DAT470/DIT066
Computational techniques for large-scale data

Assignment 1
Deadline: 2025-04-20 23:59

Problem 1 (12 pts)

In this problem, we practice some fundamental UNIX commands and solve simple tasks of analyzing a bunch of files. Each subproblem is worth 1 point. In each subproblem, write in your report the command(s) you used to solve this task, in addition to presenting answers to the questions presented in that subproblem (if any).

(a) Log in to `minerva`.

(b) Download the Linux kernel version 6.13.5 sources from `https://cdn.kernel.org/pub/linux/kernel/v6.x/linux-6.13.5.tar.xz`.

(c) Extract the contents of the tarball.

(d) Determine the size of the tarball before extraction and the size of the extracted contents in *human-readable* units.

(e) The Linux Kernel consists of C code. C code is organized into *header files* (file extension `.h`) and *source files* (file extension `.c`). Determine the total number of header and source files in the sources, respectively.

(f) Find the source file (ending in `.c`) that is the longest, that is, that has the most lines in it.

(g) In C, header files are *included* in the source code by using preprocessor directive that looks like `#include <header.h>` or `#include "header.h"`. Write a *one-liner* that determines the names of the 10 most included headers in the kernel sources. You may assume that there is always a space between the `#include` directive and the filename. The filenames are surrounded by either chevrons (`<`,`>`) or quotation marks. There may be extra text after the directive (typically comments). You only need to consider files ending in either `.c` or `.h`.

(h) The file `meps.csv` (available on Canvas) contains semicolon-separated information about the Members of the European Parliament, since it was founded. Download the file from Canvas to your local computer. Then, copy the file into your home directory on Minerva over SSH.

(i) Determine how many MEPs Sweden had in the 10th (current) European parliament.

(j) Determine the name of the largest parliamentary group in the 10th (current) European parliament.

(k) Determine the names and birth dates of the oldest and youngest MEPs (by date of birth, with relation to *today*, not when they served) ever to serve in the European parliament.

(l) Determine the fraction of female MEPs out of all MEPs who ever served. *Note:* Each MEP counts distinctly for each term, so if the MEP served for four terms, they appear in the data four times; we will keep this interpretation and then count them four times, as we are interested in the gender ratio of people serving in the parliament, not the ratio of unique people.

Problem 2: Information of computers (12 pts)

In this assignment, we practise using Slurm by collecting information about different computers in the cluster.

Choose one computer out of the following four: `io`, `europa`, `ganymede`, and `callisto`. Also, choose one computer out of the following two: `uranus` and `neptune`.

Construct a single shell script (`.sh` file) that you can run locally on `minerva` (the login node) and using Slurm on the two computers you chose above. The script shall collect the following information about the systems:

- The model of and the clock frequency of the CPU
- The number of physical CPUs (sockets in use), the number of cores, and the number of hardware threads
- The instruction set architecture of the CPU
- The cache line length
- The amount of L1, L2, and L3 cache
- The amount of system RAM
- The number of GPUs and model of the GPU(s)
- The amount of RAM on the GPU(s)
- The type of filesystem of `/data`
- The total amount of disk space and the amount of free space on `/data`
- The version of the Linux kernel running on the system and the GNU/Linux distribution and its version running on the system
- The filename and the version of the default Python 3 interpreter available on the system (globally installed)

Carefully read the guides and instructions at `https://git.chalmers.se/karppa/minerva/`. Include the information you gathered in your report, preferably as a table with a column for each of the three computers.

Hints

- The following commands are probably useful (not an exhaustive list and valid alternatives exist):
  - `cd`
  - `cpuinfo`
  - `cat`
  - `cut`
  - `df`
  - `du`
  - `find`
  - `getconf`
  - `grep`
  - `lsb_release`
  - `lscpu`
  - `lshw`
  - `mkdir`
  - `nvidia-smi`
  - `scp`
  - `sftp`
  - `sed`
  - `sha1sum`
  - `sort`
  - `ssh`
  - `tail`
  - `tar`
  - `tee`
  - `uname`
  - `uniq`
  - `wc`
  - `wget`
  - `xargs`
- You cannot log in to any of the nodes, except `minerva`; you must execute the code using Slurm.
- Read the documentation on the GitLab page carefully.
- Slurm commands can also be directly included on the command line; for example, to run `foo.sh` on the node `callisto`, you can issue the command `sbatch -w callisto foo.sh`.

Returning your assignment

Return your assignment on Canvas. Your submission should consist of a report that answers all questions as PDF file (preferably typeset in LaTeX) called `assignment1.pdf`. In addition, you should provide the code you used in Problem 2 as `assignment1_problem2.sh`. Do *not* deviate from the requested filenames.