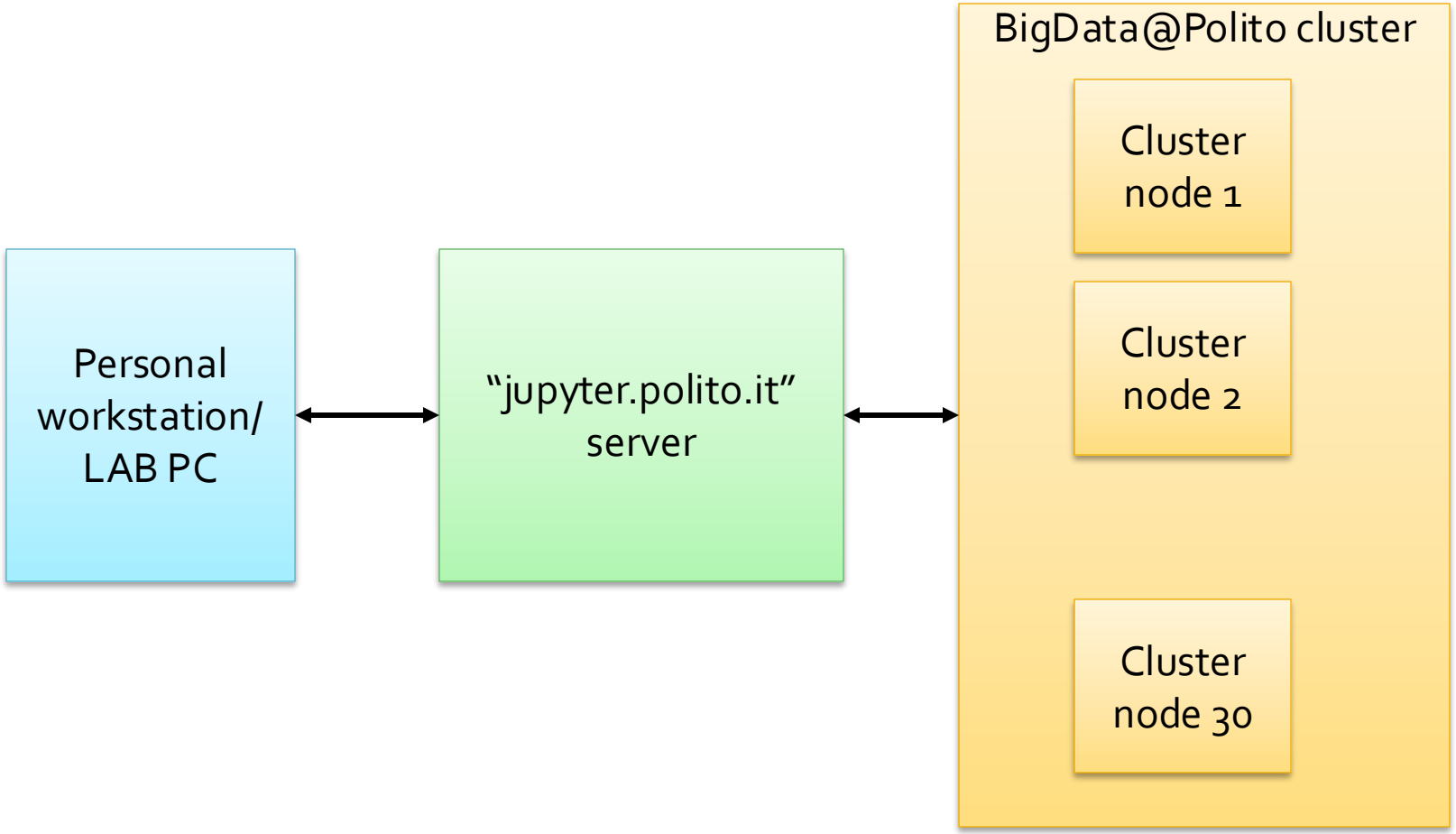
Execution of MapReduce applications

The BigData@Polito environment + Jupyter

The BigData@Polito environment

- The BigData@Polito cluster has
  - A set of servers running Spark
  - Hadoop is executed in a Single-node/Standalone mode
    - i.e., it is not executed using many servers in parallel
  - There is a Jupyter-based service to run your applications

The BigData@Polito environment



The BigData@Polito environment – Execute an application (1)

- Execute a MapReduce Application on the cluster (i.e., submit a MapReduce job on the cluster)
  - Log into jupyter
    - https://jupyter.polito.it
  - Copy the jar file containing your application from your personal workstation (or the workstation of the LAB) to the "distributed" file system of the Jupyter server
    - Drag and Drop from your PC to the Jupyter web page
  - Copy the input data of your application from the "local" file system of your personal workstation to the local/distributed file system of the Jupyter server

The BigData@Polito environment – Execute an application (2)

- Open a terminal in Jupyter
- Use the hadoop command from the opened terminal to submit the job
  - Specify the name of the jar file, the name of the input data, the name of the output folder, and the parameters/arguments of the application
  - Example

    hadoop jar Exercise1-1.0.0.jar it.polito.bigdata.hadoop.exercise1.DriverBigData 2 ex1_data ex1_out

The BigData@Polito environment – Execute an application (3)

- Open a terminal in Jupyter

This command executes an application on the cluster

- Exercise1-1.0.0.jar: jar file containing the code of the MapReduce application
- it.polito.bigdata.hadoop.exercise1.DriverBigData: driver class
- This application has three parameters
  - Number of instances of the reducer
  - Input folder
  - Output folder

- Example

```
hadoop jar Exercise1-1.0.0.jar
it.polito.bigdata.hadoop.exercise1.DriverBigData 2
ex1_data ex1_out
```