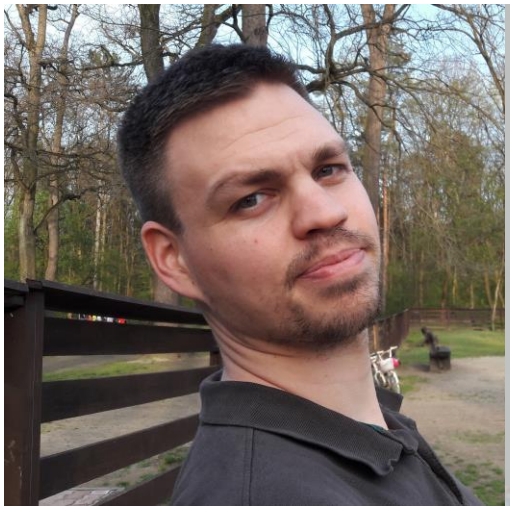



Dr. Hegedűs Péter, Dr. Ferenc Rudolf

Nagyméretű adatbázisok

Jelen tananyag a Szegedi Tudományegyetemen készült az Európai Unió támogatásával.

Projekt azonosító: EFOP-3.4.3-16-2016-00014

Szegedi Tudományegyetem
Cím: 6720 Szeged, Dugonics tér 13.
www.u-szeged.hu
www.szechenyi2020.hu

Nagyméretű adatbázisok bevezetés

Összefoglalás

Ebből az olvasóleckéből megtudhatjuk, mit is értünk pontosan a Big Data fogalma alatt. Megismerjük azokat a jellemzőket, amik megkülönböztetik a hagyományos módszerekkel feldolgozható adatokat a valóban hatalmas, dedikált módszereket igénylő Big Data forrásoktól. Áttekintést kap az olvasó arról, hol keletkezik Big Data, illetve arról, hogy milyen haszonnal jár ezen hatalmas adathalmazok feldolgozása. Valamint megismerjük a Big Data adatok három alapvető típusát is.
A lecke fejezetei:

- 1. fejezet: **Big Data fogalma, hatalmas mennyiségű adatok forrásai (olvasó)**
- 2. fejezet: **Big Data jellemzői: a 4V (olvasó)**
- 3. fejezet: **Big Data adatok típusai (olvasó)**

Téma típusa: **elméleti**
Olvasási idő: **50 perc**

📖 1. fejezet
Nagyméretű adatok, vagyis Big Data

A "Big Data" kifejezés igen népszerűvé vált napjainkra, amikoris hatalmas mennyiségben keletkeznek adatok nap mint nap, óráról órára, vagy akár percről percre. Hogyan keletkeznek ezek az adatok? Néhány példa:

- **New York-i tőzsde:** tőzsdeforgalmi adatok kb. napi 1 TB adatot tesznek ki

- **Social media**: napi 500+ TB új data kerül be a Facebook adatbázisaiba nap mint nap (képek, videók, üzenetek)

- **Szenzor adatok:** egy repülőgép meghajtója akár 10+ TB adatot generálhat 30 percnyi repülés alatt, ami napi több ezer repülés során Petabájt méretűre hízik

Elsőre azonban nem feltétlenül világos, hogy miért kellene máshogy kezelnünk az ilyen nagy méretű adatokat, és miért nem dolgozzuk fel őket a hagyományos módon. Ilyen hatalmas mennyiségű adat nem tárolható a megszokott módon, pl. nem fér el egy gépen, egy adatbázisban, nem tölthető be memóriába, feldolgozása rengeteg időt venne igénybe, stb. Más módszerek és megközelítés kell, elosztott adattárolás, párhuzamosított feldolgozás, stb.

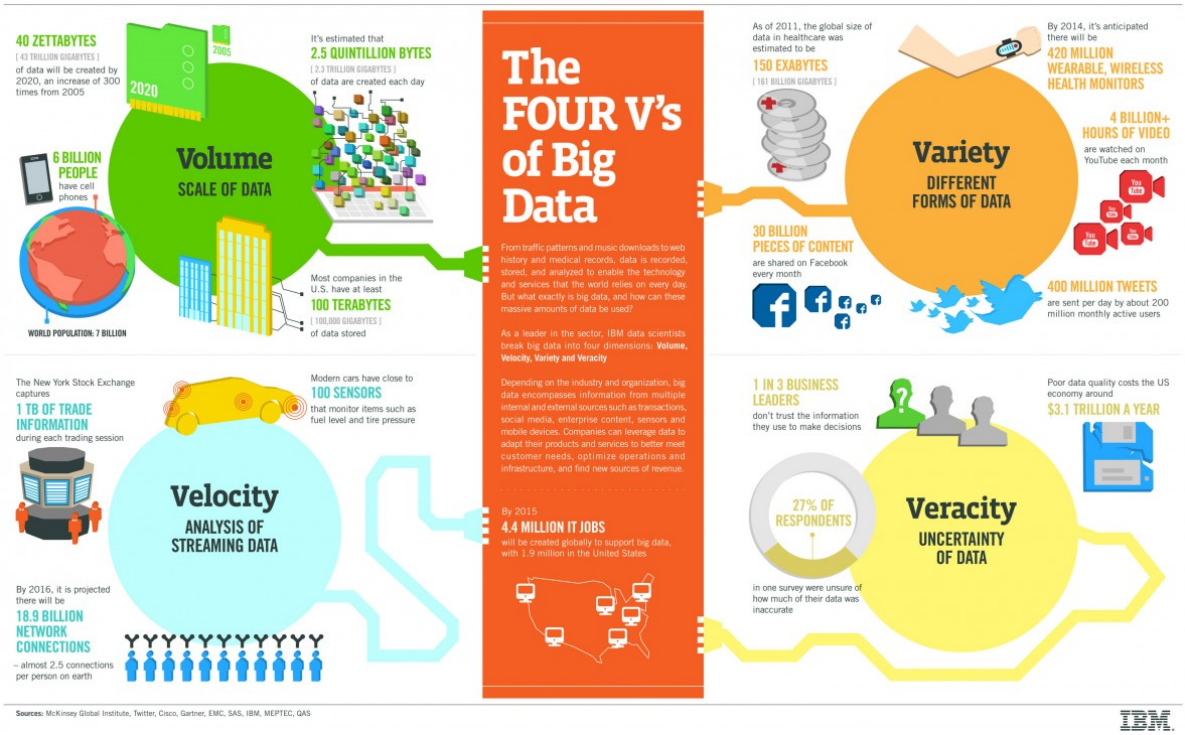
Big Data feldolgozásának előnyei

Ha olyan nagy kihívás és teljesen új technológiát igényel, mi haszna pontosan a Big Data feldolgozásának? A hatalmas méretű adathalmazok feldolgozása számos haszonnal járhat, néhány ezek közül:

- *Prediktív analitika alkalmazása* - üzleti döntéshozatal, működés optimalizálható a keletkező hatalmas mennyiségű adat feldolgozásával
- *Ügyfél elégedettség javítása* - célzott hirdetések és felhasználói visszajelzések alapján felhasználói élmény javítása
- *Irreleváns adatok kiszűrése* - a Big Data technológiák és eszközök segítségével kiszűrhetők a több forrásból érkező, irreleváns adatok, így javítva az üzleti hatékonyságot
- *Ügyfél reakciók felderítése* - szolgáltatások, termékek fogadtatása, népszerűsége hogyan alakul, mely részek fejlesztésére kell fókuszálni az időt és pénzt

📖 2. fejezet
Big Data "definíciója", a 4V jellemző

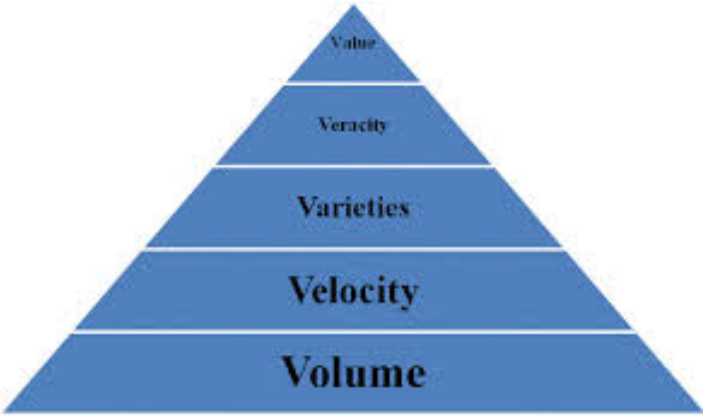
Természetesen a **Big Data** definiálása korántsem egyszerű. Ami egy kisebb cégnek már Big Data (pl. 10 TB), egy hatalmas cégnek lehet teljesen hagyományos mennyiségű adat. A határvonal nem éles a hagyományos és Big Data között. Ráadásul, egy adathalmaz kizárólag a mérete miatt még nem minősül Big Data-nak. Kezdetben 3 alapvető tulajdonsággal jellemezték a Big Data-t (mivel mind V betűvel kezdődik angolul, ez lett a **3V**), amihez folyamatosan újabb V-k adódtak hozzá (jelenleg a **4V** és az **5V** a meghatározó, de van sokkal több V is).



Ezek pedig a következők:

- **Volume:** az adatmennyiség puszta mérete. A Big Data egyik alapvető jellemzője a hatalmas méret, ami nélkül nem beszélhetünk Big Data-ról.
- **Velocity:** az adat generálásának sebessége. Big Data esetében nem csak hatalmas mennyiségű az adat, de nagyon gyorsan nő, azaz egyre több adat keletkezik. Az ilyen folyamatos és nagy mértékű adatáramlást az adatforrások felől (pl. üzleti folyamat, szenzorok, hálózat, social media) tudni kell kezelni. Ahhoz, hogy egy Big Data alkalmazás sikeres legyen, ezen keletkezett adatokat a megfelelő időn belül tudni kell feldolgozni.
- **Variety:** ez a jellemző arra utal, hogy az adatok Big Data esetén heterogének, azaz sok különböző forrásból és sok különböző formátumban jelennek meg. Míg régen alapvetően adattáblák és adatbázisok tették ki az adatok nagy részét, mára már e-mailek, fotók, videók, eszköz monitorozó mérések, PDF-ek, hang fájlok mind mind a Big Data részét képezik. Ezek miatt Big Data esetén a különböző formátumú adatok tárolását, keresését és feldolgozását is nagy kihívás megoldani.
- **Veracity:** az adatok minőségét jellemzi, azt hogy mennyi bizonytalanság (zaj) van az adathalmazban. A magas minőségű (high veracity) adathalmaz sok értékes rekordot tartalmaz, amelyek feldolgozása üzleti értéket képez, míg az alacsony minőségű (low veracity) adatnál sok az értéktelen rekord, a zaj. Értéktelen lehet egy rekord ha pl. mérési hiba miatt hiányzik, vagy teljesen valótlan értékeket tartalmaz, de akkor is, ha egyszerűen nem képez értéket az adott üzleti megoldás szempontjából (pl. redundáns adat).

Leggyakoribb kiegészítése ezeknek az 5. V, a **Value**. Egy hatalmas adat akkor értékes, ha abból valami üzletileg hasznos információt ki tudunk nyerni. Amennyiben egy adathalmaz ezzel a tulajdonsággal nem rendelkezik, Big Data eszközökkel történő feldolgozása hiábavaló technikai bravúr marad.



Ezek után kicsit precízebben definiálhatjuk mi is az a Big Data. Az olyan adatot, amelyik nagy méretű (high volume), gyorsan gyarapodik (high velocity) és nagyon heterogén az összetétele (high variety) fejlett eszközökkel és módszerekkel tudjuk csak feldolgozni, hogy értékes információt nyerjünk ki belőle. Az adat fenti jellemzői miatt azt területet, ami ezen adathalmazok tárolásával, feldolgozásával és elemzésével foglalkozik Big Data-nak nevezzük.

Egy a Gartner által használt definíció a következő:

```
“Big data is high-volume, velocity, and variety information assets that demand
cost-effective, innovative forms of information processing for enhanced insight
and decision making.”
```

📖 3. fejezet
Big Data adatok típusai

Most hogy sikerült definiálnunk mit értünk Big Data alatt, nézzük meg, hogy milyen típusai lehetnek:

- **Strukturált (structured)**
- **Nem strukturált (unstructured)**
- **Részben strukturált (semi-structured)**

Strukturált adatok

Az olyan adatokat nevezzük strukturáltnak, melyek egy előre ismert, fix formátum alapján érhetők el, tárolhatók és dolgozhatók fel. Teljesen rendszerezett információ, amely minden komolyabb erőfeszítés nélkül betölthető és feldolgozható programok által. Azonban a strukturált adatok mérete is olyan sebességgel nő, hogy napjainkban már zettabájtos nagyságrendekről beszélhetünk, aminek kezelése túlmutat a hagyományos megközelítésen. Tipikus példa strukturált adatokra a relációs adatbázisokban tárolt adat (előre definiált mezőkkel rendelkező táblázatokba rendezett rekordok).

> *Do you know? 10^{21} bytes* equal to *1 zettabyte* or *one billion terabytes* forms *a zettabyte*.

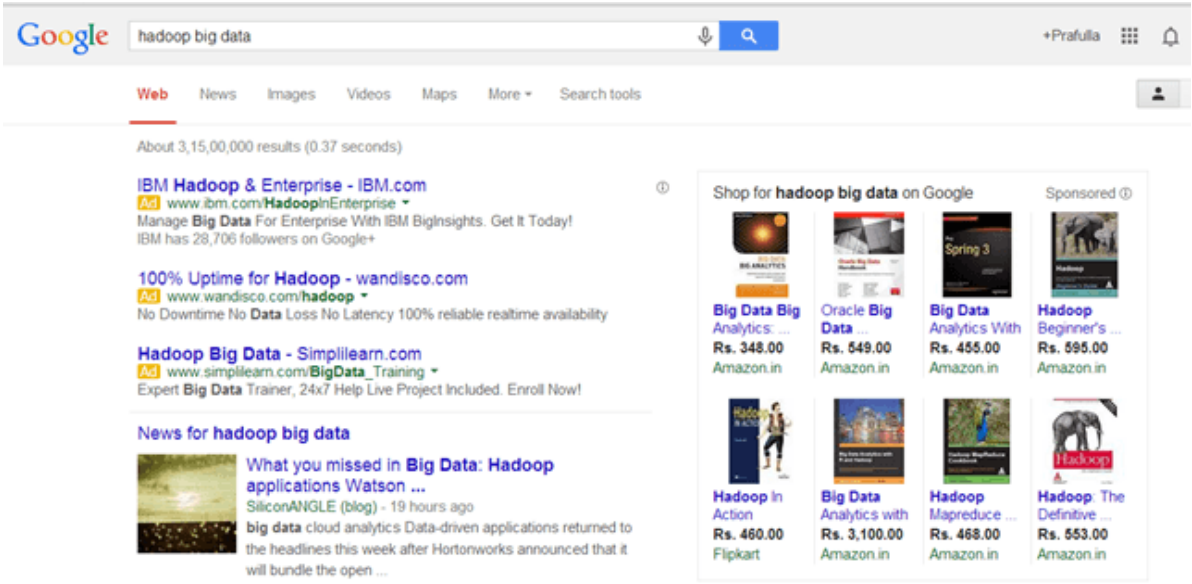
Egy példa ilyen strukturált adatra az alábbi munkavállalókat leíró táblázat:

Employee_ID	Employee_Name	Gender	Department	Salary
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Nem strukturált adatok

Minden adat, aminek a formátuma/struktúrája ismeretlen nem strukturált adatnak minősül. Az ilyen adatok esetében a hatalmas adatmennyiség mellett azok feldolgozása és belőlük értékes információ kinyerése önmagában is komoly kihívás. Tipikus nem strukturált adatok például a képek/videók, social media bejegyzések, e-mail üzenetek, stb.

Egy példa nem strukturált adatra az alábbi Google keresés eredménye:



Részben strukturált adatok

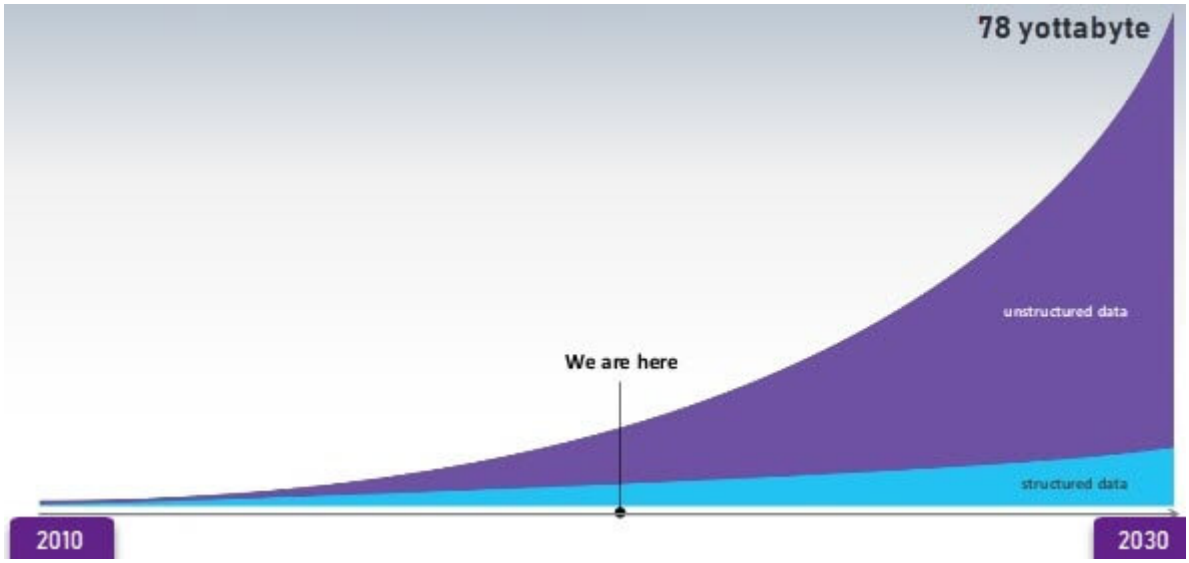
Azon adatok, amelyek nem esnek a fenti két kategória egyikébe sem egyértelműen, részben strukturált adatnak minősülnek. Az ilyen adatok ugyan nem rendelkeznek formális adat struktúra leírással, mégis tartalmaznak olyan meta-információt az adatelemekhez (pl. címkék, leírók), amelyek segítenek a rekordok csoportosításában, feldolgozásában. Azaz alapvetően strukturált adatokról beszélünk, de a struktúrájuk nincs explicit módon megadva, mint pl. egy relációs adattábla definíció (viszont jó eséllyel a megfelelő adatfeldolgozási minták kikövetkeztethetők). Az olyan adatok is részben strukturáltnak minősülnek, melyek strukturált és nem strukturált elemeket is tartalmaznak vegyesen.

Egy példa részben strukturált adatra a következő XML részlet (az egyes XML elemek tartalma, például a név nem strukturált):

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R. Latuza</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>S. Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Adatok eloszlása típusuk szerint

Ha valaki azon tűnődik, vajon milyen arányban fordul elő strukturált és nem strukturált adat a világban, az alábbi ábra igazán szemléletes lehet:



Jól látható, hogy tíz évvel ezelőtt egyértelműen a strukturált adatok domináltak, ami szépen lassan átbillent a nem strukturált adatok felé, miközben a teljes adatmennyiség volumene is töredéke volt a mainak. Napjainkban sokkal több nagy adatforrás nem strukturált, mint strukturált (pl. web alkalmazások napló fájljai, tranzakció történet fájlok). Ez az arány az előrejelzések szerint még markánsabban el fog tolódni a nem strukturált adatok irányába.

Ellenőrző kérdések

1. Mik a Big Data alapvető jellemzői? Mitől lesz egy nagyméretű adathalmaz Big Data?
2. Sorolj fel néhány olyan alkalmazást, amelyek tipikus forrásai Big Data adathalmazoknak!
3. Miért kíván a hagyományostól eltérő megközelítést, módszereket és eszközöket a Big Data feldolgozása?
4. Mik a Big Data feldolgozásának lehetséges előnyei?
5. Milyen típusú adatok lehetnek egy Big Data adathalmazban? Példákat is mondj rájuk!

Önellenőrző quiz: https://forms.gle/NdZ1boPQJcudr6yG9

Referenciák

[1] https://www.guru99.com/what-is-big-data.html#3

[2] https://github.com/AlessandroCorradini/University-of-California-San-Diego-Big-Data-Specialization

[3] https://gist.github.com/wagnerjgoncalves/35a51f7a8e9f87db929c6d789d1d97ed

[4] https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples/

[5] https://www.bbva.com/en/five-vs-big-data/