Facebook Analytics – Technical Report

The goal of this project is to evaluate and analyse the performance of three separate clusters from the Facebook data centre in Altoona- database, web, and Hadoop. The major purpose or need of the project cannot be determined just based on the project description during the initial analysis because the analysis may vary depending on the datafile collected. A process is followed to determine the project's demand, allowing us to focus on the analysis with the available datafiles and then finish the approach to the findings. The natural inclination towards the backend of the Facebook and huge data centers instigated me to choose this project and this paper[1] helped me to understand the basic structure and process behind the data center.

The CRISP-DM technique is utilised to proceed with the findings, allowing us to first complete the Business understanding and then the Data understanding. In terms of the project's business description, it is to assess the Supercomputer's performance. The data is collected and evaluated using the CRISP-DM approach to see if the datasets are connected to one another. I were able to analyse and draw conclusions from the database and Hadoop Cluster because I were unable to obtain the dataset for the web server cluster.

The data pre-processing step is considered before moving on to the analysis section. Both the database and the Hadoop cluster had 273 compressed bz2 files each. I aggregated all 273 files from a cluster into a single data frame after decompression. There were corrupted values in the succeeding variables after the data frames were entirely merged, and these values were cleaned from the dataset. To aid our study, the dataset was categorised using each parameter variable, starting with the timestamp, and ending with the interdatacenter. The timestamp values were originally in Unix format and then transformed to BST values.

I decided to classify each cluster into three – intra cluster and intra datacenter, inter cluster and intra datacenter, inter cluster and inter datacenter. I can identify traffic with the timestamp and identify the time where the machine-to-machine traffic increases in a day. The top 10 source and destination IP address combinations which contribute a lot of traffic in both clusters. Then the analysis carried on in the same manner with the rack, pod, host prefix, L4 port. Then I deep dive into the analysis with the IP protocol number. With the dataset, I can find that there are three IP protocols used across both clusters. TCP, UDP and ICMP-IPV6 are three protocols used in which contributes more to the traffic is TCP followed by UDP and ICMP-IPV6. I further drilled down the machine with the classifications of host prefix in sync with the IP protocol. There were three levels of classification I considered for this analysis. The first level was the cluster, the second was with the IP protocol and the last was the classification I introduced with the cluster and the datacenter.

For these classification of IP protocols, I used normal usage of table format to show the number of requests and replies used across. Other than that, I used graphical plots to depict each part of the analysis.

I cannot fully utilize the complete dataset to continue the analysis of discovering the observations with a high runtime for this event. As a result, this approach can be applied to future research. As explained earlier, there are some outliers which still needs some more information to proceed with the analysis. With the structured abstract, I concentrated on the packet length transmitted across with a way to balance co-location of application on servers[2]. In this analysis, the pre-processing phase is crucial. It took a long time to set up the data to begin with the analysis. Before settling on the final dataset, many attempts at creating the entire dataset were made. After obtaining the ideal dataset, the entire study appeared to be basic and straightforward. These pre-processing, analysis, and other stages were made easier by using the project template from the beginning, and they were made even easier by employing the CRISP-DM approach. Each pre-processing step is saved separately in the munge folder of the project template, and analysis file is stored separately in the source folder of the project template. The project template is important in simplifying the analysis process and making it easy to grasp if the code changes for future analysis. The use of a project template and the CRISP-DM methodology pays off because they reduce time and provide clear information about how the analysis will be conducted. Similar initiatives could be conducted in a better version in the future with this technique and specific focus on data pre-processing, as this adds to the experience.

References

1. https://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p123.pdf
2. https://engineering.fb.com/2014/08/08/production-engineering/making-facebook-s-software-infrastructure-more-energy-efficient-with-autoscale/