BIG DATA COMPUTING (Classes A and B, proff. Pietracaprina and Silvestri) 2021-2022

User guide for the cluster on CloudVeneto

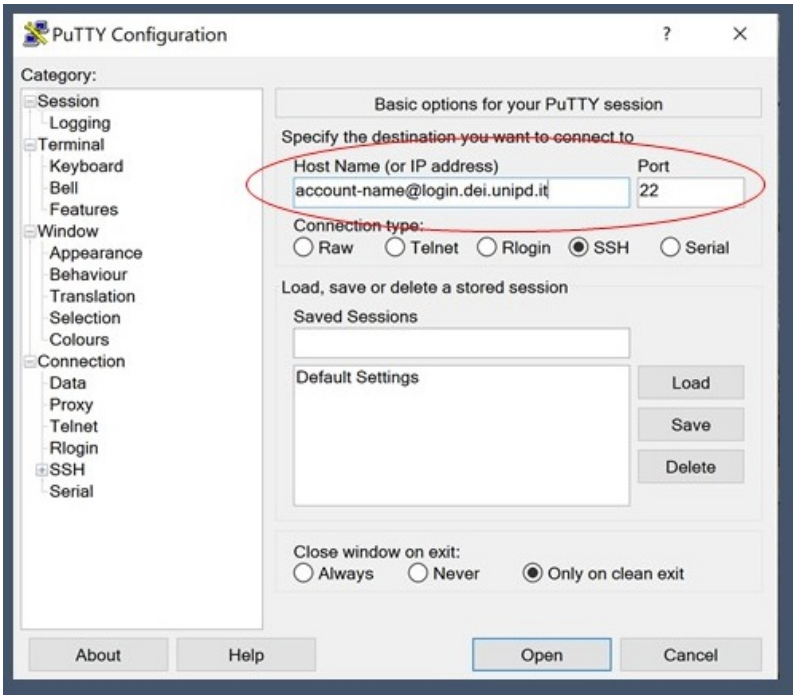
CloudVeneto is a cloud infrastructure at UNIPD. On the cloud you have access to a cluster of 10 machines, each equipped with 8 cores and 16 GB of RAM. One of the 10 machines, called *frontend*, has a coordinating role, while the remaining 9 machines (called *minion-i* with i from 1 to 9) run the Spark program.

You have direct access only to the frontend, from which you will run your jobs. On the frontend, your group has an account (groupXXX where XXX is your 3-digit group number). Initial *default passwords* have been communicated separately and you will be asked to change it immediately upon the first access.

Access to the frontend

The frontend must be accessed through the following 2 steps. Skip Step 1 if you are doing the access from one of the virtual machines that we provided or from a machine connected to the unipd network (*eduroam is considered part of the unipd network*):

- (**STEP 1**) If you are connected to a non-unipd network (e.g., you are working from home) do **remote login to one of the group members' account on a unipd machine**: for example, if you have an account (e.g., *account-name*) at DEI you must remote login to *account-name@login.dei.unipd.it*. If you have an account with other departments, ask your department's system administrators on which machine to connect. Remote login can be done as follows:
    - **Linux and MacOS users** must use the native SSH client. Open a terminal window and type the command **ssh account-name@login.dei.unipd.it** Then, you will be asked the password to enter.
    - **Windows users.** You must install an SSH client: use Putty. To access the frontend, execute Putty, once installed, and the following GUI shows up where you have to fill the boxes as indicated.



- (**STEP 2**) Once you entered a machine on the unipd network, type the following command on the terminal window: **ssh -p 2222 groupXXX@147.162.226.106**, where XX is your group number. You will be asked the password to enter. If this is your first access, you provide the default password and will be asked to change it before proceeding. Note that if your machine is already connected to the unipd network, if you are on Linux/MacOS you can directly type command ssh -p 2222 groupXXX@147.162.226.106, while if you are on Windows you can use Putty indicating groupXXX@147.162.226.106 as Host Name and 2222 as port. Contact us if you experience any problem accessing the cluster.

Datasets

On the CloudVeneto cluster you have read-only access to a number of datasets with varying sizes which have been already uploaded and are hosted in the HDFS (Hadoop Distributed File System). In fact, on the cluster there are two co-existing file hierarchies: the Operating System one, which is used during normal operation, and the HDFS, which stores data to be used as input for Spark jobs.

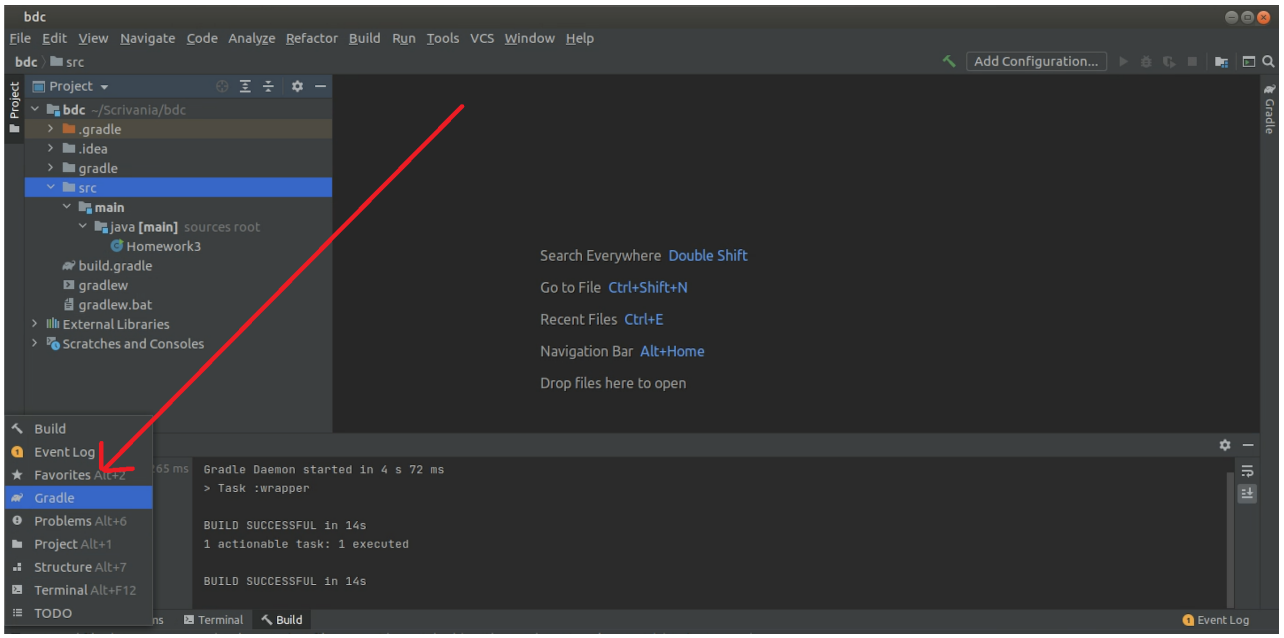
The datasets are stored in the read-only directory **/data/BDC2122**. To interact with HDFS there is a dedicated command, unsurprisingly called hdfs. In particular, to list the contents of directory /data/BDC2122 use the following command: **hdfs dfs -ls /data/BDC2122**

Uploading and running jobs

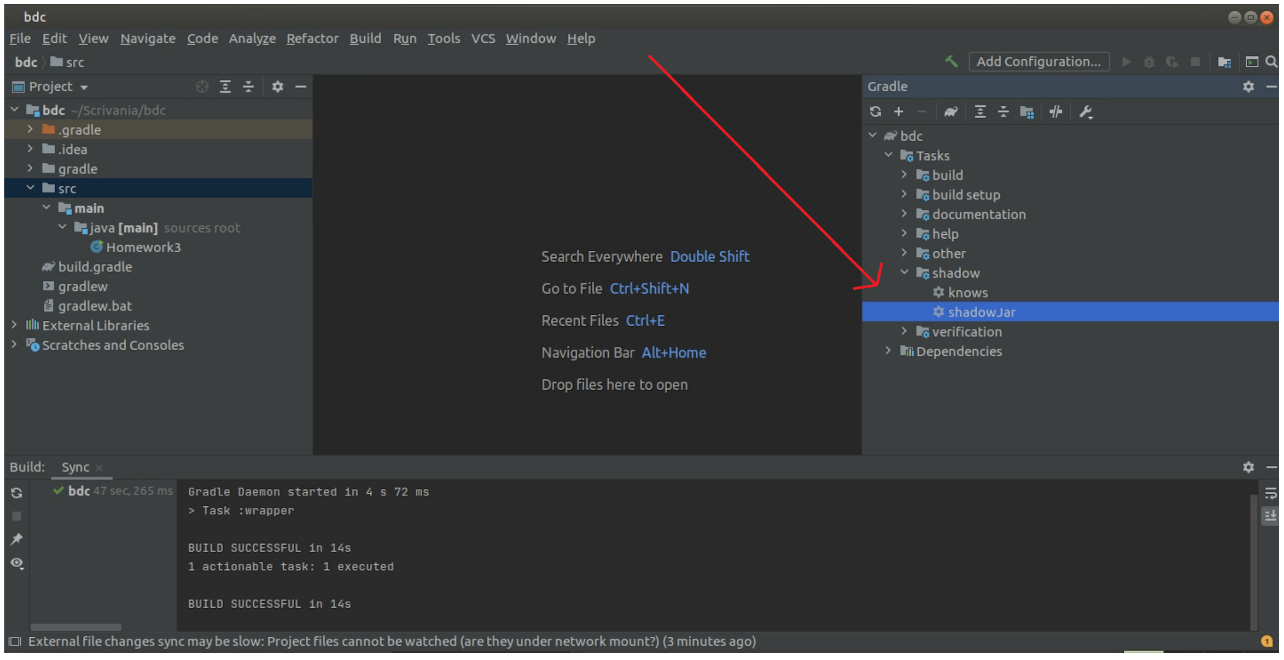
Update for Java users working on their PCs with gradle 6: if you installed IntelliJ IDEA (Community Edition) 2020, you may have to **update the build.gradle** as follows. Check what version of gradle you are using. The version is the name of a directory found inside BDC/.gradle, where BDC is the directory that you created on your computer for Homework 1. If it is version 6, execute the following steps.

- In the BDC directory that you created on your computer for Homework 1, **substitute the old *build.gradle* with the one which you must** download here.
- After susbtituting the build.gradle, start IntelliJ IDEA and if it asks you to "**import changes**", do so.

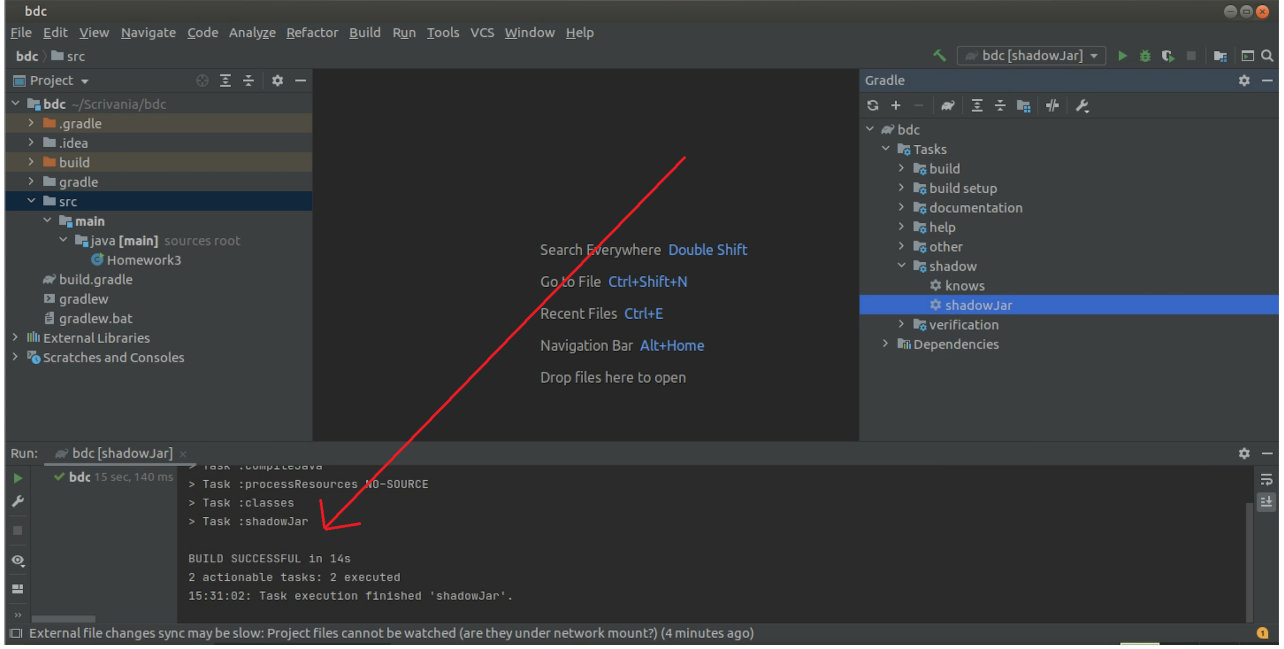
Uploading jobs (all Java users). You must pack your code in a jar file suitable for execution on the cluster. In Intellij IDEA, open the gradle panel from the menu in the bottom-left corner (see image below)



Then, on the menu that will pop-up on the right-hand side do a double click on shadowjar (see image below)



After a few seconds, the jar file will be created (see image below). **Let us know if you have problems generating the jar file.**



The jar file contains your code and all of its dependencies. If the root directory for your homeworks is BDC (as suggested in Homework 1), the name of the jar file will be BDC-all.jar and you will find it in the directory BDC/build/libs. Now you must upload BDC-all.jar to the group's account on the cluster (e.g., groupXY). To do so you must use again one of the group members' account on a unipd machine (e.g, account-name@login.dei.unipd.it) and do the following:

- Transfer BDC-all.jar to account-name@login.dei.unipd.it: you can use scp (on Linux and MacOS) or pscp (on Windows, installed along with Putty). Alternatively, for Windows users, in IntelliJ you can open the embedded terminal (select "Terminal" from the menu in the bottom-left corner) and type the command **scp build\libs\BDC-all.jar account-name@login.dei.unipd.it:.**
- Connect to account-name@login.dei.unipd.it and from there type the command: **scp -P 2222 BDC-all.jar groupXXX@147.162.226.106:.** Please, in this last transfer, *make sure you use the option -P 2222 with capital P.*

Note that, if you are doing the access from one of the virtual machines that we provided or from a machine on the unipd network, you can directly copy the jar file from your machine to 147.162.226.106 via scp (last bullet above).

Uploading jobs (Python users). You must upload your program (e.g., GxxxHW3.py) to the group's account on the cluster (e.g., groupXXX). To do so you must use again one of the group members' account on a unipd machine (e.g, account-name@login.dei.unipd.it) and do the following:

- Transfer GxxxHW3.py to account-name@login.dei.unipd.it: you can use scp (on Linux and MacOS) or pscp (on Windows, installed along with Putty).
- Connect to account-name@login.dei.unipd.it and from there type the command:**scp -P 2222 GxxxHW3.py groupXXX@147.162.226.106:.** Please, in this last transfer *make sure you use the option -P 2222 with capital P*.

Note that, if you are doing the access from one of the virtual machines that we provided or from a machine on the unipd network, you can directly copy the .py file from your machine to 147.162.226.106 via scp.

Running jobs (Java users). Suppose that on the cluster's frontend you uploaded a jar file named BDC-all.jar which includes your Homework 3 program GxxxHW3.java containing class GxxxHW3. In order to run this program, login to the frontend (as explained before) and type the following command

spark-submit --num-executors X --class GxxxHW3 BDC-all.jar argument-list

Running jobs (Python users). Suppose that on the cluster's frontend you uploaded your Homework 3 program GxxxHW3.py. In order to run the program, login to the frontend (as explained before) and type the following command

spark-submit --num-executors X GxxxHW3.py argument-list

Note that by default Spark runs Python 2. If your code requires Python 3, you can invoke spark-submit as follows:

spark-submit --conf spark.pyspark.python=python3 --num-executors X GxxxHW3.py argument-list

Command-line options and parameters (both Java and Python users).

- Option **num-executors** sets the total number of executors used by the application (i.e., workers in MapReduce terminology) to the specified value (X in the example). Each executor will run on a core with 2 GB of RAM. In the homework, we will give you an upper limit to the value X that you can specify, which is lower than the maximum parallelism available in the cluster to ensure a fair sharing of the

resources among the student groups.

- **argument list**: depends on the program that you are running. To pass one of the preloaded files as an argument to the program specify the path /data/BDC2122/filename

More details on the commands hdfs and spark-submit can be found in this guide.

Last modified: Monday, 16 May 2022, 5:10 PM

◄ Assignment of Homework 3: DEADLINE June 13, 23.59pm

Jump to...

Slides on Homework 3 presented in class (update 31/05/22) ►