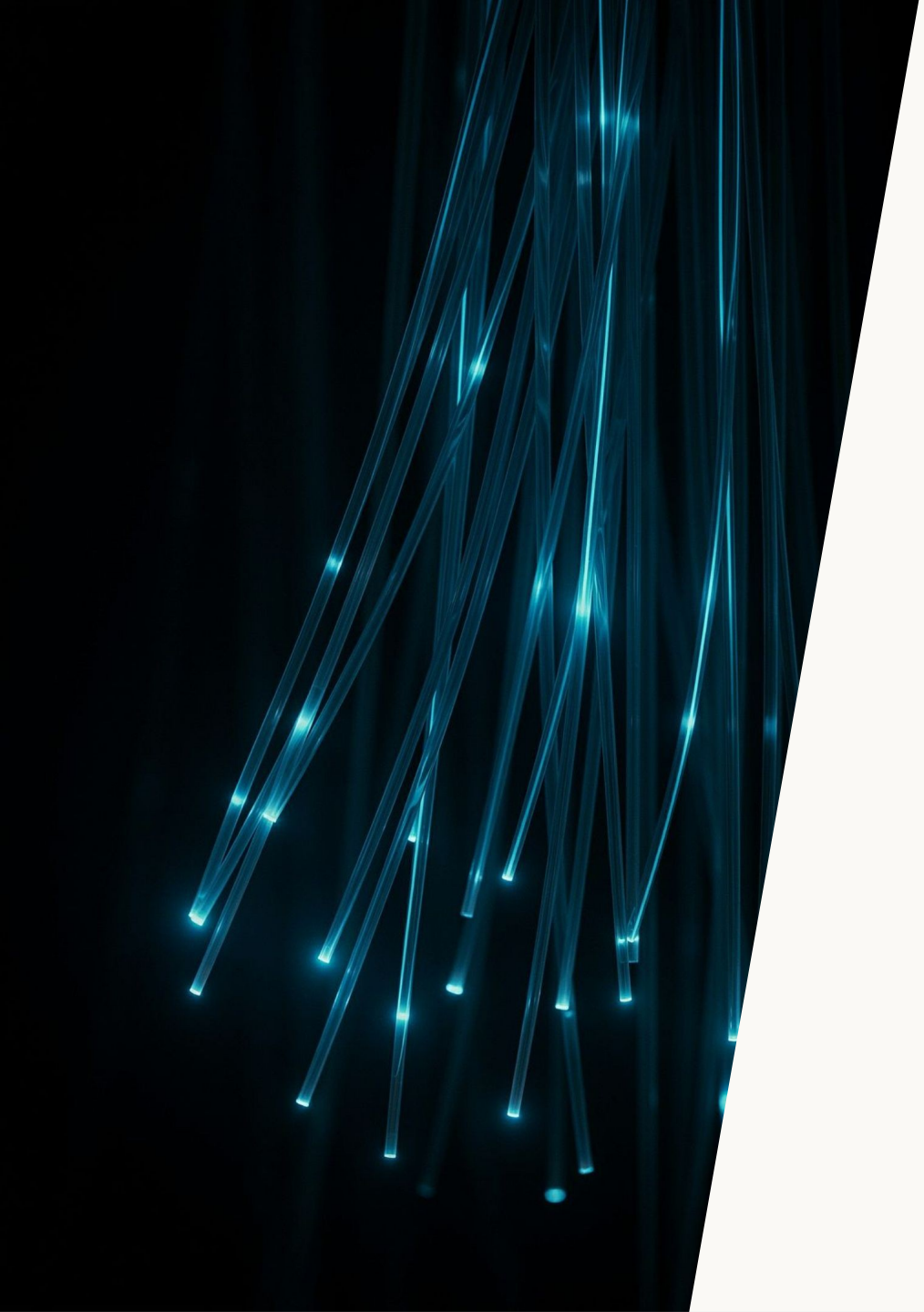
СӘТБАЕВ УНИВЕРСИТЕТІ

SATBAYEV UNIVERSITY

Статистика және деректерді талдау

Ержан Асел Ануаркызы,Ph.D., «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры

Assel.yerzhan@satbayev.university



№2. Деректерді жинау және дайындау

Деректер әлеміне саяхат

Деректердің рөлі

Заманауи қоғамда деректер — жаңа мұнайға теңелген стратегиялық ресурс. Медицина, білім, бизнес, мемлекеттік басқару — барлық салада шешімдер деректерге негізделеді. Дұрыс жиналған және өңделген ақпарат ұйымдарға бәсекелестік артықшылық береді.

Бұл курстың жоспары

- Деректер базасының негіздері мен терминологиясы
- Деректерді жинау стратегиялары мен әдістері
- Өңдеу құралдары: SQL, триггерлер, процедуралар
- Заманауи аналитика және визуализация
- Деректер қауіпсіздігі және этика



Деректер базасының негіздері

Деректер базасы (ДҚ) ұғымы

ДҚ — белгілі бір тақырыпқа немесе мақсатқа байланысты ұйымдасқан деректер жиыны. Ол ақпаратты тиімді сақтауға, іздеуге және өңдеуге мүмкіндік береді. ДББЖ (Деректер базасын басқару жүйесі) арқылы деректерге қол жеткізу басқарылады.

Негізгі терминдер

Өріс — деректердің бір атрибутын сақтайтын ұяшық (мысалы: аты, жасы). **Жазба** — бір объект туралы толық ақпараттың жиыны (кестедегі жол). **Кесте** — бірыңғай типтегі жазбалардың реттелген жиыны. Бұл үш элемент ДҚ-ның негізгі конструктивтік бірліктері болып табылады.

ДББЖ жүйелері

Oracle, Microsoft SQL Server, PostgreSQL, MySQL — заманауи кәсіби деректер базаларын басқару жүйелері. Олар деректердің тұтастығын, қауіпсіздігін және қол жетімділігін қамтамасыз ету үшін күрделі механизмдерді пайдаланады.

Деректерді ұйымдастыру түрлері

Реляциялық (кестелік) ДҚ

Деректер кестелер түрінде ұйымдастырылады. Кестелер арасындағы байланыс кілттік өрістер арқылы жүзеге асырылады. SQL тілін пайдаланады. Мысал: Oracle, MySQL, PostgreSQL.

Иерархиялық модель

Деректер ата-бала қатынасы бойынша ағаш тәрізді ұйымдастырылады. Әр баланың тек бір атасы болады. IBM IMS жүйесі бұл модельдің классикалық мысалы болып табылады.

Желілік модель

Иерархиялық модельден ерекшелігі — бір баланың бірнеше атасы болуы мүмкін. Күрделі байланыстарды бейнелеуге ыңғайлы. CODASYL стандартына негізделген жүйелерде қолданылады.



Деректерді жинаудың мақсаты мен стратегиясы

Ақпараттың сапасын арттыру

Деректерді жинаудың басты мақсаты — шешім қабылдауға жарамды, нақты және сенімді ақпараттық базаны қалыптастыру. Сапалы деректер бизнес-процестерді оңтайландырады, тиімділікті арттырады және болжамды аналитиканың негізін жасайды.

- Деректердің толықтылығын қамтамасыз ету
- Ақпараттың өзектілігін сақтау
- Жүйелі жаңартуды ұйымдастыру

Қателіктерді азайту

Деректер жинау стратегиясы қателіктерді бастапқы сатыда болдырмауға бағытталуы тиіс. Автоматтандырылған тексеру, деректерді форматтау ережелері және кіріс деректерін валидациялау жүйелері ақпараттың жоғары сапасын кепілдендіреді.

- Автоматтандырылған валидация
- Қос тексеру механизмдері
- Деректерді тазалау алгоритмдері

Жобалаудың маңызы

Тұжырымдамалық модельдеу

Жобалаудың алғашқы кезеңінде жүйенің жалпы концепциясы анықталады. Мұнда субъектілер, олардың атрибуттары және өзара байланыстары бейнеленеді. ER-диаграммалар тұжырымдамалық модельдеудің негізгі құралы болып табылады.

Бизнес-процестерді талдау

Тиімді деректер базасын жасау үшін ұйымның бизнес-процестерін терең зерттеу қажет. Қандай деректер жиналады? Кім пайдаланады? Қандай есептер қажет? Осы сұрақтарға жауап жобалаудың бағытын айқындайды.



Үлкен ұйымдардағы деректер архитектурасы

Интегралданған жүйелер

Үлкен корпорацияларда деректер әртүрлі бөлімдер мен жүйелер арасында үздіксіз ағып жатады. ERP, CRM, SCM жүйелері бірыңғай деректер ортамен интеграцияланады. Мұндай архитектура мәліметтердің дублануын болдырмайды және жалпы жүйе тиімділігін арттырады.

Таратылған деректер қоры

Физикалық тұрғыдан әр жерде орналасқан, бірақ логикалық тұрғыдан бірыңғай жүйе ретінде жұмыс істейтін ДҚ-лар. Таратылған жүйелер үздіксіз жұмысты, жоғары қол жетімділікті және географиялық икемділікті қамтамасыз етеді.

- Репликация және синхронизация
- Отказды өңдеу мүмкіндігі
- Жүктемені теңестіру

Жобалау кезеңдері: Түсініктен жүйеге

Тұжырымдамалық дизайн

ER диаграммаларын және модельдерді құру.

Логикалық дизайн

Логикалық схема және нормализация жасау.



Талаптарды талдау

Қажеттіліктер мен мақсаттарды анықтау.

Физикалық енгізу

Жабдық пен бағдарламаны орналастыру.

Деректер базасын жобалау бірнеше кезеңнен тұратын итерациялық процесс. Алдымен талаптар анықталады, содан соң тұжырымдамалық модель жасалады, логикалық схема қалыптастырылады және ақырында физикалық орналастыру жүзеге асырылады.

ER-диаграмма: Мән мен байланысты модельдеу

Негізгі субъектілер

ER-диаграмма (Entity-Relationship) — деректер моделін визуалды түрде бейнелеудің стандарты. Жүйедегі негізгі субъектілер: **Қызметкер** (аты, лауазымы, жалақысы), **Бөлім** (атауы, орналасуы, бюджеті), **Жоба** (атауы, мерзімі, мәртебесі). Әр субъект тікбұрышпен белгіленеді.

Қатынастар мен шектеулер

Субъектілер арасындағы байланыстар ромб белгісімен көрсетіледі. **Кардиналдылық** байланыстың санын анықтайды: бір-ден-бірге (1:1), бір-ден-көпке (1:N), көп-тен-көпке (M:N). Шектеулер деректердің бүтіндігін (integrity) қамтамасыз ету үшін қажет.

Деректер моделінің түрлері

Реляциялық модель

Артықшылықтары: қарапайым сұраныстар тілі (SQL), деректердің жоғары тұтастығы, транзакциялық механизмдер (ACID). Банктер, мемлекеттік жүйелер, денсаулық сақтау — ең кең тараған модель.

Иерархиялық модель

Ағаш тәрізді құрылым. Жылдам оқу мүмкіндігі, бірақ күрделі байланыстарды бейнелеу қиын. XML деректер форматы иерархиялық модельдің заманауи түрі болып саналады.

Постреляциялық модель

NoSQL деректер базалары (MongoDB, Cassandra) — реляциялық моделдің шектеулерін жеңу үшін жасалған. Үлкен деректер және икемді схемалар үшін оңтайлы. Горизонталды масштабталу мүмкіндігі бар.



Физикалық жобалау

Аппараттық платформа

Физикалық жобалау кезеңінде серверлер, сақтау жүйелері және желі инфрақұрылымы таңдалады. Дискілік жүйелердің типі (HDD, SSD, NVMe), жады мөлшері және процессор архитектурасы деректер базасының өнімділігін тікелей анықтайды.

Өнімділікті оңтайландыру

Индекстер, партициялау және кэштеу механизмдері сұраныс жылдамдығын күрт арттырады. Дұрыс таңдалған физикалық параметрлер жүйенің тиімділігін бірнеше есеге жақсартуы мүмкін. Өнімділік тесттері (benchmarking) міндетті түрде жүргізіледі.

Ресурстарды басқару

Процессор мен жадыны тиімді пайдалану

Процессор (CPU) ресурстарын оңтайлы бөлу үшін параллель сұраныстарды өңдеу, жоспарлаушы (query planner) параметрлерін баптау қажет. Оперативті жады (RAM) деректердің кэшін сақтайды, бұл дискіге сілтемелерді азайтып, жылдамдықты арттырады.

Дискілік жүйелердің рөлі

Деректердің тұрақты сақтауы дискілік жүйелер арқылы жүзеге асырылады. RAID массивтері деректердің қауіпсіздігін қамтамасыз ету үшін деректерді бірнеше дискіде параллель сақтайды. SSD технологиясы оқу/жазу жылдамдығын механикалық дискілермен салыстырғанда 10-100 есе арттырады.

Деректерді өңдеу құралдары

SQL тілі

Structured Query Language (SQL) — реляциялық деректер базасымен жұмыс істеудің стандартты тілі. **DDL** (Data Definition Language) — кестелер мен индекстер жасайды. **DML** (Data Manipulation Language) — деректерді қосу, жою, өзгерту операциялары. **DCL** — қол жетімділік құқықтарын басқарады.

Сұраныстарды құру принциптері

Тиімді SQL сұранысы: нақты мақсатты айқындайды, қажетсіз JOIN операцияларынан аулақ болады, индекстерді дұрыс пайдаланады. SELECT, FROM, WHERE, GROUP BY, HAVING, ORDER BY — сұраныстың негізгі операторлары. Оңтайлы сұрау деректер базасының жүктемесін едәуір азайтады.



Сақталынатын процедуралар

Компиляцияланған кодтың тиімділігі

Сақталынатын процедуралар (Stored Procedures) — сервер жағында алдын ала компиляцияланып сақталатын SQL блоктары. Бір рет компиляцияланып, қайта-қайта орындала береді. Желілік трафикті азайтады, өйткені бір команда бойынша күрделі логика орындалады. Деректер базасы жағында бизнес-логиканы орталықтандыруға мүмкіндік береді.

PL/SQL және Transact-SQL

PL/SQL (Oracle) — SQL тілін процедуралық мүмкіндіктермен кеңейтеді: айнымалылар, шарттар, циклдер, ерекше жағдайларды өңдеу. **Transact-SQL** (Microsoft SQL Server) — Microsoft-тың SQL кеңейтімі. Екеуі де күрделі бизнес-логиканы деректер базасы деңгейінде тікелей орындауға мүмкіндік береді.

Триггерлер: Автоматтандырылған реакция

1

Триггер дегеніміз не?

Триггер — кестеде белгілі бір оқиға орын алғанда (INSERT, UPDATE, DELETE) автоматты түрде іске қосылатын код блогы. Ол деректер базасы жүйесімен тікелей байланысқан және пайдаланушының ешқандай іс-қимылынсыз орындалады.

2

Деректерді жаңарту және бақылау

Триггерлер арқылы: деректердің өзгерту тарихын автоматты жазу, байланысты кестелерді синхронды жаңарту, деректердің дұрыстығын тексеру жүзеге асырылады. Аудит журналдарын жүргізу — триггерлердің классикалық қолданысы.

3

Бизнес-логиканы сақтау

Бизнес-ережелерді триггерлер арқылы деректер базасы деңгейінде бекіту оларды қолданба кодынан тәуелсіз ете отырып, жүйенің тұтастығын кепілдендіреді. Мысалы, тауар қоры азайғанда автоматты тапсырыс беру.

Деректерді енгізу және редакциялау

Формальды енгізу әдістері

Деректерді енгізу формалары (Input Forms) пайдаланушыға ыңғайлы интерфейс ұсынады. Пайдаланушы тек қажетті өрістерді толтырады, жүйе автоматты валидация жүргізеді. Бұл процесс қателіктер санын күрт азайтады.

Деректер типін анықтау

Әр өрістің дұрыс деректер типін таңдау өте маңызды: INTEGER, VARCHAR, DATE, BOOLEAN, DECIMAL. Дұрыс тип сақтау кеңістігін үнемдейді, іздеу жылдамдығын арттырады және деректер тұтастығын қамтамасыз етеді.

Топтық импорт

CSV, Excel, XML файлдарынан деректерді жаппай импорттау мүмкіндігі уақытты үнемдейді. ETL (Extract, Transform, Load) процестері сыртқы деректер көздерінен автоматтандырылған жүктеуді қамтамасыз етеді.



Пішімдеу және сапаны бақылау

Ұяшықтарды теңестіру және пішімдеу

Деректерді стандарт бойынша пішімдеу тұтастықты қамтамасыз ету үшін өте маңызды. Мысалы, телефон нөмірлері +7 (XXX) XXX-XX-XX форматында, күндер YYYY-MM-DD форматында сақталуы тиіс. Пішімдеу ережелері автоматты түрде қолданылса, адами қателіктер минимумға дейін азаяды.

Деректердің тұтастығын қамтамасыз ету

Тұтастық шектеулері (Integrity Constraints): **Бастапқы кілт** (Primary Key) — жазбаның бірегейлігін қамтамасыз ету. **Сыртқы кілт** (Foreign Key) — кестелер арасындағы байланыстың дұрыстығы. **NOT NULL** — маңызды өрістердің бос қалмауы. **UNIQUE** — қайталанбайтын мәндерді кепілдеу.



Қазіргі заманғы аналитика

Сапалық және сандық деректер айырмашылығы

Сандық деректер — өлшемдер, мөлшерлер, статистика (мысалы: сату көлемі, температура). Математикалық операциялар мен статистикалық талдауға жарамды. **Сапалық деректер** — сипаттамалар, категориялар, пікірлер. Контент-анализ, кластерлеу арқылы зерттеледі. Екі тип бірге жиі пайдаланылады.

Жоспарла-Істе-Тексер-Әрекет ет (PDCA) циклі

Деминг циклі ретінде белгілі PDCA — үздіксіз жетілдіруді деректерге негіздейтін басқару моделі. **Жоспарла**: мәселелерді анықта. **Істе**: шешімді сынап көр. **Тексер**: деректерді талда. **Әрекет ет**: жетілдіруді жүйелендір. Бұл цикл деректер аналитикасын бизнес-жетілдіруге тікелей байланыстырады.

Аналитикалық сипаттама: Ғылыми негіздер

Дереккөздерді жүйелеу ережелері

Зерттеу жұмысында деректер мен дереккөздерді дұрыс сипаттау — ғылыми нәтижелердің сенімділігінің негізі. Библиографиялық сипаттама стандарттары (ГОСТ, APA, MLA, Chicago) дереккөздерді бірыңғай форматта ұсынуды талап етеді. Дұрыс сілтеме авторлық құқықты қорғап, зерттеудің қайталануына мүмкіндік береді.

Кітап құрамдас бөліктерін сипаттау

Кітаптың жеке тараулары, мақалалары немесе бөлімдерін сипаттауда: автордың аты-жөні, тараудың атауы, кітаптың атауы, редактор, баспа, жыл, бет аралығы міндетті элементтер болып табылады. Аналитикалық сипаттаманың логикалық тізбегі мен элементтерінің дұрыс реті ғылыми мәтіннің академиялық деңгейін анықтайды.

Электронды ресурстарды жинақтау

1 Сандық тасымалдағыштардағы деректерді стандарттау

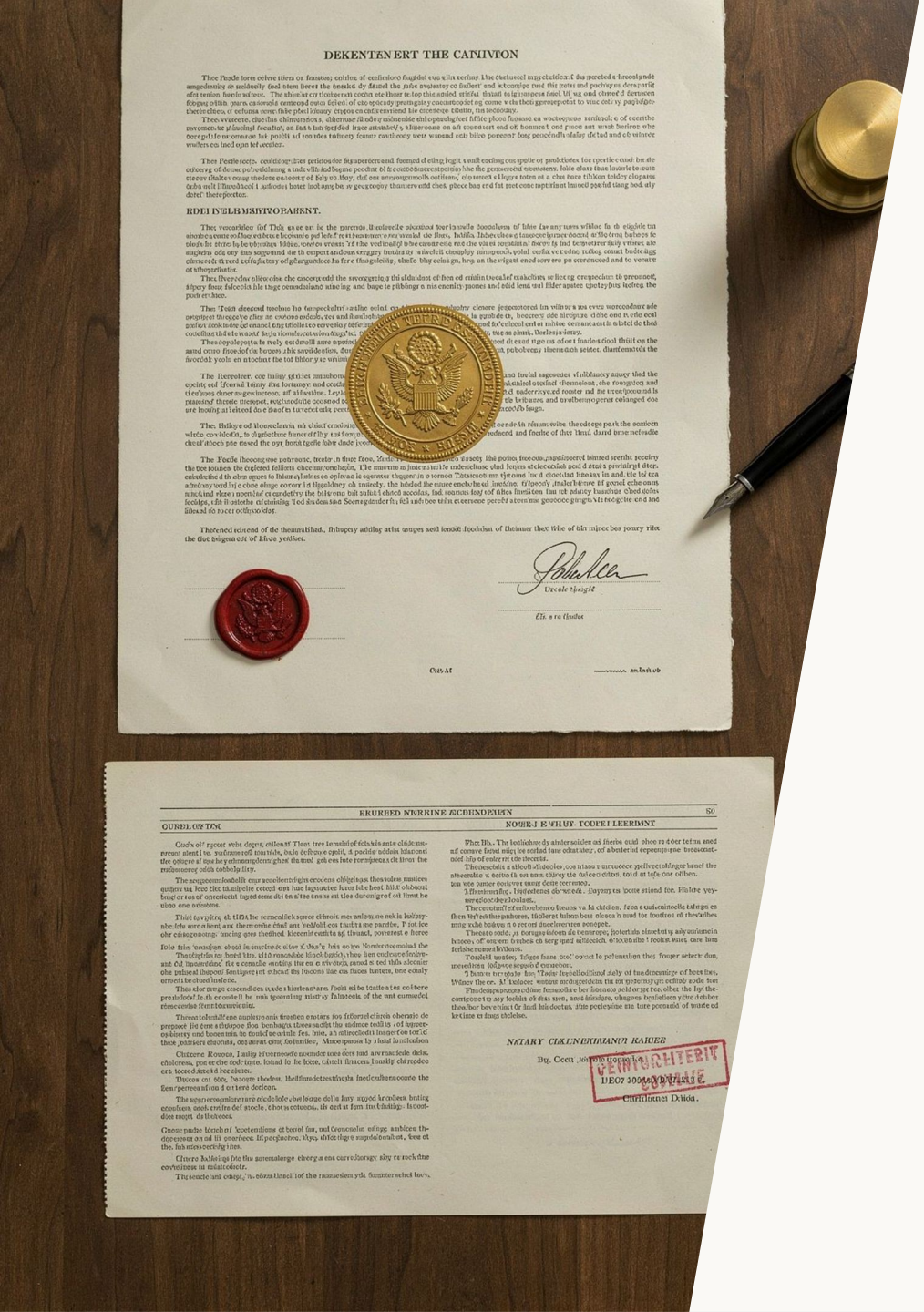
CD-ROM, DVD, USB тасымалдағыштарындағы ресурстарды сипаттауда физикалық формат пен қол жеткізу тәсілі міндетті түрде көрсетіледі. ISO 690-2 халықаралық стандарты электронды дереккөздерді сипаттаудың жалпыға ортақ ережелерін белгілейді.

2 Метадеректерді толтыру және сақтау

Метадеректер (автор, атауы, жыл, URL, қол жеткізу күні) ресурстың дәл идентификациясын қамтамасыз етеді. Онлайн ресурстар үшін соңғы қол жеткізу күнін көрсету міндетті, өйткені сілтемелер өзгеруі немесе жойылуы мүмкін.

3 Желілік ресурстарды библиографиялау

Интернет ресурстарын (сайт, онлайн-мақала, электронды кітап, деректер базасы) сипаттауда DOI нөмірі, тұрақты URL немесе архив сілтемесі (web.archive.org) ресурстың тұрақты қол жетімділігін кепілдендіреді.



Деректерді басқарудың нормативтік негіздері

Стандарттар мен регламенттер

Құжат айналымы халықаралық (ISO 15489), ұлттық және салалық стандарттармен реттеледі. Бұл стандарттар ресімдеу тәртібін, сақтау мерзімін және деректерге қол жеткізу ережелерін анықтайды.

Мемлекеттік деңгейдегі рәсімдеу

Қазақстанда мемлекеттік органдардың деректерді рәсімдеуі "Электронды үкімет" инфрақұрылымы арқылы жүзеге асырылады. ҚР Заңдары, Үкімет қаулылары деректерді жинау мен сақтаудың міндетті талаптарын бекітеді.

Құқықтық күш

Электронды құжаттардың заңды күші ЭЦҚ (электронды цифрлық қолтаңба) арқылы қамтамасыз етіледі. Тиісті рәсімделген электронды құжат қағаз нұсқасымен бірдей заңды статусқа ие болады.

Құжаттар эволюциясы: Куәліктен дереккөзге

1

Латынша Documentum

Бастапқы мағынасы — "үйрету, нұсқау". Ортағасырлық Еуропада ресми жазбаша куәлік ретінде қолданылды.

2

XVII–XIX ғасырлар

Мемлекеттік басқарудың өсуімен құжаттың мәні өзгерді. Бюрократиялық жүйелер нығая бастады, іс-қағаздар рәсімделді.

3

XX ғасыр

Типографика, ксерокс және компьютерлер құжаттарды жаппай өндіруге мүмкіндік берді. Кеңсе жұмысы стандартталды.

4

XXI ғасыр

Электронды деректер базалары қағаз құжаттарын алмастырды. Цифрлық трансформация — деректердің жаңа өмірі.

Іс-қағаздарды жүргізудің қазақстандық моделі

Бөкей хандығының кеңселік жүйесі

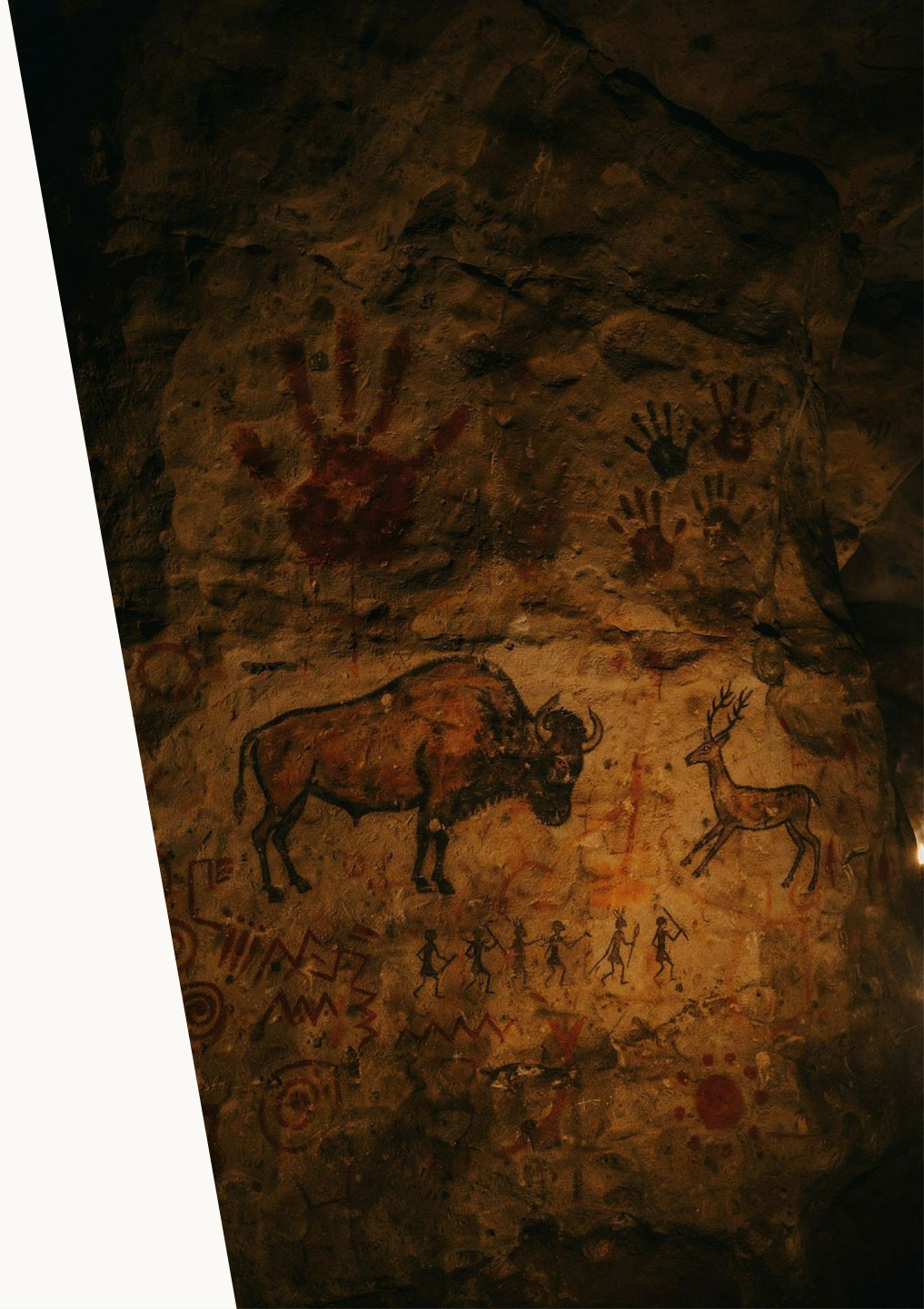
XIX ғасырдың басында Бөкей хандығы (1801–1845) Еділ мен Жайықтың арасында өз кеңселік жүйесін қалыптастырды. Хан кеңсесінде іс-қағаздар жүргізілді, жарлықтар мен жарғылар ресімделді. Бұл — қазақ мемлекеттік басқарудағы деректерді ұйымдастырудың тарихи алғашқы ірі тәжірибесі.

Тілдер мен жазу жүйелері

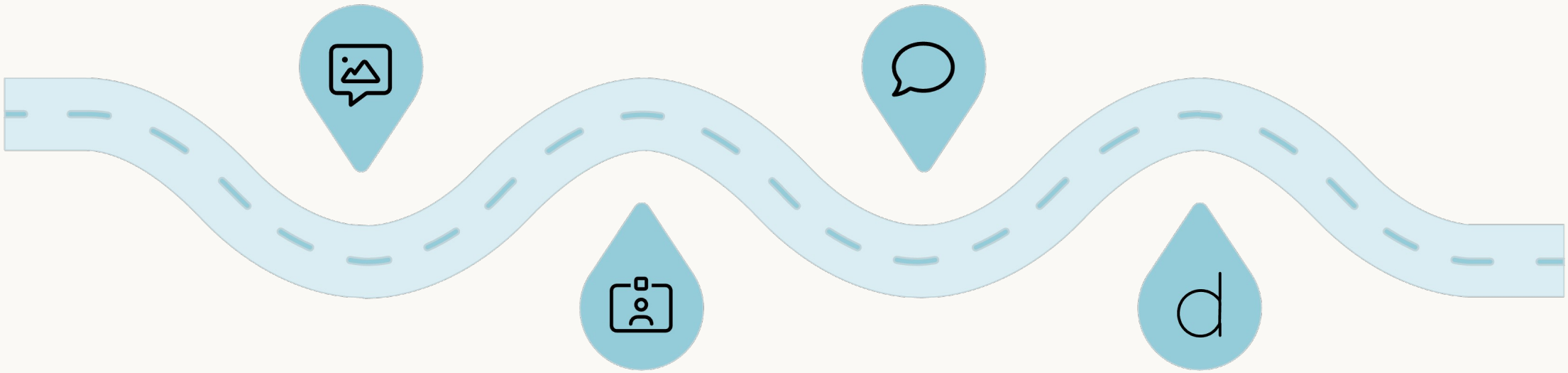
Бөкей хандығы кезінде іс-қағаздар **түркі-шағатай** (жазба әдеби тіл) және **орыс тілдерінде** жүргізілді. XIX ғасырдың аяғынан бастап қазақ тілінде ресми іс-қағаздардың пайда болуы ұлттық деректану мен іс жүргізудің тарихи кезеңін бастады. Бұл тілдік әртүрлілік деректерді интерпретациялауда ерекше дағды талап етеді.

Деректердің ертедегі тарихы

Адамзат өз тарихының басынан ақпаратты сыртқы ортаға шығаруға ұмтылды. Тасқа қашалған суреттер, саздан жасалған таблеткалар, папирусқа жазылған мәтіндер — бәрі де деректерді сақтаудың алғашқы тәсілдері. Пиктография (сурет жазуы) — ақпаратты алғашқы жүйелеу тәжірибесі: нақты объектілер суреттер арқылы бейнеленіп, оқиғалар тіркелді. Бұл адамзаттың танымдық эволюциясының ерекше белесі болды.



Жазу жүйелерінің эволюциясы



Пиктография

Суреттер заттарды білдірді, көне дәуір

Идеография

Белгілер сөздер мен мағыналарды білдірді

Буын жазуы

Символдар дыбыстарды және буынды білдірді

Алфавиттік

Әріптер фонемаларды білдіріп, тиімді сақтау

Жазудың дамуы — адамзаттың ақпаратты сақтау тиімділігін арттыруға деген ұмтылысының айғағы. Пиктографиядан идеографияға, одан буын жазуына және ақырында дыбыстық (алфавиттік) жазуға өту мыңдаған жылдар бойы жүрді. Алфавит — аз ғана таңбалармен барлық тілдік ақпаратты кодтайтын ең тиімді жүйе.



Ғылыми деректерді жинау

Экспедициялық материалдарды өңдеу

Далалық экспедициялардан жиналған деректер (фольклорлық жазбалар, этнографиялық материалдар, диалектілік сөздіктер) ғылыми өңдеуден өтіп, жіктеледі, транскрипцияланады және архивтеледі. Стандартты өңдеу методологиясы материалдардың болашақта салыстырмалы зерттелуін мүмкін ете отырып, ғылыми мұраны сақтайды.

Фольклорлық мұраны архивтеу

Ауызша дәстүр мұраларын жинау, жазып алу және архивтеу — ерекше ғылыми дағдыны талап ететін жұмыс. ЮНЕСКО-ның Материалдық емес мәдени мұра конвенциясы осы деректерді халықаралық стандарт бойынша сақтауды міндеттейді.

Қолжазба қорларын цифрландыру

Сарғайған қолжазбаларды жоғары ажыратымдылықты сканерлеу, OCR (оптикалық таңбаларды тану) технологиясын қолдану арқылы цифрлық форматқа аудару — мәдени мұраның жоғалып кетуін болдырмайтын заманауи тәсіл.

Деректерді жоғалту қаупі: Идеологиялық кедергілер

1947 жылғы қаулының салдары

Кеңес дәуірінде идеологиялық цензура ғылыми деректердің жоғалуына тікелей алып келді. 1947 жылғы қаулы этнографиялық, фольклорлық және тарихи зерттеулердің бірқатарын "буржуазиялық" деп жариялап тоқтатты. Ондаған жылдар бойы жиналған деу деректер жарияланбай архивтерде жатты немесе мүлде жойылды.

Деректерді сақтаудың маңызы

Осы тарихи тәжірибе деректерді сақтау саяси және идеологиялық қысымнан тәуелсіз болуы тиіс екенін дәлелдейді. Заманауи деректерді басқару стандарттары ғылыми мұраны қорғауда маңызды рөл атқарады. Архивтерді сақтау — ұлттық зерде мен ғылыми мұраны болашаққа жеткізудің кепілі.



Өндірістік деректер және технология

Технологиялық сызбалар

Өндірістік процестердегі деректер технологиялық нұсқаулық карталарда сақталады. Температура режимдері, уақыт параметрлері, шикізат мөлшерлері — бәрі дәл жазылады. Бұл деректер өнімнің сапасын қайта қалпына келтіруге мүмкіндік береді.

Органолептикалық деректер

Тамақ өнімдерінің сапасын бағалауда органолептикалық тәсілдер (дәм, иіс, түс, консистенция) қолданылады. Бұл сапалық деректер стандартталған шкалалар арқылы сандық мәндерге айналдырылып, деректер базасына енгізіледі.

Сапаны бақылау жүйесі

ISO 9001 стандарты бойынша өндірістік деректерді үздіксіз жинау, талдау және жетілдіруге пайдалану — заманауи сапа менеджментінің негізі. Деректер негізінде ақаулардың себебін анықтап, алдын алу шаралары жасалады.

Деректерді цифрландыру және архивтеу

Дәстүрлі құжаттардан цифрлық форматқа

Мұрағаттар мен кітапханаларда сақталған миллиондаған қағаз беттер цифрлық трансформация арқылы жаңа өмірге ие болуда. Жоғары ажыратымдылықты сканерлеу, OCR технологиясы, метадеректер жүйесі және бұлтты сақтау арқылы тарихи деректер іздестіруге ыңғайлы, жоғалтуға берік форматта сақталады. Бұл процесс мұрағат жұмысын түбегейлі өзгертті.

Ғылыми тәжірибелер мен нәтижелер

Ұлттық кітапхана, Орталық мемлекеттік архив, университет кітапханалары жаппай цифрландыру жобаларын жүргізуде. Europeana (Еуропа), HathiTrust (АҚШ), JSTOR (ғылым) — табысты халықаралық цифрлық архивтер мысалы. Ашық қол жетімді цифрлық архивтер ғылымды демократияландырады.



Мәдени мұраны цифрландыру: "Талап" қоғамының ізімен

> "Мәдени мұраны сақтау — бүгінгі ұрпақтың болашаққа берер ең маңызды деректер мұрасы"

1922 жылы құрылған "Талап" қоғамы қазақ мәдениетін зерттеу мен сақтауда ерекше рөл атқарды. Қоғам мүшелері жиналған фольклорлық, этнографиялық және тарихи материалдар бүгінде цифрлық деректер базаларына енгізілуде. Тарихи қолжазбаларды заманауи ДҚ-ға кіріктіру технологиясы XML, Dublin Core метадеректер стандарттарын, жоғары ажыратымдылықты цифрлық сканерлеуді қолданады. Бұл процесс тарихи ақпараттың жоғалмауын кепілдендіреді.

Деректерді басқарудың жаһандық стандарттары

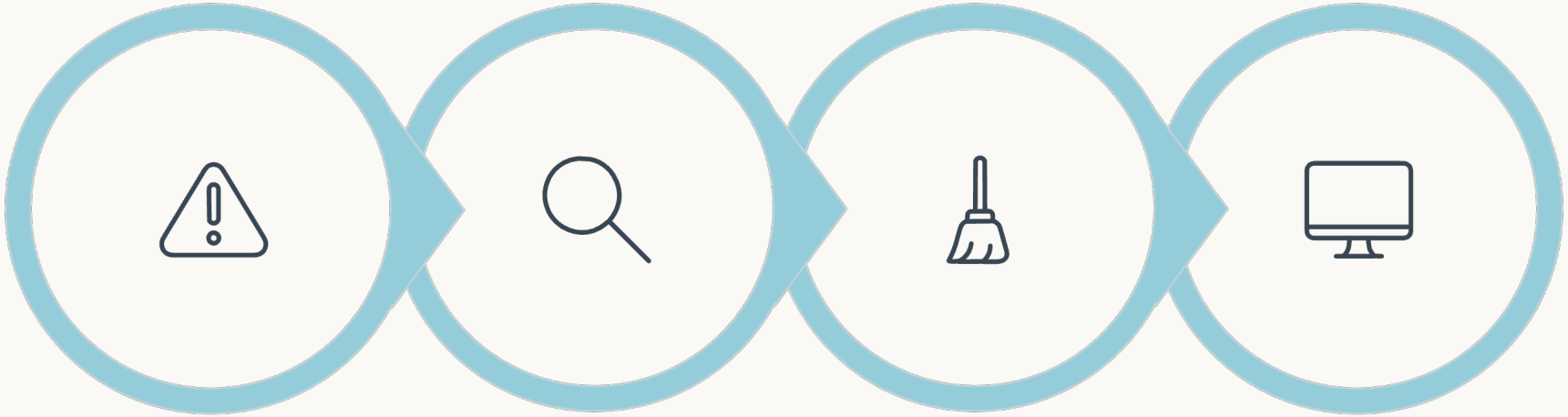
Метадеректер (Metadata)

Метадеректер — деректер туралы деректер. Файлдың атауы, жасалу күні, авторы, форматы, тақырыбы — бәрі метадеректер. Dublin Core, MARC 21, Schema.org — халықаралық метадеректер стандарттары. Дұрыс метадеректер жүйесі миллиондаған деректер арасынан қажетті ақпаратты секундтар ішінде табуға мүмкіндік береді.

Халықаралық алмасу форматтары

JSON (JavaScript Object Notation) — жеңіл, оқуға ыңғайлы формат. API арқылы деректер алмасуда кең қолданылады. **XML** (eXtensible Markup Language) — иерархиялық деректерді сипаттауға арналған күшті формат. Мемлекеттік жүйелер мен мұрағаттарда XML-ді кең пайдаланады. Екі формат та платформа-тәуелсіз.

Деректердің сапасын бақылау және тазалау



Қателерді анықтау

Түпкі себепті талдау

Тазалау және түрлендіру

Растау және бақылау

Шикі деректерді (raw data) тікелей талдауға болмайды. Деректерді тазалау (Data Cleaning) процесі жетіспейтін мәндерді толтыруды, дубликаттарды жоюды, форматтарды стандарттауды және аномалияларды анықтауды қамтиды. Сапалы аналитика тек сапалы деректерге ғана негізделе алады.



Үлкен деректер (Big Data) архитектурасы

Деректер көлемінің жарылысы

2025 жылға қарай жаһандық деректер көлемі 175 зетабайтқа жетеді деп болжануда. Интернет, мобильді құрылғылар, IoT сенсорлары — деректердің экспоненциалды өсуінің негізгі қозғаушы күштері. Дәстүрлі деректер базалары мұндай масштабты өңдей алмайды.

Data Warehouse (Деректер қоймасы)

Құрылымдалған деректерді кіріктіруге арналған орталықтандырылған репозиторий. Алдын ала өңделген, жіктелген деректер стратегиялық есептерге негіз болады. Oracle DW, Amazon Redshift, Google BigQuery — танымал платформалар.

Data Lakes (Деректер көлдері)

Барлық типтегі шикі деректерді (құрылымдалған, жартылай құрылымдалған, құрылымдалмаған) алдын ала өңдеусіз сақтайтын репозиторий. Икемді, масштабталатын, бірақ жақсы басқаруды талап етеді. Apache Hadoop, Amazon S3 — негізгі платформалар.

Бұлтты технологиялар және деректерді сақтау

Бұлтты инфрақұрылымға көшу

Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform — дүниежүзілік бұлтты есептеу алыптары. Орталықтандырылған серверлерден бұлтқа көшу шығындарды азайтады, масштабталуды жеңілдетеді және географиялық икемділікті қамтамасыз етеді. Кез-келген жерден, кез-келген уақытта деректерге қол жеткізу мүмкіндігі бизнес процестерін түбегейлі өзгертті.

Артықшылықтар мен қауіпсіздік

Бұлтты сақтаудың артықшылықтары: капиталдық шығынсыз операциялық модель, автоматтандырылған резервтік көшіру, жоғары қол жетімділік (99.99% SLA). Қауіпсіздік мәселелері: деректердің егемендігі (data sovereignty), шифрлау талаптары, бірнеше провайдерге тәуелділік (vendor lock-in) қаупі. Гибридті бұлт моделі оңтайлы шешім ретінде кеңінен қолданылады.



Деректер қауіпсіздігі және құпиялылық

Деректерді шифрлау әдістері

Заманауи шифрлау алгоритмдері деректерді рұқсатсыз қол жеткізуден қорғайды. **AES-256** — симметриялық шифрлаудың алтын стандарты. **RSA** — ашық кілтті шифрлау. **TLS/SSL** — желі арқылы деректер тасымалдаудың қауіпсіздігін қамтамасыз ету протоколы. Шифрлаусыз жіберілген кез-келген ақпарат ұсталып оқылуы мүмкін.

GDPR нормативтік талаптары

Еуропалық Одақтың GDPR (General Data Protection Regulation) регламенті дербес деректерді өңдеудің жаһандық стандартын белгіледі. Негізгі принциптер: мақсатты шектеу, деректерді минимизациялау, дұрыстық, сақтауды шектеу, тұтастық пен құпиялылық. GDPR бұзылған жағдайда айыппұл жылдық айналымның 4%-на дейін жетуі мүмкін.

Киберқауіпсіздік: Деректерді жоғалту мен ұрлаудың алдын алу

Рұқсатсыз қол жеткізуді болдырмау

Көп факторлы аутентификация (MFA), рөлдік қол жеткізу бақылауы (RBAC), нөлдік сенім архитектурасы (Zero Trust) — заманауи қорғаныс стратегиялары. Ішкі қауіптер (insider threats) сыртқы шабуылдарға қарағанда жиі кездесетіні статистикамен дәлелденген. Мінез-құлықтық аналитика аномалияларды ерте анықтауға мүмкіндік береді.

Резервтік көшіру жүйелері

Backup жүйесі деректердің апаттан кейінгі қалпына келуін (Disaster Recovery) қамтамасыз ету үшін өте маңызды. 3-2-1 ережесі: 3 көшірме, 2 әртүрлі медиа, 1 сыртқы орынды сақтау. RPO (Recovery Point Objective) және RTO (Recovery Time Objective) SLA шеңберінде анықталуы тиіс.

Деректерді визуализациялау өнері

Визуализация құралдары

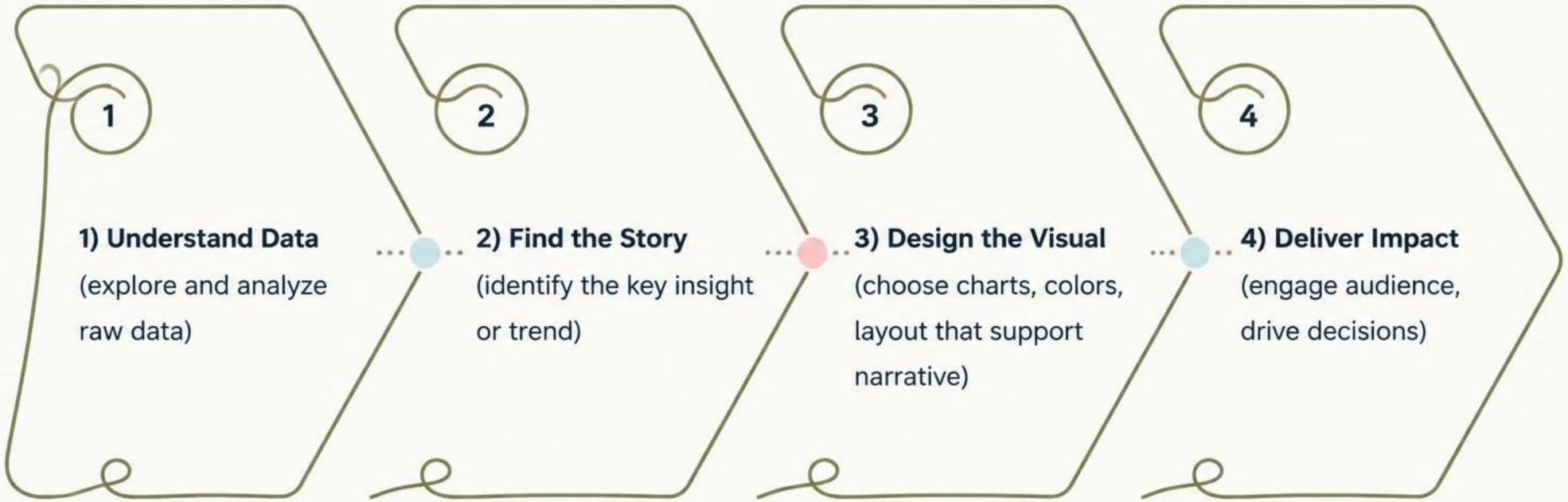
Power BI (Microsoft) — бизнес-аналитиктер үшін кеңінен қолданылатын BI платформасы. **Tableau** — деректерді визуализациялаудың де-факто стандарты, drag-and-drop интерфейсі. **Grafana** — нақты уақыттағы метрикаларды бақылауға арналған. **D3.js** — веб-визуализацияның жоғары икемді кітапханасы. Дұрыс құрал деректерді шешім қабылдаушыларға тиімді жеткізеді.

Ақпараттық панельдер (Dashboard)

Dashboard — бизнестің маңызды метрикаларын бір экранда жинайтын визуалды интерфейс. Тиімді dashboard: нақты мақсатты аудитория үшін жасалады, 5-7 негізгі KPI-ды ғана көрсетеді, нақты уақытта жаңарады және бұрғылауға (drill-down) мүмкіндік береді. Шешім қабылдаушылар мыңдаған жолдық есептерді оқымай-ақ жағдайды бір қарағанда түсінеді.



Инфографика: Деректерді баяндау тәсілі



Деректерді баяндау (Data Storytelling) — деректерді тек сандар ретінде емес, адамды сендіретін оқиға ретінде ұсыну өнері. Мидың 90%-ға жуығы ақпаратты визуалды түрде өңдейді. Дұрыс инфографика немесе визуализация кешенді деректерді секундтар ішінде түсінуге мүмкіндік береді. Психологиялық тұрғыдан оқырман санасына бейнелер сандарға қарағанда тереңірек жетеді.

Деректерді автоматтандырылған жинау (IoT)

Заттар интернеті (IoT)

IoT (Internet of Things) — физикалық объектілерді интернетке қосып, олардан нақты уақытта деректер жинайтын технология. Бүгінде 15 миллиардтан астам IoT құрылғысы жұмыс істейді. Температура, қысым, орналасу, жылдамдық — сенсорлар мыңдаған параметрлерді секундына жинайды.

Смарт-құрылғылар

Өнеркәсіпте: станоктардың тозуын болжайтын предиктивті техническое обслуживание жүйелері. Тұрмыста: смарт-термостаттар, фитнес-браслеттер, смарт-есептегіштер. Ауыл шаруашылығында: топырақ ылғалдылығы мен ауа-райын бақылайтын агро-сенсорлар. Барлығы бірге — деректердің үздіксіз ағынын жасайды.

Автоматтандырылған өңдеу

Edge Computing технологиясы деректерді сенсор жанында тікелей өңдейді, орталық серверге тек маңызды ақпаратты жібереді. Бұл желілік жүктемені азайтып, реакция уақытын минимизациялайды. MQTT, AMQP протоколдары IoT деректер ағынының стандартты хаттамалары.



Деректердің жасанды интеллектпен (AI) синтезі

AI көмегімен деректерді тазарту

Машиналық оқыту алгоритмдері жетіспейтін мәндерді болжап толтырады, форматтық қателерді автоматты анықтайды және аномалиялық деректерді жалаушалайды. Адам сағаттарымен жасайтын жұмысты AI секундтарда орындайды. AutoML инструменттері деректерді тазалауды толық автоматтандырады.

Болжамды аналитика

Тарихи деректер негізінде болашақтағы оқиғаларды болжау — болжамды аналитиканың мақсаты. Клиенттің кетуін (churn prediction), сұраныс болжамын (demand forecasting), жабдықтың істен шығуын алдын алуды (predictive maintenance) машиналық оқыту модельдері жоғары дәлдікпен болжайды.

Жасанды деректер (Synthetic Data)

Нақты деректерден статистикалық қасиеттері сақталған, бірақ жеке ақпарат жоқ жасанды деректер жасалады. Медицинада пациент деректерін ашпай AI моделін үйретуге мүмкіндік береді. GAN (Generative Adversarial Networks) жасанды деректер жасаудың жетекші технологиясы.



Он сегізінші бөлім: Деректер этикасы және жауапкершілік

Деректерді жинаудағы моральдық шектеулер

Деректерді жинау жай техникалық мәселе емес — ол этикалық жауапкершілікті де қамтиды. Адамның білімімен және келісімімен жинау (Informed Consent), мақсаттан тыс пайдаланбау, кемсітушілікке алып келетін алгоритмдерден аулақ болу — деректер этикасының негізгі принциптері. Алгоритмдік бейтараптылық (algorithmic fairness) зерттеушілердің маңызды зерттеу саласы.

Цифрлық із және мөлдірлік

Интернеттегі кез-келген іс-әрекет цифрлық із қалдырады: браузер тарихы, орналасу деректері, сатып алу тарихы, әлеуметтік медиа белсенділігі. Бұл деректер кімге тиесілі? Кім пайдалана алады? Мөлдірлік (transparency) принципі пайдаланушылардың өз деректерін бақылау құқығын қорғайды. "Ұмытылу құқығы" (Right to be Forgotten) GDPR-дың маңызды нормасы.

Мәліметтерді демократияландыру

Деректерге қолжетімділік

Деректерді тек IT-мамандары ғана емес, ұйымның барлық деңгейіндегі қызметкерлері пайдаланатын мәдениет қалыптастыру — деректі демократияландырудың мақсаты. Data Mesh архитектурасы деректерді орталықтандырылған командадан доменге беру моделін ұсынады.

Өзіне-өзі қызмет аналитикасы

Self-service analytics платформалары (Power BI, Tableau, Looker) бизнес-пайдаланушыларға IT бөлімінің көмегінсіз есептер жасауға мүмкіндік береді. Бұл шешім қабылдау жылдамдығын күрт арттырады. Деректерге негізделген мәдениет (data-driven culture) — компетентивті артықшылықтың маңызды факторы.

Деректер сауаттылығы

Деректер сауаттылығы (Data Literacy) — деректерді оқу, түсіну, жасау және пайдалану қабілеті. Gartner зерттеуінше, деректер сауаттылығы болашақта "орысша оқу" немесе "математика" сияқты базалық дағдыға айналады. Ұйымдар деректер сауаттылығын дамыту үшін арнайы оқыту бағдарламалары жүргізуде.

Болашаққа көзқарас және қорытынды

Деректер — жаңа дәуірдің ресурсы

XXI ғасырда деректер мұнай мен алтынға теңелді. Ұйымдардың бәсекелестік артықшылығы олардың деректерді жинау, өңдеу және пайдалану қабілетімен тікелей байланысты. Деректер — тек техникалық актив емес, ол стратегиялық мүмкіндік.

Технологиялық эволюция

Бұлттық есептеулерден — кванттық компьютерлерге дейінгі жол тек бірнеше онжылдықты алады. Кванттық алгоритмдер шифрлауды, оңтайландыруды және AI моделдерін түбегейлі өзгертеді. Деректерді өңдеу мүмкіндіктері экспоненциалды түрде өсуде.

Үндеу: Деректен даналыққа

Деректер → Ақпарат → Білім → Даналық — бұл DIKW пирамидасының жолы. Деректерді жинаудан бастап, оларды мағыналы шешімдерге айналдырудың бүкіл жолын меңгеру — заманауи маманның басты міндеті. **Деректер жинай аласыз — даналықпен пайдаланыңыз!**

