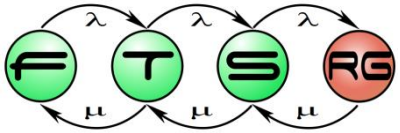
‚Big Data' elemzési módszerek

2015.09.09.



Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

A félévről

- Előadók, közreműködők
  - dr. Pataricza András
  - Dr. Horváth Gábor
  - **Kocsis Imre (op. felelős)**
  - Salánki Ágnes
  - Bolgár Bence

- ikocsis@mit.bme.hu, IB418, (+36 1 463) 2006

- 1 ZH (*terv*: 12. okt. hét), 40%
- Házi feladat
  - Kiadás: ~5. hét

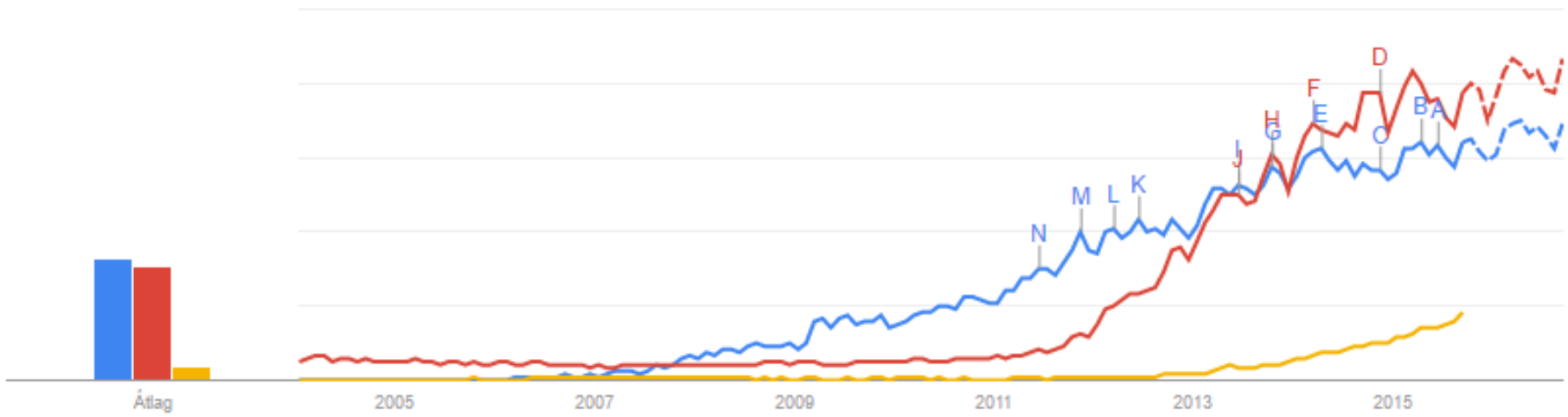
Google Trends: „Big Data”

Összehasonlítás Keresési kifejezések

hadoop	Big Data	"data science"	+ Kifejezés hozzáadása
Keresett kifejezés	Keresett kifejezés	Keresett kifejezés	

Érdeklődés idő szerint

Hírek címsorai Előrejelzés



MI AZ A “BIG DATA”?

Definíció [1]

- Adatkészletek, melyek mérete nagyobb, mint amit

- regisztrálni,
- tárolni,
- kezelni és
- elemezni tudunk

- a „tipikus” („adatbáziskezelő”) szoftverekkel.
  - Illetve a tipikus elemző szoftverekkel.

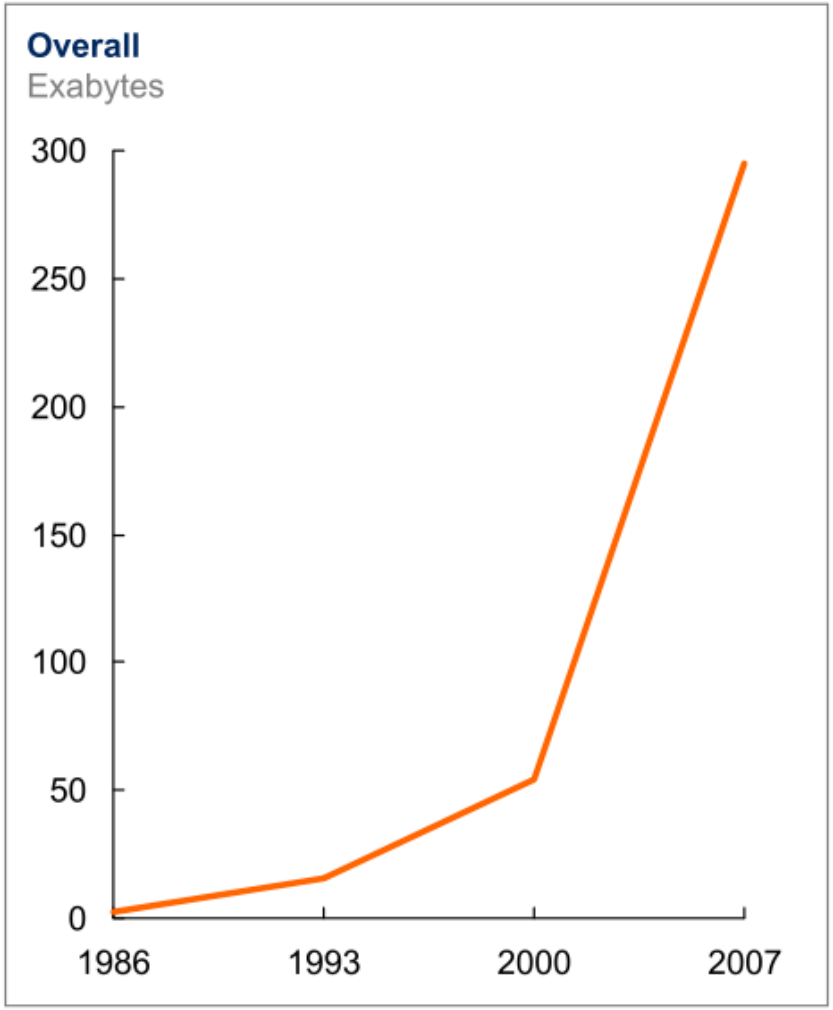
Hol van ennyi adat?

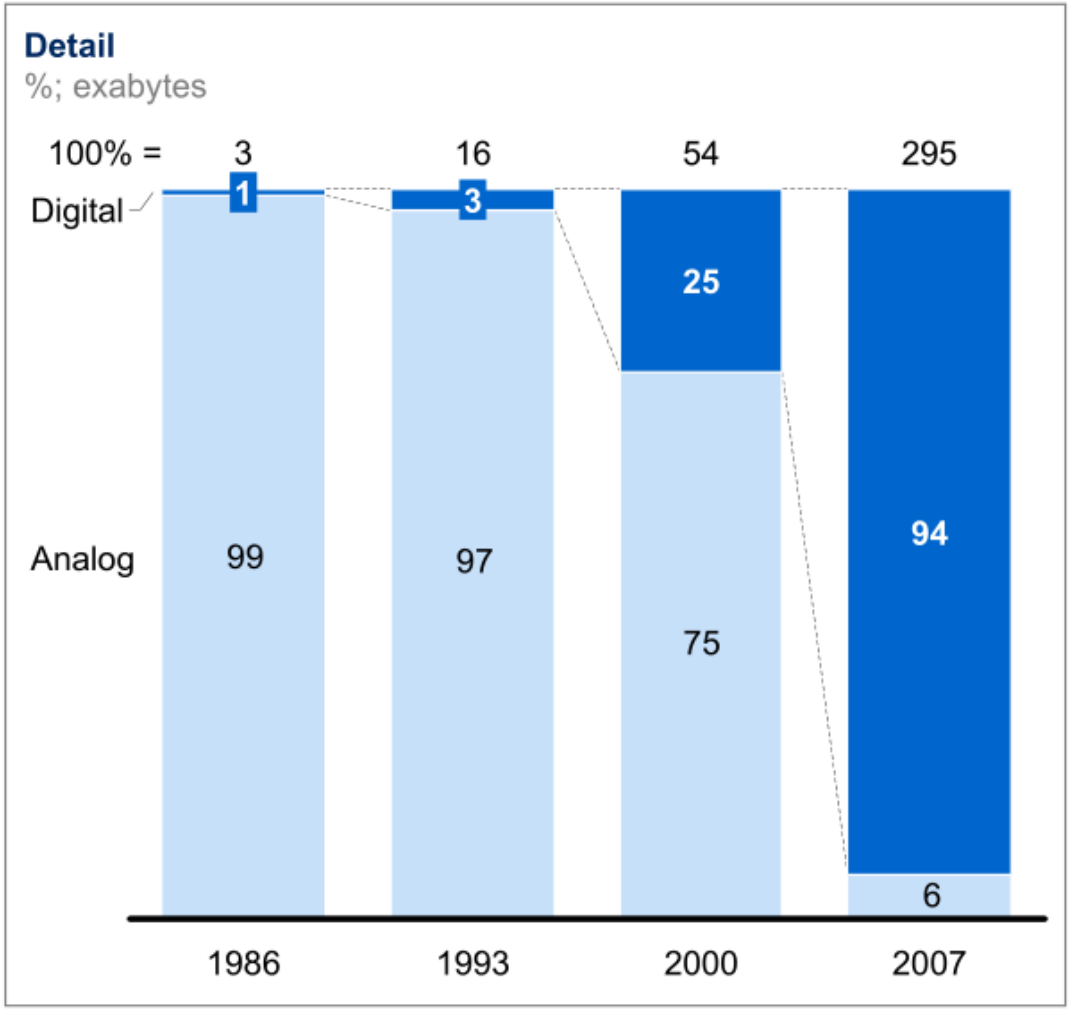
- Időben/populáción ismétlődő megfigyelések
  - Web logok
  - Telekommunikációs hálózatok
  - **Kis(?)kereskedelemi** üzletmenet
  - Tudományos kísérletek (LHC, neurológia, genomika, …)
  - Elosztott szenzorhálózatok (pl. „smart metering”)
  - Járművek fedélzeti szenzorai
  - Számítógépes infrastruktúrák
  - …
- Gráfok, hálózatok
  - Közösségi szolgáltatások

Hol van ennyi adat?

- Modern repülőgépek: ~10 TB/hajtómű/fél óra
- Facebook: 2.5 milliárd „like” egy nap
- Kollégiumi hálózat: pár GB-nyi Netflow rekord egy csendes hétvégén

Tárolási kapacitás a világon [1]

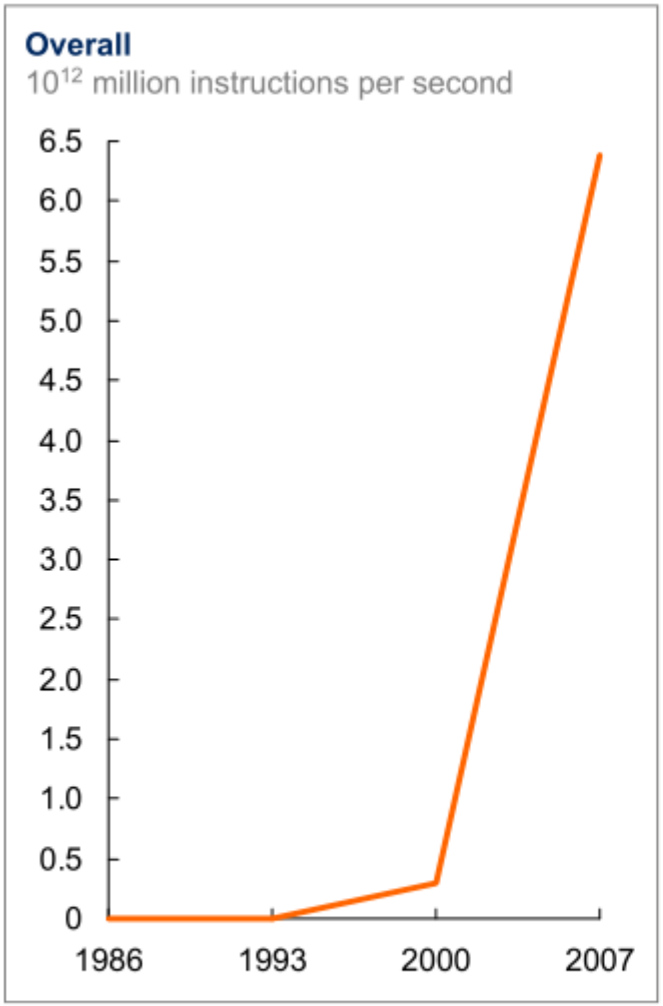


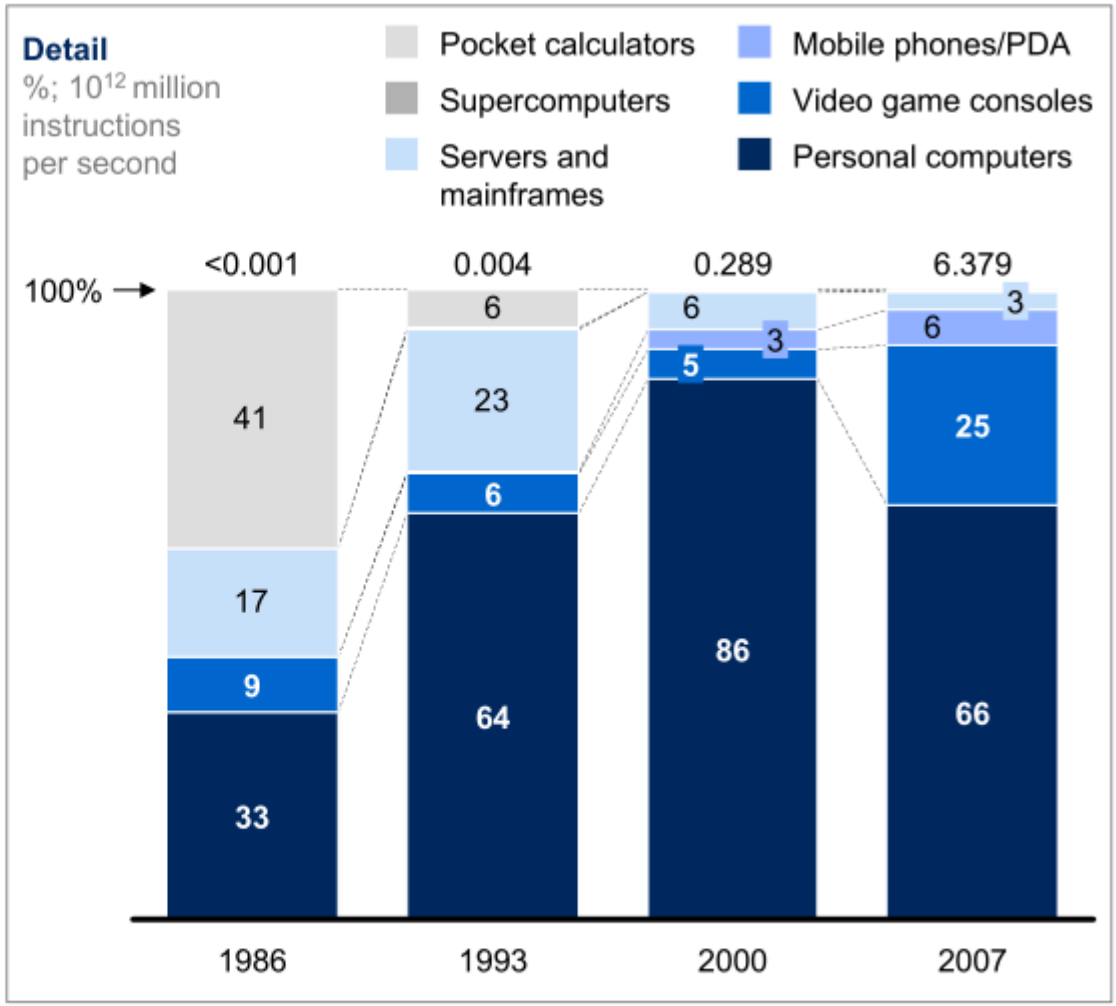


NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Számítási kapacitás a világon [1]

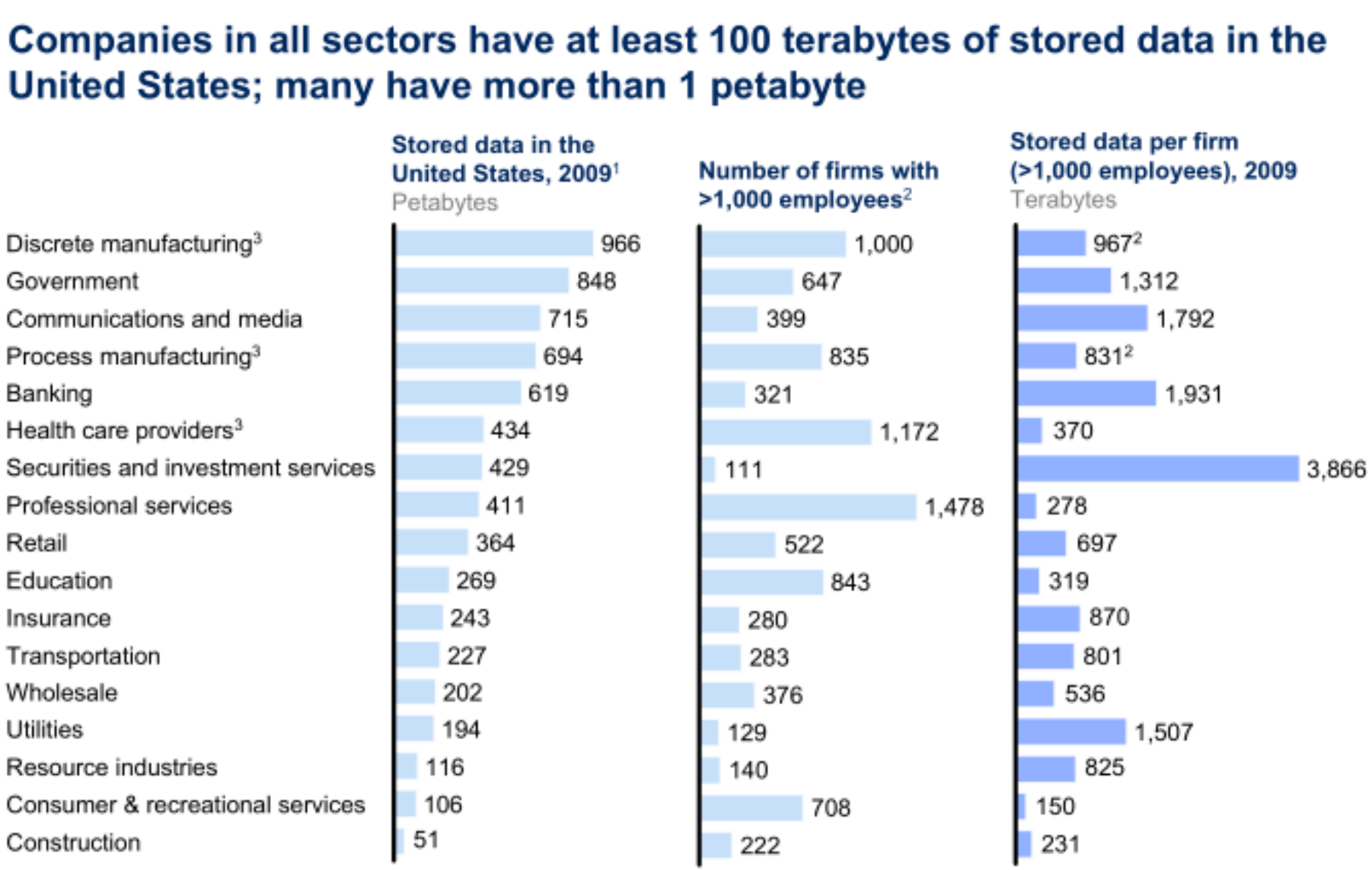




NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Nagyvállalatok által tárolt adatok [1]

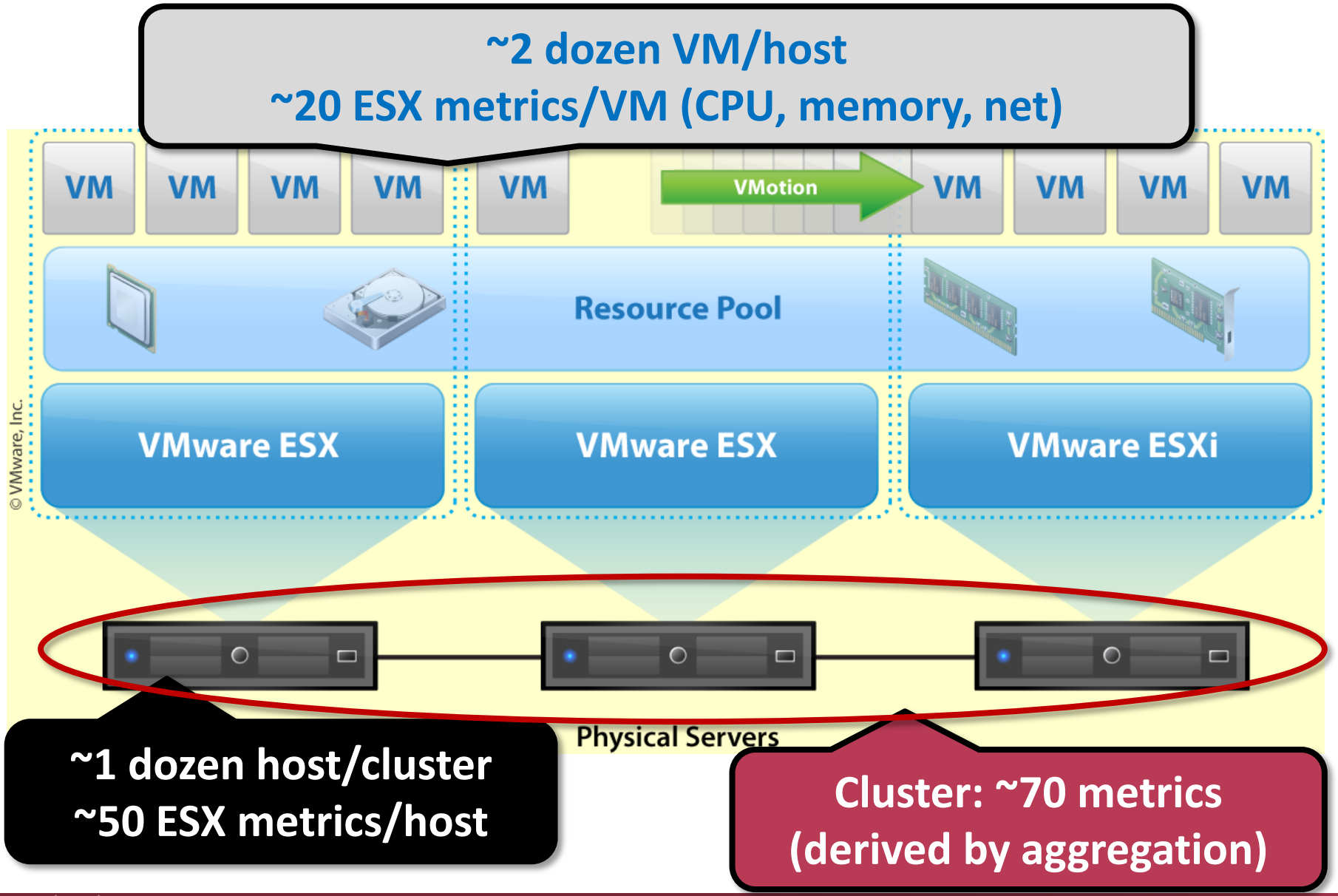


Mit kezdjünk ennyi adattal?

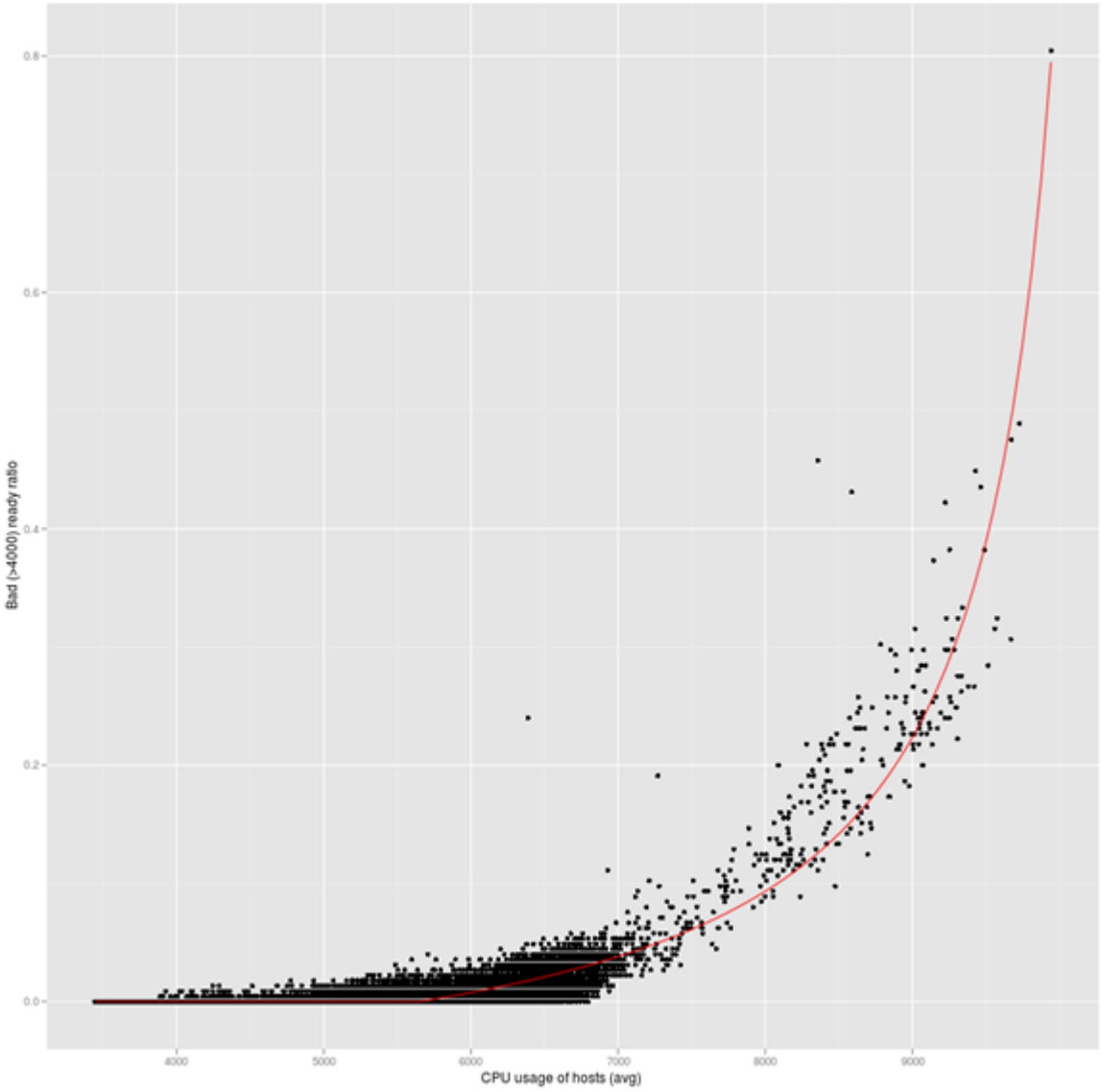
- Üzletmenet
  - Működési metrikák, előrejelzés, adatbányászat
- Szenzor-adatok
- ‚IT for IT'
  - loganalízis, diagnosztika, hibaelőrejelzés, kapacitásmenedzsment, …
- Közösségi média elemzése
  - Pl. PeerIndex
- Csalásfelderítés (fraud detection)
  - ‚Ki vesz jegygyűrűt hajnal 4-kor?'
  - N.B. ritka események; az algoritmika részben újszerű
- IBM Big Data Success Stories [8]

- …

Virtual Desktop Infrastructure: kapacitástervezés



Példa: kapacitástervezés



Alternatív definíció: Big Data jellemzők [2]

- **‚Volume'**: igen nagy mennyiségű adat
- **‚Variety'**: nagyszámú forrás és/vagy nemstrukturált/részben strukturált adatok
- **‚Velocity'**: a ‚Return on Data' (ROD) a lassú feldolgozással csökken
  - Főleg ‚streaming' problémáknál
  - Ellentéte: ‚at rest' Big Data problémák
- **‚Veracity'**: nagymennyiségű zaj
  - Pl. Twitter ‚spam'

De miért nem RDBMS (+SQL)?

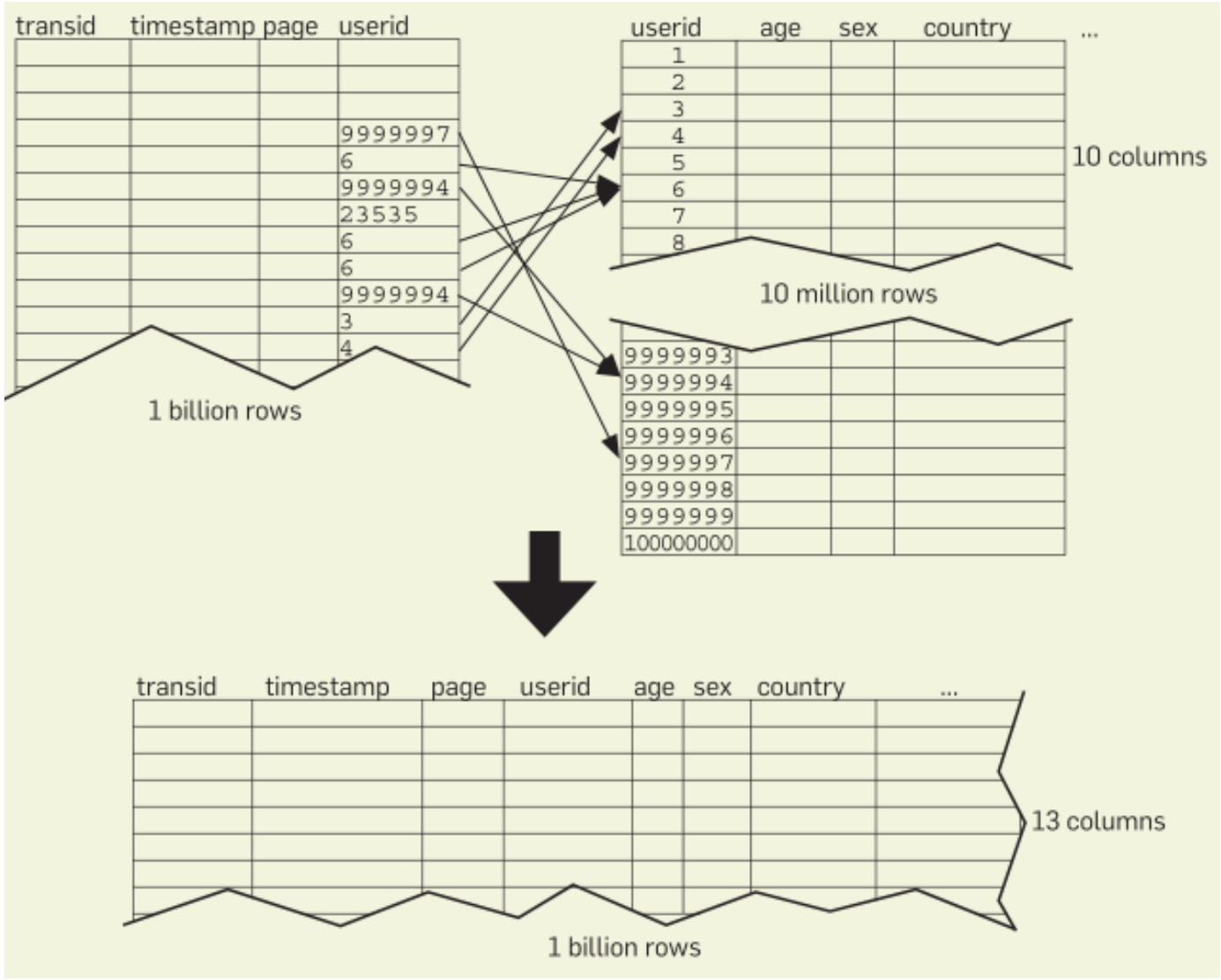
Miért nem RDBMS? Például…

- ‚Big Data’ problémáknál általában létezik természetes (részleges) rendezési szempont
  - Természetes: a nemtriviális analízisek ebben a sorrendben működnek
  - Pl. idő (idősor-analízis)

- Relációs modell: sorok sorrendje?
- Következmény: véletlenszerű hozzáférés diszkről
- Az „optimális” hozzáférési mintához képest lassú

- Mint létni fogjuk, ingyenebéd persze nincs.

A normalizált séma *igen* lassú lehet… [3]



Nagyvállalati adattárházak?

- Jellemzően igen komoly ETL
- „Válaszidő”-követelmények
  - Régi adatok aggregálása/törlése/archiválása
- Strukturálatlan adatok nem jellemzőek
- Drágák…

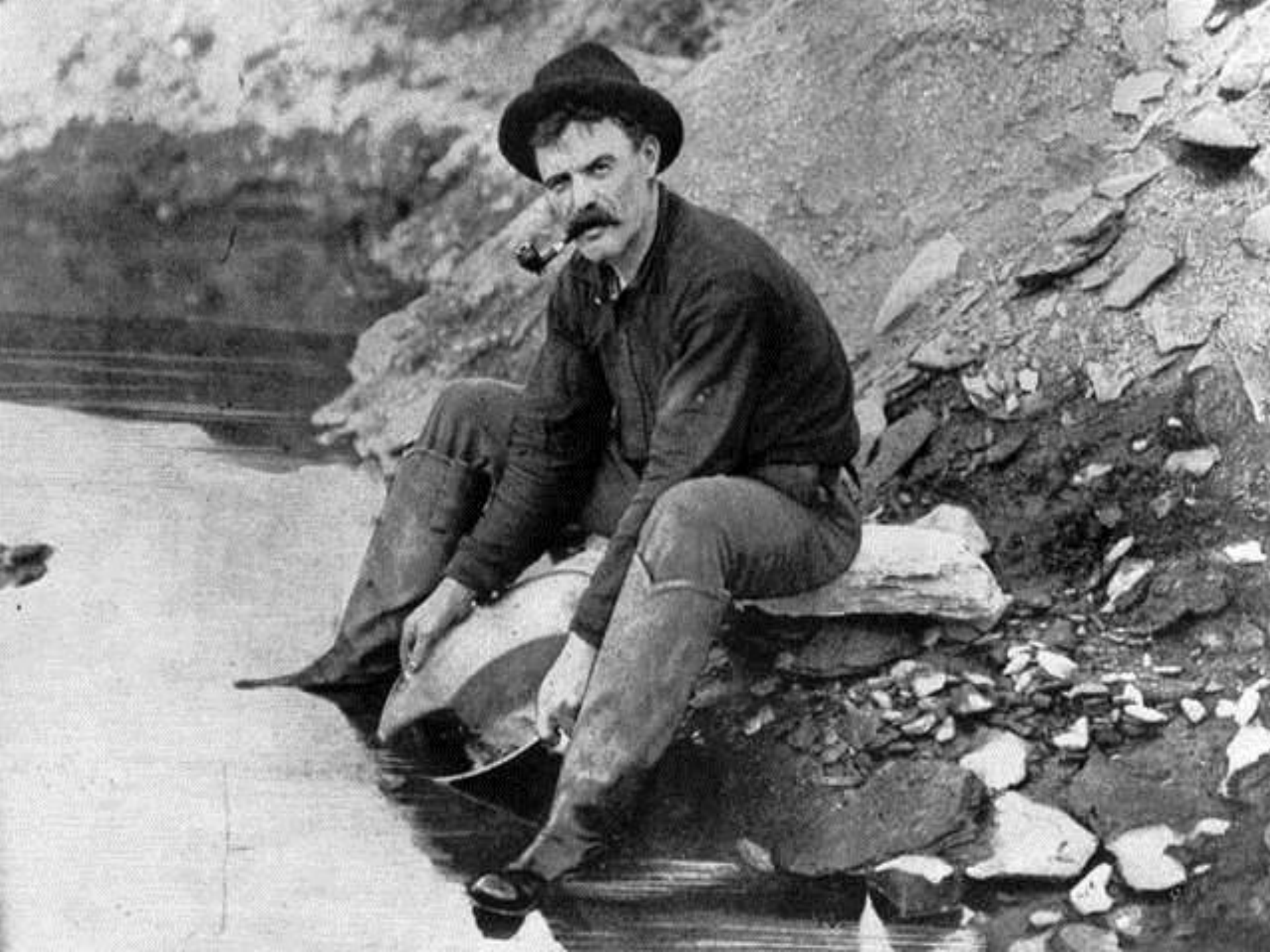
- Nem lehet későbbi analízisre „leborítani” az adatokat

Analízis eszközök?

- Példa: R
  - De lehetne SPSS, SAS, h.d. Excel is
- Kulcsrakész függvények mediántól a neurális hálókig
- De: csak memóriában tárolt adattípusok, nem hatékony memóriakezelés

Vizualizáció?

- A klasszikus megoldások erősen támaszkodnak létező tárolási és analízis-megoldásokra
- Jellemzően statisztikai leképezések
  - Önmagában Big Data problémára vezethető vissza
- Feltáró adatanalízis (EDA): GPU támogatás?



Big Data probléma

- „At rest Big Data”
  - Nincs update
  - „Mindent” elemzünk
- Elosztott tárolás
- **„Computation to data”**

„Not true, but a very, very good lie!”
(T. Pratchett, Nightwatch)

Elosztott számítástechnika

- Big Data: a ma alkalmazott stratégia COTS elosztott rendszerek alkalmazása
  - Kivételek vannak; lásd IBM Netezza
- 8 db nyolcmagos gép jóval olcsóbb, mint egy 64 magos
- Modern hálózati technológiák:
  - Memóriánál lassabb
  - Helyi diszk áteresztőképességénél/válaszidejénél nem feltétlenül!
- A *tárolás* és a *feldolgozás* is elosztott
  - Lehetőleg egy helyen legyen azért

Felhő számítástechnika

A „számítási felhők” egy modell, amely lehetővé teszi a hálózaton keresztül való, kényelmes és széles körű hozzáférést konfigurálható számítási erőforrások egy megosztott halmazához.

Amazon Web Services

Compute

Amazon Elastic Compute Cloud (EC2)

Amazon Elastic MapReduce

Auto Scaling

Content Delivery

Amazon CloudFront

Database

Amazon SimpleDB

Amazon Relational Database Service (RDS)

Deployment & Management

AWS Elastic Beanstalk

AWS CloudFormation

E-Commerce

Amazon Fulfillment Web Service (FWS)

Messaging

Amazon Simple Queue Service (SQS)

Amazon Simple Notification Service (SNS)

Amazon Simple Email Service (SES)

Monitoring

Amazon CloudWatch

Networking

Amazon Route 53

Amazon Virtual Private Cloud (VPC)

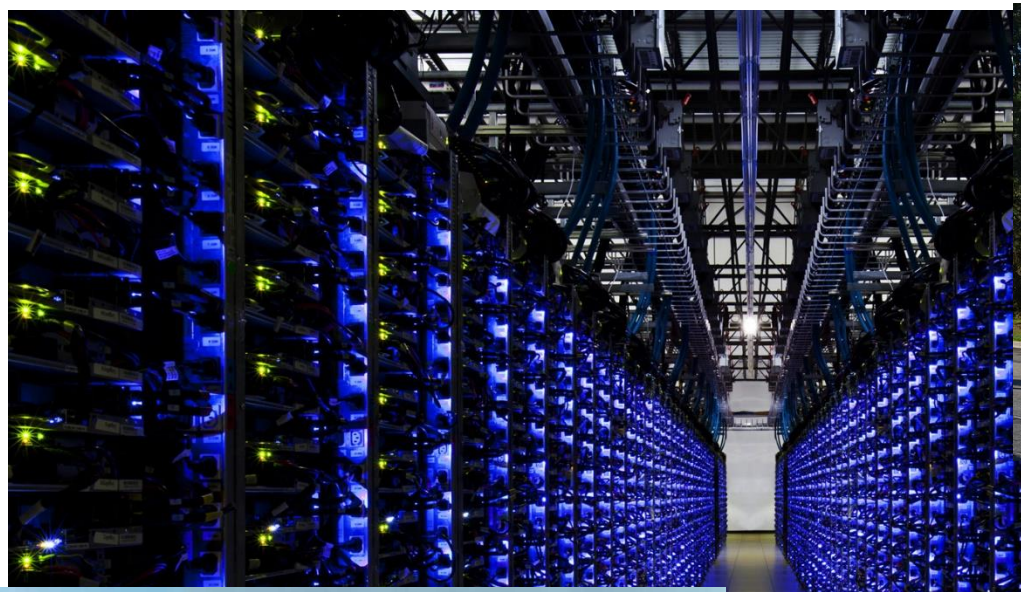
Elastic Load Balancing

Payments & Billing

Amazon Flexible Payments Service (FPS)

Amazon DevPay

Szolgáltatói oldalon…



Alapvető kérdések

- Elosztott platformon párhuzamosítás szükséges
- Hatékony feldolgozáshoz továbbra is referenciális lokalitás kell
- Bár a feldolgozás „közel vihető az adathoz”, az adatterítés logikája befolyásolja a teljesítményt
  - Pl. csak egy csomópont dolgozik

Node 1

timestamp	sensor	reading
19990101000000	1	
19990101000015	1	
19990101000030	1	
⋮	⋮	⋮
20081231235930	1	
20081231235945	1	
19990101000000	2	
19990101000015	2	
19990101000030	2	
⋮	⋮	⋮
20081231235930	2	
20081231235945	2	
19990101000000	3	
⋮	⋮	⋮
20081231235945	100	

Node 1

timestamp	sensor	reading
19990101000000	1	
19990101000000	2	
19990101000000	3	
⋮	⋮	⋮
19990101000000	1000	
19990101000015	1	
19990101000015	2	
19990101000015	3	
19990101000015	4	
⋮	⋮	⋮
19990101000015	1000	
19990101000030	1	
19990101000030	2	
⋮	⋮	⋮
19991231235945	100	

Node 2

timestamp	sensor	reading
19990101000000	101	
19990101000015	101	
19990101000030	101	
⋮	⋮	⋮
20081231235930	101	
20081231235945	101	
19990101000000	102	
19990101000015	102	
19990101000030	102	
⋮	⋮	⋮
20081231235930	102	
20081231235945	102	
19990101000000	103	
⋮	⋮	⋮
20081231235945	200	

Node 2

timestamp	sensor	reading
20000101000000	1	
20000101000000	2	
20000101000000	3	
⋮	⋮	⋮
20000101000000	1000	
20000101000015	1	
20000101000015	2	
20000101000015	3	
20000101000015	4	
⋮	⋮	⋮
20000101000015	1000	
20000101000030	1	
20000101000030	2	
⋮	⋮	⋮
20001231235945	1000	

Big Data == Hadoop?

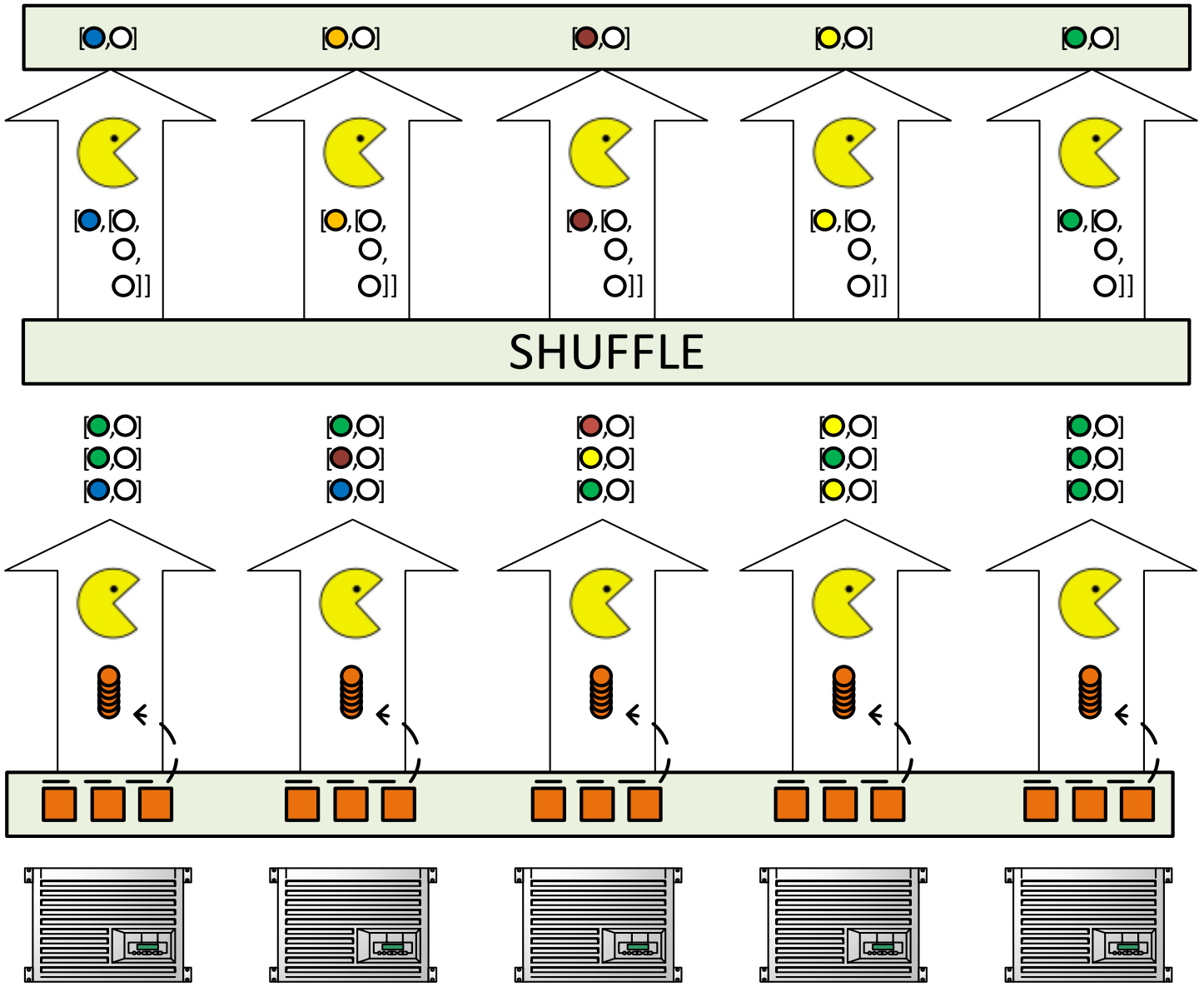
- Google MapReduce és GFS → Apache Hadoop
- Nyílt forráskódú, Java alapú keretrendszer
- Hadoop Distributed File System (HDFS)
- MapReduce programozási paradigma
- Ráépülő/kiegészítő/kapcsolódó projektek: Cassandra, Chukwa, Hbase, Hive, Mahout, Pig, ZooKeeper…

MapReduce

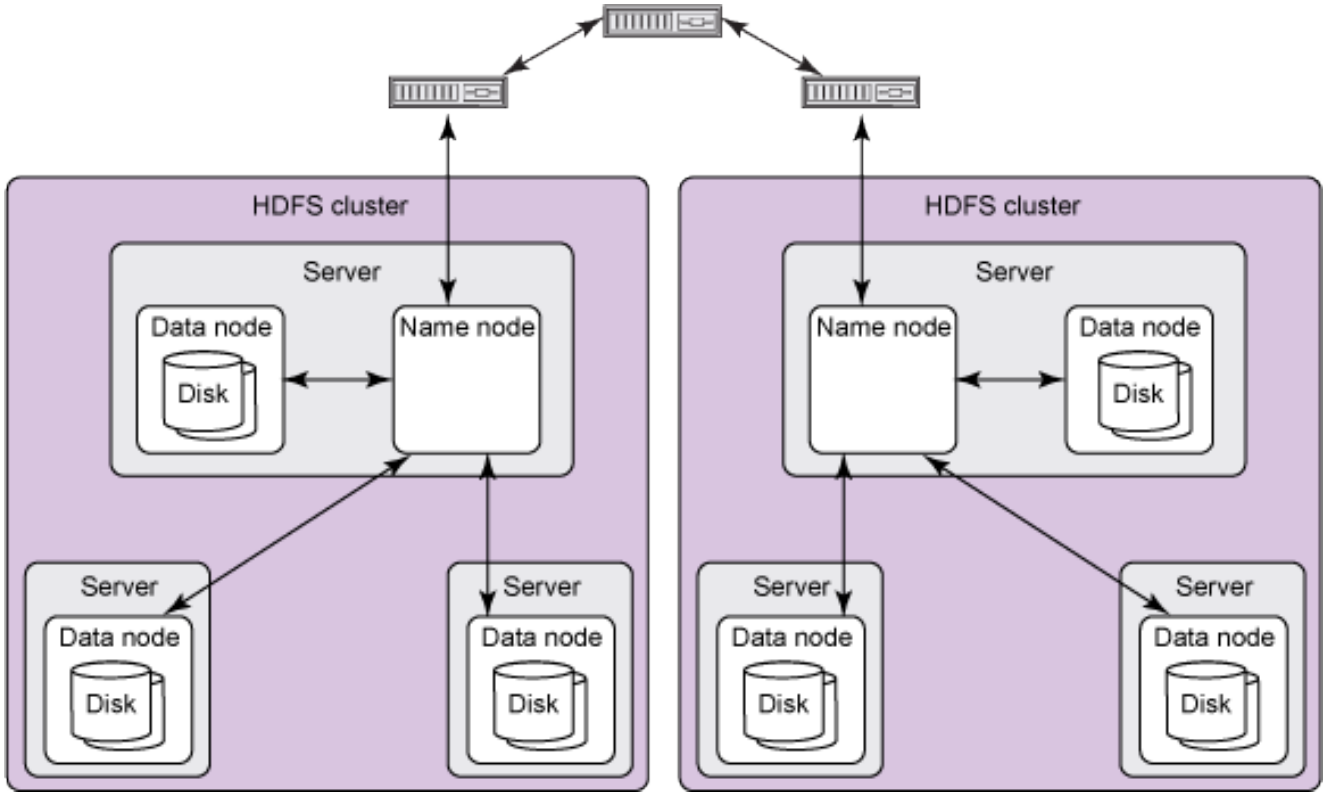
„Reduce”

„Map”

Distributed File System

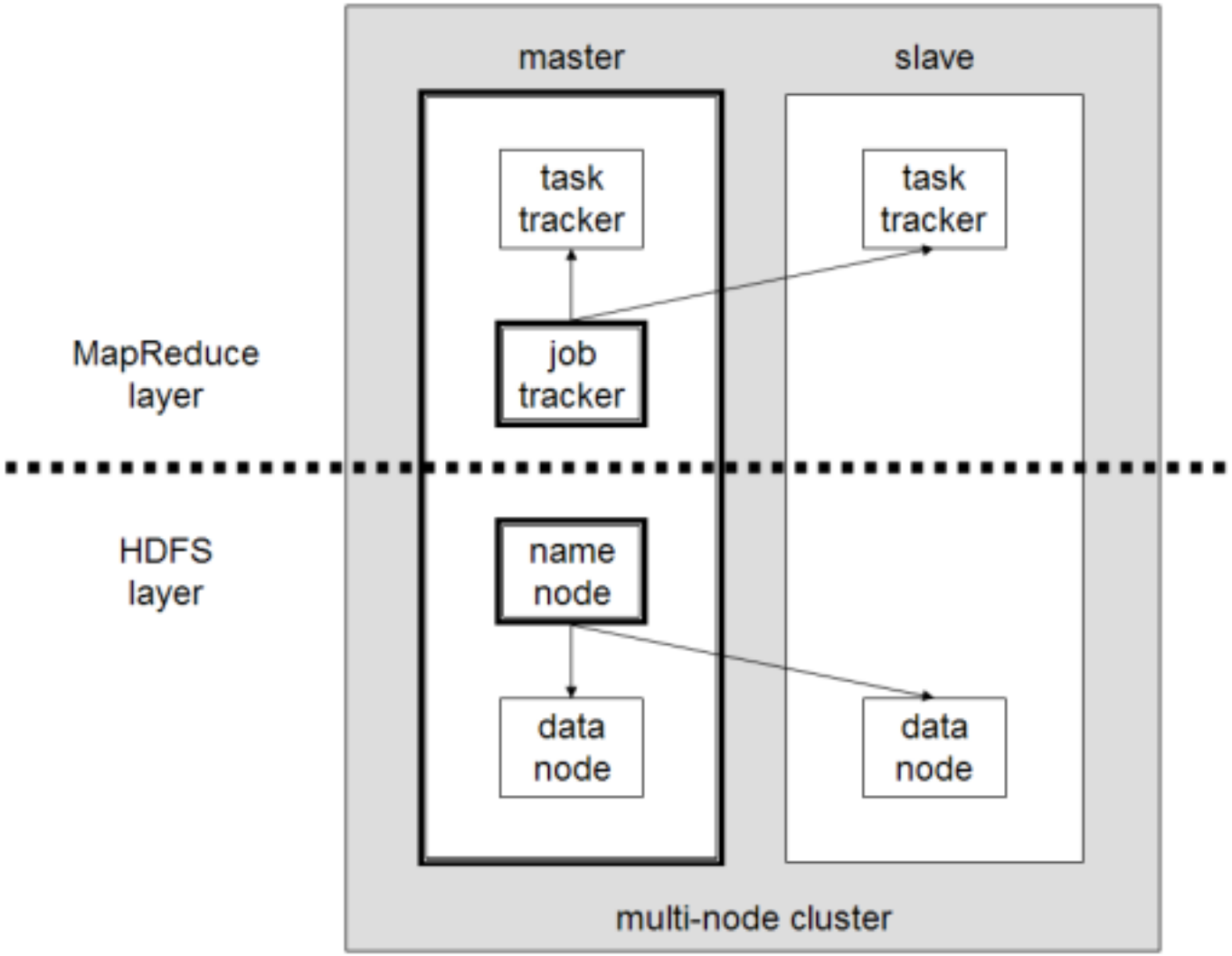


HDFS

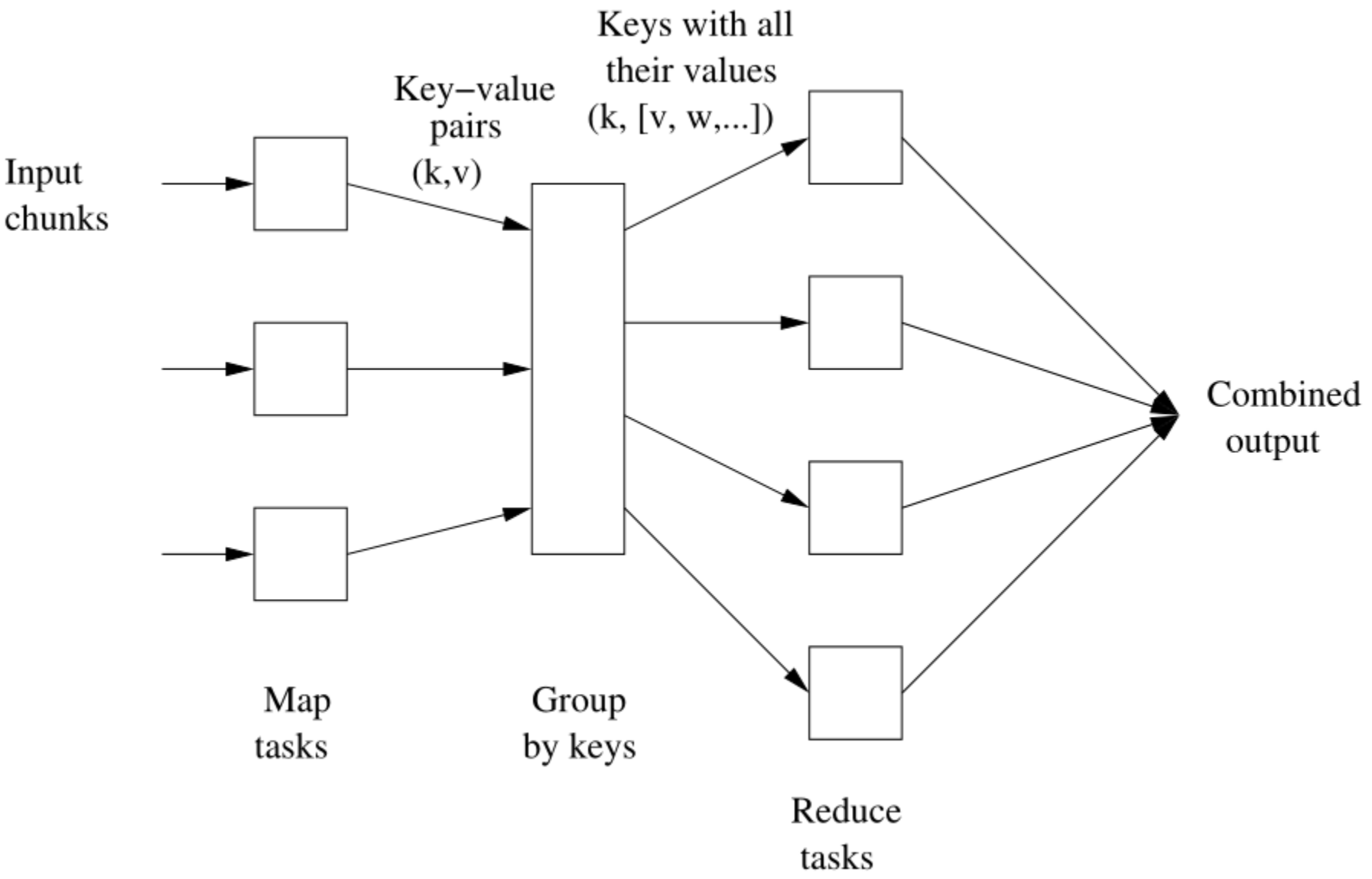


~Klasszikus állományrendszer
Nagy (64MB) blokkok, szétterítve és replikálva

Hadoop


master
slave
task tracker
task tracker
job tracker
MapReduce layer
HDFS layer
name node
data node
data node
multi-node cluster

MapReduce [6]


Input chunks
Key–value pairs (k,v)
Keys with all their values (k, [v, w,...])
Combined output
Map tasks
Group by keys
Reduce tasks

MapReduce: szavak számolása szövegben [7]

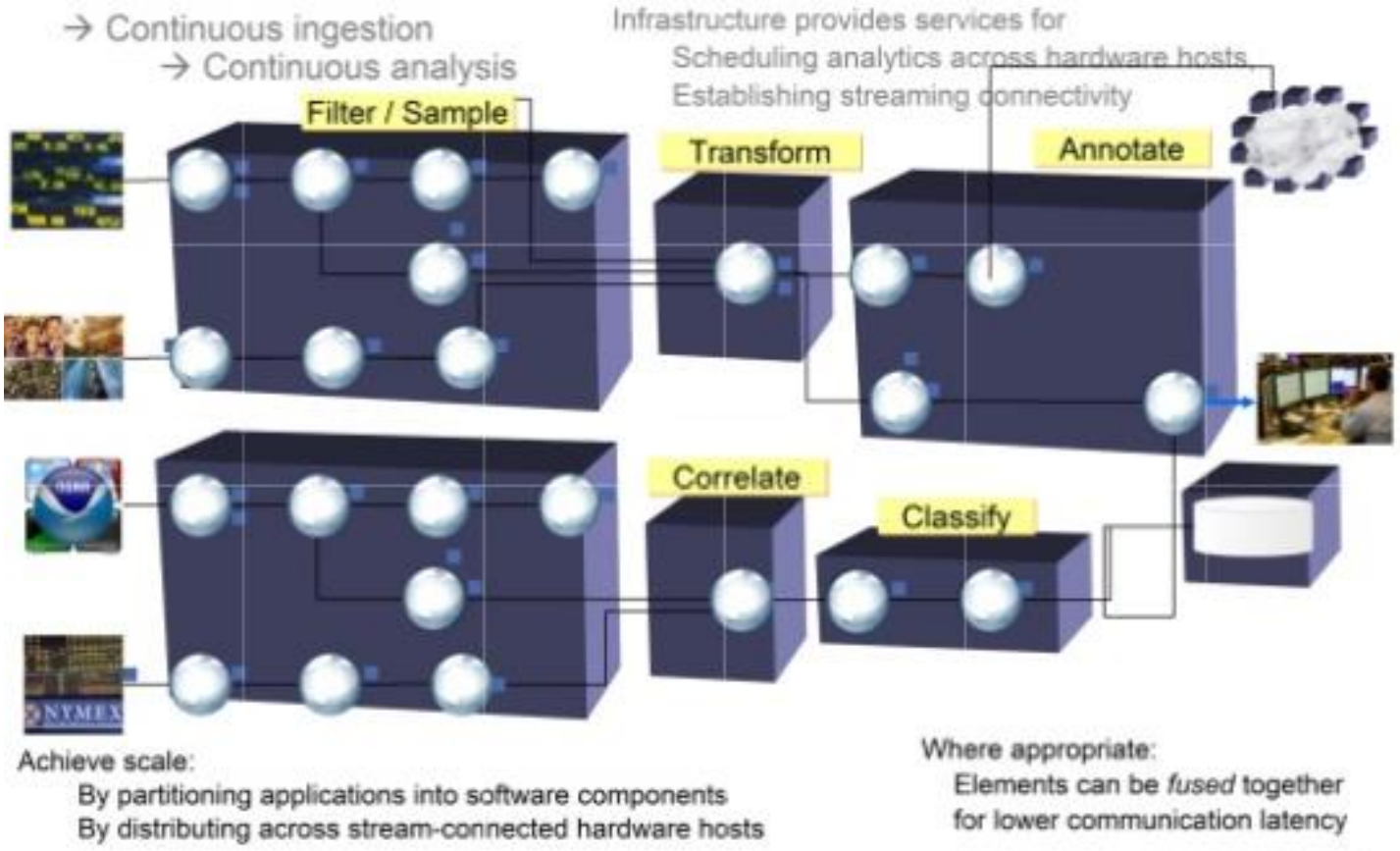
```
map(String input_key, String input_value):
  // input_key: document name
  // input_value: document contents
  for each word w in input_value:
    EmitIntermediate(w, "1");


reduce(String output_key, Iterator intermediate_values):
  // output_key: a word
  // output_values: a list of counts
  int result = 0;
  for each v in intermediate_values:
    result += ParseInt(v);
  Emit(AsString(result));
```

MapReduce, mint párhuzamosítási minta

- Számos probléma jól megfogalmazható MapReduce szemléletben
  - Mátrix-mátrix és mátrix-vektor szorzás
  - Relációalgebra
  - Korreláció
  - …
- Ezekről később beszélünk
  - Sokat ☺

Big Data =/= Hadoop (ökoszisztéma)

- Adatfolyamok!
  - Hadoop (**eredetileg**): batch & ‚at rest'



Big Data =/= Hadoop (ökoszisztéma)

- Elemző eszközök kiterjesztései
  - ‚File backed'
  - Adatbázis-integrált
  - Vitatható, hogy ‚igazi' Big Data-e
- Célhardver
  - IBM Netezza
- Gráfproblémák kezelése
  - Nem csak paraméterbecslés és tulajdonságvizsgálat; mintaillesztés is

Tentatív tematika kivonata

- Adatelemzési alapozás
- R
- Felderítő adatelemzés
- MapReduce algoritmika
- Mintavételezés
- Gépi tanulás (szemelvények)
- Folyamfeldolgozás
- ZH
- Beszámoló-előadások

Lehetőségek [1]

Illetve: tessék körbenézni Budapesten.

Források

- [1] Manyika, J., Chui, M., Brown, B., & Bughin, J. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Retrieved from http://www.citeulike.org/group/18242/article/9341321
- [2] Zikopoulous, P., Deroos, D., Parasuraman, K., Deutsch, T., Corrigan, D., & Giles, J. (2013). *Harness the Power of Big Data*. McGraw-Hill. Retrieved from http://medcontent.metapress.com/index/A65RM03P4874243N.pdf
- [3] Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, *52*(8), 36. doi:10.1145/1536616.1536632
- [4] http://www.ibm.com/developerworks/library/wa-introhdfs/
- [5] Borkar, V., Carey, M. J., & Li, C. (2012). Inside "Big Data management." In *Proceedings of the 15th International Conference on Extending Database Technology - EDBT '12* (pp. 3–14). New York, New York, USA: ACM Press. doi:10.1145/2247596.2247598
- [6] Rajaraman, A., & Ullman, J. D. (2011). *Mining of Massive Datasets*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139058452
- [7] http://research.google.com/archive/mapreduce-osdi04-slides/index.html
- [8] IBM Big Data Success Stories. ftp://ftp.software.ibm.com/software/data/sw-library/big-data/ibm-big-data-success.pdf