Batch Anomaly Detection

Lightning Talks Session 2

Eya ABID

Supervisors: Martin ADAM & Jaroslava Schovancova

15 / 09 / 2022
**Topic**

**Problem**
HTCondor batch system monitoring data is too “big” to be monitored to catch anomalies the traditional way.

**Solution**
Collectd metrics + HTC job data to spot the options for anomaly detection
Overview

Progress

- MONIT’s raw historical data was a challenge to deal with
- Data is collectable, clean, consumable
- Some anomaly detection techniques are applied on the metrics data
Steps

Data collection and manipulation

_Pyspark_ within the _SWAN_ services

Data mining

ML techniques are used to get a better view and understanding of the Data,
Job Data

Procedure to create time-series from JobEvent data failed
HW Data

- Network, CPU, Memory, and various metrics usage across the different collectd plugins for each hostgroup

- We tried using the ADMON Python API, but it was restricting our data manipulation process, we got back to manual processing.
Data Samples
Algorithms and results

*IsolationForest*

predicted 17 anomalies in one day for a single hostgroup

*OneClass SVM*

more than 30 anomalies in one day for a single hostgroup
Results
Algorithms and results

*K-means + PCA*

predicted 1 anomaly in one day for a single hostgroup
**Algorithms and results**

**Autoencoders - Tensorflow**

- Since we don’t have a solid ‘labeling’ to our data, we can not establish a good anomaly detection judgment.
- One more problem was the uniformity of the data - overfitting
- But the model was quite able to have a good encoding-decoding accuracy.
Next steps

- Get a solid ground truth for the job data
- Merge the datasets for better anomaly results
- Finetune the algorithms
QUESTIONS?

eyaabid@insat.u-carthage.tn

LinkedIn: https://www.linkedin.com/in/eya-abid/