Big Data eszközök

Emődi Márk

emodi.mark@sztaki.hu

Bemutatkozás

Emődi Márk

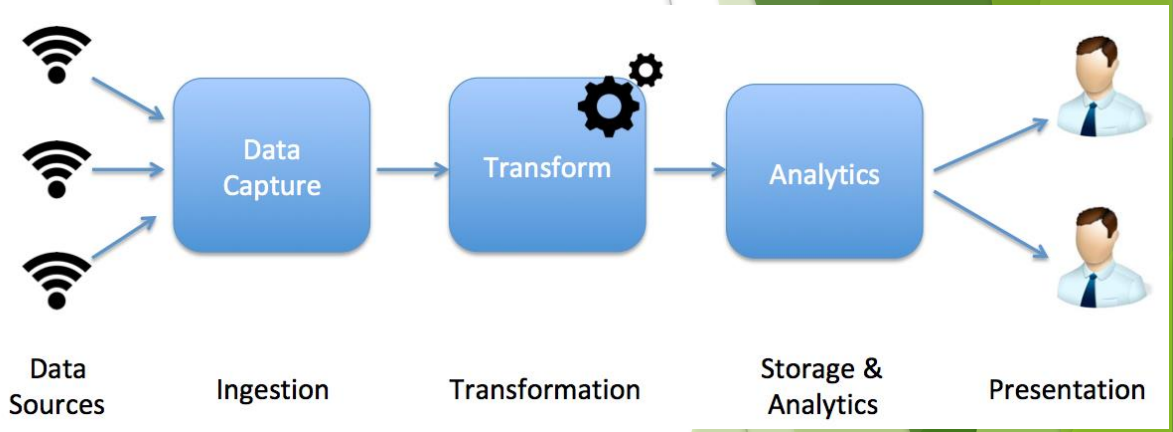
SZTAKI PERL – Szoftverfejlesztő, kutató

ÓE NIK – Tanszéki mérnök/doktorandusz

emodi.mark@sztaki.hu

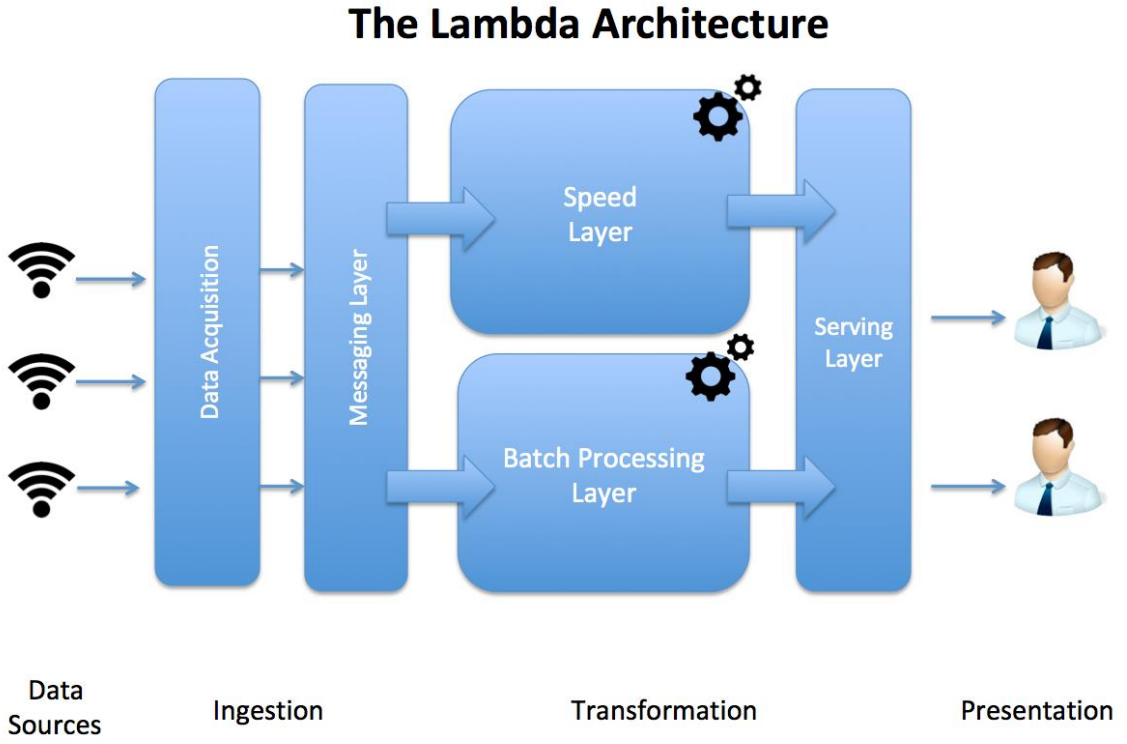
Big Data rendszerekkel szembeni elvárások

- Követelmény:
  - Adatokon végzett utasítások
    - Átalakítás/Elemzés/Vizualizálás
  - Dinamikus erőforrás allokálás igényeknek megfelelően
    - Elosztott működés több erőforrás bevonásával
  - Hiba esetén gyors helyreálló képesség



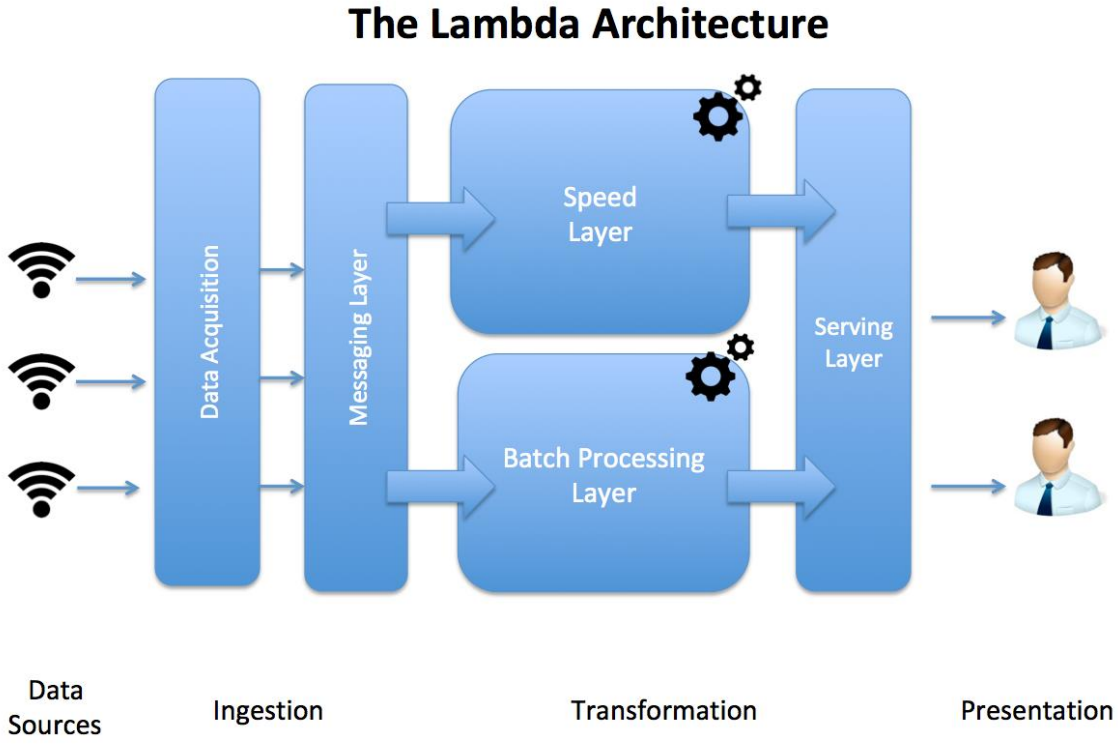
Big Data rendszerekkel szembeni elvárások

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



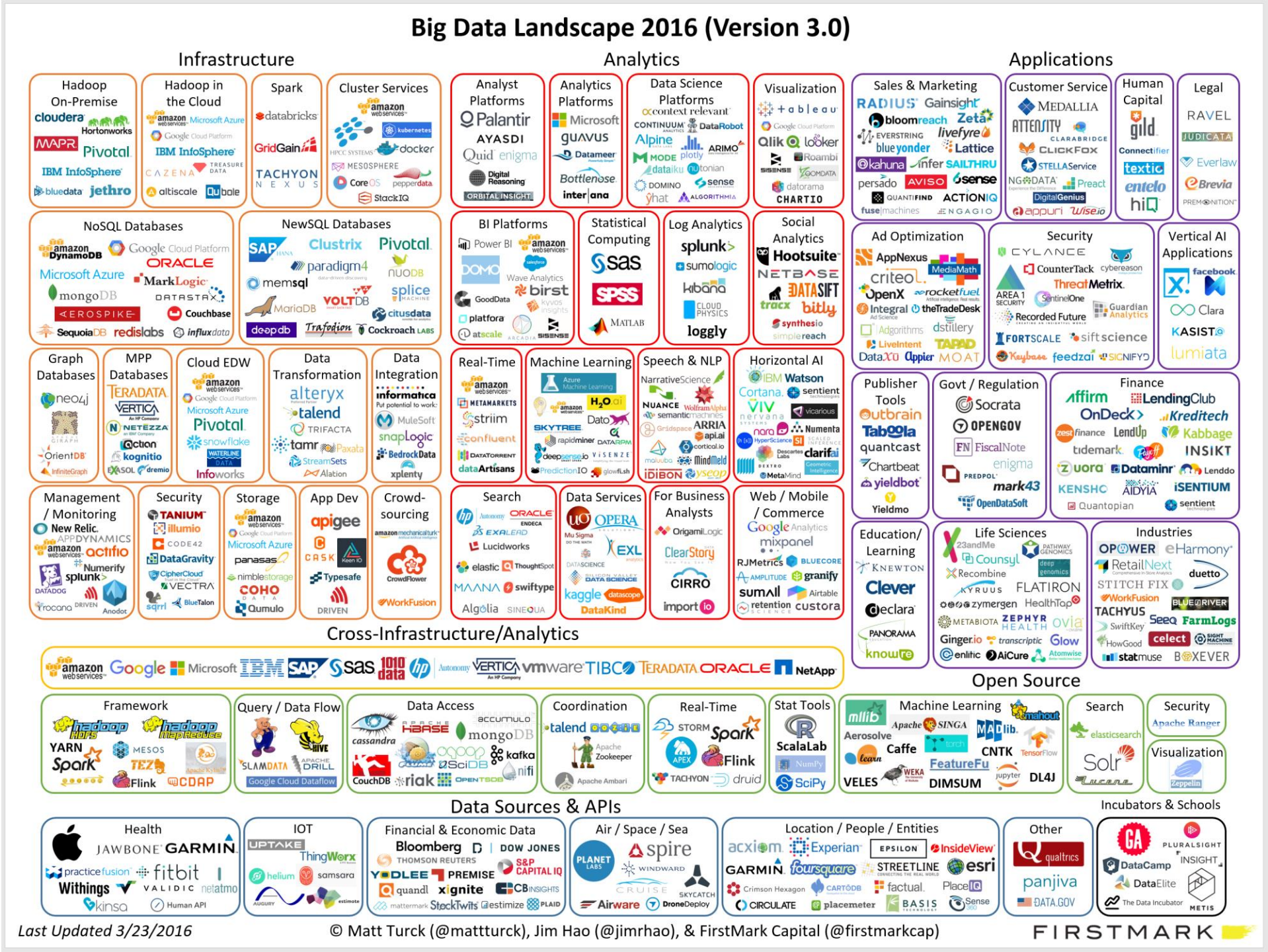
Big Data rendszerekkel szembeni elvárások

- Követelmény:
  - Használati esetek:
    - „Near” real time adatfeldolgozás
      - Nagy mennyiségű bejövő adat gyűjtése
      - Közel valós idejű feldolgozása
      - Adatfolyamon érkező valós idejű adat
      - Pl. szenzoradatok feldolgozása, naplófájlok feldolgozása, események feldolgozása (közösségi média, pénzügyi szektor, ...)



Big Data

- Big Data rendszerek

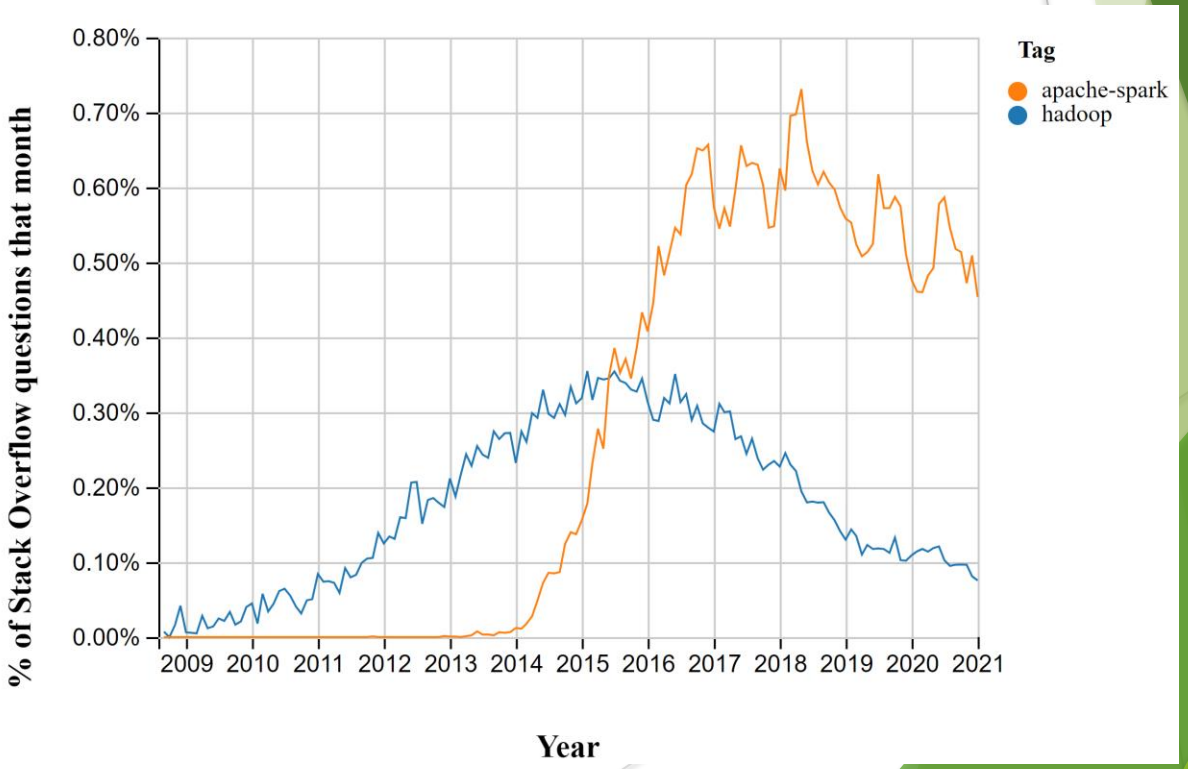


Big Data framework

- Felhasználói kérdések eloszlása az adott hónapban (Stackoverflow)
- 2014 – 2016: megnövekedett felhasználási esetek

  Új igények jelentek meg:

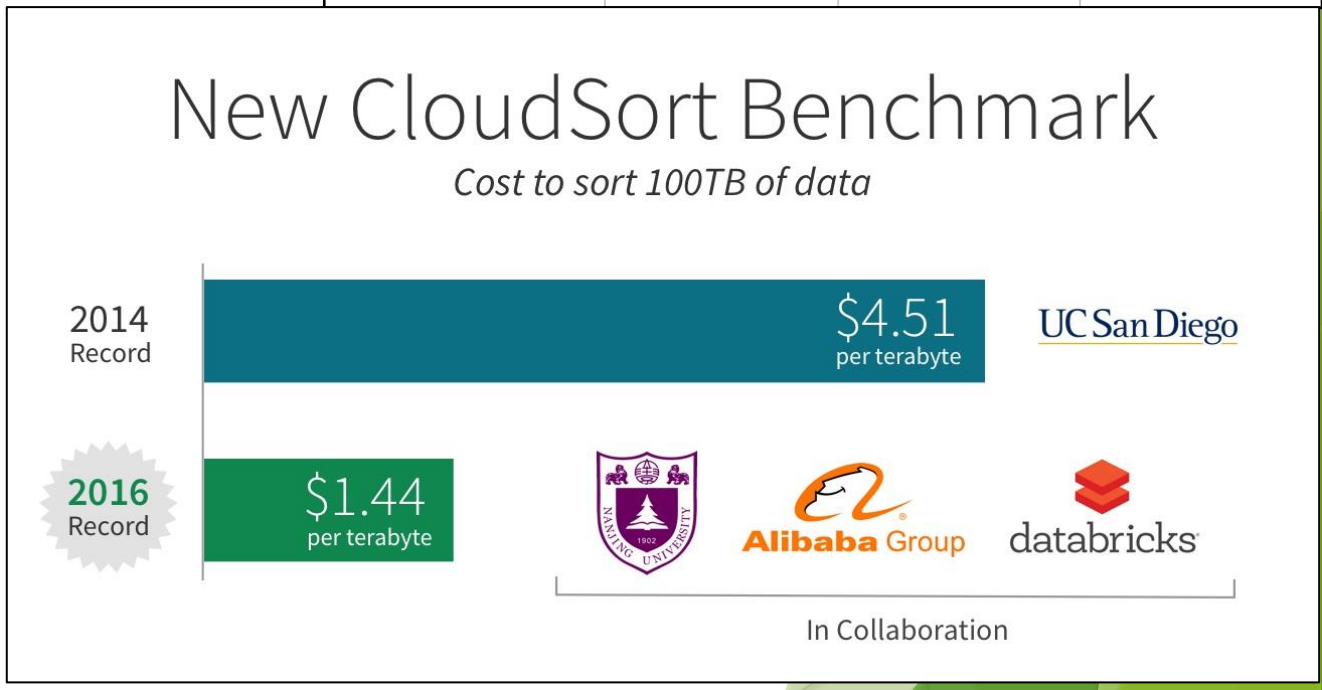
  - Valós idejű adatfeldolgozás (adatfolyamok)
  - Gépi tanulás



Apache Spark

	Hadoop MR Record	Spark Record	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400 physical	6592 virtualized	6080 virtualized
Sort rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Sort rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min

- Nyílt forráskódú
- Általános célú
- Hibatűrő
- Elosztott keretrendszer
- 2014: Új világrekord (2014 Gray Sort)
  - 100 TB adat rendezés
  - 3 x gyorsabb rendezés 10 x kevesebb gépszám mellett
- 2016: Új világrekord hatékonyságban
  - 1,44 $ / TB rendezési költség a korábbi 4,51 $ / TB helyett



https://databricks.com/blog/2016/11/14/setting-new-world-record-apache-spark.html

Apache Spark

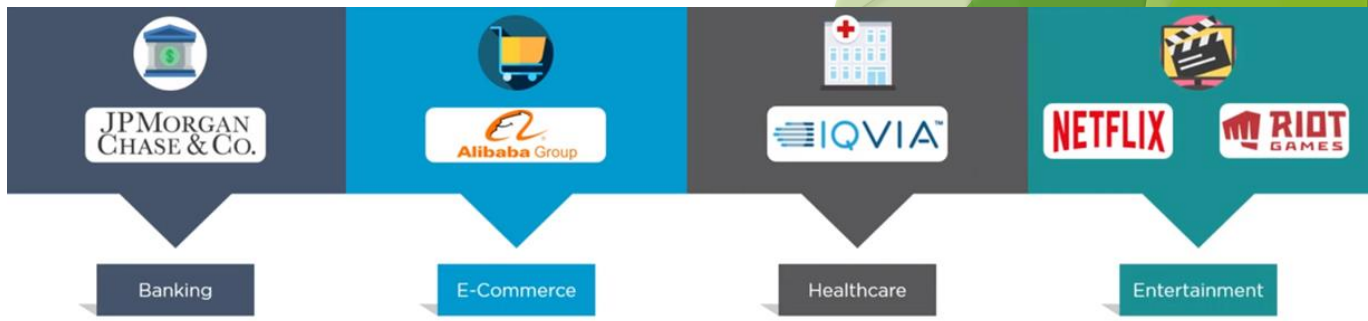
- Adatok memóriából történő használata (szemben a korábban ismertetett Hadoop architektúrával)
- Fejlesztők, elemzők azonos rendszeren tudnak dolgozni
- Széles körben használt (Baidu, eBay, Yahoo, …) (https://spark.apache.org/powered-by.html)

Apache Spark funkcionalitás

- Nagyobb teljesítmény érhető el, mint Hadoop rendszerek alatt
  - pl. Lineáris regresszió alatt 100x növekedés
- Scala, Python, Java és R programozási nyelv támogatás
- Lusta kiértékelés: a kiértékelések késleltetése, amíg nincs rá szükség
- „Valós idejű” feldolgozás alacsony késleltetés mellett
  - Memóriában tárolt adatok gyors hozzáférése
- Széles körű tárolóegység támogatása (HDFS, Apache Cassandra, Apache HBase, Apache Hive, …)
- Gépitanulás támogatása

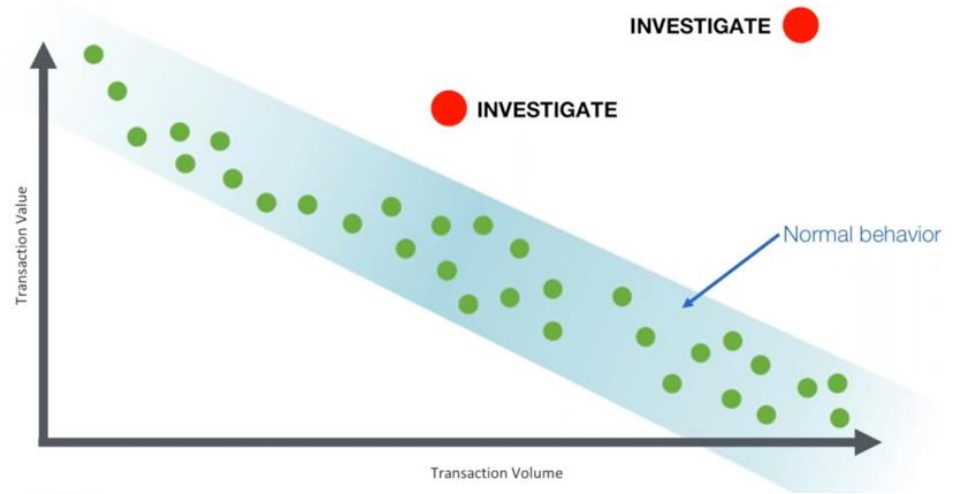
Apache Spark – Ismert felhasználási esetek

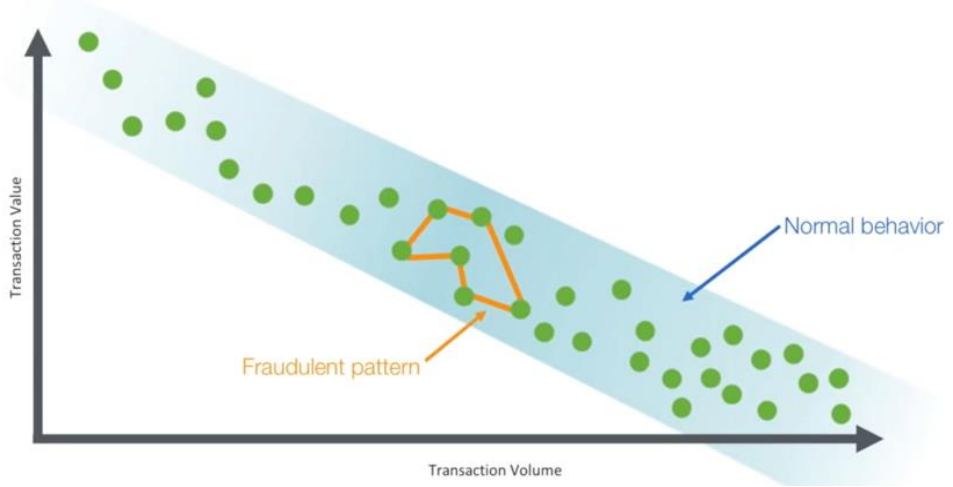
- Közösségi oldalaknál
- Pénzügyi cégeknél (bankok, biztosítók)
- Szórakoztató ipar / kereskedelem
- Egészségügy
  - Lehetséges egészségügyi betegségek felderítése (Betegadatok és gyógyszerfogyasztási előzményekből)
- Tudományos alkalmazás
  - Társadalomtudományi alkalmazás:
    - Szöveg alapú klasszifikáció
  - Biológiai alkalmazás

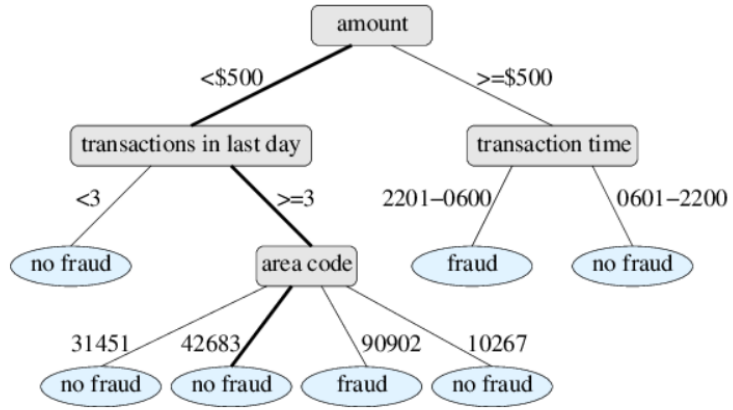


Ismert felhasználási esetek

- A csalás technikái és stratégiái folyamatosan fejlődnek az idő múlásával, ezért a csalók gyakran egy lépéssel a vállalatok előtt járnak
  - Klasszikus megközelítés
    - Az átlagtól eltérő értékek vizsgálata
  - Komplex megközelítés
    - Az átlagos minták vizsgálata, és lehetséges csalások keresése a mintákon belül
- Nagy adathalmaz -> felügyelt gépi tanulás -> döntési fák, random forest







Köszönöm a figyelmet!