



MAPREDUCE PARADIGMASI VA BIG DATA BILAN ISHLASH TEXNOLOGIYALARI

Muallif: Madaminov Uktamjon Ataxanovich¹

Affiliyatsiya: Abu Rayhon Beruniy nomidagi Urganch davlat universiteti dotsenti, p.f.f.d.
(PhD)¹

DOI: <https://doi.org/10.5281/zenodo.17350555>

ANNOTATSIYA

Ushbu maqolada katta hajmli ma'lumotlarga ta'rif berilgan. Katta hajmli ma'lumotlarga misollar keltirilgan va ularni qayta ishlash texnologiyalarining bir nechta tahlil qilingan. Katta hajmli ma'lumotlarni qayta ishlovchi texnologiyalardan parallel qayta ishlash metodi, qayta taqdim etilgan tanlov metodi va modellar kombinatsiyasi metodlarining afzalliklari va kamchiliklari ko'rib chiqilgan. Ushbu metodlar ma'lumotlarni qayta ishlovchi taqsimlangan texnologiyalar bilan solishtirilgan. Ushbu maqolada yana katta hajmli ma'lumotlarni qayta ishlash prinsiplaridan bir qanchasi ko'rib chiqilgan. Ulardan gorizontol o'lchovlar prinsipi, teskari barqarorlik prinsipi, ma'lumotlarning joylashuvi prinsiplari ko'rib chiqilgan va tahlil qilingan.

MapReduce modeli ham yuqoridagi uchta prinsipga asoslanib ish olib boradi. Bu ham kompyuter klasteridagi katta hajmdagi ma'lumotlarni qayta ishlash uchun taqsimlangan ma'lumotlarni qayta ishlash modeli hisoblanadi. Ushbu maqolada bir nechta metod va modellar tahlil qilinib afzalliklari va kamchiliklari ko'rib chiqilgan.

Kalit so'zlar: Big Data, MapReduce paradigmasi, gorizontol o'lchovlar, ma'lumotlarning joylashuvi, teskari barqarorlik.

KIRISH

MapReduce paradigmasi:

- Big Data-ma'lumotlar 100 Gb dan ortiq bo'lsa;
- Big Data-bu Excelda qayta ishlab bo'lmaydigan ma'lumotlar;
- Big Data- bitta kompyuterda qayta ishlanmaydigan ma'lumotlar.

Katta hajmli ma'lumotlar- bu turli xil strukturalangan va strukturalanmagan ma'lumotlarni qayta ishlash metodlari, usullari va yechimlarining uzluksiz o'sib borishidir. Katta hajmli ma'lumotlar deganda ma'lumotlar miqdori emas balki, ularni qayta ishlash metodlari tushiniladi. Quyidagi ba'zi bir ma'lumotlar manbai katta hajmli ma'lumotlar manbasiga misol bo'la oladi:

- internetda foydalanuvchi harakatlarining qaydlari;
- transport kompaniyalari uchun avtomobillardan kelayotgan GPS signallar oqimi;
- barcha bank mijozlarining tranzaksiyalari;
- katta chakana savdo tarmog'ida barcha xaridlar haqida ma'lumot va h.k.

METODOLOGIYA

Katta hajmli ma'lumot manbalarining soni jadal o'sib bormoqda, ya'ni ularni qayta ishlash texnologiyalari tobora ommalashib bormoqda.

Katta hajmli ma'lumotlarni qayta ishlash prinsiplari. Big Data ta'rifiga asosanib, bunday ma'lumotlarni qayta ishlashning asosiy prinsiplarini shakllantirish mumkin. Bular:

Gorizontol o'lchovlar. Katta ma'lumotlarni qayta ishlash uchun imkon qadar katta ma'lumotlarni qayta ishlashni qamrab oluvchi har qanday tizim kengaytirilishi mumkin. Bunda ma'lumotlar hajmi ikki barobar ortadi, klasterdagi komponentalar soni ham ikki marta oshadi va jarayon davom etadi.

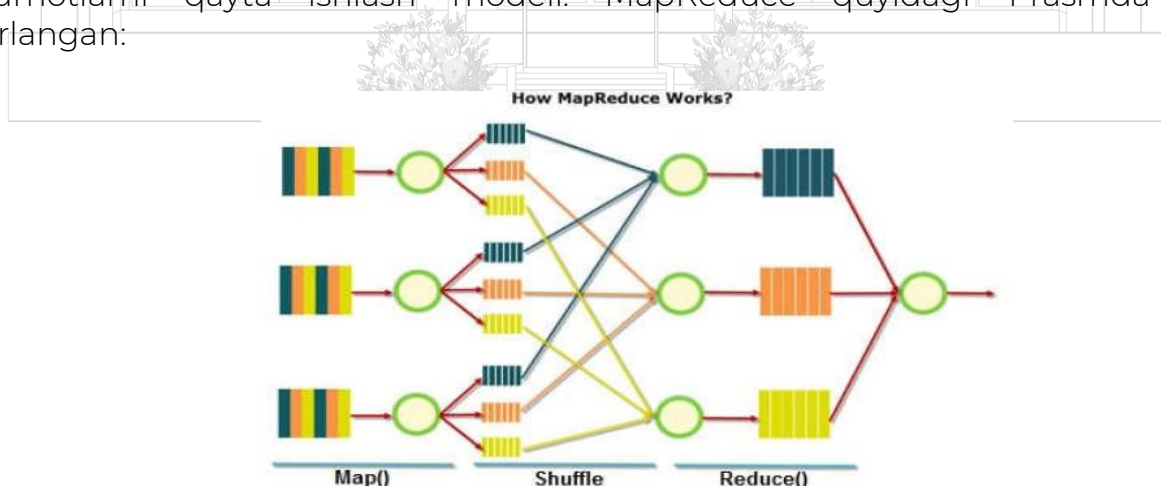
Teskari barqarorlik. Gorizontol o'lchovlar prinsipida klasterdagi komponentalar soni ko'p bo'ladi. Misol uchun, Yahooning Hadoop klasterida 42 000 dan ortiq mashinalar mavjud (turli tashkilotlarning klaster o'lchamiga qarash uchun foydalanish mumkin). Bu shuni anglatadiki, ushbu mashinalarning ba'zilar muvaffaqiyatsizlikka uchraydi. Katta ma'lumot bilan ishlash usullari bu kabi xatolarning mavjudligini hisobga olishi va hech qanday jiddiy oqibatlarisiz ishlashi kerak.

Ma'lumotlarning joylashuvi. Katta taqsimlangan tizimlarda ma'lumotlar ko'p sonli mashinalarga tarqatiladi. Ma'lumotlar bir xil serverda joylashgan bo'lsa va boshqasida qayta ishlansa - ma'lumotlar uzatish qiymati uni qayta ishlash narxidan oshib ketishi mumkin. Shuning uchun, BigData yechimlari uchun eng muhim prinsip tamoyillaridan biri axborotni joylashish tamoyili bo'lib, iloji bo'lsa, ularni saqlaydigan bir mashinaning o'zida ma'lumotlarni qayta ishlash mumkin [3].

TAHLILLAR

Katta ma'lumot bilan ishlashning barcha zamonaviy usuli bu uchta printsipga amal qiladi. Bu prinsiplar bilan ishlaydigan modellarni ko'rib chiqamiz.

MapReduce. MapReduce - Google tomonidan kompyuter klasterlarida katta hajmdagi ma'lumotlarni qayta ishlash uchun taklif qilingan taqsimlangan ma'lumotlarni qayta ishlash modeli. MapReduce quyidagi 1-rasmda aniq tasvirlangan:



1-rasm. MapReduce modelida ma'lumotlar uch bosqichda amalga oshiriladi

Map bosqichi. Ushbu bosqichda ma'lumotlar foydalanuvchi tomonidan belgilaydigan map() funksiyasidan foydalaniladi. Ushbu bosqichning vazifasi ma'lumotlarni qayta ishlash va filtrlashdir. Bunda har bir kirish yozuviga maxsus

funksiya qo'llaniladi. Map () funksiyasi bitta kirish yozuviga qo'llaniladi va bir qator kalit-qiyamat juftlarini hosil qiladi.

Shuffle bosqichi. Bu bosqich foydalanuvchiga sezilmaydi. Bu bosqichda map() funksiyasining chiqishlari "savatcha"larga bo'linadi. Keyinchalik bu "savatcha" lar Reduce() bosqichi uchun kirish ma'lumotlari hisoblanadi.

Reduce() bosqichi. Shuffle bosqichida hosil bo'lgan qiymatga ega bo'lgan har bir "savat" reduce() bosqichiga o'tadi. Reduce() bosqichi foydalanuvchi tomonidan belgilanadi va bitta "savat" uchun yakuniy natijani hisoblaydi. Reduce () funksiyasi tomonidan qaytarilgan barcha qiymatlarning to'plami MapReduce vazifasining yakuniy natijasidir.

NATIJALAR

MapReduce haqida ba'zi qo'shimcha ma'lumotlar: funksiyaning ishlashi mustaqil ravishda boshlanadi va parallel ravishda turli xil klientli mashinalarda ishlashi mumkin. Shuffle parallel saralashni ifodalaydi, shuning uchun u turli xil klientli qurilmalarda ishlashi mumkin. 1-3-bandlar gorizontali o'lchovli printsiptini bajarishga imkon beradi.

MapReduce-bu hamma vaqt ma'lumotlarni ko'chirib oladi va qachonki murojaat bo'lsa, zudlik bilan javob qaytaradi [6].

Katta hajmli ma'lumotlar aslida ilmiy-tadqiqotga aloqador sohalarda avvaldan ham mavjud bo'lgan bo'lsada, lekin so'nggi yillardagina ushbu fenomen ko'proq tilga olinadigan bo'ldi. Bunga sabab bugungi kunda yangi texnologiyalar, qurilmalar va aloqa vositalarining tezkor rivojlanishi, ijtimoiy tarmoqlarning keng miqyosda yoyilishi ta'sirida insoniyat tomonidan ishlab chiqilayotgan ma'lumotlar miqdori keskin tarzda ortib borishidir. Hozirgi kunda insoniyat tomonidan ma'lumot har 10 daqiqada yaratilmoqda. Bu turdagi katta hajmli ma'lumotlar bilan ishlashda yangidan-yangi algoritm va texnologiyalar ishlab chiqilmoqda va amaliyotda keng qo'llanilyapti. Katta hajmli ma'lumot tushunchasi katta o'lchamdagi ishlarni bajara oladigan operatorlarga nisbatan ishlatiladi. Misol uchun, «Facebook» kompaniyasi soatiga 10 milliondan ortiq yangi rasmlar joylashi bilan faxrlanishi mumkin. Katta miqdordagi ma'lumotlarni tahlil qilish inson his etish imkoniyatidan tashqarida bo'lgan qonuniyatlarni aniqlashda yordam beradi. Bu esa kundalik hayotimizdagi barcha sohalar – hukumatni boshqarish, tibbiyot, telekommunikatsiya, moliya, transport, ishlab chiqarish va boshqa sohalarni yanada yaxshilash, ularning imkoniyatlarini oshirish, muammolarga muqobil yechimlar izlab topish imkonini yaratadi. Salmoqli ma'lumotlardan foydalanish axborot mutaxasissi uchun qo'shimcha vazifalar va talablar yaratishdan tashqari, yangi ma'lumotlarni topish (Data mining) va ma'lumot bo'yicha mutaxassis (Data scientist) kabi kasblarning paydo bo'lishiga olib keladi. "Data mining" – biron qonuniyatni topish maqsadida ma'lumotlarni tahlil qilishga aytiladi [8].

"Data science" tushunchasiga ma'lumotlar omborini loyihalash va raqamlangan ma'lumotlarni qayta ishlashning barcha metodlari kiradi. "Data science" salmoqli ma'lumotning biznes nuqtai nazaridan hozirgi zamonaviy o'rindoshi hisoblanadi. Bugungi kunda elektron hukumat tizimi butun dunyoda, shu jumladan, O'zbekistonda keng rivojlanayotgan tizim hisoblanadi. Bu tizim fuqarolar, tadbirkorlik sub'yektlari va davlat hokimiyati organlariga avvaldan shakllangan davlat xizmatlarini ko'rsatishda zamonaviy axborot-kommunikatsiya texnologiyalaridan foydalanishni ko'zda tutadi [5].

XULOSA

Xulosa qilib shuni aytish mumkinki, bu uchta prinspga hozirgi zamonaviy barcha modellar amal qiladi. Bu prinsplar asosida ishlaydigan MapReduce modeli taqsimlangan tarzda ishlaydi. Bu modelda asosan axborotning joylashuv prinsipi asosiy rol o'ynaydi. Agarda axborot joylashuvi prinspi ishlamasa bunday vaziyatda ma'lumotni uzatish ma'lumotni qayta ishlashdan ko'ra qimmatroq bo'ladi. Ushbu ilmiy tadqiqot ishidakatta hajmli ma'lumotlarga misollar orqali aniq nazariy va amaliy yondashuvlar keltirib o'tildi. Big datani qayta ishlovchi texnologiyalardan parallel qayta ishlash usullari, tanlov metodlari va modellar kombinatsiyasi metodlarining afzalliklari hamda kamchiliklari keltirib o'tildi. Ushbu metodlar ma'lumotlarni qayta ishlovchi taqsimlangan texnologiyalar bilan solishtirilgan. Ulardan gorizontal o'lchovlar prinspi, teskari barqarorlik prinspi, ma'lumotlarning joylashuvi prinsplari ilmiy-amaliy jihatdan tahlil qilindi.

FOYDALANILGAN ADABIYOTLAR RO'YXATI

1. Franks, B. Taming the Big Data Tidal Wave Finding Opportunities in Huge Data Streams with Advanced Analytics / Bill Franks, 2012. – 45 p.
2. Gantz, J. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East - United States / J. Gantz, D. Rainsel // IDC Country brief, 2013. – 16 p.
3. Hadoop and Big Data [Электронный ресурс] – Режим доступа: <http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html>.
4. Golnar Assadat Afzali, Shahriar Mohammadi. Privacy Preserving Big Data Mining: Association Rule Hiding. 2016. – 10 p. <http://www.jist.ir/Article/139504261512112857>
5. Качалов Д.Л., Мишустин А.В., Фархадов М.П. Институт проблем управления РАН имени В.А. Трапезникова. Современные методы обработки больших данных в крупномасштабных системах. // Труды ИСП РАН. 2012. – 143-156 с.
6. Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, 2011. – 101-104 pp.
7. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, 2012. – 1-24 pp.
8. O'razmatov T.Q. Nurmetova B.B. Katta hajmli ma'lumot bilan ishlash prinsiplari. "Iqtisodiyotning tarmoqlarini innovatsion rivojlanishida axborot-kommunikatsiya texnologiyalarining ahamiyati" respublika ilmiy-amaliy anjumani. – Farg'ona, 2019. – 52-54 b.