Дәріс 1. BigData тарихы. Деректерді сақтау, өңдеу құралдары, BigData пайдаланатын компаниялардың түрлері

Үлкен деректердің пайда болуы

Егер деректер болмаса, үлкен деректер болмас еді. Деректер түсініктіктің іргетасы болып табылады. Кейде деректер-ақпараттық-танымдық тізбек пирамида ретінде көрінеді, бұл ретте базада деректер мен жоғарыдағы білімдер бар. Ақпарат деректер негізінде құрылады. Біз байланысты деректер топтарын жинап, осылайша әлем туралы мағлұмат аламыз немесе айналамыздағы кеңістік туралы маңызды ақпарат аламыз. Бұл дәрістегі сөздер деректер болып табылады. Ақпарат — сөйлемдерге біріктірілген сөздер, абзацтарға бөлінген сөйлемдер, ал абзацтар мәтінге бөлінеді. Ал білім ақпараттан келіп түсті. Білім - ақпаратты пайдалану мақсатында түсіндіру: кітап оқып, ақпаратты өңдеп, пікір қалыптастырасыз, өз идеяларыңызды ойлап табасыз, шара қолданасыз.

Деректер сандар жиыны да болуы мүмкін, ол өз кезегінде әр түрлі тәсілдермен, мысалы, кестеде ұсынылуы мүмкін.

Егер сіз аңшы болсаңыз, онда сіз білесіз, мысалы, немесе сізге ең жақын орманда үйректердің қашан болатыны туралы ақпарат іздейсіз. Белгілі бір аумақтағы балық санын айлар бойынша жариялайтын арнайы басылымдар мен сайттар бар.

Осы ақпаратты пайдалана отырып, үйрек аң аулауға немесе балық аулауға қашан бару керектігі туралы шешім қабылдайсыз.

Көп сан тек қазіргі заманда ғана кездесетін сияқты көрінсе де, тарих бізді тастап кеткен мәтіндер мен жылнамалардан тек аз сандарды ғана көруге болады, солай емес. Оксфорд университетінде шамамен 5000 жыл болатын артефакт бар. Онда Перғауын Нармердің Ніл дельтасының батысындағы Ливанды жеңгені туралы баяндалады. Египетте 400 мың бұқа мен 1422 мың ешкі басып алып, 120 мың тұтқынды алып жатқаны сипатталған. Египеттің өлілер кітабында да жүз мыңдаған, миллиондаған адам айтылады. Сол кезең үшін бұл өте үлкен деректер.

Үлкен деректермен қиындықтар халық санағын жүргізуге байланысты пайда болды. АҚШ-тың алғашқы санағы 1790 жылы жүргізілді. Ол кезде АҚШ халқының саны 4 млн-нан небәрі 4 млн-ға дейін болған—3 929 326 адам, оның ішінде құлдар. 2010 жылы өткізілген соңғы санақ кезінде ел халқының саны қазірдің өзінде 308 745 538 адамды құрады. АҚШ Конституциясының 1-бабына сәйкес санақ он жылда кемінде бір рет өткізілуге тиіс. Ол «0» санымен аяқталатын жылдарда өткізіледі. 1790-1840 жылдар аралығында оны шерифтер жүргізді, ал 1840 жылы Санақ бюросының алғашқы орталық кеңсесі пайда болды.

Ал әр жолы санаққа қатысқан адамдар олардың алдына қойылған міндет сәтсіздікке ұшырайды деп ойлаған. Және мұның бәрі деректердің көлеміне байланысты. Олардың саны үнемі өсіп отырды, деректерді өңдеу және сақтау қажет болды, қол жетімді де ыңғайлы құралдар жетіспеді.

Алғашқы жылдары, әрине, бәрі қолмен жасалды. Адамдар кестелерді өздері аударып, оларға деректерді енгізіп, компьютерлердің көмегінсіз санап, қателерді болдырмау үшін бірнеше рет қайта есептеді. Кейде бір санақтан алынған деректер келесі санаққа дейін толық талданбаған! Ал олардың арасындағы кезең он жыл болды! Ал келесі санақ оған жауапты шенеуніктер үшін одан да қорқынышты болды, себебі халық саны жыл сайын өсіп, әрбір кейінгі санақта алдыңғысына қарағанда деректер көп болды.

Мәселе механизацияның көмегімен шешілді. 1890 жылы Герман Холлерит (1860 — 1929) деректерді өңдеу үшін халық санағында алғаш рет электр табуляция машинасын қолданды.

Табулятор - тесіп өткен карталарда жазылған сандық және алфавиттік ақпаратты автоматты түрде өңдеуге (қосындылауға және санаттауға) арналған электромеханикалық машина. Нәтижелері қағаз таспада немесе арнайы карталарда басылады. Ал электрондық компьютерлер пайда болғанға дейін бүкіл әлемде табуляторлар қолданылды. Табуляторлар қосуда және алып тастауда өте тиімді болды. Көбейту мен бөлу қиынырақ болды: ол қосудың бірнеше рет қайталануын талап етті немесе Алу. Көптеген өнертабыстар қол еңбегін жеңілдету үшін пайда болды. Табулятор солардың бірі. Перфокарталарға арналған идея АҚШ халық санағы бюросында жоғары лауазымды шенеунік болған Герман Холлериттің болашақ әке-шешесі Джон Биллингстен келді. Әрине, баласы көлікті ойлап тауып, тарихта өз атын қалдырған. Таблицаны жасаушы ретінде ғана емес, IBM-нің шөбеті ретінде де. Ол өзінің табуляциялық машиналарын шығаратын компания құрды, кейін оны сатты, және ол *Халықаралық бизнес-машиналардың* немесе *IBM-нің* құрамына кірді, қазір әлемдегі ең ірі аппараттық, бағдарламалық қамтамасыз ету, IT-қызметтерді, консалтингтік қызметтерді өндірушілер мен жеткізушілердің бірі.

Электрондық компьютерлер 40-жылдардың аяғында пайда болды. Қазіргі кезде санақ деректерін (үлкен деректер) өңдеу толық автоматтандырылған, дегенмен интервьюерлер деректерді жинаумен әлі де айналысып келеді. Өңдеу сол ережелер мен жоспар бойынша жүргізіледі. Алғашқы бітікші мұны армандай алмады!

Біз компьютер, есептеу техникасы, технологиялық төңкеріс дәуірінде өмір сүріп жатырмыз. Қазіргі заманғы технологиялар ақпаратты өңдеуге және сақтауға мүмкіндік береді.

Қазіргі заманғы технология бұрын-соңды жасалмағандай ақпаратпен айла-шарғы жасауға мүмкіндік жасады. Қазіргі заманғы машиналар деректерді өңдеп қана қоймай, оны ақпаратқа айналдыруға қабілетті. Біз интернетке, жаңа коммуникациялық жүйелерге қол жеткізе аламыз, кез келген уақытта әлемнің басқа бөлігіндегі адамға хабарласа аламыз. Ал тез және құрылымданбаған орасан зор көлемдегі деректермен жұмыс істейтін жүйелер оны өңдеу мен талдаудың жаңа технологияларынсыз жұмыс істей алмады. Бұл технологиялардың іске қосылғанға дейін үлкен көлемдегі деректермен жұмыс істеу іс жүзінде мүмкін болмады. Әдетте іріктеме қолданылды. Мысалы, халық санағын алыңыз. Ұйымдастырушылар барлық халықтың атынан үздік деп санаған шағын топ іріктемеге мұқият іріктеліп, талданды. Ол бүкіл халықтың жағдайын айқын көрсетті деп есептелді. Әрине, нәтижесінде қате пайда болды, бірақ ол кезде деректердің үлкен ағымына төтеп берудің ең жақсы құралы болды.

Терминдер мен сипаттамалар

«Үлкен деректер» терминін «Nature» журналының *редакторы Клиффорд Линч монеталаған*.Термин 2008 жылдың 3 қыркүйегінде арнайы шығарылым жарық көргенде дүниеге келді, оның тақырыбы «Деректердің үлкен көлемімен жұмыс істеуге мүмкіндік ашатын технологиялар ғылымның болашағына қалай әсер ете алады?» Осы арнайы мәселеде редакциялық алқа деректерді өңдеудің жарылу қаупінің өсуі және оның алуан түрлілігі, сондай-ақ осы құбылыстың технологиялық келешегі туралы материалдар жинады. Саннан сапаға көшу туралы алыпсатарлық орын алған.

Ағылшын термині *Big Data* қазіргілерге ұқсас түрде тұжырымдалды: *ірі бизнес* (ірі бизнес) — ірі корпорациялар; *Big Oil*, АҚШ-тың ірі мұнай компаниясы *Big Pathé* (сөзбе-сөз «үлкен есім») — цейлондық, белгілі тұлға.

Термин академиялық ортада пайда болып, пайда болғаннан кейінгі алғашқы жылы тек ғылыми деректердің көлемі мен алуан түрлілігінің өсуі туралы сөз қозғағанда қолданылды. Бірақ 2009 жылы ол бизнес-баспасөзге ағып кетті және өте тез кең тарады. 2010 жылы деректерді үлкен өңдеу проблемасына ғана қатысты алғашқы өнімдер мен шешімдер пайда болды. 2011 жылы ірі IT-жеткізушілердің көпшілігі қазірдің өзінде үлкен деректер тұжырымдамасын қолданды (мысалы, *IBM, Microsoft).*

Бұл тақырыпта жеке зерттеулер де пайда болды. Сондай-ақ, 2011 жылы ақпараттық технологиялар инфрақұрылымындағы No 2 тренд (виртуализациядан кейін) үлкен деректер аталды. 2013 жылы үлкен деректер деректер ғылымын зерттейтін американдық жоғары оқу орындарының оқу бағдарламаларына енгізілді. 2015 жыл үлкен деректерді жаппай практикалық қолдануға көшу жылы болып саналады.

Үлкен деректер ұғымы қазіргі ақпараттық қызметтің кез келген саласында дерлік қолданылады. Ең алдымен, бұл, әрине, IT-сала, сондай-ақ жарнама, сауда және маркетинг, мобильді технологиялар. Олар банк, телекоммуникация, энергетика, логистика, өнеркәсіп, мемлекеттік басқару салаларында қолданылады. Оларды бірінші болып метеорологтар қолданды.

Деректер көлемі үнемі өсіп келеді, Интернет барлық жерде бар, сондықтан кез келген бизнес бұл технология туралы ойлануға мәжбүр.

Бұл қазіргі заманғы ақпараттық кеңістіктің негізгі элементі. Ғаламдық ақпараттық салада жеке тұлға, адамдар тобы, жалпы адамзат, бизнестің түрлі салаларының компаниялары, үкіметтер жасап жатқанның барлығы дерлік орын алады.

Сіздердің жұмыстарыңыз, бос уақытыңыз, сауда-саттықтарыңыз, саяхаттарыңыз – бәрі де үлкен деректерге байланысты. Сіз алып, электрондық хаттар жіберіп жатырсыз, телефон қоңырауларын жасап, қоңырау шалып жатырсыз, Интернетті таң қалдырып жатырсыз, сол арқылы ақпарат биттерін алып, жіберіп, үлкен деректер жүйесінің ішінде тұрсыз. Қаржылық операциялар Интернет желісінде жүріп жатыр. Әлеуметтік желілерде жариялағанның бәрі World Wide Web сайтында қалады. Бұл деректер жоғалып кетпейді. Қазіргі адам үлкен деректерден қашып құтыла алмайды. Жеке тұлғаның физикалық тұрғыдан қабілетсіздігі және өзі табатын ақпараттық салада болып жатқан процестерді түсінуге уақыты болмайды.

Қазіргі уақытта үлкен деректер тәулігіне 100 Гб-тан астам деректер ағынына жатады.2003 жылы әлемде 5 эксабайт деректер болды (1 эксабайт 1 млрд гигабайтқа тең). 2015 жылы, 6,5-тен астам зеттабайт (1 зеттабайт = 1024 экзабайт) болды. 2020 жылға қарай 40-4 зеттабайт деректер болжанады. Ал 2025 жылға қарай бұл көлем 10 есе өседі. Global Big Data Revenue 2017 — US$ 150,8 млрд. Олардың көлемі соншалықты үлкен, мұндай үлкен көлемдегі деректерді стандартты бағдарламалық-аппараттық құралдармен өңдеу өте қиын, ал кейде жай ғана мүмкін емес.

Үлкен деректердің *негізгі сипаттамаларынан басқа — көлем, жылдамдық, сорт*, кейінірек пайда болған тағы төрт сипаттама бар. Бұл *құндылық, шынайылық, өміршеңдік* және *өзгергіштік*. Қанша V-ге қарамастанүлкен деректерді сипаттау үшін қолданылады, физикалық көлемнің негізгі еместігі әрқашан ерекшеленеді

немесе Үлкен деректердің айқындаушы сипаттамасы . Басқалары үлкен деректерді өңдеу және талдау міндетінің күрделілігін түсіну үшін де маңызды және қажет. Себебі, кез келген бизнес экономикалық тұрғыдан өміршең болуы тиіс, сондықтан «құндылық» жиі кездеседі

үлкен деректердің мінездемесінде, физикалық көлемді дәстүрлі танудан кейін *екінші* V болып шығады.

Big Data - бұл деректердің орасан зор көлемін талдау үшін жаңа технологиялық мүмкіндіктердің пайда болуымен байланысты әлеуметтік-

экономикалық құбылыс. Қарапайым адамдарға неге керек? Елестетіп көріңізші, сіз әдетте саудаға шығатын супермаркетте қандай да бір жұмбақ себептермен барлық өнімдер мен тауарлар аралас. Торт сүттің қасында болып шықты, ал шампунь қосылған нан, ет қосылған алма, шырын қосылған балық. Бәрін орнына қоюға көмектесетін үлкен деректер, Дұрыс өнімді тауып, жарамдылық мерзімі мен құнын анықтаңыз. Үлкен деректерді пайдалана отырып, белгілі бір өнімге ненің пайдалы, неге зияны бар екенін, ол үшін қандай ауруларды тұтынбау керектігін және ол үшін керісінше қажет екенін де біле аласыз. Супермаркетте не бары және не үшін екені туралы толық ақпарат, үлкен деректер бар. Деректердің орасан зор көлемі нақты адам оны одан әрі қолдану үшін қажетті нақты ақпаратты ала алатындай етіп өңделеді. Бұл деректерді басқару жеке тұлға, компания, қала, ел, әлем проблемаларын шешу болып табылады.

Дәстүрлі құралдарды пайдалана отырып, әртекті және тез ағатын ақпараттың орасан зор көлемін өңдеу мүмкін емес. Үлкен деректерді өңдеу мен талдаудың жаңа заманауи құралдары адам, тіпті ескі құралдар көре алмайтын заңдылықтарды көруге мүмкіндік береді.

Бұл біздің өміріміздің барлық салаларын – өндірісті, сатуды, телекоммуникацияны, тіпті мемлекеттік басқаруды оңтайландыруға септігін тигізеді. Үлкен деректер сізге бәсекелестік артықшылық береді.

Айталық, несие карталарының балансын білгіңіз келеді. Сұрауды өңдеу үшін секундтың бір бөлігі қажет. Бұл қазіргі заманғы ақпарат нарығының жылдамдығы. Үлкен деректер оларды талап етеді. Қазіргі заманғы технологиялар орташа пайдаланушыға гигабайт ақпаратты қалтасында және үйде сақтауға мүмкіндік береді, бизнес және мемлекеттік органдар бұрын ұнамсыз ауқымдағы деректерді жинауға, өңдеуге және талдауға мүмкіндік береді. Бұл қазіргі заманғы технологиялар – *Big Data технологиялары арқылы мүмкін болады*.

Деректердің көпшілігі кәсіпорындардың есебінен қалыптасады, жыл сайын ол барған сайын маңызды активке айналады, қауіпсіздіктің рөлі арта түсуде. Үлкен деректер үш дереккөзден келіп түседі. Біріншісі – Интернет, яғни әлеуметтік желілер, бұқаралық ақпарат құралдары, әр түрлі веб-сайттар, форумдар мен блогтар. Екіншісі – корпоративтік мұрағаттар. Үшіншісі – әр түрлі аспаптар мен құрылғылардың көрсеткіштері. Ең бастысы , осы орасан зор көлемдегі ақпаратты өңдеу мен талдауды үйрену. Себебі, үлкен деректер үнемі өзгеріп отыратын сурет. Егер үлкен деректерді дұрыс тапсаңыз, ол әрқашан іріктемені сүйемелдейтін дәлсіздіктерді еңсеруге көмектесіп қана қоймай, ғажайып мүмкіндіктер береді: өткеннің деректері қазіргі кездегі деректерге қосылады және бұл жақын болашаққа ең жақсы төтеп беруге көмектеседі. Себебі, дәстүрлі статистикалық талдауға қарағанда, үлкен деректерді барлық бағыттар мен трендтерді есепке алу үшін үнемі жаңартып отыруға болады.

Жоғарыда айтылғандай, синоптиктердің болжамынша, маусымдылықты біледі және осы факторды ескереді, бірақ үлкен деректер көптеген факторлар мен вариацияларды ескеруге мүмкіндік береді. Жаңа партияларды немесе деректер жиынын немесе деректер қатарын қосып, олардың қысқа мерзімді болжамдар жасауға көмектесетінін көруге болады. Мысалы, сатуды болжау төрт маусым мен мерекені бұрыннан есепке алған.

Бірақ қазір белгілі бір күндері ауа райының сатылымға қалай әсер ететінін көруге болады. Ал бұл ауа райына тікелей қатысы бар өнімдерге (қолшатыр немесе резеңке етік сияқты) ғана емес, шұжықтар мен ашықхаттарға да қатысты. Мұның бәрін физикалық тұрғыдан үлкен деректердің, сондай-ақ оны өңдеу мен талдау технологияларының болуы арқылы тексеруге болады. Егер қандай да бір фактордың сатуға айтарлықтай (немесе қандай да бір) әсер ететінін көрсек, оны тиісті күндерге

арналған сату болжамдарында ескеріп, сұранысты қанағаттандыру үшін қолдан келгеннің бәрін жасаймыз.

Үлкен деректер болжам жасауға көмектеседі, тек жыл мезгілдері мен күндеріне ғана емес, сонымен қатар сауда-саттық өтетін жерлерге де байланысты. Олар қандай өнімнің қай салада үлкен сұранысқа ие екенін зерттеуге мүмкіндік береді.

Мысалы, хаггис (қой оффшорының сұлы жармасымен, пиязбен, шұжық гильзасымен қоспасы) — Шотландияның ұлттық тағамы. Оны Шотландиядағы дүкендерден сатып алуға болады, АҚШ-та хаггис дәстүрлі түрде жейтін көптеген шотланд иммигранттары бар жерде сатылады. Бірқатар мемлекеттерде ол ешқашан естімеген. Желедегі эллин — лондондық тағам, ол Англияның басқа бөліктерінде және басқа елдерде танымал емес.

Қара пудинг Ұлыбритания мен Ирландияда танымал. Ал сатылымдар жергілікті сұранысқа негізделген «майда реттелген» болуы мүмкін.

Үлкен деректермен жұмыс істеудің үш негізгі талабы бар – қуатты компьютерлер, интернетке қосылу және дұрыс алгоритм. Сізде деректердің физикалық тұрғыдан керемет көлемі болуы мүмкін, бұл деректердің көп нүктелерімен үлкен байланыс болуы мүмкін, бірақ деректер мен қосылым жеткіліксіз. Бұл тіпті пайдалы емес. Олармен жұмыс істеу керек, ал адам деректердің аз мөлшерін бір уақытта өңдей алады. тіпті ең керемет математик. Тек оларды өңдей алмайсың. Бізге компьютерлік бағдарламалардың көмегі қажет, атап айтқанда, алгоритмдер қажет. Алгоритм — тапсырманы орындау үшін іс-әрекеттердің бірізділігі. Оксфорд ағылшын сөздігінде сөз көне грек тілінен «сан» үшін, тек «арифметика» сияқты келеді дейді. Дегенмен тағы бір нұсқасы бар, әрі дұрыс сияқты, себебі «әл» басталатын көптеген (егер бәрі болмаса) сөздер араб сөздерінен алынған. Бұл жағдайда алгоритм IX ғасырда өмір сүрген өзбек математигі, астрономы, философы және тарихшысы Мұхаммед әл-Хваризми есімінен шыққан деп есептеледі.Ең алдымен ол математик, ал арқасында

Ол үшін алгебра өз бетінше ғылымға айналды. Оның шығармалары бірнеше ғасыр бойы Еуропа университеттерінде математика бойынша негізгі оқулықтар болды. Латынданған нұсқасында оның есімі *Алгоризми* немесе *Алгоризмус сияқты естіледі.*Атауы ортақ зат есімге айналды, сондықтан еуропалық математиктер қатаң белгіленген ережелер бойынша кез келген есептеуді атай бастады. Кейінірек тұжырымдама кез келген қызмет саласында нәтижеге қол жеткізу тәртібін сипаттайтын нұсқаулар жиынтығына дейін кеңейтілді.

Бірақ сөздің шығу тегіне қарамастан, деректермен жұмыс істеуге мүмкіндік беретін процедуралар мен ережелердің жиынтығын білдіреді. Әр түрлі деректер жиынтығына бірдей рәсімдер мен ережелер қолданылуы мүмкін. Көптеген компьютерлік бағдарламаларға алгоритмдер жатады, бірақ алгоритм компьютерді қажет етпейді, алгоритмді қамтымайтын компьютерлік бағдарламалар да бар. Қарапайым алгоритмнің мысалы - Fibonacci сандары. Сандардың бұл тізбегі ғажайып ұзын, ал ол үшін алгоритм өте қысқа әрі қарапайым: әрбір дәйекті сан алдыңғы екі санның қосындысына тең. Яғни нұсқау болады: екі санды алып, келесі мәнді алу үшін қатардағы соңғы санды алдыңғысына қайта қосыңыз.

Егер үлкен деректер туралы айтатын болсақ, алгоритмдер өте күрделі болуы мүмкін. Бірақ олар әлі күнге дейін жүйеге деректерді талдауға немесе қалыптастыруға мүмкіндік беретін процедуралар мен ережелерден тұрады. Үлкен деректер жүйесі деректерге қол жеткізу, өңдеу және манипуляциялау үшін қолданылатын алгоритмдер сияқты ғана жақсы. Алгоритм бейтарап. Ол деректердің нені білдіретініне мән бермейді, тек біз сұрағанымызды ғана жасайды. Бірақ біз, үлкен деректерді пайдаланушылар ретінде, біздің жорамалдарымызға өте мұқият болуға және алгоритмнің не істеп жатқанын нақты білуіміз керек. Ең бастысы – нәтижелерді дұрыс

түсіндіру. Алгоритм жүйені пайдаланушылар туралы дұрыс жорамалдар жасау үшін оны әзірлеушілерге және деректерден жасалуы мүмкін тұжырымдар туралы дұрыс жорамалдар жасауға байланысты. Дұрыс емес жорамалдар көбінесе шешім қабылдау кезінде сәтсіздікке ұшырау себептері болып табылады.

Мысалы, бағдаршам қызыл түске айналғанға дейін сырғанап кетеді деп ойлайсыз. Егер ұшулар арасында бір сағат болса, ұшақтарды уақытында ауыстыра аласыз деп жорамалдайсыз. Бірақ қызыл жарық сіз ойлағаннан ертерек келіп, апатқа ұшырайсыз, немесе ұшақ кешігіп, әуежайда келесі байланыстырушы рейсті күтуге тура келеді. Сіз өз өміріңізде алғаш рет адаммен кездесіп, олардың сыртқы келбеті мен киімі негізінде ғана кейінірек ақталмайтын кейбір жорамалдар жасайсыз.

Адамдар не істей алатыны және жасай алмайтыны туралы үнемі жорамалдар жасап отырады, бұл адам табиғаты, бірақ бұл жорамалдар шығармашылық пен жаңа идеялар жолына түседі. Ал алгоритм жасаушылар да деректердің шектеулері мен оның қалай қолданылатыны туралы жорамалдар жасайды. Ал алгоритмдерге түзетулер енгізудің жолы болмаса, онда бұл жорамалдар деректерді дұрыс түсіндіріп, пайдалануды қиындата түседі.

Тағы бір мысал – «2000 жылғы проблема» деп аталатын мәселе. Компьютерлер 2000 жыл әлі де өте алыс болған 60-шы жылдары айтарлықтай кең тарады деп санауға болады. БАҒДАРЛАМАЛЫҚ ЖАСАҚТАМА ЖАСАУШЫЛАР XX ғасырда жылды білдіру үшін тек соңғы екі санды ғана жиі қолданған. Сәйкесінше, көптеген жүйелер жылды «19» деп бастады деп болжаған. Яғни келесі ғасыр тоғысында мұндай жүйелер 2015 жыл 1915 жыл деп болжауға болар еді. Бұл қаржы бағдарламалары мен процестерді бақылау жүйелерінің жұмыс істеуінде елеулі іс-қатерлерге әкеп соғуы мүмкін. Бағдарламалар 2000 жылы жұмысын мүлдем тоқтатуы мүмкін еді.

Мәселе бағдарламалық жасақтаманы әзірлеушілер ғасыр тоғысында не болуы мүмкін екенін ойламағандықтан туындады. Көп күш жұмсалды (қазір олар бұдан да көп нәрсені айтады) және кейбір мәліметтер бойынша 300 миллиард доллардан астам қаржы жұмсалды. Бірақ проблема дер кезінде анықталып, тиісті дайындық, тестілеу және алдын алу жұмыстары жүргізілді. Дегенмен, қазір пайда табу мақсатында оны «үркітті» деп айтқан. Әрине, ұшақтар мен банк жүйелерін басқаратын жүйелер тексерілуге тиіс еді, бірақ негізгі офистік бағдарламалық қамтамасыз ету емес. Қалай болғанда да үлкен де, кіші де сәтсіздіктер болған жоқ, бұл ең бастысы. Бірақ қазір тағы бір жайт – дұрыс емес жорамал туралы, керісінше, салғырттық туралы айтып отырмыз: сандардың «19»-дан «20»-ға ауысуы назарға алынбады. Бұл жай ғана әзірлеушілерге болған жоқ.

Ал осыған ұқсас жағдай үлкен деректер жүйелерімен де орын алуы мүмкін. Үлкен деректерге арналған алгоритмдерді әзірлеушілердің алдында үлкен жұмыс күтіп тұр. Мен олардың «2000 жылғы проблеманы» есте сақтағанын, барынша көп тест өткізгенін, мүмкіндігінше көп жорамалдарды тексергенін, сондай-ақ түзетулер енгізуді жеңілдеткенін қалаймын. Себебі, бірдеңе сөзсіз сырғанап кетеді, түзету енгізу керек болады, сондықтан бұл түзетудің құны 300 млрд доллар емес, арзан болсын. Сондықтан салдары туралы ойланайық [1].

Үлкен деректер технологиясы

Деректердің орасан зор көлемі адам оларды одан әрі тиімді қолдану үшін нақты және қажетті нәтижелер алуы үшін өңделеді.

Шын мәнінде, Big Data проблеманы шешуші және деректерді басқарудың дәстүрлі жүйелерінің баламасы болып табылады.

McKinsey мәліметі бойынша Big Data-ға қолданылатын талдау әдістері мен әдістері:

- Деректерді өндіру;
- Краудсорсинг;
- деректерді араластыру және біріктіру;
- машиналық оқыту;
- жасанды нейралды желілер;
- Үлгіні тану;
- болжамды аналитика;
- Имитация;
- Кеңістіктік талдау;
- статистикалық талдау жүргізу;
- Аналитикалық деректерді визуализациялау.

Деректерді өңдеуге мүмкіндік беретін көлденең масштабтау үлкен деректерді өңдеудің негізгі принципі болып табылады. Деректер есептеу түйіндері бойынша таратылады, ал өңдеу өнімділігі бойынша тозусыз жүреді. Сондай-ақ Мак-Кинси қолданыс контекстінде реляциялық басқару жүйелері мен Business Intelligence-ті қосты.

Технология:

- NoSQL;
- MapRedeuce;
- Хадооп;
- R;
- Аппараттық ерітінділер.

Үлкен деректер: қосымшалар мен мүмкіндіктер

Әртүрлі және тез ағатын цифрлық ақпарат көлемін дәстүрлі құралдармен өңдеуге болмайды. Деректердің өзін талдау адам көре алмайтын белгілі бір және мүмкін емес заңдылықтарды көруге мүмкіндік береді. Бұл біздің өміріміздің мемлекеттік басқарудан бастап өндіріс пен телекоммуникацияға дейінгі барлық салаларын оңтайландыруға мүмкіндік береді.

Big Data-Based Solutions: Sberbank, Beeline және басқа да компаниялар

Beeline абоненттер туралы орасан зор деректерге ие, оларды олармен жұмыс істеу үшін ғана емес, сонымен қатар сыртқы консалтинг немесе IPTV аналитикасы сияқты аналитикалық өнімдерді жасау үшін де пайдаланады. Beeline дерекқорды сегменттеп, деректерді өңдеу үшін HDFS және Apache Spark және Rapidminer және Python-ды пайдалану арқылы клиенттерді алаяқтық пен вирустардан қорғады.

Немесе Сбербанкті олардың «AS SAFI» деп аталатын ескі ісімен еске түсірейік. Бұл банк клиенттерін анықтау үшін фотосуреттерді талдайтын және алаяқтықтың алдын алатын жүйе. Жүйе 2014 жылы енгізілген, жүйе компьютерлік көрудің арқасында стеллаждардағы веб-камералардан сол жерге жететін деректер базасынан алынған фотосуреттерді салыстыруға негізделген. Жүйе биометриялық платформаға негізделген. Осының арқасында алаяқтық фактілерінің саны 10 есеге азайды.

Қазақстандағы үлкен деректер

2020 жылғы 4 наурызда Мемлекет басшысы Қасым-Жомарт Тоқаев цифрландыру мәселелері жөнінде кеңес өткізді. Оның барысында «Smart Data Ukimet» ақпараттық-талдамалық жүйесі көрсетілді. Бүгінгі таңда 10 жеке іс енгізілді және мемлекеттік органдардың 41 ақпараттық жүйесі қосылды. Жыл соңына дейін олардың санын 100-ге дейін ұлғайту және барлық негізгі әлеуметтік-экономикалық көрсеткіштерді алу мен көрсетуді автоматтандыру жоспарланып отыр.

Мемлекеттік органдармен бірлесіп талдамалық шешімдерді құру және 50-ге жуық жеке істі іске асырумен 15 бағытты қамту бойынша жұмысты жалғастыру жоспарланып отыр.

«SmartDataUkimet – бұл бізге азаматтық қорғаныстың барлық ұйымдарынан және түрлі көздерден ақпарат жинауға мүмкіндік беретін жоба, осылайша жасанды интеллект алгоритмдерін пайдалана отырып, мемлекет үшін талдамалық және әлеуметтік-экономикалық болжамдар жасауға бірегей мүмкіндік береді. Ең бастысы, ол делдалдарсыз ақпарат алуға мүмкіндік береді. Яғни, ең дәл, тікелей және әлдеқайда жылдам», - деп атап өтті «Ұлттық ақпараттық технологиялар» АҚ басқарма төрағасы Әсет Тұрысов.

Айта кетейік, ірі деректер талдауы ел экономикасын дамыту мен оңтайландырудың міндетті шарты болып табылады. Қазіргі кезеңде талдамалық деректер негізінде қабылданған басқарушылық шешімдер ұйымдастырушылық та, экономикалық та елеулі нәтижелерге қол жеткізуге мүмкіндік берді. Бүгінгі таңда денсаулық сақтау, қаржы, білім беру, әлеуметтік қорғау салаларында 10-нан астам жағдай іске асырылды. Жалпы экономикалық әсер 187 млрд. теңгеден астамды құрады. [2]

Қазақстандық нарықтың келешегі қандай?

IBM болжамы бойынша 2020 жылға қарай осы саладағы мамандар үшін 700 мыңнан астам бос жұмыс орны болады. Батыста трансформация басталды, егер қазір ТМД-дағы Big Data термині АТ мамандары арасында кеңінен танымал болса, нарық көшбасшылары классикалық талдаушыларға арналған оқыту бағдарламалары мен тренингтерімен бейімделе бастады. Бұдан басқа, дәстүрлі лауазымдардың функционалы кеңейеді, қызметкерлер үлкен деректермен сауатты жұмыс істеуді үйренеді, демек, осы технология ұсынатын жаңа артықшылықтарды пайдаланады. Нәтижесінде біз барлық қызметкерлер, тек IT-бөлімдердің өкілдері ғана емес, Big Data-мен жұмыс істеу әдістерін игеретіндей болып шығамыз.

Мысалы, келер жылы компаниялардың 62%-ы машиналық оқытуды және Big Data-ны талдаудың негізгі әдістерін енгізуді жоспарлап отыр, сондықтан да ұйымдар қызметкерлерді осы өзгерістерге бейімдеудің неғұрлым тиімді жолдарын іздеуі қажет.

Деректерді өңдеудің үлкен технологиялары енді ғана ішкі нарыққа шығып жатыр, шын мәнісінде қазір ешкім де ірі іске асыру мен нәтижелерді мақтан ете алмайды. Мемлекет тарапынан цифрландыру бойынша үлкен жұмыс жүргізілуде, оның дәлелі – «Цифрлық Қазақстан», «ақылды қала» және басқа да мемлекеттік бағдарламалар. Шетелдік компаниялар үшін тосқауыл елдегі халық санының аздығы, яғни іске асырудың сапасы, саны мен өзін-өзі ақтау кезеңі айтарлықтай артып келеді.

Қазақстанда мемлекет деректерді өңдеу бойынша ірі тапсырыс беруші болып табылады, тәуелсіздік алған кезеңнен бастап елде деректердің үлкен көлемі жиналды, оны өңдеу және бақылау және елден тыс жерлерде бәсекелесу үшін пайдалану қажет.

Ірі іске асырудың бірі Қаржы министрлігінде өтеді. Бұл сала экономика мен ЖІӨ статистикасы мен болжамы үшін стратегиялық маңызды болып табылады. Біздің елімізде бейресми экономика көлемі 26%, Ресейде – 39%, Украинада – 46%, ал

Әзірбайжанда экономиканың 67%- дан астамы көлеңкеде тұр. Стратегиялық міндеттердің бірі ақ нарық құру үшін шағын және орта бизнестің өркениеті болып табылады. Дәлірек талдау және тезірек нәтижелер алу үшін үлкен деректерді талдау таптырмас болып табылады.

Денсаулық сақтау министрлігі өркениеттің базалық деңгейіне кірісті. ӘӘШБЕА БҚ (Денсаулықтың электрондық паспорты) медициналық картаның тарихы бар бірыңғай деректер базасын құрады. Егер ауруханаға барсаңыз, онда мәліметтер компьютерге енгізіледі. Деректер файлын толық цифрландыру кезінде дәрігерлердің жұмысын болжауға және жетілдіруге болады.

Қазір Білім және ғылым министрлігінің дерекқоры eGov платформасындағы басқа мемлекеттік органдардың деректер базасымен біріктірілген. Министрлікте барлығы 73 мемлекеттік қызмет көрсетілген. Оның 25-і автоматтандырылған. Ұлттық білім беру деректер базасын (БҰДАН әрі - НОДБ) енгізу процесі жүріп жатыр, ол білім беру саласындағы бастапқы статистикалық деректерді жинау мен өңдеудің бизнес-процестерін автоматтандыруға арналған SEA (e-learning system) кіші жүйесі болып табылады. ҰБТ әкімшілік есептер үшін деректерді жинауды автоматтандырды, олар қолмен толтырылып, «білім беру ұйымы – білім беру бөлімі – білім бөлімі – Білім басқармасы – Қазақстан Республикасы Білім және ғылым министрлігі» тізбегі бойынша жиналды. Міндеттері: ведомстволық статистиканы бастапқы көздерден (білім беру ұйымдарынан) автоматты режимде жинау; деректерді сақтау және өңдеу; әкімшілік есептілікті дайындау; Қазақстан Республикасы Білім және ғылым министрлігінің құрылымдық бөлімдерін жұмыс істеу үшін қажетті статистикалық деректермен қамтамасыз ету. ҰБТ оқушылардың толық есебін қамтамасыз етеді; қайталауды жою арқылы респонденттердің дәйексіз мәліметтерін анықтайды; білім беру ұйымдарының паспорттарын толтыру тәртібін жеңілдетеді; статистикалық деректердің тарихи қатарын қалыптастырады; Ad hoc есептерін жасауға мүмкіндік береді. Ол 2020 жылға қарай толық іске асырылатын болады.

Сондай-ақ, бизнес үлкен мәліметтерге қызығушылық танытып отыр, банктер үлкен қызығушылық танытып отыр, жалпы, деректерді сақтайтын кез келген ірі немесе орта бизнес. Бұған «Қазпошта» АҚ мен «Қазақтелеком» мысал бола алады, олар қазірдің өзінде қызығушылық таныта бастады.

Егер болашаққа көз жүгіртсек, деректердің үлкен нарығы өте перспективалы. Республикада сұраныс жыл сайын артып келе жатқан мамандар жоқ. Junior Data Scientist нарығындағы орташа жалақы айына 200 мың теңгеден 500 мың теңгеге дейін ауытқиды[3].

Сілтемелер :

1. Жай үлкен деректер. Санкт-Петербург қаласы, Страта Публ., 2019 жыл. - 148 б.

2, Қазақстан Республикасында үлкен деректерді жинаудың, талдаудың және болжаудың бірыңғай орталықтандырылған құралы іске асырылды. Https://profit.kz/news/57188/V-RK-realizovan-Edinij-centralizovannij-instrument-sbora-analiza-i-prognozirovaniya-Big-Data/ (қол жетімді: 12.09.2020).

3, Ахметов С. Қазақстандағы үлкен деректер: Ірі тапсырыс беруші, кадрлар және перспективалар / Капитал туралы, 09.08.2018 жыл. URL: https://kapital.kz/tehnology/71257/big-data-v-kazakhstane-o-krupnom-zakazchike-kadrakh-i-perspektivakh.html (қол жетімді: 12.09.2020).