

# Virtuális Obszervatórium

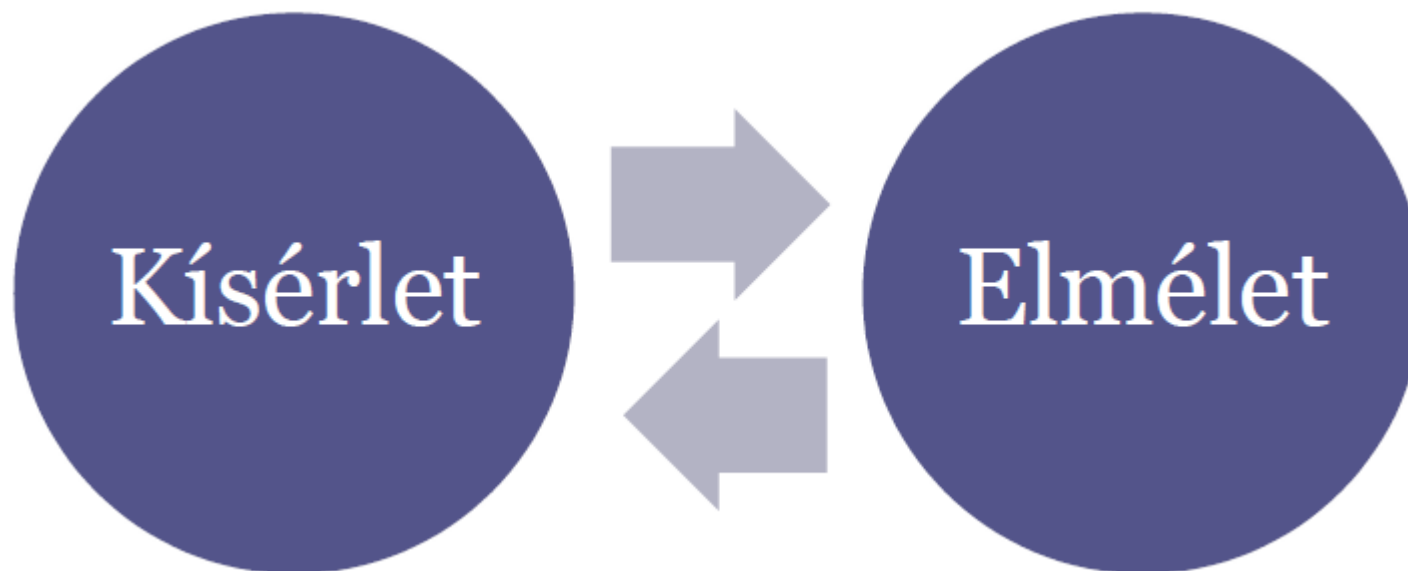
Gombos Gergő

# Áttekintés

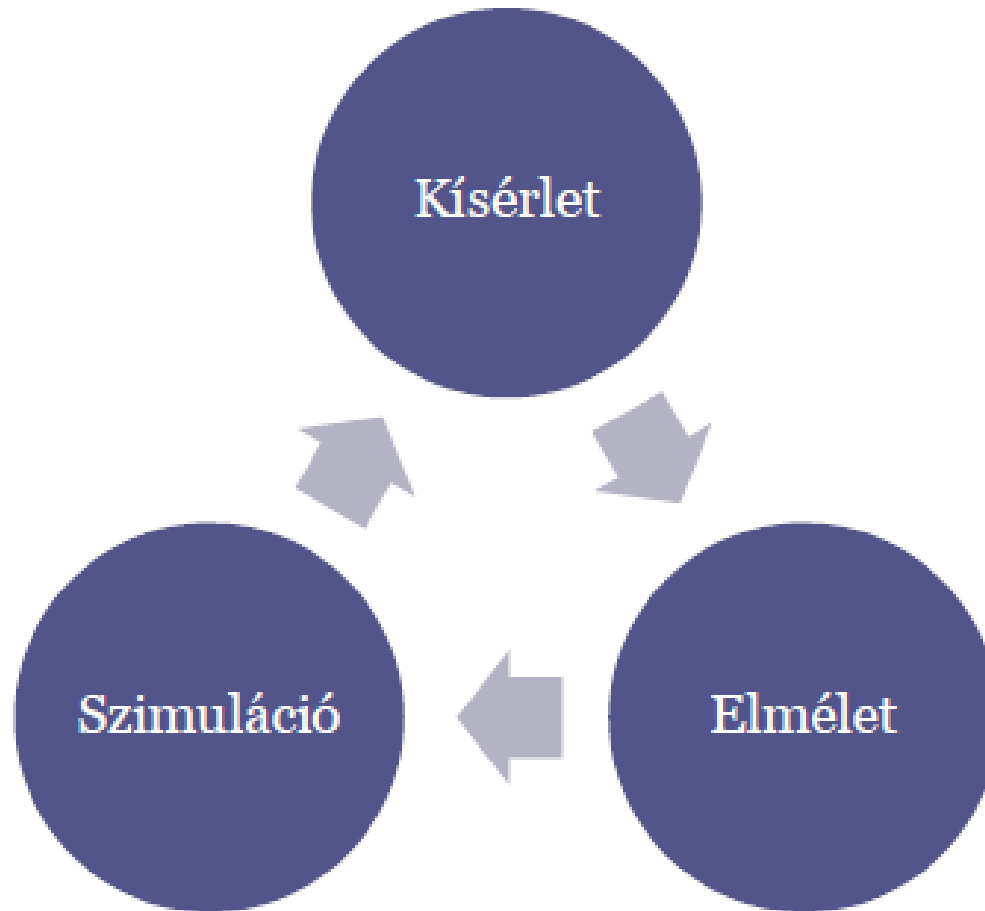
- Motiváció, probléma felvetés
- Megoldások
- Virtuális obszervatóriumok
- NMVO
- Twitter VO

# Motiváció

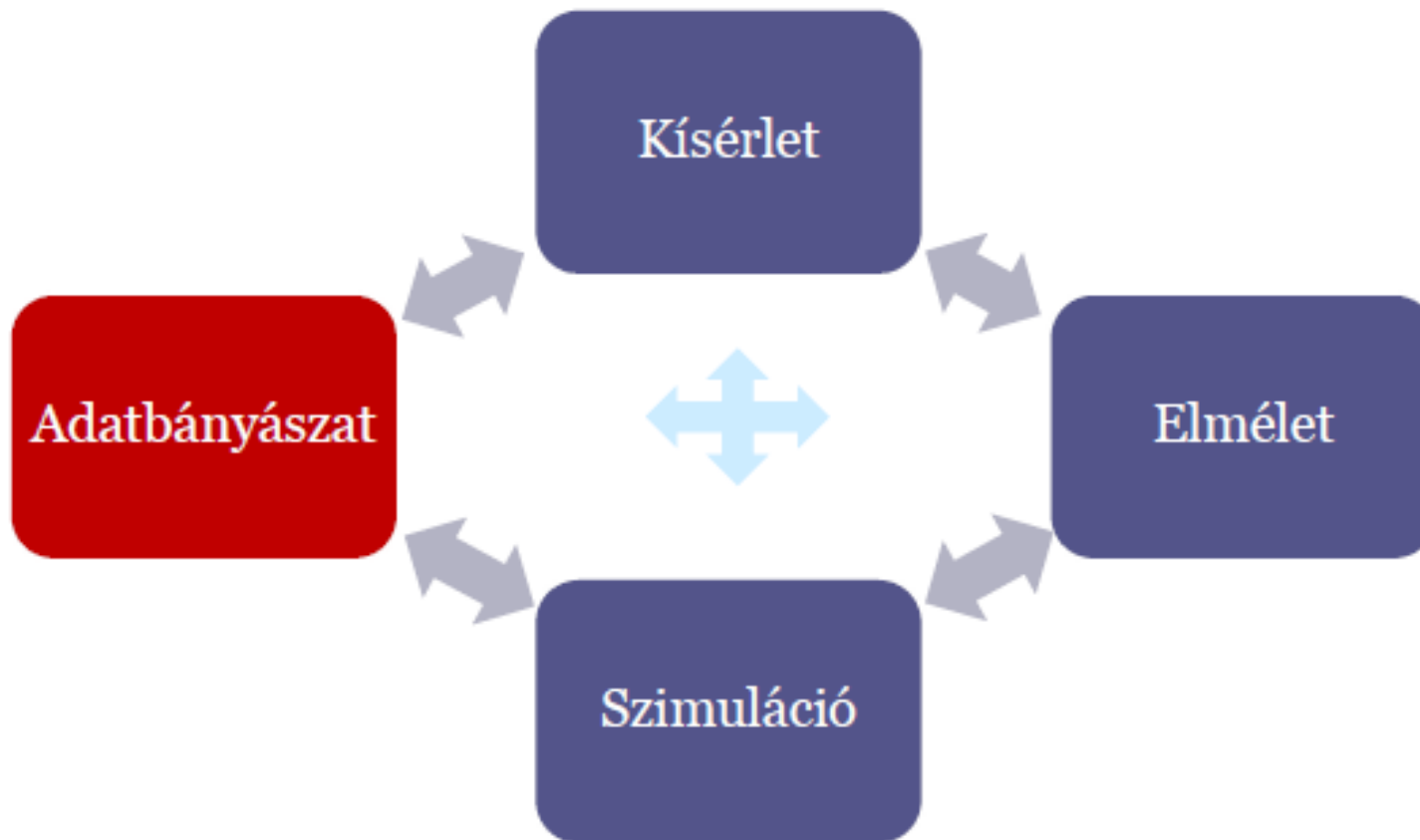
- Tudományos módszer fejlődése



# 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.
  - Az eredményeiket meg tudják osztani más felhasználókkal.



# 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. Kiéhezheti a többi klienst.

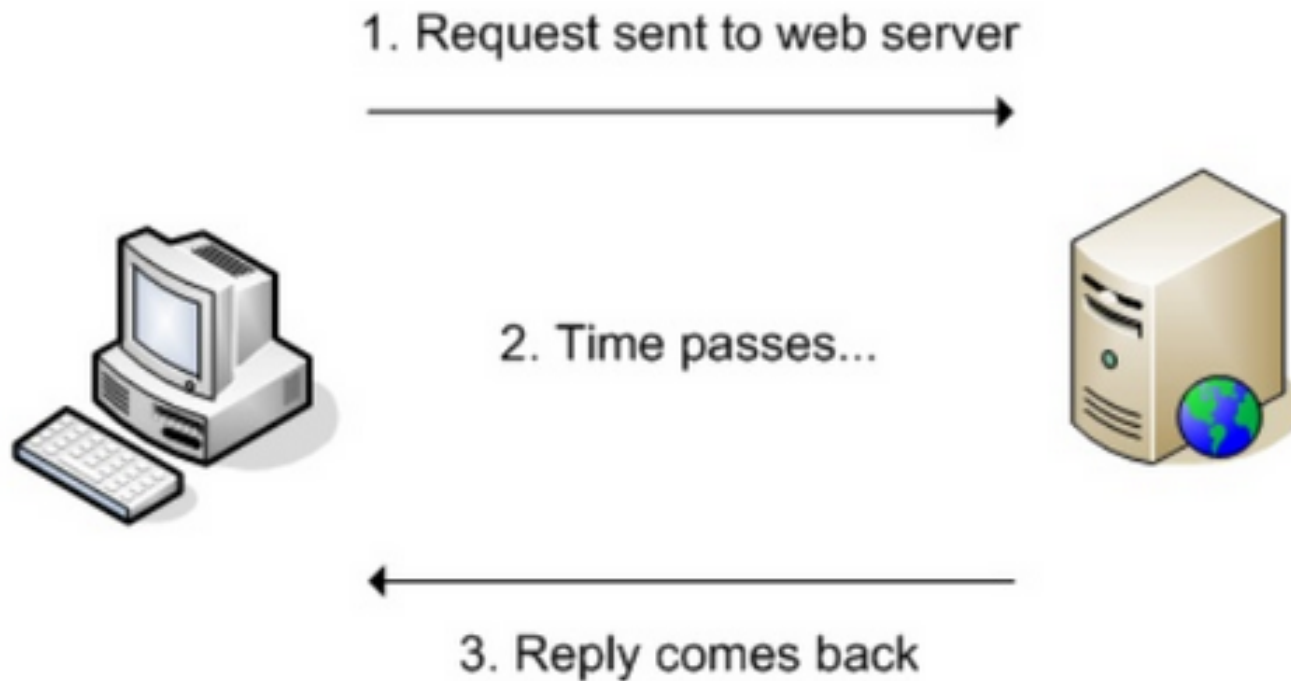
# Egyszerű megoldás

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

# Egyszerű megoldás

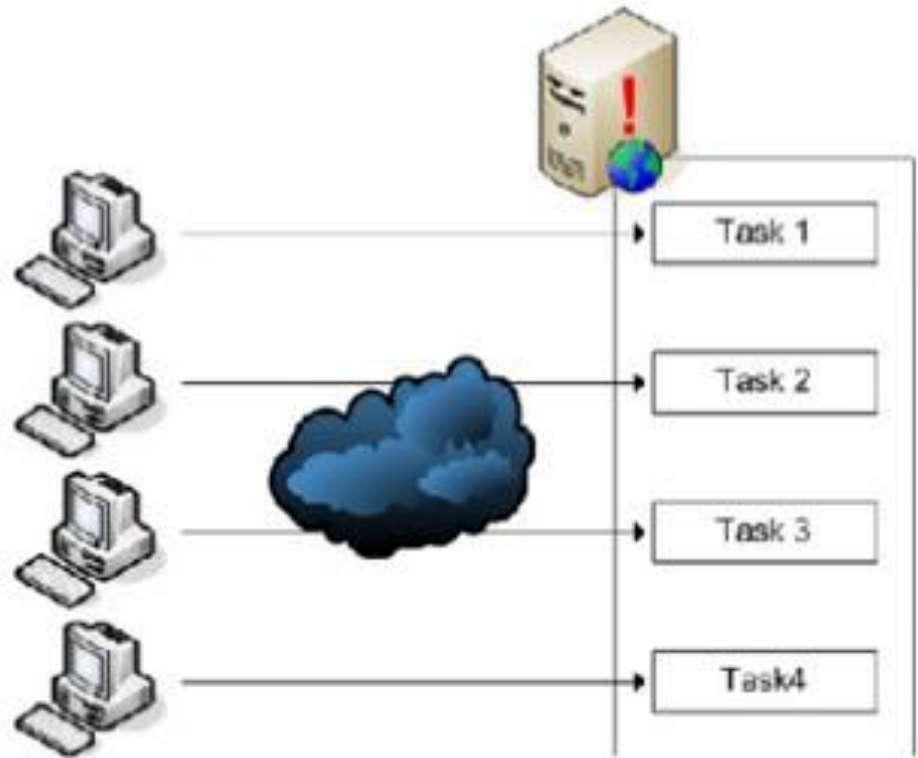
- Kliens --- Alkalmazás szerver --- DB szerver
- Csökkenti a terhelést a DB-n
- Szinkron megoldás
- Probléma:
  - Lassú válasz, timeout.

# Szinkron megoldás



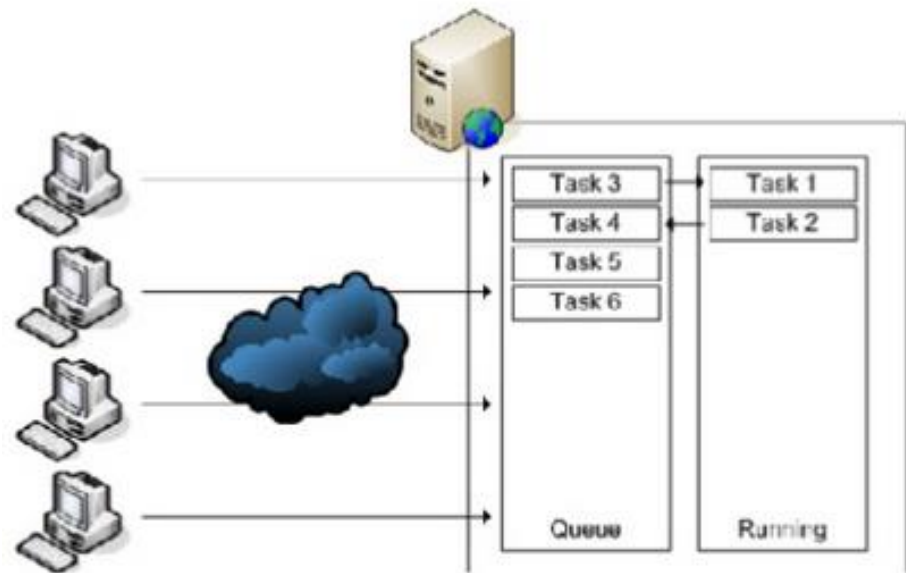
# Szinkron működés

- Szerver terhelés nagy

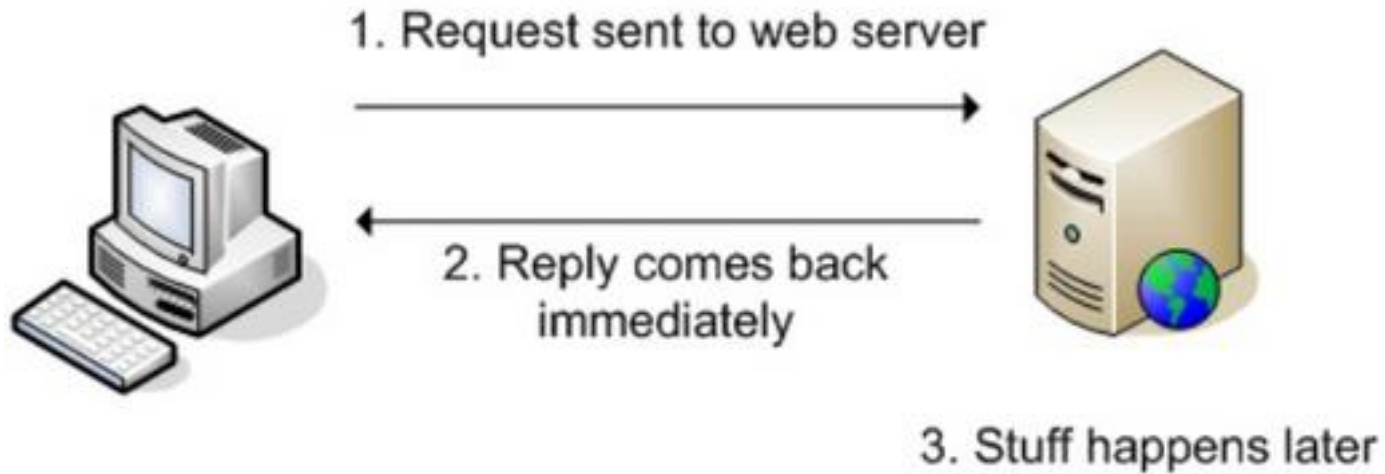


# Egyszerű javított megoldás

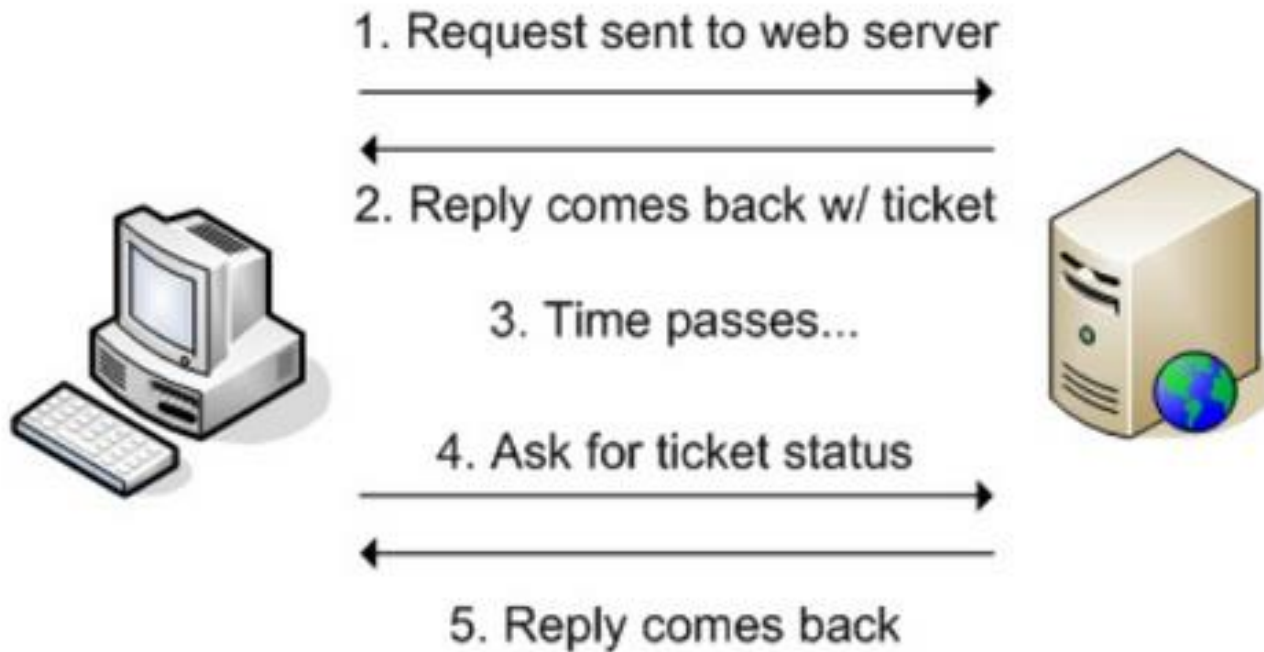
- Aszinkron
- Ticket-rendszer
- Eredmény később
- Sorba állítja a kéréseket.



# Aszinkron megoldás



# Aszinkron megoldás





# Mi a VO (Virtuális Obszervatórium)?

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 Observatórium?



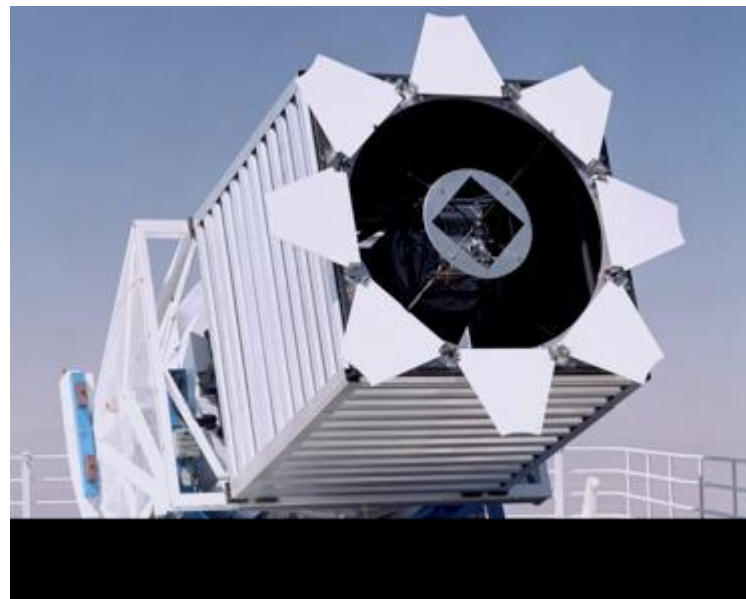
Galileo Galilei

Edwin Hubble



# Miért Observatórium?

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

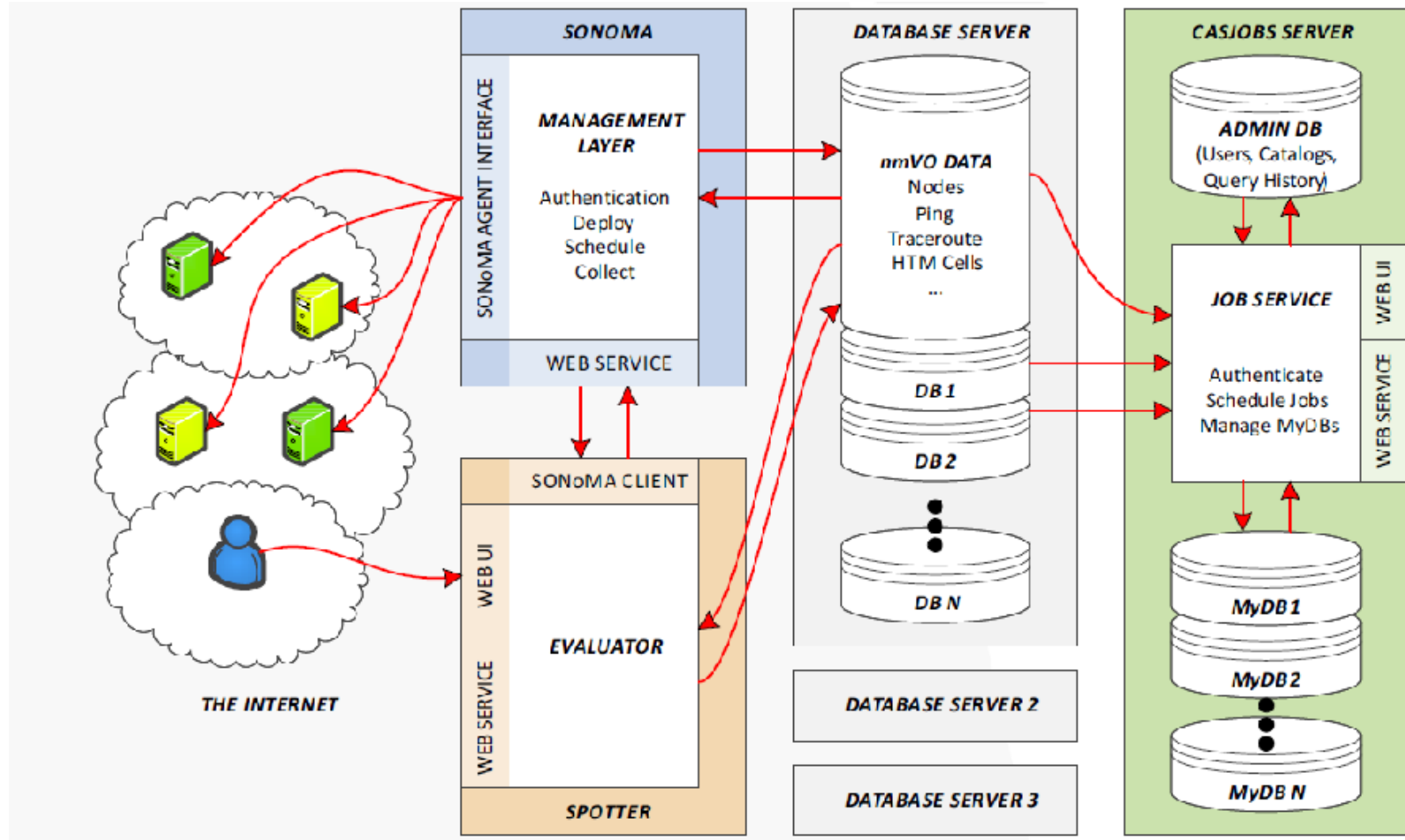


Teleszkóp           -> Digitális adatok  
Detektorok       -> Számítógépes programok

# 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





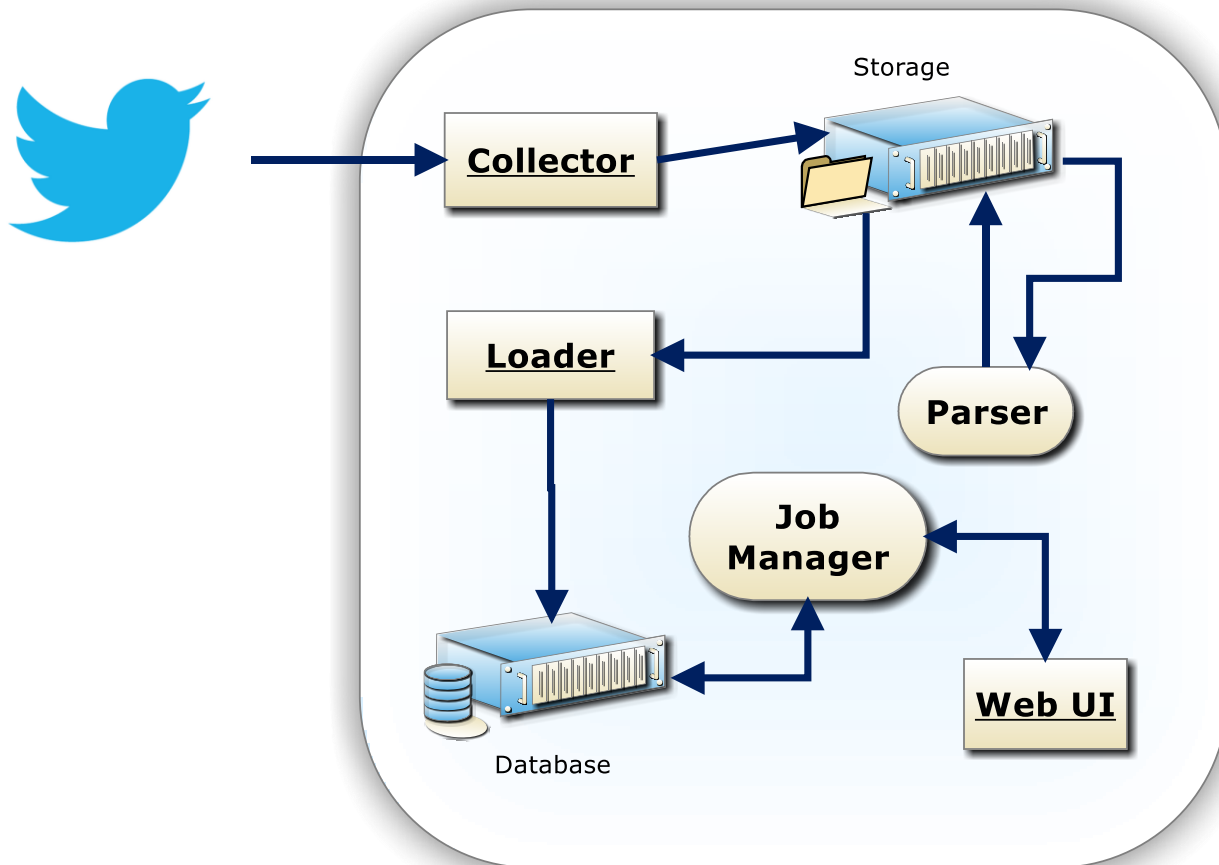
# NMVO

- 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



# Collector

- Sample API
- Napi ~15GB 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 (~15GB)  
~6 óra

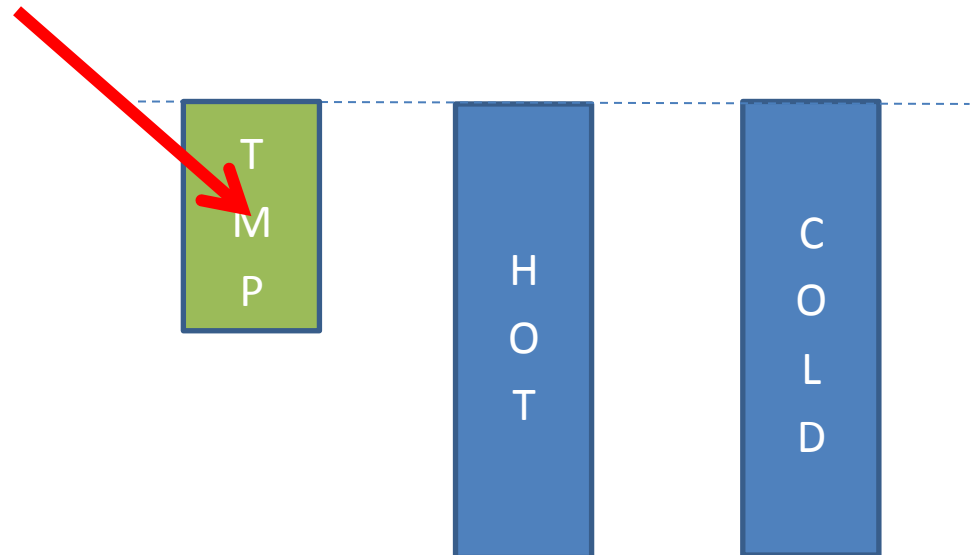


# Merge

- 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

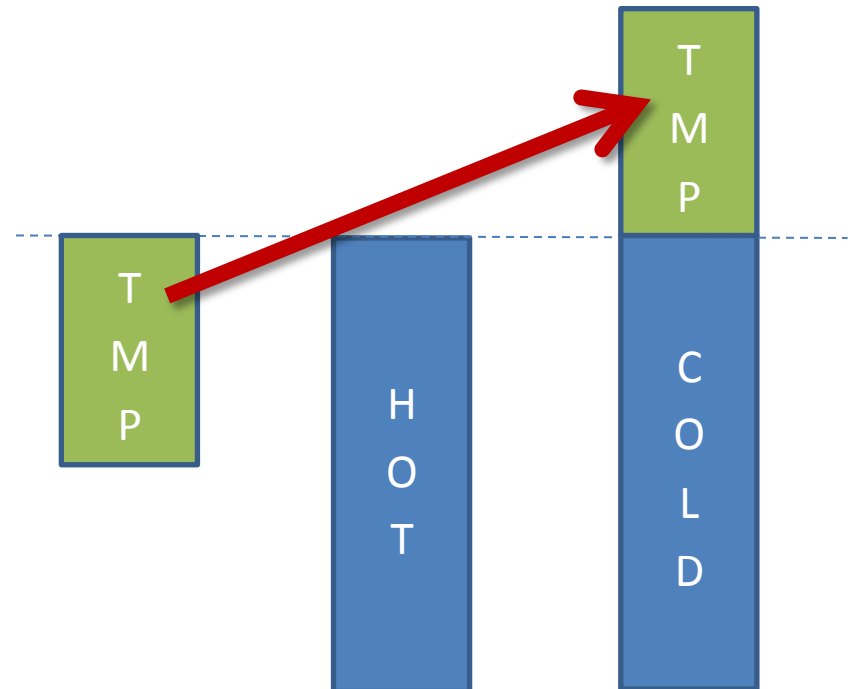
- „Hot” table
  - Webes elérés
- „Cold” table
  - Betöltéshez
- Duplikátumok eltávolítása





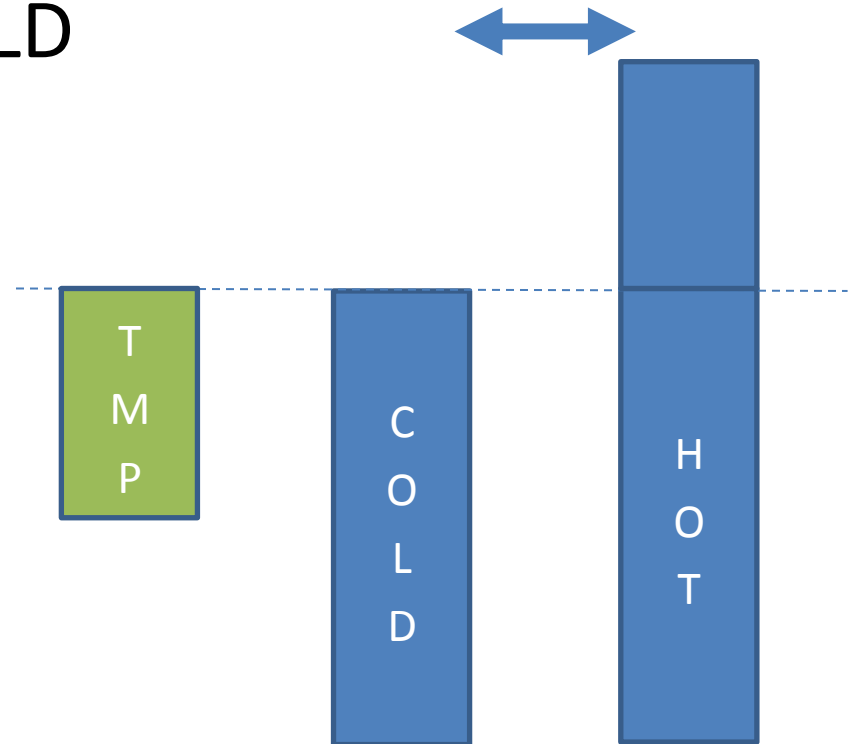
# Csere előtt

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



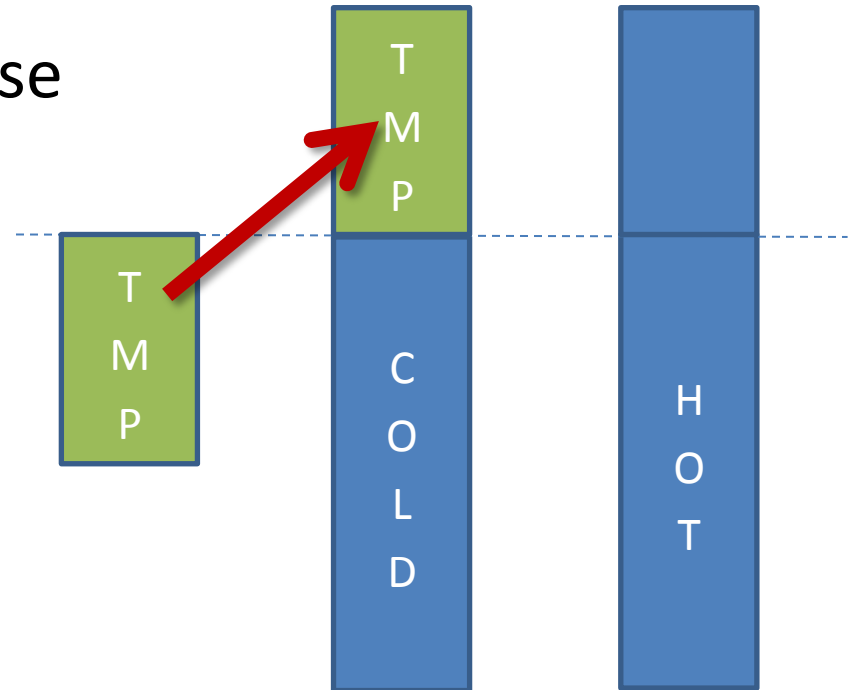
# Csere

- Átnevezés COLD → HOT
- Átnevezés HOT → COLD



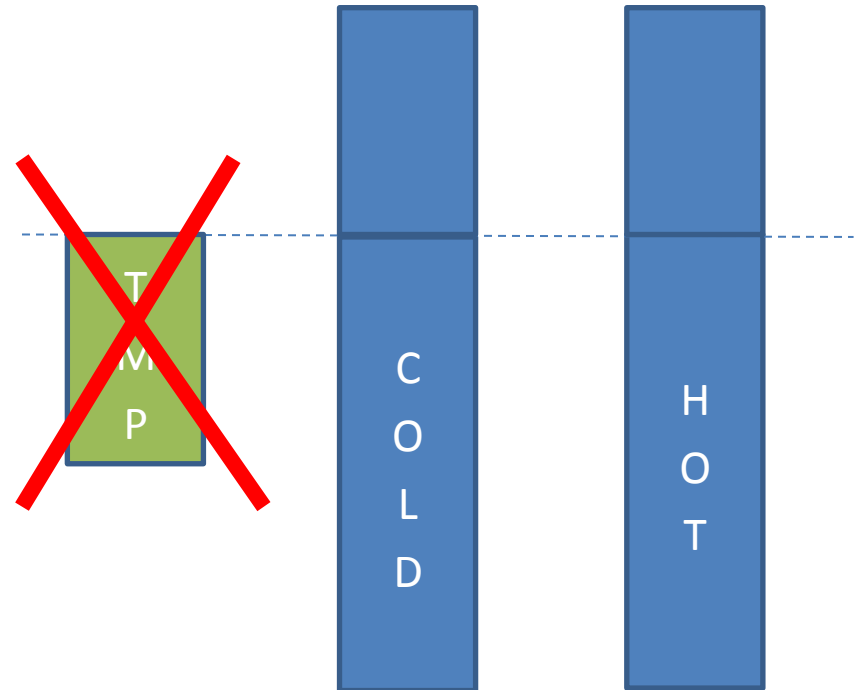
# Csere után

- 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

