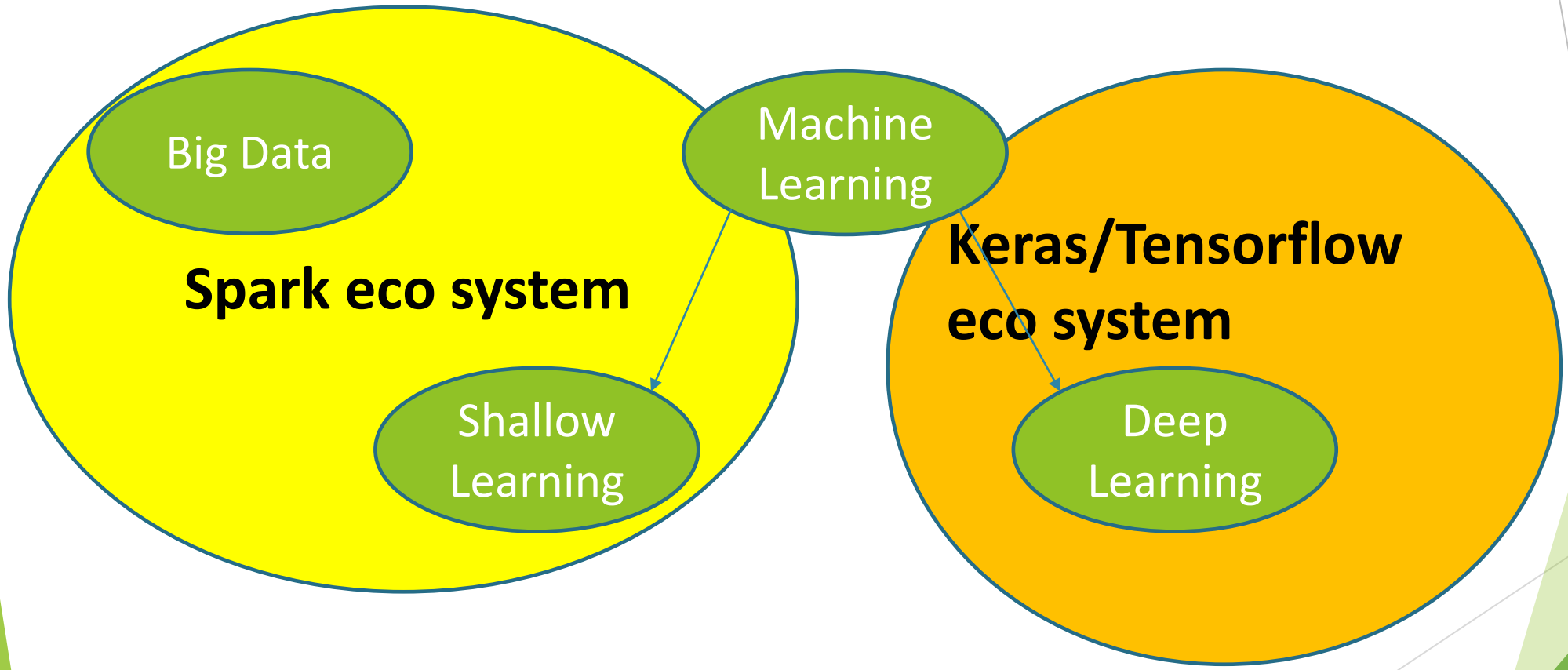
Big Data és ML eszközök, platformok

Farkas Attila

farkas.attila@sztaki.hu

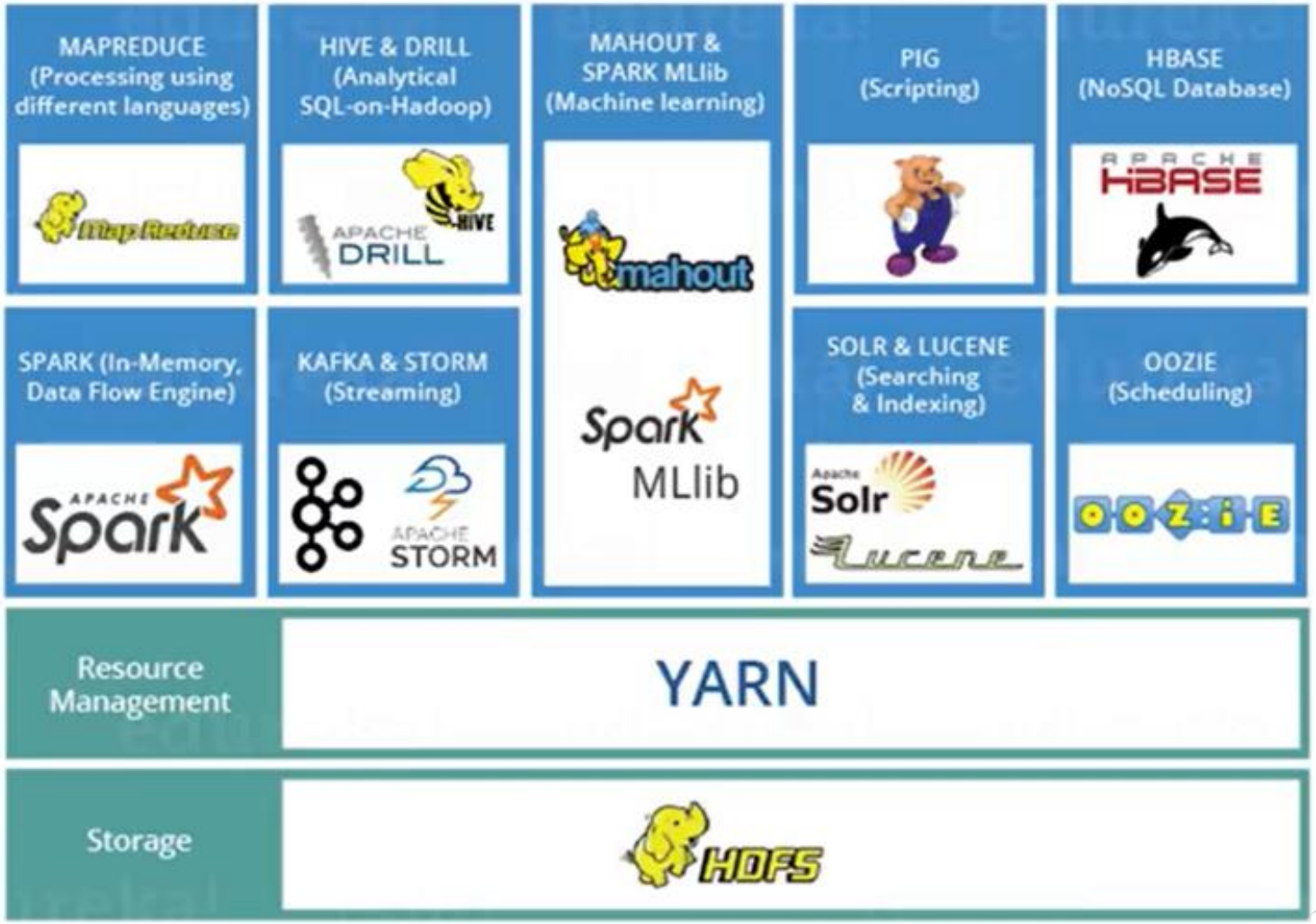
MI fő területei és támogatott keretrendszerek



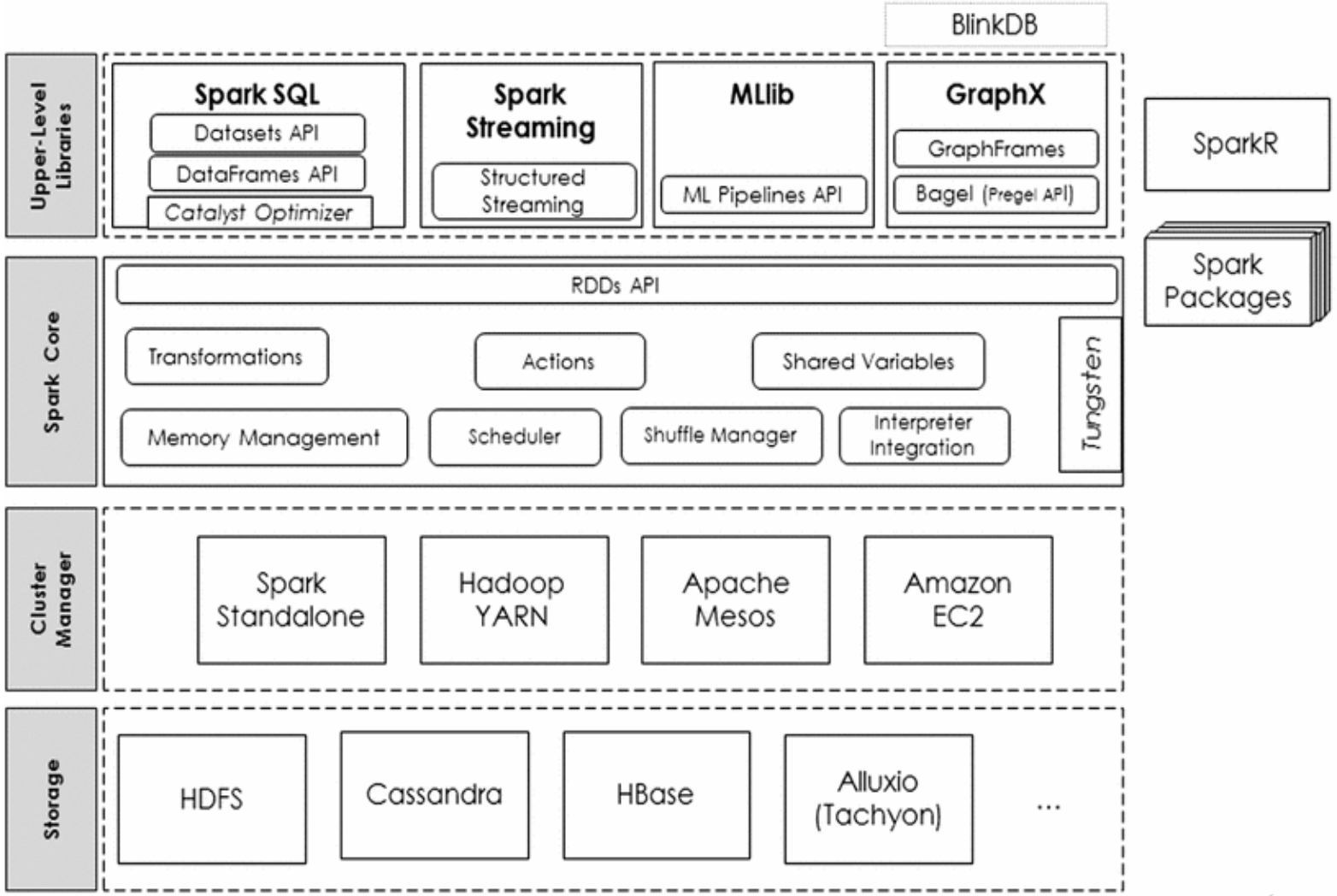
Big Data keretrendszerek

- Hadoop keretrendszer
- Spark keretrendszer

Hadoop keretrendszer



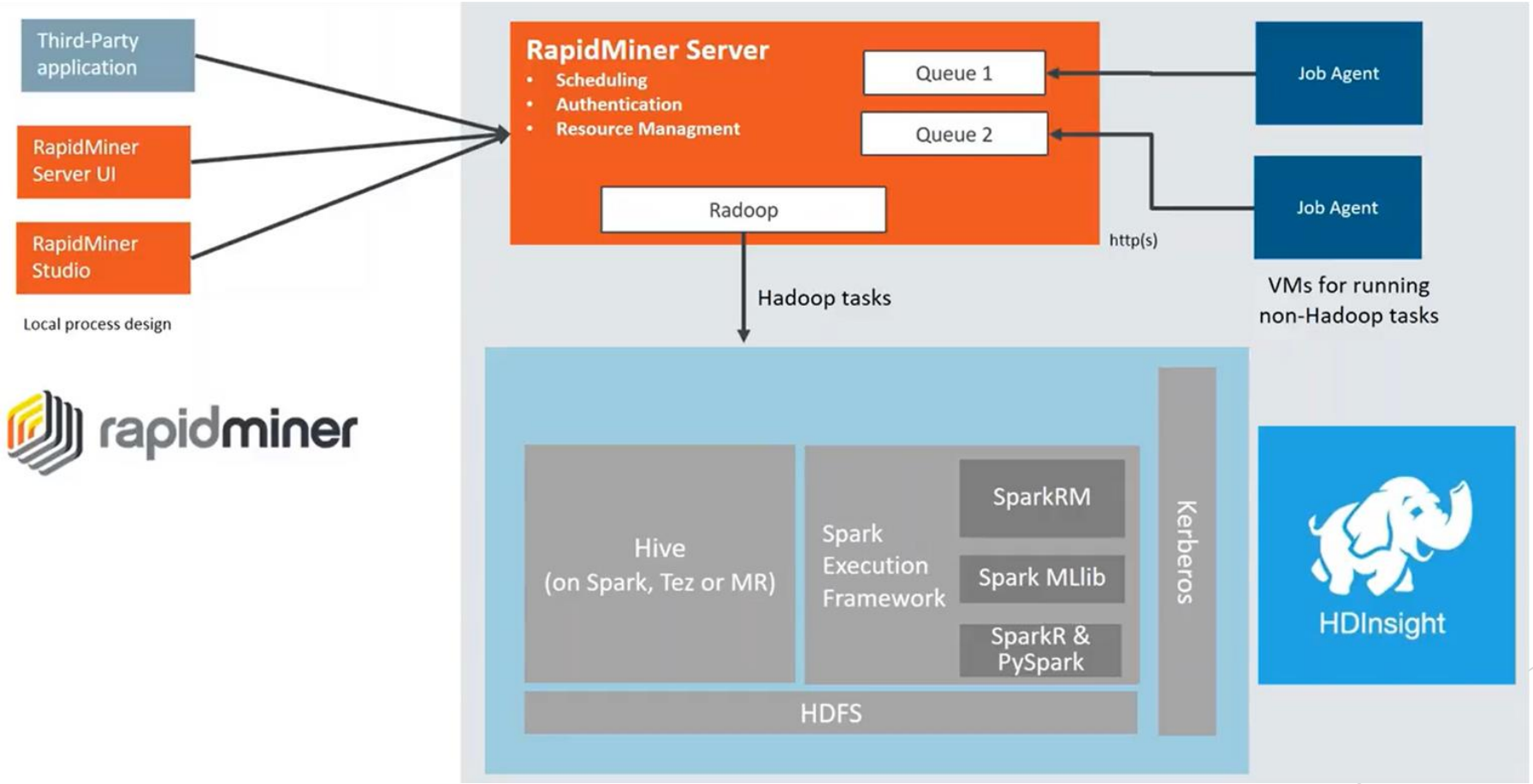
Spark keretrendszer



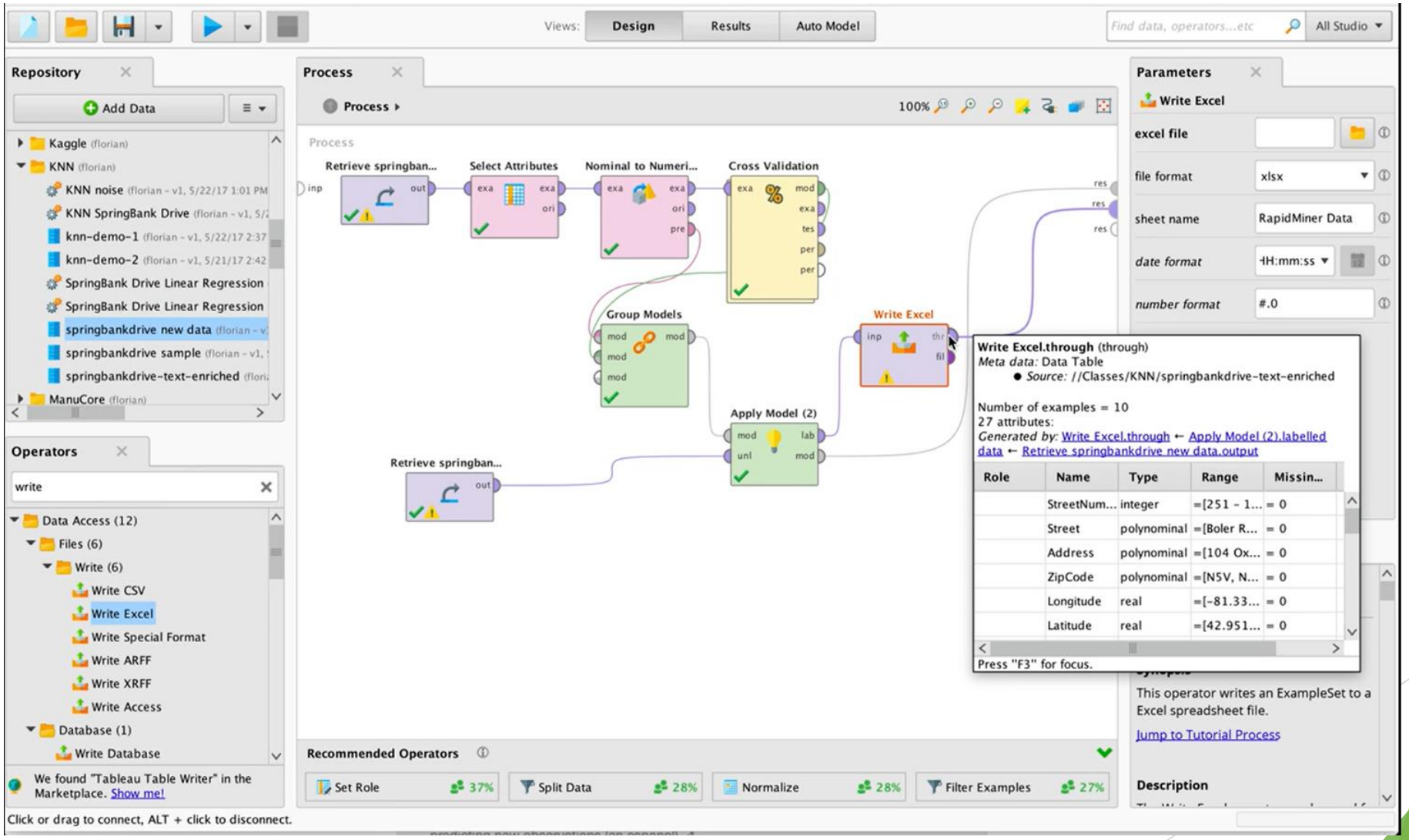
Rapidminer

- IDE (Integrated Development Environment)
- Funkciógazdag front-end workflow készítéshez
- Radoop backend használatával Spark és Hadoop támogatás
- Neurális hálózatok létrehozása, telepítése és menedzsmentje is támogatott

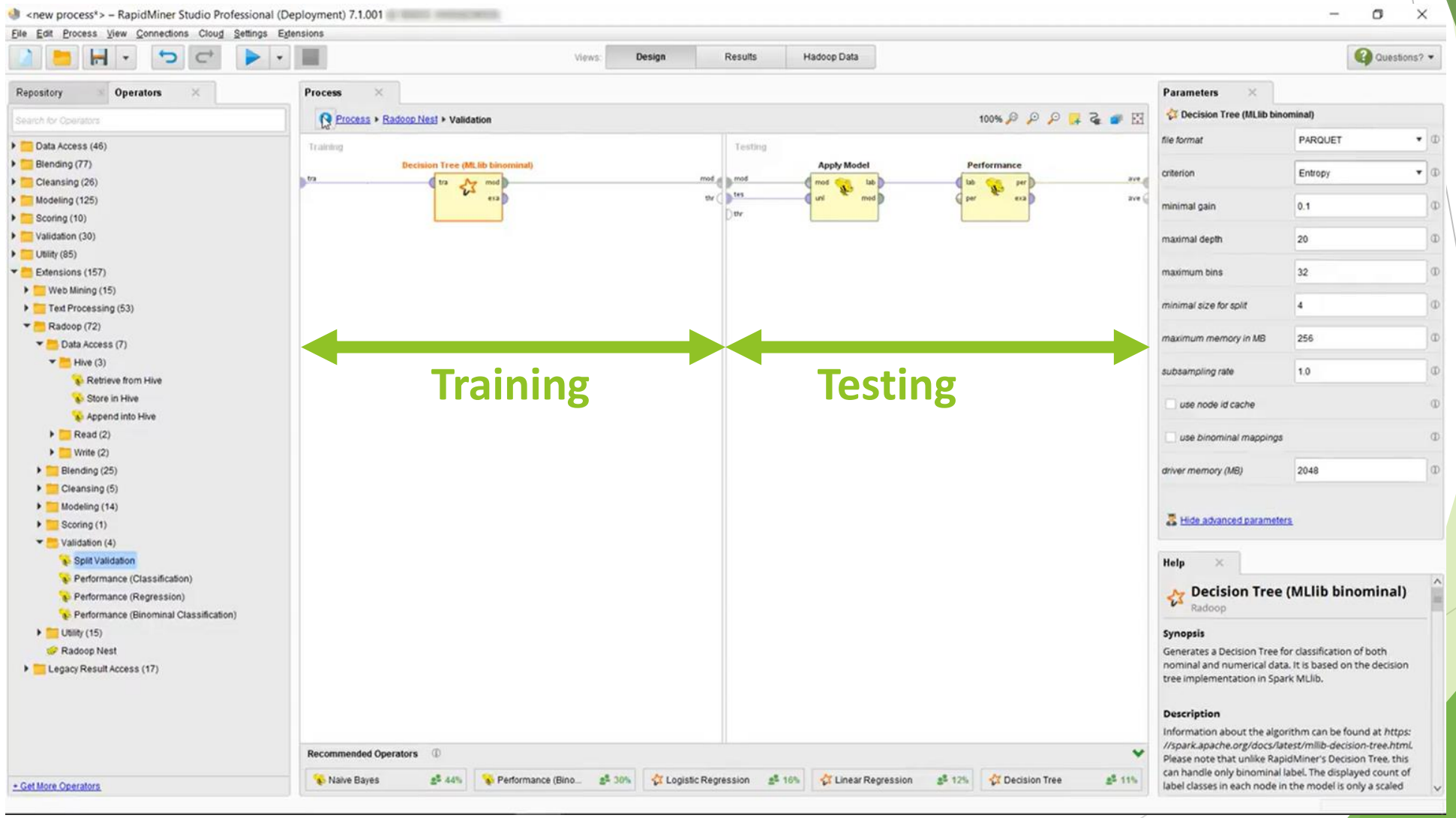
Rapidminer és Spark integráció



Rapidminer Studio



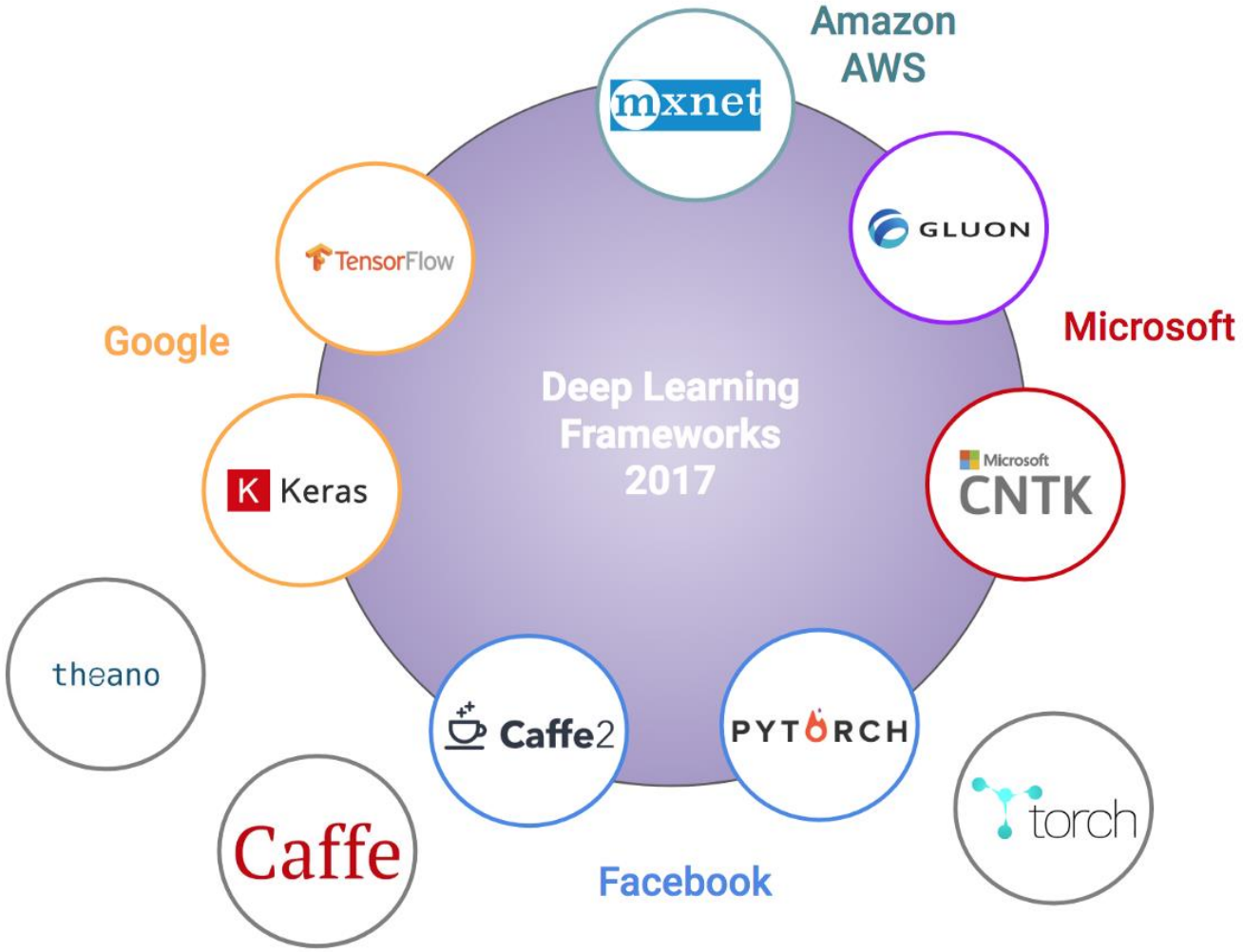
Rapidminer Studio #2



Deep Learning for Java (DL4J)

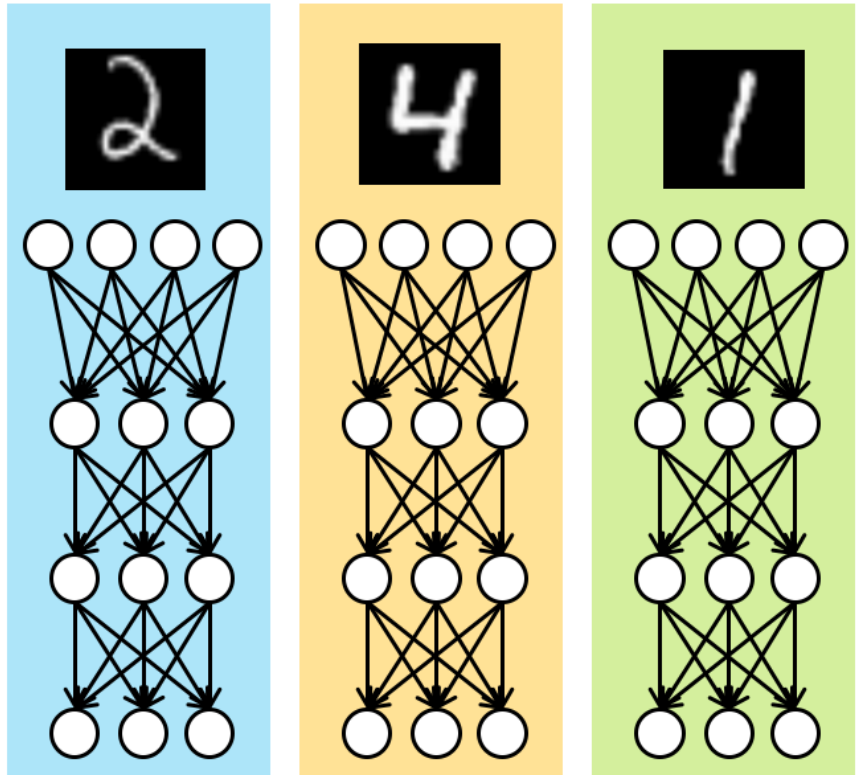
- Támogatott nyelvek:
  - Java and Scala
- Apache Hadoop és Spark támogatás elosztott CPU vagy GPU környezetben
- Széleskörű neurális háló modell támogatás
- Java nem elterjedt eszköz mély tanulás területén, így nehezen integrálható más keretrendszerekkel
- A legfőbb előnye, hogy Java alapú rendszerekhez jól illeszthető gépi tanulási feladatokhoz

Mély tanulási keretrendszerek

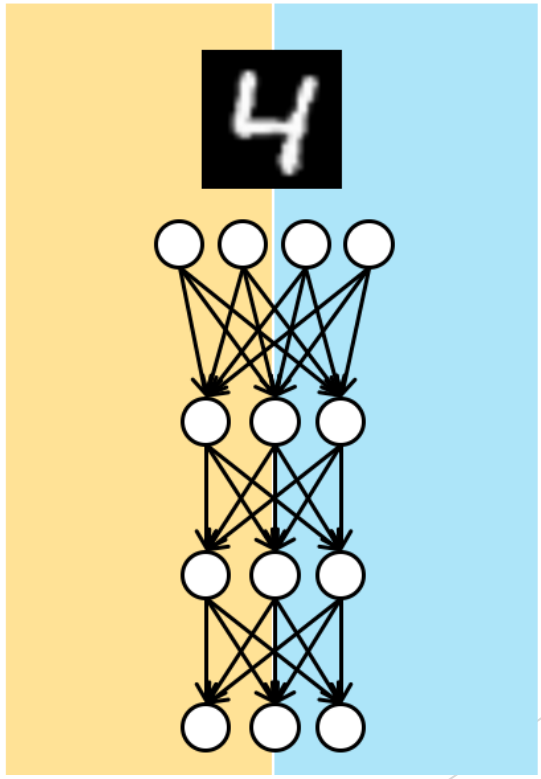


Párhuzamosítási stratégiák

Adat párhuzamos



Modell párhuzamos



Apache MXNet

- Nyílt forráskódú mély tanulási keretrendszer
- Széleskörű programozási nyelv támogatás
- Átjárhatóságot biztosít prototípus és produkciós rendszerek között
- NVIDIA CUDA alapú GPU támogatás
- Adat és modell párhuzamos elosztott tanítás
- Központosított kulcs-érték tár az elosztott tanítás szinkronizálásához

The Microsoft Cognitive Toolkit (CNTK)

- Nyílt forráskódú mély tanulási keretrendszer produkciós szintű elosztott mély tanuláshoz
- Microsoft fejlesztés – Azure felhő támogatás
- A neurális hálózat számítási gráfként való reprezentálása
- Széleskörű programozási nyelv támogatás
- NVIDIA CUDA alapú GPU támogatás
- MPI használata állomások közötti kommunikációra

PyTorch

- Nyílt forráskódú gépi tanulási keretrendszer prototípus és produkciós rendszerekhez
- Facebook fejlesztés
- Python alapú implementáció
- Tensort definiál a mátrix műveletek GPUn történő végrehajtásához
- NVIDIA CUDA alapú GPU támogatás
- Adat és modell párhuzamos elosztott tanítás
- MPI és Gloo alapú állomások közötti kommunikáció

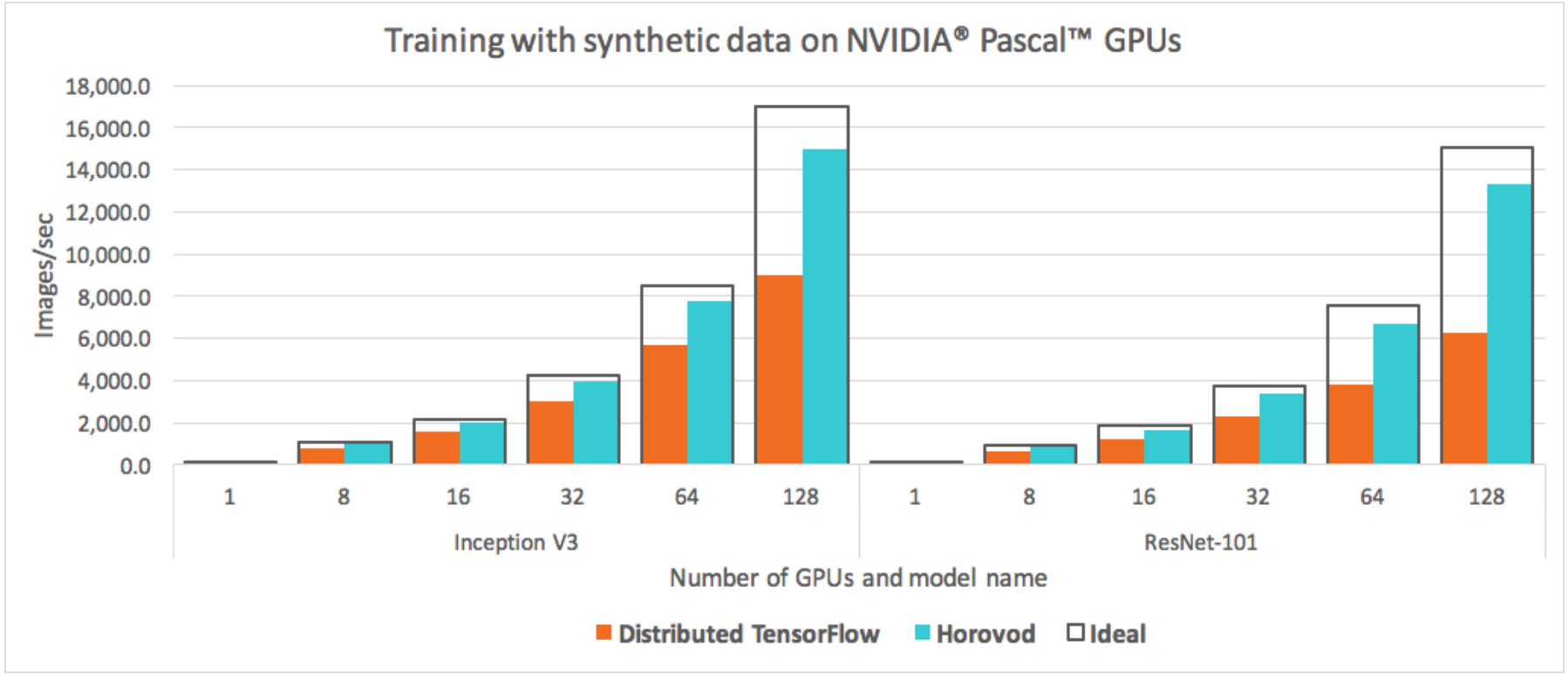
TensorFlow

- Nyílt forráskódú gépi tanulási keretrendszer
- Google fejlesztés
- Számítási gráf alapú műveletvégzés
- Tensort (n-diemnziós mátrix) definiál a matematikai műveletekhez
- CPU, GPU és TPU támogatás
- NVIDIA CUDA alapú GPU támogatás
- TensorFlow 2.0 verziótól a Keras integrálásra került
  - Magasabb szintű programozási környezet
  - Zökkenőmentes CPU és GPU közötti válltás
- gRPC protokollt használ állomások közötti kommunikációhoz

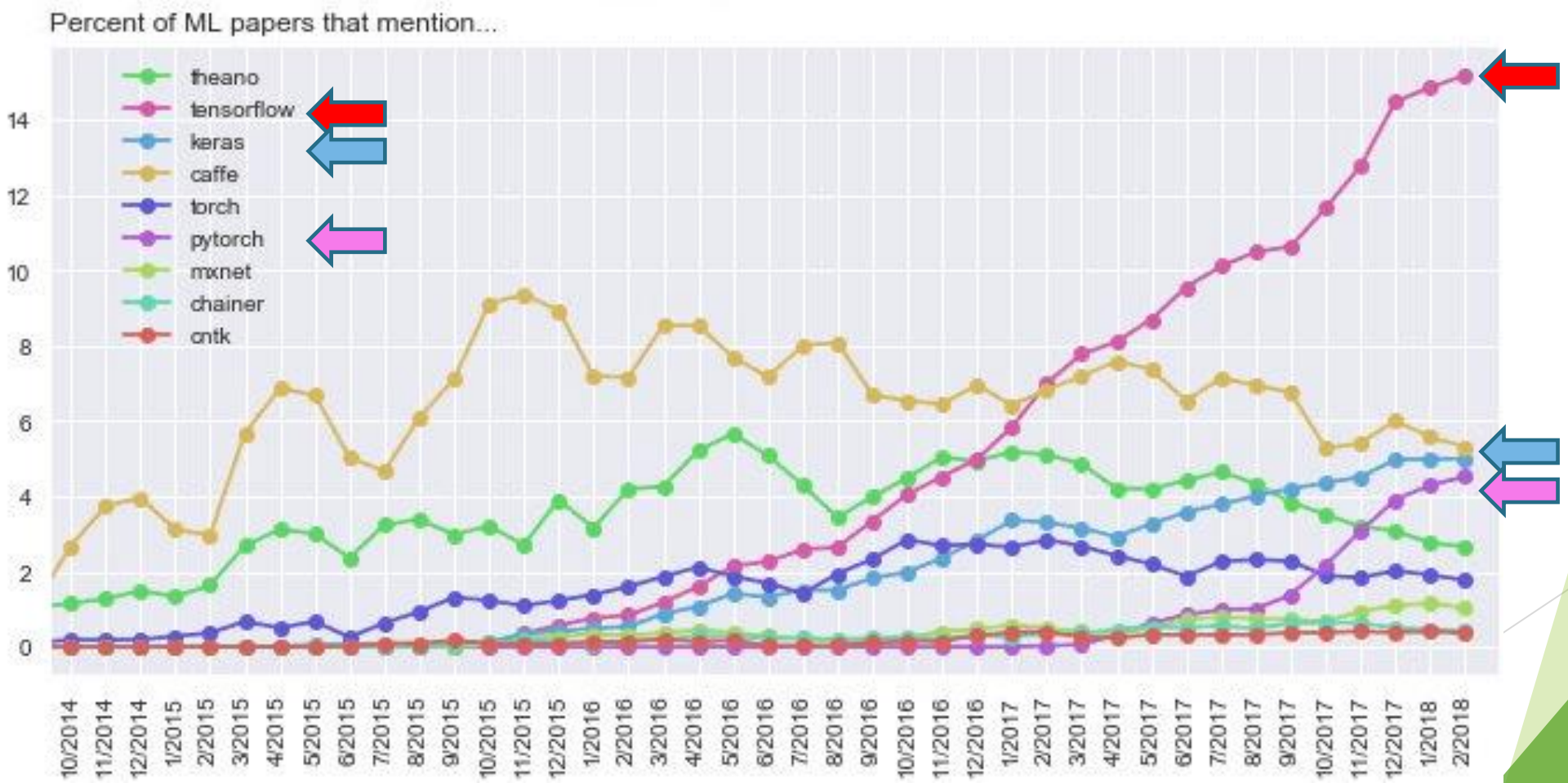
Horovod

- Nyílt forráskódú elosztott mély tanulási keretrendszer
- Uber fejlesztés
- TensorFlow, Keras, PyTorch, Apache MXNet és Spark támogatás
- Könnyen alkalmazható elosztott tanítás
- Adat párhuzamos végrehajtás
  - Modell párhuzamos végrehajtás egy állomáson belül
- MPI alapú állomások közötti kommunikáció
- Baidu eredeti megoldásának továbbfejlesztése
- Ring-allreduce stratégia
  - 2*(N-1) kommunikáció minden tanítási lépés után
  - Allreduce N*(N-1) kommunikációt generál lépésenként

Horovod elosztott tanítás



Mély tanulási keretrendszerek népszerűsége



Összegzés

- Adat feldolgozás és Big Data feladatok
  - Hadoop (disk intenzív műveletek)
  - Spark (in-memory műveletek)
- Mély tanulási feladatok
  - TensorFlow (Keras)
  - PyTorch
- Elosztott mély tanulási feladatok
  - Horovod

Köszönöm a figyelmet!