VI. Big Data a közigazgatásban

Szádeczky Tamás

DOI: 10.36250/00732.06

A fejezet célkitűzése

A fejezet célja a korábban már megismert, a relációs adatmodellen és az SQL nyelven alapuló hagyományos adatkezelésen túlmutató, nagyon nagy mennyiségű adat kezelési módjának megismerése. Ennek keretében foglalkozunk az adatok forrásaival, a lehetséges adatmodellekkel és feldolgozási technológiákkal, a Big Data és a Dolgok Internete alapvető kérdéseivel.

A fejezet feltételezi az adatbázis-kezelés elméletének ismeretét.

Az itt leírtak elsajátításával a hallgató kritikailag lesz képes értelmezni a fenti fogalmakat, és betekintést nyer a közigazgatási adatkezelés problémáiba az elkövetkező évtizedben.

1. Nagy mennyiségű adat kezelése és feldolgozása

Egy ország legnagyobb adatkezelője maga a kormányzat. Az állami adatkezelés magával az állammal egyidős. Már az ókorban is végeztek népszámlálásokat annak felmérésére, hogy kiket lehet egy háborúban besorozni. A másik jellemző adatkezelés, hasonlóan már az ókortól, az adóügyi adatok gyűjtése, kezelése. Ahogyan az állam fejlődött, úgy feladatai is bővültek. Így a modern állam foglalkozik az állampolgárok szociális biztonságával, egészségüggyel, oktatással és sok más területtel. Az állam elsődleges feladatai mellett érdekelt a kutatás-fejlesztésben, a meteorológiában és a hírközlésben, csak hogy pár tevékenységet említsünk. Szélesedő tevékenységi köre és az egyre jobb technológia egyre több adat kezelését teszi lehetővé. Az adat kezelése mellett viszont rendkívül fontos a feldolgozás képessége. A szocialista állam például csak a belső elhárítási tevékenységi körében hihetetlen mennyiségű adatot gyűjtött az állampolgárok mindennapjairól, ezek az adatok máig papíron és mágnesszalagon állnak az Állambiztonsági Szolgálatok Történeti Levéltárának polcain. Ennek oka, hogy a 20. században az extenzív adatgyűjtéshez már minden adott volt, viszont az adatok hatékony kiértékelése és az abból levonható következtetések megállapításához még nem állt rendelkezésre a szükséges eszközrendszer.

Az a technológia és megközelítésmód, amely a hihetetlen mennyiségű adat költség- és teljesítményhatékony tárolását és kiértékelését teszi lehetővé, a 2000-es években vált elérhetővé. Ezeket a szervezési és műszaki megoldásokat nevezzük *Big Datának*. Ez tehát nem konkrét eszközt vagy alkalmazást, hanem egy szemléletmódot és az ahhoz kapcsolódó

Szádeczky Tamás (2020): Big Data a közigazgatásban. In Sasvári Péter szerk.: *Informatikai rendszerek a közszolgálatban I.* Budapest, Dialóg Campus. 113–126. DOI: https://doi.org/10.36250/00732.06

technikai megoldásokat jelenti. Fontos figyelembe venni, hogy az informatikában mindig feltűnnek (majd néha eltűnnek) különböző divatos kifejezések és technológiák, amelyekről csak 5–10 év távlatában derül ki, mennyire jól alkalmazhatók a mindennapi életben. Kritikával kell tehát kezelnünk a Big Data fogalmát is, hiszen manapság túlzottan sokszor használjuk, és ezért hajlamosak lehetünk túldimenzionálni jelentőségét. A technológiai óvatosság mellett látnunk kell, hogy van mögötte műszaki tartalom, tehát érdemben foglalkozhatunk vele, hogyan lehet alkalmazni a közigazgatási informatika területén.

Hasonlóan az utóbbi években felkapott téma a Dolgok Internete (Internet of Things – IoT), amely a különböző eszközökbe épített egyszerű számítógépeket, mikrovezérlőket vagy egyszerűen csak szenzorokat (érzékelőket) jelenti. Az IoT-eszközök amellett, hogy intelligenssé vagy okossá tehetnek hagyományos termékeket, adatokat gyűjtenek és továbbítanak. Ezek a kifejezések persze teljesen hamisak, hiszen itt nem a mesterséges intelligencia (artifical intelligence – AI) tényleges kutatási eredményeinek hasznosításáról beszélünk, csak arról, hogy valamilyen eszköz adatokat gyűjt a fizikai világból, és továbbítja az interneten, esetleg az eszközt vezérelni lehet az internetről, például a mobiltelefonunkkal. Látnunk kell, hogy ezek nem az elmúlt évek műszaki újításai, hiszen az automatikában, illetve telemetriában több évtizede alkalmazzuk ezeket a technológiákat. A nagy változást a gyártási költségek csökkenésével a tömegtermelésben való alkalmazhatóság jelenti. Ma már nemcsak a százmillió forintos ipari robotot lehet ezzel a technológiával szerelni, hanem a másfél millió forintos okoshűtőt vagy a negyvenötezer forintos okosvízforralót is. Azt, hogy ez fantasztikus újítás vagy csak digitális sznobizmus, döntse el az Olvasó, de hogy van rá kereslet, az látható.

Ami viszont adatfeldolgozási szempontból érdekesebb, az az IoT legegyszerűbb megvalósítása, a szenzor. A valamilyen egyszerűen programozható platformot alkalmazó (például Arduino, lásd *32. ábra*) mikrokontroller-alapú, tízezer forintos panellel már sokféle érzékelési, adatgyűjtési és vezérlési feladat megoldható.

És itt jutunk vissza a Big Data problémájához. Az olcsón elkészíthető, könnyen telepíthető szenzorok nagyon nagy mennyiségű adatot tudnak gyűjteni, lehetővé téve ezzel a kiber-fizikai rendszerek fejlesztését, az okosvárosokat és további, még nem ismert lehetőségeket. Az így kinyert adatok feldolgozásának problémája viszont megegyezik a fent ismertetettekkel.

A mennyiségek szemléltetése végett az *1. táblázatban* látható, milyen nagyságrendekről beszélünk, közelítő példákkal.

1. táblázat

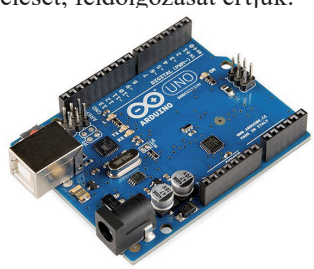
Adatmennyiségek nagyságrendje

Mennyiség	Nagyságrend	Közelítőleg minek felel meg
1 bájt = 8 bit	10^0	Egy karakter
1 kilobájt = 1000 bájt	10^3	Egy oldal szöveg
1 megabájt = 1 000 000 bájt	10^6	Egy regény
1 gigabájt = 1 000 000 000 bájt	10^9	Egy mozifilm
1 terabájt = 1 000 000 000 000 bájt	10^{12}	Egy személyi számítógép adathordozója

1 petabájt = 1 000 000 000 000 000 bájt	10^{15}	Egy globális kiskereskedelmi cég napi vásárlási adatai
1 exabájt = 1 000 000 000 000 000 000 bájt	10^{18}	Az összes valaha létezett emberi nyelv összes szava
1 zettabájt = 1 000 000 000 000 000 000 000 bájt	10^{21}	Teljes éves internetforgalom; egy globális kiskereskedelmi cég összesített vásárlási adatai; egy kontinens meteorológiai adatai
1 yottabájt = 1 000 000 000 000 000 000 000 000 bájt	10^{24}	Hírszerző szervek videófelvételei

Forrás: a szerző szerkesztése

A váltószámok tekintetében egyébként a legújabb IEC-szabvány szerint 1 kilobájt = 1000 bájt (decimálisan), 1 kibibájt (kilo-bináris bájt) = 1024 bájt, ugyanis a merevlemezeknél a decimális számrendszert használják (aminek egyébként a számítástechnikában nincs különösebb értelme). A kilo és a további prefixumok viszont az SI szerint ezres váltószámúak. Az *1. táblázat* esetében az érthetőség kedvéért maradunk a decimális rendszernél. További, nem hivatalos (nem SI-) prefixumok, amelyekkel találkozhatnak még: 10^{27}: Xenotta; 10^{30}: Shilentno; 10^{33}: Domegemegrotte. Big Data alatt a fenti táblázatból a terabájtos és annál nagyobb mennyiségű adat kezelését, feldolgozását értjük.



1. ábra

Arduino Uno R3

Forrás: Wikimedia.org (A letöltés dátuma: 2018. 01. 16.), By SparkFun Electronics from Boulder, USA – Arduino Uno – R3, CC BY 2.0

2. Adattárház

Az *adattárház* olyan speciális adatbázis, amely az adatokat lekérdezési, elemzési műveletekre optimalizált szerkezetben tárolja (szemben a hagyományos adattárolási, tranzakciókövetési céllal), a kiszolgált vezetési szintek igényeinek megfelelően aggregált adatokat is tartalmaz (ami egy hagyományos adatbázisból teljesen hiányzik), és különböző forrásokból nem tranzakciónként, hanem adott periódusonként és az adatértékek történetiségének megőrzésével frissítődik (ezzel szemben a hagyományos adatbázis nem őrzi a történetiséget, csak az aktuális állapotot).

Egy adattárház egy adott célra készül, ami lehet például marketingcélú, úgymint az értékesítések elemzése. Emellett integrált, tehát több forrásból gyűjt adatokat, amelyeket azonos formára hoz az összehasonlíthatóság érdekében. Az adattárház nem felejt, ugyanis az ide bekerülő adatok többé nem módosulnak és hosszú időn át megmaradnak, hiszen célja a trendek vizsgálhatósága, a változások elemzése. Alkalmazásával olyan kérdéseket tudunk megválaszolni, mint például: sikerült-e teljesíteni a negyedéves tervet? Sikeres volt-e a diszkontakció? Mekkora volt egy adott cikk forgalma a dél-alföldi régióban az elmúlt két hónapban? Milyen forgalmi adatok várhatók a győri üzletekben február és március hónapban? Mely boltok esetében volt kiemelkedően magas vagy alacsony a forgalom az átlagos forgalomhoz képest? Milyen termékeket vásárolnak gyakran együtt az ügyfelek (árukapcsolás)?

Két architekturális megközelítése van: az Inmon- és a Kimball-modell, William H. (Bill) Inmon és Ralph Kimball számítástechnikus után. Az Inmon-paradigma szerint az adattárház az általános üzleti intelligenciarendszer (BIS) egyik eleme. A vállalatnak egy adattárháza van, amelybe a különböző adatpiacokból (data mart) kerülnek az adatok. Az adattárházban az adatokat a harmadik normálformában (3NF) tárolják (lásd a korábbi adatbáziskezelés-elméleti tanulmányokban). Célja a döntéstámogatás. A Kimball-paradigma szerint az adattárház a vállalati adatpiacok halmaza. Az adatok mindig a többdimenziós (multidimenziós) adatmodellben vannak.

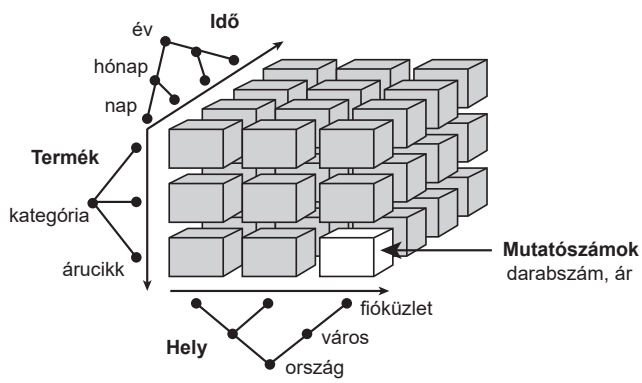
Az online analitikus feldolgozás (online analytical processing – OLAP), amely az adattárházakban történő lekérdezési műveletek végrehajtásának módja a többdimenziós adatmodellben, Edgar F. Codd nevéhez és 1993-ban megjelent tanulmányához kötődik.

Az OLAP-kritériumok – amelyek egyben meghatározzák az adattárházakkal szemben támasztott igényeinket – a következők:

1. multidimenzionális fogalmi nézet – a vállalatot többdimenziós jellegűnek tekintjük, például a nyereséget régiónként, termékenként, időtartamonként vagy forgatókönyv szerint (például tényleges, költségkeret vagy előrejelzés szerint) lehet megtekinteni. A többdimenziós adatmodellek lehetővé teszik a felhasználók számára az adatok egyszerűbb és intuitív manipulálását, beleértve a szeletelést (slicing) és a kockázást (dicing);
2. átláthatóság – ha az OLAP a felhasználói szokásos táblázatkezelő vagy grafikus csomag részét képezi, akkor ennek átláthatónak kell lennie a felhasználó számára. Az OLAP része lehet egy nyíltrendszer-architektúrának, amely beágyazható a felhasználó által kívánt helyen anélkül, hogy hátrányosan befolyásolná a gazdagép funkcionalitását. A felhasználó nem lehet kiszolgáltatva az OLAP-ba érkező forrásadatoknak, amelyek lehetnek homogének vagy heterogének;
3. elérhetőség – az OLAP-eszköznek képesnek kell lennie arra, hogy saját logikai struktúráját alkalmazza a heterogén adatforrások elérése érdekében, és elvégezzen minden olyan átalakítást, amely ahhoz szükséges, hogy egységes képet jelenítsen meg a felhasználó számára. Az eszköznek (és nem a felhasználónak) kell azzal foglalkoznia, hogy honnan származik a fizikai adat;
4. állandó lekérdezési teljesítmény – az OLAP-eszköz teljesítménye nem eshet jelentősen a dimenziók számának növekedése miatt;
5. kliens/szerver-architektúra – az OLAP-eszközök szerverösszetevőjének kellően intelligensnek kell lennie ahhoz, hogy a különböző ügyfelek minimális erőfeszítéssel

csatlakoztathatók legyenek. A kiszolgálónak képesnek kell lennie a különböző adatbázisok közötti adatok leképezésére és összevonására;
6. általános dimenzió – minden adatdimenziónak meg kell egyeznie struktúrájában és működési képességeiben;
7. dinamikus ritkamátrix-kezelés – az OLAP-szerver fizikai struktúrájának optimális, ritkamátrix- (nem teljesen kitöltött mátrix) kezelést kell biztosítania;
8. többfelhasználós üzemmód támogatása – az OLAP-eszközöknek konkurens (versenyző) elérést és frissítési hozzáférést, integritást és biztonságot kell biztosítaniuk;
9. korlátozás nélküli keresztdimenziós műveletek – a számítástechnikai eszközöknek lehetővé kell tenniük a számítás és az adatok manipulálását tetszőleges számú dimenzión keresztül, és nem korlátozhatják az adatcellák közötti kapcsolatot;
10. intuitív adatkezelés – konszolidációs útvonalban történő adatmódosulást, például a lefúrást (drilling down) vagy a kizoomolást (zooming out) az analitikai modell celláin történő közvetlen cselekvés útján kell végrehajtani, és nem kell egy menüt vagy egy nehezen elérhető funkciót használni a felhasználói felületen;
11. rugalmas jelentések – a jelentéstételi lehetőségeknek bármely olyan információt tartalmazniuk kell, amit a felhasználó meg szeretne tekinteni;
12. korlátlan dimenziószám és aggregációs szint – a támogatott adatdimenziók száma minden célból gyakorlatilag korlátlan. Minden általános dimenziónak lehetővé kell tennie egy adott konszolidációs útvonalon belül lényegében korlátlan számú, felhasználó által meghatározott aggregációs szintet.

Az OLAP estében többdimenziós (multidimenziós) adatmodellt alkalmazunk. Ez az az adatmodell, amely úgy tárolja az adatokat, hogy könnyen le lehessen kérdezni a különböző adatok közötti kapcsolatokat. A többdimenziós adatmodell az adatokat adatkockában tárolja, erre látható példa a *2. ábrán.* A kocka értelmezése igényel némi elvonatkoztatást, de belátható, hogy így végezhetjük el a legkönnyebben a lekérdezéseket. Az adatkocka jellemzői az éleihez rendelt dimenziók, amelyek az elemzés szempontjából lényeges nézőpontok, valamint az adatkocka celláiban tárolva a tények számértékei (például értékesítési adatok), amelyek az elemezni kívánt mennyiségek valamilyen mértékegységben meghatározva.



2. ábra

Eladási adatok háromdimenziós adatkockája

Forrás: Sidló 2004

3. Adatbányászat

Az adatbányászat a nagy adatkészleteken történő rendezés folyamata a minták azonosítására, kapcsolatok kialakítására és a problémák megoldására az adatok elemzése révén. Az adatbányászati eszközök lehetővé teszik a vállalkozások számára a jövőbeni trendek előrejelzését. Az adatbányászat során az összerendelési szabályokat a gyakori *ha-akkor* minták elemzésével hozza létre, majd a támogatási és a megbízhatósági kritériumokat használva megtalálja az adatok legfontosabb kapcsolatait. Támogatás az, hogy az elemek milyen gyakran szerepelnek az adatbázisban, míg a megbízhatóság az, hogy hány *ha-akkor* állítás bizonyult helyesnek. További adatbányászati paraméterek a szekvencia- vagy útvonalelemzés, az osztályozás, a fürtözés és az előrejelzés. A szekvencia- vagy útvonalelemzés-paraméterek olyan mintákat keresnek, ahol egy esemény egy másik későbbi eseményhez vezet. A sorrend a tételek csoportjainak rendezett listája, amely egy általános adatstruktúra, és számos adatbázisban megtalálható. Az osztályozási paraméter új mintákat keres, és az adatok szervezésének megváltozásához vezethet. Az osztályozási algoritmusok az adatbázison belüli egyéb tényezőkön alapuló változókat jelölik.

Az adatbányászat négy szakasza:

1. adatforrások feltérképezése – ezek az adatbázisoktól a hírekig terjedhetnek, jellemzően problémadefinícióra használjuk őket;
2. adatfeltárás/-gyűjtés – ebben a fázisban történik a mintavételezés és az adatok átalakítása;
3. modellezés – a felhasználók modelleket készítenek, tesztelik, majd értékelik azokat;
4. modellek alkalmazása – cselekvés a modell eredményei alapján.

A fürtözési paraméterek megkeresik, és vizuálisan dokumentálják a korábban ismeretlen tények csoportjait. A fürtözött csoportok egy objektumkészletet tartalmaznak, és aggregálják őket a hasonlóságuk alapján. A felhasználó különböző módokon tudja megvalósítani a fürtöt, amely különbséget tesz az egyes fürtözési modellek között. Az adatbányászat paraméterei olyan adatmintákat segíthetnek fedezni, amelyek a jövőre vonatkozó észszerű előrejelzésekhez vezethetnek, amelyet más néven prediktív elemzésnek is nevezünk.

Az adatbányászati technikákat számos kutatási területen használják, beleértve a matematikát, a kibernetikát, a genetikát és a marketinget. Noha az adatbányászati technikák a hatékonyság növelésére és a vevői viselkedés előrejelzésére szolgálnak, ha helyesen használják, akkor az üzleti vállalkozás versenyelőnyt szerezhet a konkurenciájával szemben a prediktív elemzés alkalmazásával. Az ügyfélkapcsolat-kezelésben alkalmazott adatbányászati megoldás a webes bányászat (web mining), amely a hagyományos adatbányászati módszereket vegyíti a webes technikákkal. A webes bányászat célja az ügyfelek viselkedésének megértése és az adott webhely hatékonyságának értékelése. További adatbányászati technikák közé tartozik a multitask tanulási szokásokon alapuló mintaosztályozás, az adatbányászati algoritmusok párhuzamos és skálázható végrehajtásának biztosítása, a nagy adatbázisok bányászata, a relációs és komplex adattípusok kezelése, valamint a gépi tanulás. A gépi tanulás az adatbányászatban lehetővé teszi öntanuló rendszer létrehozását.

Általánosságban elmondható, hogy az adatbányászat előnyei olyan rejtett minták és kapcsolatok feltárásának képességéből származnak, amelyek felhasználhatók arra, hogy előrejelzéseket készítsenek a vállalkozásokra nézve. A konkrét adatbányászati előnyök

a céltól és az iparágtól függően változnak. Az értékesítési és marketingosztályok például az ügyfelek adatait bányászva a konverziós arányt (elfogadott ajánlatok arányát) tudják növelni, vagy sikeresebb marketingkampányokat tudnak létrehozni. A múltbeli értékesítési mintákra és ügyfélkísérletekre vonatkozó adatbányászati információk felhasználhatók előrejelzési modellek készítésére a jövőbeni értékesítések, új termékek és szolgáltatások számára. A pénzügyi ágazatban működő vállalatok adatbányászati eszközöket használnak kockázatmodellek kialakítása és a csalások felderítése érdekében. A feldolgozóipar adatbányászati eszközöket használ a termékbiztonság javítása, a minőségi kérdések azonosítása, az ellátási lánc kezelése és a műveletek javítása érdekében.

4. Nem relációs (NoSQL-) adatbázisok

Az adatok feldolgozása előtt olyan módon szükséges az adattárolást elvégezni, amely lehetővé teszi rendkívüli mennyiségű adat hatékony kezelését. Rendkívüli mennyiség alatt több terabájt vagy e fölötti mennyiségű adatot értünk. Gondoljunk például a világ összes Facebook-profiljára, a Google keresési adatbázisára vagy Magyarország összes hírközlési szolgáltatójának forgalmára. Az alkalmazott adattárolási módnak lehetővé kell tennie például a hatékony keresést a földrajzilag szétosztott adatbázisokban is. Ezek a problémák lehetetlenné teszik például a hagyományos adatbázis-kezelő eljárások alkalmazását. Ennek megoldására következő generációs adatbázisokat kellett kifejleszteni.

A következő generációs (NoSQL-) adatbázisok jellemzően (de nem feltétlenül) nem relációs adatmodell alapján épülnek fel, akár földrajzilag is elosztottak, nyílt forráskódúak és horizontálisan skálázhatók. Fejlesztésük 2009-ben kezdődött, és azóta is folyamatos. Szemben a relációs adatbázisok ACID-követelményével, az NoSQL-nél ez nem elvárás. Az ACID – azaz az atomicitás (Atomicity), a konzisztencia (Consistency), az izoláció (Isolation) és a tartósság (Durability) – az adatbázis-kezelő rendszer tranzakciófeldolgozó képességének alapeleme, amely nélkül az adatbázis integritása nem garantálható. Ebben az esetben az adatbázis visszaállíthatatlanul sérül. Az NoSQL esetén ehelyett a BASE- (Basically Available, Soft state, Eventually consistent) követelményeket támasztjuk az adatbázis felé. Más megfogalmazásban ez az Eric Brewer által megfogalmazott CAP-tétel, miszerint az elosztott rendszerek tulajdonságai a konzisztencia (consistency), a rendelkezésre állás (availability) és a partíciótolerancia (partitions), amelyből egyszerre csak kettő tulajdonságot garantál a rendszer. Hasonlóképpen a minőségháromszöghöz (minőség–ár–gyorsaság). Ahogy a tábla is hirdeti: Cégünk *olcsón, jól* és *gyorsan* dolgozik. Ön ezek közül kettőt választhat!

Egy elosztott rendszer akkor konzisztens, ha egy adatlekérdezés eredménye bármilyen adatcsomópontban, bármilyen időpillanatban megegyezik, tehát minden esetben ugyanarra az eredményre jutunk. Egy elosztott rendszer rendelkezésre áll, ha egy kérésre minden működő csomópont válaszol. Egy elosztott rendszer partíciótoleráns, ha egy feltett kérdésre hálózati partíció kiesése esetén is helyes választ ad. Nem várható el ez a tulajdonság a teljes hálózat működésképtelensége esetén.

Az NoSQL-adatbázisokat jellemzően négy csoportra bontják, de a csoportokon belül a különböző termékeknek különböző képességeik, funkciói vannak. A különböző típusok között sok átfedés van, és általánosságban elmondható, hogy szemben a hagyományos adatbázisokkal mindegyiket vízszintesen osztják ki és vízszintesen skálázzák.

A *kulcs–érték- (key–values) tároló* a legegyszerűbb tárolási mód. Ebben az adatmodellben egy kulcshoz egy érték pár tartozik. Lehetséges művelet a beillesztés (insert), a lekérdezés (fetch), a frissítés (update) és a törlés (delete). A megoldáshatékonyság skálázható és hibatűrő, de csak akkor gyors, ha ismerjük a kulcsot, és a lekérdezés is csak ez alapján működik. A rekordok különböző csomópontok (node) között vannak szétosztva. Ilyen adatbáziskezelő a Berkeley DB, az Amazon Dynamo, a Hyperdesk, az SILT (Small Index Large Table), a Simple DB, a Redis és a Riak. Ilyen megoldást alkalmaz az Instagram és a Twitter is.

Az *oszlopalapú tároló (column store)* nagy, több gépen szétosztott adatmennyiséget tárol és dolgoz fel. Gyakolatilag egy hagyományos adatbázis elforgatásával készül. Például egy hagyományos adatbázis sorai a következők:

SQL-adatbázis	Oszlop1	Oszlop2	Oszlop3
Sor1	Magyary	Zoltán	1888
Sor2	Egyed	István	1886
Sor3	Concha	Győző	1846

Ezzel szemben a fenti rekordokat egy oszlopalapú adatbázisban a következőképp tároljuk:

Oszlopalapú adatbázis	Oszlop1	Oszlop2	Oszlop3
Sor1	Magyary	Egyed	Concha
Sor2	Zoltán	István	Győző
Sor3	1888	1886	1846

Ez sokkal gyorsabb lekérdezést és adatfeldolgozást tesz lehetővé, a valamilyen összefüggésben (esetünkben vezetéknevek, keresztnevek, születési évek felosztásban) tárolt adatok tekintetében. Ezeket jellemzően nagy teljesítményű, sebességkritikus adatelemzéseken használják. Ilyen adatbáziskezelő a Cassandra, a Hbase, a Voldemort, a Scalaris és a Memcached. Ilyen megoldást alkalmaz a Netflix és a Spotify is.

A harmadik típus a *dokumentumtároló (document store),* amely gyakorlatilag a kulcs–érték-tárolási móddal egyezik meg, de megengedi a beágyazott dokumentumok kezelését is. Ez a megoldás a félig strukturált adatok feldolgozását, úgymint tartalomkezelő, keresőrendszer vagy más lazán kapcsolódó adat tárolására alkalmas. A félig strukturált adatok kezelését jellemzően a JSON (JavaScript Object Notation) vagy az XML (Extensible Markup Language) nyelven valósítják meg. Gyors írás, jó lekérdezési idők jellemzik, de fő előnye a séma rugalmassága. Ilyen adatbáziskezelő a MongoDB, az OrientDB, a CouchDB, az Azure DocumentDB és a RethinkDB.

A negyedik – és egyben a legkomplexebb – típus a *gráfadatbázisok (graph stores).* Ezek a kapcsolatokra fektetik a hangsúlyt. A matematikai alapja a gráfok alkalmazása, amely a csomópontok és a rajtuk értelmezett összeköttetések (élek) halmaza. Az adatokat a csomópontokban tároljuk, és az azok közötti egy- vagy kétirányú kapcsolatokat az élekben (metaadatként) tároljuk. Ilyen adatbáziskezelő például az InfiniteGraph, az AllegroGraph és a Neo4j.

Az NoSQL-rendszerek előnyei, hogy olcsók, megvalósításuk könnyű. Az adatbázis replikált és particionált is lehet, könnyű megvalósítani az adatbázis szétosztását. Nincs szükség

sémára, skálázható az adatbázis. Gyors műveleteket tudunk végezni nagy adatbázisokon is. Az adatbázisra igaz a CAP-elv (lásd fentebb). Hátrányai viszont, hogy új rendszerek, és így előfordulnak hibák. Az adatokban, mivel általában több helyen is megtalálhatók, inkonzisztencia léphet fel, tehát két különböző helyen tárolt, de elvileg azonos adat eltér egymástól. Nincs szabványos séma, lekérdező és nyelv. Nehéz a komplikált struktúrákat megvalósítani. Nincs garantált támogatás, de hatalmas a választék.

5. Memóriaalapú adatkezelés

Az 1980-as években 64 kilobájt (pontosabban ma már kibibájt) operatív tár (tetszőleges hozzáférésű memória, Random Access Memory – RAM) szinte mindenre elég volt. Ma már a százharmincegyezerszerese számít normálisnak. Ez a folyamat nemcsak igény, hanem lehetőség is volt, tehát a memóriaárak is jelentősen csökkentek ebben az időszakban. Az olcsó RAM lehetővé teszi, hogy ne csak a hagyományos feladatára használjuk, hanem akár adatbázisokat is tölthetünk bele. A RAM-ban történő műveletek nagyságrendekkel gyorsabbak, mint a háttértárolón (HDD vagy SSD) végzett műveletek, cserébe viszont a mai árak mellett is rendkívül drágán, de hatékonyan tudjuk megvalósítani az adatkezelést. Memóriaalapú adatbázison (IMDB) tehát olyan adatbázisrendszert értünk, ahol a tárolt és kezelt adatok elsődleges példánya a RAM-ban található, szemben a lemezalapú adatbázisokkal (Disk-Resident DataBase – DRBD), ahol az a lemezes alapú háttértáron található. Az esetleges biztonsági másodpéldányok lehetnek lemezes vagy egyéb nem felejtő tárolón. Az elsődleges vagy munkapéldány a logikai adatelem azon példánya vagy példányainak összessége, amelyen a tranzakciós műveleteket végezzük. Célszerű ebben az esetben új adatszervezési és -kezelési elvek alapján működő adatbáziskezelő rendszereket használni. Így tipikusan célszerű alkalmazni a fent ismertetett NoSQL-megoldásokat, ugyanis a memóriaalapú adatbáziskezelők oszlopalapú formátumban kezelik az adatokat a hagyományos soralapú formátummal szemben. A hagyományos adatbázisok sokkal jobban tudják kezelni a tranzakciókat, viszont az oszlopalapú adatkezelés pont a lekérdezéseket gyorsítja meg, ami a Big Data esetében számunkra kiemelkedő jelentőségű. A legnagyobb memóriaalapú adatbáziskezelést lehetővé tevő adatbáziskezelő rendszer a SAP HANA, a Microsoft SQL Server In-Memory OLTP és az Oracle Database In-Memory, amelyek még a hagyományos adatbáziskelő funkciók kiegészítéseként tartalmazzák a memóriaalapú adatbáziskezelés lehetőségét. Így ezek a rendszerek támogatják a hagyományos soralapú (SQL nyelvű) feldolgozást is, így a nagy mennyiségű tranzakció kezelése ugyanúgy megoldható velük, mint az adatok szűrése, csoportosítása, összegzése, amelyet az oszlopformátumú adatkezelés tesz lényegesen gyorsabbá. A memóriaalapú adatkezeléshez hatalmas memóriával, gyors be- és kimenti interfészekkel rendelkező, nagy teljesítményű szerverekre van szükség.

A memória-adatbázisok alkalmazásának vannak előnyei és hátrányai is. Bizonyos feladatok elvégzéséhez egyszerűbb algoritmusokat alkalmazhatunk, mint a diszkalapú DRBD-rendszerekben, míg másokhoz bonyolultabb, újszerű megoldásokra van szükség.

A konkrét feladat és az arra vonatkozó követelmények ismeretében lehet eldönteni, érdemes-e IMDB-t használni. Ez a rendszerszervező mérnök feladata. A kizárólagos funkciójú adatbáziskezelők mellett rendelkezésre állnak hibrid rendszerek is, amelyek az adatbázis egy részét IMDB-, másik részét DRDB-elven kezelik, mindezt a programozói interfész

felől teljesen transzparens módon teszik. Ez jellemző a fent említett piacvezető termékek esetében is. E rendszerek bizonyos korlátok között képesek az egyes relációk hozzáférési tapasztalatai alapján változtatni az adatok helyét az IMDB- és a DRDB-alrendszer között.

6. A Big Data alkalmazása és veszélyei

Ha az adattárolás már sikerült, a következő lépés az adatok kezelése, feldolgozása és kiértékelése. Ez hasonlóképpen nehézségek elé állítja az adatkezelő szervezetet. Ennek megoldására a fentiekben leírt módszereket alkalmazó, új megközelítésű technológiát kell használnunk. A Big Data technológia célja, hogy rendkívül nagy mennyiségű adat esetén biztosítsa az adatok gyűjtését, kezelését, viszonylag gyors visszakereshetőségét, feldolgozását. Emellett az adatok nagy mennyisége, változatossága és komplexitása jellemző. Többdimenziós (multidimenziós) adatmodell szerint épülnek fel, ahogy az OLAP tekintetében már kifejtettük. Ebben a modellben az adatok alapegysége a tenzor, amely gyakorlatilag egy többdimenziós mátrix. Ennek kezelésére speciális, tenzoralapú számítási eljárásokat alkalmazunk.

A Big Data tekintetében a közigazgatás, a vállalatok, az intelligens hálózatok és az egyéni felhasználók által világszerte és napi szinten előállított óriási adatmennyiséget értjük. Az adatok forrása többféle lehet: a mobilinternet használatából, gépek közötti kommunikációból és szenzorok használatából (lásd az 1. fejezetben) is származhatnak. Ez a fokozatosan növekvő digitalizáció egyre növekvő, hatalmas adatmennyiségeket eredményez. Ha ezt a rengeteg adatot strukturálni lehet(ne), és kielemezni, akkor az azokból nyert rengeteg információ hasznosulni tud(na) közjavak vagy gazdasági haszon formájában. A Big Data elemzést végző szervezet változatos technológiák és eszközök segítségével igyekszik a különféle adatokat szisztematikusan feldolgozni és strukturálni, ráadásul mindezt másodpercek alatt, és amennyire lehetséges, automatizált módon. E tevékenység célja a kapcsolatok felismerése és a minták elemzése, ami ideális segítség lehet olyan modell-előrejelzések összeállításához, amelyek előrevetítenék, hogy milyen irányban alakulnak bizonyos folyamatok, a piac, hogyan javíthatók a köz- vagy magánszolgáltatások, folyamatok és struktúrák, és hogy mit szeretnének az állampolgárok, vásárlók vagy ügyfelek. Ideális esetben ennek révén a szervezetek olyan helyzetbe kerülhetnek, hogy proaktív döntéseket hoznak ahelyett, hogy fáziskéséssel reagálnak az egyes eseményekre.

A megfelelően felhasznált Big Data magyarázatot adhat az állampolgárok viselkedésére, segítséget nyújthat helyzetek, piacok felméréséhez, javíthatja a tájékoztatást és az értékesítési kampányokat, támogatást adhat a jogalkotásnál, árképzésnél, és optimalizálhatja az ellátási folyamatokat. A pénzintézetek valós időben értékelhetik ki az aktuális piaci fejlemények kockázatait, és ennek megfelelően alakíthatják termékpalettájukat. A közlekedési központok, útfenntartók figyelmeztetéseket küldhetnek a dugókról közvetlenül a kapcsolódó utakon közlekedőknek, és így hozzájárulhatnak a biztonságosabb és környezetbarát közlekedés kialakításához.

Minden, az ezekhez szükséges bemeneti adat adott. A Big Data analitika az az eszköz, amely segít a hatalmas adattömeget összegyűjteni, integrálni és elemezni, majd az állampolgárok, illetve felhasználók számára felhasználhatóvá tenni.

A Big Datának három alapvető jellemzője van. Ez a „3V”:
- mennyiség (Volume),
- sebesség (Velocity),
- változatosság (Variety).

Egyes gondolkodók szerint ezeket további két elemmel kell kiegészítenünk:
- igazságtartalom (Veracity),
- érték (Value),

és „5V”-ről érdemes beszélni.

A mennyiség (Volume) a másodpercenként generált adatok nagy mennyiségére utal. Gondoljunk csak az összes e-mailre, Twitter-üzenetre, fényképre, videóklipre és szenzoradatra, amelyet minden pillanatban megosztunk. Nem terabájtokról, hanem exabájtokról vagy zettabájtokról beszélünk. A Facebookon keresztül naponta 10 milliárd üzenetet küldünk, emellett a gombra kattintunk 4,5 milliárd alkalommal, és naponta 350 millió új képet töltünk fel. Ha a világon generált összes adatot az idők kezdetétől a 2000. évig összeadjuk, akkor ugyanazt azt a mennyiségű adatot kapjuk, amelyet minden percben generálunk! Ezáltal az adatkészletek túlságosan nagyok ahhoz, hogy tároljuk és elemezzük a hagyományos adatbázis-technológiákkal. A Big Data technológiával az adatokat elosztott rendszerek segítségével tárolhatjuk és használhatjuk, ahol az adatok részei különböző helyeken tárolódnak, amelyeket csak a hálózatok és a szoftverek kapcsolnak össze.

A sebesség (Velocity) az új adatok generálásának sebességét és az adatátviteli sebességet jelenti. Gondoljunk csak a közösségi médiában szereplő percek alatt terjedő üzenetekre, a hitelkártya-tranzakciók ezredmásodpercek alatt történő csalásellenőrzésére vagy az online kereskedőrendszerek pillanatok alatt meghozott döntésre a részvények vásárlásáról vagy eladásáról. A Big Data technológia lehetővé teszi számunkra, hogy elemezzük az adatokat, miközben azokat generáljuk akár anélkül is, hogy adatbázisokat hoznánk létre.

A változatosság (Variety) a jelenleg már használható különböző típusú adatokra vonatkozik. A múltban olyan strukturált adatokra összpontosítottunk, amelyek jól illeszkednek táblákba vagy relációs adatbázisokba, például pénzügyi adatok (termék vagy régió szerinti értékesítés). A világ adatainak 80%-a jelenleg strukturálatlan, ezért nem lehet könnyen elhelyezni táblázatokba vagy relációs adatbázisokba – gondoljunk fotókra, videókra vagy közösségimédia-posztokra. A Big Data technológiák segítségével mostantól különböző típusú adatok, többek között üzenetek, közösségimédia-kapcsolatok, fotók, szenzoradatok, videók vagy hangfelvételek kombinálhatók hagyományos, strukturált adatokkal.

Az igazságtartalom (Veracity) az adatok valótlanságára vagy megbízhatóságára utal. A Big Data sokféle formájával a minőség és a pontosság kevésbé szabályozható, például Twitter-üzenetek hashtagekkel, rövidítésekkel, elgépelésekkel és társalgási nyelvben használt kifejezésekkel. A Big Data analitikai technológia lehetővé teszi számunkra az ilyen típusú adatokkal való munkát. A mennyiség sokszor pótolja a minőség vagy pontosság hiányát.

Az érték (Value) azt jelenti, hogy képesek vagyunk adatainkat értékké alakítani. A szervezetnek látnia kell, hogy mit akar elérni a Big Data technológiákkal, és ezekkel hogyan fog értéket (például közjavakat, közérdekű információt, veszélyjelzést) teremteni.

Hogyan kapcsolódik egymáshoz a számítási felhő (cloud) és a Big Data? Úgy, hogy felhő nélkül nincs Big Data. Az internet és a felhőszolgáltatások, magánéletünk fokozódó

digitalizációja és a közigazgatási, valamint üzleti folyamatok elektronizálódásának térhódítása egyszerre teszi szükségessé és lehetővé a Big Datát. A felhőalapú számítástechnika az egyetlen lehetőség, hogy támogassuk a Big Data hihetetlenül nagy információsinfrastruktúra-igényeit, mivel az hatalmas tárolókapacitást, valamint nagy teljesítményű szervereket és adatbázisokat kínál. Tehát a közigazgatásban is a közigazgatási (privát) felhő kialakítása lehet majd az ideális megoldás. A Big Data segítségével olyan társadalmi és gazdasági folyamatokat lehet majd előre jelezni, amelyek alapvetően meg tudják változtatni az állam szerepköreit. A jogalkotás követőszerep helyett proaktívvá válhat, a Gazdasági Versenyhivatal még a kartellezés előtt felismerheti ennek közvetlen veszélyét, vagy éppen az állampolgárok szociális hálózatokon történő kommunikációja alapján fel lehet mérni az igényeiket, kiváltva ezzel például egy nemzeti konzultációt. Ezek csak példák voltak, nyilván az állam, mint a legnagyobb adatkezelő, rendelkezésére álló adatokból vagy az általa megszerezhető adatokból számtalan hasznos felhasználás adódhat.

Korábban, a „kevés adat" korában a személyes adatok és magánszféra védelme volt előtérben. Egyes kutatók szerint a Big Data korában már cselekvési és döntési szabadságunkat kell félteni, hiszen nem kizárt, hogy idővel olyan döntésekhez is felhasználják ezeket a rendszereket, amelyek csak előrejelzéseken alapulnak. Példa lehet erre a tudományos-fantasztikus irodalomból már ismert viselkedéselemzés-alapú büntetés: a potenciális bűnelkövetőről ezt előre megállapítjuk, így büntethető még a bűncselekmény elkövetése előtt. Ez rosszabb esetben az állam információs túlhatalmához, jogköreinek káros túlburjánzásához vezethet. Önnek mint potenciális közigazgatási vezetőnek a Big Data előnyeit és hátrányait is ismernie kell, hogy megfelelő döntést hozhasson a jövőben.

Összefoglalás

Az adatokat hagyományosan a „kevés adat" korában, a 20. században strukturáltan, adatbázisokban tároltuk. Az adatbázisokról, adatmodellekről korábban már más tantárgy keretében tanultak, ezért ezzel itt nem foglalkoztunk. Még mindig a hagyományos strukturált adatok kezelése tekintetében felmerült ezek összegyűjtése. Az adattárház egy olyan speciális adatbázis, amely az adatokat lekérdezési, elemzési műveletekre optimalizált szerkezetben tárolja (szemben a hagyományos adattárolási, tranzakciókövetési céllal). Az online analitikus feldolgozás (online analytical processing – OLAP), az adattárházakban történő lekérdezési műveletek végrehajtásának módja a többdimenziós adatmodellben. OLAP-kritérium a multidimenzionális fogalmi nézet, az átláthatóság, az elérhetőség, az állandó lekérdezési teljesítmény, a kliens/szerver-architektúra, az általános dimenzió, a dinamikus ritkamátrixkezelés, a többfelhasználós üzemmód támogatása, a korlátozás nélküli keresztdimenziós műveletek, az intuitív adatkezelés, a rugalmas jelentések, valamint a korlátlan dimenziószám és aggregációs szint. Az adattárházak alapját képezik az adatbányászatnak, amely a nagy adatkészleteken történő rendezés folyamata a minták azonosítására, a kapcsolatok kialakítására és a problémák megoldására az adatok elemzése révén.

Az adatbányászat négy szakasza az adatforrások feltérképezése, az adatfeltárás, a modellezés és a modellek alkalmazása. A nagy mennyiségű adat hatékony kezelésére NoSQL-adatbázisokat érdemes használni. A következő generációs (NoSQL-) adatbázisok jellemzően (de nem feltétlenül) nem relációs adatmodell alapján épülnek fel, akár földrajzilag is elosztottak,

nyílt forráskódúak és horizontálisan skálázhatók. Ezek négy kategóriába sorolhatók: kulcs–érték-tárolók, oszlopalapú tárolók, dokumentumtárolók és gráfadatbázisok. Hatékonyabb adatkezelés valósítható meg, ha az adatbázisokat nem a háttértárolón, hanem az operatív tárban tároljuk és kezeljük. Ezt hívjuk memóriaalapú adatkezelésnek (In-Memory DataBase). A fenti technológiák felhasználásával és továbbfejlesztésével beszélhetünk Big Datáról. Big Data az a technológia és megközelítésmód, amely a hihetetlen mennyiségű adat költség- és teljesítményhatékony tárolását, valamint kiértékelését teszi lehetővé. Megközelítéstől függően három vagy öt alapvető jellemzője van: mennyiség (Volume), sebesség (Velocity), változatosság (Variety), valamint ezeket kiegészítve az igazságtartalom (Veracity) és az érték (Value). A Big Data nagy lehetőségeket és nagy társadalmi veszélyeket hordoz magában.

Fogalmak

- adatpiac (data mart)
- adattárház (data warehouse)
- ACID-követelmények
- BASE-követelmények
- Big Data
- CAP-tétel
- Dolgok Internete (Internet of Things – IoT)
- dokumentumtároló (document store)
- érték (Value)
- gráfadatbázis (graph store)
- igazságtartalom (Veracity)
- kulcs–érték (key–values) -tároló
- memóriaalapú adatkezelés (In-Memory DataBase)
- mennyiség (Volume)
- mesterséges intelligencia (artifical intelligence – AI)
- NoSQL
- online analitikus feldolgozás (online analytical processing – OLAP)
- oszlopalapú tároló (column store)
- sebesség (Velocity)
- tetszőleges hozzáférésű memória (Random Access Memory – RAM)
- változatosság (Variety)

Áttekintő kérdések

1. Mire alkalmazhatók az adattárházak, és miben különböznek az adatbázisoktól?
2. Melyek az új generációs (NoSQL-) adatbázisok jellemzői?
3. Röviden ismertesse az új generációs (NoSQL-) adatbázis típusait!
4. Mi a memóriaalapú adatkezelés lényege?
5. Mi a Big Data öt fő jellemzője?
6. Milyen előnyei és hátrányai lehetnek a Big Data alkalmazásának?

Felhasznált irodalom

Codd, E. F. – Codd, S. B. – Salley, C. T. (1993): *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate.* Ann Arbor (US–MI), Codd & Associates.

Sidló Csaba (2004): *Összefoglaló az adattárházak témaköréről.* Budapest, ELTE. Elérhető: http://scs.web.elte.hu/Work/DW/adattarhazak.htm (A letöltés dátuma: 2018. 01. 10.)

Ajánlott irodalom

Bertot, John Carlo – Gorham, Ursula – Jaeger, Paul T. – Sarin, Lindsay C. – Choi, Heeyoon (2014): Big Data, open government and e-government: Issues, policies and recommendations. *Information Polity,* Vol. 19, No. 1–2. 5–16. DOI: https://doi.org/10.3233%2Fip-140328

Csurilla Károly (2016): *Adatbázisok, NoSQL, adattárház, adatbányászat.* Budapesti Metropolitan Egyetem.

Gyurkó György (2008): *Üzleti alkalmazások és üzleti rendszerekben alkalmazott IT megoldások.* Budapest, Budapesti Gazdasági Főiskola.

Loews, Bart (2015): What are the main differences between the four types of NoSql databases (KeyValue Store, Column-Oriented Store, Document-Oriented, Graph Database)? *Quora.com.* Elérhető: www.quora.com/What-are-the-main-differences-between-the-four-types-of-NoSql-databases-KeyValue-Store-Column-Oriented-Store-Document-Oriented-Graph-Database (A letöltés dátuma: 2018. 01. 20.)

Marr, B. (2015): Why only one of the 5 Vs of Big Data really matters? IBM Big Data and Analytics Hub. *Ibmbigdatahub.com,* 2015. 03. 19. Elérhető: www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters (A letöltés dátuma: 2018. 01. 20.)

Marton József Ernő (2012): *Memória-adatbázisok.* Budapest, BME-VIK TMIT.

Mergel, Ines (2016): Big Data in Public Affairs Education. *Journal of Public Affairs Education,* Vol. 22, No. 2. 231–248. DOI: https://doi.org/10.1080%2F15236803.2016.12002243

Quittner Pál – Baksa-Haskó Gabriella (2007): *Adatbázisok, adatbázis-kezelő rendszerek.* Debrecen, DE AMTC AVK. Elérhető: http://miau.gau.hu/avir/intranet/debrecen_hallgatoi/tananyagok/jegyzet/25-Adatbazisok.pdf (A letöltés dátuma: 2018. 01. 01.)

Rouse, M. (2019): Data mining definition. *Techtarget.* 2019. 02. Elérhető: http://searchsqlserver.techtarget.com/definition/data-mining (A letöltés dátuma: 2019. 12. 11.)