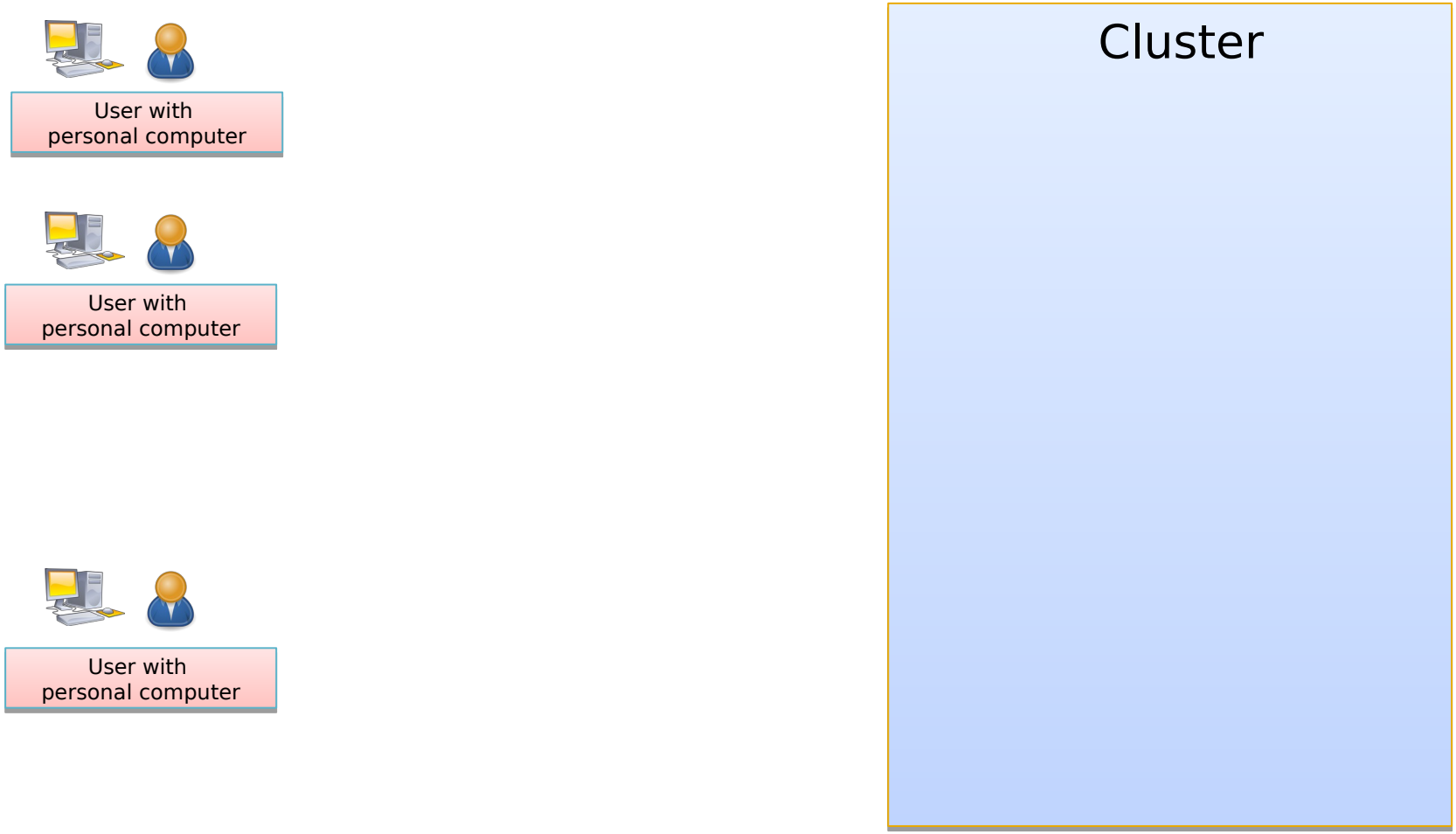
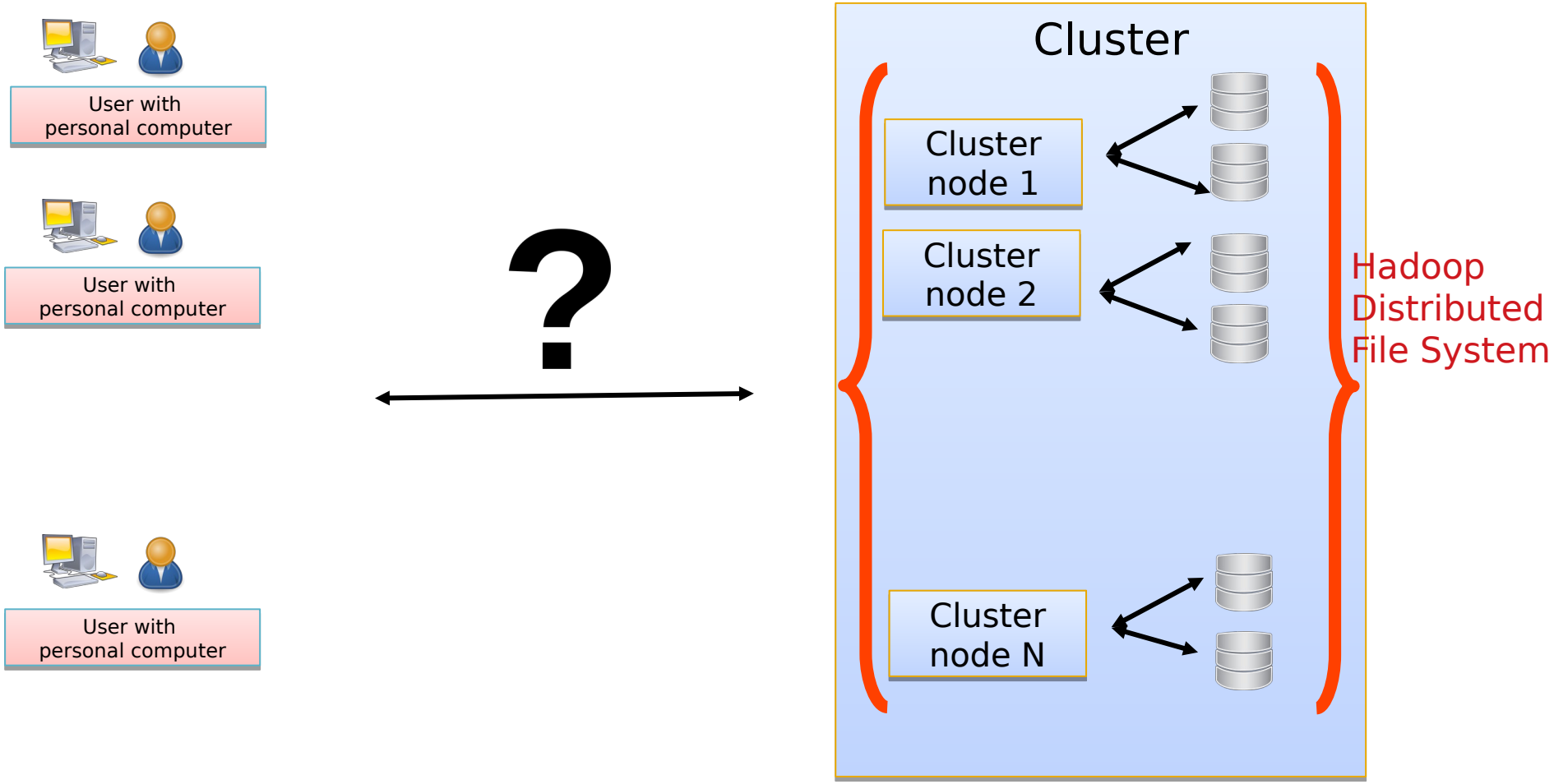
Big data for internet applications

Big Data cluster environment

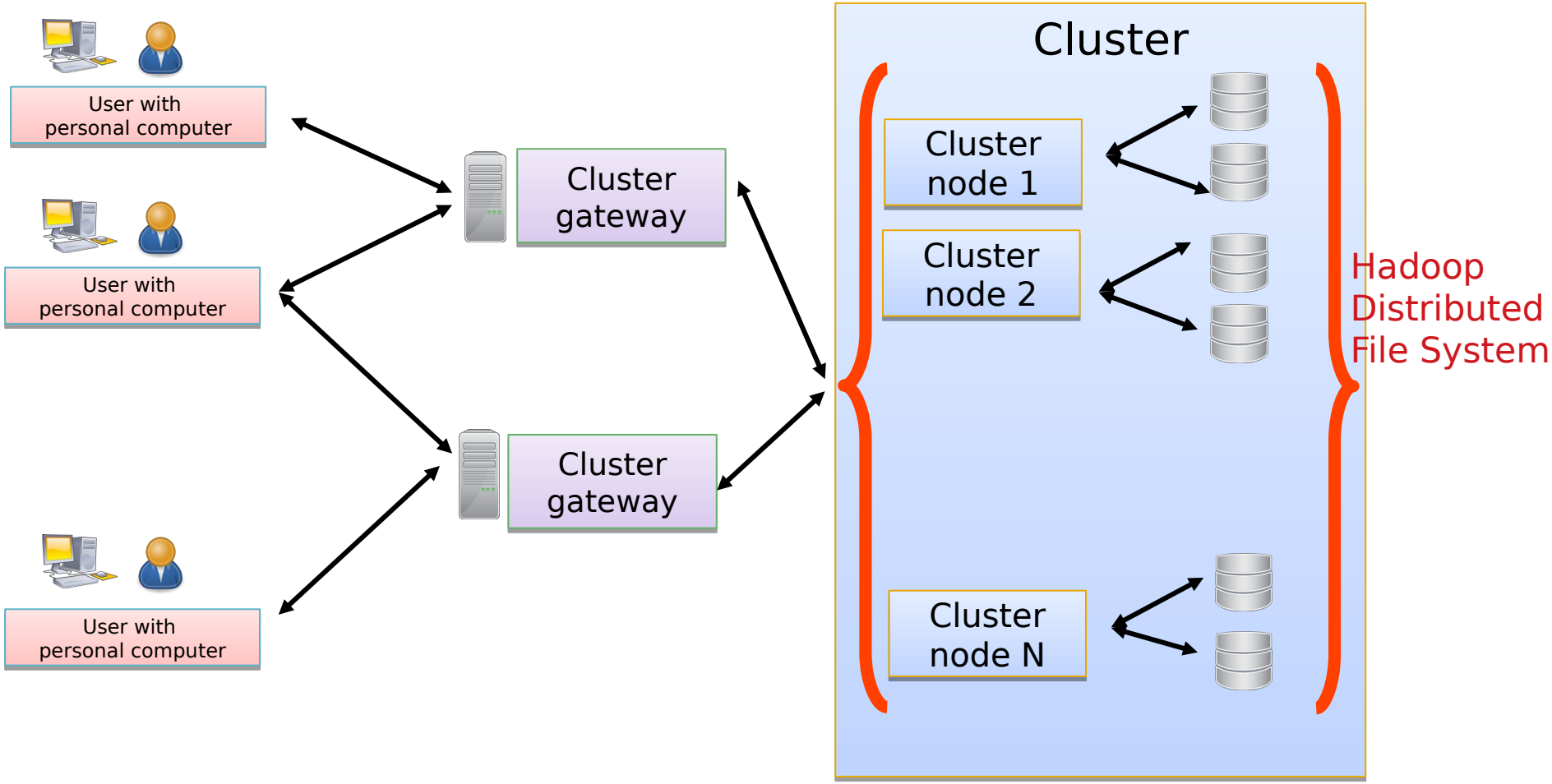
Big Data cluster environment



Big Data cluster environment



Big Data cluster environment



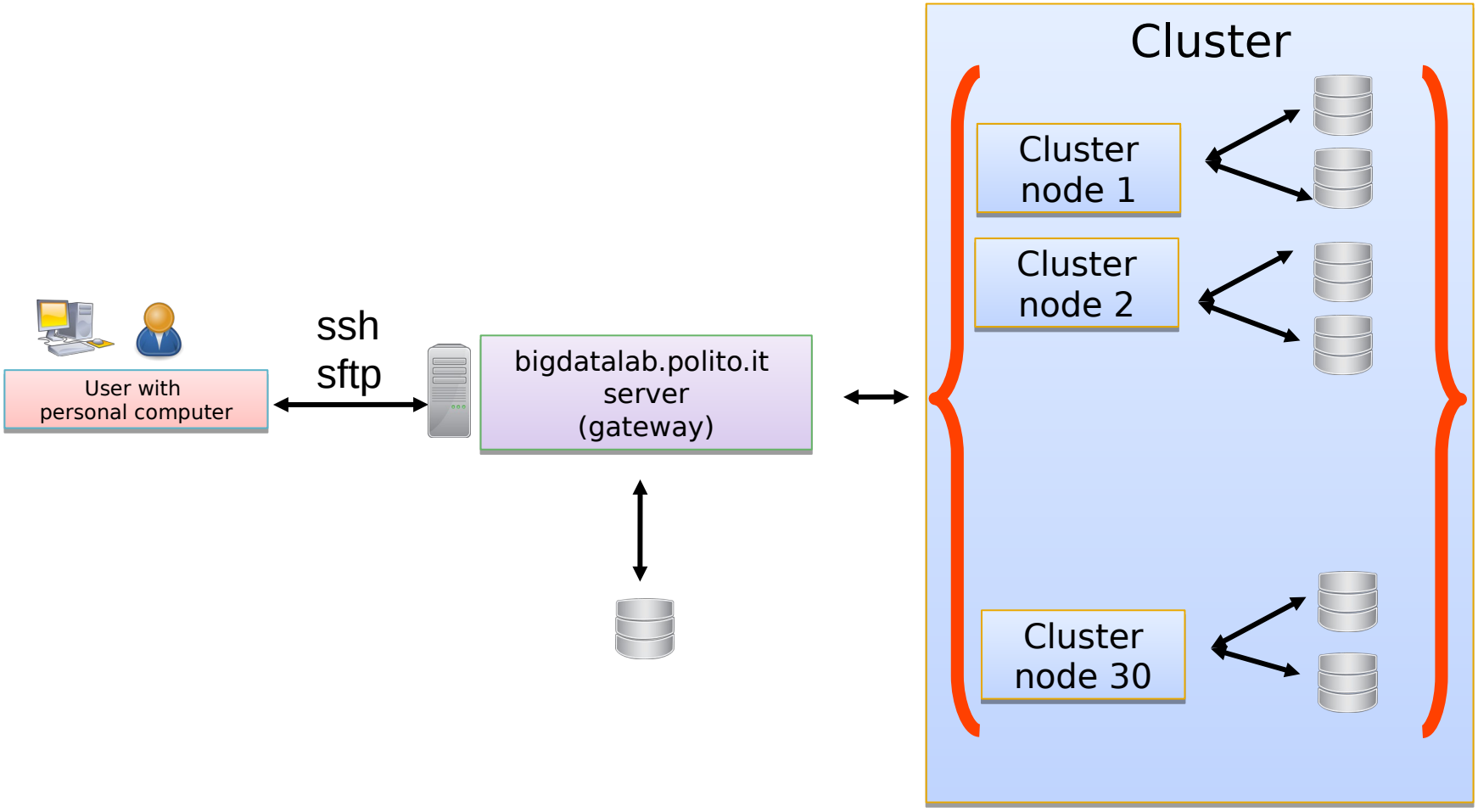
The BigData@Polito environment

The BigData@Polito environment

- The BigData@Polito cluster has
  - A set of ~60 servers running Hadoop
  - Three **Access Gateway** servers used to interact with the Hadoop cluster

1)bigdatalab.polito.it
2)hue.polito.it
3)jupyter.polito.it

The BigData@Polito environment



Ssh/sftp access to our BigData cluster

bigdatalab.polito.it is an **Access Gateway** server used to interact with the Hadoop cluster.
Connecting to it through ssh protocol you can:

- Submit jobs/execute Spark-based applications
- Submit hdfs commands (Transfer files,...)
- Analyze the log files from command line
- Interact with the local file system of the gateway

Execute an application by using PySpark through ssh

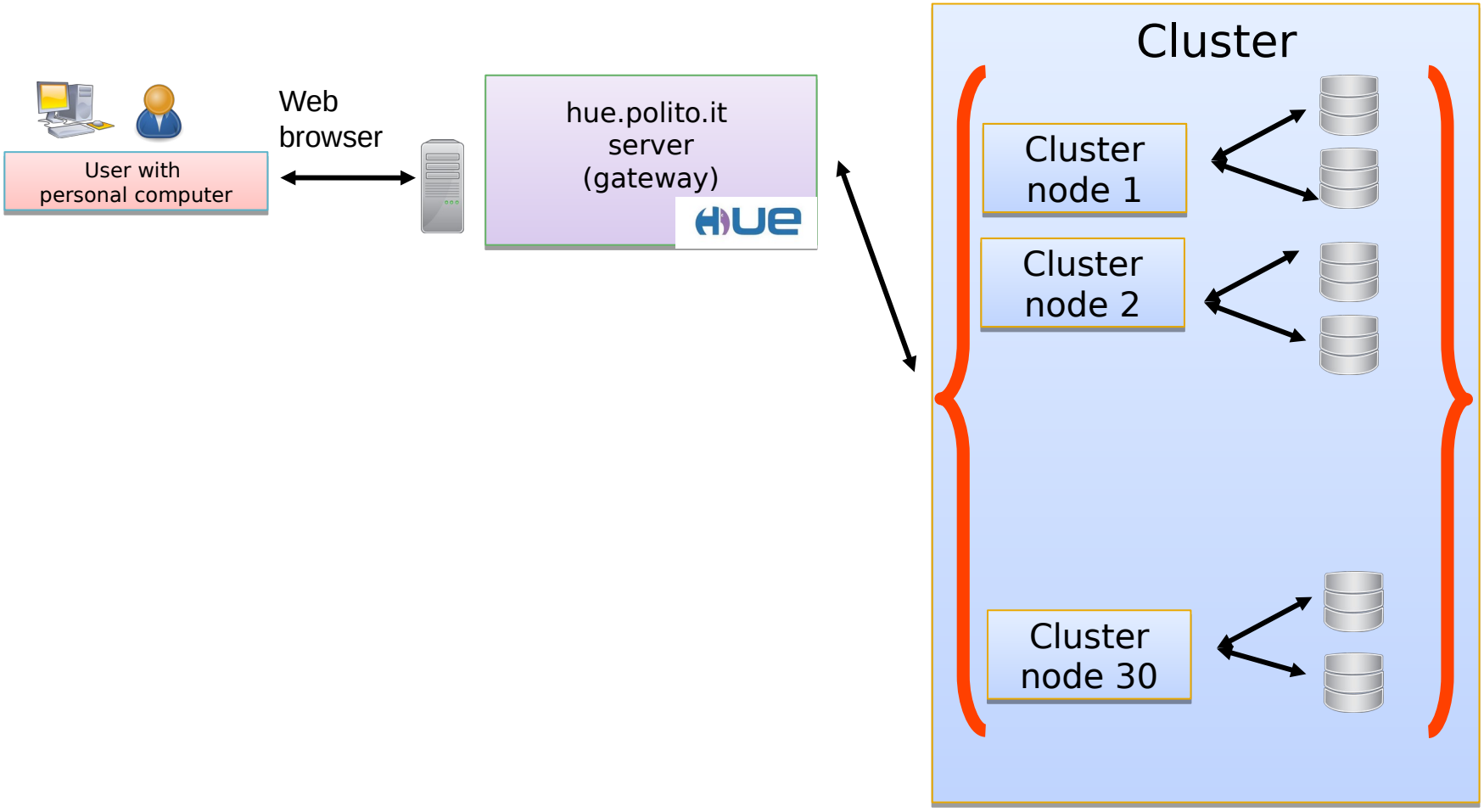
- Copy the input data of your application from the local drive of your personal workstation or from the gateway on the HDFS file system of the cluster
  - Use hdfs from command line
  - Or use the HUE web interface

- Open an interactive PySpark shell
- Write the python/spark code you want to execute and execute it step-by-step

Execute a standalone application by using a spark-submit

- Copy the input data of your application from the local drive of your personal workstation or from the gateway on the HDFS file system of the cluster
- Copy the python file containing your application from your personal workstation on the mounted file system of **bigdatalab.polito.it**
- Open a linux shell on a gateway **bigdatalab.polito.it** on which spark-submit is installed and configured
- Use the spark-submit command from the linux shell to submit your applications

The BigData@Polito environment



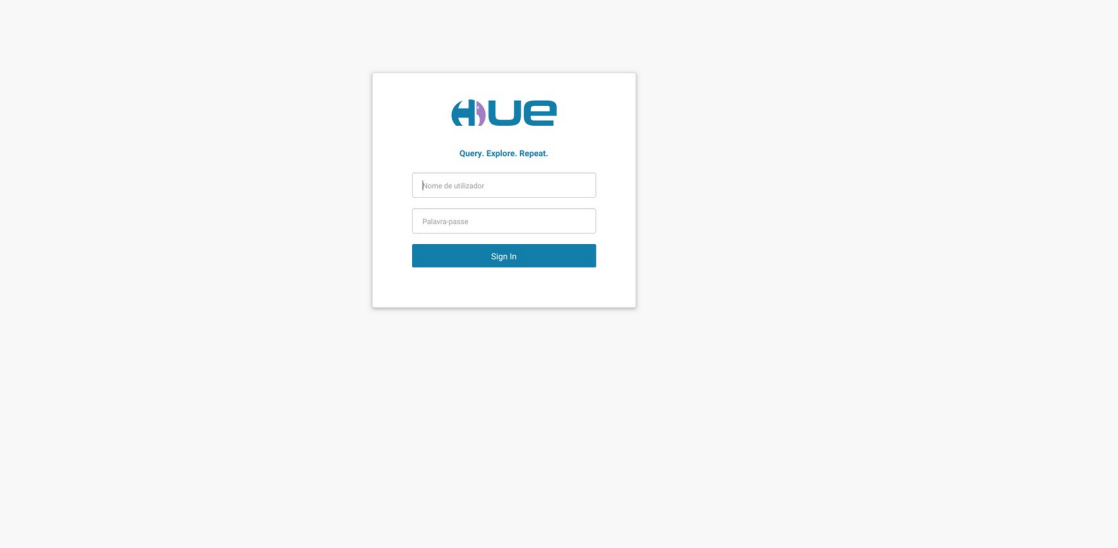
Hue hdfs/jobs web interface for our BigData cluster

hue.polito.it is an **Access Gateway** server used to interact with the Hadoop cluster. Connecting to it through a web browser you can:

- Interact with the hdfs
- Analyze your jobs on the cluster (yarn)

Hue hdfs/jobs web interface for our BigData cluster

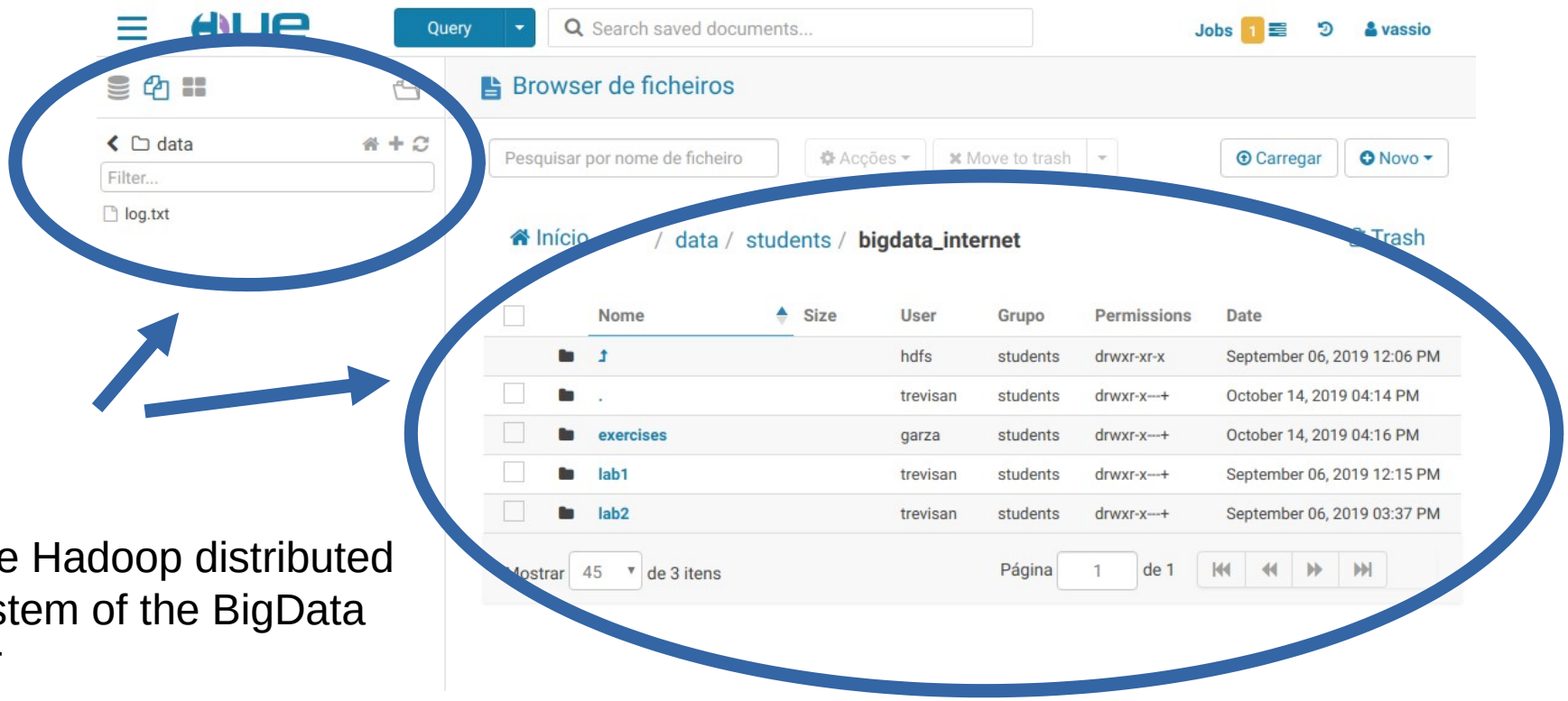
https://hue.polito.it



You should have received credential to your studenti.polito.it email

Hue hdfs/jobs web interface for our BigData cluster

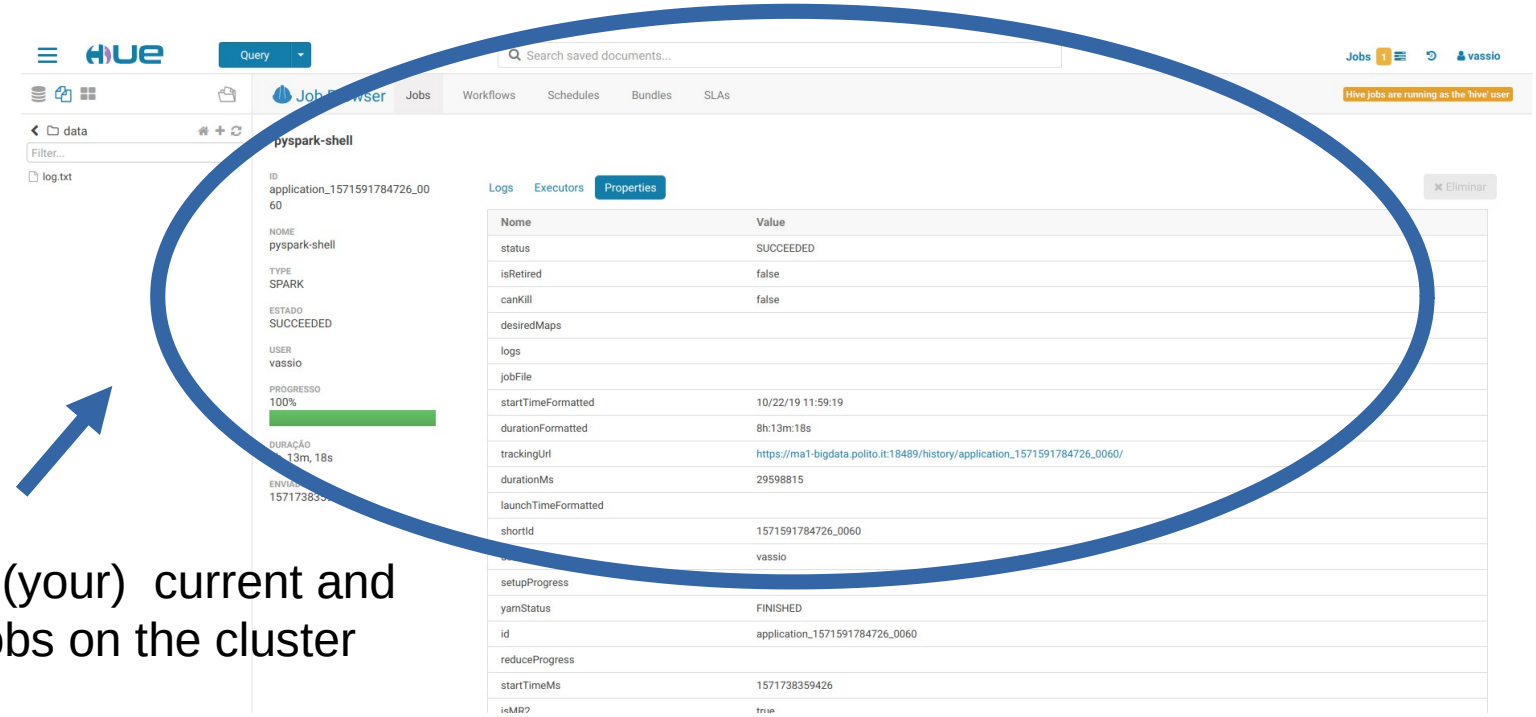
https://hue.polito.it



Browse Hadoop distributed file system of the BigData cluster

Hue hdfs/jobs web interface for our BigData cluster

https://hue.polito.it



Browse (your) current and active jobs on the cluster (yarn)

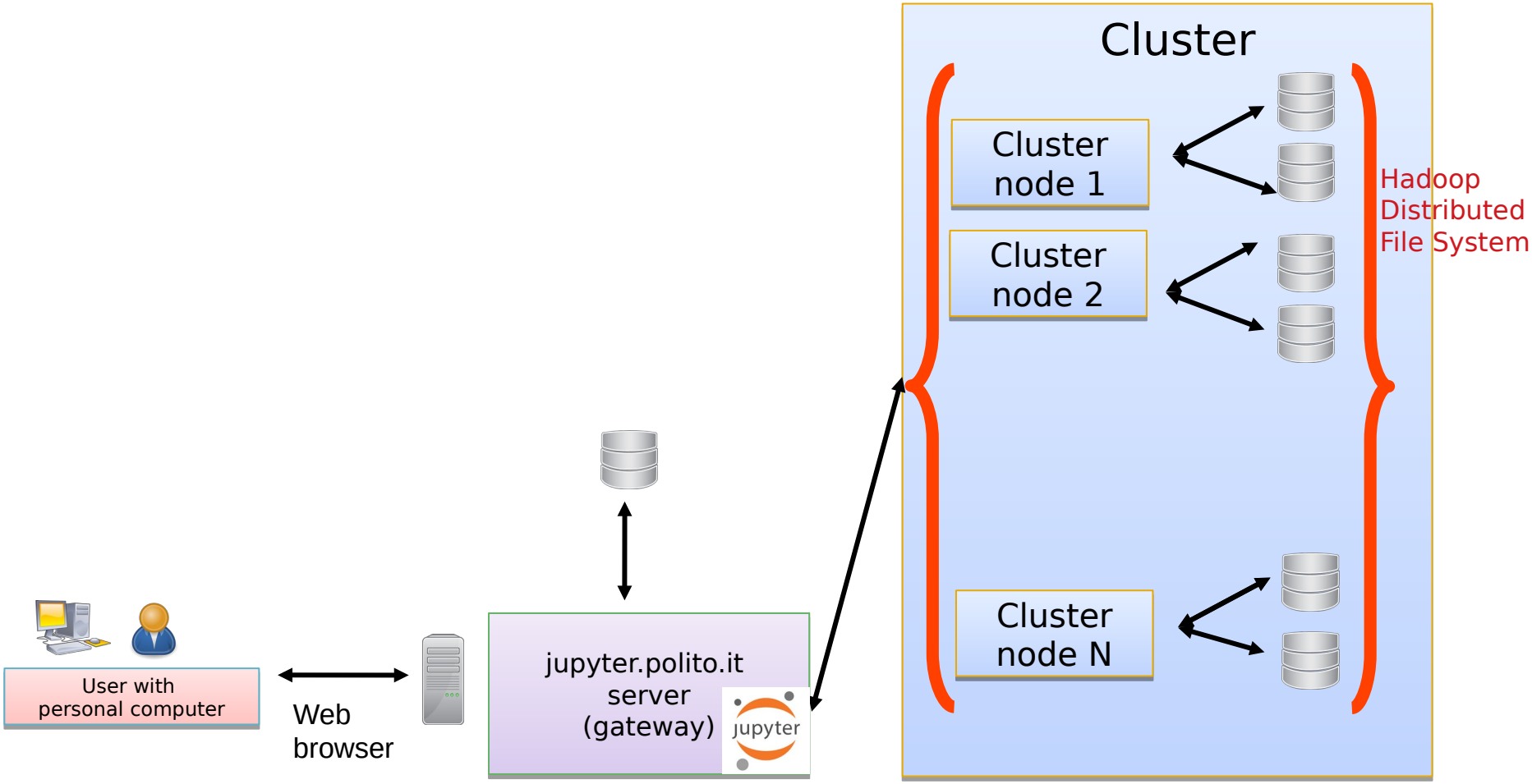
Jupyter notebooks

Jupyter notebook - browser-based interactive IDE. JSON document, containing an ordered list of input/output cells which can contain code, text (using Markdown), ...

Jupyter notebooks

- Can execute just part of a Python/Spark code
- Allows the user to include formatted text
- Can mix visualization of the results with comments and code
- Notebooks saved in a way that lets other people open them and execute the code on their own systems
- Ideal for creating reports and doing data science experiments

The BigData@Polito environment



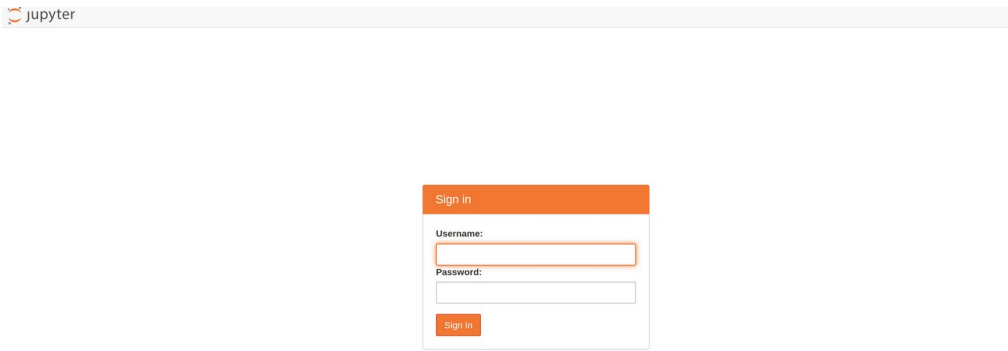
Jupyter web interface to our BigData cluster

jupyter.polito.it is an **Access Gateway** server used to interact with the Hadoop cluster.
Connecting to it through ssh protocol you can:

- **Execute interactive Spark Jupyter notebook**
- Submit jobs/execute Spark-based applications
- Submit hdfs commands (Transfer files,...)
- Analyze the log files from command line
- Interact with the local file system of the gateway

Jupyter web interface for our BigData cluster

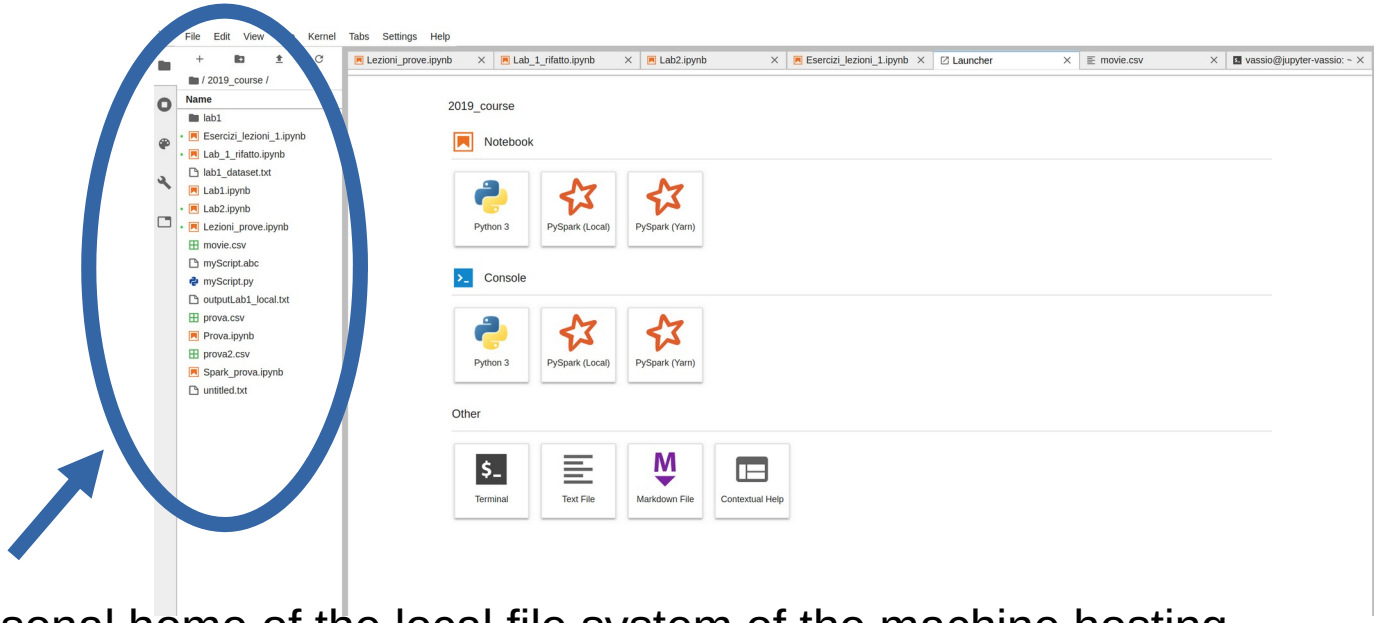
https://jupyter.polito.it/



You should have received credential to your studenti.polito.it email

Jupyter web interface for our BigData cluster

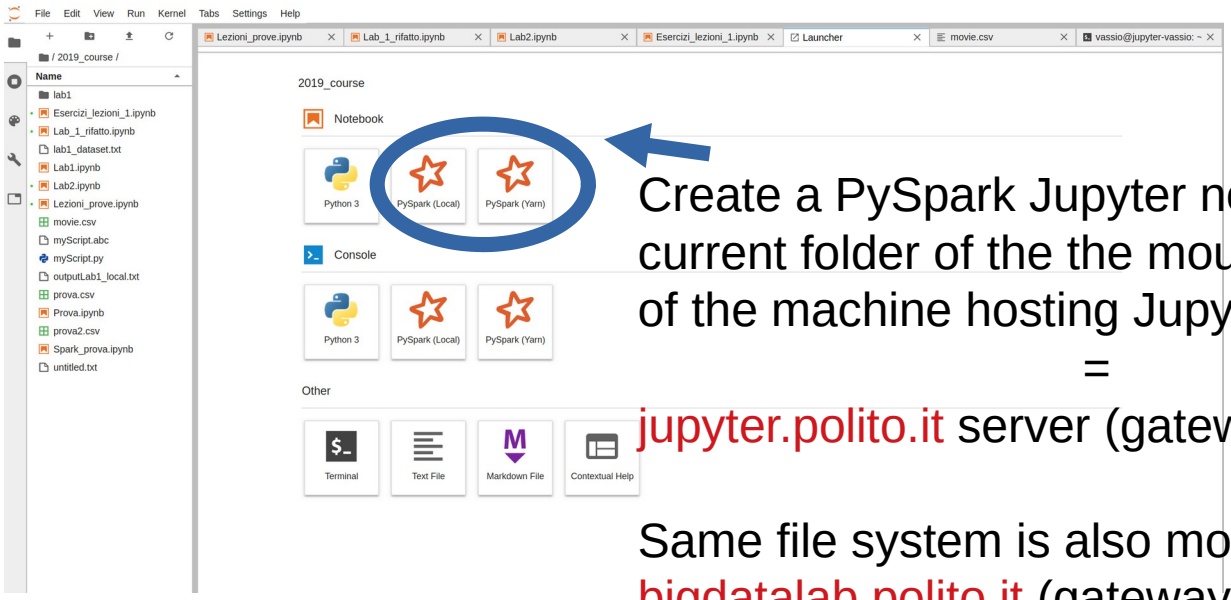
https://jupyter.polito.it/



Personal home of the local file system of the machine hosting Jupyter server = jupyter.polito.it server (gateway)

Jupyter web interface for our BigData cluster

https://jupyter.polito.it/



Create a PySpark Jupyter notebook in the current folder of the the mounted file system of the machine hosting Jupyter server

=

jupyter.polito.it server (gateway)

Same file system is also mounted on server bigdatalab.polito.it (gateway)

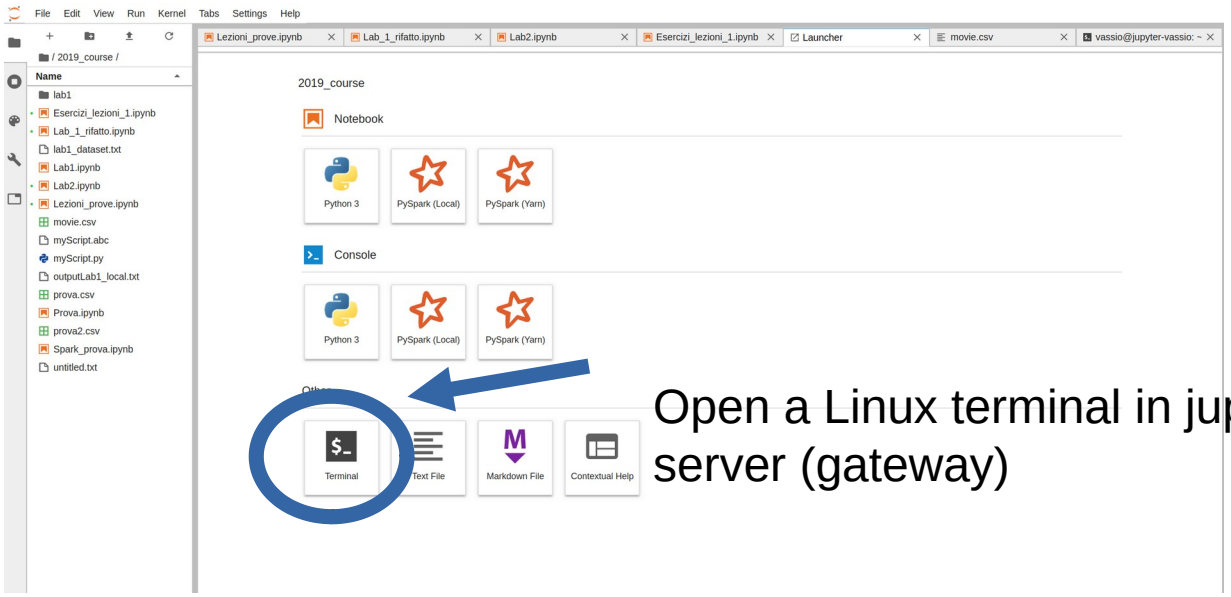
Execute an application by using a PySpark on a Jupyter notebook

- Copy the input data of your application from the local drive of your personal workstation or from the gateway on the HDFS file system of the cluster
  - Use hdfs from command line
  - Or use the HUE web interface

- 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

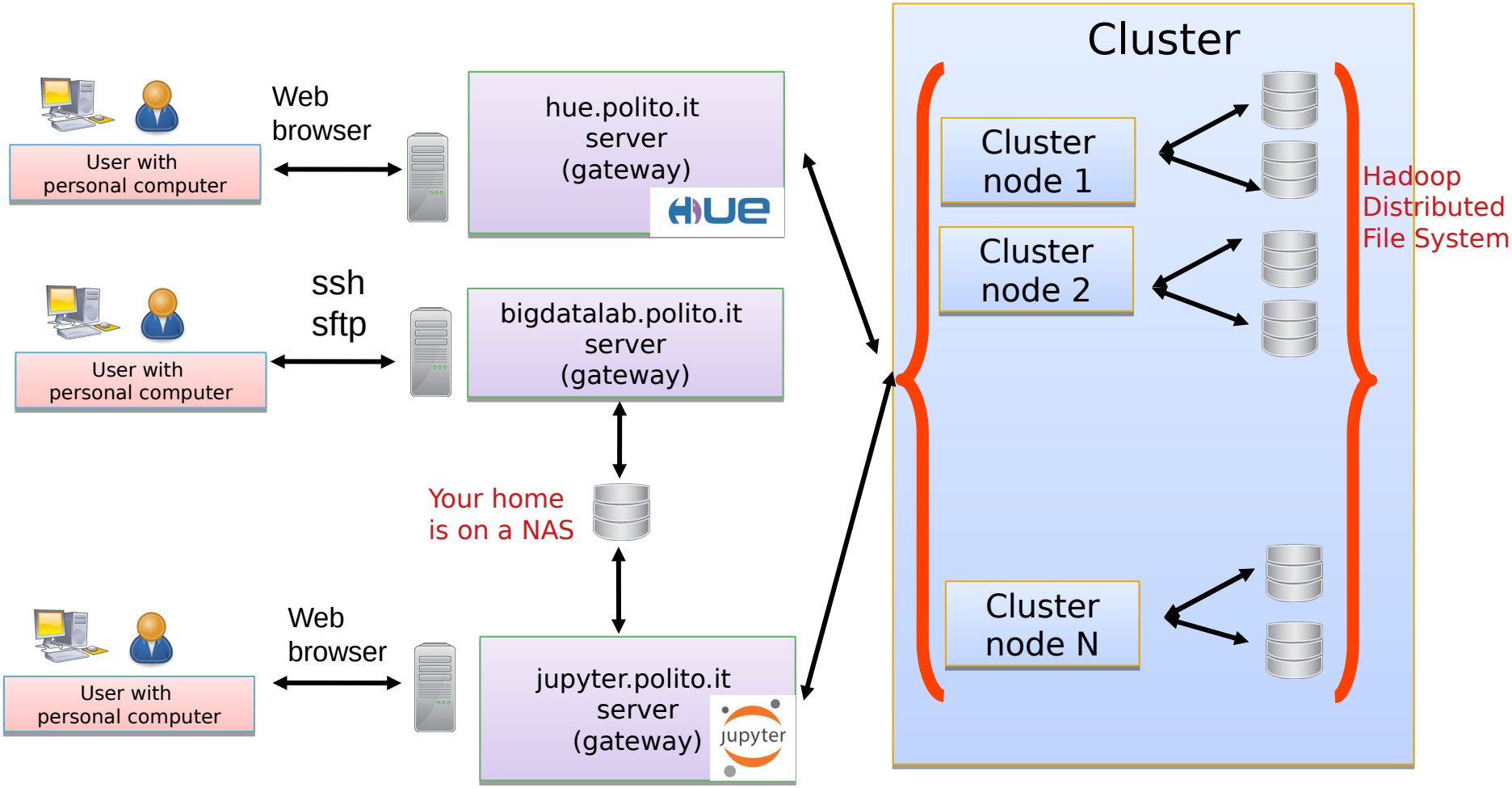
Linux shell in Jupyter web

https://jupyter.polito.it/



Open a Linux terminal in jupyter.polito.it server (gateway)

The BigData@Polito environment



Execute/develop Spark applications on BigData@Polito

- Several options
  - For developing and debugging your applications
    - Open an interactive PySpark shell
    - Open an interactive PySpark Jupyter notebook
  - For developing and executing your applications
    - Write a standalone python application by using an editor and submit the application on the cluster by using the spark-submit command from a linux shell