LECTURE-10: BIG DATA BASICS

МАЪРУЗА -10: КАТТА МАЪЛУМОТЛАР АСОСЛАРИ

Big Data – катта маълумотлар - бу катта ҳажм, тезлик, хилма-хиллик ва ишончлилик билан бошқариладиган маълумотлар тўпламидир. Бу маълумотларга ишлов беришнинг анъанавий дастурий таъминоти томонидан ҳал қилиниши мумкин бўлмаган жуда катта ёки мураккаб бўлган маълумотларни таҳлил қилиш, маълумотни мунтазам равишда чиқариб олиш ёки маълумотлар тўпламлари билан муомала қилиш усуллари билан ишлайдиган майдон ҳисобланади. Улар турли хил манбалардан - интернет, мобил қурилмалар, ижтимоий медиа, геокосмик қурилмалар, сенсорлар ва бошқа машина томонидан олинган маълумотлардан ташкил топган. МББТ ва маълумотлар омборидан фойдаланган ҳолда анъанавий маълумотларни қайта ишлаш ва тузилган маълумотларни таҳлил қилиш енди Катта маълумотлар муаммоларини хал қила олмайди. Катта маълумотлар технологиялари очиқ манбали дастурий таъминот ва оммавий равишда тақсимланган ишлов бериш платформаларини қамраб олади. Иқтисодиётнинг ўзгариши билан бир қаторда, технология асосий (mainframe) компьютер, шахсий компьютер, мижоз-сервер компьютерлари, Интернет, cloud computing, мобил компьютерлар ва ижтимоий тармоқларда ривожланмоқда.

Маълумотни қайта ишлаш хусусиятлари, маълумотларни тўплаш ва ташкиллаштиришни ўз ичига олади. Маълумотни моделлаштириш - мураккаб маълумотлар тўпламлари ва уларни визуал диаграмма намойиш этишдир. Қарор қабул қилиш учун ушбу маълумотлардан фойдаланишга ҳаракат қилаётган фойдаланувчилар учун ишнинг боришини осонлаштиради. Маълумотларни ишлаб чиқариш фойдаланувчиларга маълумотларни ҳар хил нуқтаи назардан ажратиб олиш ва таҳлил қилиш ва уларни амалдаги тушунчаларга умумлаштириш имконини беради. Бу вақт давомида тўпланган катта тузилмали маълумотлар тўпламларида айниқса фойдалидир.

Катта маълумотларни таҳлил воситалари Microsoft Access, Microsoft Excel, матнли файллар ва бошқа матн файллар каби манбалардан маълумотларни импорт қилишни ёқиши керак. Бир нечта манбалардан ва бир нечта форматдаги маълумотларни бирлаштириш имкониятига ега бўлиш, маълумотни ўзгартириш заруриятини олдини олиш орқали меҳнатни камайтиради ва тизимга тўғридан-тўғри импорт қилиш орқали бутун жараённи тезлаштиради. Экспорт имкониятлари ҳақида ҳам шундай дейиш мумкин - визуализация қилинган маълумотлар тўпламини олиш ва уларни PDF, Excel файллари, Word файллари ва .dat файллари каби экспорт қилиш аввалги жараёнларда тўпланган маълумотларнинг фойдалилиги ва ўтказувчанлиги учун муҳимдир.

Big Data Analytics воситалари фойдаланувчиларга турли хил таҳлил пакетлари ва модулларини таклиф қилади. Масалан, риск таҳлиллари ҳар қандай берилган ҳаракатлар атрофидаги ноаниқликни ўрганишдир. Ундан келажакдаги воқеаларнинг салбий таъсирини минималлаштириш учун

прогнозлаш билан биргаликда фойдаланиш мумкин. Хавфларни таҳлил қилиш фойдаланувчиларга ташкилотнинг сабр-тоқатлилиги ва хавф-хатарига аниқ жавоб бериш ва тушуниш орқали ушбу хавфларни камайтиришга имкон беради. Қарорларни бошқариш бизнес юритиш учун қарор қабул қилиш жараёнларини ўз ичига олади. Қарорларни бошқариш модуллари қарорларга фойдаланиладиган активлар сифатида қарашади. У қарорларни қабул қилиш жараёнининг қисмларини автоматлаштириш учун муҳим нуқталарда технологияни ўз ичига олади. Матнни таҳлил қилиш бу мижозлар томонидан ёзилган ёки ёзилган матнни ўрганиш жараёни. Таҳлил дастури ушбу матнда нақшларни топишга ёрдам беради ва ўрганган нарсангиз асосида бажарилиши мумкин бўлган ҳаракатларни таклиф қилади. Ушбу турдаги таҳлиллар, айниқса мижозларингизнинг еҳтиёжлари ва еҳтиёжлари тўғрисида тўғридан-тўғри ташкилотингиз билан ўзаро алоқада бўлганликлари тўғрисида маълумот олиш учун фойдалидир. Таркибни таҳлил қилиш матнни таҳлил қилиш билан жуда ўхшаш, аммо аудио, видео, расм ва ҳоказоларни, шу жумладан ҳужжатларнинг барча форматларини таҳлил қилишни ўз ичига олади. Ижтимоий медиа таҳлиллари - бу сизнинг фойдаланувчи базангизнинг ижтимоий медиада ўз брендингиз билан қандай муносабатда бўлишига қаратилган контент таҳлилининг бир шакли. Статистик таҳлиллар рақамлардан иборат маълумотлар тўпламларини тўплайди ва таҳлил қилади. Мақсад жами аҳолининг вакили бўлган умумий маълумотларнинг намунасини олишдир. Статистик таҳлил беш босқичда амалга оширилади:

Маълумотларнинг моҳиятини тавсифлаш, маълумотларни тақдим этган шахс билан боғлиқликни ўрганиш;

Уланишларни умумлаштириш учун модель яратиш;

Тўғрилигини исботлаш;

Рад этиш;

Қарорларни бошқариш учун башоратли таҳлилларни қўллаш.

Муваффақиятли бизнес учун тизимингизни хавфсиз сақлаш жуда муҳимдир. Big Data таҳлил воситалари хавфсизликни таъминлаш учун хавфсизлик хусусиятларини таклиф қилиши керак. Бундай хусусиятлардан бири бу битта тизимга кириш ёки SSO ҳам дейилади. Бу фойдаланувчиларга бир нечта дастурларга кириш учун кириш маълумотларини битта тўпламини тайинлайдиган аутентификация хизмати. У охирги фойдаланувчи рухсатларини тасдиқлайди ва бир сеанс давомида бир неча марта киришга эҳтиёжни йўқ қилади. Шунингдек, у тизимда ким нима қилаётганини кузатиб бориш учун фойдаланувчи фаолияти ва қайд ёзувларини қайд қилиши ва кузатиши мумкин. Катта маълумотларнинг таҳлил платформалари томонидан таклиф қилинадиган яна бир хавфсизлик хусусияти маълумотларни шифрлашдир. Маълумотни шифрлаш алгоритмлар ёки кодлар ёрдамида электрон маълумотни ўқиб бўлмайдиган форматга ўзгартиришни ўз ичига олади.

Hadoop катта маълумотларнинг асосий қисми ёки асосидир. Hadoop - бу арзон машиналар кластерида катта миқдордаги маълумотлар тўпламини тақсимланган ҳолда сақлаш технологияси ҳисобланади.

Катта маълумотлар базаларини сақлаш. Анъанавий МББТ катта ҳажмдаги маълумотларни сақлашга қодир эмас. Мавжуд МББТда маълумотларни сақлаш қиймати жуда катта. Бу аппарат ва дастурий таъминот учун ҳам қимматга тушади.

Турли хил форматларда маълумотларни қайта ишлаш. МББТ маълумотларни тузилган форматда сақлаш ва бошқариш имкониятига эга. Аммо реал дунёда биз маълумотлар билан тузилган, тузилмаган ва ярим тузилмали форматда ишлашимиз керак.

Маълумотлар юқори тезликда олинади. Маълумотлар ҳар куни тера-пета байтгача тартибда чиқарилади. Шундай қилиб, биз бир неча сония ичида реал вақт режимида маълумотларни қайта ишлаш учун тизимга муҳтожмиз. Анъанавий МББТ реал вақт режимида катта тезликда ишлашни таъминлай олмайди. Big data — бу структураланган ва структураланмаган маълумотларни, конкрет масалалар ва мақсадларда уларни қўллаш учун, ишлов бериш методлари, турли инстументлар ва ёндашувлардир. Структураланмаган маълумотлар - бу маълум тартибда ташкиллаштирилмаган ёки олдиндан аниқ структурага эга бўлмаган ахборот.

«Катта маълумотлар» терминини Nature журналининг редактори Клиффорд Линч 2008 йилда, дунёда ахборот хажмларининг ўсишига бағишланган махсус нашрида киритган эди. Шунга қарамасдан, албатта «Катта маълумотлар» олдинроқ хам мавжуд эди.

Мутахассисларнинг фикрича Big data даражасига кунига 100 Гб ортиқ барча маълумотлар оқими киради. Бугунда бу оддий термин остида иккитагина сўз ётади – маълумотларга ишлов бериш ва сақлаш. Замонавий дунёда Big data - катта миқдордаги маълумотларни тахлил қилиш учун янги технологиялар пайдо бўлиши билан боғлиқ ижтимоий-иқтисодий феномен.

Инсон аниқ ва унга керакли бўлган натижаларни олиш учун ва уларни келажакда самарали қўллаши учун катта хажмдаги ахборотларга ишлов берилади. Big data - бу муаммони ечими ва анъанавий маълумотларни бошқариш тизимларига альтернативдир.

Катта маълумотлар бу жуда хилма-хил, тез ўзгариб турадиган ёки анъанавий технологиялар, маҳорат ва инфратузилмани самарали ечиш учун катта ҳажмдаги маълумотларни қамраб оладиган технологиялар ва ташаббусларга тегишли. Аммо ҳозирги кунда янги технологиялар ёрдамида катта маълумотлар қийматини англаш жуда осон, масалан, харидорлар томонидан харид қилинадиган харидорларнинг ахлоқ тузатиш тенденцияларини аниқлаш, маҳсулотларнинг нархини белгилаш шулар жумласидандир.

Маълумотларнинг катта ечимлари битта машинадан минглаб машиналаргача бўлган ҳар бири маҳаллий ҳисоблаш ва сақлашни таклиф қиладиган hadoop-ga асосланади, бундан ташқари у "бепул" очиқ манбали

платформалар бўлиб, янги ташкилотни сотиб олишга сармоя киритишни минималлаштиришга имкон беради.

Катта маълумот технологияларининг ёрдами билан АТ-компаниялар учинчи томон маълумотларини тезкор равишда қайта ишлашга қодир.

Нақд пулни тўлдириш, тўлиқсиз ёки ноаниқ кредит лимитлари ёки нархлар тўғрисидаги маълумот мижозларга хизмат кўрсатишнинг йўқолишига олиб келади ёки даромадни камайтиради ёки хизмат нархини ошириши мумкин, катта маълумотлар технологиялари ва турли хил алгоритмларни тезроқ ишлатиш қобилияти билан маълумотлар янгиланиши мумкин. кун давомида мунтазам равишда янгиланади.

Маълумотларнинг тизимли таҳлили ёки мавжуд ҳолатларга мувофиқ тўғри бизнес қарорларини қабул қилишга олиб келадиган маълумотларнинг умумий ҳолатини баҳолаш учун ишлатилади, чунки баъзида нотўғри маълумотлар нотўғри бошқаришга олиб келса, бизнес қарорлари нотўғри маълумотларга асосланади ва у бизнес қулайди.

Катта маълумотлар учун «Уч V» деб номланувчи анъанавий аниқловчи характеристикаларни ажратиш мумкин.

Volume — физик хажмнинг катталиги.

Velocity — натижаларни олиш учун тезликнинг ошиши ва тезкор ишлов беришнинг тезлиги.

Variety — турли типдаги маълумотларга бир вақтда ишлов бериш имконияти.

Турли маълумотлар ҳажми ва тез келиб тушадиган сонли ахборотларга анъанавий инструментлар билан ишлов бериш имконига эга эмас. Маълумотлар таҳлилининг ўзи инсон кўра олмайдиган аниқ ва сезилмас қонуниятларни кўриш имконини беради. Бу бизнинг ҳаётимизда барча соҳаларни – давлат бошқарувидан то ишлаб чиқариш ва текоммуникацияларнинг оптималлаштириш имконини беради.

Катта маълумотлар технологияларидан фойдаланган ҳолда маълумотларнинг янгиланиши тезлиги корхоналарга мижозларнинг талабларига тез ва аниқ жавоб беришга имкон беради. Масалан, MetLife MongoDB-дан мижозлар маълумотларини 70 дан ортиқ турли манбаларда тезда бирлаштириш ва ягона, тез янгиланадиган кўринишни тақдим етиш учун фойдаланган. Катта маълумотлар корхоналарга рақобатчиларига қараганда ўзгаришларга тезроқ мослашишларига имкон берадиган даражада ҳаракатланишга ёрдам беради.

Катта маълумотлар технологиялари истеъмолчилар учун турли компанияларнинг "сотиб олиш" ва "сотиш" қарорларини башорат қилиш учун ишлатилади.

Search-Engine катта маълумотлар технологияларидан фойдаланган ҳолда турли хил маълумотлар базаларидан сонияларнинг сонияларида кўп сонли маълумотларни олиш. Масалан,Google MapReduce алгоритмидан берилган сўровни қидириш учун фойдаланади. MapReduce вазифани кичик қисмларга ажратади ва ушбу қисмларни тармоқ орқали уланган кўплаб

компютерларга тайинлайди ва натижани якуний натижани шакллантириш учун тўплайди. Молиявий хизматлар ташкилотлари мижозларнинг ўзаро муносабатлари тўғрисидаги маълумотларни қидириш учун катта маълумотлардан фойдаланиб, фойдаланувчиларни нозик сегментларга ажратиб олишади, бу эса тобора долзарб ва мураккаб таклифларни яратишга ёрдам беради.

Кластерлардан фойдаланиш кластер аъзолигини бошқариш, ресурсларни тақсимлашни мувофиқлаштириш ва алоҳида тугунларда ҳақиқий ишларни режалаштириш учун йечим талаб қилади. Кластерга аъзолик ва ресурсларни тақсимлаш Hadoop-нинг YARN (бошқа манбалар музокарачиси деган маънони англатади) ёки Apache Mesos каби дастурлар томонидан бошқарилиши мумкин.

GFT мақсади - катта файлларни сақлаш ва уларга кириш имконияти еди, ва умуман айтганда битта қаттиқ дискда сақланиб бўлмайдиган файлларни назарда тутилади. Ғоя бу файлларни бошқариладиган 64 МБ ҳажмдаги бўлакларга бўлиш ва ушбу бўлимларни бир нечта тугунларда сақлаш, шу билан бирга файллар тизимида сақланадиган қисмлар ўртасида харитани тузишдир.

GFT, кўпинча муваффақияциз бўлиши мумкин бўлган жуда арзон товар таркибий қисмларида ишлайди, шунинг учун муваффақияциз мониторинг ва тикланишни амалга ошириши керак. У бир вақтнинг ўзида кўплаб йирик файлларни сақлаши мумкин ва уларга икки хил ўқиш имкониятини беради: кичик тасодифий ўқишлар ва катта оқим оқимлари. Файлларни қайта ёзиш ўрнига, GFT тизимдаги мавжуд файлларга маълумотларни қўшиш учун оптималлаштирилган. GFT бош тугмаси файллар индексини сақлайди, GFT тармоқ серверлари еса бир нечта Linux тугунларида файл тизимларида ҳақиқий қисмларни сақлайди. GFT -да сақланадиган қисмлар кўпайтирилади, шунинг учун тизим сервернинг ишдан чиқишига тоқат қилиши мумкин. Текширув варақалари ёрдамида маълумотларнинг бузилиши ҳам аниқланади ва GFT ушбу ҳодисаларни имкон қадар тезроқ қоплашга ҳаракат қилади.

HTFT - да файллар блокларга бўлинади ва файлларга кириш кўп ўқийдиган ва битта-ёзувчи семантикасига мос келади. Хатоларга бардошлилик талабини қондириш учун турли хил DataNode-да файллар блокларга бўлинади ва файлларга кириш кўп ўқийдиган ва битта-ёзувчи семантикасига мос келади. Хатоларга бардошлилик талабини қондириш учун турли хил DataNode қувур линиясини яратади. (Репликация коеффициенти одатда қувурлар ичидаги DataNode сонини аниқлайди.) Кейинчалик ушбу блокга қувур линияси орқали ўтилади. Ўқиш операциялари учун мижоз блок нусхасини ушлаб турган DataNode -дан бирини танлайди ва ундан маълумот узатишни талаб қилади.

MapReduce бу харитани ёзиш ва қисқартириш функцияларидан иборат дастурий модел. Харита калит / қиймат жуфтлигини қабул қилади ва калит / қиймат жуфтликлари кетма-кетлигини ҳосил қилади. Кейин маълумотлар

гуруҳли калитларга бирлаштирилади. Шундан сўнг, қабул қилинган қийматларни бир хил калит билан қисқартирамиз ва янги калит / қиймат жуфтлигини ҳосил қиламиз.

Бажариш жараёнида Мап маълумотлари кириш маълумотлари асосида машиналарга берилади. Кейин ушбу Мап вазифалари ўз натижаларини беради. Кейинчалик, хариталаш натижаси аралаштирилади ва тартибланади. Кейин, қисқартириш вазифалари режалаштирилган ва бажарилади. Кичиклаштириш натижаси дискка сақланади.

Pythonda MapReduce кодларини кўриб чиқайлик.

Қуйидаги код - Харита функсиясидир. У STDIN- дан маълумотларни ўқийди, уни сўзларга ажратади ва сўзларни STDOUT- га (оралиқ) ҳисоблаш учун хариталар рўйхатини чиқаради. Мап скриптида сўзларнинг юзага келишининг ўртача (оралиқ) йиғиндиси ҳисобланмайди. Бунинг ўрнига, у дарҳол <word> 1 tuple чиқаради, гарчи маълум бир сўз киритишда бир неча бор пайдо бўлса ҳам. Бизнинг ҳолда, биз кейинги қисқартириш босқичига якуний суммани ҳисоблашга имкон берамиз.

Бу харитани қисқартириш ишида MongoDB ҳар бир кириш ҳужжатига (яъни сўров шартларига мос келадиган тўпламдаги ҳужжатлар) харита фазасини қўллайди. Харита функцияси калит қиймат жуфтлигини чиқаради. Бир нечта қийматга эга бўлган ушбу калитлар учун MongoDB йиғилган маълумотларни тўплайдиган ва сиқиб чиқарадиган камайтириш фазасини қўллайди. Кейин MongoDB натижаларни тўпламда сақлайди. Ихтиёрий равишда, қисқартириш функциясининг чиқиши, кейинчалик йиғиш ёки йиғиш натижаларини қайта ишлаш учун якунлаш функциясидан ўтиши мумкин.

MongoDB- да харитани қисқартиришнинг барча функциялари JavaScript-дир ва монгод жараёнида ишлайди. Харитани қисқартириш операциялари битта тўплам ҳужжатларини кириш сифатида қабул қилади ва харита босқичини бошлашдан олдин ихтиёрий тартиблаш ва чеклашни амалга ошириши мумкин. mapReduce харитани қисқартириш операциясининг натижаларини ҳужжат сифатида қайтариши ёки натижаларни тўпламларга ёзиши мумкин. Кириш ва чиқиш тўпламлари ўзгартирилиши мумкин.

Бошқа томондан, Салсите архитектураси мижозни, серверни ва таҳлил қилувчини олиб ташлайди ва оптимизаторга метадата ишлашда оғир ишларни бажаришга имкон беради. Калцит оптимизатори сўровларни оптималлаштириш учун 100 дан ортиқ қайта ёзиш қоидаларидан фойдаланади. Сўровлар реляцион алгебрадан фойдаланади, аммо рационал бўлмаган алгебрада ишлаши мумкин. Калцит сўровни бажариш учун энг кам харажатли усулни топишга қаратилган.

Катта маълумотларнинг асосий фунциялари:

- Маълумотларни қайта ишлаш:
  - Modeli

ng, ○
Data
Mining,
○ Data File
Sources, ○
File
Exporting,

- Башоратли Дастурлар (Predictive Applications);
- Таҳлил:
  - Хавф таҳлили,
  - Қарорларни бошқариш,
  - Контентни таҳлил қилиш,
  - Статистик таҳлил,
  - Башоратли таҳлил,
  - Ижтимоий Медиа таҳлили,
- Ҳисобот хусусиятлари:
- Хавфсизлик хусусиятлари;
  - Ягона кириш,
  - Маълумотни шифрлаш,
- Технологияларни қўллаб-қувватлаш:
  - A / B синов,
  - Hadoop билан интеграция.

Адабиётлар ва интернет сайтлари:

1. Радченко И., Николаев И., Технологии и инфраструктура Big data: Учебное пособие. – Санк-петербург: ИТМО, 2018.
2. https://computingforgeeks.com/data-mining-your-clicks
3. https://www.selecthub.com/big-data-analytics/big-data-analytics-requirements/
4. https://data-flair.training/blogs/hadoop-ecosystem-components/

5. https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology

6. https://medium.com/cracking-the-data-science-interview/an-introduction-to-big-data-distributed-data-processing-36654202c6ce