Questions for the course “Introduction to running R, Python, and Julia in HPC", 17-19 October 2023

THURSDAY, 19 OCTOBER 2023, R

1. Please note the following:
2. Add your questions below, numbering them continuously.
3. Be careful in case someone else is writing at the same time.
4. Please **DO NOT** delete your questions even when they have been answered, as we are planning to use them to improve our material.
5. Do not share any sensitive information as this document is accessible to anyone with the correct link.
6. You can start a **new line within a question by pressing <SHIFT> + <ENTER>**.
7. General information / summary:
8. Hostnames for login nodes (the main course project is at UPPMAX):
    a. UPPMAX – rackham
        i. **SSH**: `rackham.uppmax.uu.se`
        ii. **ThinLinc**: `rackham-gui.uppmax.uu.se`
        iii. **From webbrowser**: https://rackham-gui.uppmax.uu.se/
    b. HPC2N - Kebnekaise
        i. **SSH**: `kebnekaise.hpc2n.umu.se`
        ii. **ThinLinc**: `kebnekaise-tl.hpc2n.umu.se`
        iii. **From webbrowser:** `https://kebnekaise-tl.hpc2n.umu.se:300/`
9. Course project at UPPMAX: NAISS2023-22-914
10. Storage area for the course project at UPPMAX: `/proj/naiss2023-22-914`
11. `Course project at HPC2N: hpc2n2023-110`
12. `Storage area for the course project at HPC2N: /proj/nobackup/hpc2n2023-110`
13. Zoom ONLY for the parallel sessions: https://umu.zoom.us/j/64351885207?pwd=RVpwYnFRdlNLQkM5WWR0VWsrSzUwQT09
14. PLACEHOLDER
15. PLACEHOLDER
16. I runned this: notebook(dir="/proj/naiss2023-22-914/nazib/Julia") but got this Info: running `/sw/comp/python/3.10.8/rackham/bin/jupyter notebook
17. On HPC2N, you can copy the exercises in a tarball from /proj/nobackup/hpc2n2023-110/Exercises.tar.gz
18. On UPPMAX you can copy the exercises in a tarball from /proj/naiss2023-22-914/Exercises.tar.gz
19. To copy the file, go to your personal subdirectory under /proj/naiss2023-22-914/ (UPPMAX) or under /proj/nobackup/hpc2n2023-110 (HPC2N) and do: cp <path listed above> .
20. Example for me (bbrydsoe), on Rackham:

a. `cd /proj/naiss2023-22-914/bbrydsoe`
b. `cp /proj/naiss2023-22-914/Exercises.tar.gz .`
c. Then you need to extract the gzipped tar file: `tar zxvf Exercises.tar.gz` xzvf right?
d. Yes
e. tar xzvf Exercises.tar.gz
f. cd Exercises/
g. cd R
h.

21. I created a .Renviron file in my home directory and update it using this code: echo R_LIBS_USER=\"$HOME/R-packages-%V\" > ~/.Renviron I checked it in my texteditor and it looks fine. However, when I try to install a new package in my R session it cannot find this path still it looks at the default library path which is "/sw/apps/R/4.1.1/rackham/lib64/R/library" which I dont have writing access to.
    a. The system may not read your .Renviron unless you start a new shell or source .Renviron
    b. What is the output of .libPaths() in your R terminal?
22. Now it is these two:

[1] "/domus/h1/sarho/R-packages-4.0.4"
[2] "/sw/apps/R/x86_64/4.0.4/rackham/lib64/R/library"

It was only the second one yesterday. So does it mean that I had to log in again and it got fixed? Because I basically didnt do anything since yesterday.

    a. **A:** Yes, logout and login is the easiest way to get the system to detect the new settings.
23. There is no /proj directory in my home folder
    a. What system are you on? Rackham or Kebnekaise: rackham. Solved. I understand I need to move to /proj/naiss first.
24. How long should it take for a package to get installed from an R session? I started to install stringr and it has been several minutes now.
    a. Some take longer! But do you see progress? Cluster?
25. Could you please share the code for login to directory in reckham
    a. Login is one thing and go to a directory another, that you can do when you are logged in
    b. Log in from terminal: ssh -Y <username>@rackham.uppmax.uu.se
        i. Password
    c. Orient in the file trees on Rackham
        i. cd `/proj/naiss2023-22-914` (note the starting "/")
        ii. `ls`
            1. Did you create a user directory already, otherwise:
        iii. `mkdir <yourname>`
        iv. `cd <yourname>`
        v. `pwd`
            1. If the path looks correct you are now in your own folder/directory

d. Then follow step 20 above to get the exercises
e.

26. Q: On rackham, my current library path is not writable. R 4.0.4
    a. Do you get some more info that it will try to install at another place?
        i. Yes, “would you like to install in a personal library instead”
        ii. Good! That is the way to do it.
27. **Info**: devtools is installed on UPPMAX
28. When you use renv::init(), do you need to be in a specific R version? I mean does it affect the package versions that are later installed?
    a. The virtual environment is connected to one specific R version! Same for python
29. How can you confirm you have created .Renviron?
    I do the ‘touch... ‘ but when check by ‘ls’ there is no .Renviron - maybe you shouldn't see this there anyway?
    a. `ls -a` will list hidden (<.something>) files
    b. It is in you home folder (/home/<username>)
    c. Check with `ls -a $HOME`
30. When you load a specific R version (e.g., 4.0.4), and then you download packages from CRAN. Will the packages will be the latest package version on CRAN or the one that it is compatible with the R version you loaded?
    a. If the version of an R package available at CRAN and installable via install.packages(), which is always the latest version, is compatible with the version of R that you are running, then you will be able to install that version. If it is not compatible with the latest version of R, then it will not be available at CRAN and installable via install.packages().
    b. Note that CRAN is also not sophisticated in another way. If you try to install a CRAN package that is no longer available for the version of R you are running, you will get the message "this package is not available for your version of R" or something like that. If you try to install a BioConductor package with install.packages(), instead of BiocManager::install(), you'll get the same message. If you misspell the package name, or give it a nonsensical name, you'll get the same message.
31. What would happen to my virtual environments when my one-year UPPMAX account expires?
    a. You should move your data and files to a local backup! And possible move them to a new cluster
    b. Thanks for your response! Focusing my question a bit, i wonder whether I can, as part of my back up, also save the R packages versions I installed or were ready available when I created my virtual environment? Let's assume that my virtual environment is compatible with R 4.0.4

c.

32. During installation the selection for crane mirrors will be SWEDEN UMEA?
    a. I think that is default, but you may get a pop up sreen in ThinLinc or if you use x-forwarding in terminal
33. Why do we have some of function with – and other with – below- Example time and n

```
!/bin/bash
#SBATCH -A naiss2023-22-914 # Course project id. Change to your own
      project ID after the course
#SBATCH --time=00:10:00 # Asking for 10 minutes
#SBATCH -n 1 # Asking for 1 core
```

    a. Not sure I understand your question. This script is needed to reach the compute nodes

        why sometimes '- -' as before time and why sometimes '-' as before n i think person means

    b. For most commands, '-' (hyphen) is used for single letters options/flags, whereas two '-' is used for whole words
        i. Example:
            1. `ls -lrt` is shorthand for `ls -l -r -t`
            2. `What about the nodes you allocated in the batch script header then? Is it overwritten by the makeCluster() function of the doParallel package?`
            3. `ls --help` gives help output for the ls command
34. Rackham, sh script job section, is there difference between R < code.R and Rscript code.R?
    a. Rscript passes some arguments into R
35. I'm wondering about the makeCluster and stopCluster functions. Can you elaborate on that? Don't we already define the number of allocated nodes in the batch script header?
    a. Allocated cores are an upper limit for what you can define from R
    b. In the batch script/allocation commands you ask for some resource, a number of cores etc. Then when you run R or whatever, you say how many of the cores etc. you will use. It could vary what you need in the same session.
36. What about the nodes you allocated in the batch script header then? Is it overwritten by the makeCluster() function of the doParallel package?

A: No, you have the number of nodes and cores allocated that you asked for in the batch script, but maybe your script is not asking to use all of them. It can use fewer, just not more.

    a. They are still active and the CPU-hrs used will be based on that.
    b. You may book more cores to get more memory even though you are not running parallel work, typically 6-8 GB per core
    c. The number of cores in makeCluster should not exceed the number in your batch script (-n flag). Otherwise, the script will overload the cores if computations are expensive
    d. something that could be useful is to allocate some number of cores (-n) and then use less in makeCluster to get more RAM memory per core. Even you can allocate the entire node and use only few cores in makeCluster if your application is memory consuming
    e.

37. If you run the command squeue –u <username> and get this: JOBID PARTITION NAME USER ST TIME NODES NODELIST(REASON). It means you have no job running?
    a. Yes and no in the queue either
    b. Did you already start a job with sbatch? Perhaps it is already finished?
        i. Do `ls -lrt` and you can see the latest file and it might be slurm-XXX
    c. Yes i did. Thanks. I can see it with –out extension.
38. If you want to load an R- package for the hello.R command. Do you do it inside the R command orr in the slurm?
    a. The library lines has to be in the code
    b. Depending on cluster you may have to in the **batch script** load R_packages on UPPMAX or other modules at Kebnekaise, like `R-bundle-Bioconductor`
39. In the Rscript (`Example taken from https://github.com/lgreski/datasciencectacontent/blob/master/markdown/pml-randomForestPerformance.md library(mlbench) data(Sonar)). The question is, if you have your own data. You have to set path to the location of the data in your directory inside UPPMAX?`
    a. `Yes you have to set full paths to data. If you always run from the same place (present working directory) it can be sufficient with relative paths. If you run it from somewhere else the full path is a must`
40. Do you choose to either run rscript with 'R --no-save --quiet < hello.R' or send script to batch? Cause if copy and use whole example serial code it runs rscript – so then why send the script to batch?
    a. Sbatch lets you use the compute nodes. Otherwise it will run from the login node (if you are there). That is the difference. Did I understand your question?
    b. So one should use either or, but sbatch command if want to run on batch node and not in login node? I'm not sure if you understand, or if it's rather I who are not understanding. If look at slides, the 'send script to batch' command is stated/explained after the example serial code – I assumed this step was to come after running Rscript, but after running it its done right. So that is why I asked if you choose to do either or.
41. When trying the example serial code, but with add2.R - do I just add e.g. 2 4 after 'R --no-save --quiet < add2.R 2 4' like so, or where add arguments?
    a. `R CMD BATCH --no-save --no-restore '--args a=1 b=c(2,5,6)' test.R test.out &`
    b. From https://stackoverflow.com/questions/14167178/passing-command-line-arguments-to-r-cmd-batch
    c. `Thanks! So the code in the "UPPMAX solution" under the exercise is not correct? Because this is what I did, and did not work. By doing so, the arguments are ignored and answer is NA.`
        i. `Are you in the folder where add2 is? Yes, used 'ls' and add2.R is there.`

d. A can reproduce this error
e. Try: `Rscript add2.R 2 3`
f. Does still not work for me. When using Rscript instead, says 'No such file or directory' - but still in directory where add2.R is so it should find it.
g. Cou can "cd" to the right directory in the batch script, like after the "module loads". Like so:

```
# Load any modules you need, here R/4.0.4
module load R/4.0.4
cd /proj/naiss2023-22-914/bjornc/Exercises/
# Run your R script (here 'add2.R')
Rscript add2.R 2 3
```

h. Nope, still error:
[<user>@rackham4 R]$ Rscript --no-save --quiet < add2.R 2 3
Fatal error: cannot open file '2': No such file or directory
i. You are not in the batch script but on command line now
    i. And remove `–no-save --quiet`
j. So if I put in these commands, see below, I am still not in batch script?

```
#!/bin/bash
#SBATCH -A naiss2023-22-914 # Change to your own after the course
#SBATCH --time=00:10:00 # Asking for 10 minutes
#SBATCH -n 1 # Asking for 1 core


# Load any modules you need, here for R/4.0.4
module load R/4.0.4
```

42. This is really basic, but I mainly use Bianca and RStudio. Should I be saving my code and then running it in through the terminal with a SLURM request rather than using an interactive session with 2 cores or so and running the code as I go along? I guess it also depends on the size of the project probably.
    a. We'll talk about rstudio after break
    b. Was this answered?
43. HPC2N parallel session: https://umu.zoom.us/j/64351885207?pwd=RVpwYnFRdlNLQkM5WWR0VWsrSzUwQT09
44. How do you change the number of core if you are already in the interactive mode?
    a. You cannot change the max number of cores, but you could use something like "srun –n 2 <your commands>" to only run on 2 of them, for instance
45. (solved) Rackham; dependency between batch jobs, what's the best practice? Dependency between shells (each contain one R script), dependency between R scripts and all scripts in one shell, or Rscript <script1.R >script2.R ? E.g., a large simulation with 10^4 iterations, each require lots of memory. I divide the job to say 1000 (batch) * 10 iters (per batch) to get the total 10^4 iters. The second job requires all iters from the first job.
    a. For dependency on the job level: https://www.uppmax.uu.se/digitalAssets/560/c_560271-l_1-k_uppmax-slurm-2022.pdf
    b. For R: please help!

    c. (solved) Also, how to submit the batch, non-interactive job to the test node (may called otherwise)?
        i. Same as for interactive: `-p devcore` is the same as `p -core` whereas -p devel is the same as –p node
        ii. In the batch script: `#SBATCH -p devcore` for instance

46. How did you copy into thinlinc terminal again? shift + insert doesn't seem to work for me
    a. within ThinLinc
    b. copy: <ctrl>-<insert>
    c. paste: <shift>-<insert>
    d. On Bianca it is harder to copy between your computer and thinklinc
    e. but there is a "menu" to the left of the ThinLinc screen where there is a clipboard
47.
48. Suppose I have a long R script and I want to select some lines of the code.How can I do that?
    a. I am not sure what is meant. Could you clarify?
49. PLACEHOLDER
50. How to even enter batch mode? Feel like a missing part in instruction. Vim? Nano? Why not in instrucrtions?
    a. Do you mean how to write a batch script? Use any editor you like; vim, nano, emacs...
       - Okay thank you for clarification. This is not stated anywhere in slides (or I might be 'blind' for this instruction) so did not understand at all. Thanks!
    b. Then you submit the batch script with sbatch
    c. Yes, the batch script is an instructon that you send to the slurm scheduler. It contains both the commands that should be run (the job itself) and what allocations on the compute part of the cluster are needed
51. PLACEHOLDER
52. PLACEHOLDER
53. PLACEHOLDER
54. PLACEHOLDER
55. PLACEHOLDER
56. PLACEHOLDER
57. All recordings for the three days of the course will be uploaded to HPC2N's YouTube channel, to a playlist for the course: https://www.youtube.com/watch?v=aCSMEGglTG8&list=PL6jMHLEmPVLxdCllWrzseYkGoCCXjrukM
58. Evaluation survey for today: https://forms.office.com/e/05mVUtbns9
59. Evaluation survey for the whole course: https://forms.office.com/e/dnTnevNTeA