

BIG DATE TIZIMI HAQIDA UMUMIY TASNIF VA TUSHUNCHA

Tolibjonov Xurshidbek Hayotjon o'g'li
Qo'qon universiteti 4-bosqich talabasi,
Meliboyev Azizjon Ikromjon o'g'li
Qo'qon universiteti o'qituvchisi

Annotatsiya: Siz ushbu maqolada Big data atamasi haqida to'liq ma'lumot olasiz. Big data kelib chiqish tarixi va uning qo'llanish sohalari haqida aniq ma'lumotlar aks etgan. Dunyoda raqamlashgan ma'lumotlarning yillar bo'yicha ortib borishi va Hadoop loyihasining ishlash tizimi va ahamiyati haqida ma'lumotlar bilan ham boyitilgan.

Kalit so'zlar: Big data, hadoop, raqamli ma'lumotlar, vositalar.

Bizning yashab turgan hayotimizda shaxsiy va turmush tarzimiz, o'qishimiz, ishimiz uchun zarur bo'lgan ma'lumotlar juda ko'p. Ularni saqlash uchun bizga inson xotirasiga qo'shimcha ravishda zamonaviy qurilmalar zarur. Oddiygina insonlarning tug'ilishidan boshlab uning esda qolarli xolatlari tarixga muxrlash uchun ham bizga mobil telefonimiz yoki qo'shimcha ravishda xotira kartalari zarur. Bir inson yoki biror insonlar jamoasi, oilasini ma'lumotlarini saqlash uchun mobil telefon xotirasi yetar, lekin katta tashkilotlar, korxonalar va kompaniyalarni ma'lumotlarini saqlash uchun yuqorida aytib o'tilganlar juda kichiklik qiladi. Shunday vaziyatda Big data xizmatidan foydalanishga to'g'ri keladi.

Big data (*katta ma'lumotlar*) - juda katta hajmdagi bir jinsli bo'lmagan va tez tushadigan raqamli ma'lumotlar bo'lib, ularni odatiy usullar bilan qayta ishlab bo'lmaydi. Ba'zi hollarda, katta ma'lumotlar tushunchasi bilan birga shu ma'lumotlarni qayta ishlash ham tushuniladi. Asosan, analiz obyekti katta ma'lumotlar deb ataladi.

Mutaxassislar fikriga ko'ra, «Big Data», ya'ni salmoqli ma'lumot atamasi kuniga 100 gb.dan ko'p ma'lumot tushadigan oqimlarga nisbatan qo'llanilgan ekan. Keyinchalik ma'lumotlarning keskin ko'payishi oqibatida bu tushuncha keng qamrov kasb eta boshlagan. Ushbu atama odatda (terabayt, ekzabayt va petabaytlar darajasidagi) katta hajmdagi ma'lumotlarga nisbatan qo'llaniladi.

Salmoqli ma'lumot atamasi ilk bor matbuotda 2008 yili paydo bo'lgan, o'shanda «Nature» jurnali bosh muharriri Klifford Linch o'z jurnalida katta hajmdagi ma'lumotlar bilan ishlovchi texnologiyalar yordamida ilm-fan kelajagini rivojlantirish mavzusida maqola chop etgan. 2009 yilgacha bu atamaga faqat ilmiy tahlil nuqtai nazaridan yondashilar edi, ammo matbuotda shu mavzuda yana bir necha maqolalar chop etilganidan so'ng «Big Data» tushunchasidan keng foydalana boshlandi. 2010-yilda salmoqli ma'lumotlarning o'sib borayotgan muammolarini hal etish uchun dastlabki urinishlar boshlandi. Ulkan axborotlar oqimidan foydalanganda xavflarni kamaytirish uchun mo'ljallangan dasturiy mahsulotlar ishlab chiqishga kirishildi. "Katta ma'lumot" ning ko'pi ma'lumotlarni yig'ish emas. Endi "Katta ma'lumot" haqiqiydir. Google katta ma'lumotlarga ega va CERN ham katta ma'lumotga ega. Aksariyat odamlar, ehtimol, yo'q. Ma'lumotlar faqatgina do'konga 1000 kompyuterga kerak bo'lganda boshlanadi. Hadoop kabi katta axborot texnologiyalari ham haqiqiydir. Ular har doim mantiqiy ravishda ishlatilmaydilar (hadoop klasterlari 100 dan ortiq tugunlarni ishlatishdan bezovta bo'lmanglar, chunki siz yaxshi tanlangan bo'lmagan kumelenmemie mashinalari yaxshi ishlashi mumkin). Lekin, albatta, odamlar bunday dasturiy ta'minotni yozadilar. Lekin amalga oshirilayotgan ishlarning aksariyati ma'lumotlarni yig'ish emas. Bu Extract, Transform, Load (ETL) , shuning uchun ma'lumotlar omborini almashtirish . Strukturani, indekslarni va tezlatilgan so'rovlar bilan ma'lumotlar bazasini ishlatish o'rniga ma'lumotlar faqat hadoopga tushiriladi va nima qilish kerakligini anglab etganingizda, siz barcha ma'lumotlarni qayta o'qiydiz va kerakli ma'lumotlarni chiqarib olasiz, tranzit

qilishingiz va uni Excel jadvaliga yuklang. Chunki tanlov, ekstraktsiya va konvertatsiya qilishdan so'ng, odatda u "katta" emas.

Dunyoda raqamlangan ma'lumotlar hajmi eksponent bo'yicha o'sib bormoqda. IBS kompaniyasining ma'lumotlariga qaraganda, 2003-yilda 5 eksabayt (1 eksabayt - 1 milliard gigabayt) ma'lumot yig'ilgan ekan. 2008-yilda u 0.18 zettabayt (1 zettabayt = 1024 eksabayt) gacha, 2011-yilga kelib 1.76 zettabayt, 2013-yilda 4.4 zettabaytgacha yetibdi. 2015-yilning mayida dunyoda yig'ilgan raqamlangan ma'lumotlar hajmi 6.5 zettabaytdan oshib ketibdi. 2020-yilga kelib insoniyat 40-44 zettabayt raqamli ma'lumot hosil qilar ekan.

IBS mutaxassislarining fikriga ko'ra, 2013-yilda yig'ilgan ma'lumotlar massivining atiga 1.5%i qandaydiy axborot qiymatiga ega bo'lgan ekan. Baxtga qarshi, hozir dunyoda katta ma'lumotlarni qayta ishlash texnologiyalari bo'lib, ular yordamida juda katta ma'lumotlar massividan insonlarga kerak, qiziq bo'lgan, foydali ma'lumotlarni ajratib olish mumkin bo'ladi.

Katta ma'lumotlarning ko'pgina marketing va'dalari o'tkazilmaydi. Twitter sizning reklamangizdan ko'ra ko'pchilik kompaniyalar uchun juda kamroq tushuncha ishlab chiqaradi (agar siz teenie rockstar bo'lmasangiz); va Twitter-ning foydalanuvchi bazasi juda baqongizdir. Bunday qarama-qarshilikni bartaraf etish qiyin va juda tajribali statistik xodimlarga muhtojdir.

Katta ma'lumotlarni qayta ishlash texnologiyalarini dastlabki o'rganish uchun javob beradi - sizni kurs bilan oson va aniq tanishtiradi. Axborotning ko'pligi qanday ta'sir qilganligi to'g'risida tushuncha beradi kundalik hayot va uning barcha sohalarini: ilm-fan, biznes, tibbiyot va boshqalar. Ko'plab rasmlarni o'z ichiga oladi, shuning uchun u ko'p harakat qilmasdan qabul qilinadi.

Shuningdek, "oddiydan murakkabgacha" printsiplari bo'yicha katta ma'lumotlar bilan qanday ishlashni tushuntirib beradigan Big Data to'g'risidagi kitob yangi kelganlar uchun ham foydalidir. Dastlabki bosqichda ko'plab muhim jihatlarni qamrab oladi: ishlov berish, vizualizatsiya, OLAP, shuningdek ma'lumotlarni tahlil qilish va tasniflashning ba'zi usullari.

Python dasturlash tilidan foydalangan holda katta ma'lumotlardan foydalanish va ular bilan ishlash bo'yicha amaliy qo'llanmalar muhandislik talabalari uchun ham, bilimlarini chuqurlashtirishni istagan mutaxassislar uchun ham javob beradi. Shuningdek deyarli barcha Big data bilan ishlaydigan ochiq code kutubxonalarini Python dasturlash tili qo'llab quvvatlaydi.

Hadoop - bu bir vaqtning o'zida minglab tugunlarda harakatlarni amalga oshiradigan taqsimlangan dasturlar bilan ishlash uchun maxsus yaratilgan loyiha. U bilan tanishish katta ma'lumotlarning amaliy qo'llanilishini batafsilroq tushunishga yordam beradi.

Katta ma'lumotlar, "Katta ma'lumotlar" bir necha yillardan beri shaharning axborot va marketing matbuotida munozaraga aylanib bormoqda. Va bu aniq: raqamli texnologiyalar zamonaviy inson hayotiga kirib, "hamma narsa yozilgan". Hayotning turli sohalariga oid ma'lumotlar hajmi o'sib bormoqda va shu bilan birga, ma'lumotlarni saqlash imkoniyatlari ham o'sib bormoqda.

Aksariyat ekspertlar ma'lumotlar o'sishini tezlashtirish ob'ektiv haqiqat ekanligiga qo'shiladilar. Ijtimoiy tarmoqlar, mobil qurilmalar, o'lchash moslamalari ma'lumotlari, biznes ma'lumotlari - bu juda katta hajmdagi ma'lumotni keltirib chiqaradigan manbalarning bir necha turlari. Izlanishlarga ko'ra IDC raqamli olam 2012 yilda nashr etilgan, kelgusi 8 yilda dunyoda ma'lumotlar miqdori 40 Zb (zettabayt) ga etadi, bu sayyoramizning har bir aholisi uchun 5200 Gb ga teng.

Muammolar: Muammo shundaki, Hadoop lineer muammolar uchun yaxshi, lekin ma'lumotlarning ko'pchiligi lineer emas va shuningdek chiziqli bo'lmagan algoritmlar katta ma'lumotlarga ega emas; chiziqli vaqtli yaqinlashuvlarni diqqat bilan o'rganib chiqing va

aniqroq ma'lumotni yo'qotish bilan yo'qotishingiz kerak, bu esa kichikroq ma'lumotlarga ishlov berish orqali yo'qotishingizdan kichikroq bo'lishi kerak. Ushbu savdo-sotiq muammosining yaxshi namunasi k-vositadir. K - aslida (asosan) chiziqli muammo hisoblanadi; shuning uchun Xadoopda biroz harakat qilish mumkin. Yagona iteratsiya chiziqli bo'lib, agar siz yaxshi dasturga ega bo'lsangiz, u katta ma'lumotlarga to'g'ri keladi. Shu bilan birga, yaqinlashuvlar soni ma'lumotlar majmui o'lchamlari bilan ham ortib boradi va shuning uchun u chiziqli emas. Biroq, bu "vositalar" ni topish uchun statistik usul bo'lgani uchun, natijalar aslida ma'lumotlar majmui o'lchamlari bilan ko'p emas. Shunday qilib, k-vositalarni katta ma'lumotlarda ishlatishingiz mumkin bo'lsa, u juda ko'p ma'noga ega emas - faqat ma'lumotlarning namunasini olishingiz mumkin, k-vositalarining juda samarali yagona nodali versiyasini ishlatasiz va natijalar xuddi shunday yaxshi bo'lsin. Qo'shimcha ma'lumotlar sizga aniq bo'lishi kerak bo'lmagan qiymatning aniqligi bilan izohlanadi. Bu juda ko'p muammolarga tegishli bo'lgani uchun, Hadoop'ta haqiqiy ma'lumotlar konida ochilishi ko'rinmaydi. Har kim buni qilishga urinadi va ko'plab kompaniyalar bu narsalarni sotadilar. Lekin bu katta bo'lmagan versiyadan ancha yaxshiroq ishlamaydi. Biroq, mijozlar buni sotib olishni istasa, kompaniyalar bu funktsiyani sotadilar. Va sizga grant oladigan bo'lsa, tadqiqotchilar bu haqda maqolalar yozadilar. U ishlaydimi yoki yo'qmi. Bu hayot: Bularning barchasi ishlaydigan bir nechta holatlar mavjud. Google qidiruvi - Cern. Lekin, shuningdek, tasvirni tanib olish (lekin Hadoop dan foydalanmasdan, GPU klasterlari u erga borishning usullari kabi ko'rinadi) yaqinda ma'lumotlar hajmining ko'payishidan foydalandi. Ammo bu holatlarning har qandayida sizda toza ma'lumotlar mavjud. Google indeksni hamma narsa; Cern har qanday qiziqarli ma'lumotlarni olib tashlaydi va faqatgina qiziqarli o'lchovlarni tahlil qiladi - Spamni Cern-ga yuboradigan spamerlar yo'q va tasvirni tahlil qilishda siz internet-kameradan yoki tasodifiy tasvirlardan emas, agar shunday bo'lsa, ularni vakolatli ma'lumotlar sifatida emas, tasodifiy tasvirlar sifatida baholaysiz).

Salmoqli ma'lumotlar muammosi shundaki, o'n yillar davomida yig'ilgan turli axborotlar hamon har qanday tizim uchun o'ta muhim va ochiq bo'lib qolmoqda. Yana bir eng katta muammosi ularni qayta ishlash xarajatlaridir. Bunga qimmat asbob-uskunalar, katta miqdordagi axborotlarga xizmat qo'rsatuvchi malakali mutaxassislar ish haqi xarajatlarini ham qo'shish mumkin. Shubhasiz, asbob-uskunalar muntazam yangilanib turishni talab etadi, ular ma'lumotlar hajmi oshishi jarayonida ish samaradorligini yo'qotmasligi kerak. Uchinchi muammo yana qayta ishlash talab etiladigan ko'p sonli axborot bilan bog'liq. Masalan, tadqiqotlar 2-3 martada emas, balki ko'p sonli o'rganishlarda natija beradi, chunki qaysidir hodisa bo'yicha haqiqiy ta'sirga ega bo'lish uchun ma'lumotlarni umumiy oqimdan ajratish va ob'yektiv baho berish juda qiyin.

Axborotni yo'qotish muammosi. Ehtiyot choralari bir martalik oddiy zaxira ma'lumotlarni cheklamaslikni, saqlovning kamida 2-3 ta zaxirasini tayyorlashni talab etadi. Ammo, hajmning oshib borishi zaxiralashni yanada murakkablashtiradi - AT-mutaxassislar mazkur muammoning optimal yechimini topishga harakat qilishmoqda.

Xulosa o'rnida aytish mumkinki, texnologiyalardan yashirinib ololmaymiz. "Big Data" va undan foidalanadigan AI texnologiyalari dunyoni o'zgartiryapti, asta-sekin shahrimizga, uyimizga va gadjetlarimizga kirib bormoqda. Bunga yaqqol misol "Big Data" va Mashina o'rganish AI texnologiyalari asosida ishlab chiqilgan ChatGPT platformasini ko'rishimiz mumkin. Bu texnologiyaning imkoniyatlari dunyoni hayron qoldirdi. Texnologiyalar sayyorani qanday tezlikda egallaydi - aytish qiyin. Ammo, bir tushuncha aniq - «Klinika» serialida Bob Kelso aytganidek, urf (moda)ni ushlab qol yoki eskicha yashab o'tib ket.

Foydalanilgan adabiyotlar ro'yhati:

1. Hilbert va Lopez, "Dunyoning ma'lumotni saqlash, aloqa qilish va hisoblash uchun

texnologik imkoniyatlari", Science, 2011 Global.

2. Sagiroglu, Seref, and Duygu Sinanc. "Big data: A review." 2013 international conference on collaboration technologies and systems (CTS). IEEE, 2013.

3. N.Rahmonov. X.Tolibjonov. Raqamli iqtisodiyot va Sanoat 4.0 da ilg'or davlatlar. "O'zbekiston umidli yoshlari" konfirensiya. 2021. 2-mart.

4. KHurshidbek T., Shakhzodbek Y., Shoxruxmirzo A. PLATFORM COMPETITION IN BANKING //YOUTH, SCIENCE, EDUCATION: TOPICAL ISSUES, ACHIEVEMENTS AND INNOVATIONS. – 2022. – T. 1. – №. 1. – C. 43-46.

5. Shakhzodbek E., KHurshidbek T., Shoxruxmirzo A. INNOVATIVE TECHNOLOGIES IN THE DEVELOPMENT OF DISTANCE LEARNING //BARQARORLIK VA YETAKCHI TADQIQOTLAR ONLAYN ILMIY JURNALI. – 2022. – C. 453-454.

7. KHurshidbek T., Shakhzodbek Y., Shoxruxmirzo A. DIGITAL BANKING MANIFESTO 2.0 //YOUTH, SCIENCE, EDUCATION: TOPICAL ISSUES, ACHIEVEMENTS AND INNOVATIONS. – 2022. – T. 1. – №. 1. – C. 55-59.

8. Shakhzodbek E., KHurshidbek T., Shoxruxmirzo A. CHRONOLOGY OF FINANCIAL TECHNOLOGIES //TA'LIM VA RIVOJLANISH TAHLILI ONLAYN ILMIY JURNALI. – 2022. – C. 213-215.