CSC_8634

Harvey Yuan 0077439

10/01/2022

Aim and Objective

The aim of this analysis is to explore and see if there is any correlation between the number of packets exchanged by the database and the when they were exchanged during the day. By reviewing when the most frequent number of packets are received by the database, we are able to understand the load on them over an average day. By doing so, it will enable the data center to sustain the continued loads. This will also allow for engineers realise faults should any arrive due to an increased reception of packets. In addition, this will allow the engineers to decide when the capacity of the data center needs to be increased due to an increased in the average load.

Technologies and Libraries Used

Throughout this project, there were a number of different libraries which were used with the main library being "ProjectTemplate".

R version 4.1.2

R is a language used for statistical computing and graphics.

RStudio 2021.09.01 Build 372

RStudio is an IDE that is used for R. RStudio includes an in-built console and terminal that allows for direct code executing and also a pane which allows for files, plots and packages management. Different libraries can be installed through this pane.

ProjectTemplate version 0.10.2

ProjectTemplate (PT) is a system that allows the user to automate medial parts of a data analysis project such as the organisation of the project files and processing data (Darke, n.d.).

This project utilises PT heavily to organise and mange the project files throughout this analysis. The "data" folder stores the raw data that will be analysed as part of this project.

PT also stores pre-process scripts, which will run as soon as the project has been loaded by PT which is stored in the "munge" folder. Similar to the munge folder, the "src" folder stores scripts that can be manually run.

dplyr version 1.0.7

The dplyr library allows for the user to use a set of verbs to write code for common data manipulation steps. It allows the user to use familiar words when scripting analysis and pre-processing scripts (such as filter), both for easy understanding of the code and proof reading (Hadley Wickham, n.d.).

ggplot2 version 3.3.5

ggplot is a system for creating graphics (where the data is a selected data frame). The data is either provided by the user or created in script, which allows for said user to iteratively add new layers, components and functionality. ggplot will be used to demonstrate the analysis as part of this investigation.

data.table version 1.14.2

data.table is an extension of the data.frame package in R. It allows for quick manipulation of data, especially where adding/updating columns of large datasets, which was a main factor in this project.

Scales version 1.1.1

The scales package allows for ggplot to create axis label which override the default breaks which allows them to be easily read.

Data Understanding and Preparation

As part of this project, there were a number of different data sets from 3 of Facebook's production cluster (Database, Web and Hadoop Servers). Each cluster contains a compressed dataset in tsv format which includes the following variables:

- timestamp
- packet length (1)
- anonymized(2) source (src) IP
- anonymized(2) destination (dst) IP
- anonymized source (src) L4 Port
- anonymized destination (dst) L4 Port
- IP protocol
- anonymized source (src) hostprefix (3)
- anonymized destination (dst) hostprefix (3)
- anonymized source (src) Rack
- anonymized destination (dst) Rack
- anonymized source (src) Pod
- anonymized destination (dst) Pod
- intercluster
- interdatacenter

However, as the aim and objective of this analysis is to investigate if there is any correlation between the number of packets exchanged by the database and when they are exchanged, not all of the data provided will be useful towards this analysis. Within the provided clusters, each file contained a vast number of datasets. As "packet sampling does not disturb the anomaly size when measured in volume metric" (Daniela Brauckhoff, 2006), it was decided that a small sample of the ~270 data subsets from Cluster A would be used. From the list, the first, last and two random subsets were chosen to be a part of this investigation.

As the datasets were large in size, they were compressed and as such, required decompression in order to be loaded onto the platform.

Upon inspection of all of the variables of the raw dataset, only the "timestamp" column and the "packet length (1)" column will be used as these were the only columns which had the required information for this particular study. The timestamp within the subset is in the Unix Epoch format, so in order to wrangle this data, it was important to convert this into a human readable format. This will also help with the plots. Once the data was transformed into a human readable format, it was found that the data was from 1st of October 2016 07:00 AM to 2nd of October 2016 07:00 AM. This meant that, for the purpose of this investigation, the time period perfectly aligned to what was being investigated.

All four subset of data were combined into one main data subset to ensure the remaining wrangling required only one set of code. The Epoch time was converted to the human readable format of yyyy-mm-dd HH:MM and from here, got their own columns to aide in further easier data wrangling.

Since the data only spanned across two days, the data was filtered to for the 1st and 2nd to have their own sets (in retrospect, this was not required as the remaining data wrangling did not require the dates to be

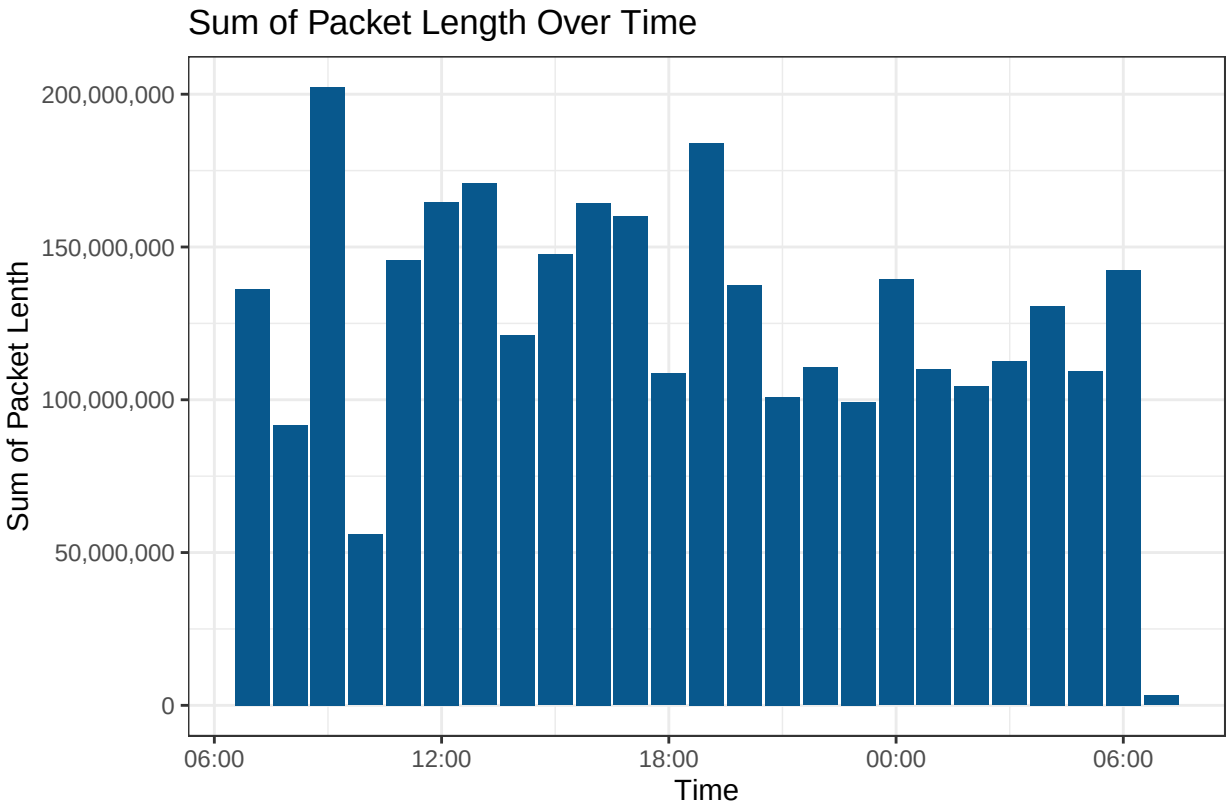
separate). Each filtered dataset was then sorted by Hour, Mins and Secs and then grouped up by Date and Hour. They each had their own data frames created to ensure the data was still being correctly processed.

Once the data was grouped, the next step was to wrangle the packet length column. With the packet length, because Facebook uses TCP segmentation offload, the packet length could be longer than the maximum of 65535 bytes. For this analysis, the packet length column will be analysed in two different ways. One method in which the packet lengths will be analysed will be by summing the total length for each hour and plotting the total length on a graph against the time. Although the sum of the packet length will have an arbitrary meaning, it has an intrinsic value of displaying showing the size of the packets over the time period. Another method in which the packet length will be analysed is by counting the individual packets over each hour. This method will show how busy each hour of the day is and highlight when the peak times which the most packets are received and offloaded. Should there be a spike or dip in the daily average of packets, engineers will be able to use this data and analyse and determine the cause.

As such, two data frames were created from the originally loaded data. The first data frame contains the Date Time grouped by the Hour and the packet length summed, whereas in the second data frame, the individual packet lengths is counted.

Analysis and Results

With the objective of this analysis being to investigate the correlation of the exchange of packets by the database, the sum packet length and count of packets lengths were plotted against each hour to show the increased and decrease of traffic load throughout the day. This will allow for a visual representation of how many packets are received and offloaded during peak and off-peak times.



Graph 1

From the above plot, the data suggests that the Altoona Data Center received the largest packets from approximately 11:00 until 13:00 and then once again from approximately 16:00 until 17:00 with the exception

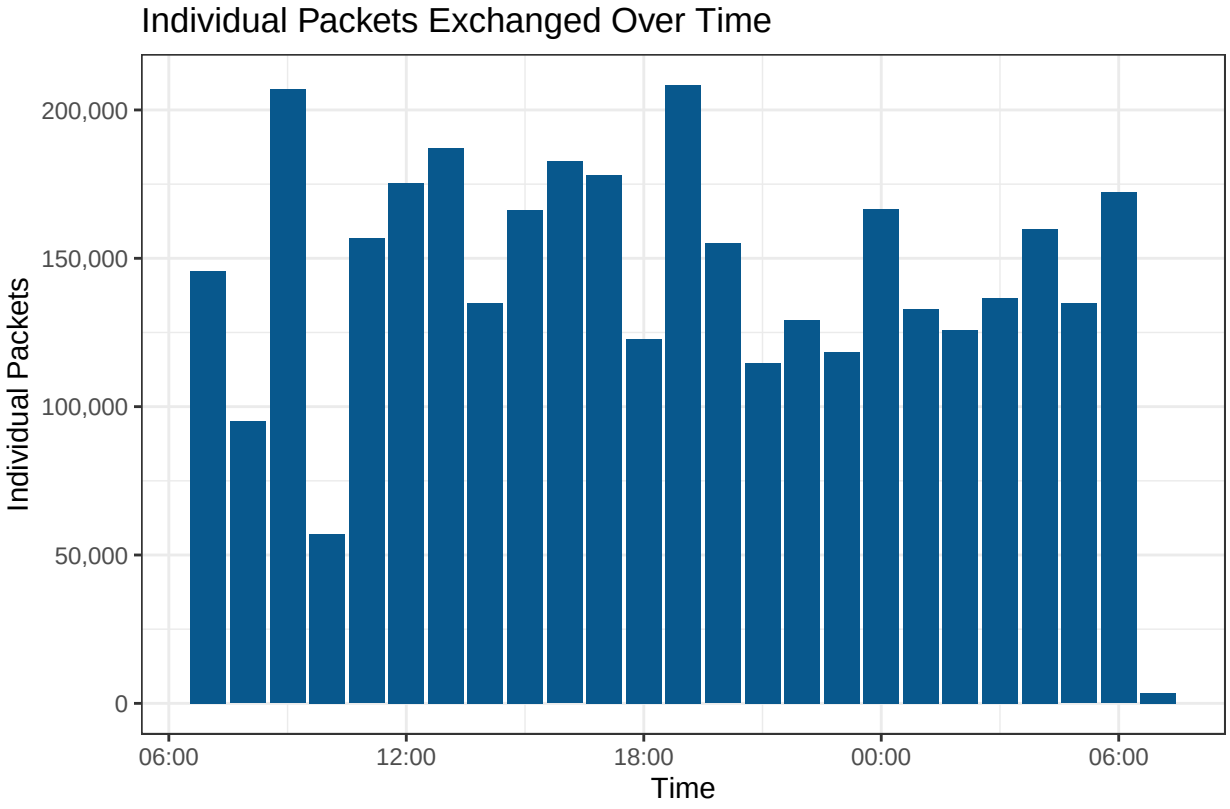
of 09:00 and 19:00 having abnormally high peaks. With the overall mean of the sum packet lengths across the day to be ~143,000, the average over the peak times are as follows:

Table 1: Average Packet Length at Peak times

Peak Times	Sum Packet Length (Nearest 1000)
11-13	173,000
16-17	180,000

The peak times from the table above shows that from 11:00-13:00, the sum packet length was approximately 19% higher and from 16:00-17:00 the total was approximately 23% higher. The results would suggest that most of the traffic was generated around lunchtime and also when people left work/school.

As mentioned, there are two peaks which appear within this data, once at 09:00 and once at 19:00. This makes sense as the morning peak suggest that much of the traffic is being generated when most people just starting their work and also when they arrive home.



Graph 2

Graph 2 shows the individual packets over time for each hour over the 24 hour period. The graph mirrors Graph 1 in terms of where the peaks are. This reinforces what is said previously that a large number of packets are being exchanged around lunch time and when people arrive home.

Although both of the graphs show that there is a peak around the aforementioned peak times, the data does not show this as pronounced as originally thought. Before the plots were created, the expectation was to see a plot similar to a bimodal distribution with peaks around morning and late afternoon. This was based off the assumption that most people only visit Facebook around the peak times.

Summary

As the aim of this investigation was to explore the the effect of packets the packets and the time of day they were received, I believe that the plots clearly show this, however, if I were to repeat this project, I would spend more time and delve deeper into the packets themselves and investigate where they arrive from and offloaded to. Once the region of the packets have been determined, they can be separated into their own data frames and evaluated further. By doing so, it will allow for more accurate plotting of the packets over time graphs as it would allow the packets from each region to be plot on it's own graph and therefore show a more evident peak in the increase in activity without more active regions communicating with the data center.

An initial issue I can foresee with this method of analysis is the legal requirements to keep the IP address anonymous, this issue was raised in "The Effect of Packet Sampling on Anomaly" paper which stated that due to those restrictions, it would "make it difficult to collect such detailed packet-level data" (Daniela Brauckhoff, 2006).

Another foreseeable hurdle with further evaluation of the packets would be the decryption of the anonymous IP addresses and sorting by region (if hypothetically the legal requirement of data protection was not in place). It might be difficult to decrypt different versions of IP between the different protocol numbers.

This project has taught me a lot regarding time formatting and time series data. I found that without outside help, the Unix Epoch format for time was not easy to understand and not as intuitive as many other standard time formats, however, since having hands on experience with the format, I can understand and appreciate why the particular format is used as timestamps.

To conclude, I believe that the results of the plot demonstrate the increase and decrease in traffic to the Facebook Altoona Data Center across over the course of a day which will allow for the engineers to realise any faults or unusual traffic. However, I would like to have delved deeper into the originating region of the packets and plot a graph similar to the ones produces in this investigation with the addition of global region.

Bibliography

Darke, P. (n.d.). Reproducible data science techniques in actuarial work. Retrieved from https://philipdarke.com/reproducible-actuarial-work/exercise1 (Last accessed 21st of January 2022)

Daniela Brauckhoff, B. T. (2006). The Effect of Packet Sampling on Anomaly.