FACULTY OF TECHNOLOGY, POLICY AND MANAGEMENT
ENGINEERING SYSTEM AND SERVICES SECTION
ICT TRACK

SEN163A – Fundamentals of Data Analytics
Formative Assignment - Large-scale Internet Data Analysis

dr. Jacopo De Stefani - Joao Pizani Flor
Based on the material by T.Fiebig

February 28, 2022

Learning Objectives

- Identify fundamental pitfalls inherent to data analysis and identify whether they exist in case-studies.
- Propose and develop suitable data mining pipelines for different case-studies.
- Learn to efficiently load large data sets.
- Combine and execute data queries to find relevant data.
- Analyze a given dataset to answer a given research question.

Disclaimer

All characters and other entities appearing in this work are fictitious. Any resemblance to real persons or other real-life entities is purely coincidental.

Introduction

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.

To guide your work, GNI Bank has proposed the following plan of action:

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 are:

- **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 **01-March-2022** 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:** Datasets containing information about the ASNs and the corresponding probes. Available on Brightspace.

You will report your advice through a **Jupyter Notebook**, including both your code and considerations on your findings. The use of figures, formulas, tables and pseudo-code to support your analysis is strongly encouraged. In case you used specific python libraries, you will need to include these libraries in a `requirements.txt` file. A template for the **Jupyter Notebook** and an example of `requirements.txt` file are available on the assignment page.

Evaluation criteria

The final grade for this assignment will be calculated based on the following criteria:

- **Quality of the report** - 35%
  - Reasonable formatting of the document and used citation appropriately
  - Reasonable formatting of the document and used citation appropriately
  - Use of proper English (typos, grammar)
  - Code script deliverable
  - Code quality
  - Problem Description
  - Dataset Description
  - Limitations
  - Use of figures and tables to summarize results.
  - Conclusion/Action recommendations
- **Functional tasks a. to e.** - 65%

Rules for the assignment delivery

To be read carefully !

1. The assignment must be developed in groups of 4 students.
2. The assignment must include the **name** and **student id** of all the students.
3. The assignment must be submitted in **Brightspace** as a **Zip file** containing:
   - A jupyter notebook (`.ipynb`) containing your code and the discussion of the results
   - *(If you used additional libraries)* A file `requirements.txt` containing, one per line, the additional employed libraries to solve the assignment.
4. You have to follow the following constraints:
   - Upload of a file `Group_X.zip` on the Brightspace page of the course, where `X` should be replaced by your actual group number./
   - Date: **Friday 18 February 2022**
   - Time: **Before 18:00**

   After this deadline the assignment will be considered as late and **no feedback will be provided**.
5. Use the feedback you will receive to improve your work for the final version due on **12 April 2022 18:00**.
6. **Knock-off criteria:**
   - Missing names and id on the document/document name.
   - Code not executable: Missing libraries - Errors.