Adatelemzési platformok, Április 24., 8. elméleti óra

Big data

Saját definíciók

- Sok adat, sok dimenzióban.
- Elosztott párhuzamos rendszerek, architektúrák
- Nagy (hyped) sebesség, sokaság, sokszínűség

Hivatalosabb definíciók:

1. Hangsúlyosabban része a definíciónak az adat 'túl nagy' mennyisége
    1. Excelben: 1 millió sor
    2. Memóriában: kiinduló adathalmaz 2-6 GB körül mozog
    3. Adatbázis: 200-600 GB adat körül
    4. Disk: 1TB (SSD sebesség: 500 MB/sec, az egész adat sima beolvasása kb fél óra)
    5. Hálózat/felhő: itt nehezen értelmezhető, ott inkább a net a keresztmetszet
2. Volume, Velocity, Variety
3. Big data eredetmonda:
    1. Business Intelligence, Data mining
    2. Nagy piaci koncentráció: Oracle, IBM, SAS, Microsoft, SAP
    3. 2007: Google publikálja a kivezetett MapReduce technikát
    4. Ebből fejlődött ki a Hadoop (open source, JAVA), amit ingyen tudtak használni olyanok, akiknek sok adatuk volt de nem akartak költeni nagy gépekre. "Big Data" kvázi szinonímája volt a Hadoop használatnak.
    5. ~2014: Újságírók
    6. "Big data" innentől már nem csak a Hadoopot jelenti.

Szembe a régebbi nagy hardver/szoftver/architektúra konfigokkal, a big data egyik igérete, hogy egyszerű, lehasznált pc-kkel is meg lehet oldani nagy adatos problémákat.

Ktgelszoszlás:

- hw: 10%
- sw: 10%
- hr: ?

Nagy költségigény, leginkább akkor vezetik be, amikor már nem lehet dolgozni a meglévő eszközökkel.

Romboló innováció

- Idő/komplexitás függvényében
  - high-end igény:
    - legnagyobb szereplők igényeit szolgálják ki
    - túl sokat tudnak adni a megoldások, én egyre drágábbá válnak
  - low-end igény:
    - egyszerű probléma és a rájuk adott megoldások
    - Itt is egyre jobbak lesznek a megoldások

'Rombolás': amikor a low-end megoldások elérik a high-end igényeket. Jelenleg a Big Data megoldások ilyen szinten vannak, de ezek open source megoldások és az esetek nagy részében nem működnek. Ezért is igen nagy a HR igénye a Big Data-nak.

Működő open source megoldások nagy részét alapvetően a nagy tech cégek finanszírozzák, és a technológia függőség miatt ez problémás is lehet (pl airbnb- kiugrik egy használt technológiából).

MapReduce

1. Virtuális input fájl

9 betű:

ABR
CCR
ACB

HDFS: Hadoop Distributed File System

2. Automatikusan szétsplitteli a fájlt:

- ABR
- CCR
- ACB

Split: blokkokra bontja az adatot

3. Három példányban tárolja a blokkokat, lehetőleg három különböző fizikai eszközön

4. 'Map' fázis: ezt nekünk kell megírni, kulcs-érték párok listáját produkálja

- ABR:
  - A, 1
  - B, 1
  - R, 1
- CCR:
  - C, 1
  - C, 1 # Fontos, hogy nem C, 2 kombináció, mivel az lassabb.
  - R, 1
- ACB:
  - A, 1
  - C, 1
  - B, 1

Ezeket a kereséseket elosztottan tudja csinálni.

5. Shuffle and Sort: Kulcsonként összegyűjti a dolgokat

- I.
  - A, 1
  - A, 1
- II.
  - B, 1
  - B, 1
- ...

A szándék, hogy minél kevesebb dolgot átküldeni a neten

6. Reduce: megnézni, hogy milyen hosszúak a listák

- A, 2
- B, 2
- C, 3
- R, 2

Az igéret, hogy ha van elég vas alatta, bármekkora is a fájl, akkor folyamatosan ki tudja számolni. Ha esetleg kiesik egy gép, tudja folytatni a számítást. Itt nincs 'körkörös' várakozás, csak a fenti pontok függvényében van függőség. Az adatok fizikailag nem mozognak, maga a kód viszonylag kicsi és

ezért inkább azt küldik körbe. Ráadásul viszonylag fix ideig tart egy lekérdezés, sőt néha a heurisztikák miatt több adattal a lekérdezés kevesebb ideig tart.