Course: Stream Analytics

Team Size : At Most 2 students per team

November 30, 2024

Objective: Fetch, analyze and classify the data provided in the link given below:
`https://tools.netsa.cert.org/silk/referencedata.html#`

With the above data your task is to achieve following objectives:

- Retrieve the data stream that represents only the TCP traffic flowing through the network. To do this, you should become familiar with the SiLK suite, a tool for analyzing the network and its components.
- Classify nodes based on the amount of TCP traffic they process. For instance, you can group nodes that handle TCP traffic within a specific range, such as 10-100 packets per second, into one class. Use any two algorithms, say Very Fast Decision Trees (VFDT) and On Demand Classification, discussed in the lectures to perform this classification. During the demonstration, you should demonstrate the accuracy of your classification by sending a query that retrieves data from at least two different ranges.
- Detect the anomaly in the fetched data. Anomalies are defined as the nodes which send large number of TCP packets per sec, say 1K per sec. You are free to define your own threshold based on your observation in the given data stream.

Methodology: Before you start you assignment, you should make yourself comfortable with different tools involved in the SiLK suite. For this you should commence your learning process with the understanding basics of it, such as its retrieval commands for a small dataset, etc. I would suggest to refer to the following book for the same.
`https://tools.netsa.cert.org/silk/analysis-handbook.pdf`

Post this, you can proceed with the installation of the SiLK suite. For which you can refer to the following link
`https://tools.netsa.cert.org/silk/silk-on-box-deb.html`

For this assignment I would strongly suggest to use Ubuntu OS, with version ¿ 18.04, as most of the packages are developed such that they are compatible to it.

Expected Output:

- Screen shot of terminal showing the summary of TCP packet in SiLK suite.
- Graphical representation of the classification and the anomaly detection. You can use python library, such as pyplot, to plot the result. Furthermore, you are free to plot any other available tool to plot the answer.

SUBMISSION INSTRUCTIONS:

1. The assignment should be done in a group consisting of a maximum of *TWO* students.
2. The code should follow programming etiquette with appropriate comments.
3. Add a **README** file which includes a description of the code and gives detailed steps to compile and run the code.
4. Name your files as **config**.csv/txt and **outputfile.txt**.
5. Zip all your code as a single file with the name **rollno1-rollno2.tar.gz** and upload it to Google class room. Only one group member should upload the file.
6. Please note that you won't be allowed to make any changes to the code once submitted. You can only make changes to the topology file. Any changes to the code will result in 20% penalty of total marks. The penalty will increase with the amount of changes done.