayta ishlash va tahlil qilish samaradorligini oshirishga qaratilgan bo’lib, tadqiqot natijalari odatiy yakka hisoblash tizimlaridan foydalanish kutilgan natijalarni bermasligini ko’rsatmoqda.

Ma’lumotlarning katta hajimda ekanligini va uni yagona kompyuter yoki serverda qayta ishlash imkoniyati pastligi hisobga olinsa, eng yuqori natija beruvchi yondashuv taqsimlangan hisoblash mexanizmlari orqali ma’lumotlarni yig’ish va qayta ishlashdir. Ma’lumotlarning katta hajimda ekanligi sababli taqsimlangan hisoblash tizimlarida jarayonlarni to’g’ri tashkil etish tizim samaradorligiga ijobiy tasir etadi. Ya’ni Big Datani qayta ishlashda taqsimlangan hisoblash tizimining samaradorligi tizimning bir nechta xususiyat (omil)lariga bog’liq bo’ladi. Shunday taqsimlangan hisoblash tizimi samaradorligiga chambarchas bog’liq bo’lgan muhim omillardan biri - bu tizim abonentlari

o’rtasidagi jarayonlararo aloqa va sinxronizatsiya. Ayniqsa, katta hajimdagi turli xillik xususiyatiga ega ma’lumotlarni bir serverdan ikkinchisiga to’qnashuvlarsiz va yo’qotishsiz uzatish tizim samaradorligining qiyin bo’g’inidir. Bu, shuningdek, taqsimlangan tizimlarda parallel ishlash jarayoniga ham tasir qiladi. Shuning uchun Big Datani qayta ishlash uchun taqsimlangan hisoblash tizimining o’zaro bog’lanish sxemasini tanlash ish samaradorligiga ta’sir qiladi. Taqsimlangan hisoblash tizimlarining bir qancha o’zaro bog’lanish sxemalari mavjud bo’lib, ulardan asosiylari 4.1-rasmda ko’rsatilgan.



4.1-rasm. Taqsimlangan hisoblash tizimlarining oddiy o’zaro bog’lanish sxemalari.

Ushbu rasmda aylana hisoblash tugunini, chiziq esa tugunlar orasidagi to’g’ridan-to’g’ri aloqa kanalini ifodalaydi chiziqli o’zaro bog’lanish sxemasi - hisoblash tugunlari bitta chiziqqa joylashtirilgan va ulangan (1-a rasm). Marshrutlash oddiy va topologiyani rekursiv sifatida ko’rish mumkin. Biroq, qo’shni bo’lmagan har qanday ikkita tugun o’rtasidagi aloqa boshqa tugunlarning yordamiga muhtoj, har qanday oraliq tugundagi nosozlik butun tizimni buzadi. Sxema oddiy va arzon, ammo Big Datani qayta ishlash uchun yuqori ishlash yoki ishonchlilikni yarata olmaydi hamda tizim miqyosi kattalashgani sari unumdorlik tez pasayadi. Trassali o’zaro bog’lanish sxemasida har qanday ikkita tugun o’rtasida to’g’ridan-to’g’ri ulanish mavjud (1-b rasm). O’zaro ulanish umumiy trassa orqali amalga oshiriladi. Bu bog’lanishlardagi murakkablikni sezilarli

darajada kamaytiradi. Biroq, trassadan katta ma’lumotlarning oqib o’tishini hisobga olsak har bir abonent ma’lumotlarni uzatishda aloqa kanalini uzoq vaqt band qiladi. Natijada bir vaqtda uzatilgan xabarlar to’qnashuvi sababli o’z manziliga yetib bormaydi. Halqali o’zaro bog’lanish sxemasi - bu chiziqli o’zaro bog’lanish sxemasining ikki uchi o’rtasida qo’shimcha ulanish bilan takomillashtirilgan sxema (1-c rasm). Bu o’zaro ulanish masofasini 2 martaga kamaytiradi. Biroq, asosiy xususiyatlar hali ham bir xil, Big Datani qayta ishlash uchun yuqori ishlash yoki ishonchlilikni yarata olmaydi hamda tizim miqyosi kattalashgani sari unumdorlik tez pasayadi. Yulduzli o’zaro bog’lanish sxemasi, barcha abonentlarni birlashtiruvchi markaziy abonentga ega (1-d rasm). Har bir aloqa kanali faqatgina ikkita abonentga xizmat qilganligi uchun ma’lumotlar to’qnashuvi kuzatilmaydi. Bundan tashqari o’zaro ulanish masofasi 2 ga teng bo’lib, markaziy kommutator tugunining yordami bilan kollektiv aloqani oson amalga oshirishni qo’llabquvvatlaydi va rekursiv kengayish imkonini beradi. Lekin ma’lumotlar hajmining kattaligi, ularni qayta ishlash va abonentlar o’rtasidagi sinxronzatsiyani amalga oshirish vazifalarining ko’pligi sababli markaziy kompyuter doimo yuklanish bilan ishlaydi va markaziy kompyuterning har qanday nosozlikka uchrashi o’zaro aloqani yo’qotadi. Bu Big Datani qayta ishlash ishonchliligiga zarar yetkazadi. To’liq bog’langan o’zaro bog’lanish sxemasi. Unda har qanday ikkita hisoblash tugunlari o’rtasida to’g’ridan-to’g’ri aloqa mavjud (1-e rasm). O’zaro bog’lanish masofasi 1 ga teng va har bir aloqa kanal faqatgina ikkita hisoblash tugunini bog’laydi. Natijada bular Big Datani qayta ishlash ishonchliligini oshiradi.

Bugungi dunyo ma’lumotlarga asoslangan dunyo bo’lib, hayotimizning barcha jabhalariga kirib borgan texnologiyalarning tez o’sishi natijasida ma’lumotlar katta hajmda ishlab chiqarilmoqda. Turli shakllarda ishlab chiqarilgan ma’lumotlarning doimiy hajmidan ma’noli tushunchaga ega bo’lish uchun ma’lumotlarni qayta ishlashning yangi usullari ishlab chiqilishi va takomillashtirilishi kerak. Mashinani o’rganish texnologiyalari katta hajmdagi

ma’lumotlarni qayta ishlash va undan qiymat olish uchun istiqbolli yechimlar va potentsial usullarni taqdim etadi. Katta ma’lumotlarni qayta ishlashda mashinani o’rganish usullarini qo’llash mashinani o’rganish algoritmlari va usullarining umumiy ko’rinishi taqdim etilmoqdi.

Katta ma’lumotlar deganda o’rtacha vaqt ichida an’anaviy IT, dasturiy va apparat vositalaridan foydalangan holda tushunish, qo’lga olish, boshqarish yoki tahlil qilish qiyin bo’lgan ma’lumotlar to’plami tushuniladi. Boshqacha qilib aytadigan bo’lsak, Katta ma’lumotlar relyatsion an’anaviy metodologiyalar yordamida tahlilni samarali amalga oshirishga to’sqinlik qiladigan hajmli, olish tezligi yoki formatli ma’lumotlar yoki gorizontal kattalashtirishning muhim usullaridan foydalangan holda samarali qayta ishlanishi mumkin bo’lgan ma’lumotlar sifatida tavsiflanadi. Katta ma’lumotlar tushunchasini u bilan bog’liq bo’lgan turli xil V-larni tushunish orqali aniqroq aniqlash mumkin. Bu V.lar katta ma’lumotlarni boshqarish tizimlari duch keladigan asosiy o‘lchovlar (qiyinchiliklar). Ushbu o’lchamlar quyidagicha aniqlanadi: Terabaytdan zettabaytgacha bo’lgan soniyada ishlab chiqarilgan juda katta ma’lumotlar. Uni tahlil qilish uchun tegishli vositalarni ishlab chiqish uchun saqlash va qayta ishlash modellarini qayta ko’rib chiqish kerak.

Taqsimlangan tizimlar butun dunyo bo’ylab ma’lumotlar bazalarida ma’lumotlarni saqlash va tahlil qilish uchun katta ma’lumotlarda qo’llaniladi. Bu atama talablarni qondirish uchun ma’lumotlarni yaratish va qayta ishlash tezligini anglatadi. An’anaviy tahlillar real vaqt rejimidagi ma’lumotlarga qaramlikning ortishi bilan shubhalanadi, chunki ma’lumotlar juda katta va doimiy harakatda. Ma’lumotlar turli manbalardan kelib chiqishi va turli shakllarga ega bo’lishi mumkinligi sababli, asosiy muammo - ma’lumotlar formatining mos kelmasligi. Endi ma’lumotlar bir nechta formatlarda mavjud, jumladan, tuzilgan, yarim tizimli,
tuzilmagan va hatto murakkab tuzilgan ma’lumotlar.

Ma’lumotlar formatlarining xilma-xilligi tufayli an’anaviy tahliliy usullar katta ma’lumotlarni boshqara olmaydi. Tahlil uchun ma’lumotlarni tayyorlash uchun samarali texnikani loyihalash zarur bo’lib, ular juda katta vaqt va kuch talab qiladi. Olingan ma’lumotlarning sifati sezilarli darajada farq qiladi. U ma’lumotlarning noto’g’riligini, shovqinlarini, anormalliklarini va boshqalarni ko’rsatadi. Bu tahlilning to’g’riligiga ta’sir qiladi. Haqiqiylikni saqlash tizimda nuqsonli ma’lumotlarni to’plamaydi. Qiymat haqiqatga ta’sir qilishi mumkin. O‘zgaruvchanlik katta ma’lumotlarning yangi o‘lchami tomonidan kiritilgan.“O‘zgaruvchanlik” atamasi ma’lumotlar oqimi tezligining o‘zgarishini bildiradi. Katta ma’lumotlarning tezligi ko’pincha tartibsiz bo’lib, vaqti-vaqti bilan

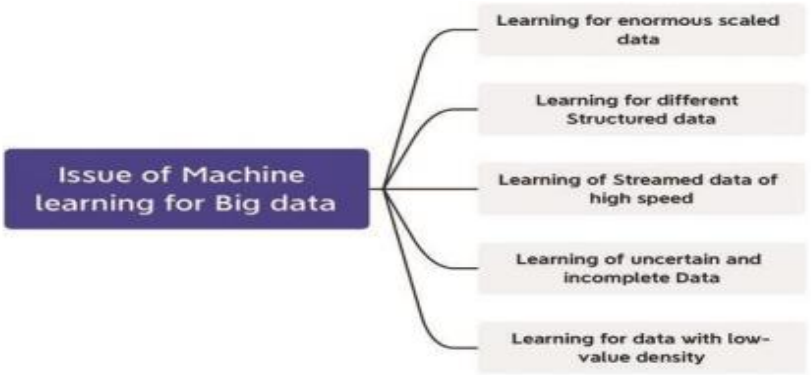
cho’qqilar va pastliklar bo’ladi.

"Ma’lumotlarning haqiqiyligi" va "ma’lumotlarning haqiqiyligi" atamalari ko’pincha xuddi shunday qo’llaniladi. Ular bir xil tushuncha emas, lekin ular o’xshashdir. Haqiqiylik deganda ma’lumotlarning to’g’riligi va ulardan maqsadli foydalanish bo’yicha aniqligi tushuniladi. Boshqacha qilib aytadigan bo’lsak, ma’lumotlarning to’g’riligi bilan bog’liq muammolar bo’lmasligi mumkin, ammo

tushunilmagan bo’lsa, u haqiqiy bo’lmasligi mumkin. Katta ma’lumotlarning o’zgaruvchanligi haqida gap ketganda, tashkilotlarda har kuni qo’llaniladigan tuzilgan ma’lumotlarni saqlash siyosatini osongina eslash mumkin. Saqlash muddati tugaganidan keyin uni osongina yo’q qilish mumkin. Qiymat Oracle tomonidan katta ma’lumotlarning belgilovchi xususiyati sifatida taqdim etilgan.

Hadoop va MapReduce bir-birini almashtirib bo’lmaydigan atamalar emas; Hadoop aslida MapReduce kontseptsiyasini amalga oshirishdir. MapReduce - bu katta hajmdagi ma’lumotlarni qayta ishlash uchun bo’lish va egallash texnikasidan foydalanadigan model. Hadoop ikkita tugundan iborat: master va ishchi, MapReduce esa ikkita asosiy bosqichni bajaradi: Map va Reduce. Asosiy

tugun kiruvchi ma’lumotlarni kichik muammolarga ajratadi, ular keyinchalik ishchi tugunlarga tayinlangan xarita bosqichida bo’ladi. Keyin barcha kichik muammolarning natijalari asosiy tugun tomonidan qisqartirish bosqichida birlashtiriladi katta ma’lumotlarni qayta ishlash uchun mashinani o’rganish yondashuvlarining eng muhim muammolarini ko’rib chiqadi.



4.2-rasm: Katta ma’lumotlarni o’rganish usullari.

Texnologik taraqqiyot tufayli biz bilan shug’ullanadigan ma’lumotlar miqdori

kundan-kunga o’sib bormoqda. 2017-yil noyabr oyida Google har kuni taxminan petabayt ma’lumotni qayta ishlashi aniqlandi va bu oxir-oqibatda ma’lumotlarning

o‘zaro bog‘liqligini tasdiqlaydi. Ma’lumotlar hajmi katta ma’lumotlarning aniq asosiy atributidir, bu esa muhim muammo tug‘diradi. Ushbu qiyinchilikni hal qilish uchun taqsimlangan va parallel ramkalar hisoblash afzal bo’lishi kerak. Hozirgi vaqtda juda ko’p turli xil ma’lumotlar mavjud. Chiziqli bo’lmagan va yuqori h o’lchovli ma’lumotlarga olib kelishi mumkin bo’lgan uchta turdagi ma’lumotlar tuzilgan, tuzilmagan va yarim tizimli ma’lumotlardir. Ushbu katta ma’lumotlar to’plamidan o’rganish juda katta muammo bo’lib, ma’lumotlarning murakkabligini oshirishga olib keladi. Natijada, ushbu to’siqni bartaraf etish uchun

ma’lumotlar integratsiyasi talab qilinadi. Muayyan vaqt oralig’ida ishni yakunlash kerak bo’lgan turli xil tadbirlar mavjud.

Katta ma’lumotlarning tezligi uning eng muhim xususiyatlaridan biridir. Agar ish ma’lum bir vaqt ichida tugallanmasa, ishlov berish natijalari o’z qiymatini o’zgartirishi mumkin, agar foydasiz bo’lsa. Natijada, katta hajmdagi ma’lumotlarni o’z vaqtida qayta ishlash juda muhim va qiyin vazifadir. Qiyinchiliklarni bartaraf etish uchun onlayn ta’lim strategiyasidan foydalanish kerak. Ilgari ma’lumotlar aniqroq bo’lgan mashinani o’rganish algoritmlariga etkazilgan. Chunki o’sha paytda natijalar to’g’ri bo’lgan. Biroq, bugungi kun ma’lumotlari turli xil manbalardan olinganligi sababli noaniq va to’liq emas. Natijada, katta ma’lumotlar tahlilida qorong’ulik mashinani o’rganish uchun muhim masaladir.

Ma’lumotlar sifatining noaniqligi va to’liqsizligini hal qilish va boshqarish muhimligini ta’kidlash uchun biz katta ma’lumotlar bilan o’rganish uchun to’rtinchi asosiy muammo sifatida haqiqatni sanab o’tamiz. Masalan, simsiz tarmoqlarda noaniq ma’lumotlar shovqin, so’nish, soya va boshqa omillar natijasida yaratilgan ma’lumotlardir. Bu qiyinchilikni yengish uchun tarqatishga asoslangan usuldan foydalanish kerak. Mashinani o’rganish asosan katta ma’lumotlar tahlilida tijorat maqsadlarida katta hajmdagi ma’lumotlardan mazmunli ma’lumotlarni olish uchun ishlatiladi.

Ma’lumotlarning qiymati uning eng muhim xususiyatlaridan biridir. Qiymat zichligi past bo’lgan katta hajmdagi ma’lumotlardan mazmunli qiymatni topish juda qiyin. Shunday qilib, bu katta ma’lumotlar tahlilida mashinani o’rganish uchun katta muammodir. Ushbu qiyinchilikni hal qilish uchun ma’lumotlarni qazib olish vositalari va ma’lumotlar bazasi bilimlarini kashf qilishdan foydalanish kerak. Ushbu texnologiyalar katta hajmdagi ma’lumotlardan muhim ma’lumotlarni olish uchun istiqbolli echimlarni taqdim etishi sababli o’ynaydi.

Machine Learning-ning katta ma’lumotlar tahlilidagi turli muammolarini ehtiyotkorlik bilan hal qilish kerak. Mashinani o’rganish bo’yicha ko’plab yechimlar mavjudligi sababli, ularning barchasi o’qitish uchun juda ko’p ma’lumotlarni talab qiladi. Mashinani o’rganish modellari aniq bo’lishi uchun tuzilgan, tegishli va aniq tarixiy ma’lumotlarni o’rganishni talab qiladi. Boshqa qiyinchiliklar ham bo’lishi mumkin, ammo bu imkonsiz emas.

Katta ma’lumotlarni o’rganish va tahlil qilish rivojlanish va kengayishda davom etadigan murakkab va juda muhim yo’nalishdir. Har yili inson tobora ko’proq ma’lumot ishlab chiqaradi va uning asosiy qismi tartibsiz shaklda bo’ladi. Shu turdagi ma’lumotlarni tahlil qilishni o’rganish, alohida ma’lumotlar to’plamlari orasidagi aloqalarni aniqlash davrimizning eng muhim vazifasidir.

Katta ma’lumotlar bilan ishlash deyarli barcha sohalarda zarur: fan, tibbiyot, biznes. Big Data ni qayta ishlash ayniqsa biznes yechimlari uchun foydalidir. Ushbu sohada saralanmagan ma’lumotlarni tezda qayta ishlash qobiliyati muvaffaqiyat omillaridan biridir. BIG DATA—bu uchta operatsiyani bajarishga mo’ljallangan texnologiyalar to’plami:

1.“Standart”ssenariyga nisbatan katta hajmdagi ma’lumotlarni qayta ishlash.

2. Juda katta hajmdagi tez keladigan ma’lumotlar bilan ishlashni o’rganish.

3.Tartiblangan va tartiblanmagan ma’lumotlar bilan parallel va turli mezonlarga ko’ra ishlay olish.

Ushbu “ko’nikmalar”insonning cheklangan idrokidan chetda qolgan yashirin narsalarni ochib berishga imkon beradi deb ishoniladi. Bu bizning hayotimizning ko’plab sohalarini: hukumat, tibbiyot, telekommunikatsiya, moliya, transport, ishlab chiqarish va boshqalarni optimallashtirish uchun misli ko’rilmagan imkoniyatlarni beradi. Jurnalistlar va marketologlar Big Data iborasini shu qadar tez-tez ishlatib kelganliklari sababli, ko’plab mutaxassislar ushbu atamani noaktual deb hisoblashadi va undan voz kechishni taklif qilishadi. Katta ma’lumotlarni aniqlovchi xususiyatlari sifatida, ularning fizik hajmidan tashqari va uni tahlil qilish murakkabligi ham sanab o’tiladi.

Mashinalarning vazifasi - tezkor tartiblash uchun kiruvchi ma’lumotlarning muhimlik darajasini aniqlash. Katta ma’lumotlar texnologiyasining ishlash tamoyili foydalanuvchini har qanday obyektyoki hodisa to’g’risida maksimal darajada xabardor qilishga asoslanadi. Ma’lumotlar bilan tanishishning maqsadi to’g’ri qaror qabul qilish uchun ijobiy va salbiy tomonlarni ko’rib chiqishga yordam berishdir. Aqlli mashinalarda bir qator ma’lumotlar asosida kelajak modeli quriladi, so’ngra turli xil variantlar simulyatsiya qilinadi va natijalar kuzatiladi.

Zamonaviy analitik agentliklar g’oyani, taxminni yoki muammoni hal etishda millionlab shunga o’xshash simulyatsiyalarni qo’llaydilar. Jarayon avtomatlashtirilgan.Katta ma’lumot manbalariga quyidagilar kiradi:

- Internet bloglar, ijtimoiy tarmoqlar, saytlar, OAV va turli forumlar;
- Arxiv, tranzaksiyalar, ma’lumotlar bazasi;
- o’qish moslamalari -meteorologik qurilmalar, uyali aloqa dachchiklari va boshqalar.

Yuqoridagi ta’riflarga asoslanib, katta ma’lumotlar bilan ishlashning asosiy tamoyillari quyidagilar:

1. Gorizontal mashtablilik. Bu katta ma’lumotlarni qayta ishlashning asosiy tamoyilidir. Yuqorida aytib o’tganimizdek, kundan kunga katta ma’lumotlar ortmoqda. Shunga ko’ra, ushbu ma’lumotlar tarqatiladigan hisoblash tugunlari sonini ko’paytirish va sifat darajasini oshirish lozim. Ma’lumotlar miqdori 2 barobar oshdi degani - klasterdagi temir miqdori 2 barobar oshdi degani.
2. Faoliyat barqarorligi. Ushbu tamoyil avvalgisidan kelib chiqadi. Klasterda ko’plab hisoblash tugunlari bo’ladi (ba’zan o’n minglab) va ularning soni ko’payishi aniq. Shuning uchun mashinaning ishlamay qolish ehtimoli oshadi. Masalan, Yahoo-ning Hadoop klasterida 42000 dan ortiq mashinalar mavjud. Ma’lumotlarning katta qismi bunday uzilishlar ehtimolini hisobga olishi va ularni sifatli saqlab turishi kerak.

3. Ma’lumotlarning lokalligi. Ma’lumotlar juda ko’p sonli hisoblash tugunlari bo’yicha tarqatilganligi sababli, agar ular bir serverda jismonan joylashgan bo’lsa va boshqasida qayta ishlansa, ma’lumot uzatish xarajatlari asossiz ravishda katta bo’lishi mumkin. Shuning uchun, ular saqlanadigan o’sha mashinada ma’lumotlarni qayta ishlash maqsadga muvofiqdir. Ushbu tamoyillar yaxshi tuzilgan ma’lumotlar uchun an’anaviy, markazlashtirilgan, vertikal saqlash modellarida mavjud bo’lganlardan farq qiladi. Shunga ko’ra, katta ma’lumotlar bilan ishlash uchun yangi yondashuvlar va texnologiyalar ishlab chiqilmoqda. Dastlab yondashuvlar va texnologiyalar to’plamiga noSQL MBBT, MapReduce algoritmlari va Hadoop loyiha vositalari kabi tuzilgan ma’lumotlarni massiv ravishda parallel qayta ishlash vositalari kiritilgan. Keyinchalik juda katta hajmdagi ma’lumotlar massivlarini qayta ishlashga o’xshash imkoniyatlarni ta’minlaydigan boshqa yechimlar va ba’zi bir qo’shimcha qurilmalar katta ma’lumotlar texnologiyalari deb nomlana boshlandi.

- MapReduce - Google tomonidan taqdim etilgan kompyuter klasterlarida taqsimlangan parallel hisoblash modeli. Ushbu modelga muvofiq, dastur klaster tugunlarida bajariladigan va so’ngra yakuniy natijaga qadar tabiiy ravishda kamaytirilgan bir xil elementar topshiriqlarning ko’p soniga bo’linadi.

Reducefunksiyasi foydalanuvchi tomonidan belgilanadi va alohida “savat”uchun yakuniy natijani hisoblab chiqadi. Reduce funksiyasi tomonidan qaytarilgan barcha qiymatlar to’plami MapReduce vazifasining yakuniy natijasidir. MapReduce haqida bir nechta qo’shimcha ma’lumotlar:

1) Mapfunksiyasining barchasi mustaqil va parallel ravishda ishlaydi. Shu jumladan klasterdagi turli xil mashinalarda ham ishlashi mumkin.
2) Reducefunksiyasining barchasi mustaqil va parallel ravishda ishlaydi. Shu jumladan klasterdagi turli xil mashinalarda ham ishlashi mumkin.
3) Shufflefunksiyasining ichki tuzilishi parallel bo’lib, u ham klasterdagi turli xil mashinalarda ishlashi mumkin.

4) Mapfunksiyasi odatda ma’lumotlar saqlanadigan o’sha mashinada qo’llaniladi -bu tarmoq orqali ma’lumotlar uzatilishini kamaytirishga imkon beradi (ma’lumotlar lokalligi tamoyili).

5) MapReduce –bu har indekslar mavjudligini va doim to’liq ma’lumotlarni skanerlash degani. Bu MapReduce juda tez javob talab etilganda juda yomon ishlashini anglatadi.

- NoSQL(Not Only SQL) - turli norelatsion ma’lumotlar bazalari va omborlari uchun umumiy atama bo’lib, ma’lum bir texnologiya yoki mahsulotga tegishli emas. An’anaviy relyatsion ma’lumotlar bazalari juda tez va bir xil so’rovlar uchun juda mos keladi va aksincha katta ma’lumotlarga xos bo’lgan murakkab va egiluvchan so’rovlarda bosim o’rtacha me’yordan oshib ketadi va MBBT danfoydalanish samarasiz bo’ladi.
- Hadoop - yuzlab va minglab tugunlarning klasterlarida ishlaydigan tarqatiladigan dasturlarni ishlab chiqish va bajarish uchun utilita, kutubxonalar va ramkalar to’plami. Bu katta ma’lumotlarning asoslaridan biri hisoblanadi.
- R - statistik ma’lumotlarni qayta ishlash va grafikalar uchun dasturlash tili. U ma’lumotlarni tahlil qilish maqsadida keng qo’llaniladi va statistik dasturlarning amaldagi standartiga aylangan.
- Apparatli yechimlar. Teradata korporatsiyasi, EMC va boshqalar katta ma’lumotlarni qayta ishlashga mo’ljallangan apparatli va dasturiy ta’minot tizimlarini taklif qilishadi.

Ushbu majmualar server klasteri va massiv parallel ishlov berish uchun boshqaruv dasturini o’z ichiga olgan o’rnatishga tayyor telekommunikatsion shkaflar sifatida yetkazib beriladi. Bunga ba’zida operativ xotirada analitik ishlov berish uchun apparatli yechimlari ham kiritiladi. Xususan, SAP kompaniyasidan Hanava Oracle kompaniyasidan Exalytics apparat va dasturiy ta’minot tizimlari kompleksi bo’lishiga qaramay, ularning operativ xotirasi miqdori bir necha

terabayt bilan cheklanadi.McKinsey konsalting kompaniyasi aksariyat tahlilchilar tomonidan ko’rib chiqiladigan NoSQL, MapReduce, Hadoop, Rtexnologiyalaridan tashqari Business Intelligencetexnologiyalari va SQL tilini qo’llab-quvvatlaydigan katta ma’lumotlarni qayta ishlashga qodir relyatsion ma’lumotlar bazasini boshqarish tizimlarini o’z ichiga oladi. McKinsey xalqaro strategik boshqaruv kompaniyasi katta ma’lumotlarga tatbiq etilishi mumkin bo’lgan 11 ta tahliliy uslublarni keltiradi.

- Data Mining uslubi(ma’lumotlarni olish, ularni intellektual va chuqur tahlil qilish) - qaror qabul qilish uchun zarur bo’lgan ilgari noma’lum, ahamiyatsiz, amaliy foydali bilimlarni aniqlash usullari to’plami. Bunday usullarga, xususan, assotsiativ qoidalarni o’qitish (association rule learning), klassifikatsiya qilish (turkumlarga ajratish), klasterli tahlil, regression tahlil, og’ishlarni aniqlash va tahlil qilish va boshqalar kiradi.
- Kraudsorsing – bu ishni mehnat munosabatlariga kirmasdan bajaradigan keng doira kuchlari tomonidan ma’lumotlarni tasniflash va boyitish.
- Ma’lumotlarni birlashtirish va integratsiya qilish(data fusion and integration) - chuqur tahlil qilish maqsadida (raqamli signallarniqayta ishlash, nutqni qayta ishlash, shu jumladan ohang tahlili va h.k) turli xil manbalardan olingan ma’lumotlarni birlashtirishga imkon beradigan texnik vositalar to’plami).
- Avtomatik ta’lim. shu jumladan nazorat ostida va nazoratsiz o’rganish - bazaviy modellardan murakkab bashoratlarni yaratish uchun statistik tahlilga asoslangan modellardan foydalanish yoki avtomatik ta’lim.
- Sun’iy neyronli tarmoqlar. Tarmoqli tahlil, optimallashtirish, shu jumladan genetik algoritmlar (genetic algorithm - Tabiatda tabiiy tanlov jarayoniga o’xshash mexanizmlardan foydalangan holda, kerakli parametrlarni tasodifiy tanlash, kombinatsiya qilish va o’zgartirish orqali optimallashtirish va modellashtirishni hal qilishda foydalaniladigan evristik qidiruv algoritmlari).

- Bashoratli tahlil. Tahlilchilar tizimga oldindan ma’lum parametrlarni o’rnatishga harakat qiladilar. So’ngra katta hajmdagi ma’lumotlarning kelib chiqishi asosida obyektning xatti-harakatlarini tekshiradilar.
- Imitativ modellashtirish(simulation) - jarayonlarni haqiqatda bo’lgani kabi tasvirlaydigan modellarni yaratishga imkon beradigan usul. Imimtatsiyani eksperimental sinovning bir turi deb hisoblash mumkin.
- Statistik tahlil - vaqtinchalik qatorlar tahlili, A/B-testlash (A/B testing, split testing - marketing tadqiqot usuli: undan foydalanganda, elementlarning nazorat guruhi bir yoki bir nechta ko’rsatkichlar o’zgartirilgan test guruhlari to’plami bilan taqqoslanadi.

Bu o’zgarishlar aniq nima yaxshilaganligini aniqlash uchun qilinadi. Tahliliy ma’lumotlarni vizuallashtirish - natijalarni olish, qo’shimcha ma’lumotlarni tahlil qilish uchun kirish ma’lumotlari sifatida ishlatish interaktivlik va animatsiyadan foydalangan holda ma’lumotlarni rasmlar, diagrammalar shaklida taqdim etish. Keng qamrovli ma’lumotlarni tahlil qilishning eng muhim bosqichi bo’lib, bu sizga tahlil natijalarini tushunarli shaklda taqdim etishga imkon beradi.

BIG DATA dan maksimal darajada foydalanish uchun faqat analitik IT-yechimlarni qo’llashning o’zi yetarli emas. Ushbu ma’lumotlarning tartibsiz manbalaridan yig’ilishini va ajratib olinishini tashkil qilish kabi ishlar muhim ahamiyatga ega bo’lib, shu maqsadlar uchun data, text, procces miningishlatiladi. Turli sohalarida katta ma’lumotlardan foydalanish. Biznes va marketingsohasidagi katta ma’lumotlar. Inson ma’lum narsa va hodisalar to’g’risida qanchalik ko’p ma’lumotga ega bo’lsa, aniq bashorat qilish ehtimoli shuncha yuqori bo’ladi. BIG DATA biznes va marketing sohasi uchun juda muhim. Biznes strategiya, marketing faoliyati, reklama tahliliga va mavjud ma’lumotlar bilan ishlashga asoslangan. Katta massivlar katta miqdordagi ma’lumotni tahlil qilishga imkon beradi va shunga mos ravishda tovar, mahsulot, xizmatni rivojlantirish yo’nalishini iloji boricha aniqroq ko’rsatadi.Masalan,

RTB kim oshdi savdosi katta ma’lumotlar bilan ishlaydi va tijoriy takliflarini hammaga emas, balki kerakli auditoriyaga samarali ravishda reklama qilish imkonini beradi.Biznes uchun foydasi:

- foydalanuvchilar va xaridorlar orasida talabga mos bo’lgan loyihalarni yaratish.
- kompaniyaning mavjud xizmati asosida mijozlar talablarini o’rganish va tahlil qilish. Hisob-kitob asosida xizmat ko’rsatuvchi xodimlarning ishini yaxshilash.
- Bloglar, ijtimoiy tarmoqlar va boshqa manbalardan olingan turli xil ma’lumotlarni tahlil qilish orqali mijozlar bazasining loyal yoki noroziligini aniqlash.
- Ko’p sonli ma’lumot bilan tahliliy ish olib borish orqali auditoriyani jalb qilish va saqlab qolish. Google Trends, Yandex va Wordstat (Rossiya va MDH uchun) texnologiyalaridan foydalangan holda mahsulotlarning ommabopligini taxmin qilish mumkin.

BIG DATA dan barcha yirik kompaniyalar -IBM, Google, Facebook va moliyaviy korporatsiyalar -VISA, Master Card, shuningdek, dunyodagi ba’zi vazirliklar foydalanadi. Masalan, Germaniyada ba’zi fuqarolar ishsizlik bo’yicha nafaqani asossiz olayotgani hisoblanib, ishsizlik bo’yicha nafaqa berish qisqartirildi. Shunday qilib, budjetga taxminan 15 milliard yevro qaytarildi. Yaqinda foydalanuvchi ma’lumotlarining tarqalishi sababli yuzaga kelgan Facebook bilan bog’liq mojaro saralanmagan ma’lumotlar hajmi o’sib borishini va hatto raqamli asrning gigantlari ham har doim konfidensiallikni to’liq ta’minlay olmasliklari ko’rinib qoldi. Masalan, Master Card mijozlarning hisob varoqlari bilan bog’liq firibgarlik operatsiyalarining oldini olish uchun katta ma’lumotlardan foydalanadi. Shunday qilib, yiliga o’g’irlikdan 3 milliard AQSh dollaridan ko’proq mablag’ni asrab qolish mumkin. O’yin sohasida katta ma’lumotlar o’yinchilarning xatti-harakatlarini tahlil qilish, faol auditoriyani aniqlash va shu asosda o’yinga qiziqish darajasini taxmin qilish imkonini

beradi.Bugungi kunda korxonalar o’z mijozlari haqida ularning o’zlaridan ham ko’proq ma’lumotga ega.

Nazorat savollar.

Katta hajmdagi ma’lumotlarni yaratish va boshqarish usullarini izohlang.

Katta ma'lumotlarni boshqarish arxitekturasini yaratishni izohlang.

Mavzuni mustahkamlash uchun savollar.

1. An’anaviy AT tizimlari Hadoop kabi katta hajmdagi ma’lumotlar texnologiyalari bilan birlashganda nimaga asos bo’lishi mumkin?

a) Katta hajmdagi ma’lumotlarni boshqarish va ma’lumotlarni qidirish

b) Katta hajmdagi ma’lumotlarni boshqarish

c) Katta hajmdagi ma’lumotlarni qidirish

d) Strukturalanmagan ma’lumotlarni to’plash va saqlash

2.Quyidagi javoblardan qaysida Hadoopni aniq va to’liq tasvirlashning xususiyatlari to’g’ri ko’rsatilgan?

a) Ochiq kodli, javaga asoslangan, taqsimlangan
b) Obektga yo’naltirilgan dasturlash asosida, ochiq kodli
c) Ochiq kodli
d) Haqiqiy vaqt (real time) rejimida ishlaydi

3.Hadoop - bu turli xil vositalar bilan ishlaydigan freymwork bo’lib, u qaysi vositalarni o’z ichiga oladi?

a) MapReduce, Hive va HBase
b) MapReduce va Google Apps
c) MapReduce, MySQL va HBase
d) D. Heron va Trumpet