21-§. Katta hajmdagi ma’lumotlarda Hadoop arxitekturasi

Reja:

1. *Katta hajmdagi ma’lumotlarda Hadoop arxitekturasi.*
2. *Katta ma’lumotlarni qayta ishlashda neyron tarmoqlarning o’rni.*

Hadoop, Kafka, Spark va TensorFlow kabi katta ma’lumotlar ekotizimidagi ochiq kodli platformalarga bo’lgan talab uning sun’iy intellekt, mashinani o’rganish, chuqur o’rganish yoki ma’lumotlar faniga bevosita murojaat qilishi tufayli kamayishi mumkin. Ammo Hadoop, NoSQL, xotira, oqim va boshqa ko’plab ma’lumotlar bazalari kabi ma’lumotlarni tahlil qilish platformalarining gibrid o’rnatilishi ma’lumotlar ko’lami va ma’lumotlar ishlab chiqarish yechimlari bozor ulushini oshiradi. 2021-yilda mamlakatlarning global yirik ma’lumotlar va biznes-tahlil bozoridagi ulushi: AQSh – 51%; Yaponiya – 5,7%; Xitoy – 5,5% Buyuk Britaniya – 5,1%; Germaniya – 4,4%; boshqalar – 28,3%. «Katta ma’lumotlar»ning kelajagi tahlilchilarni talab qiladi.

Katta ma’lumotlar hajmi o’sishda davom etar ekan, unga sho’ng’ish va amaliy tushunchalarni olish uchun o’qitilgan ma’lumotlar tahlilchilariga ehtiyoj ortib bormoqda. «Katta ma’lumotlar» tahlili moliya, hukumat va sog’liqni saqlash kabi sohalarda o’zgarishlar yaratish uchun ajoyib imkoniyatlarni taqdim etadi, shuningdek, firibgarlikning oldini olish, tabiiy ofat yuz berganda resurslarni taqsimlash yoki sog’liqni saqlashni yaxshilash orqali odamlar hayotini o’zgartirishga yordam beradi.

Hadoop - yuqori yuklangan saytlar uchun qidiruv va kontekstli mexanizmlarni amalga oshirish uchun ishlatiladi - Facebook, Twitter eBay, Amazon kabi veb loyihalarda qo’llaniladi. Shuningdek IBM, EMC, Dell, Oracle kabi dasturiy maxsulotlarda qo’llaniladi. O’ziga xos xususiyat shundaki, tizim har qanday klaster tugunlarining ishdan chiqishidan himoyalangan, chunki har bir blokda kamida bitta nusxa mavjud. Apache Software Foundation tomonidan katta

hajmli ma’lumotlarni parallel muhitda saqlash uchun ochiq manbali ramka. Ma’lumotni qayta ishlash mexanizmi bilan samarali tarqatish omboriga ega. Hadoop saqlash tizimi Hadoop Distributed File System (HDFS) nomi bilan mashhur. Ma’lumotni ba’zi mashinalar o’rtasida taqsimlaydi.

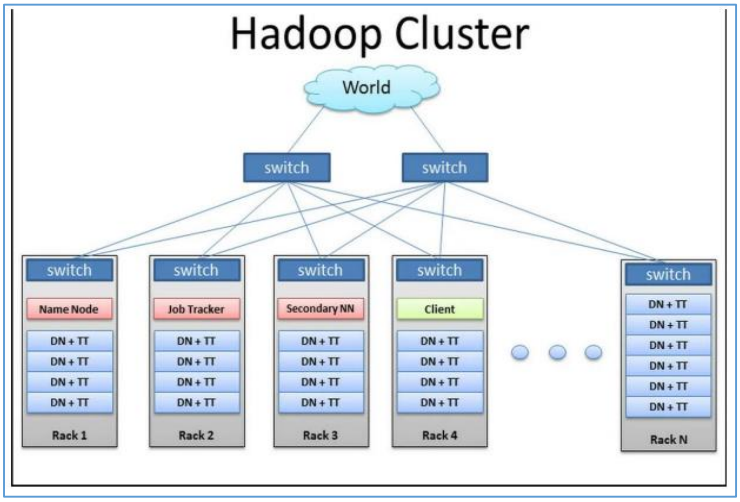
Hadoop usta-qul me’morchiligiga amal qiladi. Asosiy tugun Ism tuguni, qullar Ma’lumot tugunlari deyiladi. Ma’lumotlar barcha ma’lumotlar tugunlari o’rtasida taqsimlanadi. Hadoop da ma’lumotlarni qayta ishlashning asosiy algoritmi Map Reduce deb nomlanadi. Xaritani kamaytirish dasturlari yordamida ishlarni qul tugunlariga yuborish mumkin. Xaritani qisqartirish dasturlarini yozish uchun standart til Java hisoblanadi, lekin boshqa tillardan ham foydalanish mumkin. Ma’lumot tugunlari yoki qul tugunlari tahlil vazifasini bajaradi va natijani asosiy tugunga/ism-tugunga qaytaradi. Master-tugun/ism-tugunida xaritani ishga tushirish uchun ish kuzatuvchisi mavjud bo’lib, u qul tugunlaridagi ishlarni kamaytiradi. Qul tugunlari ma’lumotlar tugunlarida ma’lumotlarni tahlil qilishni yakunlash va natijani asosiy tugunga qaytarish uchun vazifa kuzatuvchisi mavjud.

Hadoop ning asosiy texnik xarakteristikalariga quyidagilar kiradi:

- Kengayuvchanligi: platforma petabayt ma’lumotlarni saqlash va ishlov bera olishi bilan chiziqli kengayishi mumkin;
- Ishdan chiqishga turg’unligi: barcha saqlanayotgan mlumotlar keragidan ortiq, uzilib qolgan ishlov berish masalalari qaytadan boshlanadi;
- krossplatformalik: Hadoop kutubxonalari asosan Java tilida yozilgan bo’lib, Java mashinani qo’llab-quvvatlaydigan har qanday operatsion tizim ostida ishlashi mumkin;
- Masalalarni avtomatik tarzda paralellashtirish: Hadoop texnalogiya dasturchilariga ko’rinib turadigan “shaffof” abstraksiyalar hosil qiladi. Shu bilan ularni malumotlarni paralel ishlov berish natijalrini loyihalash, boshqarish va agregatsiya qilish ishlaridan forig’ qiladi.

Hadoopdan foydalanishning afzalliklari quyidagilarda namoyon bo’ladi:

- Qayishqoqlik: strukturalangan va strukturalanmagan malumotlar tipini saqlash va tahlil qilish;
- Samaraliylik: ko’p hollarda terabayt malumotlarni saqlash va ularga ishlov berish boshqa texnologiyalarga nisbatan arzon narxga tushadi;
- Klasterni arzon xosil qilish: Hadoop klasterlarini hosil qilish uchun qimmat server apparat talab qilinmaydi.



4.7-rasm. Hadoop klasterlarini hosil qilish

- Nisbatan yengil moslashuvchanlik: Hadoop keng va aktiv rivijlangan ekosistemaga ega
- Minimal risk:platforma yadrosini noto’g’ri ishlashi bilan bog’liq risklarning minimalligi
- “Open source” litsenziyasi: Hadoop platformasini qo’llash va egalik qilishning arzon narxdaligi
- Platformadan foydalanadigan ishlab chiqaruvchilar soning ko’pligi.
- Forrester Research kompaniyasi analitiklarining fikricha, Apache Hadoop platforma barcha katta kompaniyalarning AT- infrastrukturalari uchun standart vazifani o’taydi.

Katta xajmdagi ma’lumotlarda Hadoop arxitekturasi - Hadoop va Spark kabi yangi katta ma’lumotlar texnologiyalari kompyuter klasterlari bilan ishlash va ularni boshqarishni osonlashtiradi. Hadoop minglab kompyuterlarni

masshtablashtirib, petabaytli klasterlar va saqlash joylarini yaratishi mumkin. Bu korxonalarga mavjud bo'lgan katta hajmdagi ma'lumotlardan foydalanish imkonini beradi.

Hadoop. Spark. Hive. Katta ma`lumotlarda qo`llanilayotgan Hadoop, Spark kabi yangi texnologiyalar ko`mpyuter klasterlari bilan va ularni boshqarish ishlarini osonlashtiradi (qisqartiradi). Hadoop – minglab komyuterlarni masshtablashda ma`lumotlar saqlash muxitini petabaytlar sig`adigan klasterlar yaratish imkonini beradi.

Hadoop- bu katta hajimdagi axborotlarni qayta ishlash va saqlash infrastrukturasidir. Hadoop quyidagi maqsadlarni amalga oshirish uchun qo`llaniladi:

Ishonchlilik – bu bir necha nushada ma`lumotlarni to`plamlarini yaratib, shu ma`lumotlardan qayta foydalanish mantiqini uzilish, buzulish (сбой) holatlarida vosita vazifasini bajaradi.

Rad etishga mustaxkamlik - uzilish, buzulish (сбой) holatlari avtomatik tiklasjni ta`minlash

Masshtablilik – ma`lumotlarni qayta ishlash taqsimoti kompyuterlarda klasterlanadi (bu gorizantal masshtablashtirish deb ataladi).

Universal port (Портируемость) – barcha turdagi qurilma va OT larga o`rnatish imkonini mavjudligi. Hadoop ning asosiy komponentlari

HDFS – fayl tizimini taqsimoti.

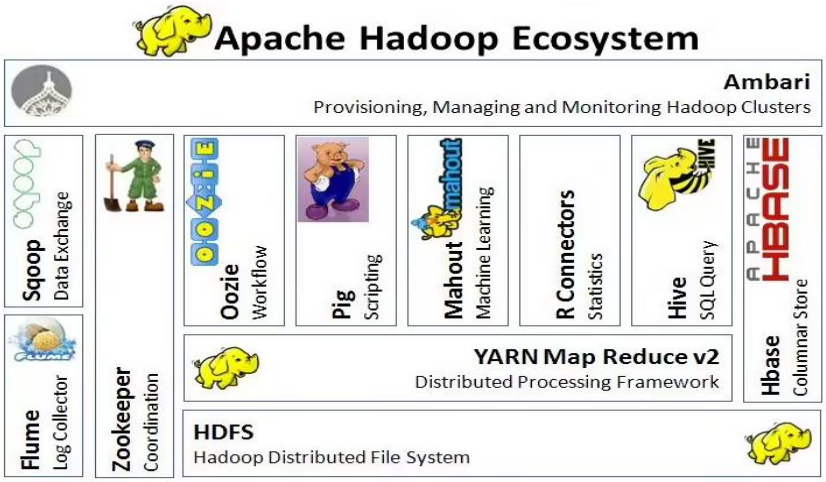
MapReduce – dastur bajarilishini katta masshtabli uslubi.

YARN – klaster resurslarni boshqarish tizimi.

Spark: MapReduce ni almashtirish orqali ishlab chiqarishni ortirish- MapReduce ning ishini kichik bir xayoliy misol bilan ko'rib chiqish mumkin. Siz o'yinchoqlar ishlab chiqaruvchi kompaniya direktorisiz. Har bir o'yinchoq ikkita rangga bo'yalgan; mijoz veb-sahifadan o'yinchoq buyurtma qilganda, buyurtma fayli ko'rsatilgan ranglar bilan Hadoop-ga joylashtiriladi.

Sizning vazifangiz bo’yoq uchun qancha idish tayyorlash kerakligini aniqlashdir. Ranglarni hisoblash uchun MapReduce uslubidagi algoritm qo’llaniladi.

Hadoop turli kompaniyalar, jumladan yirik va taniqli kompaniyalar eBay, Facebook, Amazon, IBM, AliExpress, Yahoo tomonidan qo’llaniladi. Shu bilan birga, har qanday kompaniya uchun ma’lumotlar bilan ishlashning yagona sxemasi mavjud emas, chunki barcha xizmatlarning ishi o’ziga xosdir. Shuningdek, ma’lum kompaniyalar uchun maxsus ishlab chiqilgan qo’shimcha funktsiyalar asosiy funksionallikka o’rnatiladi.



4.8-rasm. Hadoop ilovalari

Hadoop-ning asosiy ilovalari:

• yuqori yuklangan onlayn-do’konlar va veb-saytlar uchun kontekstli va qidiruv mexanizmlari;

• katta hajmdagi ma’lumotlarni saqlash va saralash, ulkan fayllar tarkibini tahlil qilish;

• grafik ma’lumotlarni tez qayta ishlash.

?	• **Nazorat savollari.**
1.	• Katta hajmdagi ma’lumotlarda Hadoop arxitekturasini izohlang.
2.	• Hadoopning dasturiy vositalar bilan integratsiyasini izohlang.

Mavzuni mustahkamlash uchun savollar.

1. Apache Hadoop - bu ...?

a) MapReduce-ning ochiq kodli ilovasi

b) MapReduce-ning shaxsiy amalga oshirilishi

c) Talend yopiq prototipi

d) Talend ochiq prototipi

2.Hadoop – bu …?

a) Klasterlarda ishlaydigan taqsimlangan dasturlarni bajarish uchun yordamchi dasturlar to’plami va dasturiy ta’minot tizimi

b) Katta ma’lumotlar bilan ishlay oladigan taqsimlangan MBBT

c) MapReduce paradigmasida ishni bajarish tili

d) Katta hajmdagi fayllarni saqlash uchun mo’ljallangan taqsimlangan fayl tizimi

3.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