Description

Write a program that can parse a file containing flow log data and maps each row to a tag based on a lookup table. The lookup table is defined as a csv file, and it has 3 columns, dstport,protocol,tag. The dstport and protocol combination decide what tag can be applied.

Sample flow logs (default logs, version 2 only).

2 123456789012 eni-0a1b2c3d 10.0.1.201 198.51.100.2 443 49153 6 25 20000 1620140761 1620140821 ACCEPT OK

2 123456789012 eni-4d3c2b1a 192.168.1.100 203.0.113.101 23 49154 6 15 12000 1620140761 1620140821 REJECT OK

2 123456789012 eni-5e6f7g8h 192.168.1.101 198.51.100.3 25 49155 6 10 8000 1620140761 1620140821 ACCEPT OK

2 123456789012 eni-9h8g7f6e 172.16.0.100 203.0.113.102 110 49156 6 12 9000 1620140761 1620140821 ACCEPT OK

2 123456789012 eni-7i8j9k0l 172.16.0.101 192.0.2.203 993 49157 6 8 5000 1620140761 1620140821 ACCEPT OK

2 123456789012 eni-6m7n8o9p 10.0.2.200 198.51.100.4 143 49158 6 18 14000 1620140761 1620140821 ACCEPT OK

2 123456789012 eni-1a2b3c4d 192.168.0.1 203.0.113.12 1024 80 6 10 5000 1620140661 1620140721 ACCEPT OK

2 123456789012 eni-1a2b3c4d 203.0.113.12 192.168.0.1 80 1024 6 12 6000 1620140661 1620140721 ACCEPT OK

2 123456789012 eni-1a2b3c4d 10.0.1.102 172.217.7.228 1030 443 6 8 4000 1620140661 1620140721 ACCEPT OK

2 123456789012 eni-5f6g7h8i 10.0.2.103 52.26.198.183 56000 23 6 15 7500 1620140661 1620140721 REJECT OK

2 123456789012 eni-9k10l11m 192.168.1.5 51.15.99.115 49321 25 6 20 10000 1620140661 1620140721 ACCEPT OK

2 123456789012 eni-1a2b3c4d 192.168.1.6 87.250.250.242 49152 110 6 5 2500 1620140661 1620140721 ACCEPT OK

2 123456789012 eni-2d2e2f3g 192.168.2.7 77.88.55.80 49153 993 6 7 3500 1620140661 1620140721 ACCEPT OK

2 123456789012 eni-4h5i6j7k 172.16.0.2 192.0.2.146 49154 143 6 9 4500 1620140661 1620140721 ACCEPT OK

For e.g. the lookup table file can be something like:

dstport,protocol,tag

25,tcp,sv_P1

68,udp,sv_P2

23,tcp,sv_P1

31,udp,SV_P3

443,tcp,sv_P2

22,tcp,sv_P4

3389,tcp,sv_P5

0,icmp,sv_P5

110,tcp,email

993,tcp,email

143,tcp,email

The program should generate an output file containing the following:

- **Count of matches for each tag, sample o/p shown below**

Tag Counts:

Tag,Count

sv_P2,1

sv_P1,2

sv_P4,1

email,3

Untagged,9

- **Count of matches for each port/protocol combination**

Port/Protocol Combination Counts:

Port,Protocol,Count

22,tcp,1

23,tcp,1

25,tcp,1

110,tcp,1

143,tcp,1

443,tcp,1

993,tcp,1

1024,tcp,1

49158,tcp,1

80,tcp,1

Requirement details

- Input file as well as the file containing tag mappings are plain text (ascii) files
- The flow log file size can be up to 10 MB
- The lookup file can have up to 10000 mappings
- The tags can map to more than one port, protocol combinations. for e.g. sv_P1 and sv_P2 in the sample above.
- The matches should be case insensitive

For anything else that is not clear, please make reasonable assumptions and document those in the Readme to be sent with your submission.

Reference for flow logs:
https://docs.aws.amazon.com/vpc/latest/userguide/flow-log-records.html

Submissions

Please upload the submission to any of the collaboration portals like GitHub that you are comfortable sharing and share the link with us. The submission should come with a readme with info on all the assumptions made, for instance, the program only supports default log format, not custom and the only version that is supported is 2.

Also, please include instructions on how to compile/run the program, what tests were done, and any other analysis you may want to share about your code/program.

Please avoid using non-default libraries or packages like Hadoop, spark, pandas etc. The idea is to be able to review and run the program on a local machine without needing to install too many dependencies / packages.