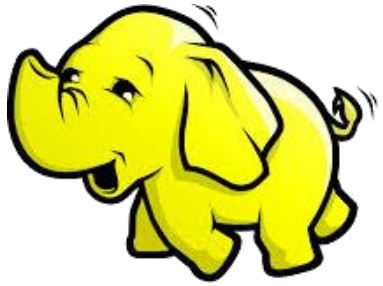
A Hadoop ökoszisztéma

Rusznák Attila

SZTAKI

Mi a Hadoop?



Az Apache szerint:

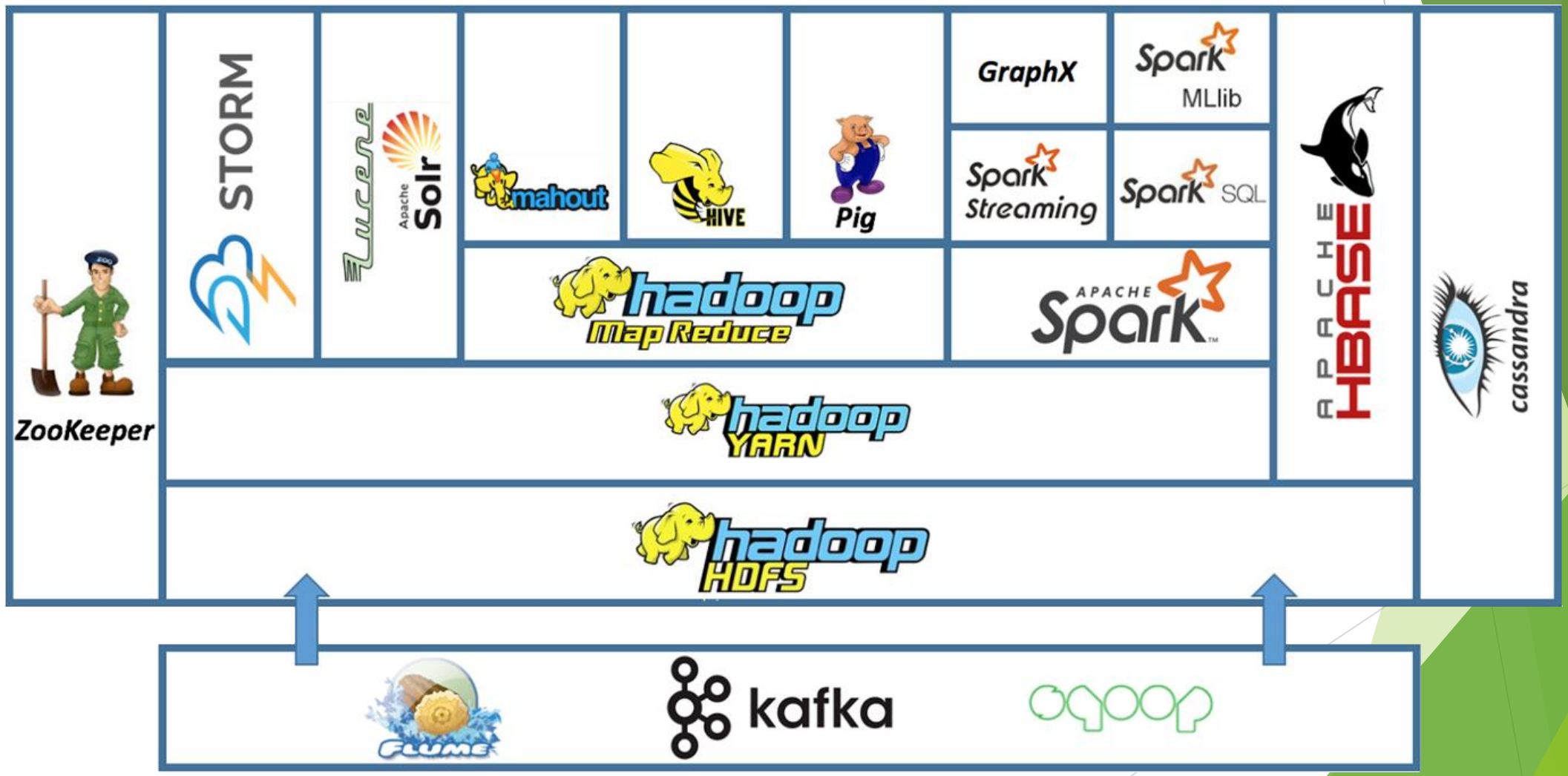
A Hadoop egy nyílt forráskódú szoftver, mely lehetővé teszi elosztott környezetben hatalmas adathalmazok feldolgozását egyszerre több gépen.

Két fő komponensre bonthatjuk fel:

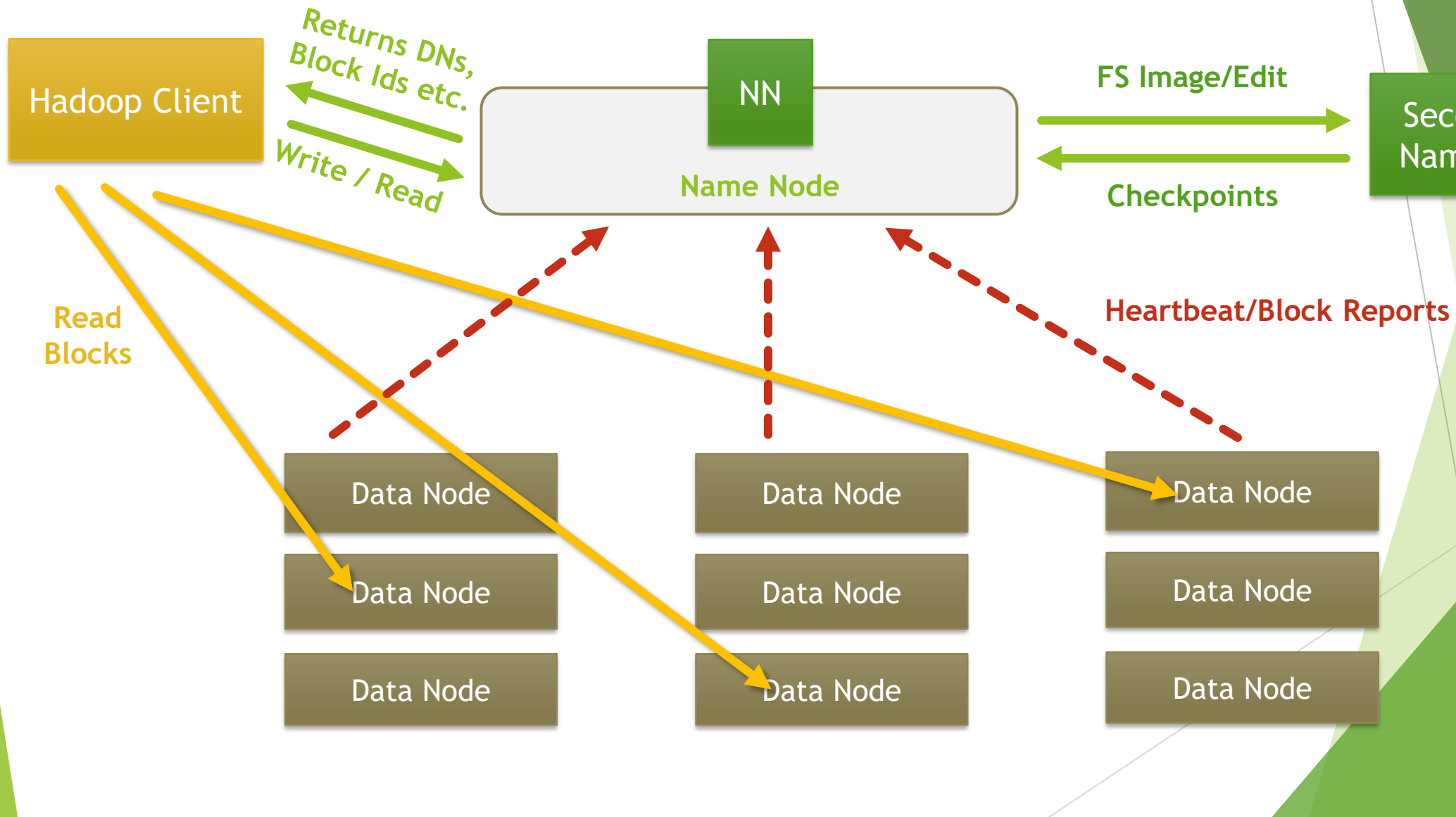
1. Core / Hadoop base
   - HDFS, Hadoop common, MapReduce, YARN
2. Az ezekre épülő különböző szoftverkomponensek
   - Hive
   - Hbase
   - Pig

A két komponenst nevezzük együttesen Hadoop ökoszisztémának.

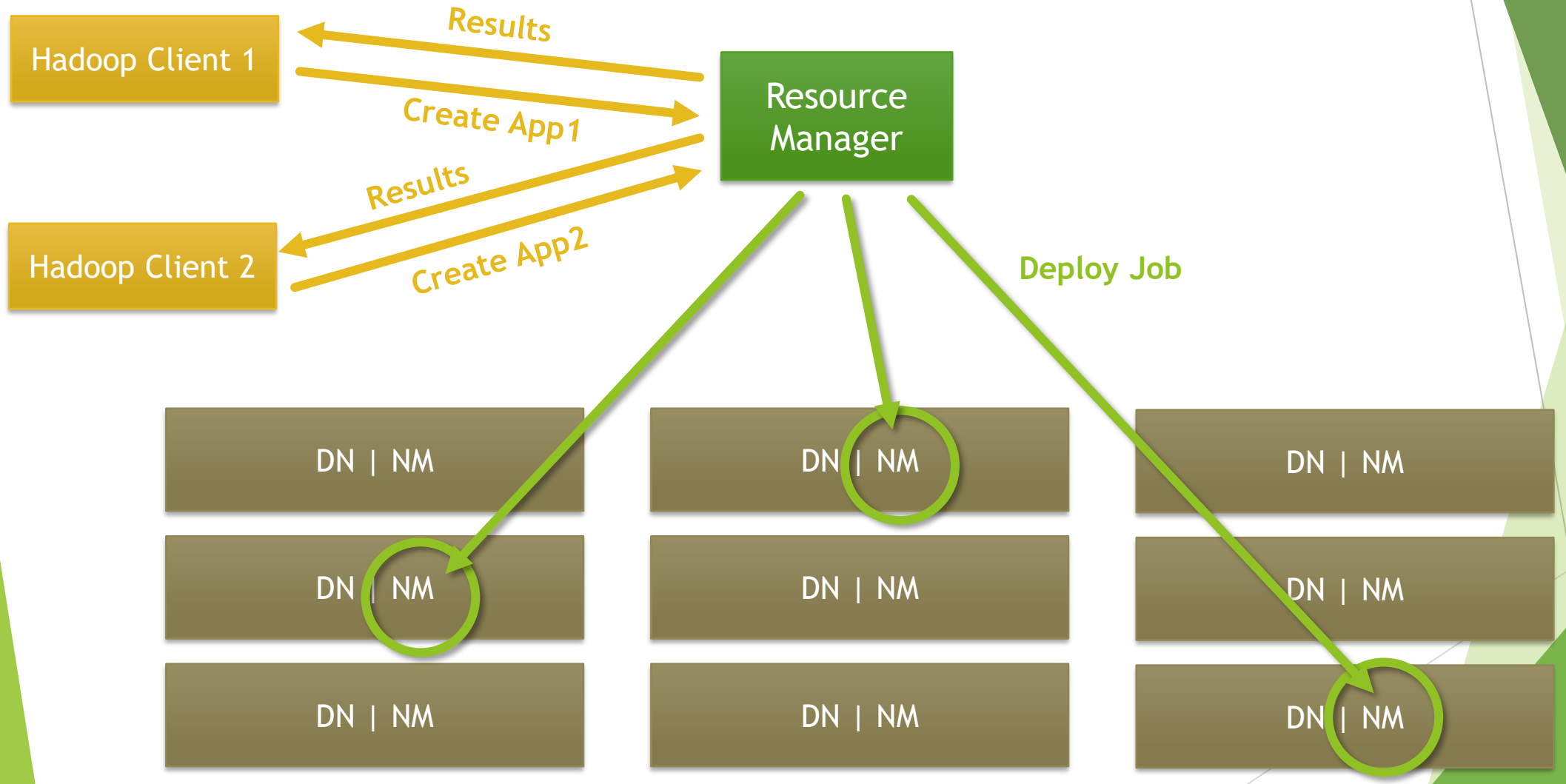
A Hadoop ökoszisztéma



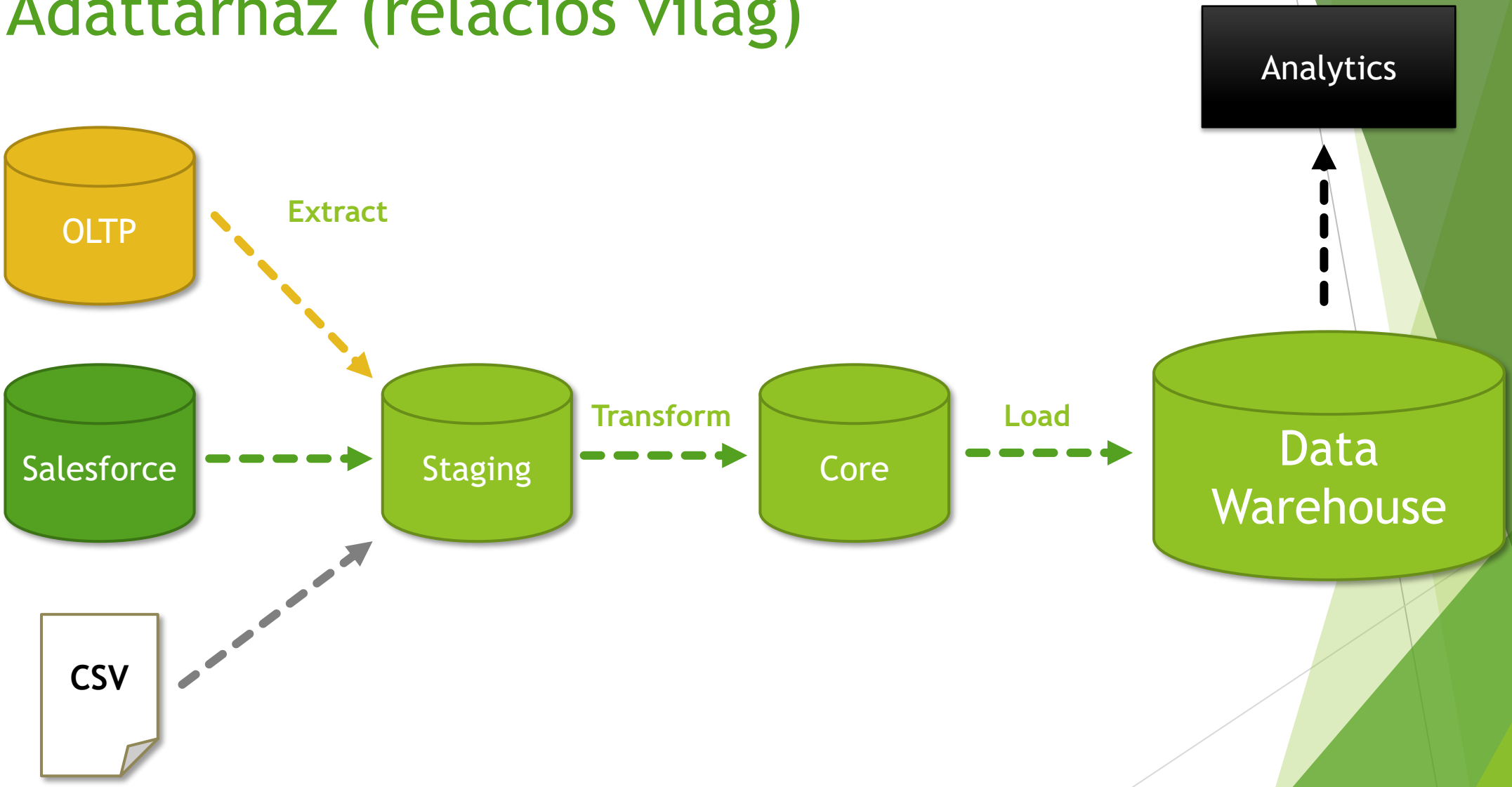
Hadoop: fájl írása és olvasása



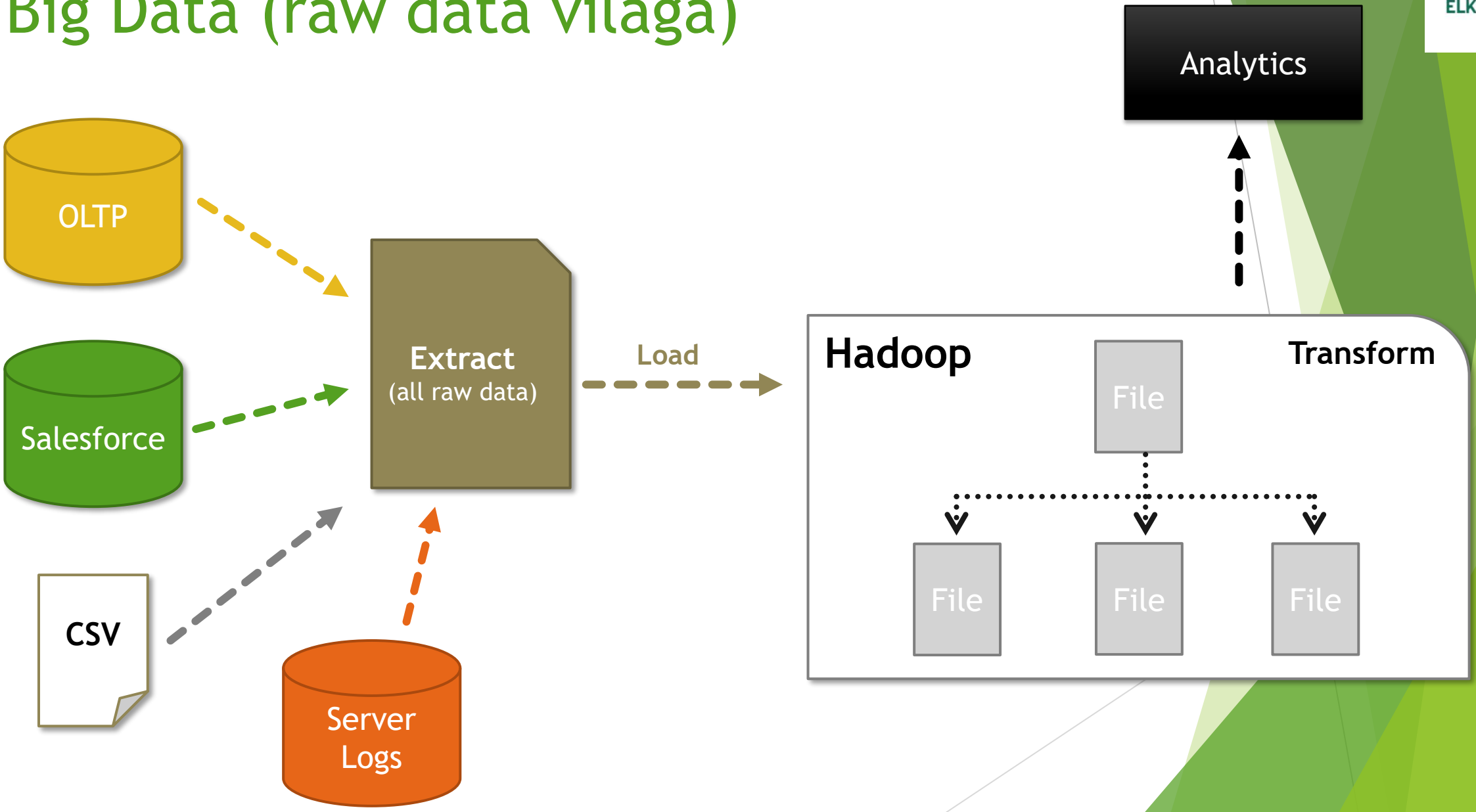
Hadoop: job futtatás



Adattárház (relációs világ)



Big Data (raw data világa)

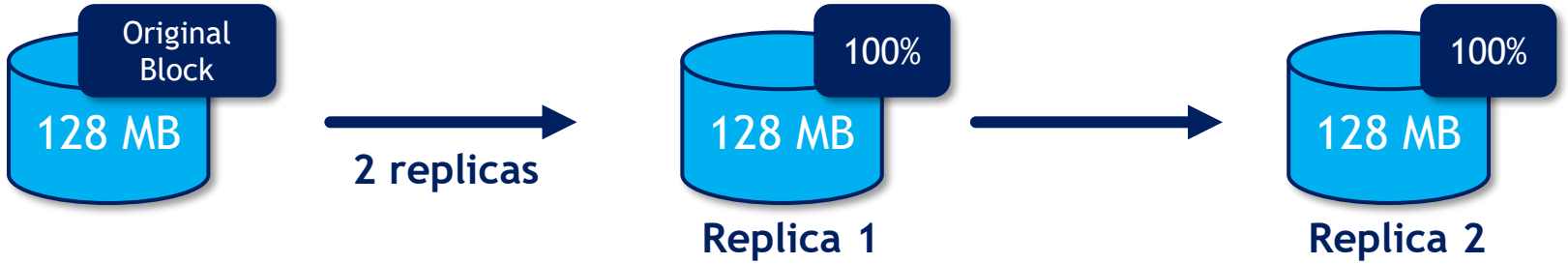


Hadoop 3

Hadoop 2.x	Hadoop 3.x
Replikáció alapú hibatűrés	Erasure coding alapú hibatűrés
A replikáció a HDFS-t 200%-ban használja fel	A replikáció a HDFS-t csupán 50%-ban használja fel
Limitáltabb skálázhatóság	Továbbfejlesztett skálázhatóság
A NameNode helyreállítása manuális beavatkozást igényel	A NameNode helyreállítás automatizálható
Támogatja a DFS, Amazon S3 és FTP tárolókat	Szinte minden tárolót támogat, ill. a Microsoft Azure Data Lake-et is
Min. Java verzió: JDK 7.0	Min. Java verzió: JDK 8.0

A Hadoop 3 csökkentett tárigényű

- A Hadoop 2.x alapértelmezetten 3 replikát tárolt a fájlokból
  - Egyrészt vannak a fájlból készült blokkok
  - A blokkokból készül további 2db. biztonsági másolat
  - Minden replikált blokk tárolási költsége 100%, így 2x100% = 200%-os tárigény



- A Hadoop 3.x új tárolási technológiát vezetett be
  - Egy új kódolás-dekódolás technológiával elérték, hogy a tárigény csak 50%-os legyen (Erasure kódolás)