DSLab

The Data Science Lab

Start your engines!

- Head over to the course webpage
  - https://dslab2019.github.io/

- If you want to try the big data platform at home (with small data)
  - Download, configure & start the HDP (not HDF) Sandbox
    - *https://hortonworks.com/downloads/#sandbox (registration required)*
    - *You can choose between the VM for VirtualBox or Docker*

Final project

- **Description**: https://dslab2019.github.io/final_project/

- Any questions?

Graded Homework #1

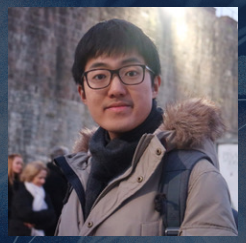
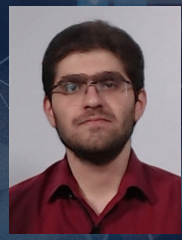
- **Solutions**: https://dslab2019.github.io/solutions/

- All questions on Mattermost

Dealing with big data

Mark



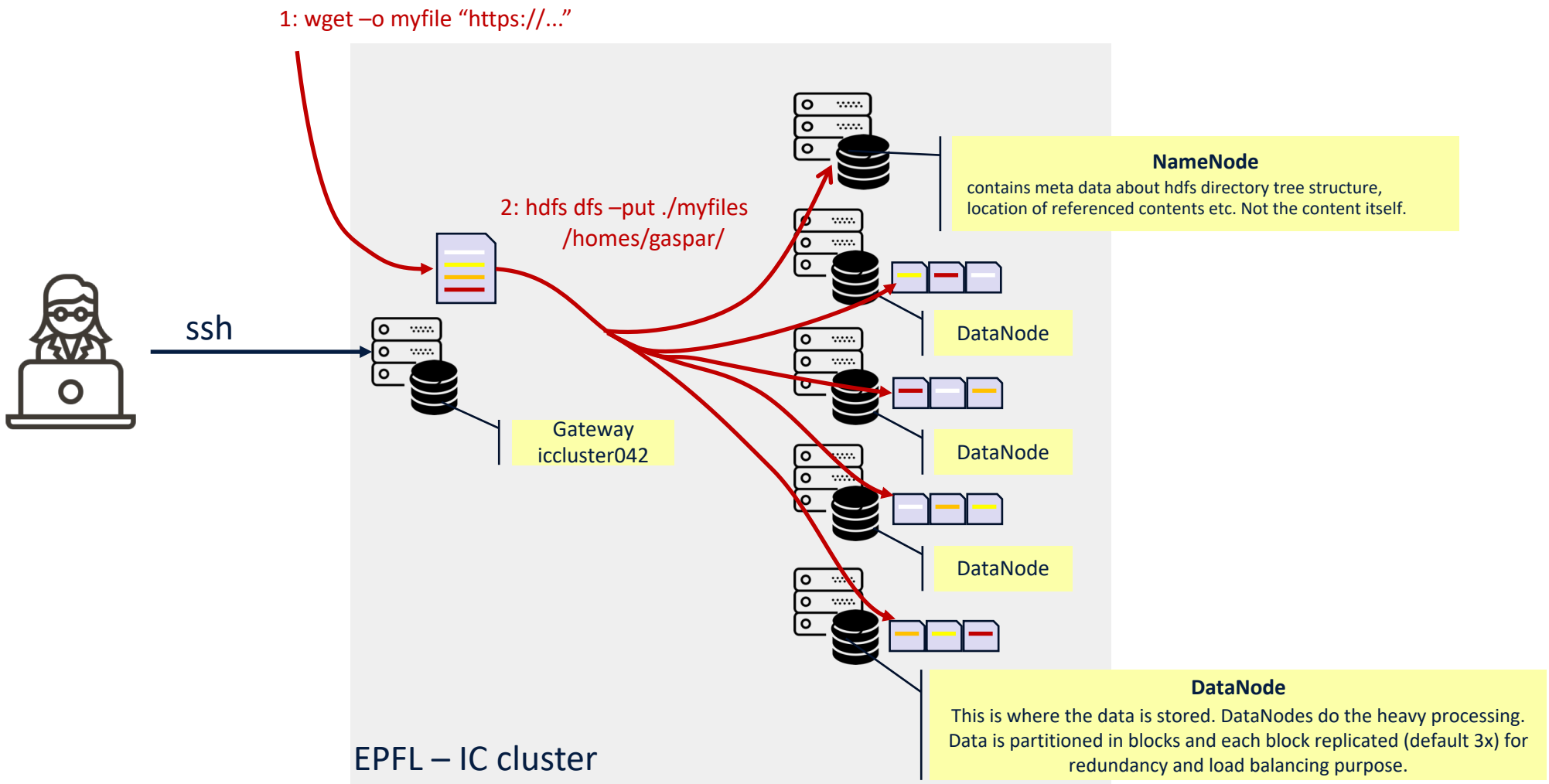

Tao, Ramtin, Mark, Apostolos, Dorina, Christine & Olivier

Spring 2019 - **week #6**

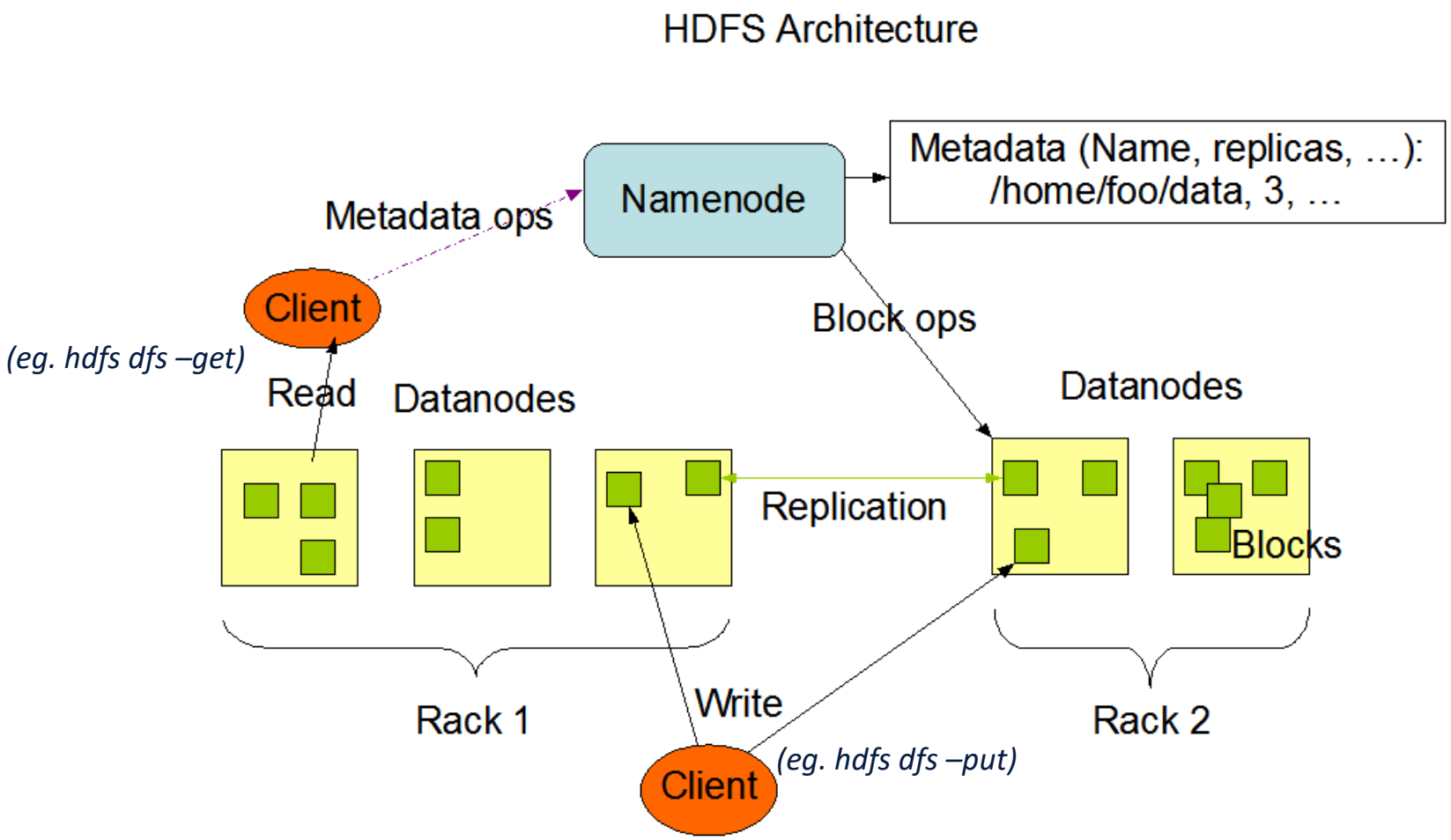
Retrospective of exercise week# 5

- First steps with Hadoop Distributed File Systems
- First steps with Hadoop Hive (Map Reduce SQL)

HDFS – under the hood

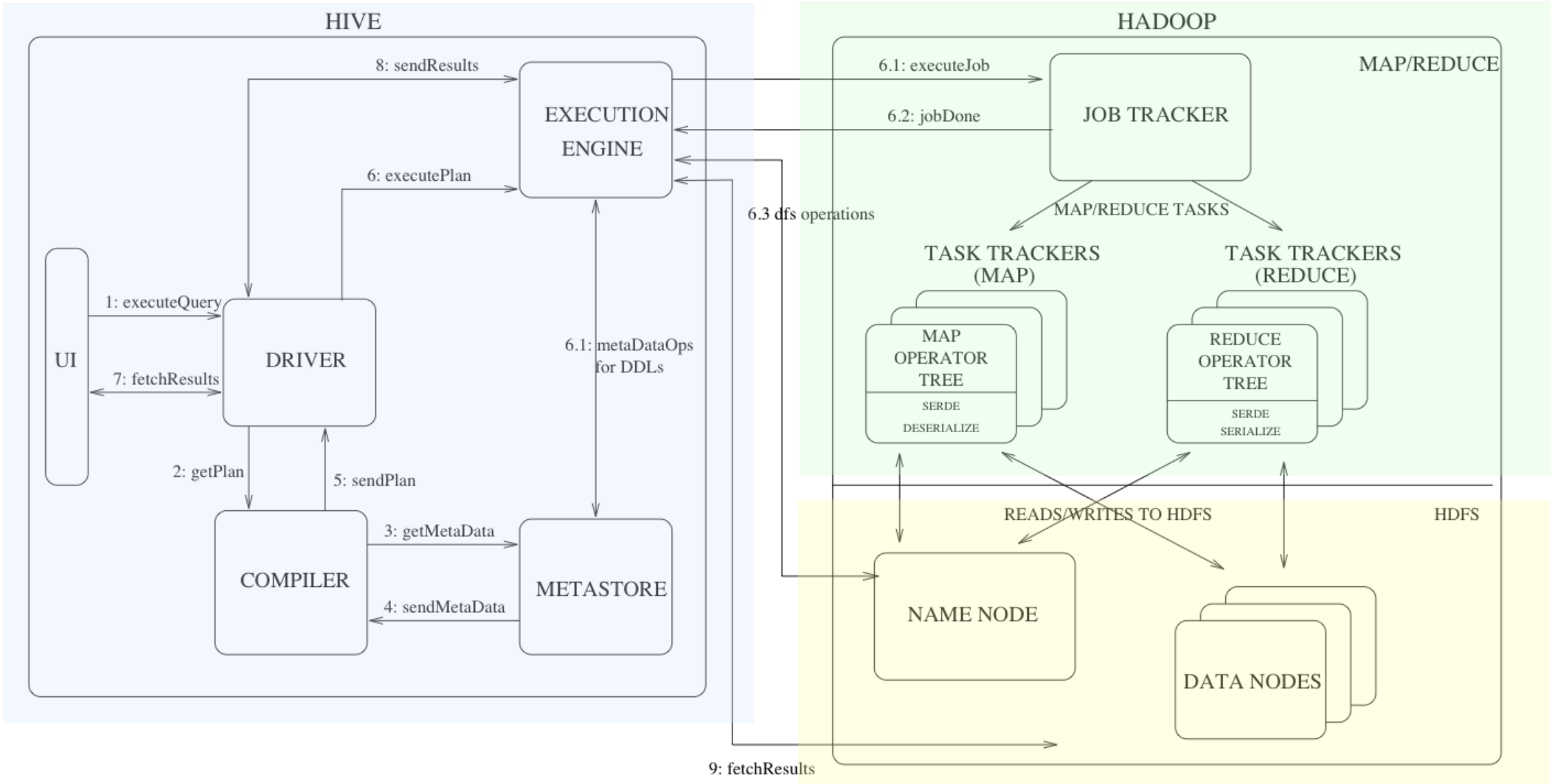


HDFS – under the hood



source: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

HIVE – under the hood



source: https://cwiki.apache.org/confluence/display/Hive/Design

Graded Homework #2

- **Description**: https://dslab2019.github.io/week6/

- Any questions?