HPC Data Management Policy

Introduction

NeSI is introducing a new data storage policy on their nobackup storage, and will delete files that haven't been accessed for the last 120 days.

Every 2 weeks, they will generate the list of files to be deleted (/nesi/nobackup/nesi00213/.policy/to_delete/latest.filelist.gz), and remind us to clean them up during the following 2-week time.

What does "access" mean?

According to NeSI, any file that has been opened or edited are regarded as "accessed".

What can we do?

1. Do nothing: Files in the list will be deleted after the grace period.
2. Backup/Relocation: We can review the list and decide which ones to keep. We can either archive them to nearline tape store or move them to elsewhere (eg. local pc, RCC etc)
3. **Fake access**: While we shouldn't "exploit" the insider information, the definition of "access" gives us an opportunity to keep the file on nobackup for at least another 120 days.

To Keep or To Delete ?

We will leave it as each user's responsibility to review the "to-be-deleted" list and remove the lines if the user wishes to keep the file in nobackup for another 120 days. The user will decide based on the criteria below.

1. What is the nature of this file? Is it science data or just some temporary data? Is it a code? (should be in /projects or $HOME in such a case) Why do we have this file in the first place?
2. How often do we need to access this data? It wasn't opened in the last 120 days - Does it mean it is outdated?
3. How easy is it to reproduce this file? Can we leave the bare minimum data and rerun the calculation? Does this file require software that is no longer available or supported (ie. old version)?

To Keep or To Delete?

```
If (Do I need this file ever again) == TRUE:
        If (Easy to reproduce) == TRUE:
                Keep the input files only and compute again when I need it
        Else:
                Keep everything //except temporary files
Else:
        If (Am I SURE???) == TRUE:
                Delete them
        Else: // Uh.... Not Sure....
                If (Easy to reproduce) == TRUE:
                        Keep the input files only and compute again when I need it
        Else:
                If (Do I know how to use Tape Storage) == TRUE:
                        Put them on tape storage
                Else:
                        Talk to Sung
```

Process

1. NeSI's fortnightly alert and latest.filelist.gz
2. SW team to produce separate lists for individual user
3. Each user edits the list, and removes lines for the files to be kept. The user runs touch_files_to_keep.py to "fake access" to the files that will be kept.
4. NeSI produces the final list of files to be deleted, which will be promptly reviewed. Unless objected, NeSI will go ahead with deletion.
5. SW team to review the process and rectify issues that emerged.

Issues

1. SW team may not necessarily have the permission to handle files produced by each user (unless correct group permission was set) : See Action 1,2 below.
2. After each user edited the list, they apply "fake access" to those files to keep. (If SW team does this, requires 1 above to be resolved)
3. If the same file appears in the to_delete list multiple times, the file becomes a candidate for migration to nearline storage. (how to do this is yet to be planned)
4. Ideally, NeSI gives the list of files to delete before the action actually takes place (ie. dry run), so that there is no surprise. (Requested NeSI and pending the response)

Actions

1. (Everyone) Include this line in your ~/.bashrc. (We need to have umask 0002 set to give other group members permission to manage your files)

```
source /nesi/project/nesi00213/share/bashrc.uceq
```

2. (Everyone) Make sure your files/directories have the correct group. If you create a new file/directory under **RunFolder**, it should have the correct group. (Sung to request NeSI to fix all group ownership & permission)

Good

```
drwxrwx---+ 6 leer nesi00213 262144 Mar 23 05:43 v20p4p8
drwxrwx---+ 6 leer nesi00213 262144 May 14 09:35 v20p4p90
```

Bad

```
drwxrwx---+ 6 leer leer 262144 May 30 05:39 v20p5p20
drwxrwx---+ 6 leer leer 262144 May 30 15:05 v20p5p21
```

If this is the case, fix the ownership and permission.

```
chgrp -R nesi00213 v20p5p20
```

```
chmod -R g+rws v20p5p20
```