Use postgres database [Linux preferred]

use psql to import, create a database first

here is an article: <https://medium.com/coding-blocks/creating-user-database-and-adding-access-on-postgresql-8bfcd2f4a91e>

Dataset and paper link:

<https://csr.lanl.gov/data/2017.html>

Paper link is given on the website above.

use the following commands in psql:

\connect database\_name;

\copy lanlfeatures FROM 'lanlfeature90days.csv' DELIMITER ',' CSV;

lanlfeature90days

-- csv file

-- host, day, featureno, feature

---- each host has 90 days of data (there are a few that dont, ignore or remove those)

---- feature is event count -- refer to the paper and my code that generates them

-- may need to consolidate

Tasks: find useful features, feature dimension reduction, anomaly detection [unsupervised]

-- PCA and other dimensionality reduction experiments

-- Search for "PYOD" that contains a bunch of anomaly detection algorithm

-- Use pyod to find anomalies

Anomaly

how??

-- create feature per day (all host) output: feature, day

?? how to generate the features, there are thousands of host

Search google how to load db dump in postgres:

Andromal [supevised + unsuprvised]

-- postgres database dump

-- please use only rows with istest=1

-- feature2 for supervised experiments -- count of syscalls in a particular order -- see my python code [try all algorithms available, svm, random forest, etc. with precision recall zscore and things like that]

-- feature1 is for supervised -- sequence classification -- now count events every 1000 calls and crate a 2d sequence per app -- classify the sequence -- lstm

-- Unsupervised -- feature1 or 2 -- can we model the benign apps so that malware can be detected from feature1 or 2 -- unsupervised algorithms or deep learning autoencoders i dont know

-- tweak my code to generate more or modified features if necessary from the field jtrace (the field may contain partial system call names, ignore those)