Big Data rendszerek

Mi a Big Data?

Nem létezik egzakt Big Data definíció.

A „big data” kifejezést a mai értelemben először *Cox* és *Ellsworth* (NASA) használta 1997-ben a szuper-számítógépes szimulációik során előállt extrém adatmennyiség feldolgozási kérdéseinek leírására.

„ Extrém nagyadathalmazok, amelyek számításigényes analizálása során mintázatokat, trendeket és összefüggé-seket lehet feltárni különösen az emberi viselkedés és interakciók terén.”

Oxford Dictionaries

Mi a Big Data?

„A big data olyan adat, ami meghaladja a hagyományos adatbázis rendszerek feldolgozási kapacitását. Az adat túl nagy, túl gyorsan mozog, vagy nem illeszkedik az adatbázis architektúra megkötéseihez. Ahhoz, hogy értéket nyerjünk ki ezekből az adatokból, alternatív feldolgozási módszert kell választani.”

E. Dumbill: „Making sense of big data”, Big Data, vol. 1, no. 1, 2013

„Big Data az, amikor az adat mérete maga is a probléma részévé válik.”

Mike Loukides, O’Reilly

A Big Data jellemzői

Gartner „3V”:

volume – terjedelem

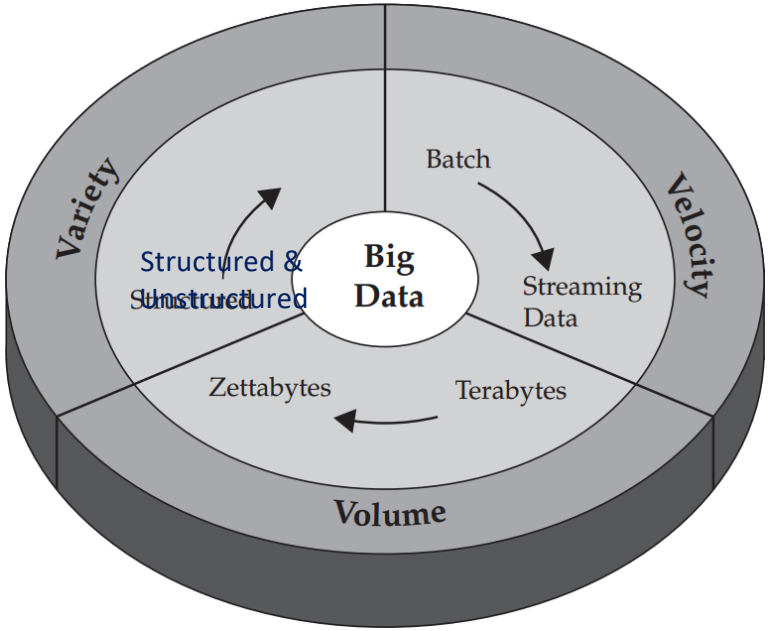
velocity – sebesség

variety – sokszínűség

IBM „4V”:

3V

veracity – valódiság



Forrás: *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill, 2011*

További fontos jellemzők: a változékonyság (variability), a megjelenítés (visualization), az értékes, felhasználható eredmény (value), az érvényesség (validity), az illékonyság, azaz az érvényesség hossza (volatility).

DeVan, 2016

Terjedelem (Volume)

Az előállt nyers adat mennyisége (akár exabájtos nagyságrend)

Az adat terjedelme nem csak tárolási kérdés (pl. milyen adat meddig kerüljön tárolásra, ritkán használt adatok periodikus törlése), jelentős feldolgozási problémákat is felvet.

Az IDC *„The Digital Universe in 2020”* tanulmánya szerint a digitális univerzum megközelítőleg 8 zettabájt méretű (2015) és a jelenlegi növekedési ráta alapján 2020-ra elérheti a 40 zettabájtos méretet is.

Facebook: exabájt kapacitású cold storage adatközpont átadása 2013. októberében

CERN Data Center: 220 PB (2013), 230 000 processzor mag, 15 000 szerver (2018)

Sebesség (Velocity)

Az adat előállásának és változásának gyorsasága, illetve az a sebesség, amivel az adatot fogadni, értelmezni és feldolgozni kell.

Jellemzően közel valósidejű vagy valósidejű/stream jellegű adatfogadásra és -feldolgozásra van szükség.

Fontos az adatok „időértéke” is:

Az adott pillanatban rendelkezésre álló adatok valós idejű elemzése lehetővé teszi jobb adatvezérelt döntések meghozását, például termékajánlás, biztonsági kockázatok elemzése stb. esetén.

Sebesség (Velocity)

Például:

A New York Stock Exchange naponta 50-60 milliárd tranzakció mellett 15 TB kereskedelmi információt gyűjt össze és tárol (csúcsértékek).

Az LHC (Large Hadron Collider, CERN) 2008-ban átadott részecskegyorsítója és ütköztetőgyűrűje, másodpercenként 1 PB adatot állít elő → ennek kb. 1%-a kerül tárolásra.

A SKA (Square Kilometre Array) Telescope az átadása után várhatóan napi 1 EB adatot fog előállítani.

Sokszínűség (Variety)

Az adatok forrásának, strukturáltságának és típusának változatossága.

Strukturált adat:

Jól definiált formában állnak rendelkezésre (pl. relációs vagy OO adatbázisokban).

A tárolt adatoknak csak kb. 15%-a strukturált.

Strukturálatlan adat:

Nem követ rögzített formátumot, sorrendet vagy egyéb szabályt.
Nem megjósolható.

Sokszínűség (Variety)

Strukturálatlan adat:

Pl.: emberek által generált tartalmak (dokumentumok, képek, hangok, videók)

Félig strukturált adat:

Heterogén forrásokból származó adatok integrációja és megosztása rendszerek között → egyre jelentősebb.

Absztrakt leírás: gráfokkal (csomópont - adatelem,
élek - relációk, címkék - attribútumok).

Pl.: XML állományok + dokumentumok

Valódiság (Veracity)

A rendszer által feldolgozott és tárolt adatok minősége, pontossága, és származásának megbízhatósága.

A 3V eredményeként nehéz az adatok valódiságát biztosítani.

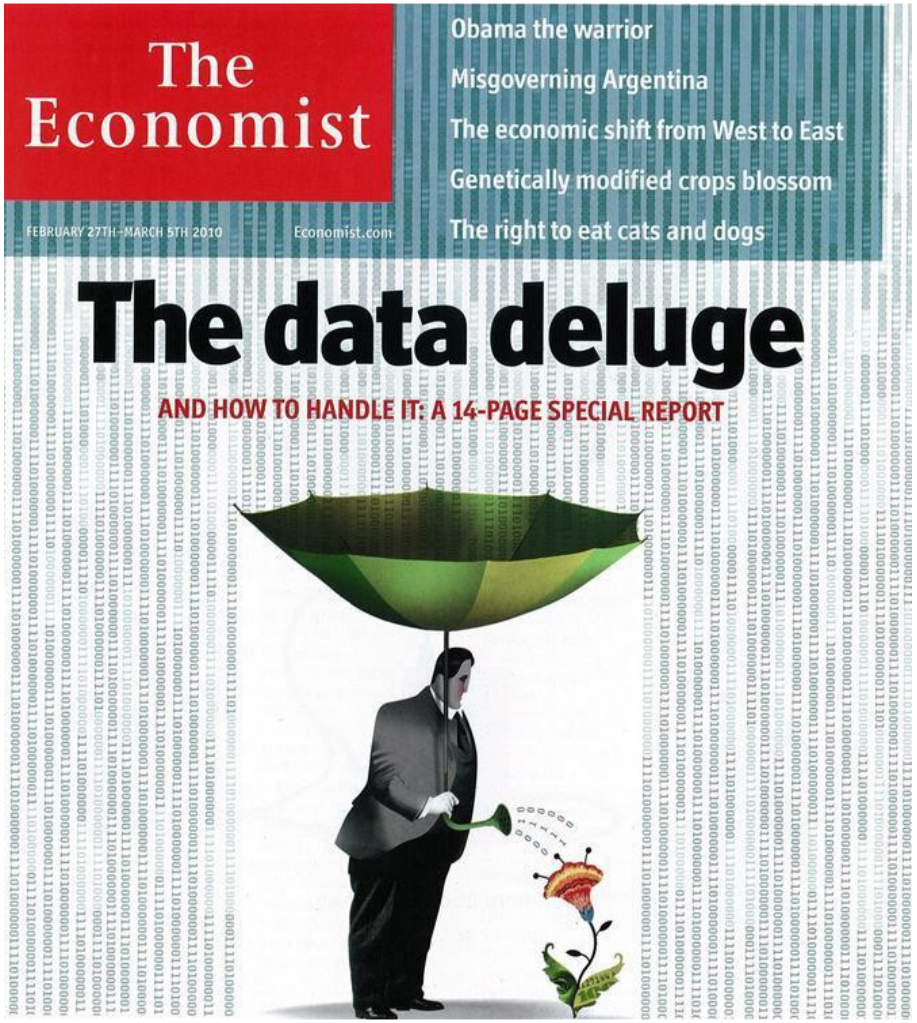
Jelentős kérdés az adatok valódiságának ellenőrzése.

Egy szakértői vélemény szerint* az USA gazdaságának évi 3,1 billió dolláros költséget jelent az adatok karbantartása és „tisztítása” (2011).

* Hollis Tibbetts - http://www.newswire.com/dirty-data-costs-the-us-economy/128732

Kezelendő problémaforrások: ellentmondásos vagy többértelmű adatok, duplikátumok, modell approximációk, feldolgozási késleltetésből eredő hibák, hamis adatok, spam stb.

A Big Data jelentősége

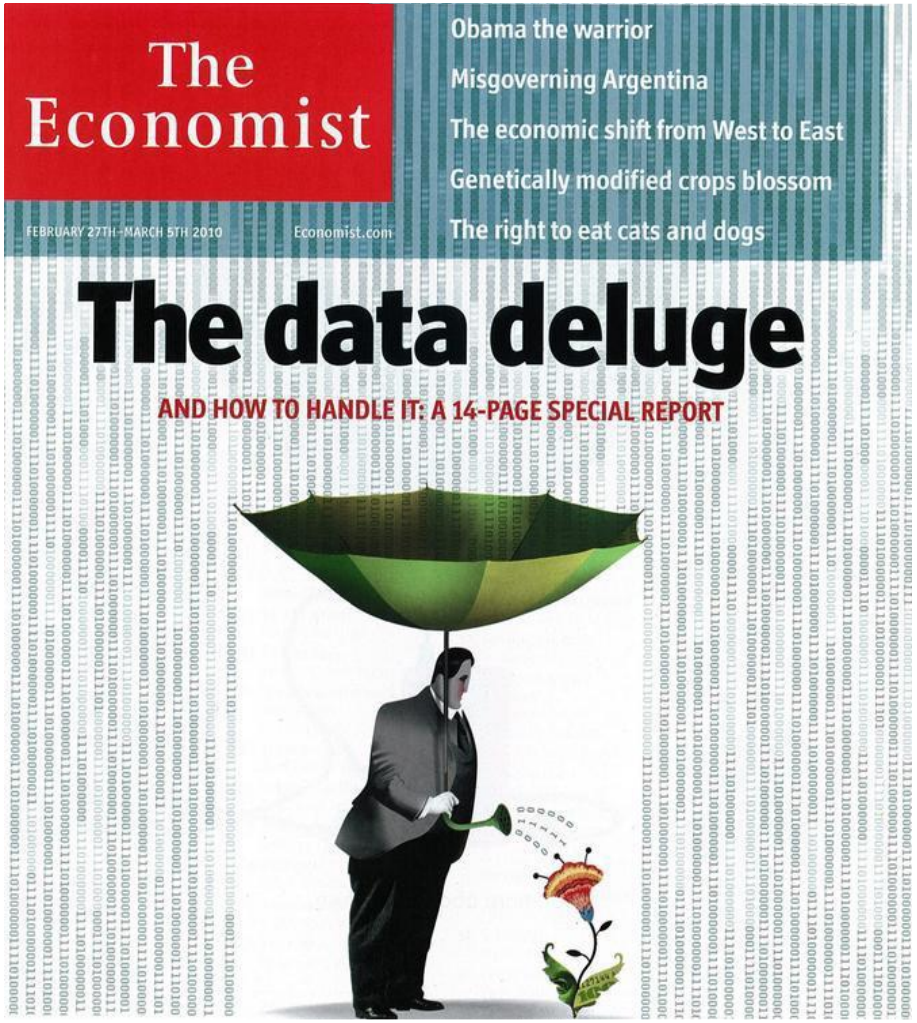


A számító- és tárolókapacitás árának csökkenésével egyre nagyobb mennyiségű adat gyűjtése és tárolása válik lehetővé.

A hálózatba kapcsolt eszközök számának jelentős növekedésével az adatok összegyűjtése egyre könnyebbé válik.

„Az adatáradat” –
The Economist, 2010. február 25.

A Big Data jelentősége



Leggyorsabban növekedő adatforrások:

Tranzakciós és rendszernaplók

Különböző eszközök szenzorai

Social media: mindenki számára lehetővé vált tartalmak (szöveg, hang, kép, videó, GPS koordináták) könnyű és gyors megosztása.

Cél: érték kinyerése a rendelkezésre álló nagy mennyiségű adatból.

A Big Data néhány alkalmazási területe

Tudományos kutatás

Az összegyűjthető és előállítható adatok mennyiségének robbanásszerű növekedése számos tudományterületen tette szükségessé a Big Data eszközök alkalmazását.

A kollaboráció és az interdiszciplinaritás egyre inkább előtérbe kerül → különböző adatok összekapcsolása.

Konkrét hipotézisek ellenőrzése mellett új tudást is ki lehet nyerni a rendelkezésre álló adatokból → hipotézismentes adatvezérelt kutatás.

Tudományos kutatás – példák

Élettudományok: A biotechnológia fejlődésével és a szekvenálás költségének csökkenésével egyre több adat áll rendelkezésre, amit a hagyományos módszerekkel már nem lehet vizsgálni.

Alkalmazási területek: genetika, metagenomika, proteomika, evolúciós biológia, farmakológia stb.

Csillagászat: Square Kilometer Array (SKA) átadása után petabájtos nagyságrendű adatok feldolgozására lesz szükség.

Ennek célja: az általános relativitáselmélet tesztelése, kozmológiai kutatás, sötét energia vizsgálata, galaxisok létrejöttének vizsgálata stb.

Egészségügy és orvostudomány

Elektronikus beteg rekord (EPR): egy beteg kórtörténete

Diagnózisok, alkalmazott gyógyszerek, terápiák leírása és eredményei, védőoltások dátuma, allergiák és érzékenységek, radiológiai felvételek, laboratóriumi és egyéb tesztek eredményei stb.

Az egészségügyi folyamatokra jellemző, hogy

általában több szervezeti egység vesz részt bennük, amelyek saját IT infrastruktúrával rendelkeznek

→ heterogén adatforrásokból kell jellemzően félig strukturált adatokat kinyerni és feldolgozni.

Egészségügy és orvostudomány

Ezekből adatbányászati eszközökkel lehet új ismeretet származtatni. Például:

- betegségek közötti összefüggések, krónikus betegségek kialakulásának okai, fertőzések terjedésének megfigyelése és előrejelzése, régiók jellemző megbetegedései (környezeti hatások feltárása) stb.

Prediktív modellek létrehozásával lehetségessé válhat valós idejű döntéstámogatást nyújtani különböző egészségügyi osztályozási problémák során, diagnózis felállításában, az egészségügyi szolgáltatás minőségének javításában és annak költségeinek csökkentésében is.

Biztonság és felderítés – adatforrások

A biztonság és felderítés területein: számos különböző forrásból származó, esetenként torzított adatok együttes elemzése

Ezeken a területeken a Big Data elsősorban gépi tanulási technológiában van jelen, az emberi munka segítésére.

Nyilvános adatok (weboldalak, blogok, tweet-ek, egyéb online elérhető adatok, nyomtatott és elektronikus média)

Kommunikációs csatornák (hang, szöveges üzenetek, e-mail, chat stb.)

Műhold-, légi felvétel, radarkövetési információk, GPS adatok

Szenzor adatok (pl. biztonsági kamerák, megfigyelőrendszerek)

Biztonság és felderítés – adatforrások

Biometriai adatok (ujjlenyomat, DNS, írisz- és arcképek, testtartás, stb.)

Strukturált és félig strukturált adatok, amelyeket cégek és szervezetek biztosítanak (repülési adatok, bankkártya adatok és banki tranzakciók naplói, telefonhívások, személyi akták, EPR, rendőrségi akták, nyomozati anyagok stb.)

A cél: jelentéssel bíró mintázatok és trendek felderítése. Ehhez olyan adatbányászati eszközökre van szükség, amelyek képesek nagyszámú fals pozitív találat nélkül ilyen mintázatok detektálására nagyon nagy mennyiségű adatban, akár közel valós időben is.

Dr. Hajdu András, Tóth János: Nagy adathalmazok elosztott feldolgozása című tananyaga alapján készült