

Big data for internet applications

Big Data cluster environment

Big Data cluster environment



User with
personal computer



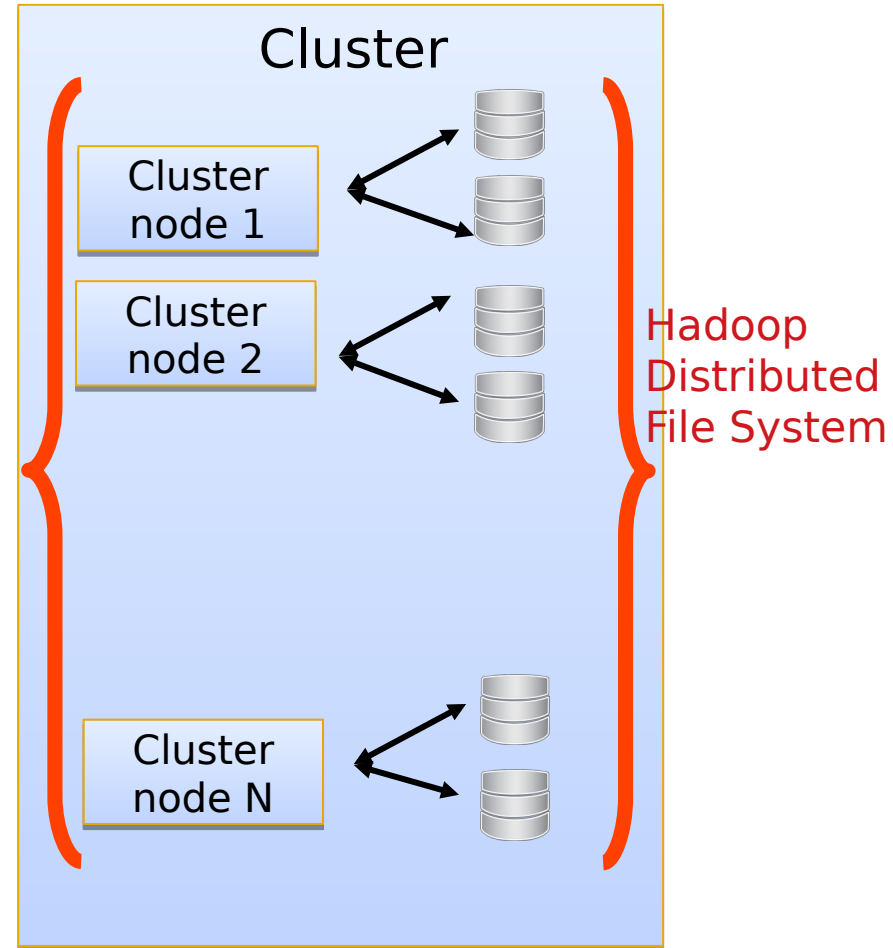
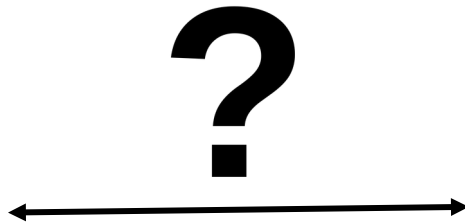
User with
personal computer



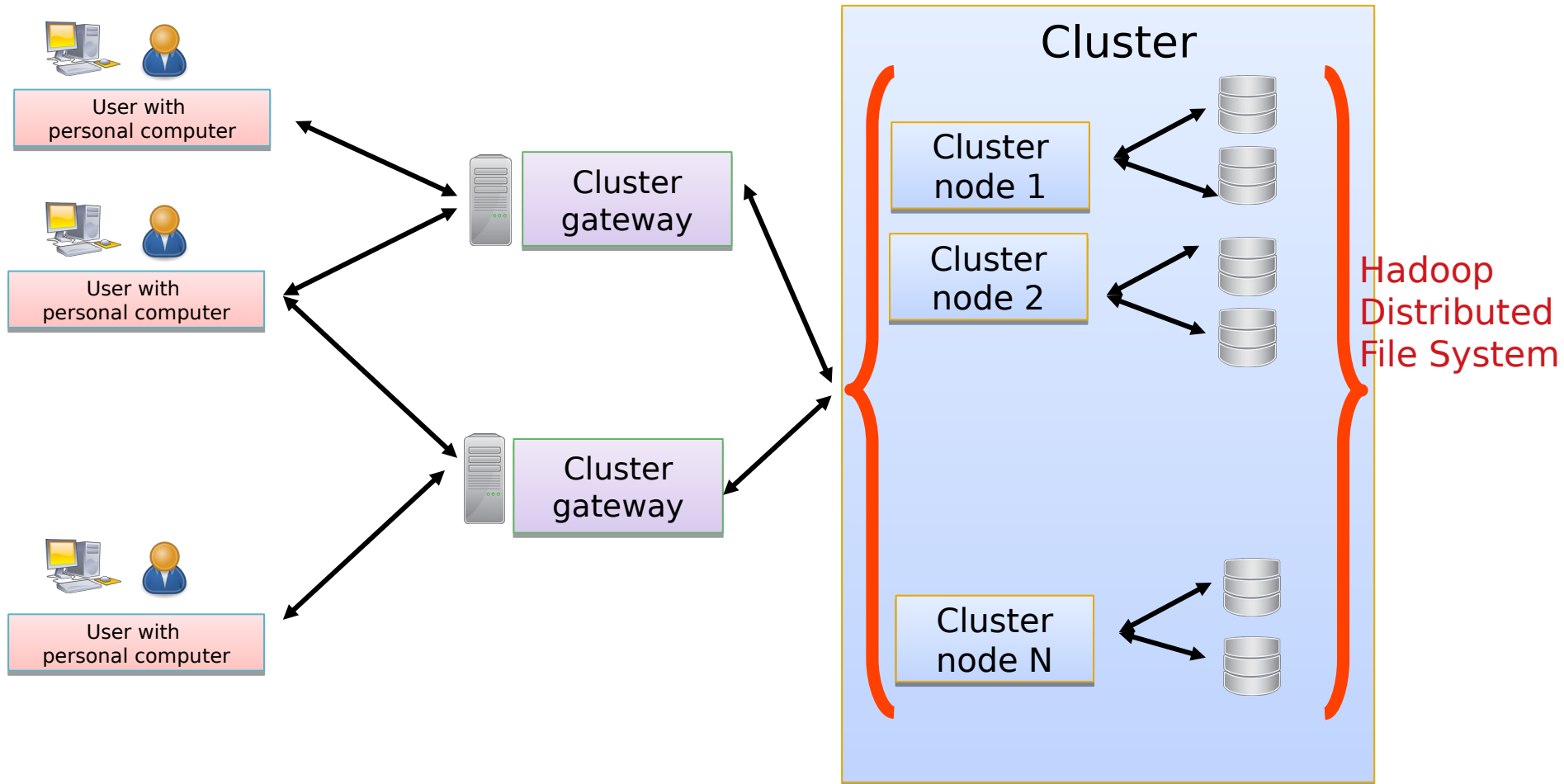
User with
personal computer

Cluster

Big Data cluster environment



Big Data cluster environment

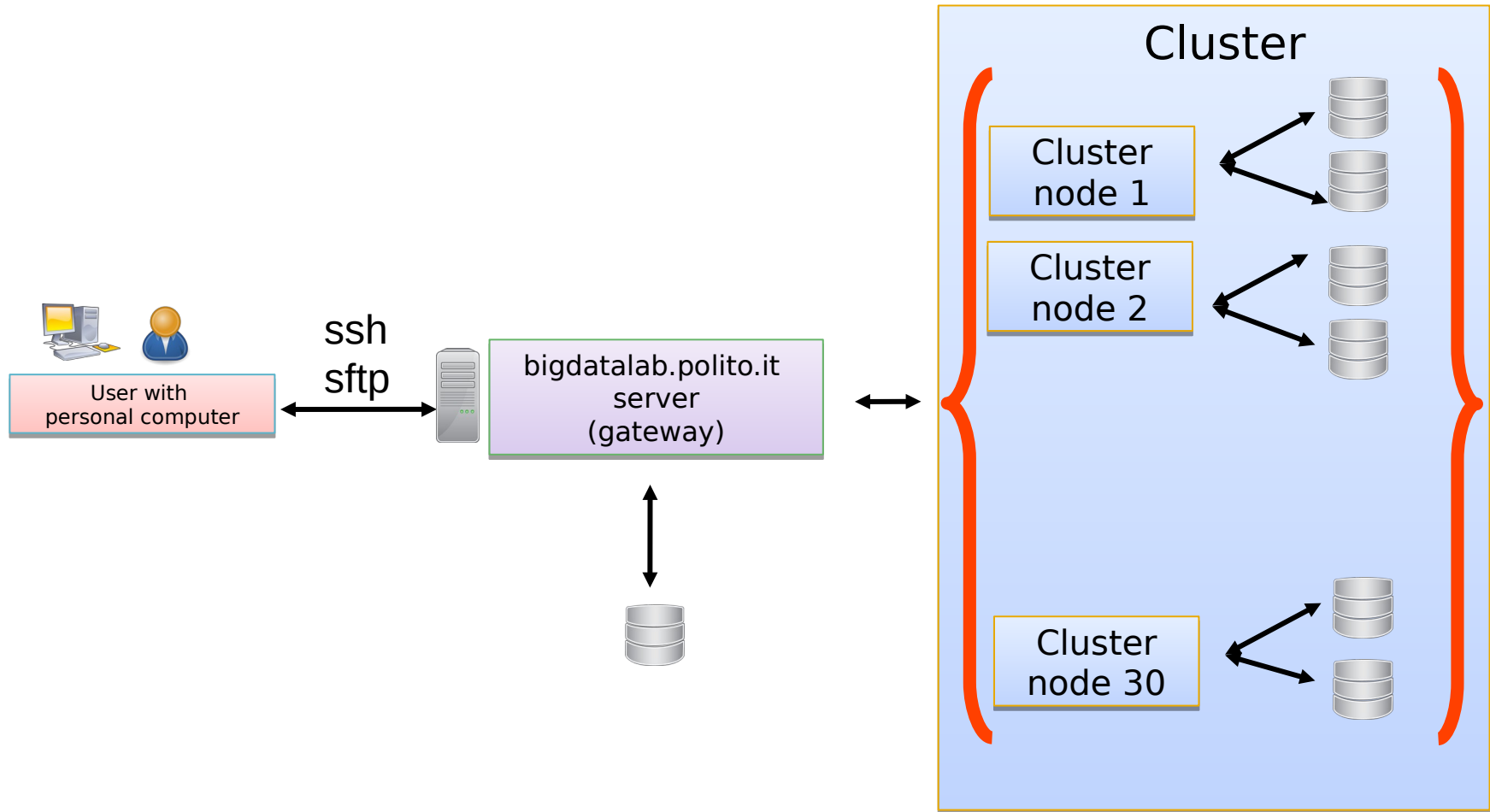


The BigData@Polito environment

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

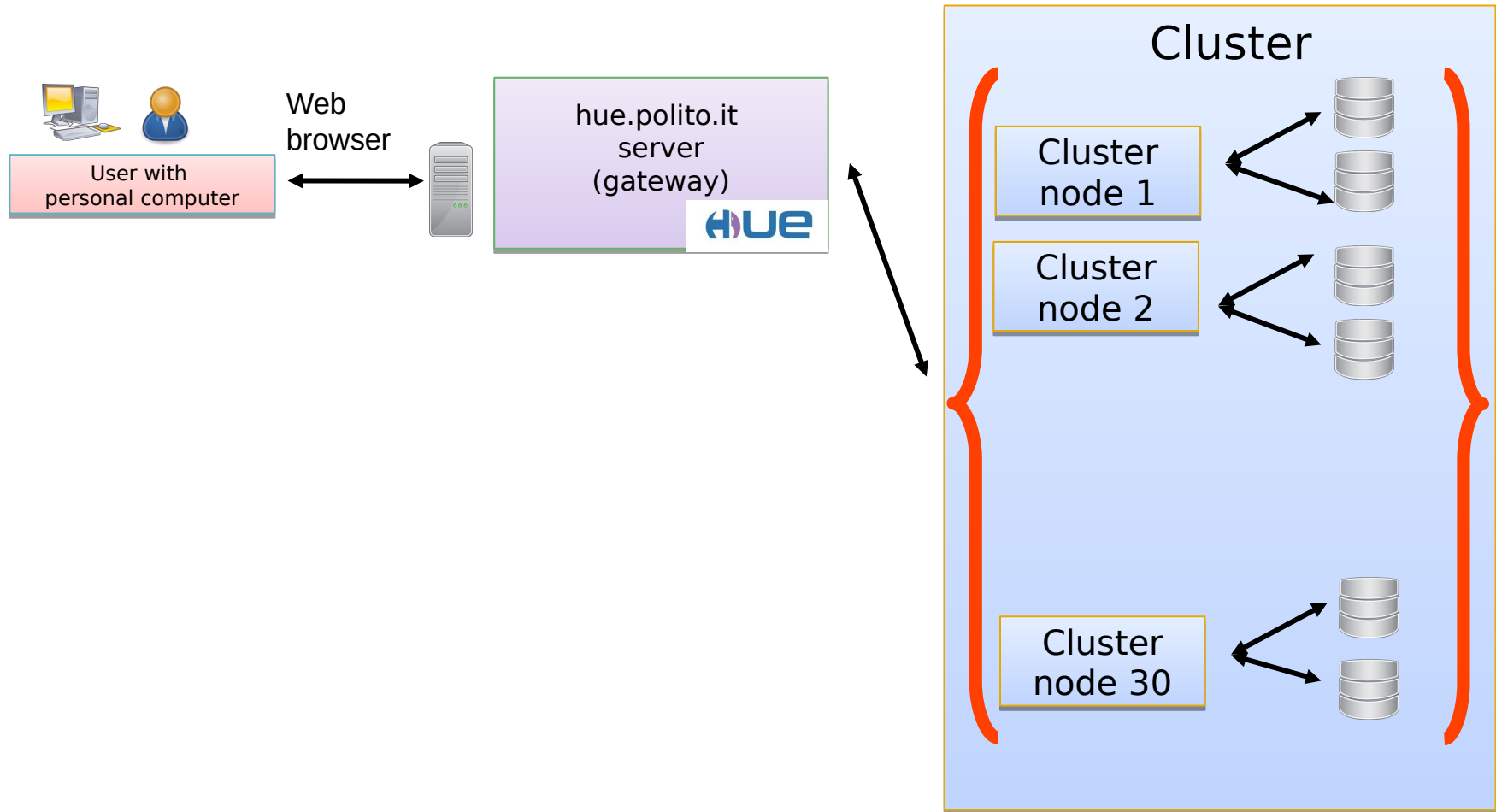
Execute an application by using PySpark through ssh

- 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

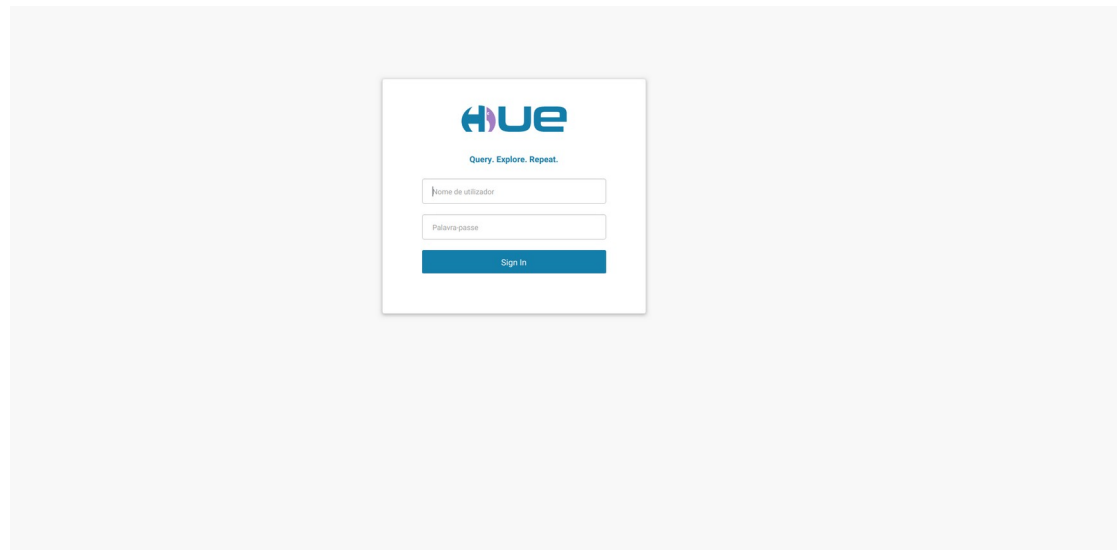
hue.polito.it is an **Access Gateway** server used to interact with the Hadoop cluster. Connecting to it through a web browser you can:

- 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>

Query

Jobs 1 vassio

Browser de ficheiros

Pesquisar por nome de ficheiro Acções Move to trash Carregar Novo

Início / data / students / bigdata_internet Trash

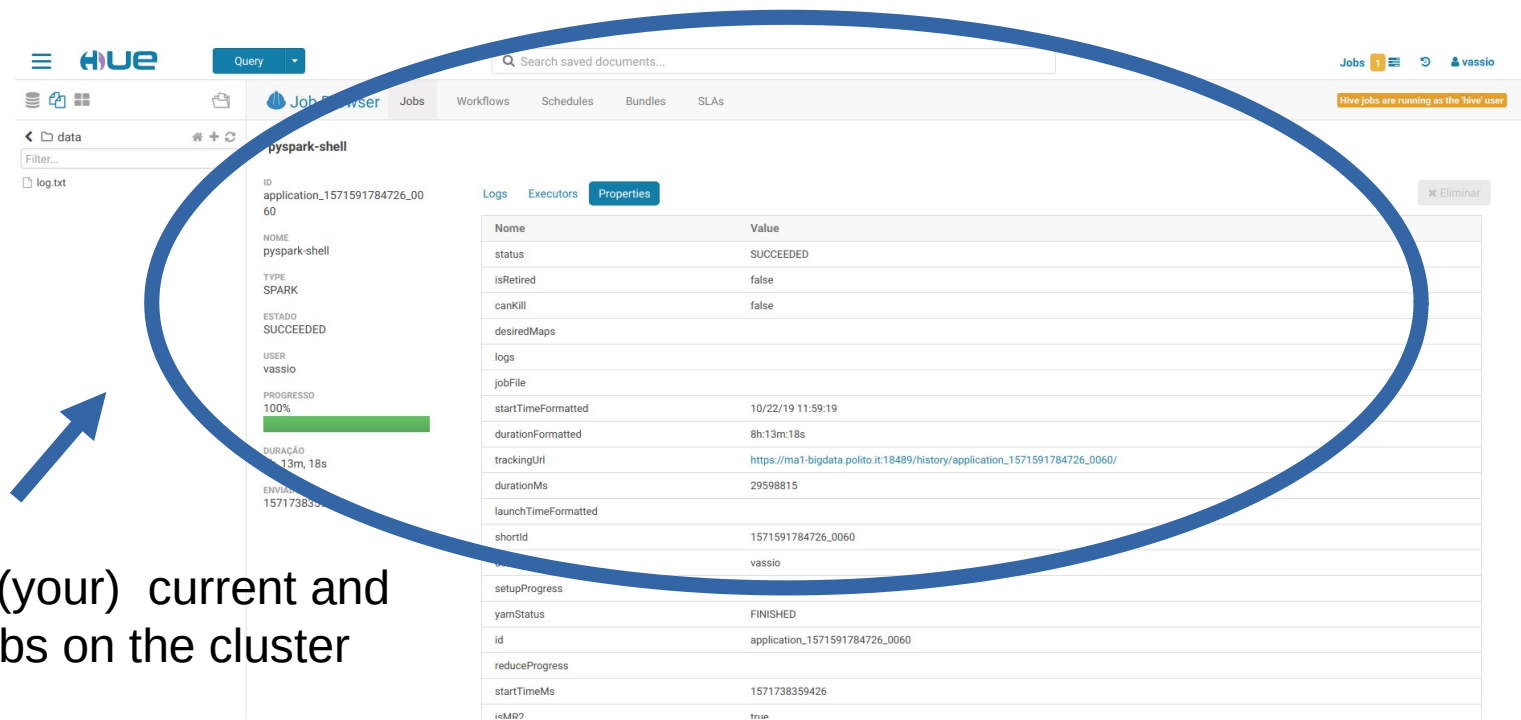
<input type="checkbox"/>	Nome	Size	User	Grupo	Permissions	Date
<input type="checkbox"/>	↑		hdfs	students	drwxr-xr-x	September 06, 2019 12:06 PM
<input type="checkbox"/>	.		trevisan	students	drwxr-xr-x	October 14, 2019 04:14 PM
<input type="checkbox"/>	exercises		garza	students	drwxr-xr-x	October 14, 2019 04:16 PM
<input type="checkbox"/>	lab1		trevisan	students	drwxr-xr-x	September 06, 2019 12:15 PM
<input type="checkbox"/>	lab2		trevisan	students	drwxr-xr-x	September 06, 2019 03:37 PM

Mostrar 45 de 3 itens Página 1 de 1

Browse Hadoop distributed
file system of the BigData
cluster

Hue hdfs/jobs web interface for our BigData cluster

<https://hue.polito.it>



The screenshot displays the Hue web interface for managing Hadoop jobs. On the left, a sidebar shows a file tree with 'data' and 'log.txt'. The main area is titled 'pyspark-shell' and shows a job with ID 'application_1571591784726_0060'. The job status is 'SUCCEEDED'. A progress bar indicates 100% completion. The job details table on the right lists various properties and their values.

Nome	Value
status	SUCCEEDED
isRetired	false
canKill	false
desiredMaps	
logs	
jobFile	
startTimeFormatted	10/22/19 11:59:19
durationFormatted	8h:13m:18s
trackingUrl	https://ma1-bigdata.polito.it:18489/history/application_1571591784726_0060/
durationMs	29598815
launchTimeFormatted	
shortid	1571591784726_0060
user	vassio
setupProgress	
yarnStatus	FINISHED
id	application_1571591784726_0060
reduceProgress	
startTimeMs	1571738359426
isMD?	true

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), ...

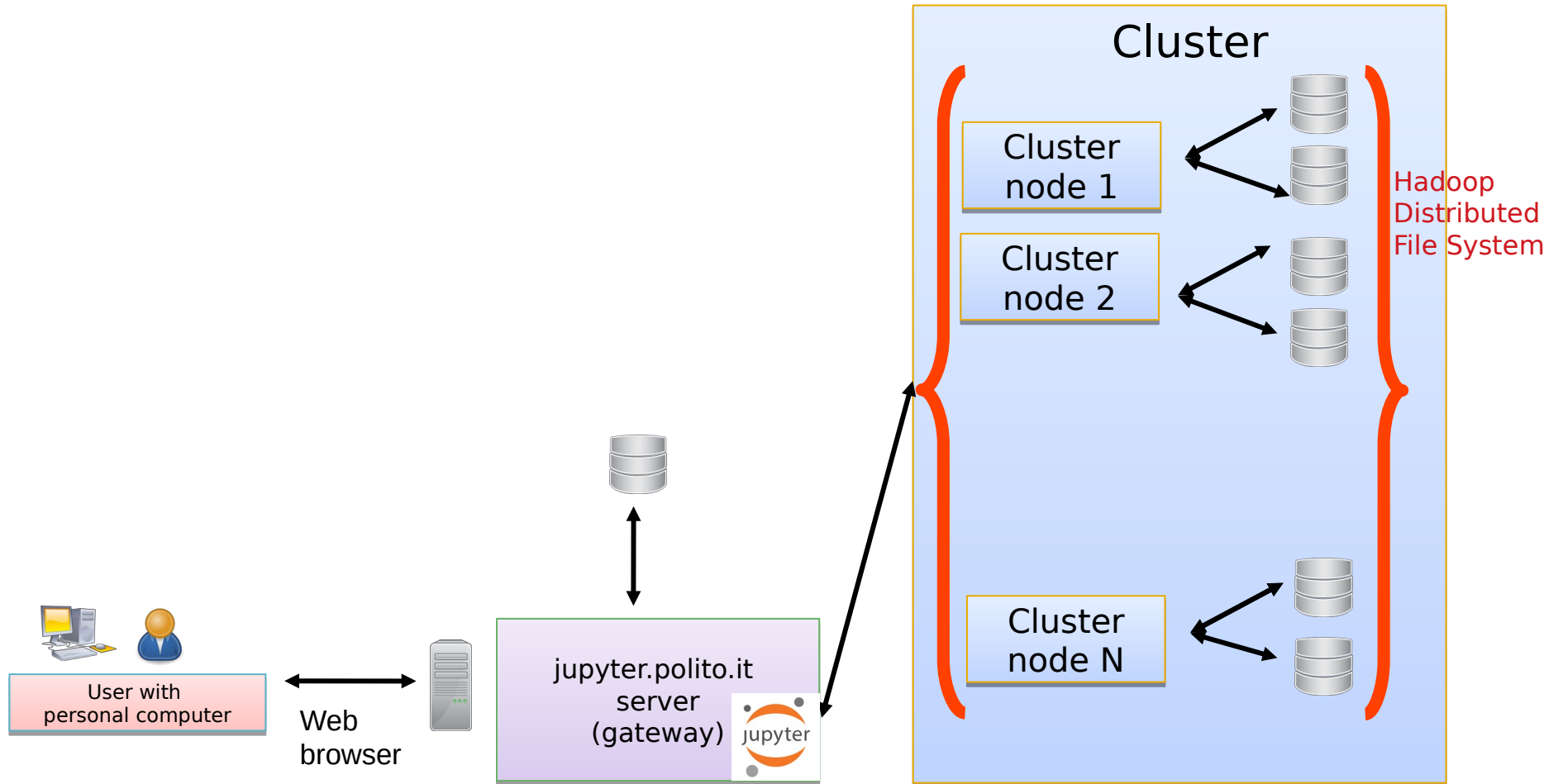


Jupyter notebooks

- 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/>



 jupyter

Sign in

Username:

Password:

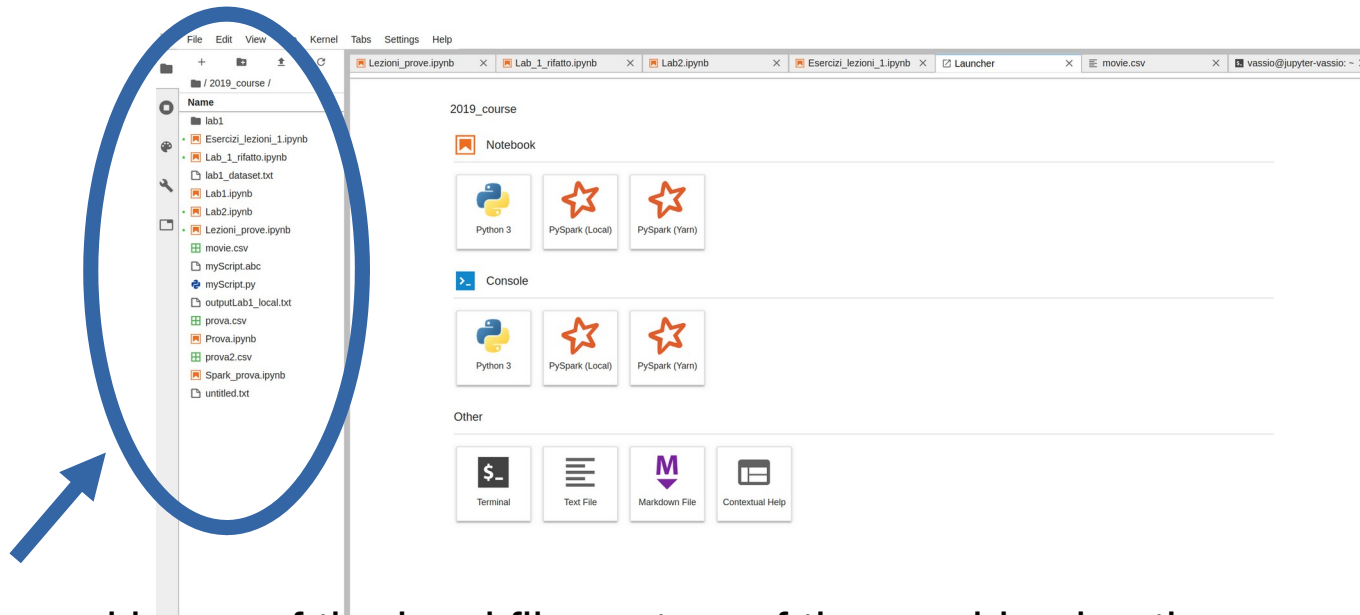
Sign in

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

Jupyter web interface for our BigData cluster



<https://jupyter.polito.it/>

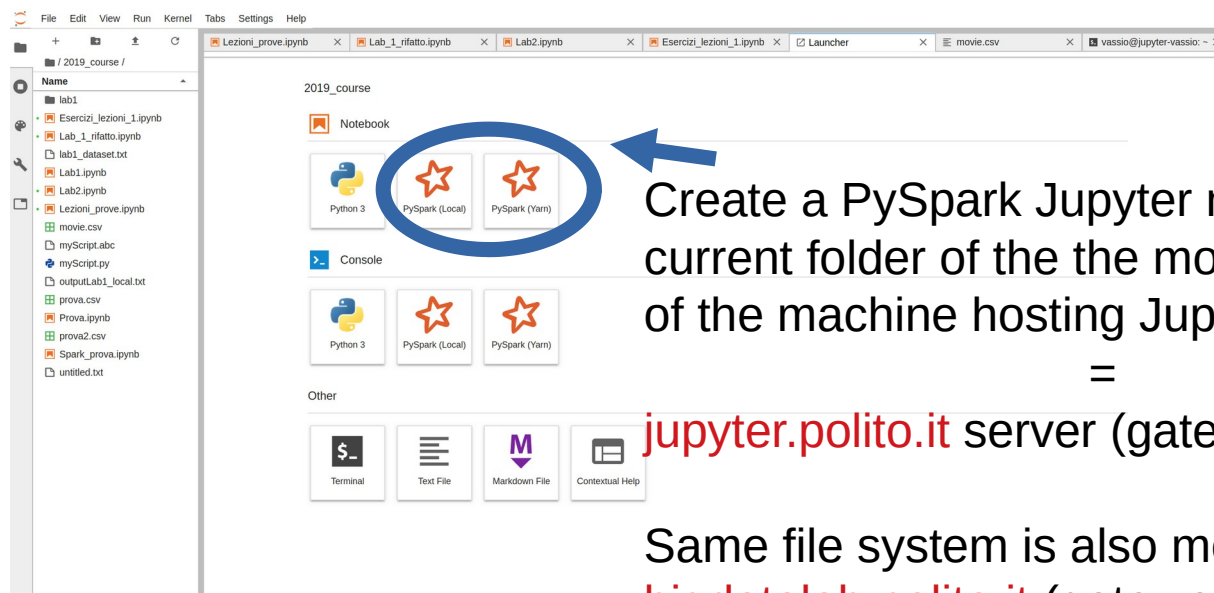


Personal home of the local file system of the machine hosting Jupyter server = jupyter.polito.it server (gateway)

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)

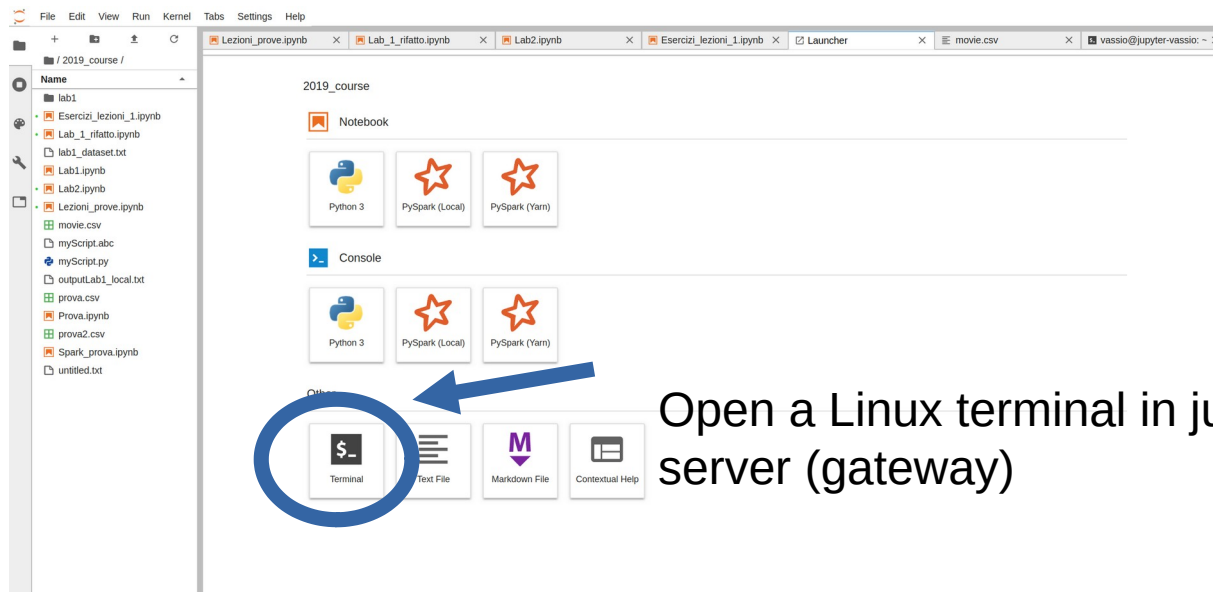
Same file system is also mounted on server bigdatalab.polito.it (gateway)

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

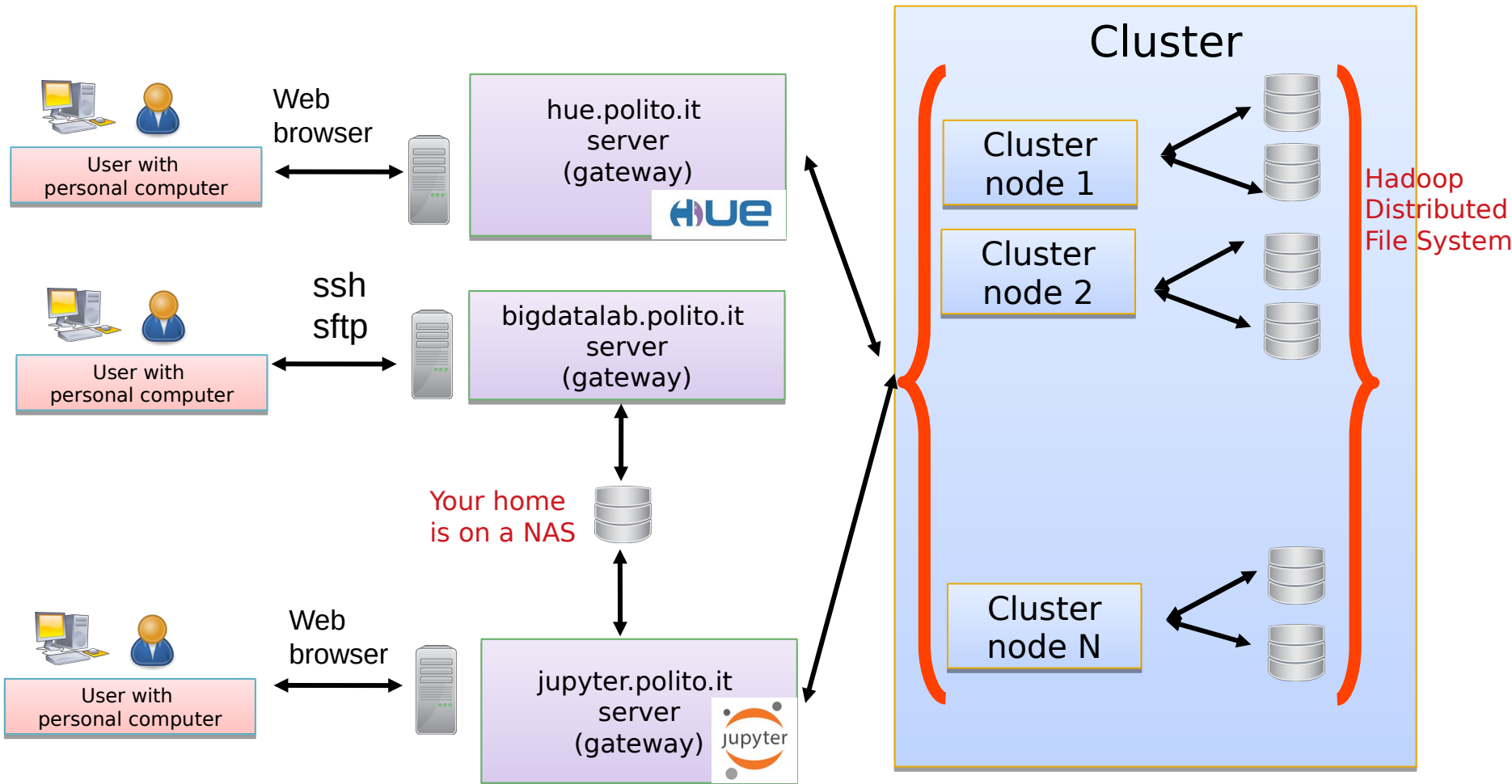
Linux shell in Jupyter web

<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