

Hadoop

Jakub Podeszwik

infoShare Academy

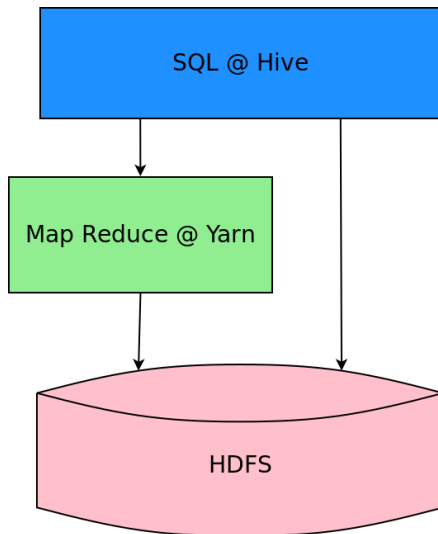
20-21.03.2019

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

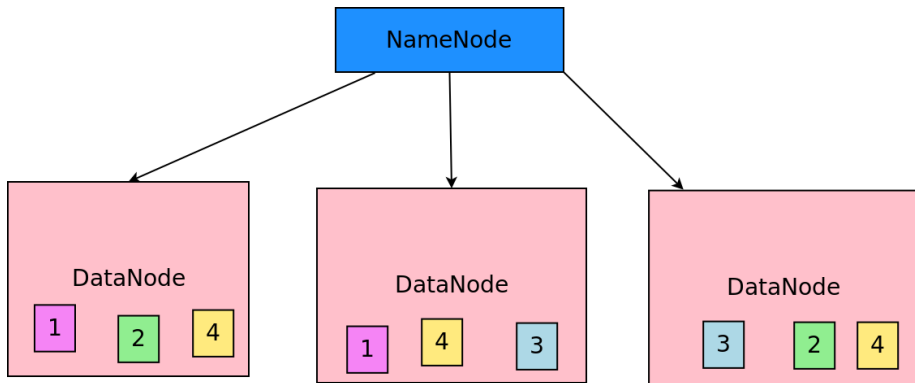
- ❶ 2003 - Google File System
- ❷ 2004 - MapReduce: Simplified Data Processing on Large Clusters
- ❸ 2006 - powstaje projekt Hadoop
- ❹ 2008 - Hadoop staje się open source'owym projektem na licencji Apache

Hadoop - zalety

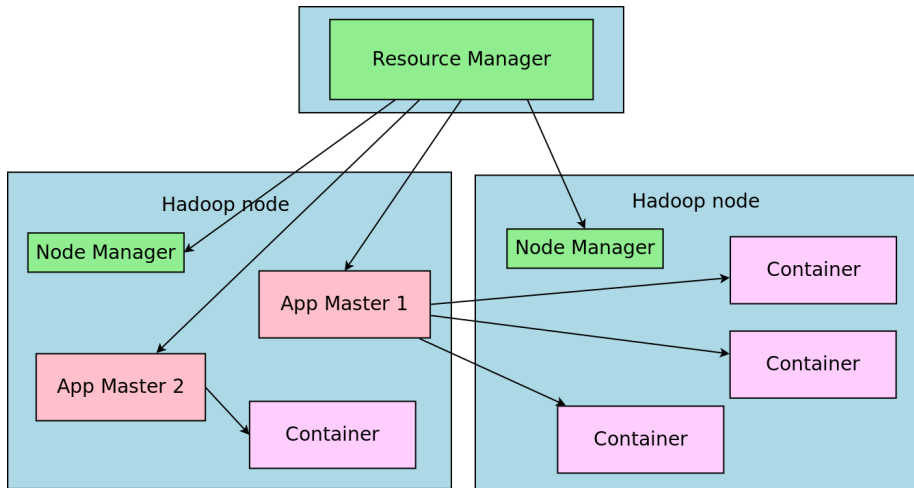
- ① umożliwia przechowywanie i analizę wielkich zbiorów danych
- ② skalowalny
- ③ odporny na problemy sprzętowe
- ④ elastyczny



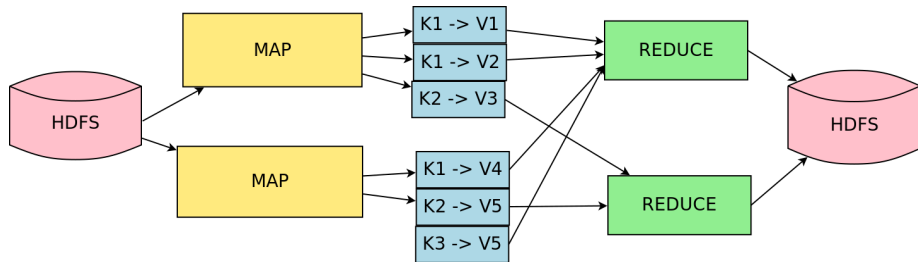
HDFS



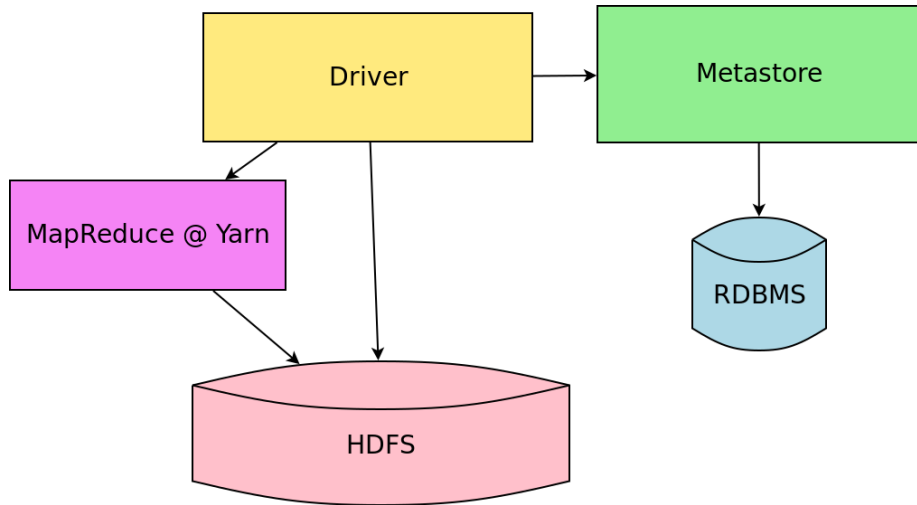
YARN



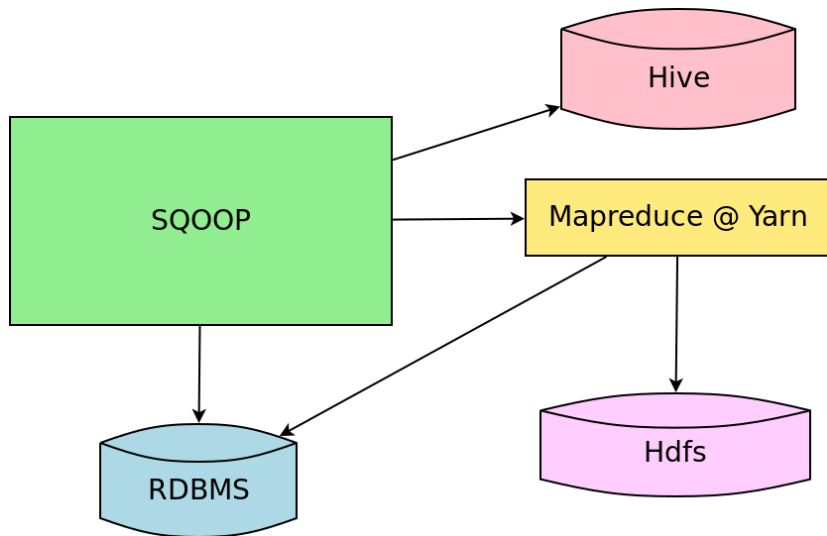
MapReduce



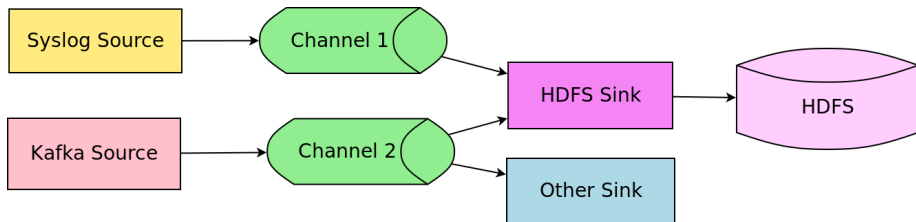
Hive

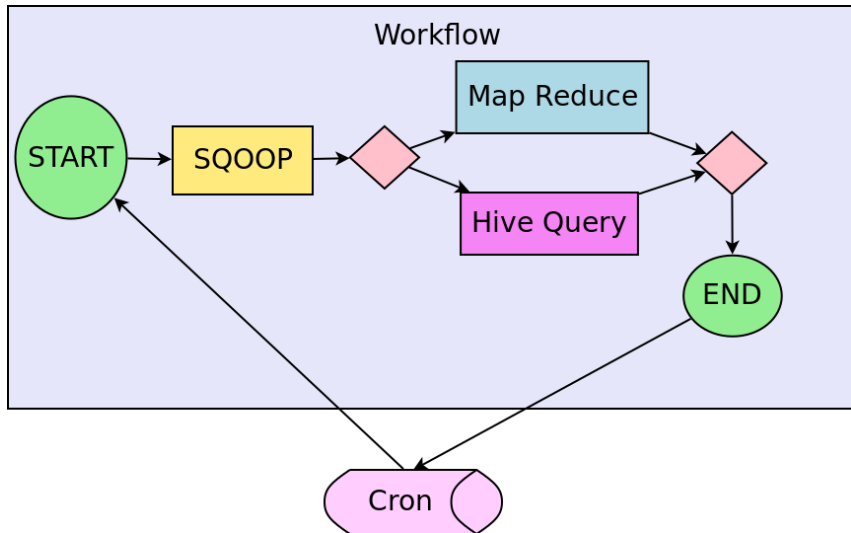


Sqoop



Flume





The screenshot displays the Hue web interface. At the top, there's a navigation bar with icons for home, search, and user profile. Below it, a search bar contains "Search data and saved documents...". The main area is divided into three panels:

- Left Panel:** A sidebar menu under "u_omain" listing various tables like "account_stnew", "brinetest1", "doc_feedback_aug", etc.
- Middle Panel:** Contains a SQL query editor. It shows a query to find JIRAs in Hue with multiple SPDC tickets linked. Below the query, there are filters for component (Hue), type (Escalation), and date (2016-10-01). The query results show 100% completion for several queries.
- Right Panel:** Displays a table titled "jira_c cases ticket" with columns for name, jira_summary_c, and tickets. It lists 10 items, starting with "CDH-45011 Improve interaction between Hue and Impala".

(<http://gethue.com/>)



Dzień 1

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

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

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

Pytania?