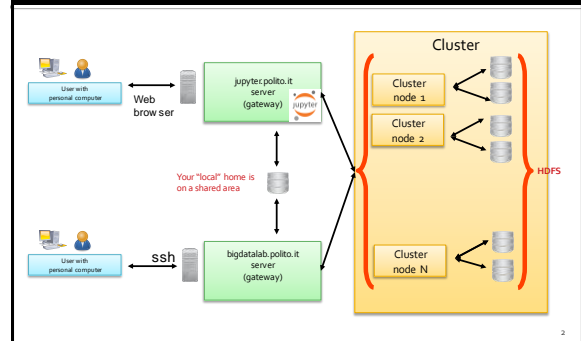


How to execute Spark applications by using Jupyter notebooks

The BigData@Polito environment



Jupyter notebooks



- Jupyter notebook
 - Browser-based interactive IDE
- Specific “notebooks” can be used to run Spark applications on the Spark cluster
 - PySpark (Local)
 - Run the application in a container (in a local instance of Spark)
 - PySpark (Yarn)
 - Run the application on the BigData@Polito cluster
- Both notebooks read/write data from/in HDFS

3

Execute an application by using PySpark on a Jupyter notebook



- Copy the input data of your application from the local drive of your personal workstation on the HDFS file system of the cluster
- Open an interactive PySpark shell by using a Jupyter notebook
- Write the python/spark code you want to execute and execute it step-by-step by using the PySpark notebook
- The result is stored in the output HDFS folder specified in your application

4

Execute an application by using PySpark on a Jupyter notebook



5