# SEN163A Fundamentals of Data Analytics

August 10, 2020

## Introduction

The Groote Nationale Investeer Bank (GNI Bank) is a large European bank, that is preparing to enter the mobile banking sector. To enter the banking sector, the bank needs to know the four best locations for hosting in the EU. To prepare a strategy they would like to make use of data analytics to find this location. In the report, first the contents off the dataset will be described. This dataset will be used to answer the four questions.

## 1 Description of the dataset

The initial dataset consist of 3 different sets of data; AS, Probe and Ripe Atlas. Every dataset has different variables which are explained below.

**AS:**

1. **ASN:** ASN is unique for every different AS.
2. **Country:** The country in which the company is located.
3. **Name:** The company name.
4. **NumIPs:** The destination IP address.
5. **Type:** The type of the AS.

**Probe:**

1. **prb_id:** The unique id which is given to a probe.
2. **ASN:** ASN number which is unique for every different AS.

**Ripe Atlas:**

1. **dst_addr:** The destination address
2. **src_addr:** The source address
3. **prb_id:**The unique id which is given to a probe.

## 2 Possible locations for hosting

First, the country codes of the EU are used to filter the AS dataset, in such a way that only the AS's of the EU are left. After that, the probes in the probe dataset that also occur in the AS dataset are filtered out. The final number of ASN's filtered out is 166. Subsequently, they have been sorted in ascending order and the first and last tree are stated in table 1 and table 2.

## 3 Filtering probes and IP's

In the first part of this assignment we've used the AS and probe data set to find the ones that can be used for hosting in the EU and have probes in the RIPE data set. The data set made is used in the second part

Table 1: Head, first rows of the data set

| | ASN | Country | Name |
|---|---|---|---|
| 4022 | AS5404 | AT | conova communications GmbH |
| 18697 | AS5430 | DE | freenet Datenkommunikations GmbH |
| 18895 | AS5521 | DE | PlusServer GmbH |

Table 2: Tail, last three rows of the data set

| | ASN | Country | Name |
|---|---|---|---|
| 15377 | AS203953 | DK | Hiper A/S |
| 54697 | AS205544 | GB | LEASEWEB UK LIMITED |
| 19599 | AS205766 | DE | Jonas Pasche |

of the assignment to find all valid entries, where the probe has hosting type the target is IPv4. In order to do so, the data set has to be filtered. The data set is filtered on EU ASN, on ’hosting’ at the type of data and probes in probe data set that occur in AS data set. Which results in the entries given in table 3.

Table 3: Valid entries

| | prb_id | ASN |
|---|---|---|
| 27 | 57 | AS20621 |
| 42 | 92 | AS12859 |
| 47 | 107 | AS25596 |
| 61 | 142 | AS12637 |
| 62 | 144 | AS12637 |

The dataset is already sorted, therefore a binary search with complexity O(logN) can be used. Our search function repeatedly divides the search function in half. It will check the intervals for the value you want to find or if the interval is empty. This means for our dataset, since the IP dataset has 174972 entries, there are at maximum log(174972)/log(2) = 17.42 iterations needed to find the range of the decimal IP-number. The number of iterations needed must be therefore lower then or equal to 17.

## 3.1 RIPE data set quality check

The RIPE data, which is in JSON, need to be quality checked. Below is an example on how to do so and the entry of RIPE database looks.

if *dst_addr* in jsonline and *avg* in jsonline and *prb_id* in jsonline and *:* not in jsonline[”dst_addr”] and jsonline[”prb_id”] in probs[”prb_id”]. Values and jsonline[”avg”] is bigger than 0:

We determined for which probe ID values the avg is bigger than zero. If there is a destination address in the data, an avg, a probe ID that is ipv4 and if the probe ID nr is in probe ID values.

’fw’: 4900, ’lts’: 19, ’dst_name’: ’192.203.230.10’, ’af’: 4, ’dst_addr’: ’192.203.230.10’, ’src_addr’: ’192.168.42.168’, ’proto’: ’ICMP’, ’ttl’: 56, ’size’: 20, ’result’: [’rtt’: 19.971325, ’rtt’: 17.799955, ’rtt’: 16.52825], ’dup’: 0, ’rcvd’: 3, ’sent’: 3, ’min’: 16.52825, ’max’: 19.971325, ’avg’: 18.0998433333, ’msm_id’: 1013, ’prb_id’: 10046, ’timestamp’: 1582157156, ’msm_name’: ’Ping’, ’from’: ’82.217.84.104’, ’type’: ’ping’, ’step’: 240

## 4 Minimum latency per country

To calculate the minimum average latency for each country-ASN combination, the following procedure is used to obtain them. After running the binary search algorithm for appropriate IP list, we obtain the file output-final-v2.txt which is read as latency3 dataset containing the columns 'IP_dest' and 'Country'. The data set ASN_IP contains the columns 'ASN', 'to_IP', 'latency' and 'ASN_IP'. These datasets are merged on IP, modified and resulting dataframe fin_df is saved as output-df-final-v2.txt. The file is read and pivoted resulting as df_final which leads to the country-AS combination. The dataset contains 27 rows and 163 columns. From the country-AS matirx, the minimum latency per country is calculated and documented in table 4.

Table 4: Minimum average latency per country

| **Country** | **Latency** |
|---|---|
| Austria | 2.715995 |
| Belgium | 1.446367 |
| Bulgaria | 12.984955 |
| Croatia | 11.126990 |
| Cyprus | 106.234629 |
| Czechia | 0.802876 |
| Denmark | 1.169748 |
| Estonia | 1.457398 |
| Finland | 1.230155 |
| France | 2.867094 |
| Germany | 3.573851 |
| Greece | 14.395791 |
| Hungary | 5.218637 |
| Ireland | 3.787307 |
| Italy | 2.601187 |
| Latvia | 0.962479 |
| Lithuania | 0.693350 |
| Luxembourg | 4.538076 |
| Netherlands | 2.034900 |
| Poland | 1.021413 |
| Portugal | 1.416076 |
| Romania | 3.888023 |
| Slovakia | 0.666608 |
| Slovenia | 11.277749 |
| Spain | 3.719392 |
| Sweden | 0.929297 |
| UK/GB/Northern Ireland | 2.178014 |

# 5 Optimal server placement

Per AS, the average latency is taken. Subsequently, ASN's corresponding country is determined and all the values of this country are summed and averaged. To obtain the average latency per country, these values are divided by number of ASN. The results are shown in table 5.

Table 5: Average latency per country

| Country | Average latency |
|---|---|
| Austria | 43.508257458062424 |
| Belgium | 28.203687076005053 |
| Bulgaria | 53.23272833293965 |
| Croatia | 48.04861101347622 |
| Cyprus | 150.6038375393496 |
| Czechia | 30.653148503835222 |
| Denmark | 34.48028741992613 |
| Estonia | 48.714345215683984 |
| Finland | 43.774014053762976 |
| France | 28.377659175186167 |
| Germany | 30.468190317461755 |
| Greece | 64.23782282901065 |
| Hungary | 41.242103736350735 |
| Ireland | 37.95326785301388 |
| Italy | 37.46602313564135 |
| Latvia | 46.45373207934853 |
| Lithuania | 45.29838262182955 |
| Luxembourg | 31.130479213132226 |
| Netherlands | 24.952344725781078 |
| Poland | 34.401717783861784 |
| Portugal | 48.62662604205458 |
| Romania | 50.631103698223995 |
| Slovakia | 41.02811037577967 |
| Slovenia | 44.22693156789418 |
| Spain | 45.532176445025435 |
| Sweden | 35.22241898782231 |
| UK/GB/Northern Ireland | 32.69725918780929 |

The matrix of Country and ASN with average latency is obtained from previous step. There are 27 countries with 163 ASN values. First, the combinations of 4 ASN is calculated. An exhaustive enumeration algorithm is created to identify the mininum latency across all countries and their respective sum of latency is calculated. By running the algorithm for the combinations, we get the latency value for all combinations. The number of combinations are 28342440. The minimum latency value is calculated and the data centres are identified. The four datacenters are as follows: (AS25151, AS30823, AS47869, AS60610)

Table 6: The ASN and country matrix for corresponding value

| | AS25369 | AS30823 | AS47869 | AS61138 |
|---|---|---|---|---|
| Austria | 16.601123 | 106.607949 | 15.568479 | 18.288481 |
| Belgium | 10.129796 | 15.705225 | 41.071473 | 23.973938 |
| Germany | 10.025316 | 19.574280 | 35.957525 | 26.185254 |
| Netherlands | 3.223430 | 25.306029 | 8.734829 | 6.325226 |
| Sweden | 0.929297 | 16.653729 | 2.597707 | 3.732385 |
| UK/GB/Northern Ireland | 8.684740 | 22.989454 | NaN | 3.831218 |

To obtain the four data centers for all countries, we replace Nan values with mean of the columns in the initial country-AS matrix and same process is repeated. In this case, the four data centers identified are: (AS25151, AS29405, AS42442, AS62282)

# 6 Conclusions and recommendations

To conclude, the data retrieved from the GNI bank consisted of three different data sets; RIPE data set, IP locations data set, AS data set and Probe data set. Before the analysis the data had to be prepared in order to come to a conclusion. The first the four data sets had to be combined to one, in order to find the number of AS's that are able to host in the EU and also had probes in the RIPE data set. Secondly, the probes which have hosting type AS and the target IPv4 is from an EU country, had to be found. Which could be done by first writing an O(log n) search function to search the IP location data set for IPv4 that are from a EU country.

Eventually, by going from only a small subset to the entire one hour files, it's possible to retrieve the minimum average latency for each country when placing one server in each country. Since the GNI bank was set to place only four servers in the EU, the most optimal locations were examined. The resulting four location where it is most optimal to place a sever are; AS25151, AS30823, AS47869,AS60610.

## 6.1 Discussion

Due to limitations in the data or analysis the results could be more optimal. One of the limitations dealt with is the fact that the data centers had to be in Europe. When taking the location in consideration a more optimal location could be found. Secondly, the option of creating new AS's could influence the outcome. The last limitation is the restriction to only place four servers. By giving an exact amount of servers that can be placed, the solution is limited and will never be the most optimal outcome.