Hogyan lesz térbeli a Big Data?

Nagy méretű téradatok elosztott feldolgozása

Olasz Angéla

DOI: <https://doi.org/10.30921/GK.70.2018.1.4>

Bevezetés

A Big Data kifejezéssel manapság gyakran lehet találkozni informatikai hírekben, magazinokban, cikkekben, blogokon, TV- és rádióműsorokban, tudományos cikkekben és könyvekben. A kifejezés gyors elterjedése talán egyszerűségének köszönhető, azonban a Big Data nem csak „nagy adatot” jelent, ennél komplexebb fogalom. A Big Data jelenség természetesen magában foglalja a rendelkezésre álló (extrém) nagy mennyiségű adatot, annak minden jellemzőjével (Laney 2001, NESSI 2012), de a mögöttes infrastruktúra technológiai megoldásait és az elemzésekhez alkalmazott algoritmusokat, eljárásokat is (Grady–Chang 2017). A térinformációs iparban ezzel szinte egy időben jelennek meg az informatikai modulokra építő fejlesztések térbeli problémák megoldására. A Big Data robbanás 2012 körül következett be, amikor minden, versenyképességét javítani kívánó IT-cég Big Data megoldást kezdett kutatni, fejleszteni, szolgáltatni. A jelenség ma is töretlen, bár egyes vélemények szerint visszaesett a népszerűsége. Ez a befektetések értékében azonban nem tükröződik, inkább a növekedés lassulása figyelhető meg (van der Meulen 2016). Az akadémiai szektort tekintve, nem tapasztalható

csökkenés; például a Scopusban¹ indexelt publikációk száma (1. ábra) minden évben jelentősen felülmúlja az előző évit². A földtudományi szakkikre szűkítve a keresést a tendencia azonos³ (2. ábra).

A növekedéséhez nagyban hozzájárul az adatok (téradatok) nagy tömegű előállításának bővülő lehetősége, új felvételezési technológiák megjelenése. Ha a rendelkezésre álló téradatok mennyiségi növekedésének ütemét tekintjük, megállapítható, hogy a tárolási és feldolgozási kapacitás egyértelműen nem képes ezt követni (Olasz 2017). Szükség van az elosztott feldolgozást támogató megoldások bevezetésére, használatára a téradatok elemzése során is. Lehetetlen a Big Data megoldásai nélkül az adatok (téradatok) összekapcsolt feldolgozása (Linked Data⁴); az eddig

használt módszerekkel nem kutatható új tématerületek vizsgálata (Holt 2017), új információk levezetése. A jelenleg használt algoritmusok nem képesek az új igények és a növekvő mennyiségű téradatban rejlő információegyüttes, gyors – akár valós idejű – kinyerésére, ezért szükség van újak kidolgozására (a meglévők adaptálása mellett) a térinformatika és a távérzékelés (beleértve a fotogrammetriát is) területén.

A Big Data definíciók továbbgondolása

A Big Data definíció megalkotása nem egyszerű feladat, tekintve, hogy a technológiai háttér folyamatosan változik. Annál is inkább, hiszen az 1997-ben a „Big Data” fogalomnak megfelelő adatmennyiség mára már kicsinek tekinthető, míg a mai adatmennyiség a jövőben szintén átértékelődik. Ebből fakadóan az a meghatározás, amely a feldolgozási kapacitás meghaladásával jellemzi a jelenséget, nem elégséges a jelenség leírására.

Az 1997-es, első fogalom-meghatározás (Cox–Ellsworth 1997) óta eltelt 20 év szerves fejlődésének ismertetése meghaladná e cikk kereteit. Az általam fellelt legutóbbi mérvadó forrás, az amerikai *Nemzeti*

használt a formalizált, strukturált és rendezett adatok vizsgálatában.

1. ábra. A Big Data kifejezés előfordulásának alakulása 2000–2016 között a Scopus adatbázisban

2. ábra. A Big Data kifejezés előfordulásának alakulása 2000–2016 között a Scopus adatbázisban, a földtudományos publikációkban

Szabványügyi és Technológiai Intézet (National Institute of Standards and Technology) által létrehozott munkacsoport (Big Data Public Working Group) munkájára támaszkodva ismertetem a Big Data fogalom főbb jellemzőit (Grady-Chang 2017):

„A Big Data terjedelmes – a mennyiség (Volume), a változatosság (Variety), sebesség (Velocity) és/vagy a változékonyság (Variability) jellemzőket tekintve – adatokból áll, amely a hatékony tárolás, kezelés és elemzés elvégzése érdekében megköveteli a skálázható architektúrát.”

A korai értelmezések nagy változáson mentek keresztül, mára a korábbi koncepció komplex, tökéletesen definiált fogalommal vált. A Big Data Analitika elnevezésből rövidített „Big Analitika” azon feldolgozó eljárások összessége, amelyek képesek a Big Data követelményeknek megfelelő adatokon hatékony módon elemzést végezni (Russom 2011). A 3. ábrán a Big Data fogalom alkotóelemeit szemléltetem (összefoglalva az áttekintett definíciókat). Megállapítható, hogy a három fő alkotóelem, az *adat*, az *analitika* és az *infrastruktúra és számítási környezet*. A 3. ábra egyes halmaiba szereplő jellemzők és elemek tükrözik a Big Data megoldások sokféleségét, melyek a feldolgozott adatokban, az alkalmazott infrastruktúrában vagy a támogatott analitikai eljárásokban nagyban eltérhetnek egymástól, mégis a Big Data technológia

szerves részei. Az *adat* alkotóelem a legjobban definiált része a jelenségnek. Legkorábban ismerték fel a rendelkezésre álló adatok megnövekedett mennyisége, sokfélesége, keletkezési ideje miatti tárolási és feldolgozási problémákat, amelynek köszönhetően a Big Data-adatok jól definiált jellemzők bevezetésével azonosításra és leírásra kerültek. Később az *analitikai* megoldások kidolgozása került a középpontba, felismerve, hogy új generációs eljárások kidolgozása szükséges (természetesen a meglévő matematikai, statisztikai algoritmusokra épülve) az értékes információk kinyerésére. A két alkotóelem együtt állította fel a mögöttes infrastruktúra (hardver) és számítási környezeti (szoftver) követelményrendszerét, és ezáltal továbbfejlődési lehetőséget kijelölve a korábban már létező rendszereknek. Az *infrastruktúra és számítási környezet* (szoftver-megoldások, keretrendszerek) alkotóelem magába foglal minden olyan megoldást, amely képes az *adat* és *analitika* alkotóelemek által támasztott igény kielégítésére. Tehát minden olyan komplex megoldás, ahol extrém nagy mennyiségű adat, gyors, elosztott feldolgozása történik. Természetesen az alkotóelemek egyes egységei folyamatosan fejlődnek külön-külön is, és az újabb megoldások kidolgozása és bevezetése kölcsönösen hat az egyes alkotóelemekre, a felhasználói igényekhez igazodva, így alkotva meg a Big Data komplex rendszerét.

A Térbeli Big Data meghatározása

A Térbeli Big Data (Geospatial Big Data, Spatial Big Data stb.) megjelenése csak idő kérdése volt, hiszen a Big Data technológia informatikai megoldásai minden tudományterületre hatást gyakorolnak. A téradatok feldolgozásában egyre nagyobb szükség van az új informatikai megoldások rövid időn belül történő adaptálására, hiszen enélkül nem leszünk képesek hatékonyan szűrni, tárolni, feldolgozni és megjeleníteni a megnövekedett adatmennyiséget. A „tradicionális” földrajzi információs rendszerek kétféle objektumleíró csoportra támaszkodnak: a raszteres és a vektoros adatmodellekre (Elek 2006). Napjainkban az új típusú szenzorok, leginkább a lézerszkennerek hatalmas tömegben állítanak elő háromdimenziós pontfelhőket, amelyek kulcsfontosságú téradatokká váltak. Ezen kívül a közösségi hálózatokon nagy mennyiségű új típusú, helyadatokkal ellátott téradat keletkezik, amelyek eltérő tárolási és elemzési módszereket igényelnek (Ivan et al. 2017).

A Big Data meghatározásához hasonlóan a Térbeli Big Data is a *mennyiség*, a *sebesség* és *változatosság* jellemzőkkel került leírásra, amely meghaladja a hagyományos térbeli számítási kapacitást (Kambatla et al. 2014, Li et al. 2015, Lee-Kang 2015). A Big Data-technológiában már jól ismert jellemzők a Térbeli Big Data-ra is jól illeszthetők, bár az egyes jellemzők hangsúlyossága eltérő. A Térbeli Big Data-technológiában is helyet kapnak további dimenziók, hiszen a *megbízhatóság* (Veracity), a *kapcsoltság* (Valence), az *érték* (Value) és a *megjelenítés* (Visualization) jellemzők kiemelten fontosak a téradatok elemzésekor is (Miller-Goodchild 2014, Lee-Kang 2015). A jellemzők hangsúlyát sokszor az aktuális alkalmazási terület definiálja (Evans et al. 2013). Az alábbi táblázatban adattárolási típusokra bontva adom meg a Térbeli Big Data tulajdonságait (1. táblázat).

A Térbeli Big Data jelenségének leírásához egyaránt figyelembe kell venni a *téradatok*, a *feldolgozási eljárások* (analitika) és az *infrastrukturális háttér és számítási környezet*

3. ábra. A komplex Big Data definíció elemei

Adattípus (formátum)	Mennyiség (Volume)	Sebesség (Velocity)	Megjelenítés (Visualization)	Analitika
Vektor (pont, vonal, poligon)	globális és/vagy nemzeti szintű felszínborítás, föld-megfigyelési, környezeti adatok, nemzeti ingatlan-nyilvántartási adat, vízfolyások, közmű, közlekedési hálózat adatai (vektoros formátumban tárolva), tematikus adatok stb.	Valós idejű, gyors reagálású adatszolgáltatás és monitoring, állandó megfigyelése a környezeti változóknak és az emberi beavatkozásoknak. korai előrejelző rendszerek (cunami, földrengés, hurrikán stb.)	Az OGC és más formátum-konverziós és adatpublikációs és -megosztó szabványoknak, nyílt technológiáknak köszönhetően a Web-GIS platformokon megjeleníthetők és elemezhetők a vektoros adatok.	átlapoló műveletek: szűrés, átkódolás, intervallumba sorolás, osztályozás, unió, merge, újraosztályozás, övezetképzés, zonális statisztika, távolságszámítások, térbeli korrelációs vizsgálatok, térbeli kapcsolás stb.
3D (pontfelhő, TIN)	mobil térképező rendszer (MMS) pontfelhő, LiDAR pontfelhő, sztereo légifelvétel-modell alapú felszínmodellek (DTM, DSM), BIM pontfelhő stb.	az időérzékeny 3D-s adatok gyors feldolgozást igényelnek (katasztrófakezelés és -szimuláció)	a 3D-s (perspektív) megjelenítés tematikus adatokkal kiegészítve, lehetőség van nyíltan publikálni (új eljárásoknak köszönhetően már a weben is képesek vagyunk pontfelhő publikálására [GPU-gyorsítás])	modellezés, félautomatikus és automatikus objektumfelismerés -kinyerés, osztályozás, szimuláció, város-/egyéb 3D-s modell-készítés és elemzés, és szimulációs repülés virtuális kameraképből, láthatósági vizsgálatok, térfogati keresztmetszeti vizsgálatok, BIM bejárhatósági elemzések stb.
Raszter (grid)	idősoros műhold- és radar-felvételek, hiperspektrális felvételek, ortofotók, globális, országos, regionális vagy helyi lefedettségben, hagyományos repülő vagy UAV felhasználásával, globális topográfia, klímaadatok, geovideo	valós idejű térbeli, idősoros földmegfigyelési adatfeldolgozás, közzététel, előrejelzés, szimuláció	A földmegfigyelési adatfeldolgozás és vizualizációs megoldások fejlesztése szükséges a gyors publikálás érdekében, regionális nemzeti és közösségi szinten.	Adatelőkészítési lépések (korrekciók, transzformációk, képfeldolgozási előkészítése a feldolgozáshoz igazodva) spektrális indexelés, főkomponens-analízis, transzformációk, osztályozás, szegmentálás, klaszterezés stb.
Gráf (csomópont, él)	közlekedési hálózatok (vízi, közúti, vasúti, légi, hajózási) útvonalainak milliárdnyi számú él- és csomópontadatainak a feldolgozása (közösségi hálón rögzített bejegyzésekkel), közműhálózatok gráfadatbázisai + támogató láncok időbélyegzővel ellátva, vonulási útvonalak, a természeti környezet mozgásainak monitoringja	mozgó objektumok valós idejű monitoringja, közlekedési döntések online támogatása azonnali, katasztrófa esetén azonnali „okos” közlekedési, szállítási adatok szolgáltatása	A hálózatelemzési, nyomkövetési és útválasztási elemzésekben a megjelenítési komponens elengedhetetlen része a valós időben történő megjelenítés.	következő generációs forgalomirányítás: (közlekedés management, adott időre történő számítások online követése, döntéstámogatás), csomópont-elkerülési számítások, útvonalak rangsorolása nyomkövetés, eco-routing, eco-parking, geofencing
Földrajzi információval ellátott média (szöveg: bejegyzések, tweetek, webnaplók, bejelentkezések, video, fénykép, geoPDF stb.)	szövegalapú hagyományos Big Data földrajzi hellyel ellátva, (helyfüggő Big Data) geo-szocio adatbázis, Geo-LoT, (koordináták, címek, földrajzi nevek stb.)	A valós idejű közösségi hálózatok adatainak elemzése gyorsabb, mint valaha a tér-információs Big Data megoldások elsődleges terepe, az emberi interakciók marketing célú nyomon követése valós időben zajlik.	A helyfüggő közösségi média, webes jelenségek monitoringja és levezetett információk elemzése és megjelenítése a Big Data elemzések egyik alapeleme, rendhagyó megoldások jelennek meg nap mint nap a vizuális analitika területén, melyek a földrajzi elhelyezkedést is integrálják.	Adatbányászat, geostatistikai technológiák alkalmazása, prediktív modellezés a Térbeli Big Analitika elemei is, mely kiterjed a webtartalom elemzésre is.

hármását. A Térbeli Big Data fogalmai az elmúlt 5 évben kerültek megfogalmazásra, közöttük azonban számos ellentmondás fedezhető fel. Jelentős számú idevágó irodalmi forrás feldolgozása után (amelyek egyenkénti ismertetésétől terjedelmi korlátok miatt most eltekintek) megállapítható, hogy a jelenleg elérhető Térbeli Big Data meghatározások csak részben érintik az egyes elemeket, például csak a nagy téradatok jellemzőire fókuszálnak. Jelenleg tehát nincs átfogó, komplex megfogalmazás a téradatok tulajdonságaiból fakadó, azok különleges kezelési igényeit kielégítő Térbeli Big Data technológia meghatározására.

Ennek áthidalására eddigi munkám eredményeképp az alábbi megfogalmazást javaslom:

A Térbeli Big Data technológia nagy mennyiségű, nagy sebességű és igen változatos téradat (vektor, raszter, pontfelhő, helycímkézett szöveges adatok) gyors és hatékony feldolgozására képes, ehhez felhasználja az innovatív Térbeli Big Analitika eszközeit (térinformatikai, raszteres képfeldolgozó matematikai-statisztikai algoritmusok összessége) és a modern Big Data infrastrukturális és számítási környezet elemeit, így lehetővé válik a téradatok összekapcsolt feldolgozása, adatcseréje, és megjelenítése. Ezáltal megvalósítja a nagy adatokon alapuló térbeli és időbeli döntéstámogatás technológiai alapját, amely különös figyelmet szentel a téradatok forrásainak megbízhatóságára és a biztonsági elvárásokra.

Az infrastruktúrával és számítási környezettel szemben megfogalmazott elvárások a téradatok elosztott feldolgozásában

Könnyen belátható, hogy egyre nagyobb számítási teljesítményt kell elérni ahhoz, hogy a Big Data feldolgozás elvégezhető legyen, ezen kívül szükség van olyan eljárásokra, amelyek segítségével képesek vagyunk az adatok szűrésére, redukálására, az értékes információ kinyerése érdekében. Az elemzéseket belátható időn belül csak elosztott környezetben lehet elvégezni, ahol az egyes munkafolyamatok

képesek párhuzamosan futni az elosztott adatokon. A Térbeli Big Data feldolgozás eszközeit a „hagyományos” Big Data technológia által kínált hardver és szoftver-architektúra kínálja, azonban a téradatok általában különleges adatkezelési és feldolgozási igényeket támasztanak.

Az alábbiakban megfogalmazom az ideális Térbeli Big Data tároló és elemzőrendszerrel kapcsolatos legfontosabb elvárásokat.

Architektúrával és keretrendszerrel szemben támasztott elvárások (fontossági sorrend nélkül)

Általános elvárások

1. **skálázhatóság:** a rendszer áteresztőképességének a rugalmas növelése. A skálázhatóság annak a biztosítása, hogy a rendszer teljesítményét a terhelés növekedésével fokozatosan növelni lehessen újabb szerverek, komponensek (CPU, RAM, HDD) hozzáadásával.
2. **hibatűrés:** a működés során bekövetkező hibák észlelése, javítása és kiküszöbölése.
3. **elosztási átlátszóság lehetősége:** a mögöttes infrastruktúra elemeinek rugalmas alakíthatósága a felhasználó profiljához igazodva.
4. **függetlenség a programozási modelltől (pl. MapReduce):** az eljárás kisebb, egyenlő méretű darabokra bontja fel a feldolgozandó adatot, amit eloszt a feldolgozó node-ok valamelyikére és végül a feldolgozott darabokat egyesíti.

Szakterület-specifikus elvárások

1. a már meglévő programok futtatásának (esetfüggő) támogatása. Lehetséges szintjei:
 - a) A meglévő algoritmusok teljes eszköztárát újra kell implementálni, és optimalizálni az új megváltozott környezetre adaptálva a masszív párhuzamos feldolgozásra. Előnye, hogy az új környezetben az elemzések gyorsabban végrehajthatók, minimális indítási felár merül fel. Hátránya, hogy a rendszer kiépítésekor nagy implementációs ráfordítással kell számolni.

b) A térinformatikai/távérzékelési algoritmusfejlesztők a megszokott kódolási környezetben dolgoznak, amelynek az új feldolgozási környezetben is elérhetőnek és futtathatónak kell lennie. Előnye, hogy nem kell újraírni a meglévő eljárásokat. Hátránya, hogy várhatóan nagyobb indítási felárral, ezáltal futási idővel kell számolni. Szükség van egy közvetítő szintre az infrastruktúra és az algoritmusok között. A korábban egy szálon futó algoritmusok párhuzamos futtatása nem mindig lehetséges.

2. a térinformatikai/távérzékelési algoritmusfejlesztők által előre megírt adatelosztási és összegyűjtési eljárások elérhetőek legyenek nem fejlesztő felhasználók számára, ajánlásokkal, téradatok és esetek (use-case) feldolgozási céljainak megfelelően. Lehetséges szintjei:
 - a) a felhasználó kiválaszthatja a meglévő eljárásokból,
 - b) a fejlesztő rendel hozzá alapbeállítást az esettől függően,
 - c) a rendszer a futtatási naplófájlok (log) elemzése (akár gépi tanulás) alapján ajánlja fel a hatékony stratégiákat.
3. optimalizált legyen sokféle adattípusra (szenzor, vektor, raszter, pontfelhő): optimalizált legyen az adatok keresésében és elérésében valamint a már feldolgozott adatokat is könnyen kereshetővé kell tenni a rendszerben (indexelés stb.).
4. a téradatok (vektor, raszter, pontfelhő, egyéb) könnyen és gyorsan legyenek betölthetőek a rendszerbe, akár publikus tárházak letöltési szolgáltatásainak automatizált elérésével, akár lokális központi adatszervereken, adathordozókon tárolt adatokra van szükség.

Az analitikával szemben támasztott elvárások

Általános elvárások

1. alapvető elvárás az analitika szállítása az adathoz, az adatmozgatások minimalizálása érdekében.

2. az algoritmusfejlesztői környezet esetfüggő követelményrendszere. Lehetséges szintjei:

- a) ne legyenek programozási (Python, Ruby stb.) és algoritmuskorlátozások a keretrendszer (R, Matlab, Python stb.) használatakor.
- b) a gyakran használt fejlesztői környezetet támogassa a rendszer, de továbbiakat ne.
- c) kötött fejlesztői környezet legyen, ahol csak szigorú kritériumok szerint lehet algoritmusokat implementálni.

Szakterület-specifikus elvárások

1. támogassa a térinformatikai és távérzékelési feldolgozókönyvtárakat, nyílt forráskódú feldolgozószoftvereket (pl.: GDAL, SNAP, OTB stb.).
2. támogassa a gyakran használt térinformatikai adattípust (formátumot).
3. lehessen az éppen aktuális térinformatikai és képfeldolgozási algoritmushoz megválasztani a raszteres adatelosztás és összegyűjtés (Tiling & Stitching) típusát a számítási csomópontokon (node).
4. A térinformatika és távérzékelési algoritmusfejlesztők által létrehozott, adaptált cél-térinformatikai és -távérzékelési algoritmusok legyenek tárolhatóak, kereshetők egy központi algoritmustárban.
5. Könnyen lehessen bővíteni az algoritmustárat.
6. A munkafolyamatokat szabadon kombinálható szervizek láncolataként lehessen előállítani, azaz az egyes szervizek bemeneti és kimeneti állományai, adatformátuma egymással kompatibilis legyen.

Elosztott számítási környezet

Az elosztott számítási környezet a Big Data technológia alapvető jellemzője, amely sokféle lehet a felhasznált megoldásokat tekintve. Az egyes számítási környezetek részletesebb bemutatását a korábbi cikkemben publikáltam (Olasz 2017), az alábbiakban a felhőalapú számítás (*Cloud computing*) bemutatására szorítkozom, annak a Big Data-technológiák fejlesztésében

és elterjedésében játszott szerepe miatt.

Felhőalapú számítás

A Big Data üzleti rendszerek kialakítását és a tudományos kutatásokat a nagy adatfolyamok megjelenése és a felhőalapú számítási környezet fejlődése indukálta. A felhőalapú számítás elterjedése kissé megelőzte a mai értelemben vett Big Data-technológiát, de később a Big Data platformok egyik legfontosabb alapjává vált (Yang et al. 2017). A Big Data és a felhőalapú szolgáltatások sokszor összekapcsoltan működnek (Hashem et al. 2015). A felhőalapú szolgáltatás lehetővé teszi a hálózaton keresztül való, kényelmes és széles körű hozzáférést, konfigurálható számítási erőforrások egy megosztott halmazához (Mell–Grance 2011). A felhőszolgáltatók további szolgáltatási szinteket határoztak meg az igénybe vehető rendszerek jellemzésére.

1. Infrastruktúra, mint szolgáltatás (*Infrastructure as a Service, IaaS*), ahol a hardvert (virtuális szervert, tárhelyet, számítási kapacitást és hálózati kapcsolatot) és alap rendszerprogramokat vehet igénybe a felhasználó. A felhasználónak kell telepítenie és felügyelnie a további, számára szükséges operációs rendszert, alkalmazásokat.
2. A második szint a platform, mint szolgáltatás (*Platform as a Service PaaS*), ahol a felhasználó feladata a meglévő operációs rendszerre telepíteni az általa használni kívánt programokat, programozási könyvtárakat, eszközöket, amelyeket a szolgáltató támogat. Lehetősége van a telepített programok felügyeletére, és esetleg a további alkalmazás-hosting beállításokat alkalmazhat.
3. A harmadik szint a szoftver, mint szolgáltatás (*Software as a Service SaaS*), ahol a felhasználó az előre kialakított környezetet használhatja. A kialakított környezet távolról (akár vékony kliens felhasználói felületen keresztül, pl. webböngészőből) elérheti különböző eszközökön keresztül (Mell–Grance 2011).

Később további szolgáltatási rétegek kerültek hozzáadásra, mint például

az adat, mint szolgáltatás (*Data as a Service, DaaS*), ahol a felhasználó megkapja a hozzáférést adatforrásokhoz is (pl. az Amazon elérhetővé tette a Landsat, Sentinel adatokat), vagy a modell, mint szolgáltatás (*Model as a Service, MaaS*) ahol a nyílt adatok mellett azok adatmodelljét és feldolgozási modelljét is szolgáltatják (Roman et al. 2009).

A felhőalapú számítást azonban nem lehet mindenkire szabni, vannak korlátozó tényezők is. Például a hálózati kapcsolat nagyban befolyásolja a felhőalapú rendszerek használatát, ha a hálózat terhelt vagy túl nagy az adatforgalom, a válaszidők nagyobbak lehetnek, mintha asztali gépünkön dolgoznánk. Azt is érdemes mérlegelni, hogy mennyire akarunk felhőszolgáltató-függők lenni, nehéz megvalósítani a szolgáltatóváltást a szabványosítás hiánya miatt (persze vannak már kezdeményezések a migráció elősegítésére pl. kormányzati felhő illetve MTA-felhő, amelyekben ideális esetben megvalósulhat a nemzetközi szolgáltató kiváltása hazai szolgáltatóval). A hazai térinformatikai szektorban az adatpolitika sajátosságai miatt a felhőszolgáltatások használata ellen leggyakrabban felvetett szempontok a *biztonság* és *adatvédelem* kérdései. Vannak olyan érzékeny adatok (akár nemzetbiztonsági, kormányzati, állami, egészségügyi nyilvántartások), amelyeket nem tanácsos a felhőben kezelni (Kovács 2015).

Big Analitika

A Big Analitika jelentősége abban rejlik, hogy az új technológia (infrastruktúra és feldolgozórendszerek) alkalmazása lehetővé teszi akár a régi (történelmi), akár az új adatok új típusú feldolgozását, olyan eljárások alkalmazásán keresztül, amelyre korábban nem volt lehetőség. A rendelkezésre álló nagy mennyiségű adatban rejlő értékek kinyerése érdekében szükség van új eljárások fejlesztésére. Természetesen az alábbi matematikai-statisztikai eljárások nem új felfedezések, azok nagy részét évtizedek óta használják a kutatók, viszont azok új környezetben történő felhasználása új eredményeket hozhat olyan adatokon, amelyek eddig nem, vagy csak

elszigetelten kerültek feldolgozásra. Valamint, a meglévő eljárások mellett lehetőség nyílik új eljárások fejlesztésére vagy a meglévők bővítésére, optimalizálására is, hiszen korábbiakban nem látott feldolgozó- és tárolókapacitás és adatforrás áll rendelkezésre. A Big Data olyan eljárásokat kíván, amely képes nagy adatmennyiség korlátozott idő alatt (akár valós időben) történő feldolgozására, megjelenítésére.

A Big Analitika általános feldolgozási folyamata hat fő lépésből áll, amelyek két fő csoportra oszthatók: az *adatkezelés* és az *analitika* csoportjaira (Labrinidis–Jagadish 2012). Az *adatkezelés* csoportban első lépésként elvégezzük az *adatgyűjtést*. Második lépésben elvégezzük az *adattisztítást*, *adatkinyerést* és *annotálást*. A harmadik lépésben az *aggregálást* hajtjuk végre.

Az adatkezelés után átlépünk az *analitika* főcsoportjába, ahol elvégezzük a *modellezést*, lefuttatjuk az *elemzést*, később *interpretáljuk* az eredményeket, végül *megjelenítjük* és *megosztjuk* az információkat.

Az elemzések a feldolgozás fajtája szerint öt típusba sorolhatók. A *leíró elemzések* (*Descriptive analytics*) a múltbeli történéseket és az összegyűjtött információkat érthető formában, általában diagramok, térképek formájában jelenítik meg. A *prediktív elemzések* (*Predictive analytics*) a rendelkezésre álló adatokból extrapolál előre jelezve a várható jelenségeket. Az extrapolációhoz használt eszközök általában statisztikai módszerek, neurális hálók vagy egyéb gépi tanulási algoritmusok. A harmadik a *feltáró* vagy *felfedező elemzések* (*Discovery analytics*) kategóriája, amely a nagy adatok gyűjteményei között keres nem várt összefüggéseket. Ez a típus teszi lehetővé, hogy a különböző forrásokból származó adatokat együttesen elemezzük, és ezáltal új következtetéseket vonhassunk le. A negyedik típus az *előíró elemzések* (*Prescriptive analytics*) típusa, amikor az összegyűjtött adatok alapján azonosítja a problémák megoldásait és előírja, hogyan lehet azokat kivitelezni (Rajaraman 2016). Ötödik típus a *kognitív elemzések* (*Cognitive analytics*), amelyben az észlelés, memória, tanulás, érvelés mentális folyamataihoz

hasonló elemzéseket vesznek alapul (IBM Sverige 2013).

Gépi tanulás (Machine Learning)

A Big Analitika szempontjából kiemelkedő jelentőségű a *Gépi tanulás*, olyan módszereket és algoritmusokat foglal magába, melyek a rendelkezésre álló adathalmazból egy modellt tanulnak, és ezt felhasználva következtetést (predikciót) tesznek lehetővé (a meglévő vagy más/újabb adatokról). Fontos eleme a *mesterséges intelligenciának* (Artificial intelligence), melynek célja a számítógépek számára empirikus adatok alapján viselkedési algoritmusok kialakítása. A gépi tanulás legfőbb jellemzője az automatikus tudásfeltárás (knowledge discovery) és intelligens döntéshozás. A gépi tanulás esetében is a feldolgozóalgoritmusok legfontosabb feladata a Big Data képes feldolgozás, az algoritmusok skálázásával.

A jelen kor kihívása, hogy a hagyományos matematikai-statisztikai alapokon működő módszereket a Big Data környezetbe tudjuk illeszteni, mind a feldolgozandó adatok jellemzőihez, mind az infrastruktúra és számítási környezethez igazodva.

Térbeli Big Analitika

A téradatok hagyományos feldolgozásának elemei a matematikai-statisztikai módszertanra épülnek. A térinformatikai, távérzékelési statisztikai és egyéb szakterület-specifikus eljárások azonban a Big Data területére érkezve kihívásokkal néznek szembe, ezért új megoldások kidolgozására van szükség, ahogyan az a hagyományos adatelemzési eljárások Big Data képes adaptációja során történik. Alább a *raszteres téradatokra*

koncentrálva példákkal illusztrálom a feldolgozási algoritmusok Big Data környezetbe történő adaptációjának nehézségeit.

Algoritmusok térbeli elemzésekhez

A szakirodalom a távérzékelési feldolgozó műveleteket azok kiterjedése alapján csoportosítja (Schowengerdt, 2007; Lein, 2011; Lillesand et al., 2014; Liu and Mason, 2016). Megkülönböztet lokális, fokális, zonális és globális algoritmusokat. A térbeli elemzések elvégzésekor felhasználhatunk több sávban rendelkezésre álló egyéb spektrális (intenzitás) értékeket (Liu and Mason, 2016). A kiterjedésük alapján jellemzett algoritmus kategóriák között felállíthatunk egy sorrendet tekintve, azok párhuzamosíthatóságának komplexitását. Az alkalmazott csoportosítás egyértelműen megadja az algoritmusok területhez igazodó adatfelosztási gyakorlatot (Nguyen, 2016; Ma et al., 2015).

A globális algoritmusok esetében a feldolgozás jól párhuzamosítható, mivel az adatfelosztás tervezhető, azonban az eredmények összesítésére mindenképpen kiegészítő algoritmusokra van szükség. A zonális algoritmusok esetében az adaptáció problémásabb, hiszen állandóan változó, előre nem ismert szomszédos területi kiterjedést kell figyelembe venni, mivel az objektumok térbeli lehatárolása az algoritmus futtatása során alakul ki. A művelet párhuzamosítható (elkülönült vagy egymást részben átfedő tartományokra), de gondoskodni kell a feldolgozott részek egyesítéséről. A fokális algoritmusok jól párhuzamosíthatók, hiszen előre ismert, szabályos szomszédsági kontextus figyelembe vétele szükséges,

4. ábra: Raszteres térbeli műveletek csoportosítása (forrás: innovativegis.com)

amelyhez egzakt geometriai felosztás rendelhető. Végül a lokális algoritmusok esetében szintén jól párhuzamosítható egységekre tudjuk bontani műveletet. Az alábbi táblázatban (2. táblázat) példaként mutatok be néhány képfeldolgozó eljárást, azok kiterjedése és párhuzamosíthatóságának megadásával együtt.

A Térbeli Big Data módszerek sajátosságai

Összességében elmondható, hogy az egyszerűbb pixelenkénti műveletek könnyebben alakíthatóak át, mint a komplex szomszédsági tulajdonságokat felhasználó eljárások. A 5. ábrán részletesen bemutatom a raszteres képfeldolgozás lépéseit és alkalmazott eljárásait, annak érdekében, hogy részletesebb képet kapjunk azok kapcsolatáról és elemeiről.

A munkafolyamat egyes állomásai természetesen nagymértékben esetfüggők, hiszen az elemzés célja (a bemenő adatok száma, típusa, felbontása, kiterjedése a rendelkezésre álló idő, a kimenő adatok felbontása és kiterjedése, az elvégzendő elemzés) nagyban befolyásolja lépések számát és a felhasznált algoritmusokat. A felvázolt lépések felhasználása, sorrendje is esetfüggő, azonban a folyamatok párhuzamosításának vizsgálatához szükséges a gyakran használt feldolgozási eljárások áttekintése. Emellett mindenképpen szükség van előre definiált adatelosztási és összegyűjtési (Tiling & Stitching) módszerek kialakítására minden egyes elemzéstípusra (lokális, fokális, zonális, globális és kevert esetek (Olasz et al. 2016)).

A Térbeli Big Data feldolgozási folyamat jelentősen eltérhet a hagyományos Big Data feldolgozásának folyamatától. Az *előfeldolgozás* során az esetspecifikus feldolgozásra alkalmas formába hozzuk a téradatokat, de sokszor ez a lépés iterálódik az integrációs és modellezési lépésekkel. Az *analitikai* fázisban is jelentős eltérés van a hagyományos Big Data folyamatához képest, hiszen szükség lehet további utófeldolgozási eljárások alkalmazására (pl. újramintázás, adatformátum-konverzió, szűrés) az adatok megjeleníthető formába

2. táblázat. Térbeli elemzések csoportosítása kiterjedés és párhuzamosíthatóság szerint

Művelet	Forrás	Kiterjedés				Párhuzamosítható		
		Lokális	Fokális	Zonális	Globális	Egyszerű	Közepes	Nehéz
radiometriai korrekciók	Lillesand et al. 2014	x		x		x		
spektrális indexek	Wang-Weng 2013	x				x		
multispektrális transzformációk	Lillesand et al. 2014	x			x		x	
hisztogramműveletek	Lillesand et al. 2014			x	x		x	
konvolúciós szűrések	Liu-Mason, 2016		x				x	
szegmentálás	Blaschke et al. 1970	x	x	x				x
tematikus osztályozás	Schowengerdt 2012				x		x	
objektumalapú képelemzés	Blaschke 2010	x	x	x	x			x
neurális háló-alapú osztályozás	Mas-Flores 2008	x	x	x	x		x

5. ábra: A téradatok feldolgozásának folyamata és algoritmusai (saját szerkesztés)

hozása vagy publikálása (pl. térképes webszolgáltatáson keresztül) előtt. Mindemellett a téradatok elemzése és

feldolgozása során olyan térbeli összefüggések figyelembe vétele szükséges, amely specializált algoritmusok

alkalmazását követeli meg, és ennek az adatok elosztására is jelentős hatása van. Ezért az automatizálás lehetősége kisebb ezekben a lépésekben. A szakterület-specifikus megoldások kidolgozása olyan speciális keretrendszert igényel, mely alkalmazkodik a téradatok feldolgozási folyamatához és a megfogalmazott igényekhez.

Következtetések

A Térbeli Big Data feldolgozás és a Térbeli Big Analitika eljárások értelmezésének kiindulási alapja a „hagyományos” Big Data technológia előnyeinek és hátrányainak megismerése.

Azon túl, hogy az egyes kutatás-fejlesztési vizsgálatok bizonyítják egy-egy alkalmazási példa létjogosultságát (ami erősen esetspecifikus elemeket tartalmazhat), szükség van a Térbeli Big Data feldolgozása és a Térbeli Big Analitika eljárásaiban bizonyos mértékű szabványosításra is, annak érdekében, hogy a szakterület alkalmazkodjon a kor kihívásaihoz, és jövőben is képes legyen feldolgozni a rendelkezésre álló téradatokat. Ennek első eleme, hogy a leggyakrabban használt képfeldolgozó algoritmusokat megvizsgáljuk a párhuzamosíthatóság szempontjából.

A bemutatott távérzékelési, képfeldolgozási eljárások természetesen nem tekinthetők teljes körű távérzékelési és térinformatikai gyűjteménynek, hiszen számos egyéb feldolgozási eljárásra is szükség van például a vektoros térinformatika, a fotogrammetria, a távérzékelés területéről.

Véleményem szerint a szakterület (haszonlóan másokhoz) paradigmaváltáson megy keresztül, mely nagyban meg fogja változtatni a jövőbeli munkavégzés eszközrendszerét.

Összefoglalás

A Big Data technológiák által kínált megoldási lehetőségek a nagy tömegű heterogén adatok feldolgozására a térinformatikai és távérzékelési (beleértve a fotogrammetriát is) alkalmazásokban kiaknázatlan lehetőséget jelentenek. Ahhoz, hogy a rendelkezésre álló lehetőségeket kihasználva értéknövelt termékeket, szolgáltatásokat tudjunk előállítani és közzétenni,

értékesíteni, elkerülhetetlen az új technológiák alkalmazása és adaptálása, szakterület-specifikus megoldásokkal történő kiegészítése.

A cikk első felében áttekintettem a Big Data fogalmait és létrehoztam a fogalmak közös elemeit, áttekinttem az adatforrásait. Kitértem Térbeli Big Data meghatározására, megfogalmaztam annak komplex definícióját, valamint meghatároztam a feldolgozására alkalmas rendszer infrastrukturális és algoritmikus elvárásait. Az alkalmazott infrastrukturális megoldások közül kiemelten ismertettem a felhőalapú számítási környezetet. A cikk második felében áttekintettem a Big Analitika és a Térbeli Big Analitika eszközrendszerét és különbségeit. Ismertettem néhány Térbeli Big Analitika képfeldolgozó eljárást, és felvázoltam párhuzamosíthatóságuk nehézségi fokát.

Irodalom

- Blaschke, T. (2010). Object based image analysis for remote sensing, *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(1): 2–16. DOI: <https://doi.org/10.1016/j.isprsjprs.2009.06.004>
- Blaschke, T. – Burnett, C. – Pekkarinen, A. (1970). Remote Sensing and Digital Image Processing. DOI: <https://doi.org/10.1016/j.patcog.2008.04.013>
- Chi Hau, C. – Pei-Gee, P. H. (2008). Statistical Pattern Recognition in Remote Sensing, *Pattern Recognition*, 41(9): 2731–2741. DOI: <https://doi.org/10.1016/j.patcog.2008.04.013>
- Cox, M. – Ellsworth, D. (1997). Managing big data for scientific visualization, *NASA Advanced Supercomputing Division* (Ames)
- Evans, H. – Hagen, C. (2013). Big Data and the Creative Destruction of Today's Business Models, *A.T. Kearney, Inc.*
- Evans, M. R. – Oliver, D. – Yang, K. – Zhou, X. – Shekhar, S. (2013). Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities, in *CyberGIS: Fostering a New Wave of Geospatial Innovation and Discovery* (Eds. Wang, S. – Goodchild, M.).
- Grady, N. – Chang, W. (2017). NIST Big Data Interoperability Framework, Definitions, Vol. 1, Version 2, Gaithersburg, MD 20899 USA: U. S. Department of Commerce, NIST Big Data Public Working Group (NBD-PWG).
- Hashem, I. A. T. – Yaqoob, I. – Anuar, N. B. – Mokhtar, S. – Gani, A. – Ullah Khan, S. (2015). The rise of “big data” on cloud computing: Review and open research issues, *Information Systems*, 47: 98–115. DOI: <https://doi.org/10.1016/j.is.2014.07.006>
- Holt, I. (2017). Geospatial Big Data, *GIM International*, 2017. június 6.
- IBM Sverige (2013). Gene Villeneuve – Moving from descriptive to cognitive analytics.
- IDC (2012). Big Data in 2020, IDC iView.
- Ivan, I. – Singleton, A. – Horák, J. – Inspektor, T. (2017). The Rise of Big Spatial Data, Springer DOI: <https://doi.org/10.1007/978-3-319-45123-7>
- Kambatla, K. – Kollias, G. – Kumar, V. – Grama, A. (2014). Trends in big data analytics, *Journal of Parallel and Distributed Computing*, 74(7): 2561–2573. DOI: <https://doi.org/10.1016/j.jpdc.2014.01.003>
- Kovács, Z. (2015). Az infokommunikációs rendszerek nemzetbiztonsági kihívásai, PhD-disszertáció Nemzeti Köszolgálati Egyetem, Katonai Műszaki Doktori Iskola
- Labrinidis, A. – Jagadish, H. V. (2012). Challenges and Opportunities with Big Data, *Proceedings VLDB Endowment*, 5(12): 2032–2033. DOI: <https://doi.org/10.14778/2367502.2367572>
- Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety, META Group Inc. Stamford, Connecticut, U.S.A.
- Lee, J.-G. – Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities, *Big Data Research*, 2(2): 74–81. DOI: <https://doi.org/10.1016/j.bdr.2015.01.003>
- Lein, J. K. (2011). Environmental Sensing: Analytical Techniques for Earth Observation, Springer Science & Business Media. DOI: https://doi.org/10.1007/978-1-4614-0143-8_2
- Li, S. – Dragicevic, S. – Anton, F. – Sester, M. – Winter, S. – Coltekin, A. – Pettit, C. – Jiang, B. – Haworth, J. – Stein, A. – Cheng, T. (2015). Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges, *ISPRS Journal of Photogrammetry and Remote Sensing* DOI: <https://doi.org/10.1016/j.isprsjprs.2015.10.012>
- Lillesand, T. – Kiefer, R. W. – Chipman, J. (2014). Remote Sensing and Image Interpretation, John Wiley & Sons Ltd.
- Liu, J. G. – Mason, P. J. (2016). Operations, Image Processing and GIS for Remote Sensing, John Wiley & Sons Ltd.:191–210. DOI: <https://doi.org/10.1002/9781118724194>
- Manyika, J. – Chui, M. – Brown, B. – Bughin, J. – Dobbs, R. – Roxburgh, C. – Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity, McKinsey & Company.
- Mas, J. F. – Flores, J. J. (2008). The application of artificial neural networks to the analysis of remotely sensed data, *International Journal of Remote Sensing*, 29(3): 617–663. DOI: <https://doi.org/10.1080/01431160701352154>
- Mell, P. – Grance, T. (2011). The NIST Definition of Cloud Computing, NIST Special Publication 800–145, National Institute of Standards and Technology.
- Meulen, R. van der (2016). Survey Analysis: Big Data Investments Begin Tapering in 2016. Gartner Inc., Stamford, Connecticut, U.S.A.
- Miller, H. J. – Goodchild, M. F. (2014). Data-driven geography, *GeoJournal*, 80(4): 449–461. DOI: <https://doi.org/10.1007/s10708-014-9602-6>
- NESSI (2012). Big Data, A New World of Opportunities.
- Nguyen, B. (2016). IQLib specification, IQmulus Network

- Olasz, A. (2017). Big Data és térbeliség, Geodézia és Kartográfia, (Vol.5): 12–21.
- Olasz, A. – Nguyen Thai, B. – Kristóf, D. (2016). A new initiative for Tiling, Stitching and Processing Geospatial Big Data in Distributed Computing Environments, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., Volume III-4: 111–118. DOI: <https://doi.org/10.5194/isprsannals-III-4-111-2016>
- Rajaraman, V. (2016). Big data analytics, Resonance, 21(8): 695–716. <https://doi.org/10.1007/s12045-016-0376-7>
- Roman, D. – Schade, S. – Berre, A.-J. – Bodsberg, N. R. – Langlois, J. (2009). Model as a Service (MaaS).
- Russom, P. (2011). Big Data Analytics, Fourth Quarter 2011, The Data Warehousing Institute (TDWI). Renton, USA
- Schowengerdt, R. A. (2007). Remote Sensing: Models and Methods for Image Processing, Elsevier.
- Schowengerdt, R. A. (2012). Techniques for Image Processing and Classifications in Remote Sensing, Academic Press.
- Wang, G. – Weng, Q. (2013). Remote Sensing of Natural Resources, CRC Press. DOI: <https://doi.org/10.1201/b15159>
- Wu, C. – Buyya, R. – Ramamohanarao, K. (2016). Big Data Analytics = Machine Learning + Cloud Computing, Big Data, Morgan Kaufmann, :3–38. DOI: <https://doi.org/10.1016/B978-0-12-805394-2.00001-5>
- Yang, C. – Huang, Q. – Li, Z. – Liu, K. – Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges,

International Journal of Digital Earth, 10(1): 13–53. DOI: <https://doi.org/10.1080/17538947.2016.1239771>

Summary

This paper presents the Big Data phenomenon, introduces the importance of new processing techniques to provide solutions to handle Big Data and Geospatial Big Data. Recently, volume and variety of available data are evolving as never before, exceeding the capabilities of traditional algorithm performance and hardware/software environment in the aspect of data management and computation (Manyika et al., 2011; IDC, 2012; Evans and Hagen, 2013). Hence, improved efficiency is required to exploit the available information derived from Geospatial Big Data. Consequently, geospatial analysis needs to be reformed to exploit the capabilities of current and emerging computing environments via new data management and processing concepts. To understand the evaluation of the

techniques, the differences and the requirements we need to go in deep into the Big Data solutions (including data, analytics and infrastructure, computing background). Existing Big Data definitions are provided and summarized within a figure to serve a complex perspective. After giving summary of existing Geospatial Big Data definitions I have provided my complex, synthesized version. Geospatial Big Analytics are introduced focusing on image processing algorithms (local, focal, zonal, and global) and their parallelization aspects.

Olasz Angéla
térinformatikus

Budapest Főváros Kormányhivatala
Földmérési, Távérzékelési és Földhivatali Főosztály, Térinformatikai Osztály
olasz.angela@bfkh.gov.hu