The Groote Nationale Investeer Bank (GNI Bank) is a large European bank and is preparing to enter the mobile banking sector. They have hired your advisory firm to prepare a strategy. Your data analytics team has been put on the task of finding the best four datacenter locations for hosting in the EU. After this assignment you are able to efficiently load large data sets, combine and execute database queries to find relevant data and use it to solve realistic problems.

a Evaluate if there are limitations in the provided datasets (AS and probe data set). If you find limitations, describe these and conjecture possible reasons, supported with data.

b With the AS and probe data set, find the number m of AS's that can be used for hosting in the EU and have probes in the RIPE data set. Sort the ASN's in ascending order and include the first and last three in your report (number, name and country).

c For a single hour in the RIPE data set: find all valid entries where the probe has hosting type AS and the target IPv4 is from an EU country. Implement this in an efficient way.

d Move from using only an hour to the full day. It is advisable to store the raw results of each file. Then, using all processed files, calculate the average latency's for each country-AS combination and store the results into one $n_{countries} \times m$ matrix. If we could place one server in each country, what would the minimum average latency be for each country? (include in your report)

e Since we are only allowed to place four servers, determine the best four datacenters based on the total latency for all countries. Report your findings and your procedure to obtain them. Also include the average latency for each country.

The data sets that should be used:

• **RIPE data set:** The Ripe Atlas ping data set contains ping measurements executed by a select number of probes to most IPv4 addresses in the world. One day of measurements are stored in 24 files (one for each hour) totalling about 192 GB in size. The Ripe dataset can be found here: `https://data-store.ripe.net/datasets/atlas-daily-dumps/`. Use the data from **20-February-2021** starting with 'ping-'. Here you can find more about the data format: `https://atlas.ripe.net/docs/data_struct/#v5000`

• **IP location data set:** The IP location dataset can be downloaded from here: `https://lite.ip2location.com/database/ip-country`

• **AS and probe data set:** You can find the data sets on Brightspace.

Write a four-pager with your advice for the GNI bank before 11-March-2021 08:00 local time. Upload your draft report and code to Brightspace. You will receive feedback on your work ASAP. Use this feedback to improve you work for the final version.