DATA SCIENCE W PRAKTYCE

Analiza dużych zbiorów danych

CZĘŚĆ I: PLATFORMA APACHE HADOOP

Tematy

- Właściwości Apache Hadoop
- Ekosystem Hadoop
- Model przetwarzania MapReduce
- Hadoop Streaming
- Hive SQL

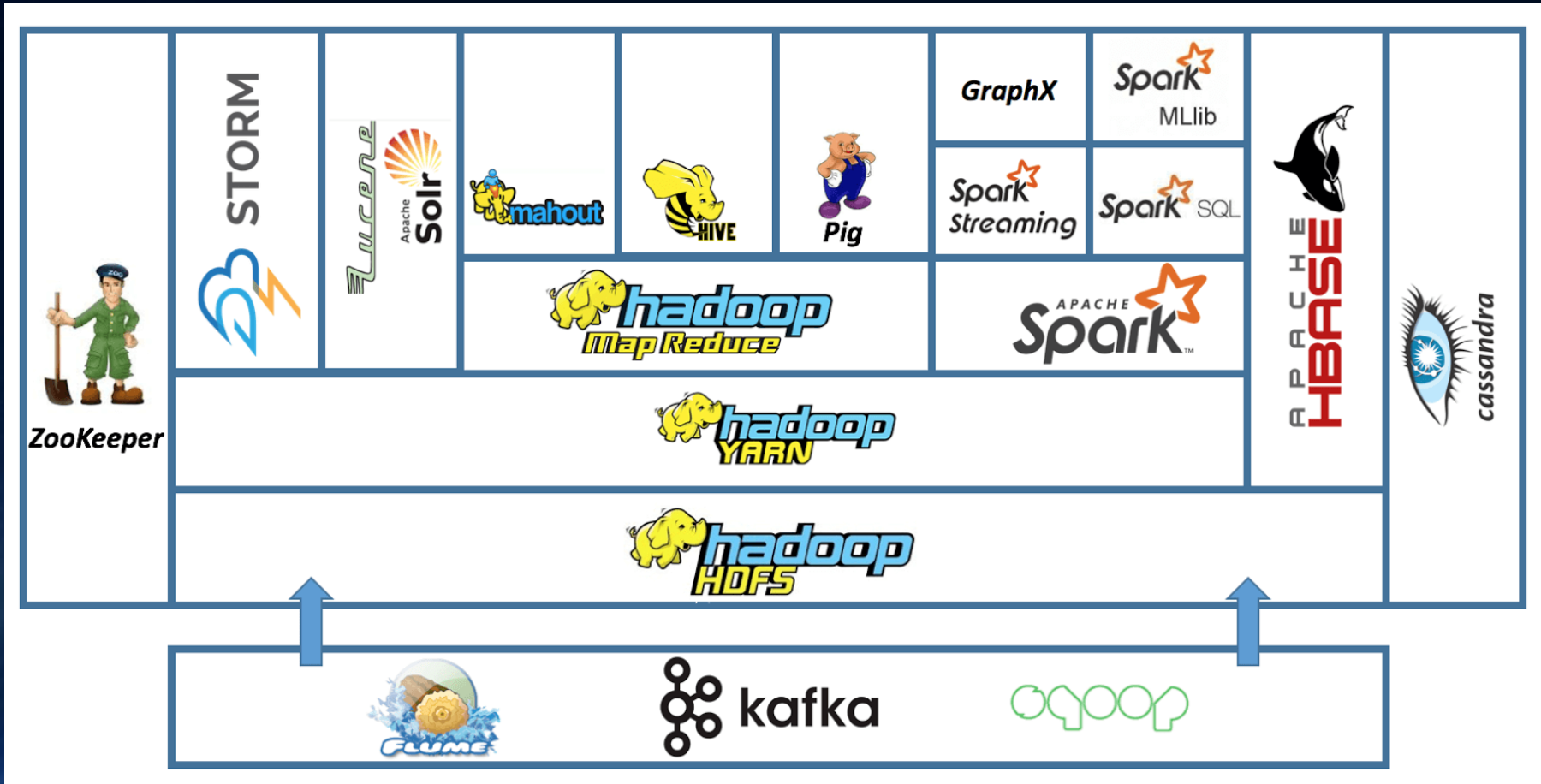
Apache Hadoop

PLATFORMA

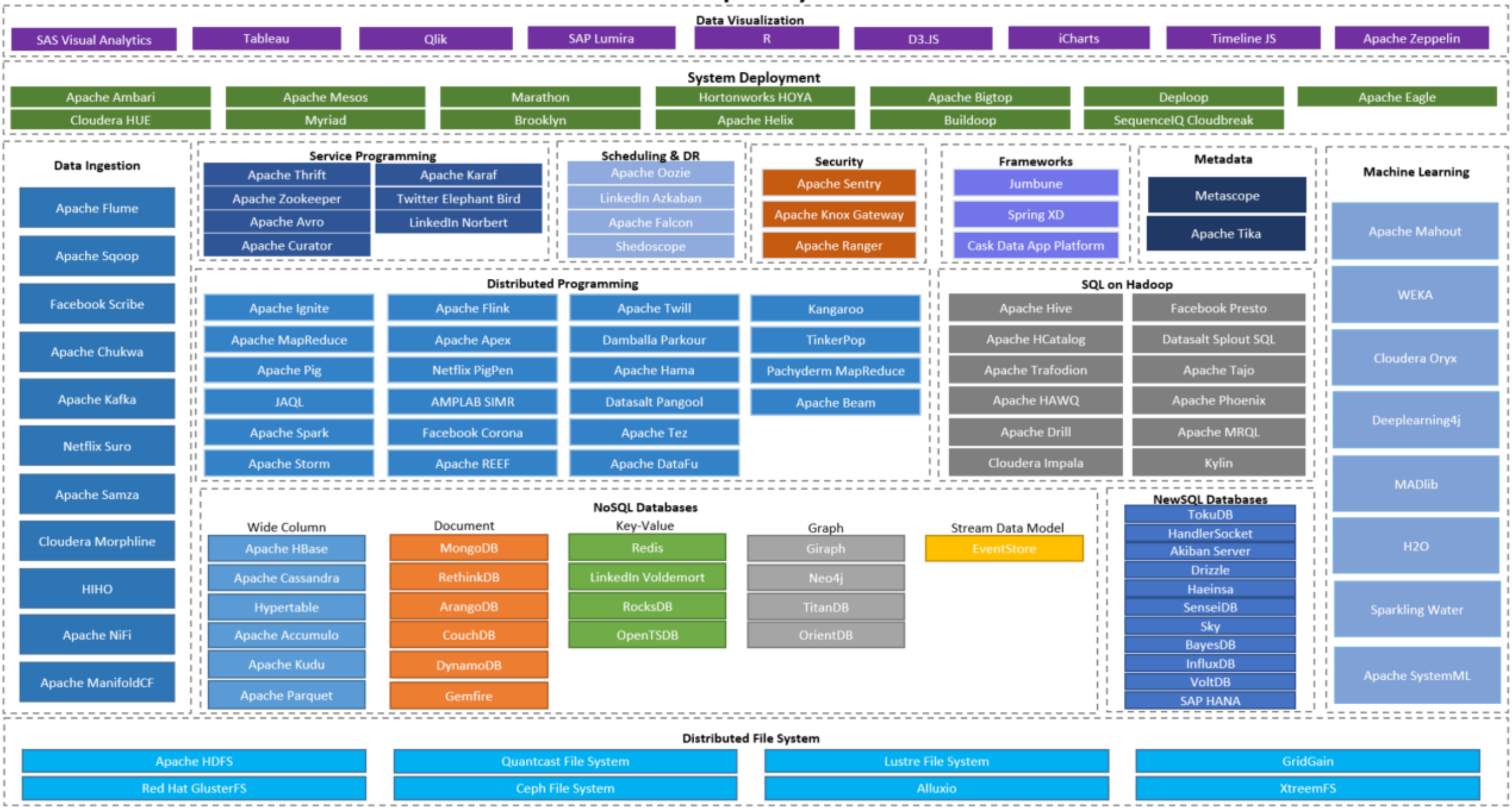
Cechy Apache Hadoop

- Open Source
- Rozwijany od 2006 roku pod szyldem Apache Foundation
  - Twórcy: Doug Cutting z Yahoo oraz Mike Cafarella
  - Stworzony na bazie dokumentacji systemu Google File System
- Przetwarzanie w rozproszonym modelu MapReduce na HDFS
- Zarządzanie zasobami klastra z użyciem podsystemu YARN

Architektura Apache Hadoop



Hadoop Ecosystem

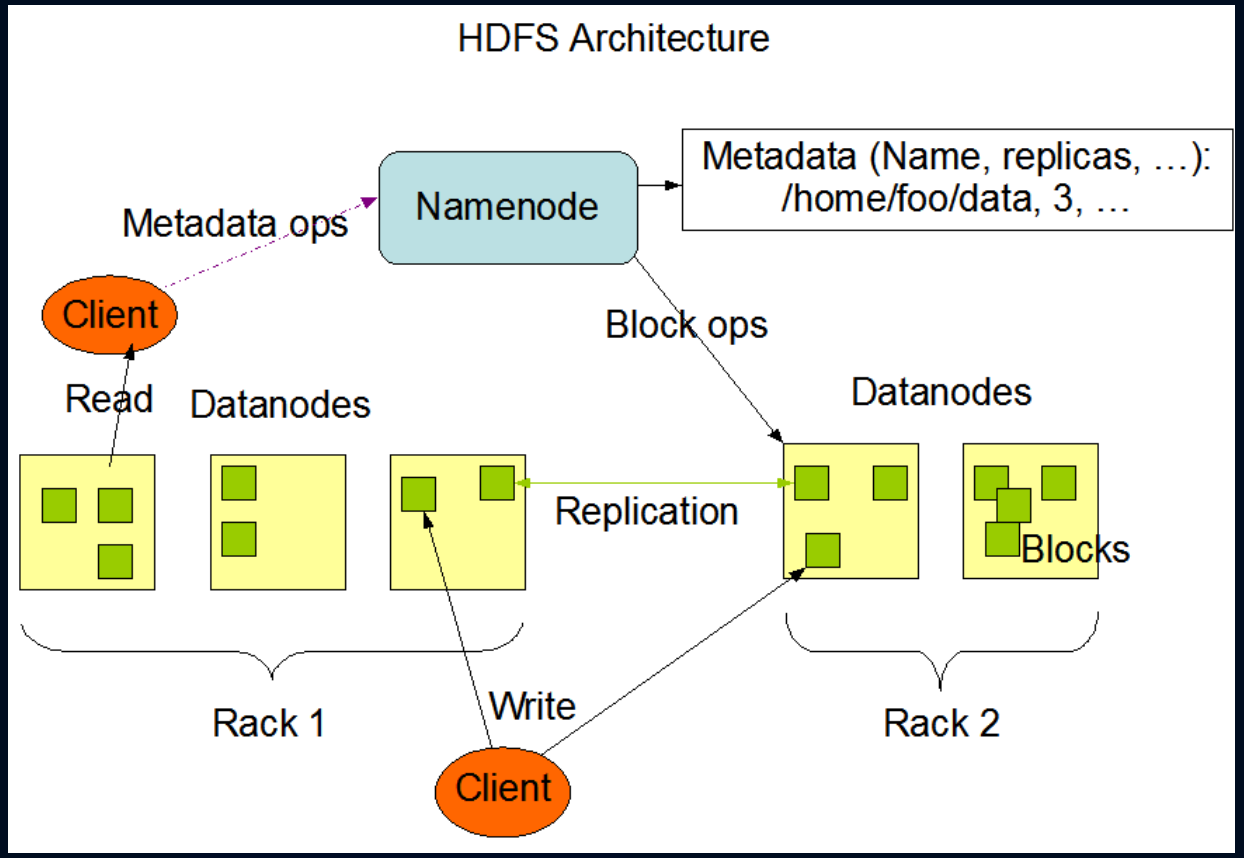


Warstwy funkcjonalne platformy Hadoop

- Warstwa sprzętowa klastra
  - Commodity hardware, niski koszt
- Warstwa rozproszonego systemu plików
  - Implementacje HDFS: Apache HDFS, IBM GPFS-FPO, Intel Luster, MapR
  - Master NameNode, slave DataNodes
  - Bloki, replikacja, wysoka dostępność (HA), odporność na awarie, skalowalność
- Warstwa zarządzania zasobami
  - Pamięć, procesory
  - Implementacje: Apache YARN, Mesos, IBM Spectrum Symphony
- Warstwa rozproszonego przetwarzania
  - Standardowo model MapReduce
  - Apache Spark, Tez, Flink, Impala, IBM BigSQL
  - ACL
- Warstwa komponentowa
  - Relacyjne podejście typu OLAP, batch, interactive, real time
  - Bazy danych NoSQL (kolumnowe, dokumentowe, key-value), obliczenia na grafach (węzły, krawędzie)
  - Przetwarzanie strumieniowe
  - Uczenie maszynowe i NLP
  - Wyszukiwanie danych
- Warstwa API
  - SQL
  - Języki programowania, biblioteki
- Warstwa wspólnych usług
  - Zarządzanie metadanymi
  - Bezpieczeństwo
  - Przepływ pracy
  - Mapowanie danych

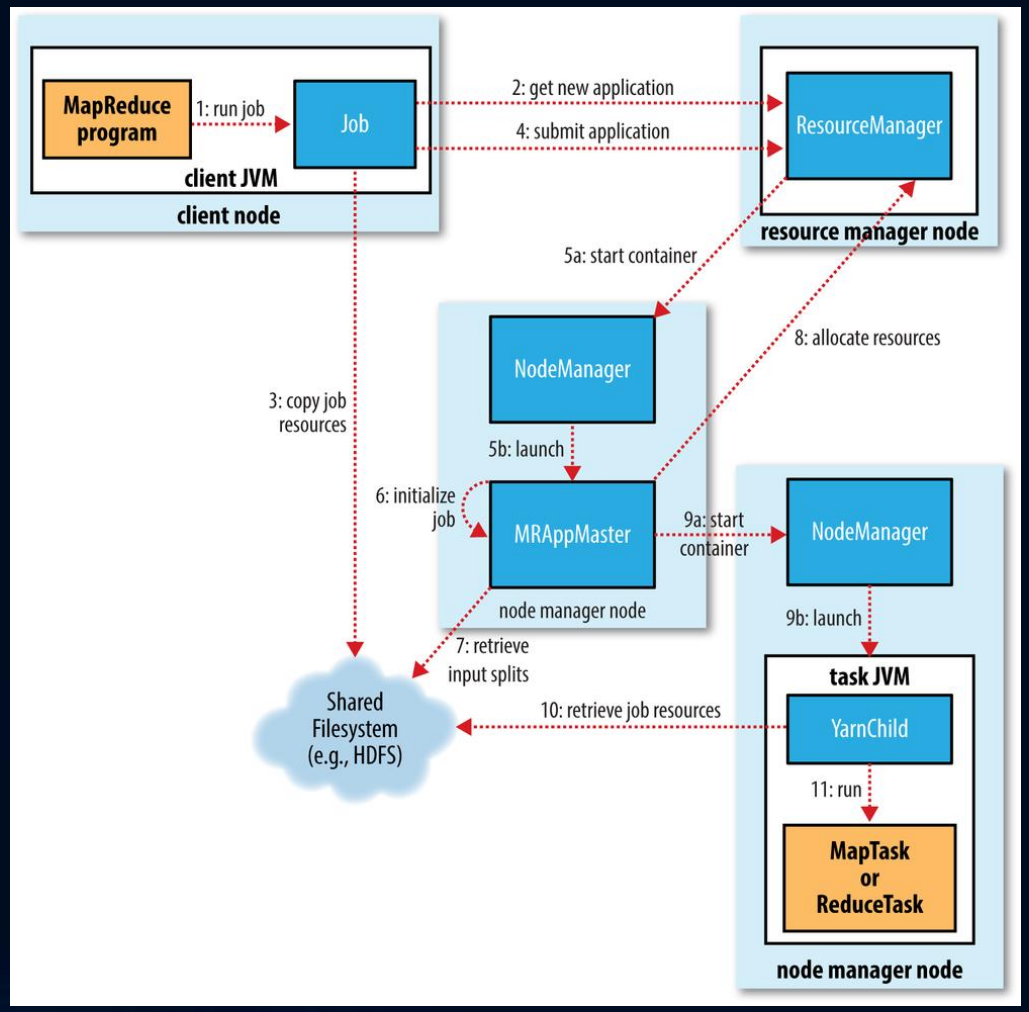
HDFS– rozproszony system plików Hadoopa

- Klaster maszyn tzw. commodity hardware
- Wydzielony zarządca NameNode
- Dane przechowywane na DataNode'ach
- Bloki danych domyślnie 128 MB oraz w 3 kopiach
- Strumieniowanie danych
- Problem małych plików
- Odporność na awarie, wysoka dostępność, skalowalność klastra

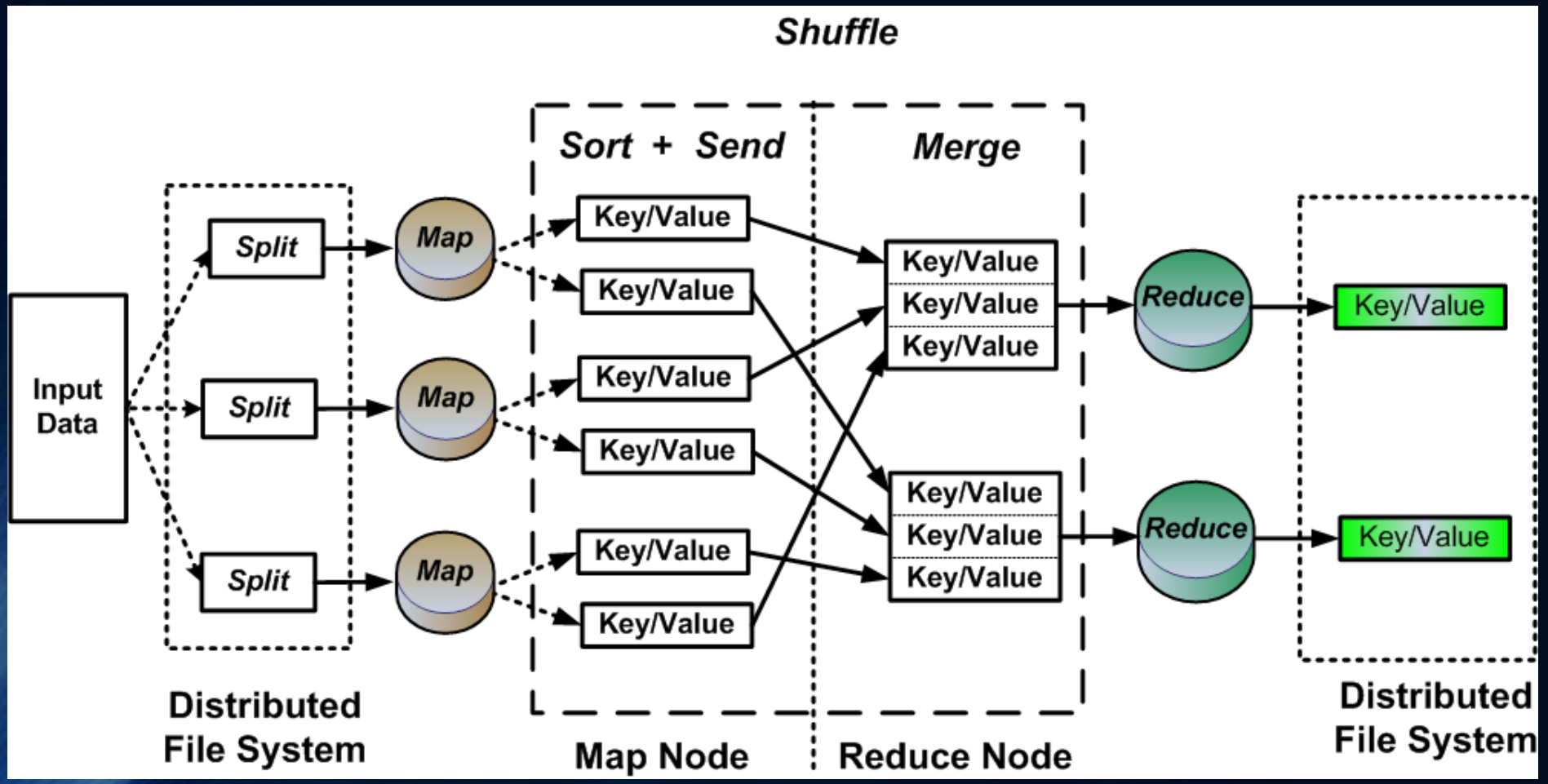


YARN – zarządca zasobów Hadoopa

- Zarządza pracą jobów
- Zarządza pamięcią oraz procesorami klastra
- Resource Manager w roli mastera
- Node Managery na slave'ach
- Dedykowany Application Master oraz Task Manager dla aplikacji
- Konteneryzacja
- Odporność na awarie



Rozproszone przetwarzanie MapReduce



Hive

- Skalowalny magazyn danych
- Stworzony przez firmę Facebook
- Dostęp do danych dowolnego formatu z użyciem SQL
- Tabele zarządzane vs zewnętrzne (managed vs external)
- Partycjonowanie danych w katalogach, by przyspieszyć dostęp do części danych
- Klastrowanie danych w tzw. wiaderka (buckets) dla optymalizacji joinów
- Serializatory i deserializatory SerDe
- User Defined Functions
- Thrift service na porcie 10000
- Dostęp poprzez Hue, Ambari oraz interaktywną konsolę CLI