Lifelines UMCG High Performance Cluster (HPC) manual

Date 09-03-2023

Dear researcher,

This document provides an instruction on how to work on the Lifelines UMCG HPC and provides best practices to perform your research on this environment. If you are familiar with this environment, you will be able to find the genetic Lifelines data in the Gearshift cluster on this location:

```
/groups/umcg-lifelines/prm03/releases/
```

The requested phenotypic data from the Lifelines catalogue can be found in the Gearshift cluster in your project folder (ovYY_XXXX) on this location:

```
/groups/umcg-lifelines/prm0*/projects/
```

You can follow the login instruction here: http://docs.gcc.rug.nl/gearshift/logins/ and use your UMCG HPC-account to login to Gearshift. In the remainder of this document you will find more information regarding project folders, performing analyses and available support.

Content

Golden rules for working on the UMCG-cluster

On the cluster you are responsible for your own data management. Please be aware that the cluster has less safety measures than the Lifelines workspace and that poor data management may lead to data loss. Adhering to the following rules will help you in your work on the cluster:

1. Never export/download/upload any Lifelines related data from the UMCG-cluster without approval from Lifelines data management. Failing to adhere to this rule violates the code of conduct and can result in a revocation of your data access rights.
2. If you need to work on data for which you cannot use the cluster, always request advice from Lifelines data management on how to proceed.
3. Always write your executable code in a job scripts and use the job scheduler to run an analysis.
4. Create separate directories within your project folder to provide a structure for your data, scripts and results.
5. Timely archive your scripts/results to your /archive/ folder on the prm-storage and remember not to store any unnecessary (large) data-files.
6. You are free to download/upload data unrelated to Lifelines to the cluster and/or your projects folder. For example, if you are building an analysis pipeline with other researchers, we encourage you to use services such as Github to sync your scripts.
7. If you have any technical questions, on for example the cluster or the job scheduler, you can contact the HPC helpdesk (hpc.helpdesk@umcg.nl) .
8. If you have any questions on the data, you can always consult the README-file, which is located in the data release folder on prm, to see who to reach out to for questions.
9. If you have any questions on access rights or data handling, you can contact Lifelines data management
10. When in doubt, contact Lifelines data management. It is always better to discuss how to proceed, than to file an incident report because of a data breach.

Project and release folders

The Gearshift cluster uses three types of drives for running analyses and storing data, the tmp (temporary), the prm (permanent) and the rsc (read-only storage cache) drive. Within the umcg-lifelines group you can find four different drives, namely **tmp01** (for working on your scripts and storing temporary data files), two prm drives **prm03 and prm02** (for archiving data and retrieving phenotypic/genetic data from Lifelines) and **rsc01** (for using the data in your analyses). Below we specify what you can find in each drive, and how you can use the different drives.

prm-drives

The Lifelines data are stored in so-called release folders. The data is located here:

```
/groups/umcg-lifelines/prm02/releases/
/groups/umcg-lifelines/prm03/releases/
```

Information on the specific data you have requested and the location of these data will be emailed to you by Lifelines data management once you get access to the data. In each release folder a README.txt file is located. Please read the README file for more information about the data present in that release folder and contact the contact persons in these README's if you have any questions about the data.

The prm03 and prm02 drives are used for permanent storage of the data, and are not mounted for analyses on the data. This means that you cannot use the data stored here for analyses. For your convenience, the data has been copied to the release folders on the rsc01-drive. If you would like to have a specific selection of the data from prm03 or prm02 available for analyses, you can copy this part of the data to your project folder on tmp01. As a researcher, you have read-only permissions to rsc01, prm03 and prm02, giving you the ability to copy the data to your own folders, but not change the data. Please be aware that copies of the data within the projects folders on tmp01 should be temporarily stored there. The space we have within the umcg-lifelines group is enough but has its limits. Please check out the next section (tmp01-drive) for more information about the project folders on tmp01.

prm0*/projects/ The prm0* project folder can be used to archive any files you would like to store permanently or can be used to access phenotypic data. The requested phenotypic data from the Lifelines catalogue is stored in this folder. The project folder name will have the following format: ovYY_XXXX. This is similar to the project code of your Lifelines application and the project folder on tmp01. The location of the project folder is:

```
/groups/umcg-lifelines/prm0*/projects/ovYY_XXXX
```

```
*is either 2 or 3
```

As a researcher, you have read-only permissions to this project folder. If you need to archive files, you can do so in the subfolder /archive/. Please keep in mind that only those files that require archiving should be stored (so no data files that can also be found on /prm0*/releases). If you have any questions on the data that you would like to store/archive, you can also contact Lifelines data management (data@lifelines.nl) for more information.

tmp01-drive

/projects/ The tmp01 project folder is a working folder in which scripts and results from analyses can be stored. A project folder has at least one researcher that is responsible for its content. Other researchers can be added to a project folder upon approval by the main researcher and principal investigator and a signed code of conduct for that particular project. The project folder name has the following format: ovYY_XXXX. This is similar to the project code of your Lifelines application. The location of the project folder is:

```
/groups/umcg-lifelines/tmp01/projects/ovYY_XXXX
```

Please place your files within the project folder of your project. If you would like to protect your files from your fellow project members working within the same project folder, you can create your own user folder within the project folder on tmp01. You can restrict access for other group members by using the command ‘chmod’.

To prevent data duplication, only copy the files you would like to modify to your projects folder and use the unmodified files from this /releases/ folder.

rsc01-drive

/releases/ Data from prm0*/releases/ have been copied to rsc01/releases. The data located here can be used for jobs. Please check if all data is available in the release folder you need for your jobs. Files on prm03/releases and prm02/releases always have original, unmodified versions.

To prevent data duplication, do not copy any of the data that has been made available on rsc01/ to your own projects-folders, unless you really have to because you would like to modify its content.

Getting started

Explore the data

If you've received access to the Lifelines group on the cluster, you can access the data you've requested through your proposal. To get started, we advise to verify if you can access the relevant release folders, found in:

```
/groups/umcg-lifelines/prm0*/releases/
```

Here you will find the available datasets that are being used by Lifelines researchers. Please read the README-files which you can find for every release dataset you have access to.

Since these data are available for all authorized researchers, you will have read-only access to these folders. When you want to modify a file, you can copy it to your own project folder on tmp01.

The Lifelines phenotype data is located in your own projects folder on prm0*:

```
/groups/umcg-lifelines/prm0*/projects/ovYY_XXXX/dataset_order_yyyymm
```

The dataset order consists of three folders:
Folder containing the data as comma separated files (.csv):

```
/results
```

Two folders containing the metadata as comma separated files (.csv):

```
/variables
/enumerations
```

For information on the phenotypic data and data structure, please read the "Lifelines Data Manual_UMCG HPC.pdf" that was sent to you as an attachment in the data access email you received from Lifelines. You can also find this manual on Gearshift on this location:

```
/groups/umcg-lifelines/prm02/documentation/manuals/
```

If you need information on the label values and/or specific variables, please check out our wiki page: http://wiki-lifelines.web.rug.nl/doku.php.

Please be aware, how tempting it may be, **do not download a file to your own environment**. While it might be easy to read a .csv data-file in Excel, you are not allowed to do so. If you do require applications that have graphical user interfaces (such as Excel, SPSS, or STATA), Lifelines can offer a Lifelines workspace (please note that costs might be charged). Please inform Lifelines data management (data@lifelines.nl) if you would like to know more about this service.

Organize your project folder

If you have familiarized yourself with the available data, you can find your own projects folder in:

```
/groups/umcg-lifelines/tmp01/projects/
```

In this folder you can save your job scripts, store data and your results. You will have write permissions in this folder, so you can create your own subdirectories and move files from a release folder or from your project folder on prm02 to your project folder on tmp01 to prepare them for analysis.

Write your first job script

Having structured your project folder, you can now write your first job script to learn more about the job scheduler. The resources on the cluster are shared between all users and to best balance this, everyone's work is being scheduled. This means that your script might not be executed instantaneously, but may have to wait on work from other users to be completed first. To learn more about a job script and the settings you can provide to the scheduler, you can go to the analysis page from the HPC Helpdesk: http://docs.gcc.rug.nl/gearshift/analysis.

Mount the data to prevent data-duplication

Having tested your job script on a small subset of the data, you may eventually want to run it on the entire dataset. To avoid having each researcher copy an entire release dataset to his own project folder, we also provide a storage for releases found in:

```
/groups/umcg-lifelines/rsc01/releases/
```

If the dataset you want to use has not yet been mounted on this storage, please inform Lifelines data management (data@lifelines.nl). They will then copy the data for you to use.

For more experienced users that can mount/stage data themselves, it is advised to create a *data staging job* that runs before your analysis job. To submit your job, also consider the quality of service levels for job priorities, please visit for more: http://docs.gcc.rug.nl/gearshift/analysis/#5-qos-ds

Install packages (e.g. for R)

Please note that you can install packages (e.g. R) yourself. Through the UMCG HPC it is possible to access the internet to download packages. This is different from working on a Lifelines workspace, in which you are dependent on Lifelines data management to install packages for you. On the UMCG HPC, you are not dependent on Lifelines data management or the HPC helpdesk to install packages. Since you do have access to the internet, please also read the section 'Reporting an incident' in this manual.

Support data access

Questions on using the cluster

The Genomics Coordination Center (GCC) HPC helpdesk can provide help to access the cluster and for requesting certain software tools. In order for researchers to work with the data on the HPC cluster bioinformatics knowledge and skills are required. Please also consult the UMCG-HPC wiki page for more information: http://docs.gcc.rug.nl/gearshift/index.html. For questions on using the cluster or if you are experiencing problems, please contact the Genomics Coordination Centre HPC helpdesk (hpc.helpdesk@umcg.nl).

Questions on phenotypic data

Questions on phenotypic data can be send to Lifelines data management (data@lifelines.nl).

Questions on genetic data

Questions on genetic data can be send to the respective researcher responsible for the release folder. Each release folder will contain a README file describing how to get in touch with the responsible researcher.

Reporting an incident

User will report any data incident (i.e. (unintended) release, security breach, etc.) involving Lifelines data within 24 hours to privacy@lifelines.nl. If you're not sure and think you might have violated the term of the code of conduct, you can also reach out to Lifelines data management to inform them about the incident. They will help to assess the necessary subsequent steps.

Cluster & data e-learning module

In order to work on the cluster you have to be knowledgeable on how to use a Linux operating system and how to work with data in a command-line interface. Because the data cannot leave the cluster, you are not allowed to download the data files to your own (local) environment and change them in applications such as Excel, SPSS or STATA. If you do require these applications, you are advised to work on a Lifelines workspace (which is a Windows environment).

To help determine if the cluster is the right environment for you to work in and to get an idea of what you can expect, you can follow our cluster & data e-learning module. Through this e-learning you can watch course videos on working with the data in the cluster and you will be tested on your Linux knowledge.

For more information please check our website: https://www.lifelines.nl/researcher/explore-lifelines/e-learning