Bevezetés a Big Data világába

Rusznák Attila

SZTAKI

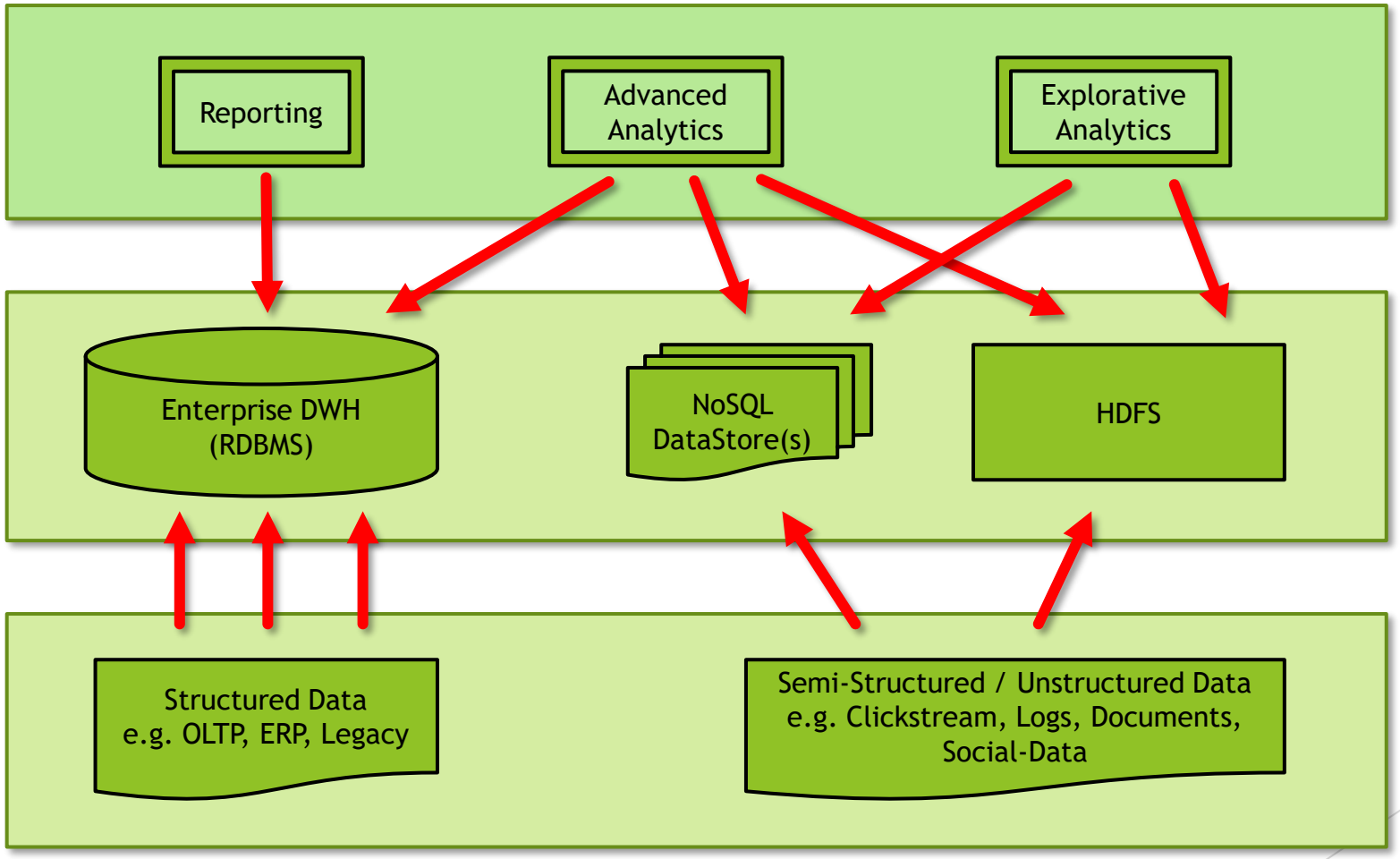
Bemutatkozás

ELKH Cloud

Rusznák Attila

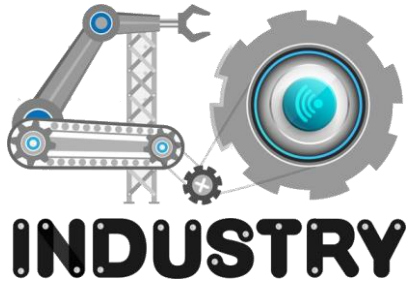
- SZTAKI PERL, fejlesztő
- Óbudai Egyetem NIK, oktató
- Szegedi Tudományegyetem, tanszéki mérnök

A Big Data és a Business Intelligence



A Big Data kialakulása

Az elmúlt 20-25 évben az alábbi technológiák változtatták meg a világunkat és hozták létre a Big Data-t:



BIG DATA

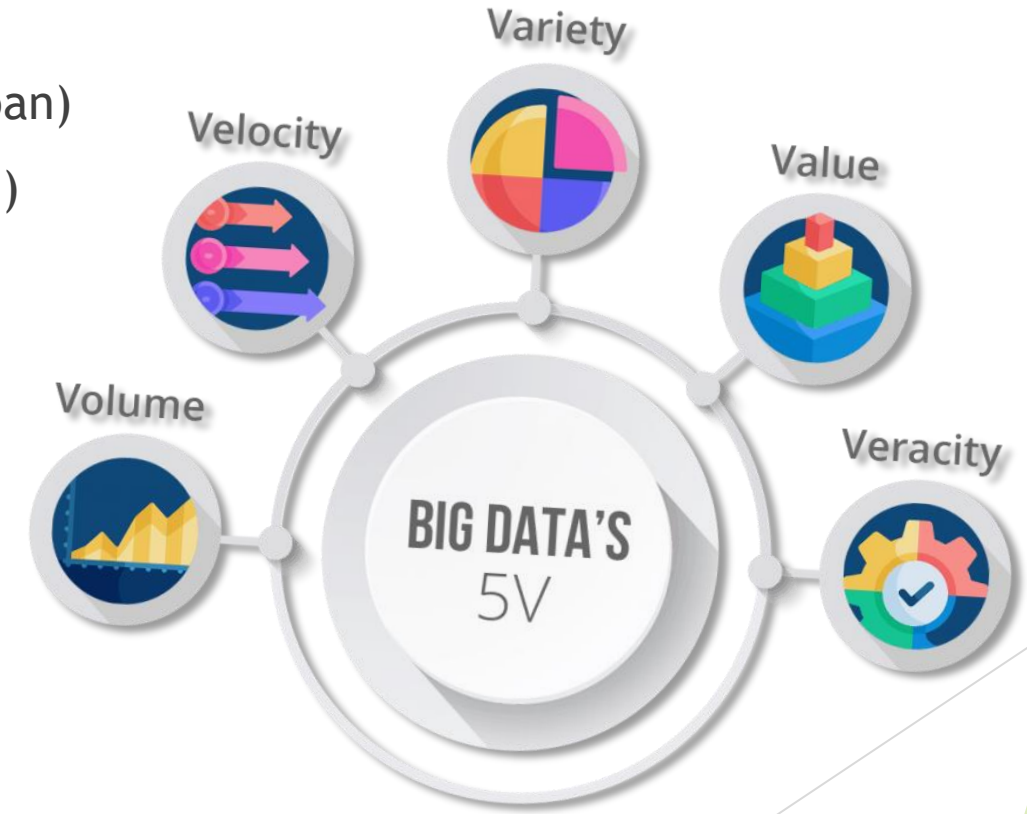
eCommerce
(C2B, C2C B2B)

Mobile WEB

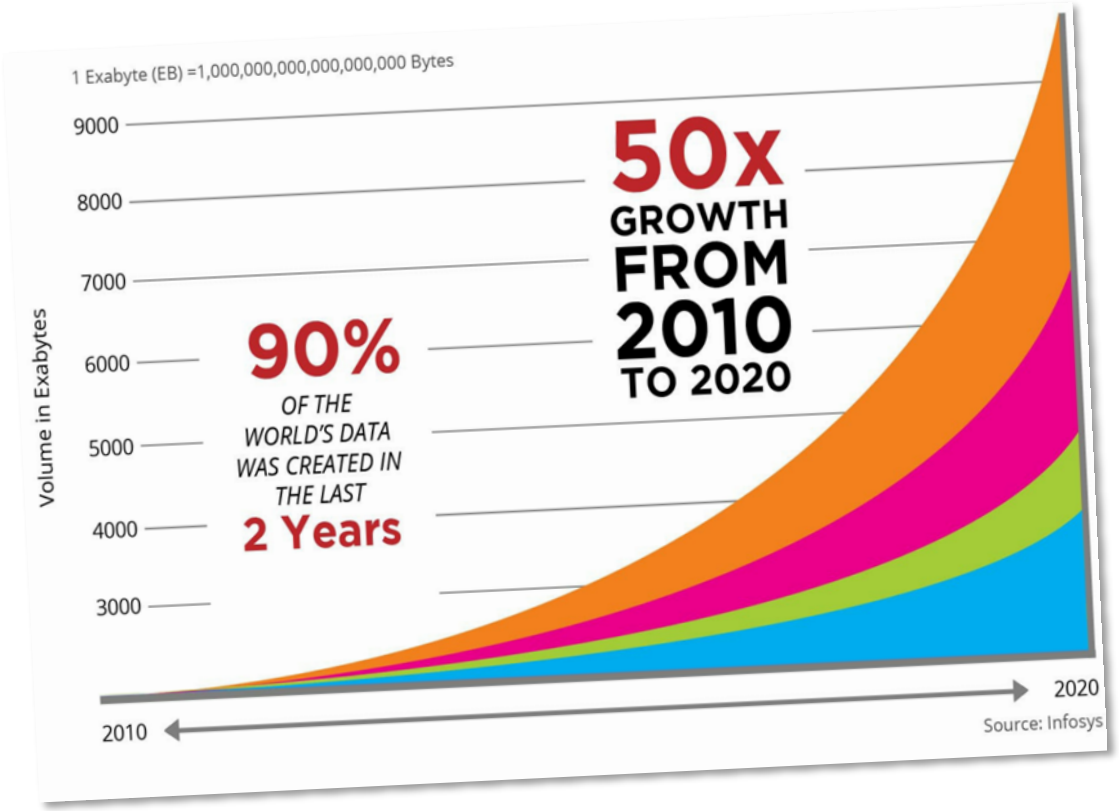
Az 5V

Néhány éve még 3V-vel definiálták a Big Data-t, ma ez már 5V:

- Volume (nagy adatmennyiség)
- Velocity (nagy sebességgel)
- Variety (különböző formátumokban)
- Veracity (eltérő adatminőségben)
- Value (értéket képvisel)



Volume: adatnövekedés



Machine generated data (IoT)

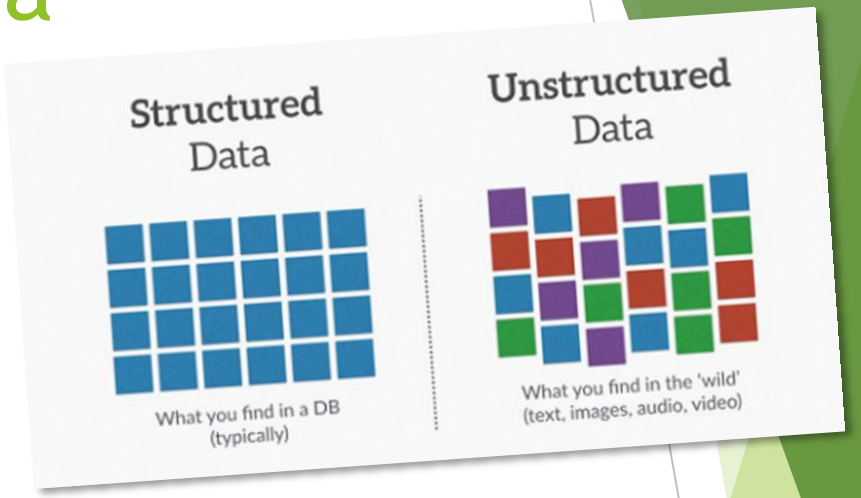
Social interactions

Human files

Transactaional data

Variety: az adatok struktúrája

- Strukturált adatok
- Félig strukturált adatok
- Nem strukturált adatok



Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
 ....
</University>
```

Structured data

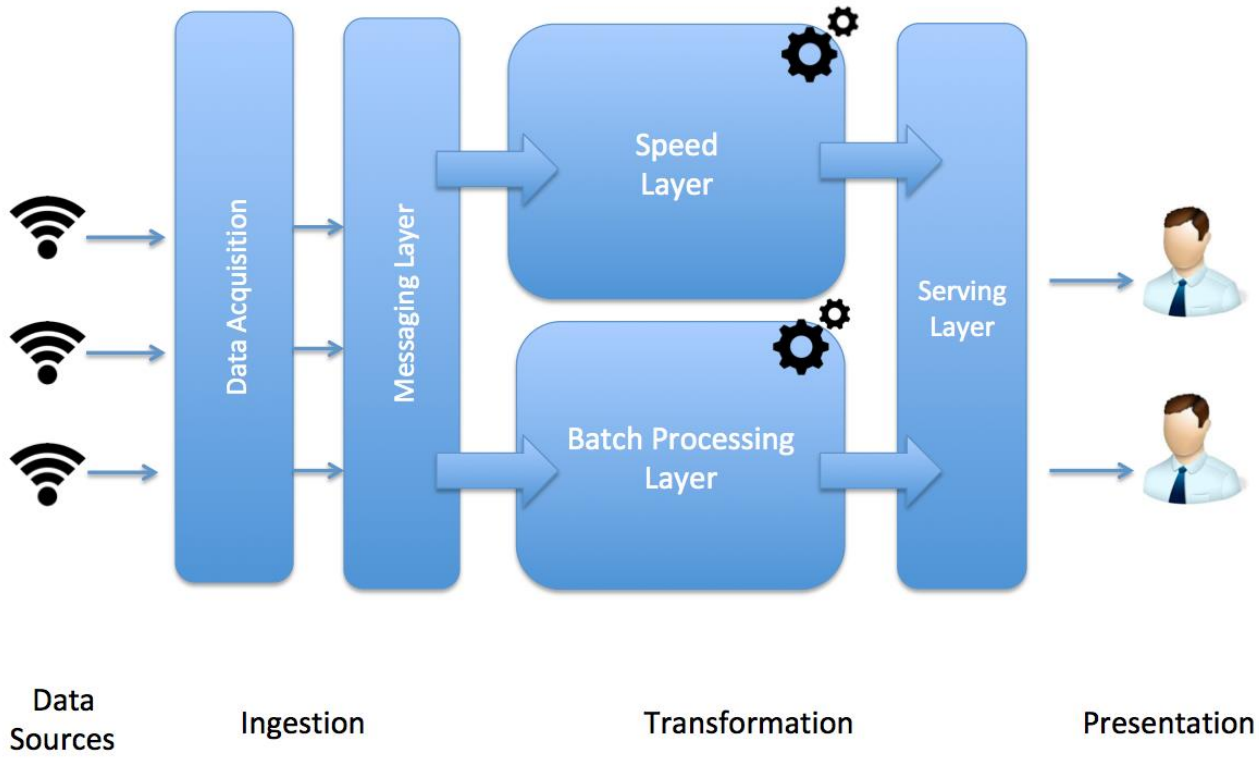
ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Velocity: az adatok sebessége

Big Data rendszerekkel szembeni elvárások

Követelmény:

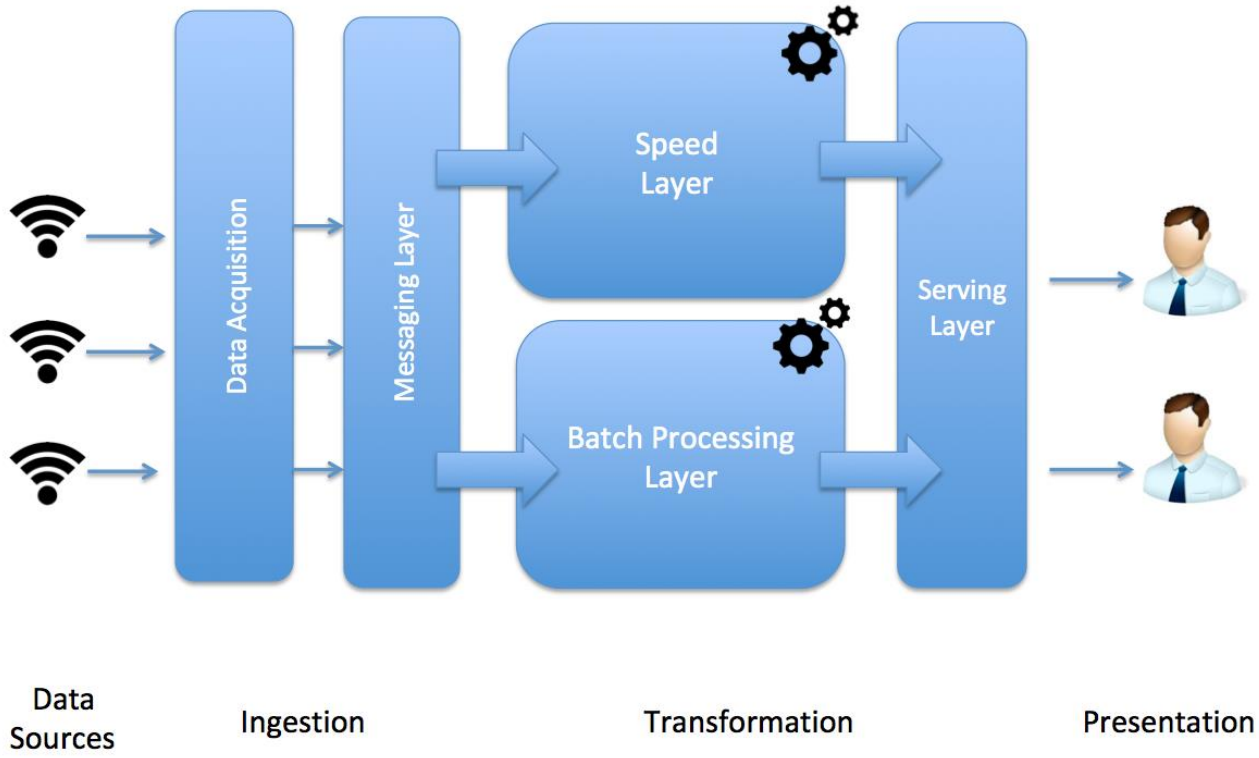
- Használati esetek:
  - „Near” real time adatfeldolgozás
    - Nagy mennyiségű bejövő adat gyűjtése
    - Közel valós idejű feldolgozása
    - Adatfolyamon érkező valós idejű adat
    - Pl. szenzoradatok kiértékelése, banki csalások detektálása



Big Data rendszerekkel szembeni elvárások

Követelmény:

- Használati esetek:
  - Kötegelt (batch) adatfeldolgozás
    - Nagy mennyiségű adaton történő kiértékelés
    - Adatforrás tároló egységből származik
    - Futási idő napokban/hetekben is mérhető

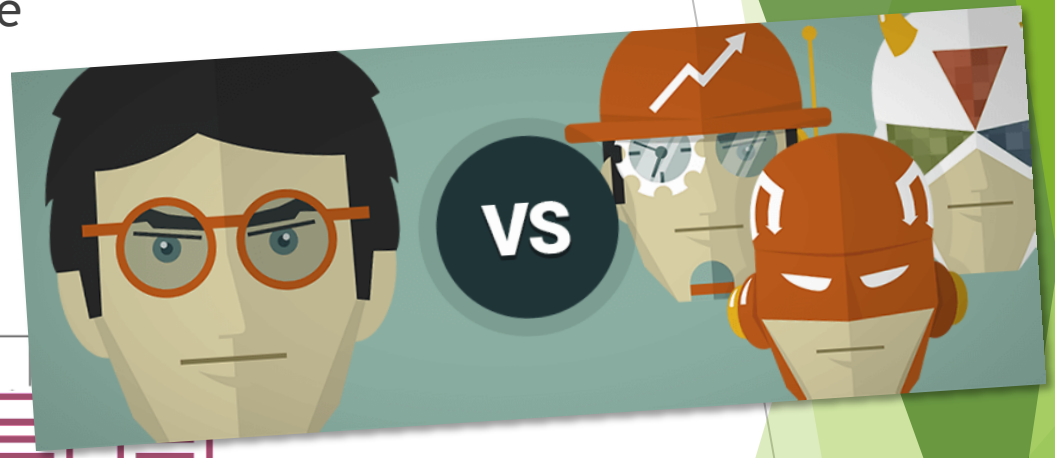


Mennyire nagy a Big Data? (2016)

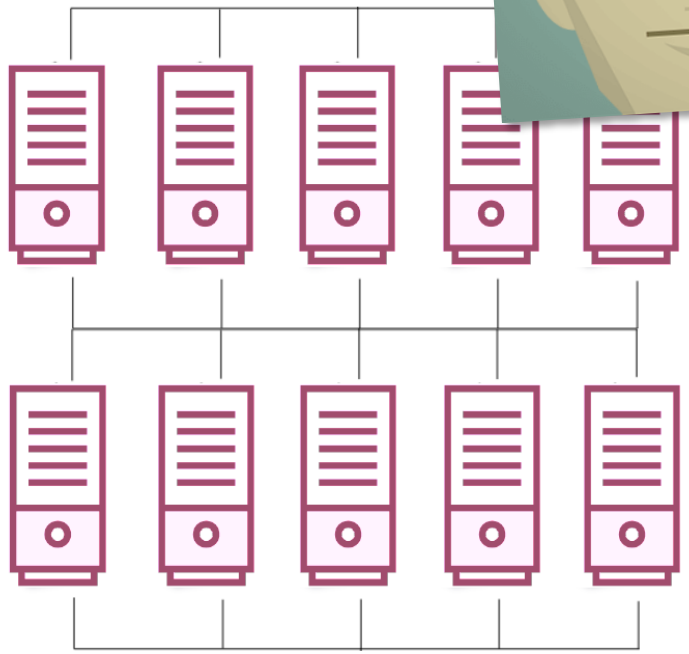
- Facebook 300 PB adat
  - 600 TB / nap
  - 1 milliárd felhasználó / nap
  - 2.7 milliárd like / nap
  - 300 millió kép / nap
- NSA: 5 EB adat
  - 30 PB / nap
  - Teljes internetforgalom 1.6%-át monitorozták
- Google: 15 EB adat
  - 100 PB / nap
  - 60 millió oldal indexelés
  - 1 milliárd egyedi keresés / hó

Számítási rendszerek fajtái

- **Monolitikus:** minden egyetlen gépbe van belesűrítve
- **Elosztott:** párhuzamosan dolgozik több gép



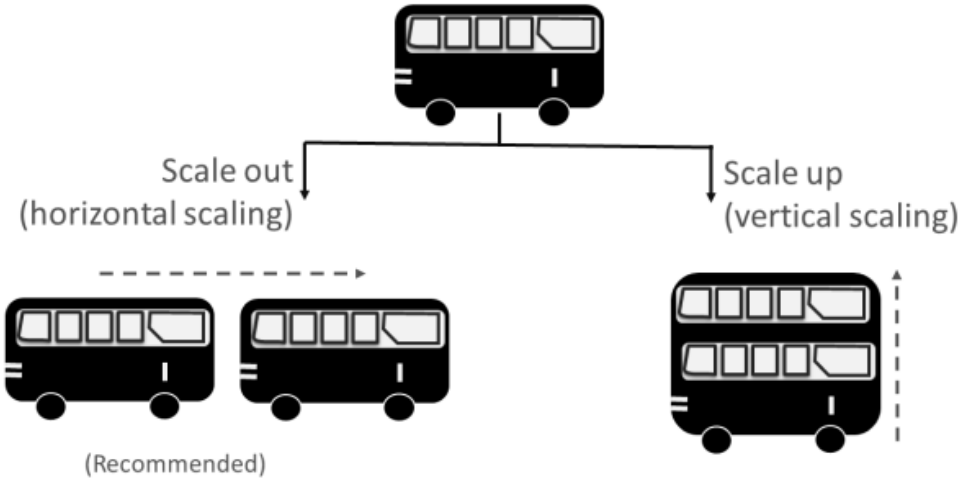
- Egy elosztott rendszer legfontosabb részei:
  - node
  - cluster

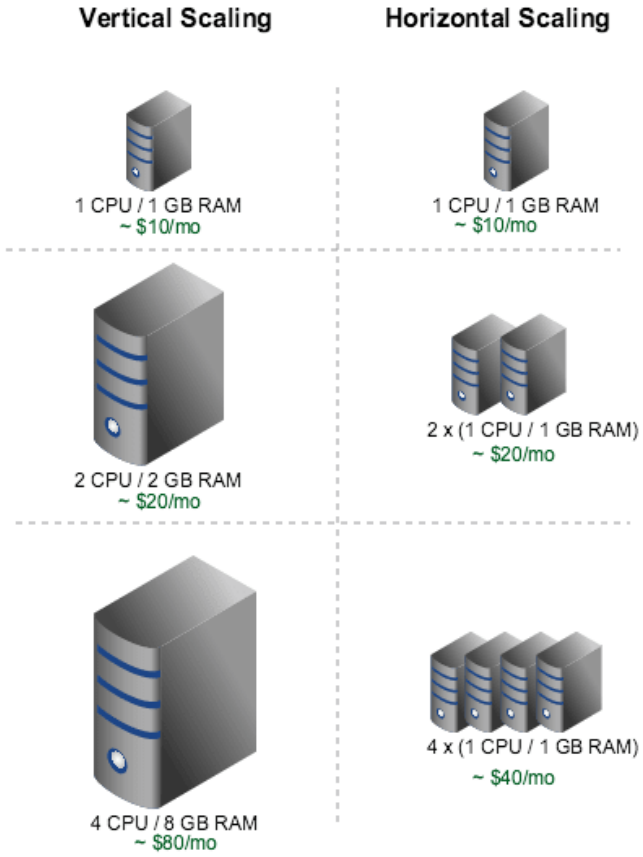


A gépek skálázhatósága

Régen nem volt ennyi adat

- Az adatok és az erőforrásigény növekszik
- A vertikális skálázás egy idő után nem opció
- Megoldás: horizontális skálázhatóság





ELKH Cloud

Egy elosztott rendszer működtetése

Szoftver kell, ami működteti ezeket a hatalmas szerverfarmokat:

- felosztja az adatokat a node-ok között
- koordinálja a számítási feladatokat
- legyen hibatűrő és visszaállítható
- allokálja a szükséges erőforrásokat