DATA CENTER SERVER OWNERSHIP TRACIKING

EGECAN ÇELİK EVGİN
CS540 PROJECT PRESENTATION
OZYEGIN UNIVERSITY

CONTENT

PROBLEM

In large data centers, server lending or team division often leads to confusion about which team is using a server.

In cases of system failures or cyberattacks, it is difficult to identify the correct team and wait for their approval before taking action, especially during the night.

SOLUTION

Objective 1

Collecting useful and meaningful data from servers.

Objective 2

Annotating and processing the data correctly

Objective 3

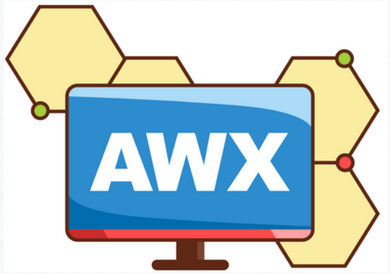
Training machine learning models and deploying them for use.

DATA COLLECTION

Ansible

- Writing an Ansible Playbook to pull virtualization environment, running service name, environment, role.
- Combining these with current data that administrators hold.
- Manually making SSH connection to thousands of servers and verifying the correctness of all information.

AWX



- Running all the commands on every server that AWX can reach in the data center.
- If they are unreachable or not allowed, manually taking the useful information.

FEATURE ENGINEERING

- Working on IP Address column to find underlying patterns on each application team groupings.

- Trying possible combination of features such as: IP Address, Name, Role, Environment, Service, Console, Cluster, OSDistro, PHY/ VIRT, Department.

MACHINE LEARNING

Used models such as Decision Tree, Random Forest, LightGBM, XGBoost, LSTM, MLP, Logistic Regression

Best Performing Model was Random Forest with 92.7% Accuracy, 90.5% Precision and 84% Recall.

STEP I

Used Graphviz to see how the Trees were formed, what were the decision functions.

STEP II

Tried SHAP and Hyperpamater Optimizations which took tens of hours.

Lastly the model was deployed using Pickle, FastAPI, HTML, CSS.

STEP III

RESULTS

With this, now over 90% of the unknown servers can be attached to the correct team and wrongly labeled servers can be detected.

ADVANCEMENT

This showed IP manipulation is key to detect server ownerships. Also Location and Server Name is so important.
All the detection can be done in few seconds now.

REALIZATION

In the second phase, there can be other information such as all the running processes and everybody who has made an SSH connection to the servers.

NEXT PHASE

THANK YOU FOR WATCHING

References:

- *https://learningnetwork.cisco.com/s/question/0D53i00000KspTcCAJ/ip-addressing-scheme-for-data-center*
- *https://research.google/pubs/machine-learning-applications-for-data-center-optimization/*
- *https://medium.com/@bhattacharyya.shilpi.sbu/data-centers-ai-and-infrastructure-b973a0a57e46*
- *https://medium.com/@mike.anderson007/optimizing-modern-data-centers-a-comprehensive-overview-of-datacenter-infrastructure-management-408df25b4b1f*
- *https://medium.com/@ismaelbouarfa/the-future-of-data-centers-innovations-sustainability-and-security-bd6596bdf929*
- *https://ieeexplore.ieee.org/document/8215725*
- *https://www.techrxiv.org/users/789959/articles/1053538-ransomware-detection-on-linux-using-machine-learning-with-random-forest-algorithm?__cf_chl_tk=HkeUHz1CB4ZnFqnNmJUZKBO2OtoKc1bgWTDtlOk4lkM-1734282337-1.0.1.1-RsTSMdLQfcGORnLRmjtLYYvlPUxRRylyhLHQ9psqvig*