Archiving data using the command-line interface

Hans van Piggelen, SURFsara

UvA HPC Course 2017: Data Management

Data Archive usage – Best practices

- Try to store files of significant size (> 1 GB) as much as possible. Smaller files will always be accepted, but will lower the performance of restoring your files from tape.

- If you have many small files, make sure to pack them using a file archiving tool like tar or dmftar.

- Try to pack your files before uploading them to the archive, possibly by using dmftar which can do this directly after packing your data.

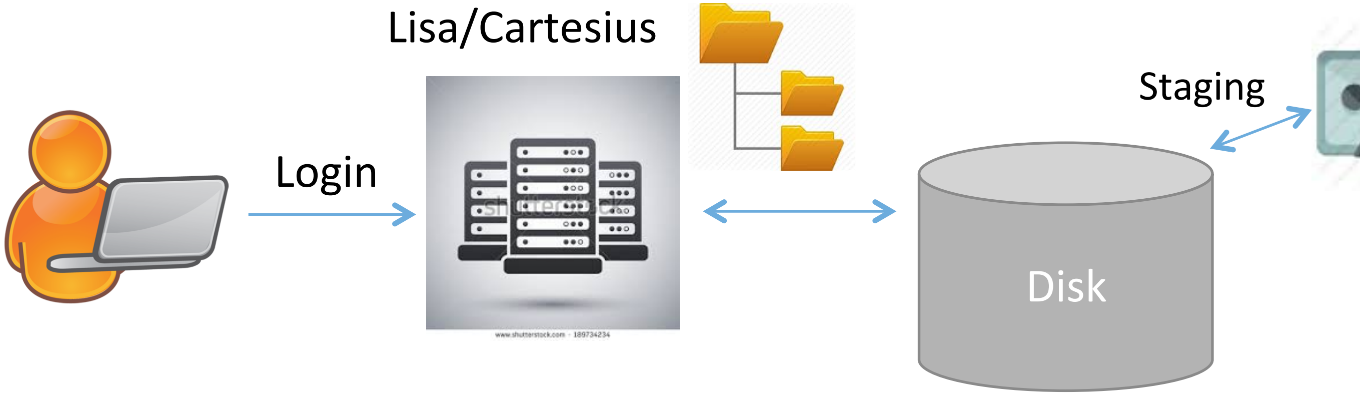
- Do not store more than 2000 files in a single folder on the archive

- Organise your files in such a way that in case the files are needed again only parts of the data needs to be restored from tape.

- Avoid storing unpacked software packages, these usually contain a lot of small files. Instead pack these as well, or refer to a specific software repository.

Archiving workflow

- In all cases:
  - User logs in to Lisa / Cartesius
  - User’s archive home folder is mounted as folder `/archive/<username>`
- Storing data:
  - Pack your data using tar or dmftar locally
  - Copy to archive
- Retrieving data:
  - Stage archived data
  - Unpack using tar or dmftar
- Direct archive access:
  - archive.surfsara.nl (CLI)



Archiving tools: tar

- A software tool for collecting many files into a single archive file or tarball
- Tarballs contain various file system parameters:
  - E.g. file name, time stamps, ownership, file access permissions, and directory organization

Archiving tools: dmftar

- Wrapper for GNU tar developed in-house by SURFsara
- Packs any file or folder to single dmftar archive (which is a folder) of one or more tarballs
- Thus contains the same information as tarballs, plus more:
  - Checksum of each tarball
  - File listing of original file and directory structure
- Understands underlying storage infrastructure: ‘tape-aware’
  - Automatically intelligently stages your archived files
- Automatically splits data into 10 GB volumes

Archiving tools: comparison

- tar:
  - Available everywhere!
  - All Linux distributions and OS X by default
  - Windows requires installation: Tar for Windows
- dmftar:
  - Only available on Lisa / Cartesius / archive
  - Automates extra tasks concerning data archiving
  - Ideal for archiving data on the Data Archive!

Archiving tools syntax

- tar syntax:

```
tar [options] [file] [pattern]
```

- dmftar syntax:

```
dmftar [task] [options] -f [[user@]remote:]dest/ [file|dir ..]
```

- Note: always use the right extension (‘tar’ and ‘dmftar’) for your archives!

Archiving tools syntax (2)

- Staging data from tape on the archive:

```
dmget -a [file]
```

- Pushing data to tape on the archive:

```
dmput [-r] [file]
```

- Wildcards are accepted!

Exercises

- Can be found here: http://bit.ly/2kjlwCB
- Consist of 5 parts of several exercises covering tar and dmftar usage
  - ±10 minutes per part
  - Including bonus exercises for each part
- Go at your own pace, solutions will be given at end of each part!
- For the experts: a bonus part if you have the time
- Your working directory on Lisa and archive: `~/data-management/`

Before you start:

- Read carefully!
- Make sure you are working in the right directory (differs per exercise)
- Good luck!

Wrap-up

- Archive your data using large tarballs or dmftar archives
- Use tar or **preferably the dmftar tools**
- Take into account the tape infrastructure of the Data Archive
  - Regarding the number of files and the individual file size

- Archiving is easy and provides safe storage of your data!