Дәріс 1

Үлкен деректерді талдаудың негізгі анықтамалары, терминдері, міндеттері

Ақпараттық технологиялар дамуының нәтижесі болып тез қарқынмен өсіп келе жатқан, электрондық түрде жинақталған деректердің өте үлкен көлемі болып табылады. Бұл ретте деректер, әдетте, әртүрлі құрылымға ие болады (мәтіндер, суреттер, бейне-жазбалар, аудио, гипермәтіндік құжаттар, реляциялық деректер қоры). Ұзақ уақыт бойы жинақталған деректер жоспарлау, болжам жасау, шешім қабылдау, процесстерді бақылау кезінде құнды ақпарат болып табылатын заңдылықтарды, үрдістерді және өзара қарым-қатынастарды өзіне қамти алады. Алайда адам біртекті емес мұндай деректердің көлемін тиімді талдауға физикалық жағынан қабілетті емес.

Үлкен деректер деректерді интеллектуалды талдаудың кеңірек әдістерінің танымалдылығының өте тез өсуіне алып келді, себебі, ақпарат одан да көп бола бастады, және ол өзінің табиғаты мен мазмұны бойынша әртүрлі және кең болды. Үлкен деректер жиынымен жұмыс істеген кезде бұдан былай салыстырмалы қарапайым және түзусызықты статистика жеткілікті болмады. Сатылымдар туралы 30 немесе 40 миллион нақты жазбаларға ие бола отырып, олардың екі миллионы бір жерде жасалғаны туралы білу жеткіліксіз. Сатып алушылардың қажеттіліктерін одан да жақсырақ қанағаттандыру үшін, сол екі миллион сатылымның белгілі бір жастық топқа жатуын түсіну және олардың орташа еңбек ақысын білу қажет. Бұл бизнес-талаптар деректерді қарапайым іздеу және статистикалық талдаудан деректердің одан да қиынырақ интеллектуалды талдауына алып келді.

Математикалық статистиканың дәстүрлі әдістері ұзақ уақыттан бері деректерді талдаудың негізгі құралы рөліне үміткер болған. Бірақ олар жаңа болжамдарды синтездеуге мүмкіндік бермейді, тек алдын-ала тұжырымдалған болжамдарды растау үшін және деректердің жедел аналитикалық өңдеуінің (online analytical processing, OLAP) негізін құрайтын барлау талдауы үшін қолданылуы мүмкін. Көп жағдайда болжамды тұжырымдау болашақ шешім қабылдау үшін талдау жүргізу кезіндегі ең қиын міндет болып табылады, өйткені деректердегі барлық заңдылықтар алғашқы көргеннен айқын болмайды. Сондықтан деректерді интеллектуалды талдау технологиялары (Data mining) зерттеу үшін және ақпараттық технологиялар саласында қолдану үшін ең маңызды және перспективті тақырыптардың бірі болып қарастырылады. Data Mining мақсаты үлкен көлемді (өте үлкен) деректердің жасырын ережелері мен заңдылықтарын анықтаудан тұрады. Себебі, адамның ақыл-ойы өзімен өзі орасан зор алқаптағы әртүрлі ақпаратты қабылдау үшін бейімделмеген. Орта есеппен адам, кейбір жеке тұлғаларды есептемегенде, тіпті шағын таңдаулар ішіндегі екі-үш өзара байланысты қабылдауға қабілетті емес. Бұл жағдайда деректерді интеллектуалды талдау астарында жаңа, дұрыс және деректердің үлкен көлемі

негізінде пайдалы болатын білімдерді анықтау процессі түсініледі. Мысалы, MIT Technology Review Data Mining-ті әлемді өзгертетін, дамып келе жатқан он технологияның бірі ретінде сипаттады. Бүгінгі күні
ғылым болжау технологияларын әзірлеуде бірталай алға жылжыды. Мамандарға болжаудың нейрондық желілер әдісі, айқын емес логика және т.б. әдістері өте жақсы белгілі. Сәйкес келетін бағдарламалық пакеттер әзірленген, бірақ олар тәжірибе жүзінде, өкінішке орай, қарапайым қолданушыға әрқашан қол жетімді болмайды.

- Компанияларда о баста жоспарлау және есеп беру жүйесі болуы тиіс. Ол мәліметтердің қорын, аналитикалық құралдарды және есептілік құралдарын қамтиды. Қажетті бағдарламалық қамтамасыз етуді ұсынатын бірнеше компания бар. Жеткізушілерді шолып шығудың жақсы тәсілдердің бірі Gartner Magic Quadrant for Analytics and Business Intelligence пайдалану болып табылады (суретті қараңыз).

Осы бейнеден көріп отырғанымыздай, ең үздік өнімді Microsoft ұсынып отыр.

Big Data клиенттерді, олардың тәртібін және артықшылығын дұрыс түсіну үшін пайдаланылады. Компаниялар өз клиенттерінің толық көрінісін алу үшін өз клиенттік базаларын әлеуметтік желілер, браузер журналдары, сонымен қатар мәтіндік аналитика мен тетіктер мәліметтері арқылы кеңейтуге ұмтылады. Бұның басты мақсаты көптеген жағдайларда болжамалы үлгілерді құру болып табылады.

Жалпы, Big data технологиясын практикалық тұрғыда пайдаланудың бірнеше мысалы бар:

- Денсаулықты сақтандыру 1. Сақтандырушылар егер адам салауатты өмір салтын ұстанса, денсаулықты сақтандырудың кейбір өнімдеріне жеңілдіктерді (немесе үстамақы) жасайды. «Салауатты өмір салты» деп спортпен шұғылдануды, дене салмағының үйлесімді индексіне ұмтылысын, темекі шегуде шектеу қою, ұйқы, демалыс және жұмыс арасындағы дұрыс теңгерімнің болуын айтады. Бұл ақпарат тұтынушылар өзімен бірге ұстап жүруге тиіс тетіктер арқылы жиналады. Мұндай өнімдер, мысалы, Скандинавияда бар, сонымен қатар Эстонияда да пайдалана бастады.

- Денсаулықты сақтандыру 2. Медбикелер қызметкерлеріне көмек көрсету. Тетік арқылы пациенттің стандартты белсенділігін бағалауға және нормалардан ауытқуларды бақылауға болады. Сондай-ақ пациент осындай құрылғыларды пайдалана отырып, өзі көмекке медбикені шақыра алады. Егер пациент аталған

аумақтан тыс шығатын болса, бұл құрылғы ол туралы хабарлай алады. (https://www.vivago.com/solutions/assisted-living/).

- Шетелде сырқаттан / жазатайым оқиғалардан сақтандыру Адам шетелге кететін кезде, оған шетелдегі сақтандыру туралы оферта жіберіледі. Литвада осындай сату арнасы іске қосылған.

- КАСКО (ЖБКАС) немесе ОГПО ВТС (КҚИ АҚЖМС) сақтандыру. Әрбір көлікте қозғалтқыш, оны айдау (жылдамдығы, бұрылуы, тежелуі, жұмыс-демалыс режимі), отынды пайдалану туралы және т.б. деректерді тіркейтін тетікті қоюға болады. Жаңа көліктерде барлық тетіктер біріктірілген. Мәліметтер көлік өндірушілері немесе телематика қызметтерін ұсынатын серверлерінде жинақталады. Бұл деректер әрбір 5-15 секунд сайын деректер базасына жіберіледі.

Ақпаратты стандарттауға және жүргізушінің профилін сипаттау үшін пайдалануға болады. Профиль сақтандыру оқиғаның тәуекелін көрсетеді. Менеджмент жүргізудің тәуекелін төмендету және сақтандыру оқиғаларын төмендету үшін (бұл сақтандыру сомасына әсер етеді) қажетті шараларды қабылдай алады.

- КАСКО (ЖБКАС) сақтандыру. Құрылғы меншік иесіне ұрлық/бұзып кіру сәтінде хабарлай алады.

Microsoft (SQL Server + SharePoint) өнімдері деректер мен есептілікті визуалдау үшін шығару, қайта өзгерту және аналитикалық құралдарға жүктеу үшін пайдаланылады. Әрбір компанияда бизнесті талдау және есептілікт үшін деректер базасы болуы тиіс. Үлкен деректер – бұл ағымдағы datawarehouse-тан тыс орналасқанның бәрі.

Big Data технологиясының басқа мүмкіндіктері:

- деректерді дайындау: ақпаратты деректер базасында сақтау процедуралары, деректерді пайдалану үшін құрылымдау.

- деректерді үлгілеу (сегментация, профильдерді құру, үлгілеу және болжам жасау).

-іс жүзінде пайдалану: сату, маркетинг немесе жеңілдіктер процедуралары үшін триггерлерді құру.

- Түрлі адамдар, әлеуетті клиенттер туралы деректерді түрлі дереккөздерінен жинаудағы этикалық жік қайда орналасқан? Мамандар, оның ішінде бұл

технологияны кеңінен қолданып жүрген Ресей мамандары да жалпы үлкен деректерді реттеу жүйесінің жоқтығы, сонымен қатар Big Data қорғау бойынша әдіснамалар мен стандарттардың жоқтығы туралы айтады. Түрлі ұйымдар өз ұсынымдары мен әдіснамаларын жариялайды, алайда ISO деңгейіне әлі де ешқайсысы жетпеген.

Сонда кез келген сауатты IT маманы адамдардың әлеуметтік желідегі профильдеріндегі олардың мәліметтерін бүлдіріп қою мүмкін бе? Мұнда ұстап тұрушы факторлары бар ма?

- Жалпы, Сіз дұрыс айтасыз. Кез келген сауатты IT маманы адамдардың әлеуметтік желілердегі профильдеріндегі мәліметтерін бүлдіре алады. Бұны қалай істеу керектігін білсең, бұл қиынға соқпайды.

Мұндағы білім жалпы ақпараттан алынатынын атап кету маңызды. Қорытынды жасау үшін деректер көлемі қажет. Әлеуметтік желілерде көптеген клиенттер бар, бірақ жеке деректердің жайылып кетуі бойынша соңғы жанжалдардан кейін оларға қалай қол жеткізуге болады?

Өмірді сақтандырушылары үшін әлеуметтік желілерден алынған деректерден гөрі, клиенттің денсаулығы туралы ақпаратты алу маңызды болмақ. Едәуір жеке деректер қажет.

Еуропада басқа компанияларға жеке деректерді жөнелтуге шектеулерді қою жоспарлануда. Клиенттің келісімі үлкен рөл атқарады. Бірақ егер қызметтерді тегін (немесе жеңілдікпен) пайдалануға болатын болса, жалпы, клиенттер деректерді өздері-ақ беретін болады.

- Қазіргі таңда Батыста Facebook-тегі американдықтардың 50 миллион профилінің деректерімен туындаған жанжал белсенді түрде талқылануда, бұл мәліметтерді Cambridge Analitica компаниясы АҚШ-та 2016 жылғы сайлаудың нәтижелеріне ықпал ету үшін жинаған. Facebook пайдаланушылардың деректерін осы жанжалға қатысты қорғаудың қолданылатын шаралар туралы жариялады. Сіз Big Data этикалық пайдалану тұрғысынан бұл жағдайды қалай түсіндірер едіңіз?

- Егер бұл шындық болса, онда бұл Big Data технологиясын қалайша пайдалану керектігі бойынша өте жақсы мысал. Идея қарапайым – клиенттің профилін дәлме-дәл жасап, қызметтер туралы нақты ақпаратты ұсыну. Бұл клиенттермен байланыс орнатудың ең тиімді тәсілдердің бірі болып табылады. Көбіне

жарнама қажет етпейтін адамдарға арнап жасалады, нүктеленген жарнама клиенттерді тарту үшін ресурстарды тиімді пайдалануға мүмкіндік береді.

Этикалық қағидалар бұзылады, өйткені «пайдалы» ақпарат табандылықпен ұсынылады. Бұл индивидтің таңдау құқығына, жеке өміріне қол сұқпаушылыққа және дербес деректерін қорғауға құқығына әсер етеді.

- Осы жылдың 25 мамырдан бастап ЕО Еуропалық Одақ елдерінде Деректерді қорғаудың жалпы қағидалары (GDPR) қолданысқа енгені туралы хабарлады. Мысалы, енді әрбір пайдалнушы осы қағидаларға сәйкес оның барлық деректерін жоюға құқығы бар, яғни «ұмытылып қалған құқығына» ие бола алады. Сонымен қатар егер оның талаптары 72 сағат ішінде орындалмаған болса, онда орындаушыларға олардың жыл сайынғы әлемдік табысынан 2% немесе 10 млн евро мөлшерінде, қайсысы көбірек болғанына қарай, айыппұл салынады. Адамның деректерін басқа елге заңсыз жіберген немесе пайдаланушы туралы деректерді қорғау қағидаларын бұзған кезде деректерді қорғау бойынша орган 20 млн евро мөлшерінде айыппұл салады. Сіз бұл жағдайды қалай түсіндіресіз?

- Бұл Регламентте 21-ғасырдың деректерді қорғаудың тәсілдемесі қолданылады. Ол жеке тұлғалардың жеке ақпараты қалай жинақталып, өңделетінін бақылауға құқықтарын кеңейтеді, сонымен қатар ұйымдар бойынша бір қатар міндеттемені белгілейді.

Бұл деректерді қорғау мәселесіне баса назар аударту керектігін білдіреді. Деректер базасына әрбір қосылуын түсіндіру қажет, онда барлық іс-әрекеттер жазылады. Бұл процедуралар белгілі бір бюрократияны енгізеді, бірақ келешекте әр түрлі болып жұмыс істеуге болмайды.

Егер қысқаша айтатын болсақ, Еуропада жақын арада күшіне енетін қағидалар және директивалар (The EU General Data Protection Regulation (GDPR):

- Жеке деректері бар серверлер елде болу керек (немесе дәл сол экономикалық кеңістікте)
- Жеке деректер клиенттің келісімісіз басқа тараптарға берілмеуі тиіс
- Тиісті қауіпсіздікті, тұтастықты және құпиялылықты қамтамасыз ету
- Деректерді өңдеу операцияларының барлығын толық жазбаларын жүргізу. Яғни деректерге не болатынын тіркеу қажет
- Дербес деректердің қауіпсіздігін қамтамасыз ету бойынша тиісті шараларды енгізу
- Қажет болған жағдайда деректерді қорғау бойынша қызметкерді тағайындаңыз
- Аудит, деректер қалай қолданылады.

Жалпы, жағдай қиын емес: әлеуметтік желілерден және басқа деректер базасынан жеке деректерге ие болғандар саяси, сондай-ақ экономикалық пайдаға қол жеткізе алады. Мүмкін, сондықтан болар Facebook, Google, LinkedIn Қытайда жұмыс істемейді (олар өздері-ақ өз клиенттері үшін ұқсас желіні жасап жүр). Сондай-ақ кейбір бағдарламалар Ресейде жұмыс істемейтін болды.