Mission Possible: Data Engineering with Unix Tools

The URL https://www.spinellis.gr/cgi-bin/oasa-history?id=YOUR_STUDENT_ID (e.g. https://www.spinellis.gr/cgi-bin/oasa-history?id=x2899999) returns a constantly updated stream of GPS location records associated with public buses in Athens. The data (more than 240 MB when uncompressed) are provided in a compressed format compatible with the gzip program. The stream's fields are: data acquisition time stamp, line number, bus number, position reporting time stamp, bus position latitude, bus position longitude.

Your mission, should you decide to accept it, is to analyze the data in order to help the bus company improve its operations, and thus improve everyday life in the city. Answer the following questions using Unix command-line filter tools, such as grep, fgrep, wc, sort, cut, tee, xargs, uniq, awk, sed, join, comm, diff, seq, head, tail, rev, tac, gzip, tr, tsort, etc. For each question provide the answer and the command(s) you used to obtain it. For multi-stage pipelines, add comments on a line before each command or after the command's name to explain what your commands do. (On the Unix shell comments start with the # character.) All commands given must produce exactly the answer you provide. (It is not allowed to interpret the data manually or with other tools.)

You are allowed (and encouraged) to consult command manual pages, course material, and Wikipedia. Given the data's large size you might want to first try out your solutions on a small subset of the data until you're satisfied with the output.

For all answers use the same version of the data file, which you will store locally (perhaps compressed) on the computer you use. Ensure that you provide your correct student-id in the URL you use for obtaining the data, otherwise your submission may not be graded. The deadline for finishing this exercise is February 28th, anytime on Earth. Do not answer any questions for which you cannot deduce your answer on your own. Note that you may be asked to repeat and explain your answers in an oral interview. Any violation of these rules or attempt to circumvent them will result in a grade of zero. At the end of this exercise you will be asked to indicate how many hours it took you to complete it and provide feedback on the exercise and the lectures. Good luck!

As always, should you be caught or killed while working on this mission, the faculty will disavow any knowledge of your actions. This page will not self-destruct in ten seconds.