Hadoop

Jakub Podeszwik

infoShare Academy

20-21.03.2019

Warsztat

- 2 dni
- 10% prezentacja
- 40% live coding
- 50% zadań głównie programistycznych

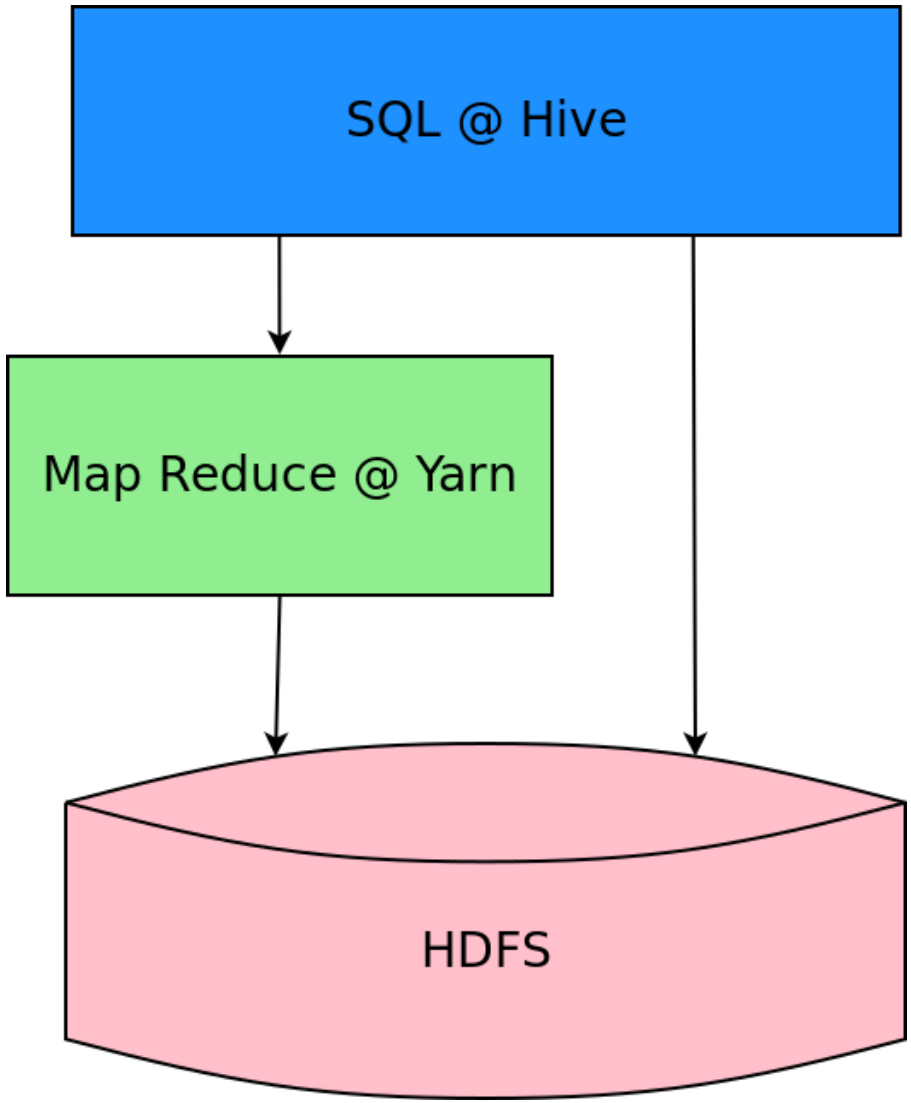
Historia

1. 2003 - Google File System
2. 2004 - MapReduce: Simplified Data Processing on Large Clusters
3. 2006 - powstaje projekt Hadoop
4. 2008 - Hadoop staje się open sourcowym projektem na licencji Apache

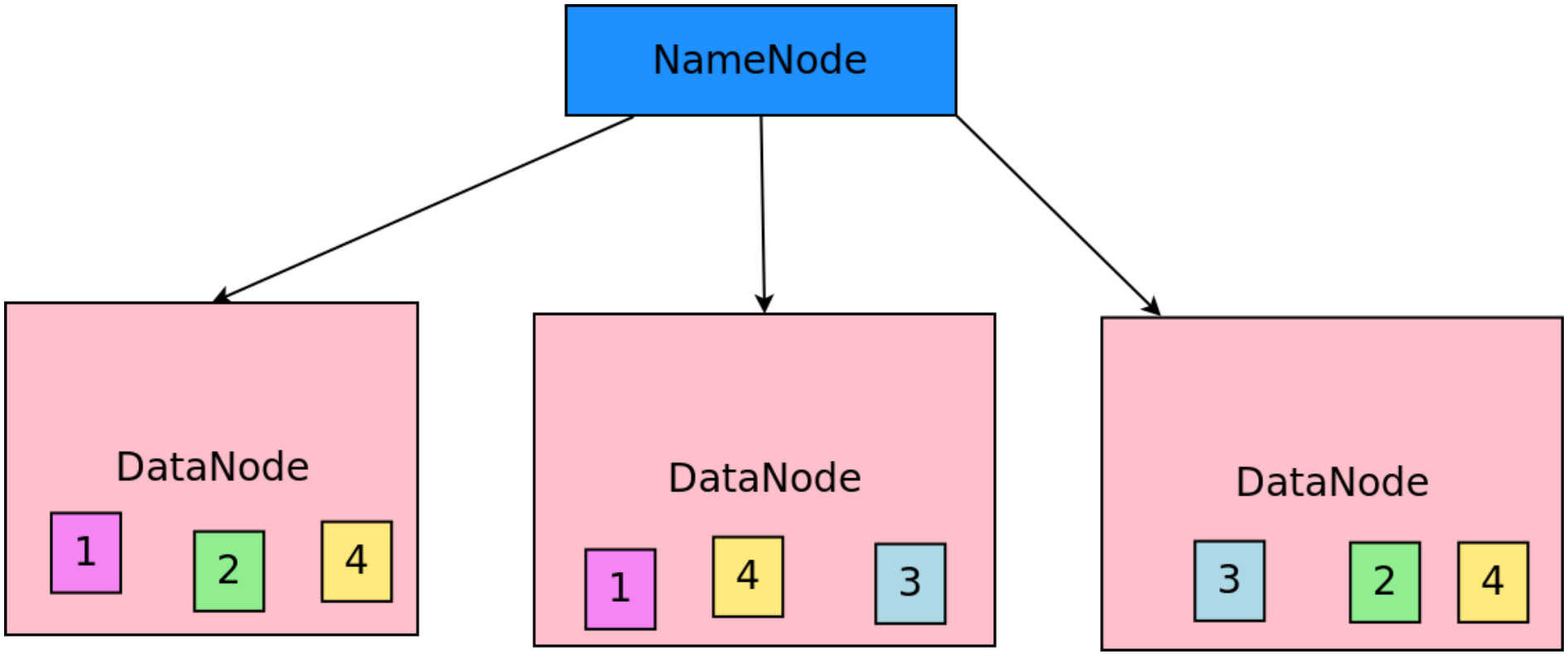
Hadoop - zalety

1. umożliwia przechowywanie i analizę wielkich zbiorów danych
2. skalowalny
3. odporny na problemy sprzętowe
4. elastyczny

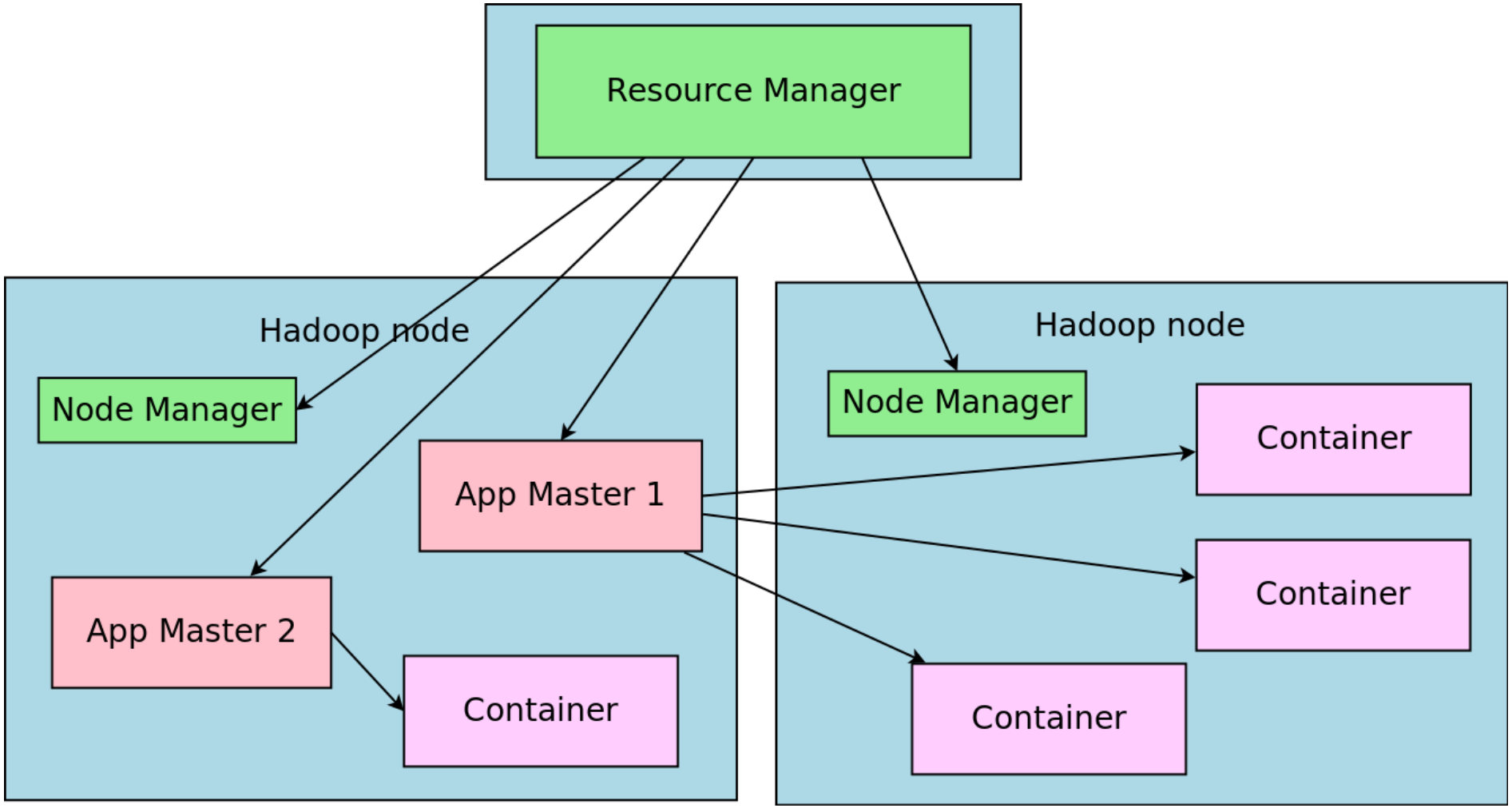
Hadoop



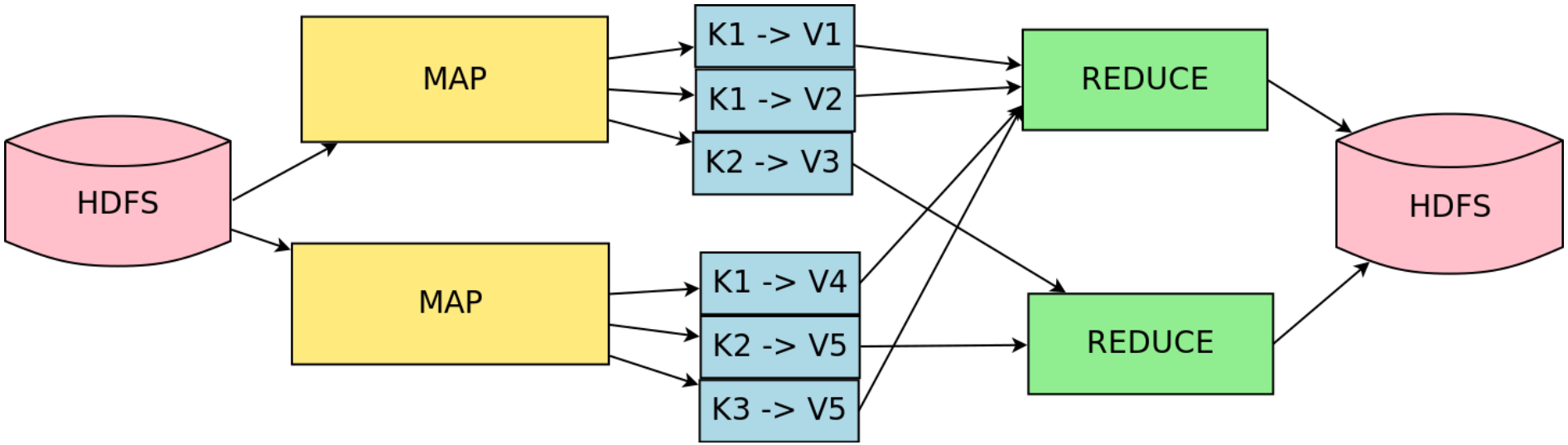
HDFS



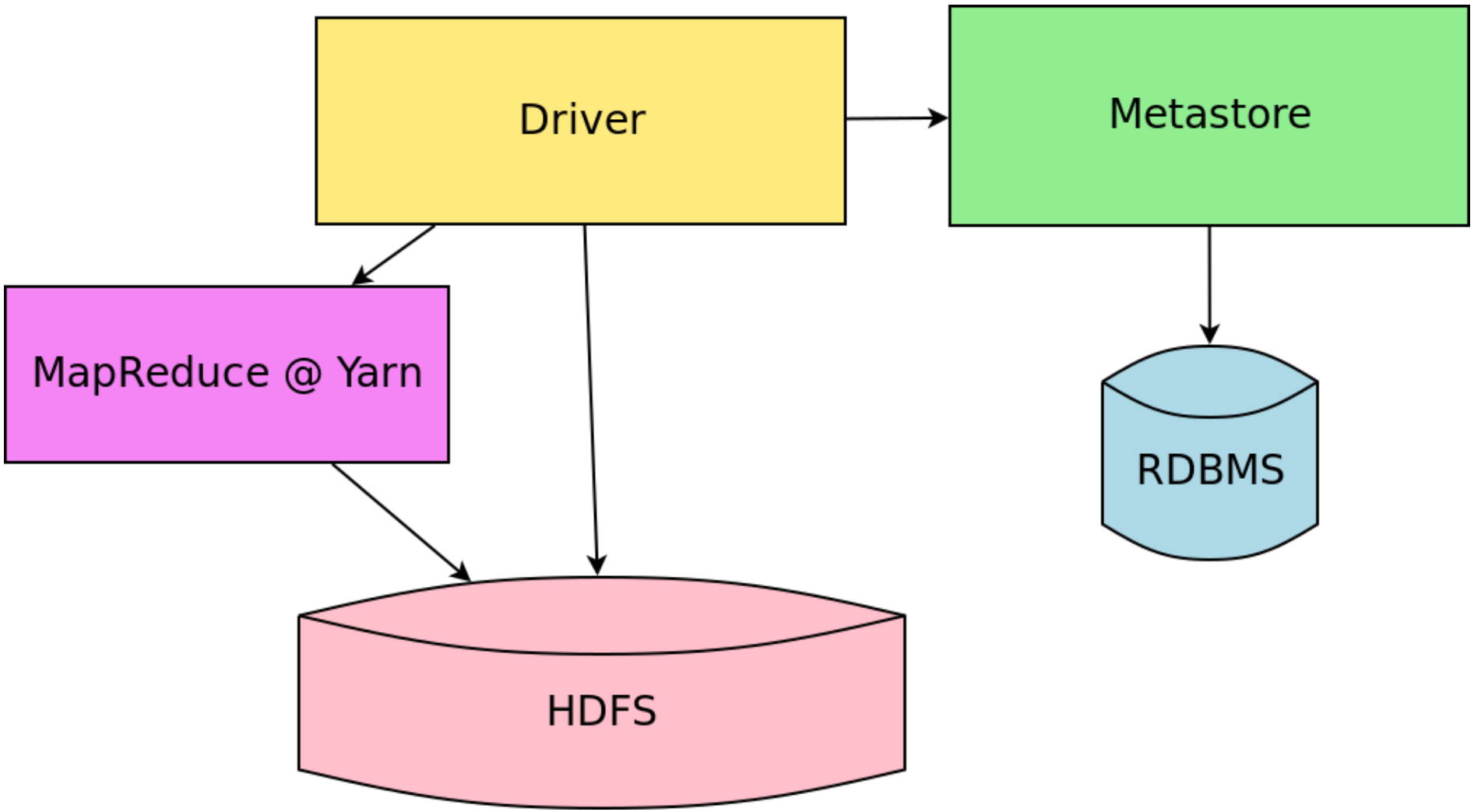
YARN


Resource Manager
Hadoop node
Node Manager
App Master 1
App Master 2
Container
Hadoop node
Node Manager
Container
Container
Container

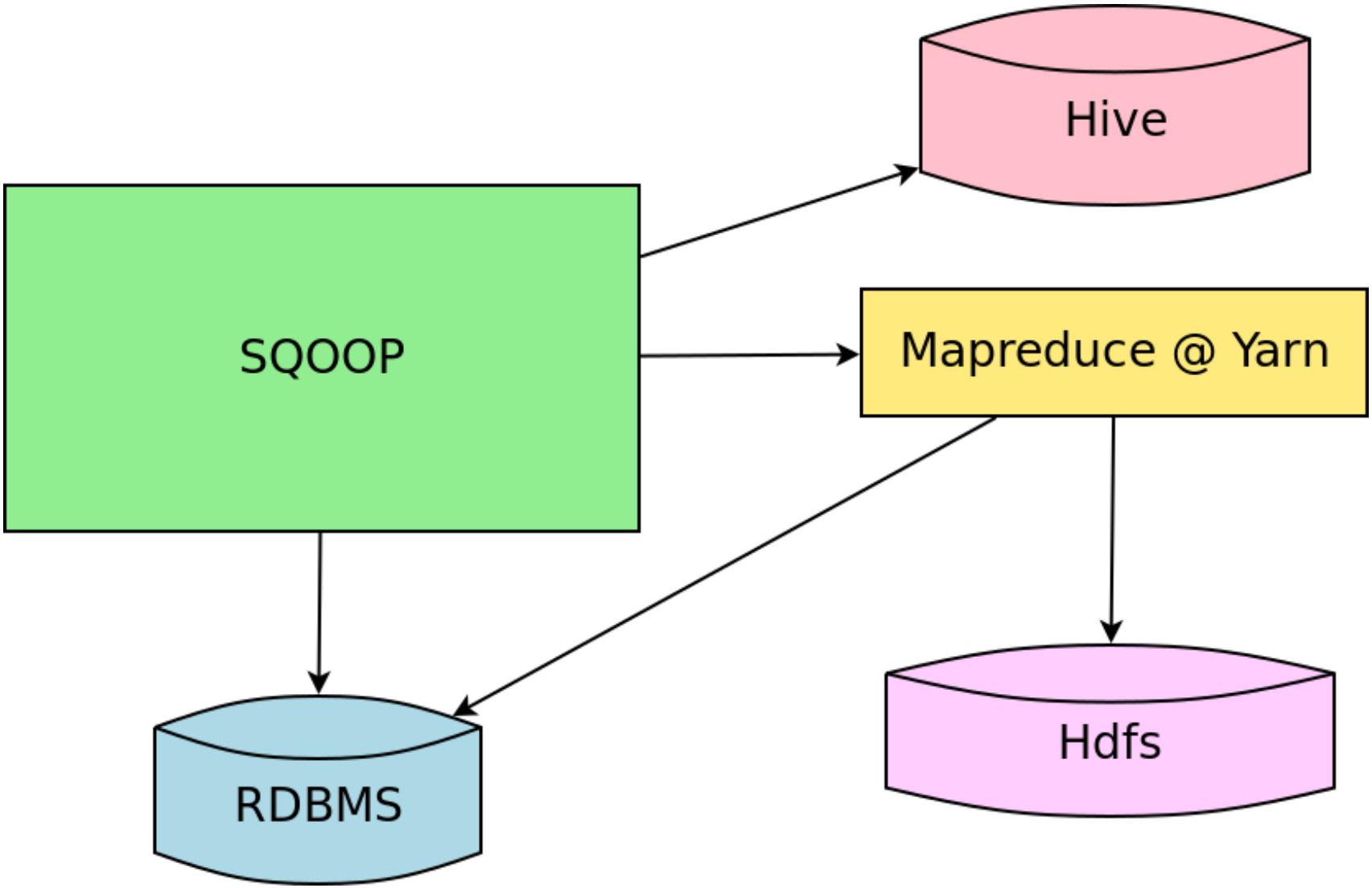
MapReduce


HDFS
MAP
MAP
K1 -> V1
K1 -> V2
K2 -> V3
K1 -> V4
K2 -> V5
K3 -> V5
REDUCE
REDUCE
HDFS

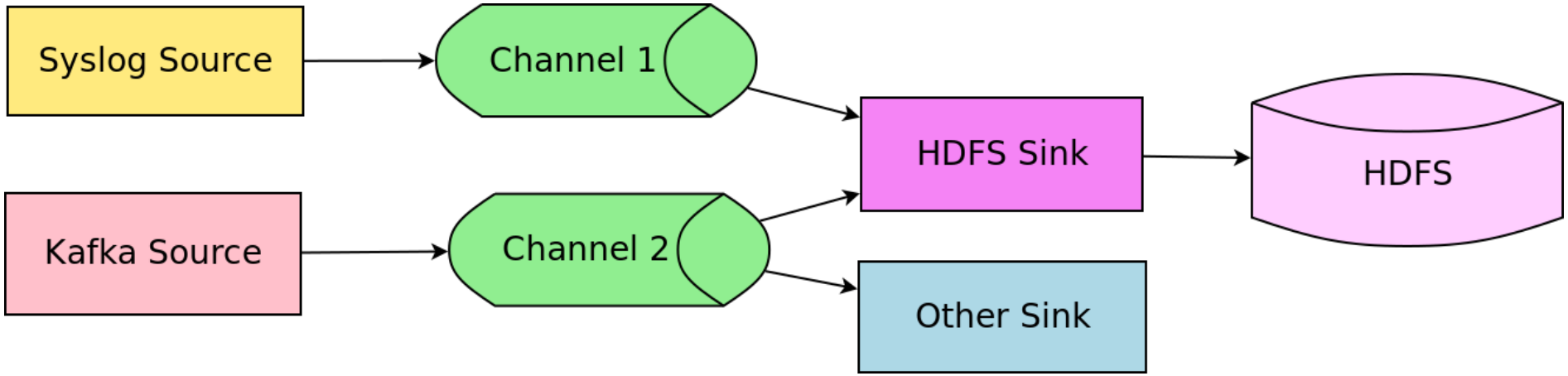
Hive



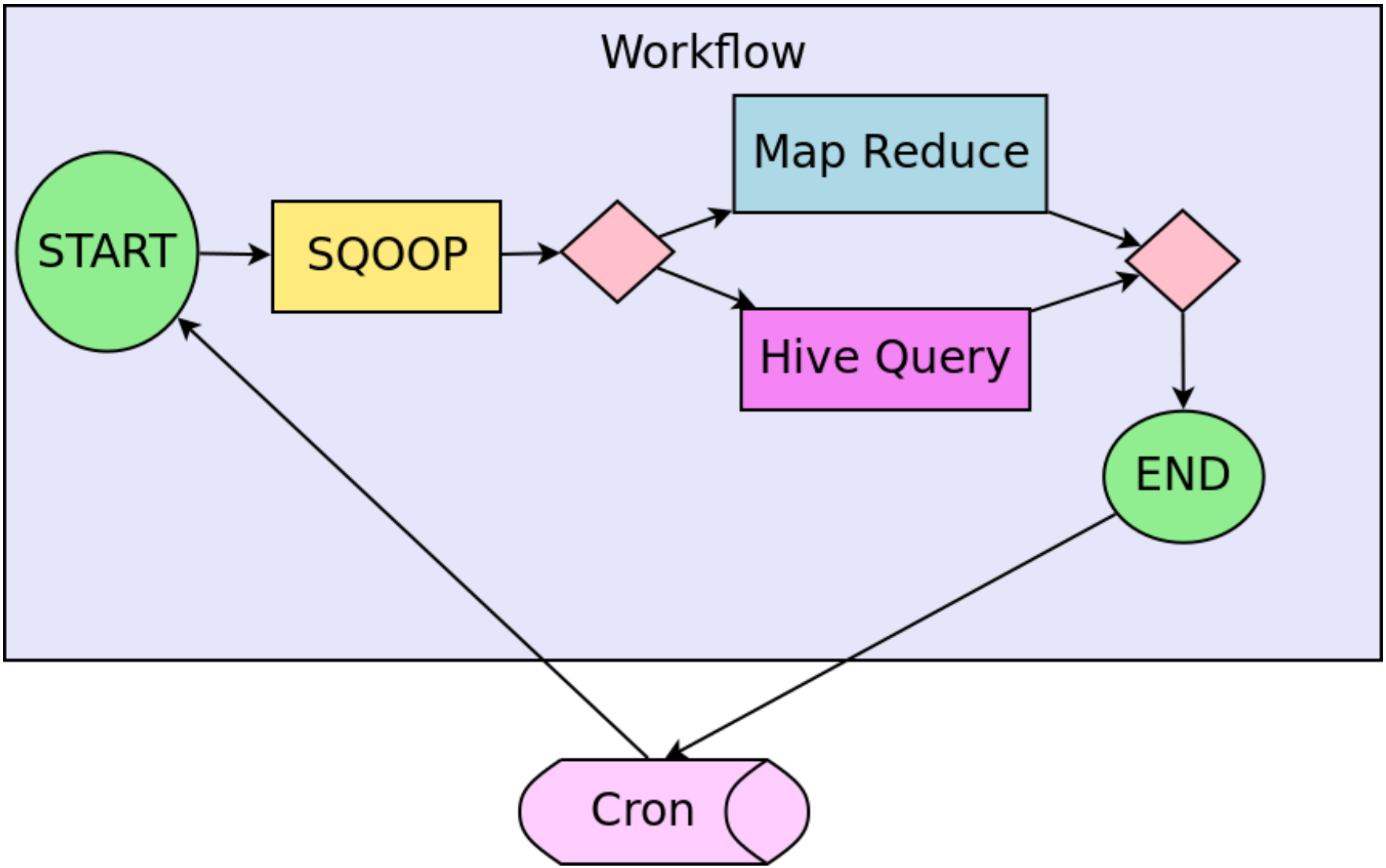
Sqoop



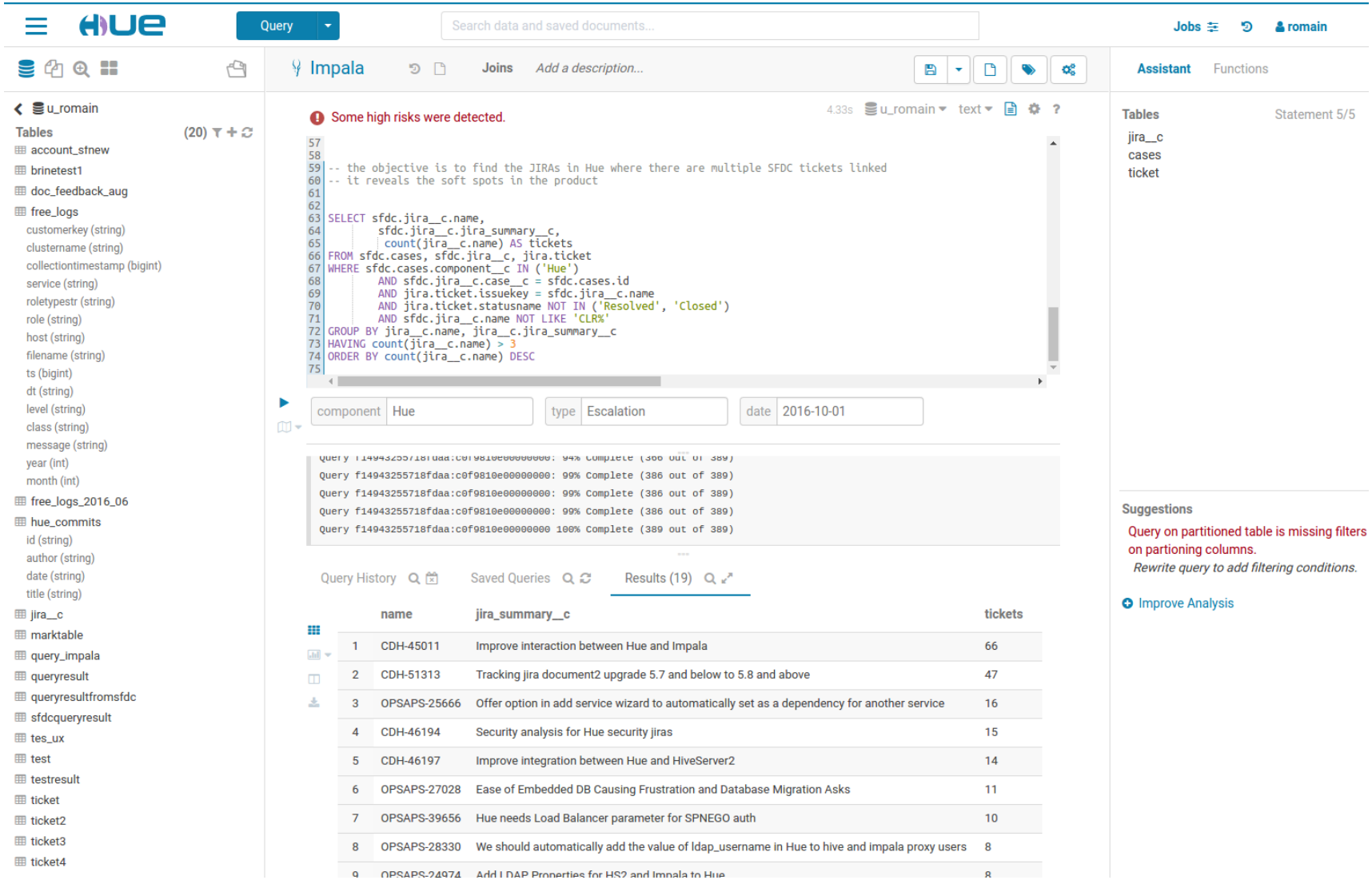
Flume


Syslog Source
Channel 1
HDFS Sink
HDFS
Kafka Source
Channel 2
Other Sink

Oozie

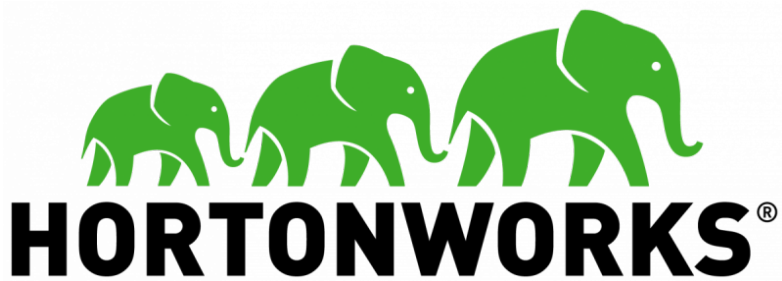


Hue



(http://gethue.com/)

Dystrybucje



Dzień 1

1. HDFS
2. MapReduce
3. Hadoop Streaming
4. Sqoop

Dzień 2

1. Flume
2. Hive
3. Hive UDFs
4. Oozie

https://github.com/infoshareacademy/hadoop-workshop-2018-03

Pytania?