JASMIN Workshop: Exercise 6: Choosing the right storage for your workflow

Scenario

You are working on a project that has a multi-step processing workflow. You want your workflow to make efficient use of available compute and storage resources.

Objectives

- Use your knowledge of storage types in use on JASMIN to identify appropriate storage for use within a typical workflow

JASMIN resources

- JASMIN storage types as described in “JASMIN Overview”
- Documentation at https://help.jasmin.ac.uk

Local resources

- Pen & paper
- Colleagues (groups of 3-5?)
- Coffee

Instructions

1. Study the design brief below
2. Consider the following questions in the design of your workflow:
    a. Where do you need to do your processing (using what compute resources?)
    b. What steps are involved?
    c. How much storage space do you need at each stage?
    d. How would you move data if you need to move it between stages?
    e. What happens when you’ve finished your processing?
3. Sketch a diagram describing your intended workflow, labelling important features
4. Be prepared to present it to the group!

Design brief: (DO NOT PROCEED TO NEXT PAGE!)

- Step 1
    - 50TB of observation data need to be processed to intermediate output products. The observation data are available in the CEDA Archive. Each intermediate output product is half the size of the equivalent input file.
    - Your supervisor wants you to use the “AweSUM-1™” processing code to do this: *“it’s efficient because several processes can write to the same output file at once”*. A log of what’s been processed would be useful.
- Step 2
    - The intermediate output products need to be processed with some code you have written yourself, to generate 1TB of final products.
- Constraints
    - Your GWS size is 60TB. Currently it’s empty. It is at path /gws/nopw/j04/awesumproj
    - You want at least 10% of the input files to be available later
    - You want to keep ALL the intermediate files for now
    - You would like the final output products to be available to future researchers

Review / Cheat Sheet / Discussion

We need to design a workflow which makes best use of available storage and processing resources, avoiding unnecessary duplication, and being as efficient as possible:

- The input data are available in the CEDA Archive. On JASMIN, this actually means that we don't need to copy the data: we can simply process it in-place as it's all available at the file system level, read-only. So that's 50TB we don't have to worry about copying or storing.
- Even if we want 10% of the input files to be available later, we don't need to copy them, as the CEDA Archive provides long-term curation of data.
- We need storage for 25TB of intermediate output products. Or do we?
    - Could we break the processing up into chunks (e.g. 10% at a time?)
    - This would reduce the need to store ALL the intermediate products at any time.
- The initial processing requires storage which supports parallel-write.
    - Does our GWS `/gws/nopw/j04/awesumproj` support this? No.
    - We could use scratch space to store intermediate output products
    - We need to use scratch space which is parallel-write-capable. Currently this is `/work/scratch` (not `/work/scratch-nompiio`). but can be in high demand, so **chunking** the processing reduces the scratch space we need.
    - If we want to keep log information of all the processing, we should write 1 log file for each process, i.e. each job, or element of a job array, if we're using job arrays. It's good practice to do this, even if, as in this case, the (scratch) storage tolerates multiple processes writing to the same file. But it's essential to have 1 log file for 1 process if using storage which does not support parallel write.
- Where will we be doing the second processing?
    - You wrote the code yourself, so you should know whether it does parallel-writes
    - How could you test it?
        - Use "`lsof`" to examine what files are open by what processes during a small test-run. (Try "man lsof" to get further information about this command)
        - Example at: https://help.jasmin.ac.uk/article/4702-faqs-storage
    - If it's "safe", you could write the 1TB output data straight to the GWS.
- What are you planning to do with your final output products?
    - If you want them to be available long-term, you need to speak to the CEDA Archive team about getting them deposited into the archive. Your GWS does NOT provide long-term curated storage, although can be a good place from which to share the data with collaborators while you're working on them.
    - CEDA Archive datasets can be cited (and even have DOIs associated with them), so when your work hits the front page of *Nature*, you (well, your supervisor) can take the credit!
    - It would be appreciated if people and projects could acknowledge their use of JASMIN
        - See https://help.jasmin.ac.uk/article/4693-acknowledging-jasmin
        - Please also take part in information-gathering exercises run by the CEDA team: these help gather stories about the impact of JASMIN which help strengthen the case with funding bodies for the ongoing availability of JASMIN.
- If either your processing code, or your output products consisted of many small (<64Kb) files, what could you do differently?
    - Consider using a "small files" (SMF) area for your group workspace. Check with your GWS manager if one exists already, but otherwise one can be requested via them. This would provide much more performant storage for things like custom Python environments, or processes which involve creating lots of small output files.
    - Another option would be to review how your output is produced, or consider writing this sort of information to a database instead of the filesystem.

Alternative approaches and best practice

- Discussion!