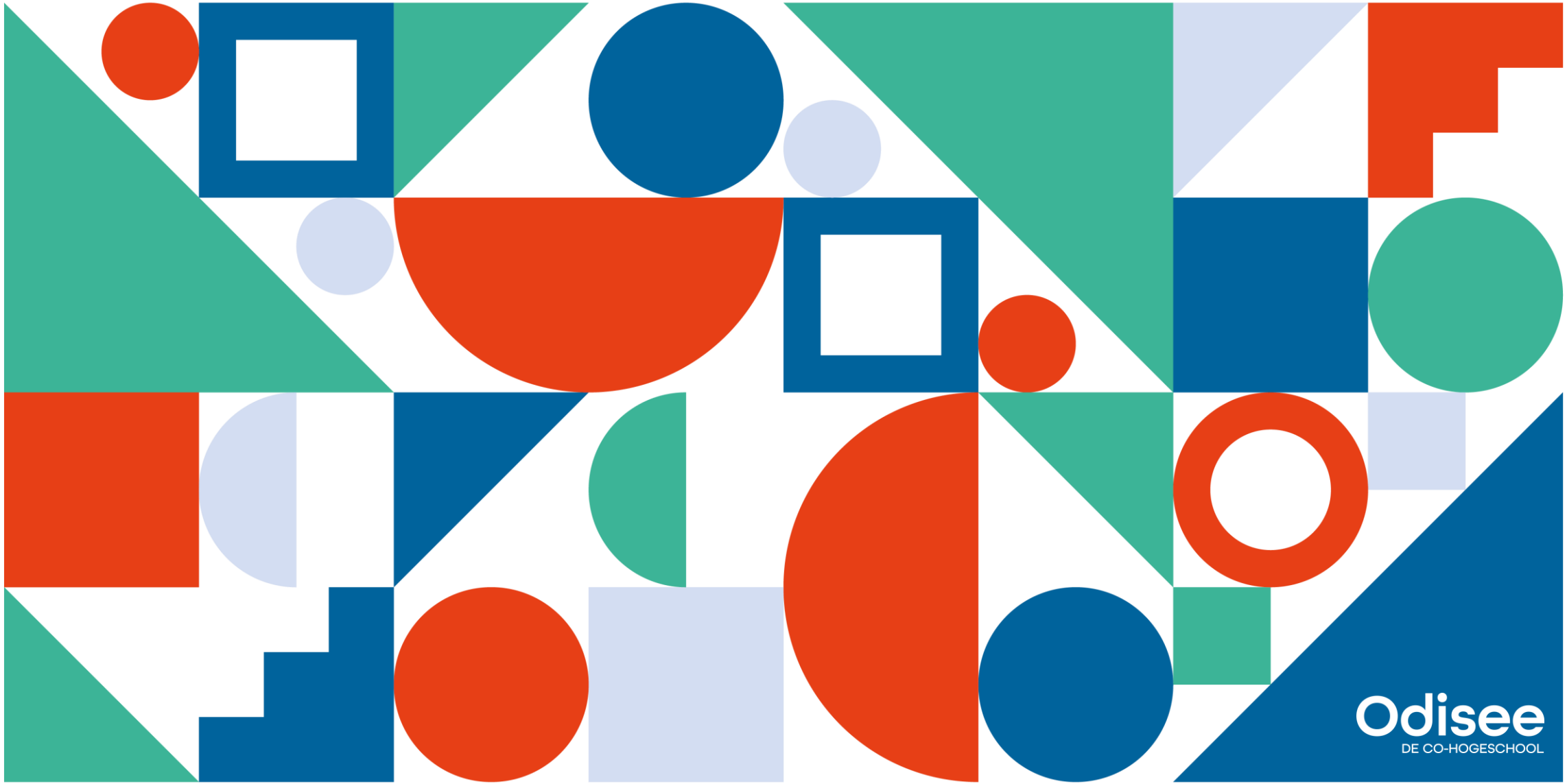

Odisee
DE CO-HOGESCHOOL

Big Data – Hadoop cluster

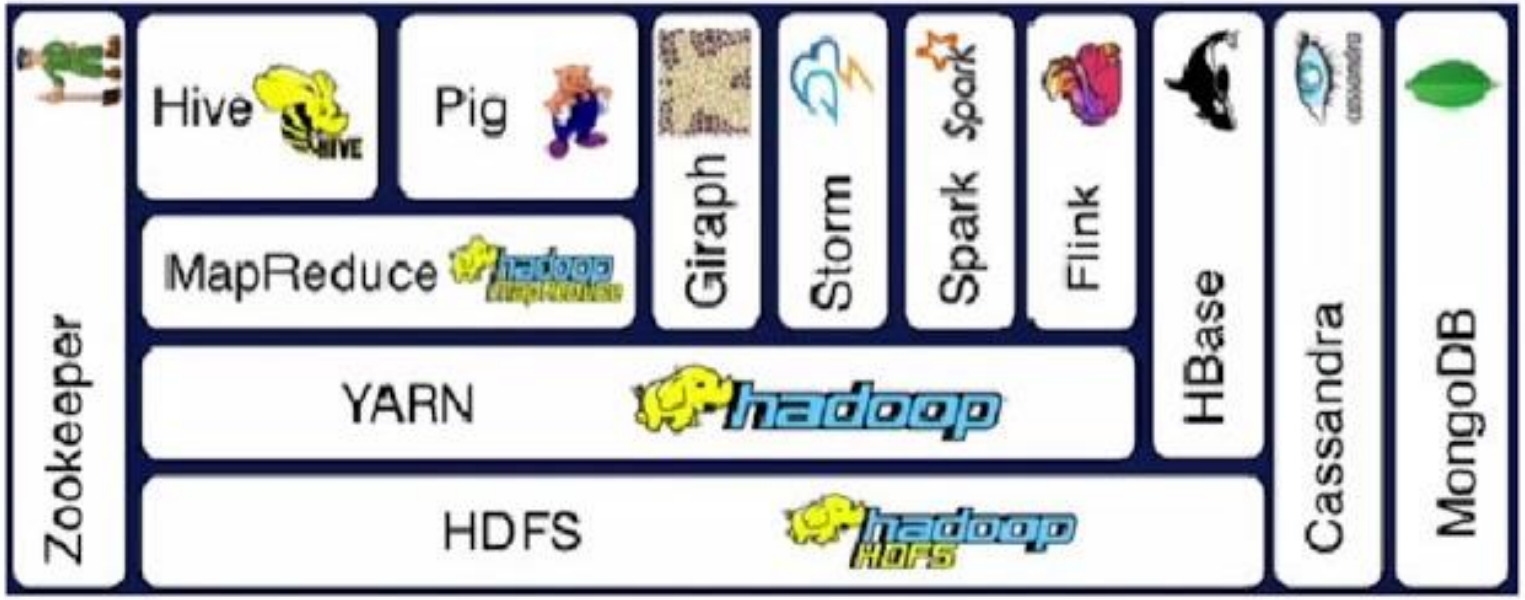
Jens Baetens

Hadoop

- Gebaseerd op Google File System (2003)
- Ontwikkeld door Apache
- Open source
- Uitgegroeid tot omgeving met veel verschillende applicaties
  - Elke applicatie runt typisch in aparte containers/servers

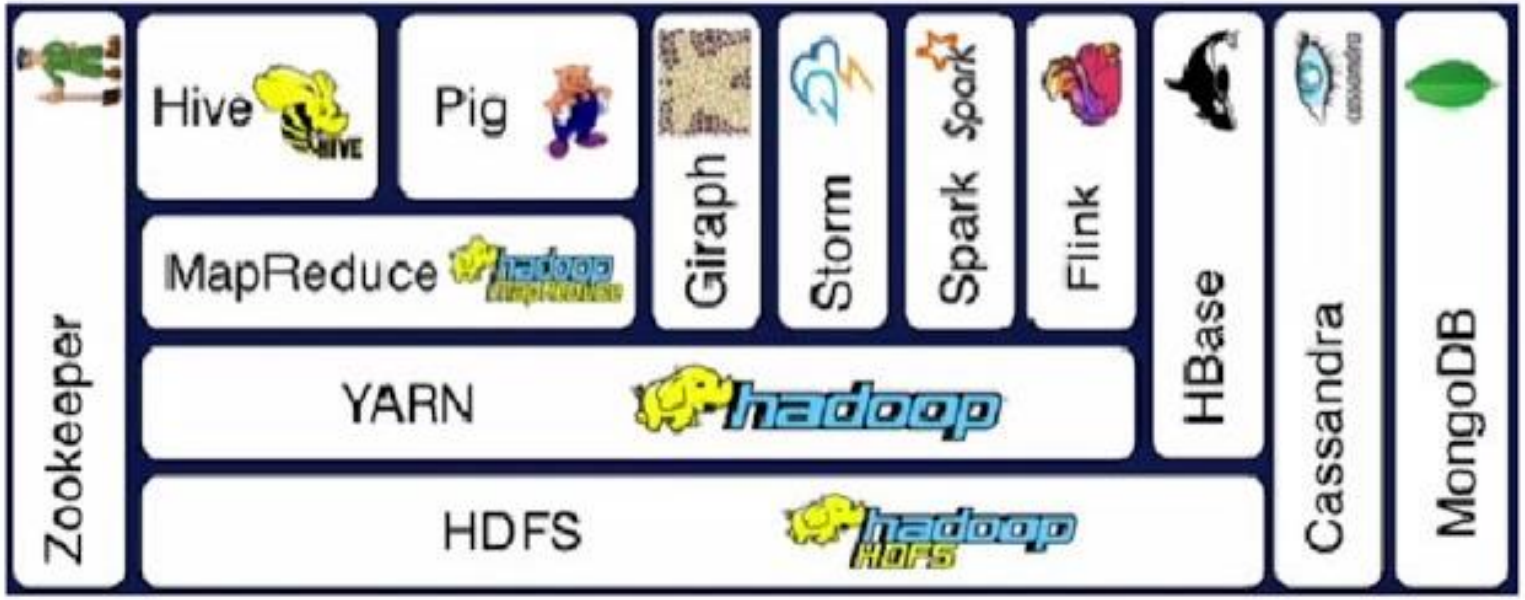
Hadoop Ecosystem

- HDFS – core functionality
- Distributed File System
- Op HDD



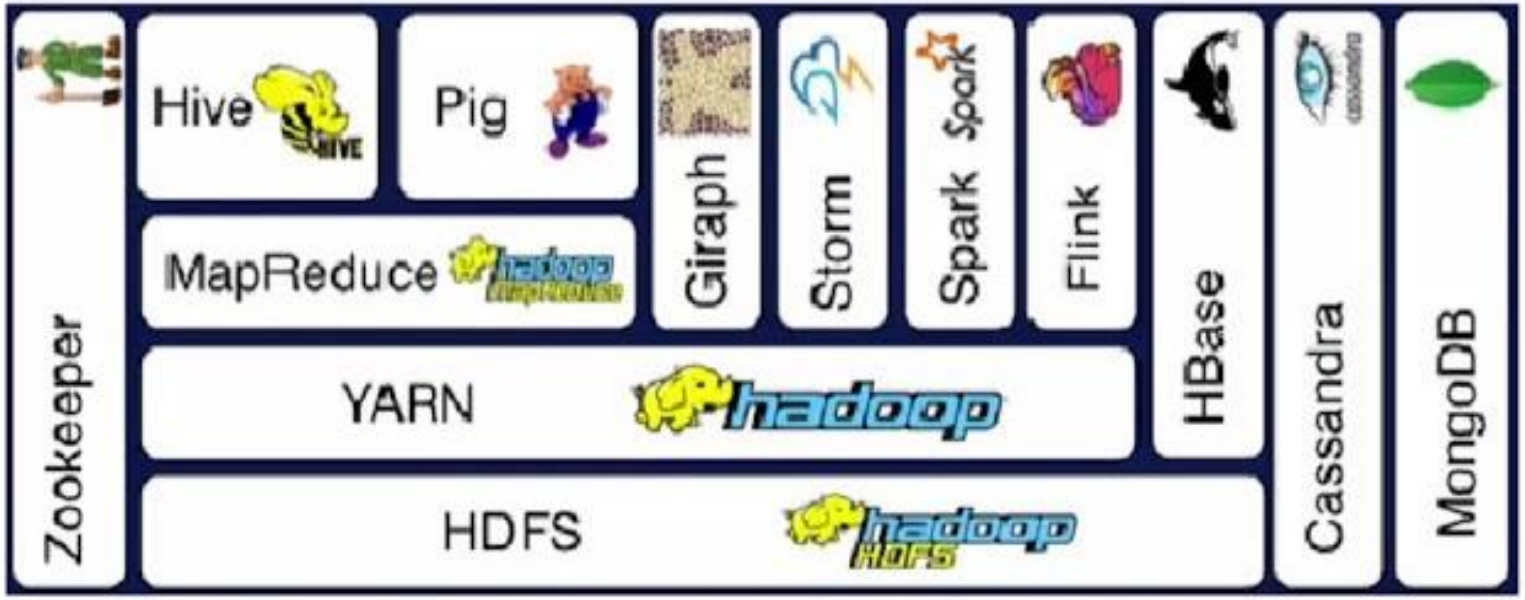
Hadoop Ecosystem

- YARN – Yet Another Resource Negotiator
- Beheer van computing power
- Welke code op welke node



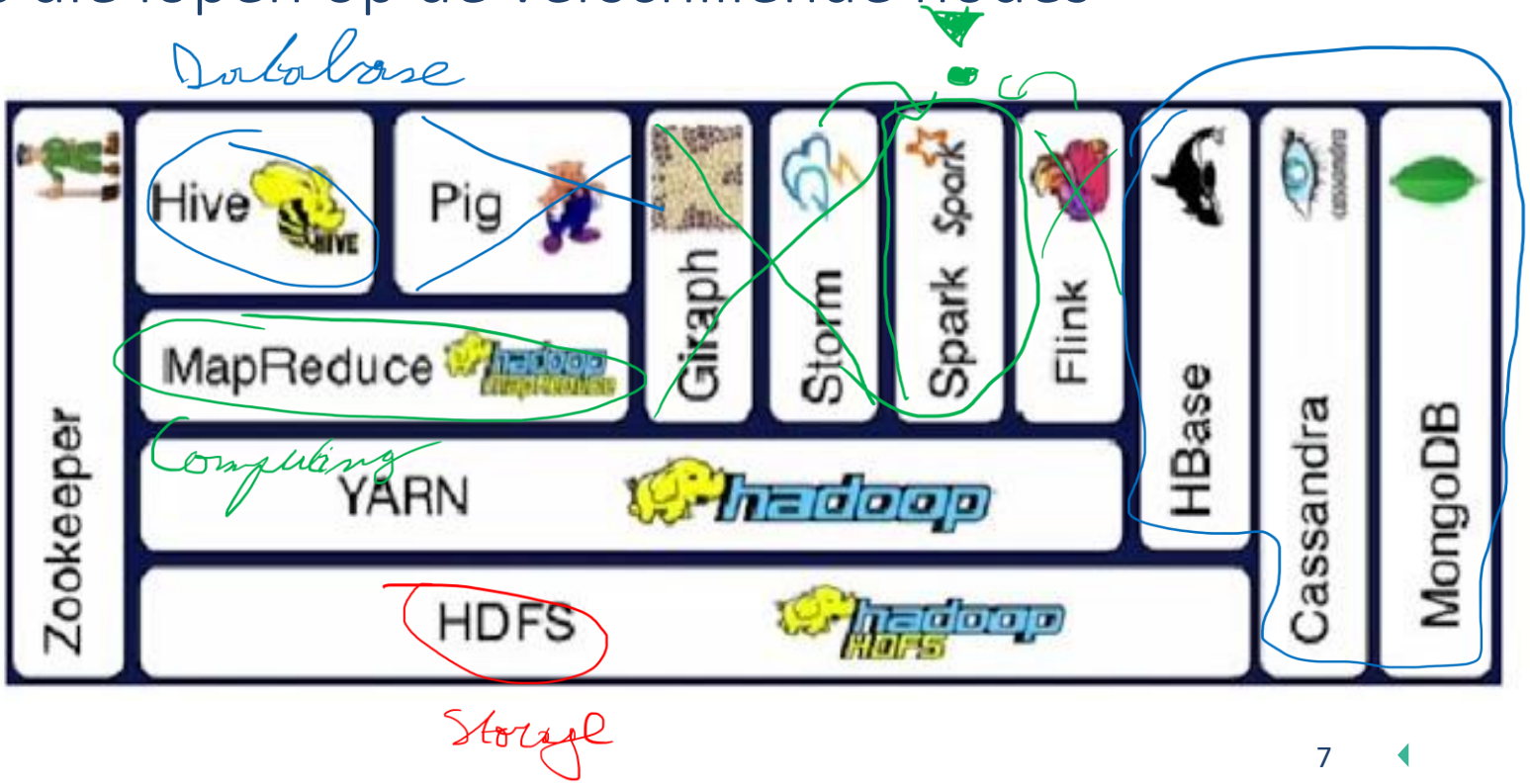
Hadoop Ecosystem

- MapReduce
- Distributed Computing
- Ontwikkeld door Google
- 2 fases
  - Mapping (Divide)
  - Reduce (Conquer)



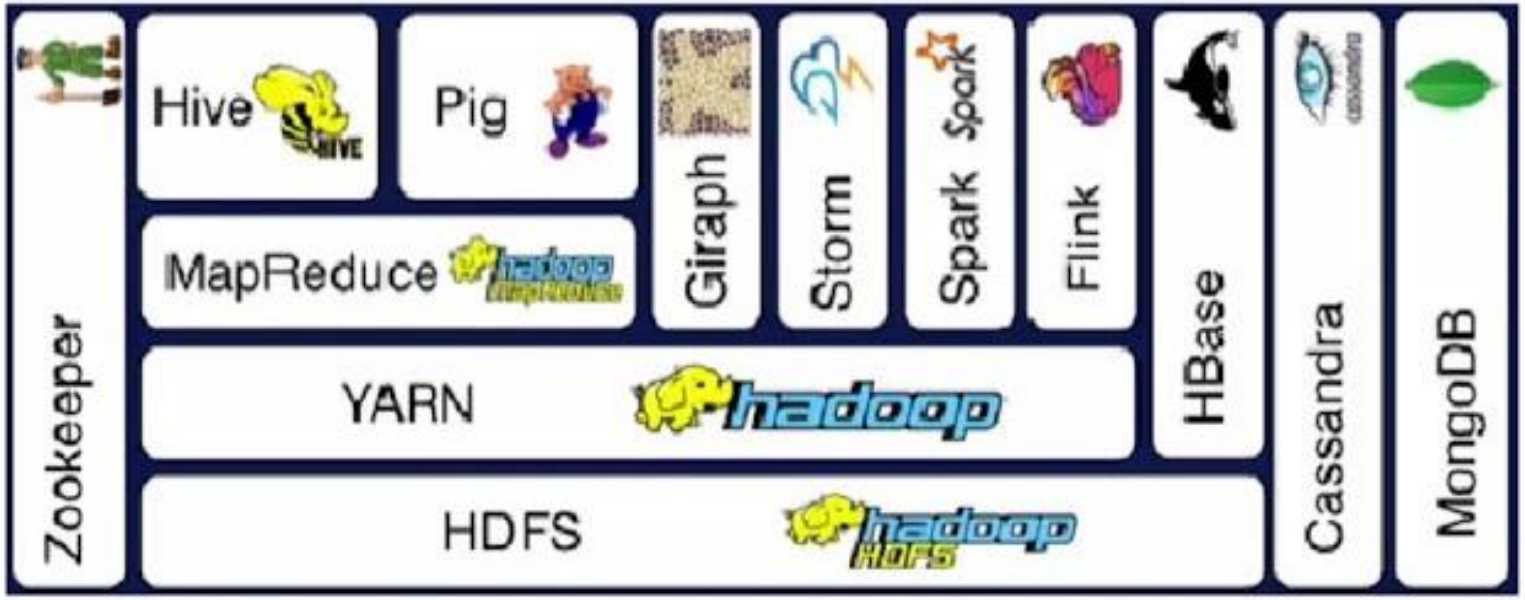
Hadoop Ecosystem

- Zookeeper
- Beheren van alle applicaties die lopen op de verschillende nodes



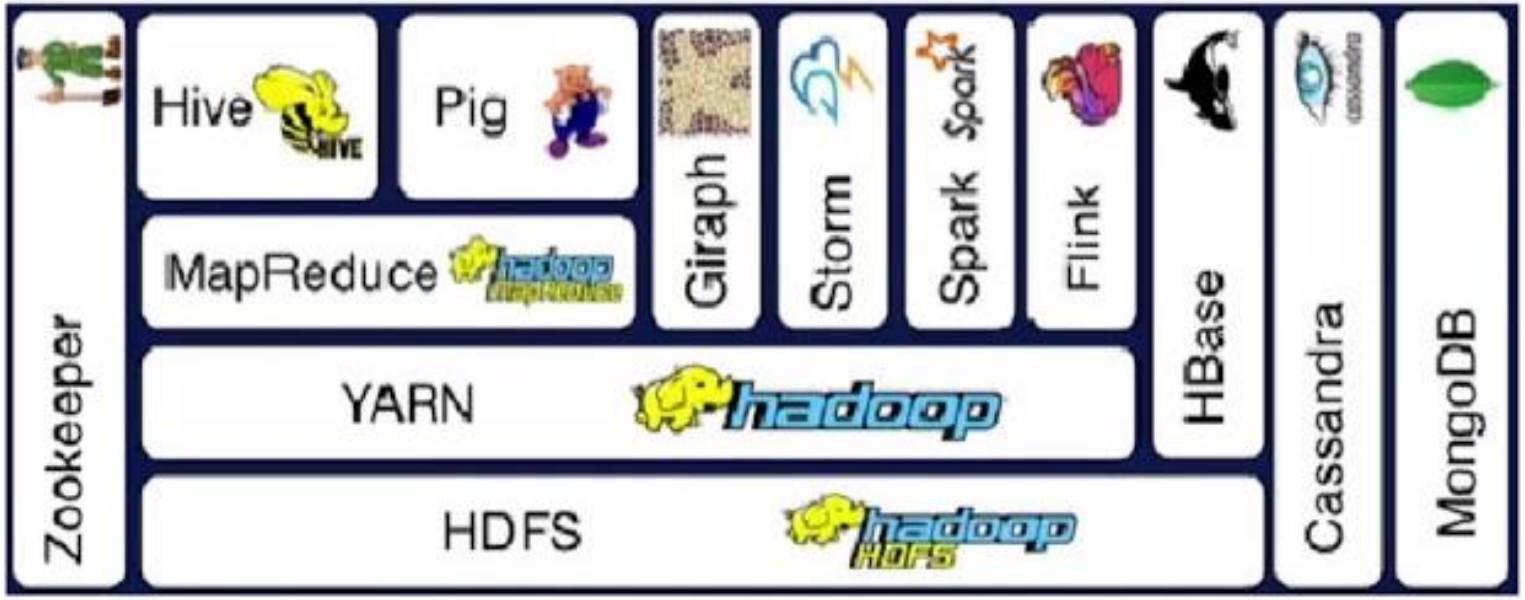
Hadoop Ecosystem

- Hive
- Distributed Datawarehouse
- Sql-like
- Queries via MapReduce



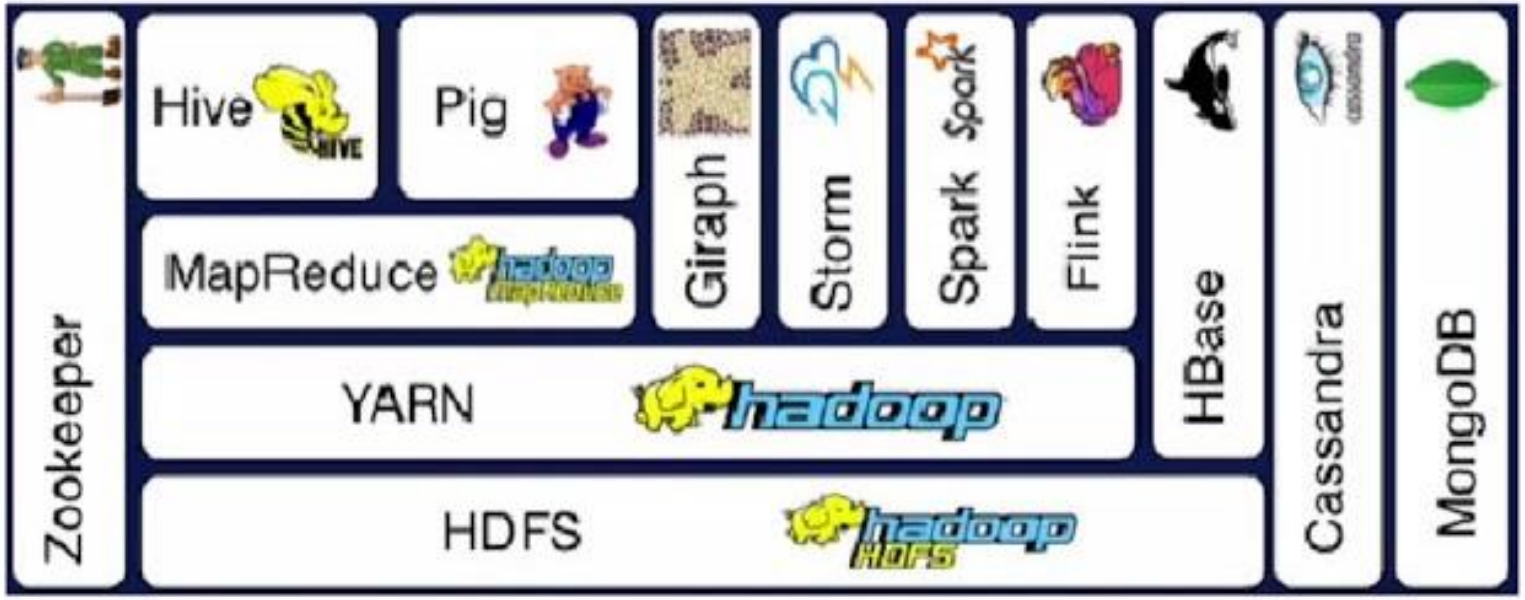
Hadoop Ecosystem

- Pig
- Data analysis
- Using MapReduce/Spark/…
- Taal: Pig Latin



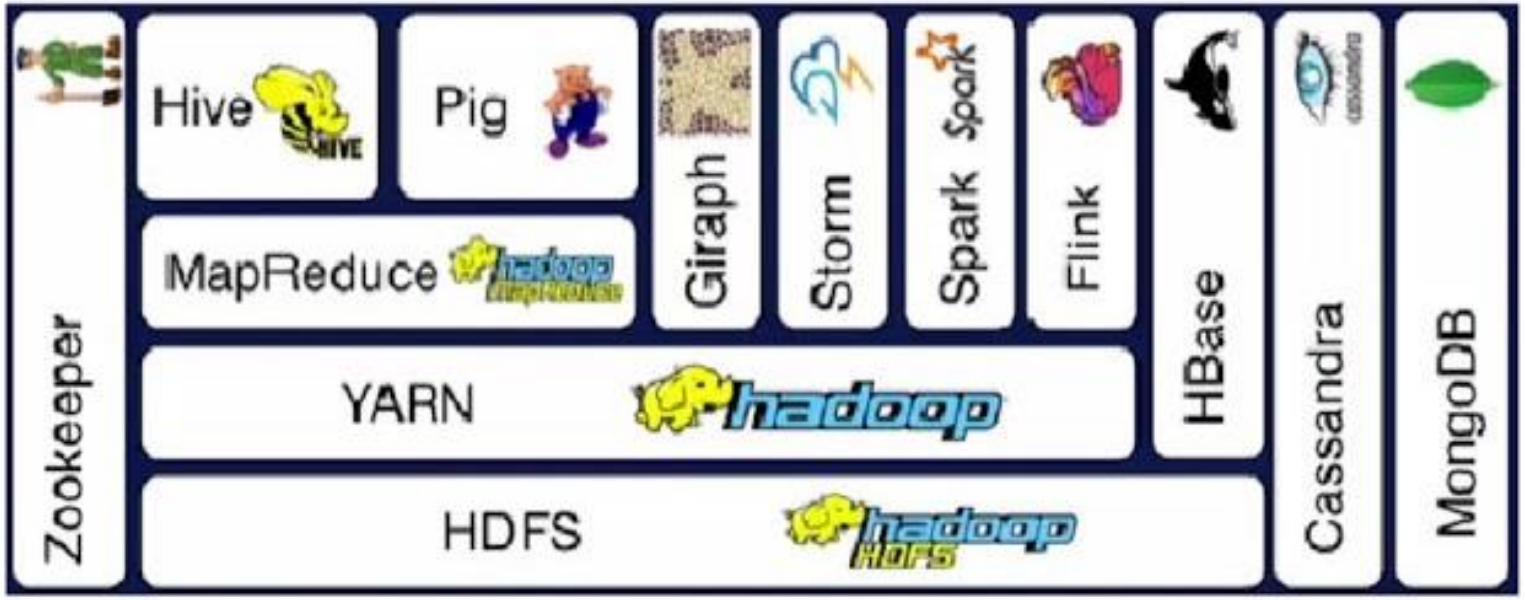
Hadoop Ecosystem

- Giraph
- Bestuderen van een graaf
- Social graph
  - Facebook
  - Twitter
  - …
- Gebruikt geen mapreduce



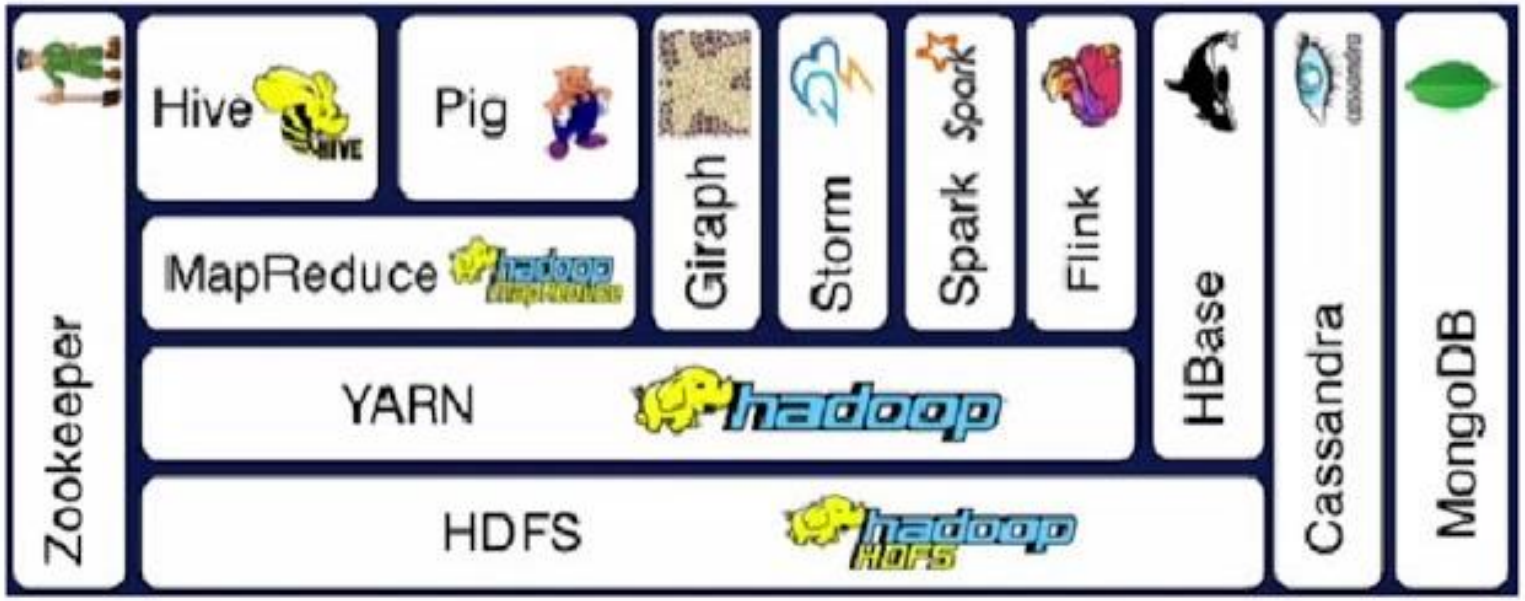
Hadoop Ecosystem

- Storm / Flink
- Verwerken van data streams – continue inkomende datastromen
  - Classificeren
  - Opslaan
  - …



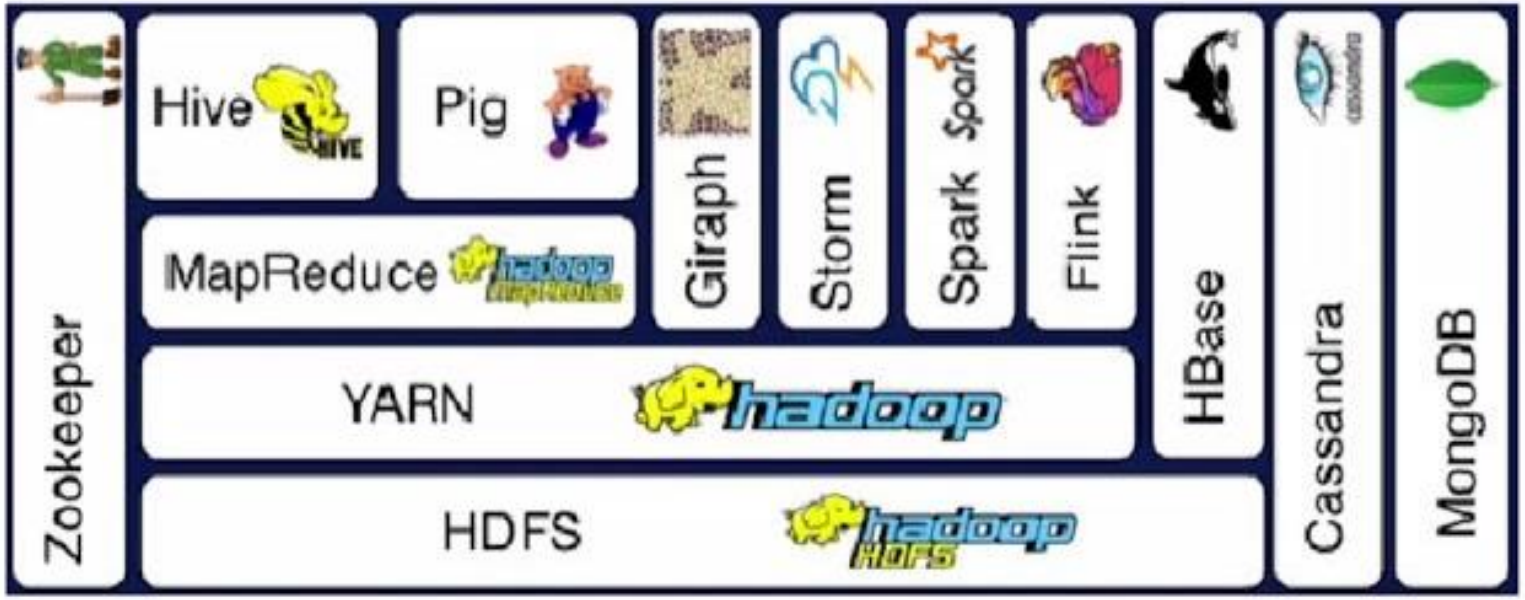
Hadoop Ecosystem

- Spark
- Alternatief voor MapReduce
- Computing in Ram
- Op Hadoop/Cloud/…
- Gebruikt voor
  - SQL (Spark SQL)
  - Streaming (Spark Streaming)
  - Machine Learning (MLlib)
  - Graph analysis (GraphX)



Hadoop Ecosystem

- HBase
- Distributed NoSQL Database
- Geen SQL maar in JAVA



Hadoop Ecosystem

- Cassandra / Mongo DB
- Maken geen gebruik van HDFS
- NoSql databases
- Stand-alone solutions