CyVerse_Workflow_Notes

stribling

2023-04-10

General set-up & workflow for shared data, analyses, and coding for the forest fragmentation working group Overall goal: Store all shared data online, develop code in a shared version-control workspace, and run code & analyses on shared data using cloud computing

Data storage: We can store data in a common folder in CyVerse

- Grace's shared folder is our current data home: "/iplant/home/mgracemcleod/Forest_Fragments/Data/"
- Our Team for this forest fragmentation working group is **"Forest Fragments"**

Data notes & questions

- I think (?) spatial shapefiles need to be stored unzipped within cyverse in order to access the files within the Cyverse's RStudio docker ("Rocker")
- It looks like we need **all** the multi-part files in order for R/Rocker to sucessfully read in a .shp file.
  - What needs/issues does this generate for data storage?
    - How do we best manage multi-part shapefiles?
    - Does storing a bunch of files in Grace's folder use up all her storage space? If so, how can we share the storage burden? Sharing with a "Team" sounds like the right solution, but I'm not currently able to share with our Team.

Code development: We can store and edit code on Github, using a shared group repo, "forest_frag".

- We can integrate the github repo with RStudio on personal computers or within CyVerse's RStudio ("Rocker")**.
  - This allows us to pull the most up-to-date code, edit it, and push code back to main repository.
  - Version control within github allows us to "back up" and find old versions of code if we need to.

Coding notes & questions

- Jenny Bryan's Happy Git and GitHub for the useR is a great resource for getting your RStudio fully integrated with GitHub.
- GitHub also allows for data storage, but I've read recommendations against using that feature. When is this feature useful?
- See additional issues raised below under "Analysis notes & questions"

Analysis Workflows: We can run analyses within the "VICE" environment of CyVerse.

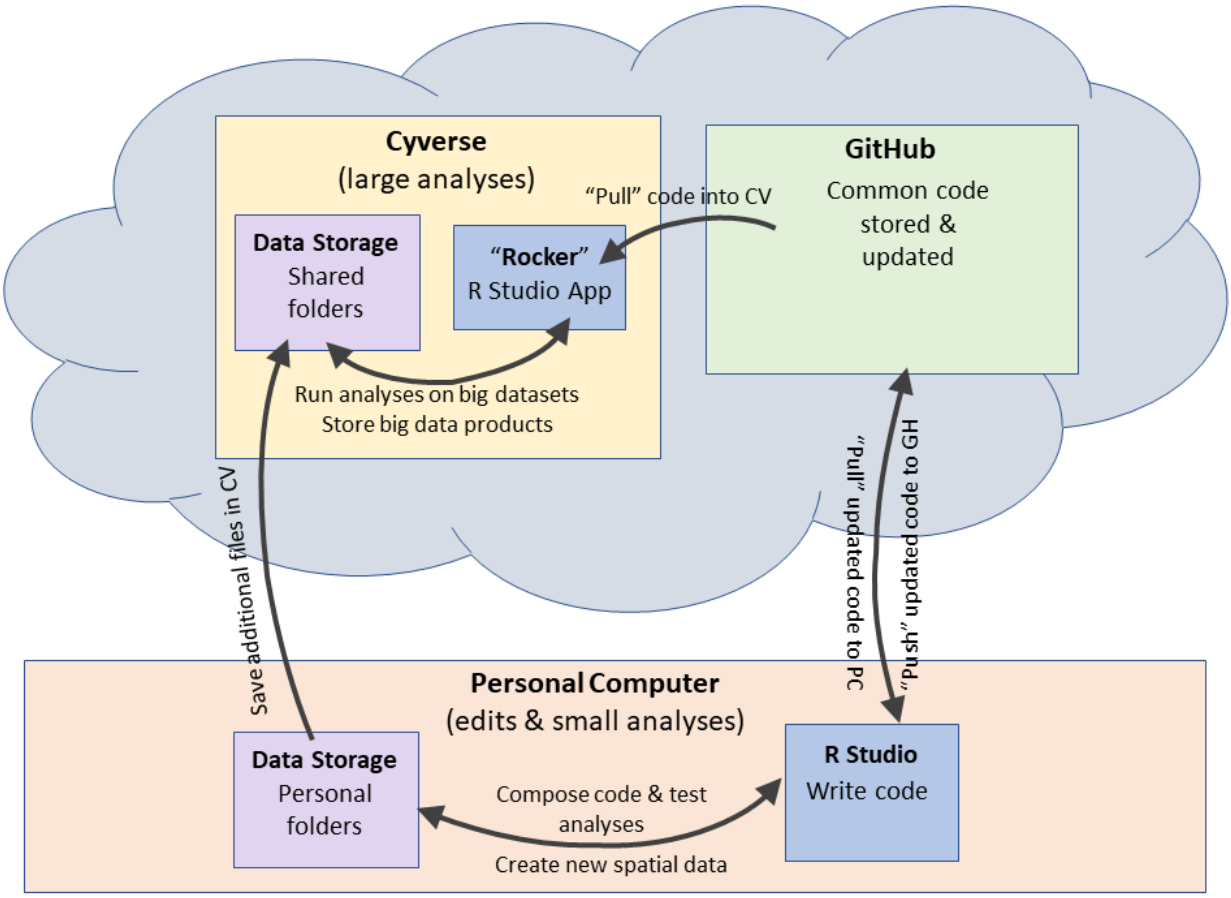
1. A Rocker R Studio Geospatial session** can be linked to the 'forest_frag' repo on GitHub to pull the code into Rocker.
2. Data for analyses is stored within CyVerse in a shared folder.**
   - ** When launching Rocker, during Step 2, when establishing analysis parameters for your input data, select the shared "Forest_Fragments" folder as your Input Folder. (You can browse to it or directly input this path: /iplant/home/mgracemcleod/Forest_Fragments)
3. Code is run through the "Rocker" app within CyVerse; should function just like RStudio.
   - ** DO choose the Geospatial version of Rocker; it seems to come pre-loaded with more packages needed for geospatial analyses. e.g. 'terra' is only available in the Geospatial version.

Analysis notes & questions

- Each Cyverse account permits 200 Core Hours of cloud computing for free.
  - A simple test run of "Rocker" within CyVerse for ~ 25 mins with some super simple code used up 13 core hours. I don't understand why; is there a way to avoid this?
  - What is the best use of this cloud computing resource, given the restrictions?
- Given limited processing hours within CyVerse, we may need to do most coding and work on personal computers, and then open CyVerse & Rocker for the big analyses, once code has been fully written.
  - This necessitates all of us downloading all files to our PCs anyway, so we can work off the cloud, tinker around, and then push it back up to the cloud and run the big analysis.
- Some troubleshooting we've encountered so far:
  - Currently unable to read in .kml files within Rocker, though the same code works on my PC - why?
  - ** Trial run of using github integrated with Rocker on 4/13/23. Successfully established a version-control project connected to our "forest_frag" repo and 'pulled' code into rocker, and ran it. **Could not commit updated code to GitHub from Rocker despite generating a token and setting github credentials with token. Error message below.

```
>>> /usr/bin/git commit -F /tmp/RtmpeCs61U/git-commit-message-17957767511.txt
Author identity unknown

*** Please tell me who you are.

Run

  git config --global user.email "you@example.com"
  git config --global user.name "Your Name"

to set your account's default identity.
Omit --global to set the identity only in this repository.

fatal: unable to auto-detect email address (got 'rstudio@a1106b86f.(none)')
```

Visualization of how all of this comes together?


Cyverse
(large analyses)
Data Storage
Shared
folders
"Rocker"
R Studio App
"Pull" code into CV
GitHub
Common code
stored &
updated
Run analyses on big datasets
Store big data products
Save additional files in CV
"Pull" updated code to PC
"Push" updated code to GH
Personal Computer
(edits & small analyses)
Data Storage
Personal
folders
Compose code & test
analyses
Create new spatial data
R Studio
Write code