Big Data az adattárházban

A párbaj folytatódik?

Néhány fontos Big Data projekt

Cég	Téma	Adat	Új-fajta	Mennyi-ség	Saját adat?	Típus	Megjegyzés
Google	Influenza előjelzés	Google	I	„big”	I	Előjelzés	
Farecast	Repjegy vásárlás optimális ideje	Internetről összegyűjtött	I	„big”	N	Előjelzés	
Xoom	Nemzetközi pénzátutalások	Saját tranzakciók (régiek is!)	N	„big”	I	Csalás (pattern)	DW-vel is elő kellett volna jönni
Chicagoi egyetem	Szumóbajnokság	Jegyzőkönyvek	N	64 ezer	N	Csalás	Hagyományos data mining ?
Barabási Albert-L.	Hálózati kutatás	Európai ország mobil hívásai egy szolgáltatónál	I	„big”	N	Pattern	
Price Stats	Fogyasztói árindex	Internetről összegyűjtött	I	„big”	N	Előjelzés	
Amazon	Vásárlói ajánlás	Internetes böngészés	I	„big”	I	Ajánlás	
Walmart	Vásárlói szokások	Tranzakciók (régebbiek is!)	N	„big”	I	Pattern	
Aviva	Hitelképességi vizsgálat	Egészségügyi és életmód adatok ???	I-N	„big”	N	Előjelzés	
Obama Big Data	Választás	Kérdőívek	?	mintavétel	I	Előjelzés	

Big Data adatok

- (Általában) nagy mennyiségű,
- Sokféle,
- Sok esetben **újfajta,**
- Nagy sebességgel keletkező,
- Strukturálatlan,
- „Kuszább” adatok, ahol
- Nem feltétlenül a cég az adatok tulajdonosa

Újfajta adatok

- Web server logok
- Internet clickstream adatok
- Böngésző kifejezések
- „Social media” tartalom
- „Social network” tevékenységek
- Email szövegek
- Egyéb interneten elérhető adat
- Felmérések eredményei
- Mobil eszközök hívásai
- Szenzorok által szolgáltatott adatok (Internet of Things)

„Kusza” Big Data adatok

Adatok „Boyle - Marriotte törvénye” DW és Big Data rendszerek esetén:

Adattisztaság * Adatmennyiség = Állandó

Elemzések fajtái

- Általában adatbányászat
  - kapcsolatok: minták, amikor egy esemény egy másikkal van kapcsolatban
  - amikor egy esemény egy későbbi másik eseményhez vezet
  - osztályozás, új minták keresése
  - klaszterezés, eddig nem ismert tények csoportjának felismerése
  - előrejelzések: olyan minták felfedezése, amelyek nagy valószínűséggel bekövetkező eseményeket jeleznek
- Szövegelemzés,
- Egyéb statisztikai elemzések

- **Kauzalitás vs Korreláció**

Újfajta eszközök

- Adatok tárolására, feldolgozására, elemzésére
- Hadoop/MapReduce
- NoSQL adatbázisok
- Elemzésre: R, R Enterprise
  Adatbázis szinten fejlesztett SQL, pl. Match Recognize, Big Data SQL

Felmerülő kérdések

- Van már DW a cégnél?
  (WALMART, AMAZON, XOOM)
- Mi a Big Data Projekt és a DW kapcsolata?
- Ha önálló projekt, akkor is sok adatot kell „hozzátölteni”, pl:Google influenza projekt:
  - Influenza statisztika időegységenként és földrajzi egységenként
  - IP cím - földrajzi egység összefüggése (Geolocation)

DW vs Big Data: a párbaj folytatódik

DW és Big Data: Inmon vs Kimball

- Inmon: Adattárháznak és Big Data-nak nincs semmi köze!

  DW gondosan megtervezett, tiszta, összefüggésében ellenőrzött adatokat tartalmaz.

- Kimball: Adattárházban helye van a Big Data-nak
  - Staging area szintjén kapcsolódnak be ezek az adatok
  - Hadoop és NoSQL rendszerek
  - Kapcsolatot a közös dimenziókon keresztül építhetünk (Termék, Ügyfél, Földrajzi egységek, Idő..)
  - CRM fontos adatai keletkeznek

DW és Big Data: Oracle

- Staging area továbbfejlesztve: Hadoop központi tára az összes input adatnak (reservoir), itt lehet különböző modelleket felállítani és innen kerülhetnek továbbtöltésre adatok

Data reservoir + Big Data SQL **Data warehouse**

Cloudera Hadoop
Adatbázis
Oracle NoSQL
Oracle R Advanced Analytics for Hadoop

Big Data Connectors

Data Integrator

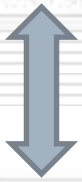
Big Data és DW összekapcsolása

Strukturálatlan clickstream adat

Big Data és DW összekapcsolása

Clickstream adatok: View/external tábla

	ts	ip	url	swid
0	2012-03-15 01:17:06	99.122.210.248	http://www.acme.com/SH55126545/VD55170364	{7AAB8415-E803-3C
1	2012-03-15 01:34:46	69.76.12.213	http://www.acme.com/SH55126545/VD55177927	{8D0E437E-9249-4D
2	2012-03-15 17:23:53	67.240.15.94	http://www.acme.com/SH55126545/VD55166807	{E3FEBA62-CABA-1
3	2012-03-15 17:05:00	67.240.15.94	http://www.acme.com/SH55126545/VD55149415	{E3FEBA62-CABA-1



category	url
books	http://www.acme.com/
movies	http://www.acme.com/SH55126545/VD55149415
games	http://www.acme.com/SH55126545/VD55163347
electronics	http://www.acme.com/SH55126545/VD55165149

DW tábla

Néhány összefoglaló gondolat

- Újfajta gondolkodás, szerepkör: „adattudós”
- Újfajta adatok bevonása
- DW-ben helye van az ilyen jellegű adatoknak
- Közös (Conformed) dimenziók használata, összekapcsolás ezen adatokkal
- DW egyéb adataihoz kapcsolás (pl. CRM)
- Tudni kell, hogy honnan származnak ezek az új adatok! (Válasz Inmon-nak)
- Strukturált DW adaton történő Big Data jellegű elemzést hova soroljuk?
- Meglevő adattárházat használjuk ki jobban!
- Szükség van hagyományos DW-re? Igen!