Virtuális Obszervatórium

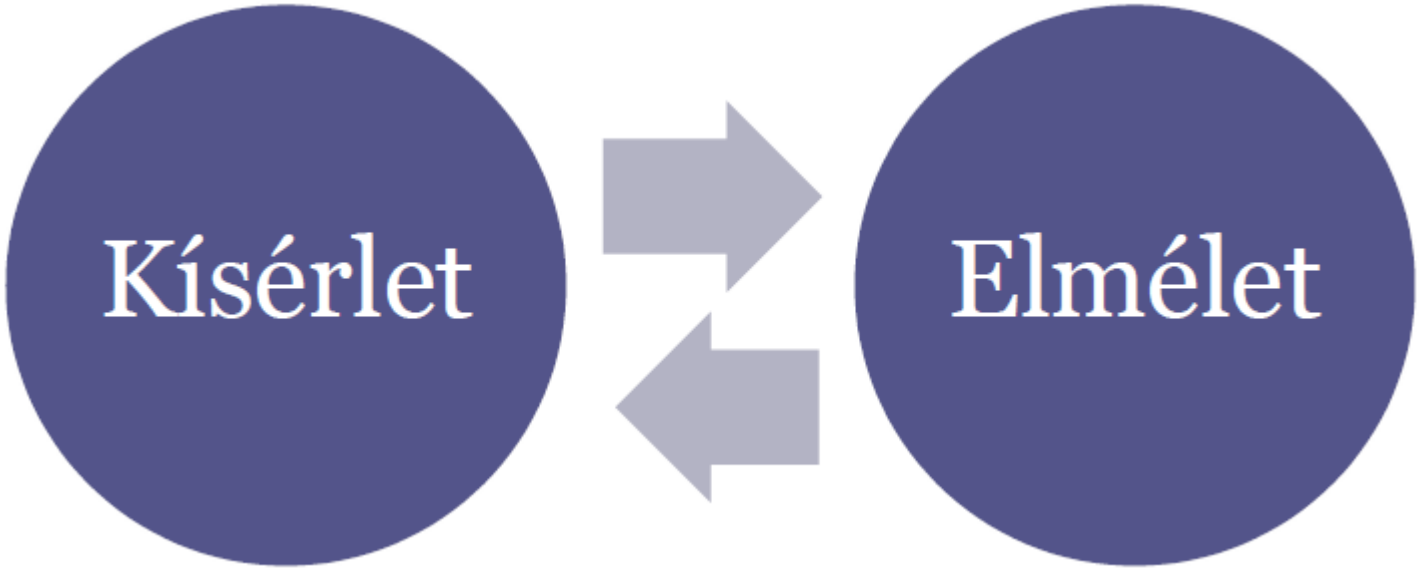
Gombos Gergő

Áttekintés

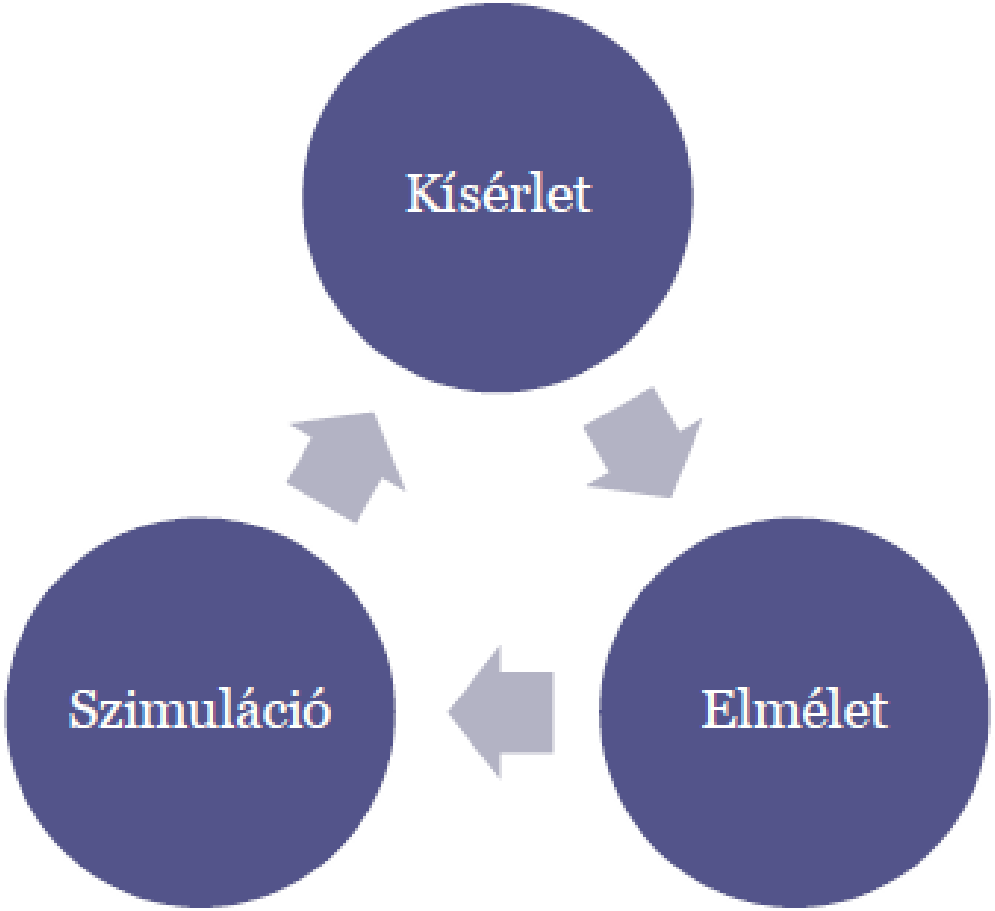
- Motiváció, probléma felvetés
- Megoldások
- Virtuális obszervatóriumok
- NMVO
- Twitter VO

Motiváció

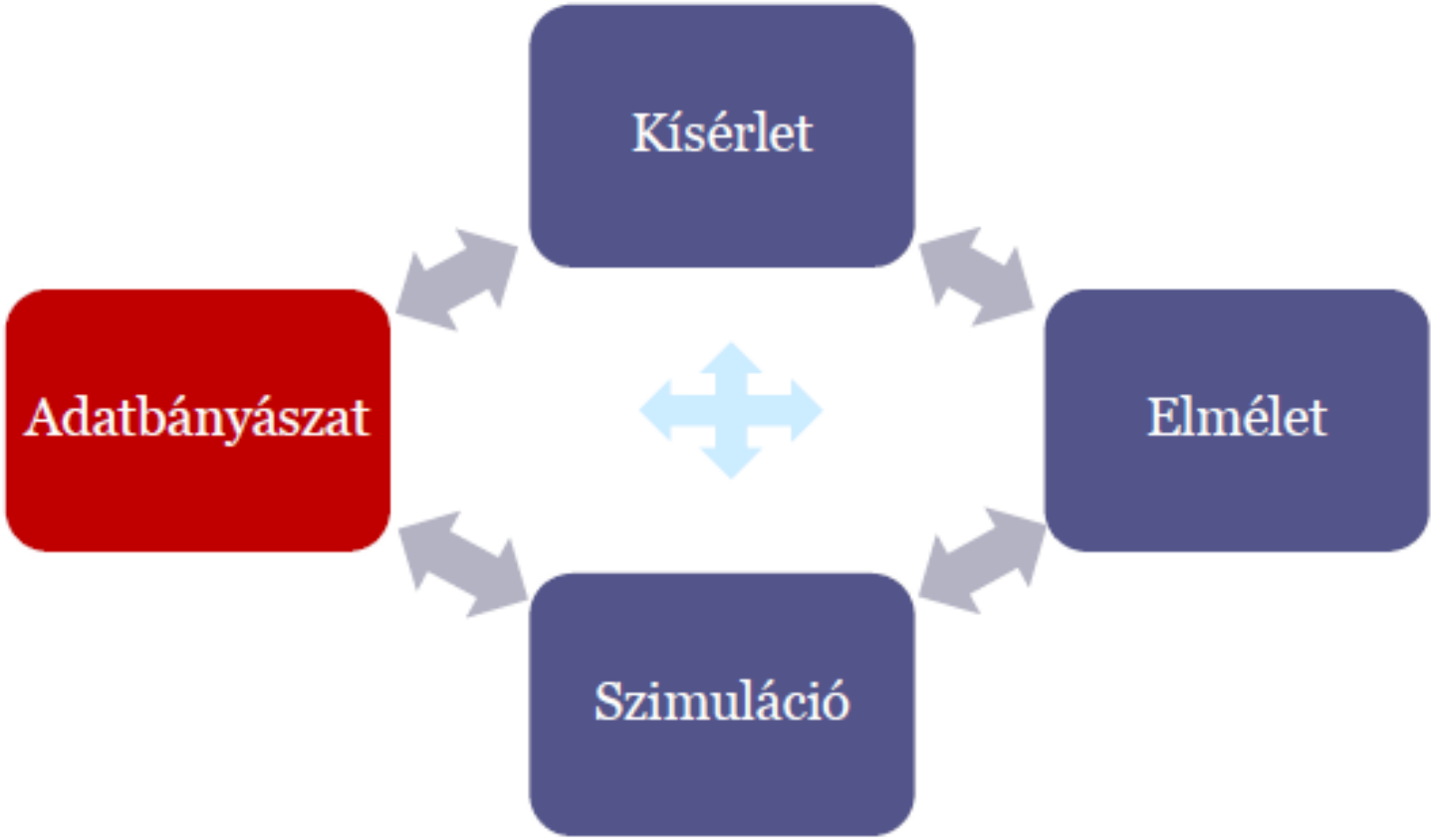
- Tudományos módszer fejlődése



Motiváció



Motiváció



Probléma

- Kicsiben működik, nagyban nehézkes
- Nagy mennyiségű mérési adatok
  - Távcsövek
  - Részecskegyorsítók
  - Szenzor hálózatok
- Méretek
  - ~PB méretű

Probléma

- Adatok elérése
  - Tárolás lemezen (lassú)
  - Felhasználói interfész

- Hogyan tudjuk elérni hatékonyan?

Feladat

- Olyan rendszert építsünk, amely
  - Képes nagy mennyiségű adatok tárolására, elemzésére.
  - Lehetőséget biztosít a felhasználóknak saját elemzések elvégzésére.

Egyszerű megoldás

- DB Kliens --- DB szerver
- Relációs adatbázis, SQL
- Felhasználók hozzáférnek
- Szinkron megoldás
- Probléma:
  - Ha olyan lekérdezés amelyre nincs megfelelő index, hosszú idő a válasz. Kiéhezteti a többi klienst.

Egyszerű megoldás

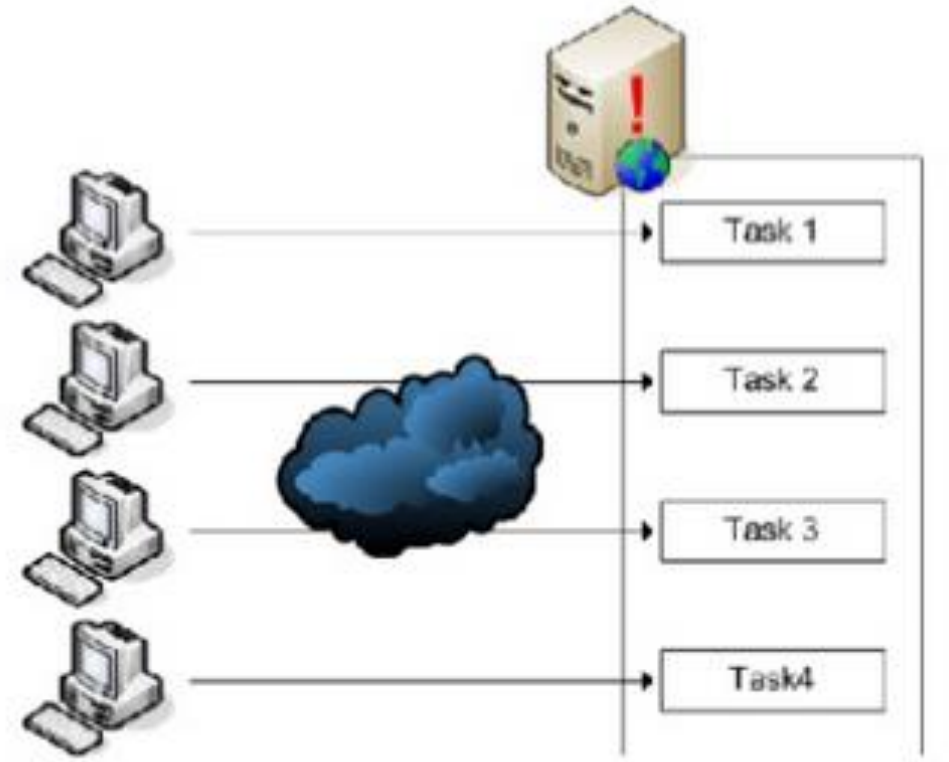
- Alkalmazás Kliens --- DB szerver
- Szinkron működés
- Alkalmazás szintjén egyszerűsödik
- Probléma:
  - Válaszidő még mindig lassú, timeout.
  - Klienst le kell tölteni. Nem érhető el mindig.

Egyszerű megoldás

- Kliens --- Alkalmazás szerver --- DB szerver
- Csökkenti a terhelést a DB-n
- Szinkron megoldás
- Probléma:
  - Lassú válasz, timeout.

Szinkron működés

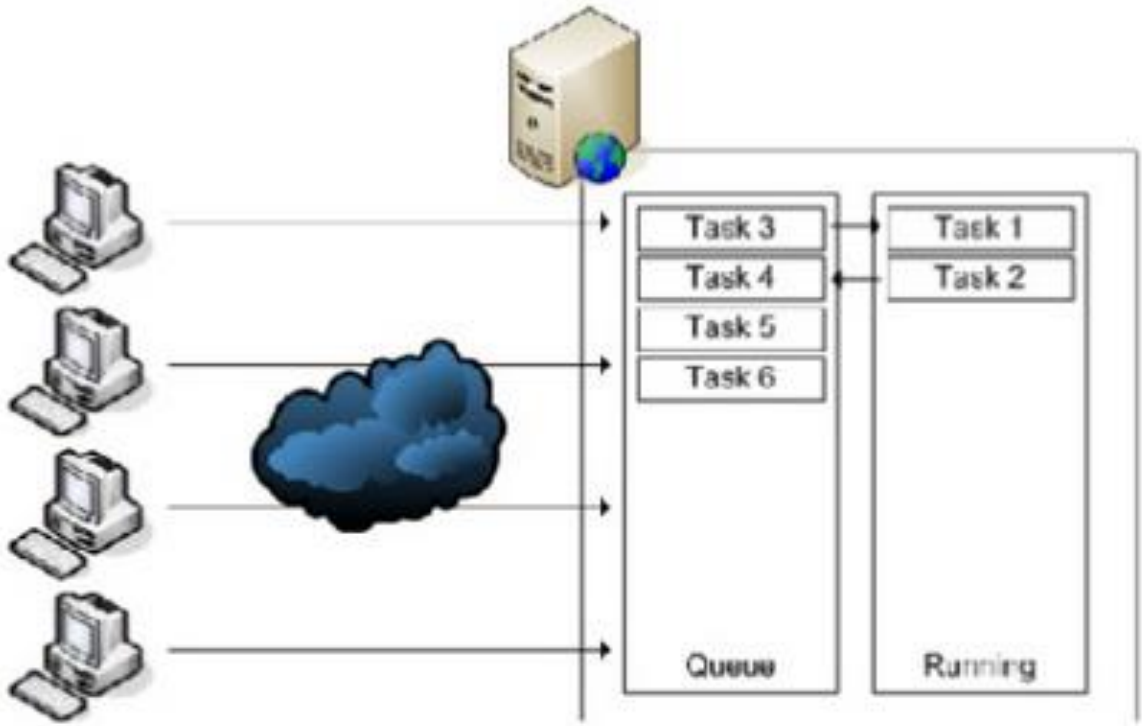
- Szerver terhelés nagy



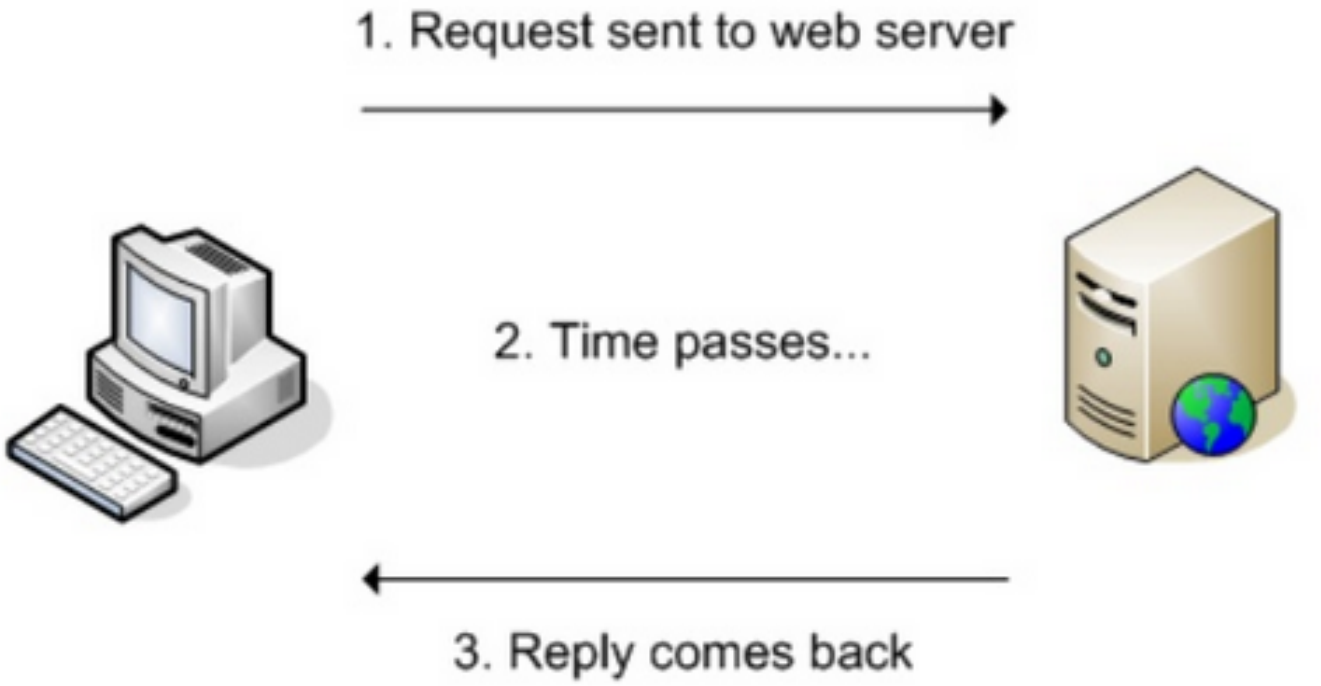
Egyszerű javított megoldás

- Aszinkron
- Ticket-rendszer
- Eredmény később
- Sorba állítja a kéréseket.

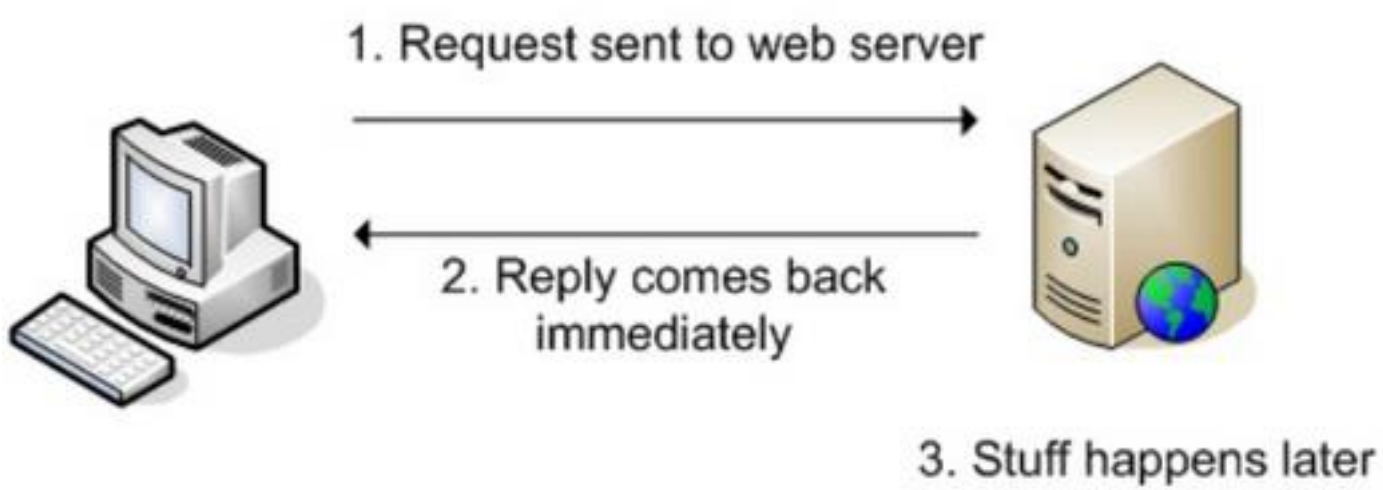
Egyszerű javított megoldás



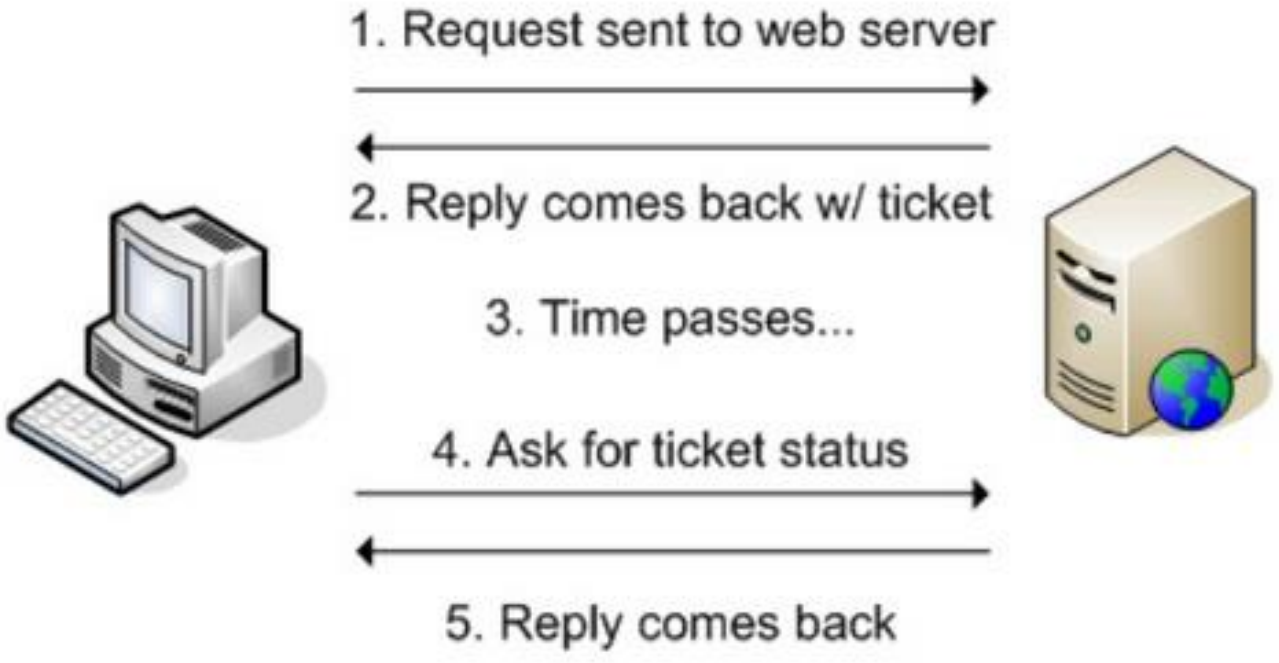
Szinkron megoldás



Aszinkron megoldás



Aszinkron megoldás



Mi a VO?

Olyan rendszer, ahol a digitális gyűjtött mérési adatokat tároljuk, elemezzük. A rendszer fő szempontja nem a válaszidő, hanem a bonyolult elemzések elvégzésének lehetősége.

VO célok

- Nagy adattömegek kezelése
- Nagy számításigény kielégítése
- Hatékony keresés, elemzés
- Kollaboráció kutatókkal
- Eredmények megosztása

Technológiák a VO-hoz

- Hardver
  - Tár- és számítási kapacitás, hálózat
- Adatbázis-technológiák
  - Adatmodellek, adatbázis-tervezés
  - Indexelés hatékony kereséshez
  - Adatelemzés, adatbányászat
- Párhuzamos, elosztott rendszerek
  - Párhuzamos feldolgozás
  - Grid technológiák
  - MapReduce technika
- Felhasználói felület, vizualizáció
  - Webes portálfelület
  - Vizualizációs technikák

VO feladatok

- Adat regisztráció
  - Metadata alapú adatforrás rögzítés
- Adat elérés
  - Regisztrációval vagy a nélkül
- Adat összekapcsolás
  - Különböző adatbázisok összekapcsolása
- Adat manipulálás
  - Adatbányászat, adat elemzés

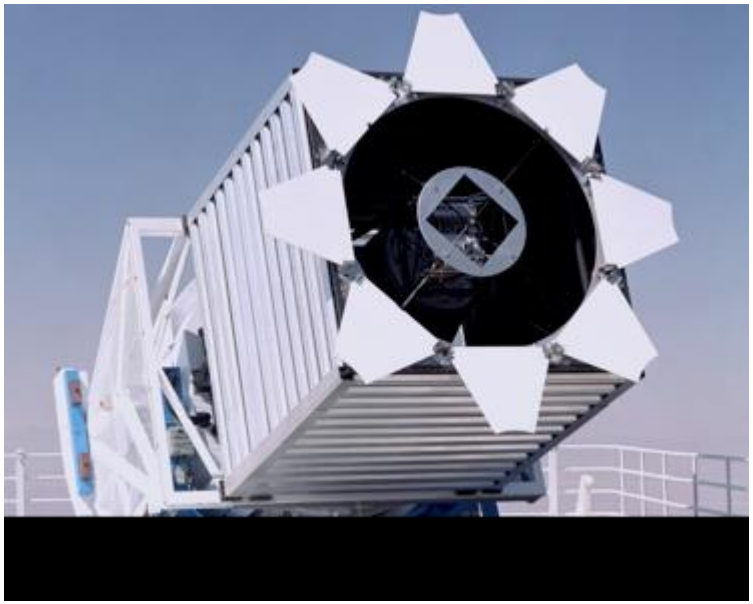
Miért Obszervatórium?

Galileo Galilei

Edwin Hubble

Miért Obszervatórium?

- SDSS
  (Sloan Digital Sky Survey)
- 2.5 m teleszkóp
- >100 TB

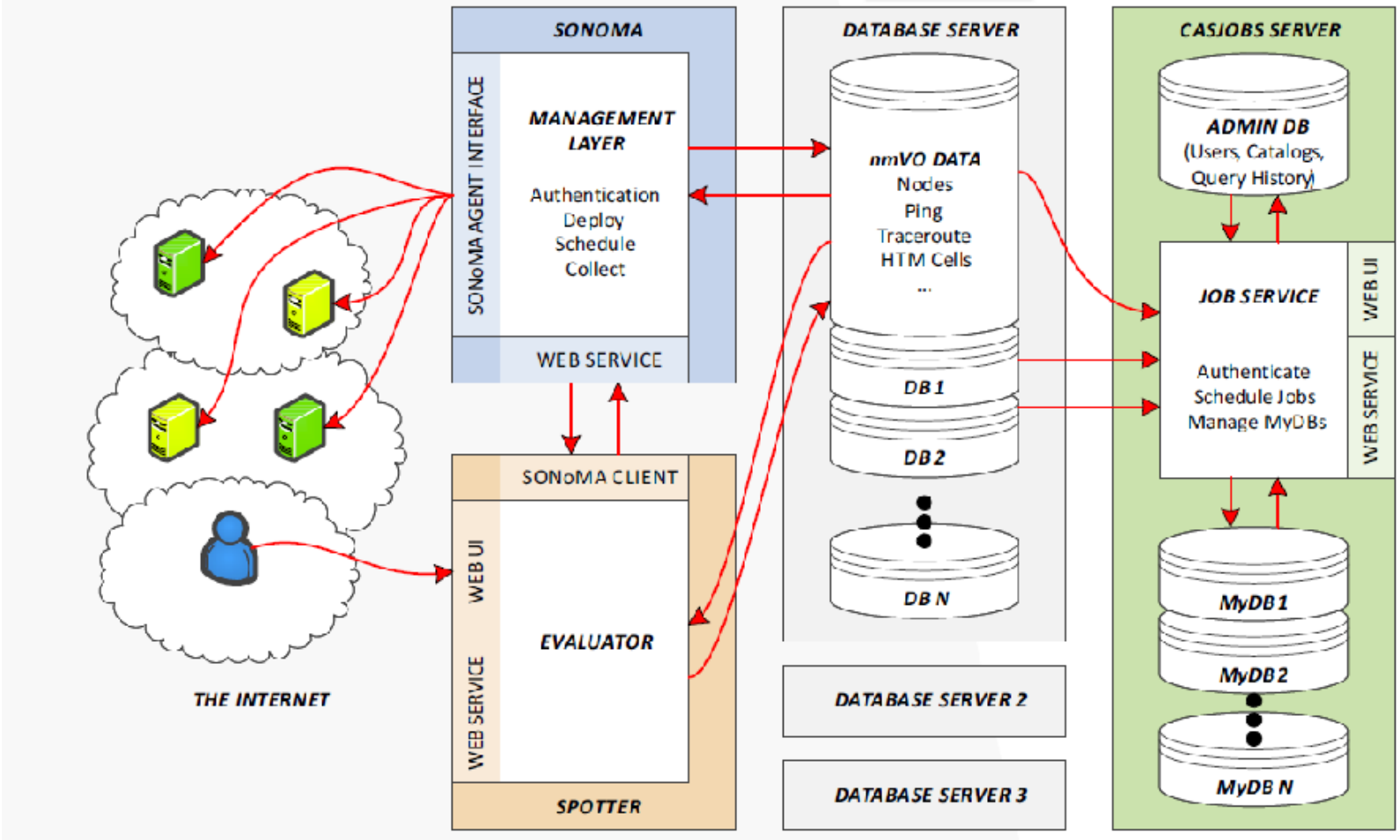


Teleszkóp -> Digitális adatok
Detektorok -> Számítógépes programok

VO-k

- SkyServer
  - Csillagászati adatok
  - http://skyserver.sdss.org
- NMVO
  - Főleg hálózati adatok, de van twitter, csillagászat
  - http://nm.vo.elte.hu/casjobs/casjobs.aspx
- (Twitter Casjobs)
  - Twitter adatok
  - http://oktnb16.inf.elte.hu/casjobs

NMVO


SONOMA
SONoMA AGENT INTERFACE
MANAGEMENT LAYER
Authentication
Deploy
Schedule
Collect
WEB SERVICE
SONoMA CLIENT
WEB UI
EVALUATOR
WEB SERVICE
SPOTTER
THE INTERNET
DATABASE SERVER
nmVO DATA
Nodes
Ping
Traceroute
HTM Cells
...
DB 1
DB 2
DB N
DATABASE SERVER 2
DATABASE SERVER 3
CASJOBS SERVER
ADMIN DB
(Users, Catalogs, Query History)
JOB SERVICE
Authenticate
Schedule Jobs
Manage MyDBs
WEB UI
WEB SERVICE
MyDB 1
MyDB 2
MyDB N

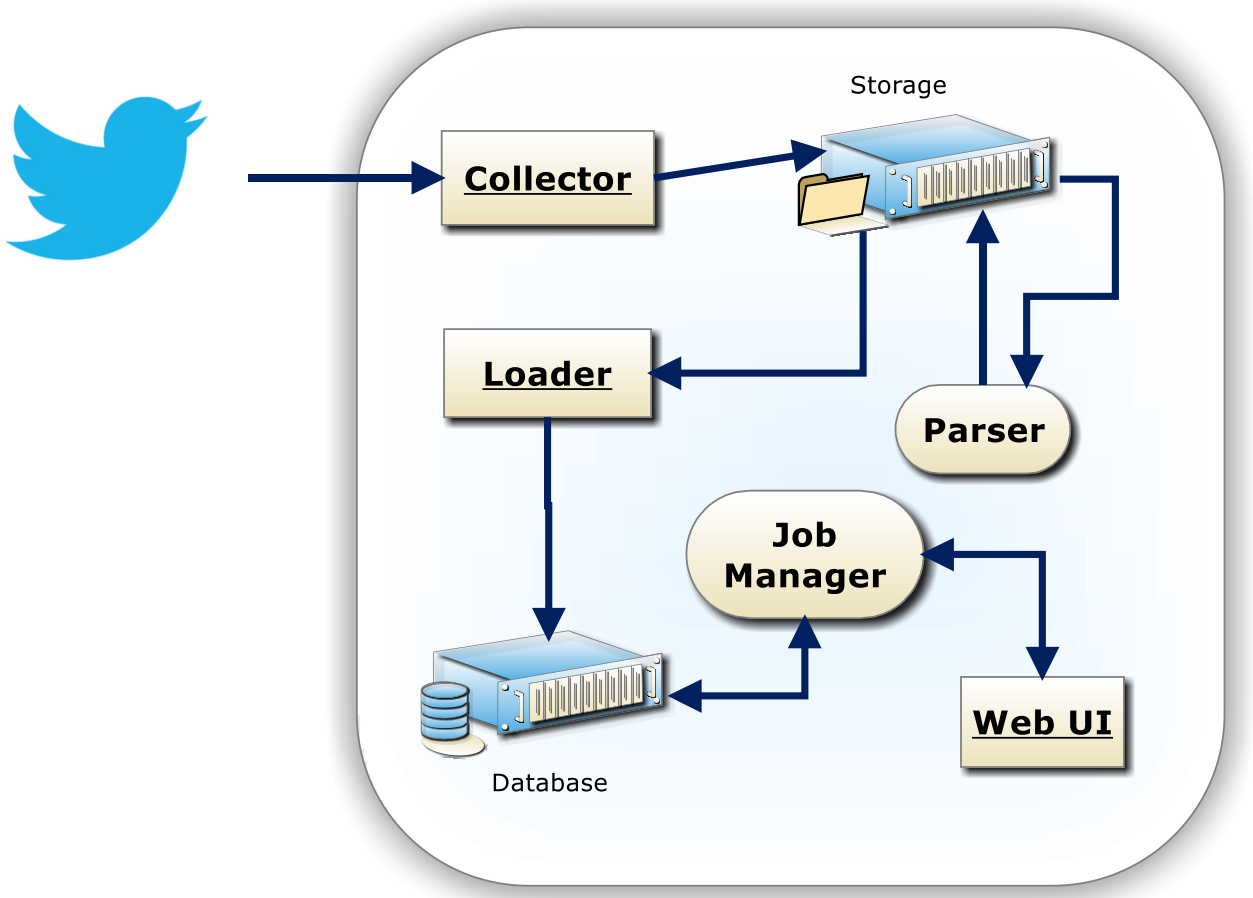
NMVO

- Gyors és lassú lekérdezési sor
- MyDB, saját adatbázis az eredményeknek
- Több adatbázis kapcsolat
- Plot
- Query plan
- Schema browser
- Csoport kezelés

Twitter VO

- Cél:
  - Twitter adatok gyűjtése, tárolása elemzés céljából

Twitter VO


Storage
Collector
Loader
Parser
Job Manager
Web UI
Database

Collector

- Sample API
- Napi ~12GB JSON adat
- Backup gyűjtő (ciklikus)
- Éles gyűjtő

Storage

- Táblák
  - Tweet
  - User
  - Hashtag
  - User Mention
  - Media
  - URL
  - Retweet

Loader

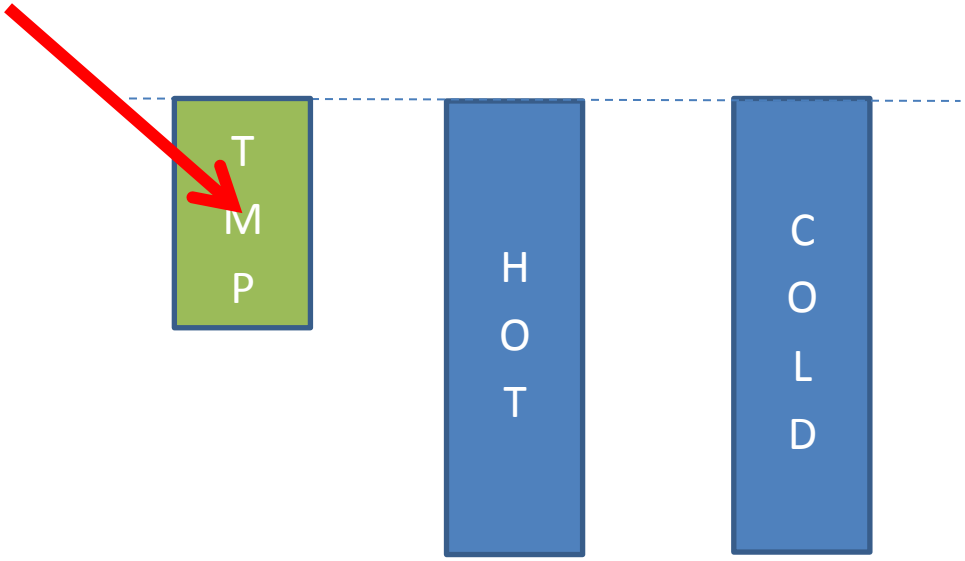
- Problémák:
  - Hálózat, I/O
- Speciális karakterek:
  - €©Д☺你好こんにちはمرحبا
- Betöltés 1 nap (~12GB)
  ~6 óra

Merge

- Retweet-ben megtalálható az eredeti tweet is
  - Nem lánc lesz a retweetekből
  - az „ős” tweet-t tartalmazza
- Szükséges a merge:
  1. Diff táblába töltünk, és az inaktív táblába merge-lünk
  2. Merge segítő indexek szükségesek.

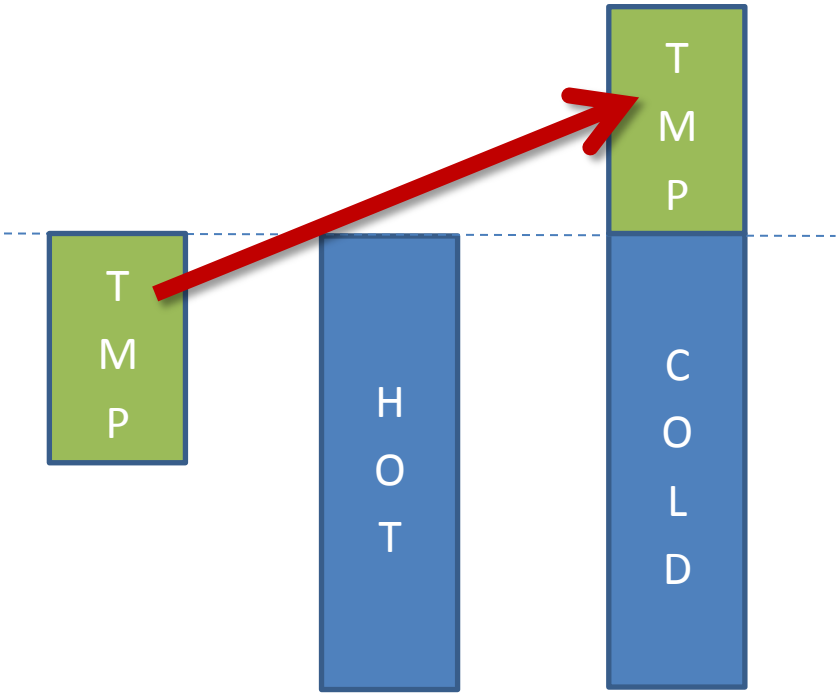
Loader

- „Hot” table
  - Webes elérés
- „Cold” table
  - Betöltéshez
- Duplikátumok eltávolítása



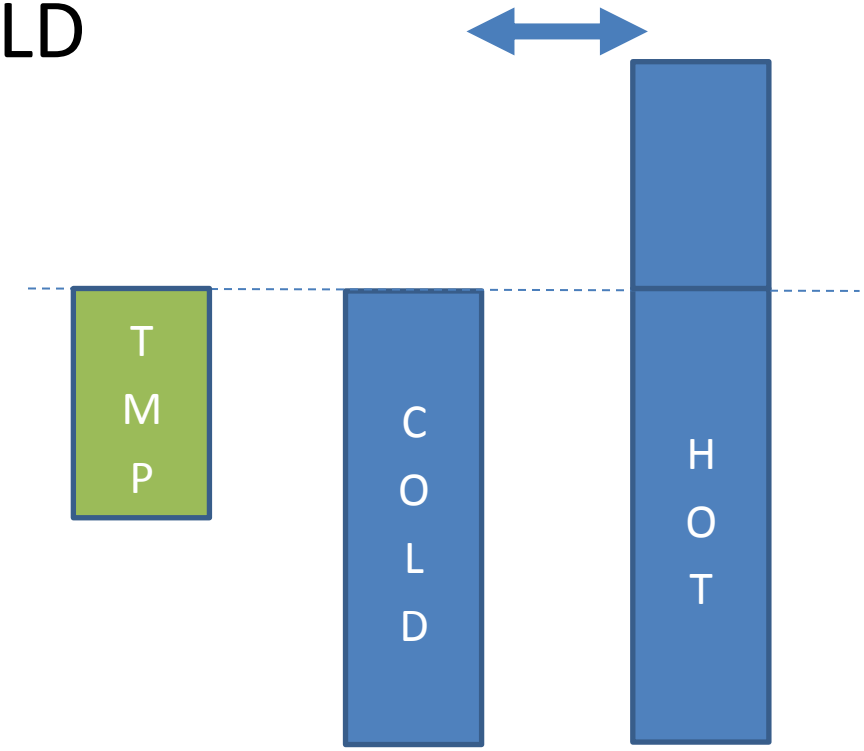
Csere előtt

- Merge TMP → COLD
  - Sorok mergelése
  - Merge indexek eltávolítása
  - Query indexek készítése



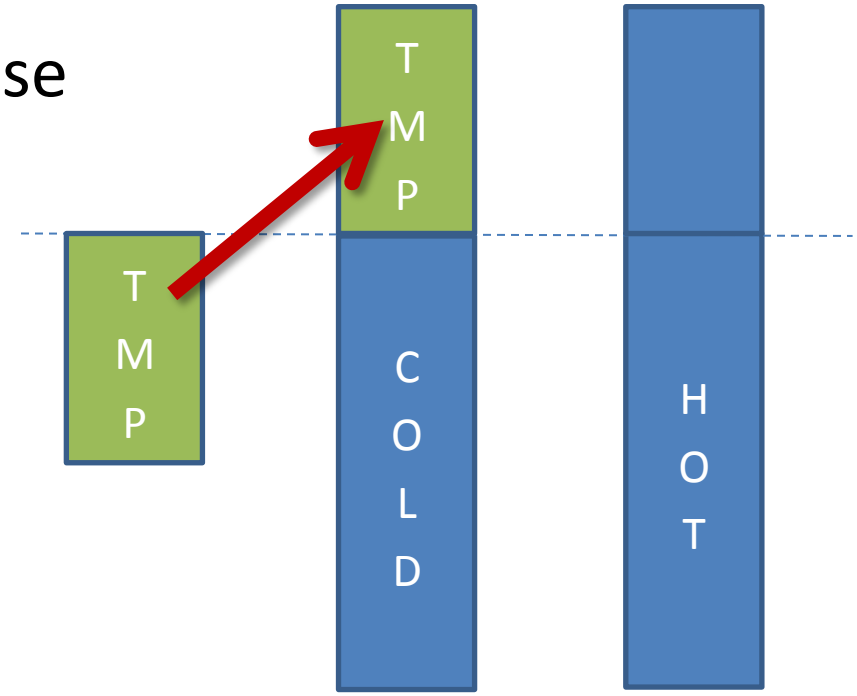
Csere

- Átnevezés COLD → HOT
- Átnevezés HOT → COLD



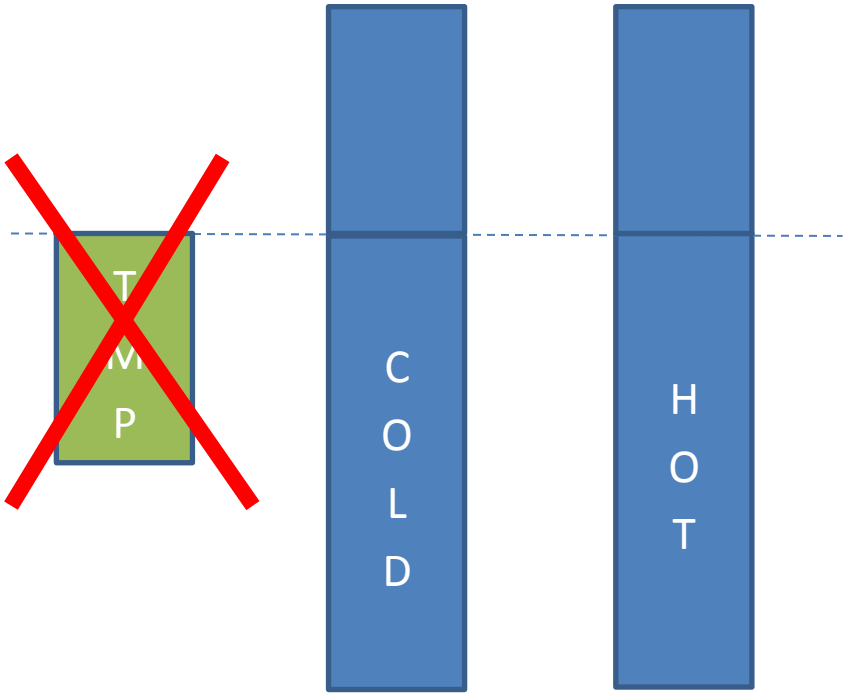
Csere után

- Merge TMP -> COLD (megint)
  - Query indexek eltávolítása
  - Merge indexek készítése
  - Sorok mergelése



Csere véglegesítése

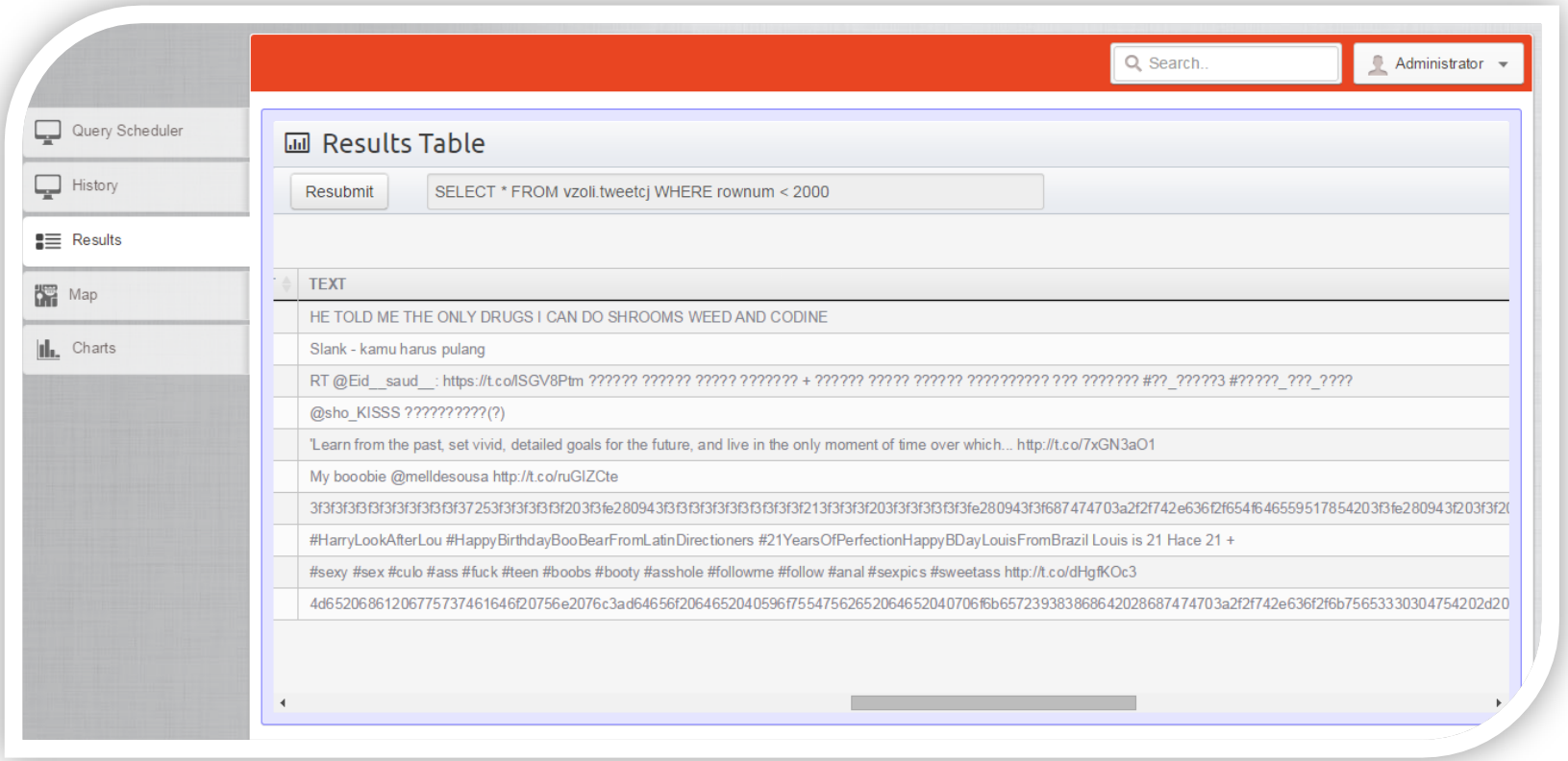
- Temp tábla eltávolítása



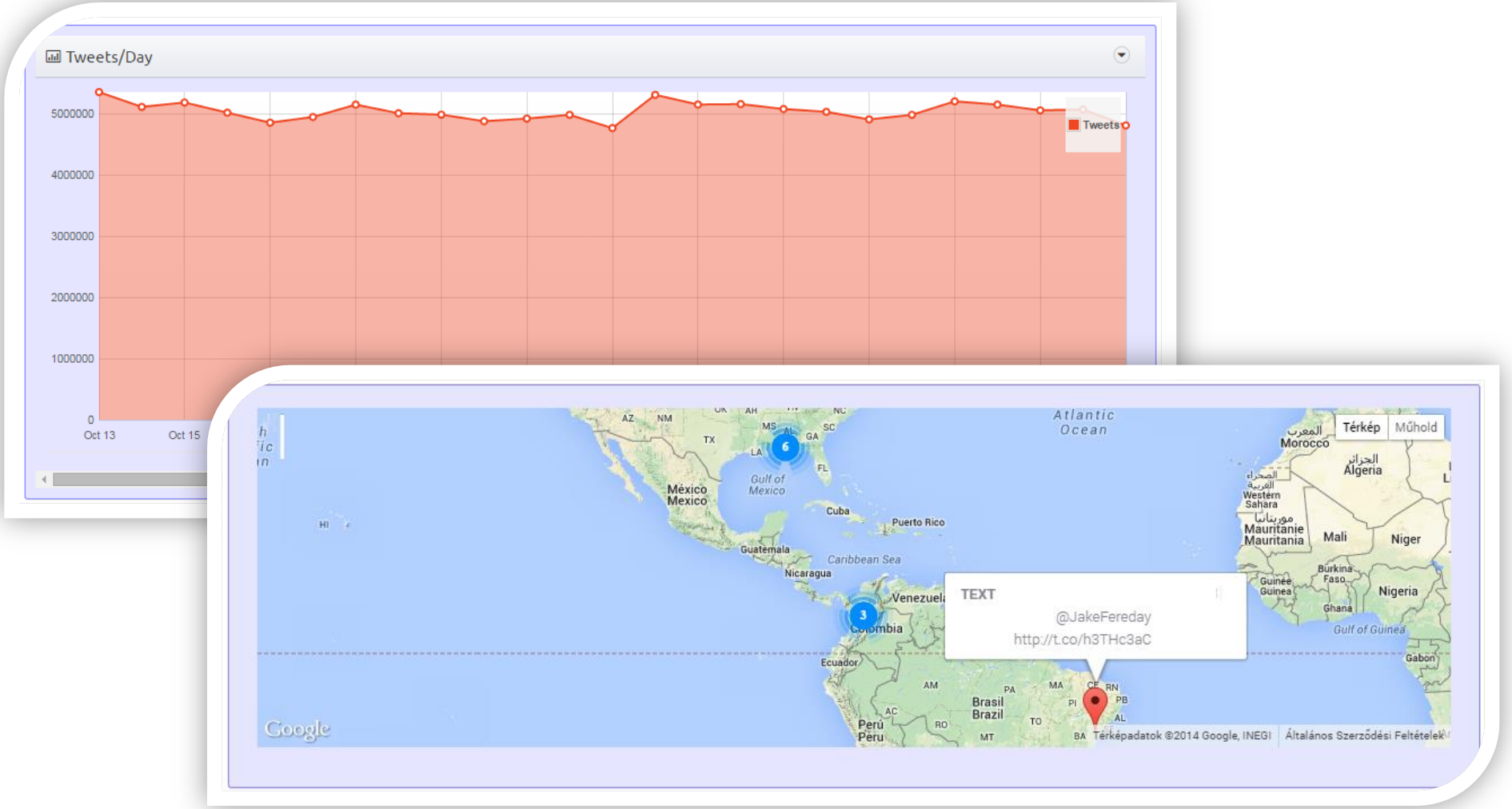
Job Manager

- Ütemező csomagok
  - DBMS_JOB
  - DBMS_SCHEDULER

Web UI


Search..
Administrator
Query Scheduler
History
Results
Map
Charts
Results Table
Resubmit
SELECT * FROM vzoli.tweetcj WHERE rownum < 2000
TEXT
HE TOLD ME THE ONLY DRUGS I CAN DO SHROOMS WEED AND CODINE
Slank - kamu harus pulang
RT @Eid__saud__: https://t.co/lSGV8Ptm ?????? ?????? ????? ??????? + ?????? ????? ?????? ?????????? ??? ??????? #??_?????3 #?????_???_????
@sho_KISSS ??????????(?)
'Learn from the past, set vivid, detailed goals for the future, and live in the only moment of time over which... http://t.co/7xGN3aO1
My booobie @melldesousa http://t.co/ruGIZCte
#HarryLookAfterLou #HappyBirthdayBooBearFromLatinDirectioners #21YearsOfPerfectionHappyBDayLouisFromBrazil Louis is 21 Hace 21 +
#sexy #sex #culo #ass #fuck #teen #boobs #booty #asshole #followme #follow #anal #sexpics #sweetass http://t.co/dHgfKOc3

Web UI


Tweets/Day
5000000
4000000
3000000
2000000
1000000
0
Oct 13
Oct 15
Tweets
Térkép
Műhold
Atlantic Ocean
México
Mexico
Gulf of Mexico
Cuba
Puerto Rico
Guatemala
Caribbean Sea
Nicaragua
Venezuela
Colombia
Ecuador
Perú
Peru
Brasil
Brazil
Morocco
Algeria
Western Sahara
Mauritanie
Mauritania
Mali
Niger
Burkina Faso
Guinée
Guinea
Nigeria
Ghana
Gulf of Guinea
Gabon
TEXT
@JakeFereday
http://t.co/h3THc3aC
Google
Térképadatok ©2014 Google, INEGI
Általános Szerződési Feltételek

Feladatok (NMVO)

http://nm.vo.elte.hu/casjobs

- Hány tweet volt 2012. december 24-én?
- Hányban szerepet az XMAS szó ezek közül?
- Hányban szerepelt a http://www.youtube.com/watch?v=z8Vfp48laS8 ?
- Hány magyar nyelvű tweet volt?
- Melyik tweetet retweetelték a legtöbbször aznap?
- Hányan retweeteltek aznap?
- (Ki,kit) retweetelt gráfnak hány csúcsa, hány éle van?

Feladatok (Twitter Casjobs)

Táblák: vzoli.tweetcj, gognaai.followers

1. Hány tweet volt 2012-12-24 napon? (count)
2. Melyik a legrégebbi tweet? (min)
3. Irassuk ki a legkorábbi tweetet (order by, rownum)
4. Legtöbbet retweetelt tweet kiiratása (max)
5. Nyelvenként hány tweet van? (group by)
6. Hány tweetben szerepelt „Obama”? (like)

Feladatok (Twitter Casjobs)

Táblák: vzoli.tweetcj, gognaai.followers

4. Nyelvenként hány tweet van? (group by)
5. Hány magyar tweet volt? (where)
6. Hány tweetben szerepelt „Obama”? (like)
7. Írjuk ki a 1021951981-es user követői, milyen nyelven tweetelnek. (join, distinct)
8. Rajzoljuk grafikonon a nyelvek eloszlását!
9. Rajzoljuk térképen az első ezer olyan tweet-et amelynek nem null a lat, lon koordinátája!