Big Data Típushibák és megoldások

Kazi Sándor
sandor.kazi@webvalto.hu

Big Data

Legyen kézzelfoghatóbb, legyenek dimenziói:

„Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." / D. Laney (Gartner) /

- **Volume** – mennyiség
- **Velocity** – adatsebesség
- **Variety** – sokféle típusú adat
- …: Veracity – igazságtartalom, tisztaság; Validity – helyesség; Variability – egyre rugalmasabb struktúrák; Value – nagy értékű; Visualization – vizualizálhatóság; …

Big Data - Minek?!

Motivációs Marshall-kereszt

- Data Science → kidobni drága (az új kor “olaja”)
- Olcsó tárolás → megtartani olcsó

Elosztott környezetek felé?

- CPU gyártás korlátai
- Diszk sebessége
- SPOF

Új eszközök: Hadoop, streaming és MQ/MB, NoSQL, NewSQL

Hardver / Szoftver

Hardver

- Elosztott környezet
  - Klaszter szemlélet
  - Középkategóriás hardver

Szoftver

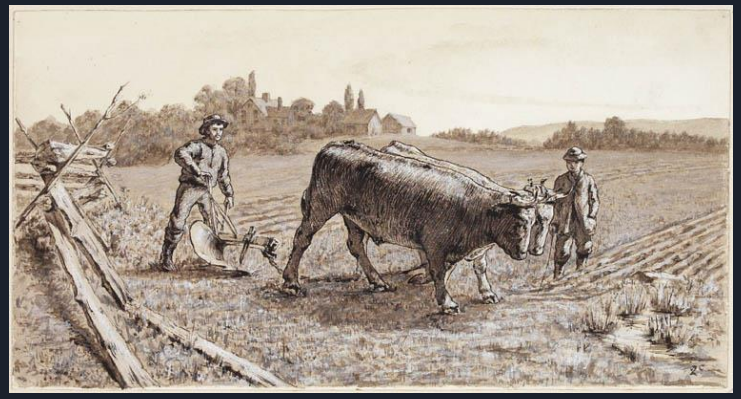
- Sok open-source eszköz
- Licensz-költségek helyett… ???
- Java dominancia, de csökkenőben
- Infrastructure as code

Mit ne? (helyette?)

Elosztottság

“In pioneer days they used oxen for heavy pulling, and when one ox couldn’t budge a log, they didn’t try to grow a larger ox. We shouldn’t be trying for bigger computers, but for more systems of computers.” /Grace Hopper/

- Örökké jó-e a scale-out felfogás?



Hardver / Szoftver

- Átültetett megoldások
  - Interfész-intuíció (SQL?)
- Környezetek:
  - DEV? TEST/UAT?
- Virtuális gépek? NAS?
- Üzemeltetés, konfigurációk, telepítés
  - Verziófrissítés
  - Infrastructure-as-code
- Emberi tényező: beletanulás?
- Proof-of-concept paradoxon
  - “Kis big data projekt”

Projekt oldalon

Mérhetőség és összehasonlíthatóság

- Mikor sikeres? Mit tudunk?
- Felhasználói elégedettség?

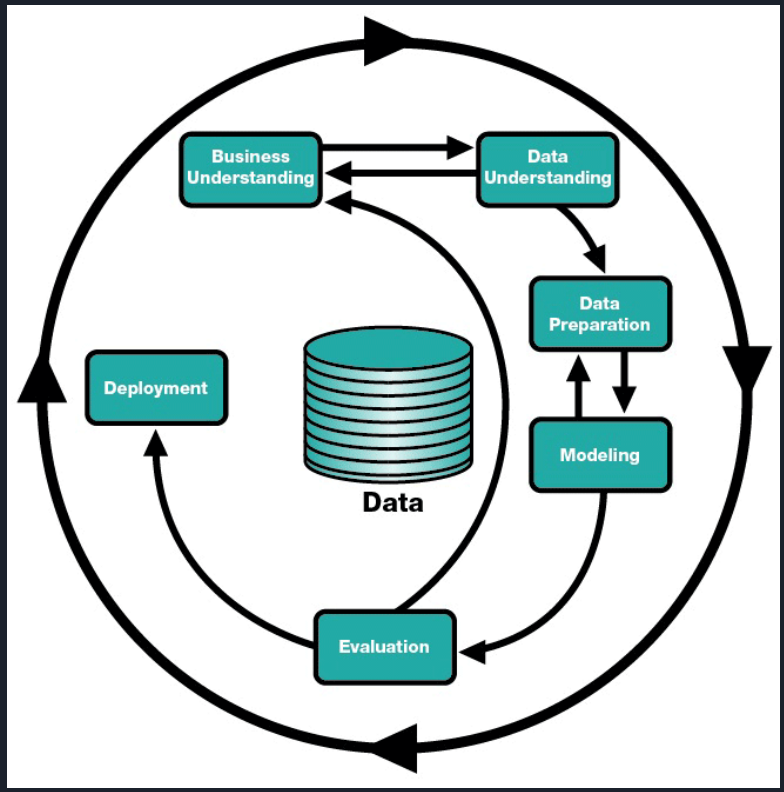
Szponzor és ráfordítás, visszajelzések

Szerepkörök tisztázása

Lefutás hossza (lásd: hitel)

Governance…

Sosincs “kész” (módszertan hiánya)



Adat oldalon

Minden adat “megvan”, ami kell?

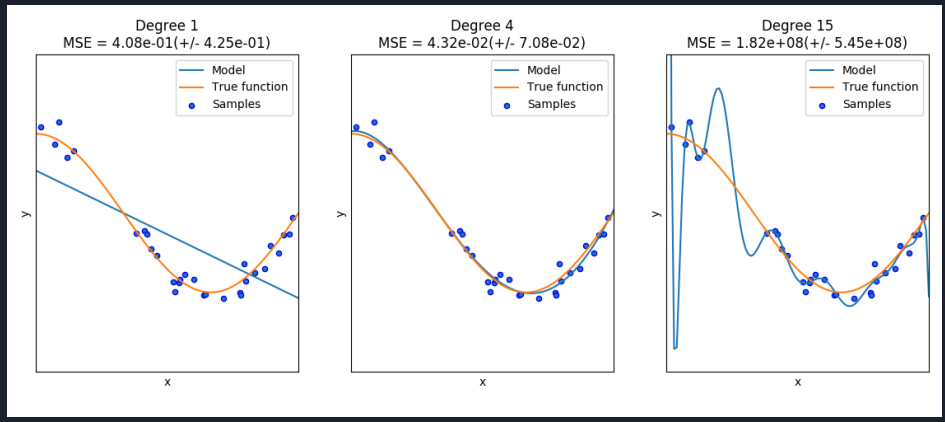
- Eseményhorizont?
- Hiányzó adatok by definition (pl.: logok)
- Adathibák
  - Hiányzó adat
  - Téves adat
  - Kiugró adat
- Hiányzó metaadatok
  - Adat, processz, történelem, stb. …
- Adatgyűjtés
  - A/B teszt?

Analitikai ugrás

1	Standard riportok	Mennyi … értéke az előző …?
2	Ad-hoc riportok	Mennyi … értéke, ha …?
3	OLAP	Miért ennyi?
4	Alertek	Milyen értéknél van gond?
5	Statisztikai elemzések	Mi történt? Mitől függ …?
6	Előrejelzés	Mi lesz ha folytatódik ez a trend?
7	Prediktív modellezés	Mennyi lesz … érték, ha …?
8	Optimalizáció	Hogyan csináljuk jobban?

Ez gyakorlatilag az adat megértésének mélységével esik egybe, ezért szinteket átugrani nem gyakran sikerül.

Gépi tanulási oldalon



“Eseményhorizont”, kauzalitás

Korreláció vs ok-okozat

“Szűklátókörűség” - túltanulás

“Örökké tartó igazság” - modell elévülése

Rossz mérőszám

“Sziszifuszi munka” - akadályozott konvergencia

Eltitkolt tudás

Köszönöm a figyelmet!