Large Scale Anomaly Detection in Data Center Logs and Metrics

Rafael Martínez
Gradiant, ICT technology centre in Spain

Since 2008, focused on technological development and knowledge transfer to industry

- +100 professionals
- 5,2M€ revenue in 2017
- 54% contracted companies
- 46% competitive public funding
- 12 European projects
Focus on...

Connectivity
communication systems and internet of things for data transmission

Intelligence
infrastructure and algorithms for extracting value from data, converting them into useful and actionable information

Security
protection of data and information systems, and protection of privacy
Focus on… Connectivity · Intelligence · Security

- Biometrics
- MGDA
- HGDA
- eLearning
- Video
- Comms
- eHealth
Large Scale Anomaly Detection in Data Center Logs and Metrics

Problem statement

- Large amounts of data
- Data flowing $24 \times 7 \times 365$
- Logs/events & metrics
- Goal: Keep systems OK
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection

ML to the rescue!
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection

1. Anomaly Detection in Metrics

2. Anomaly Detection in Logs
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection

1. Anomaly Detection in Metrics
2. Anomaly Detection in Logs

Requirements

- Run on **endless data streams**
- Provide results in **real-time**
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection

1. Anomaly Detection in Metrics
   • Discord Discovery

2. Anomaly Detection in Logs
   • LogScore

Requirements

• Run on **endless data streams**
• Provide results in **real-time**
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Metrics

Discord Discovery
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Metrics

Discord Discovery

Improvements on the original batch algorithm

- Heuristics to reduce $O(n^2)$ towards $O(n)$
  - Early abandon
  - Randomization
  - Denoise filtering

![Algorithm Description]

- 4032 points
  - subseq_len: 288
- exec_time: ~19 sec
- exec_time: ~3 sec
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Metrics

Discord Discovery

Improvements on the original batch algorithm

• Heuristics to reduce $O(n^2)$ towards $O(n)$

• Fully streaming operation
  • Supports input chunks of any size
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

LogScore

- **LogScore Engine**
  - .fit()
  - .predict()

- **Cluster patterns**

- **Anomaly score**

Example log entries:
- May 24 00:09:58 prox4 172.26.1.9 TCP_TUNNEL/200 55 - 40%
- May 24 00:10:30 prox4 1495577430.877 128.193.42.51 - 0%
- May 24 00:10:48 prox4 (squid-1): 181057 172.26.1.31 - 0%
- May 24 00:11:30 prox4 (squid-1): 60283 128.21.4.51 - 20%
- May 24 00:11:48 1495577268.046 TCP_MISS/384 378 GET - 20%
- May 24 00:11:48 TCP_REFRESH_UNMODIFIED/200 11446 GET - 0%
- May 24 00:11:48 GET http://r2.abcimg.es/resizer.php - 80%
- May 24 00:11:50 prox4 1495577430.877 128.193.42.51 - 0%
- May 24 00:11:58 prox4 (squid-1): 181057 172.26.1.31 - 0%
- May 24 00:12:00 prox4 1495577430.877 128.193.42.51 - 0%
- May 24 00:12:00 prox4 (squid-1): 181057 172.26.1.31 - 0%
- May 24 00:12:04 HIER_DIRECT/176.34.10.3 text/javac - 0%
- May 24 00:12:09 GET http://load.m.exelator.com/load?antonio.villanueva image/gif 1495577270.157 480 35 - 100%
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

**LogScore**

User bob login from 10.0.0.1
User alice login from 10.0.0.1
User jim login from 10.0.0.2
User Srv Admin login from 10.0.0.3

LogCluster intelligence
frequent words
wildcards
clustering

User *{1,2} login from *{1,1}
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

LogScore

No satisfactory results from single runs

- Too little representative clusters
- Too many outliers
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs
LogScore

\[ \text{spec} = \frac{\text{freq\_words}}{\text{freq\_words} + \text{wildcards}} \]

high spec => more informative patterns
low spec => less outliers (bigger clusters)

\[ \text{spec} = 0.6 \]
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs
LogScore
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

LogScore

```json
{
    "timestamp":"1490583167",
    "stage":"1",
    "rsupport":"25",
    "support":"295",
    "total_lines":"1183",
    "result":{
        "num_clusters":2,
        "clusters":[
            {
                "cluster_id":"0",
                "support":"489",
                "pattern":"May 1 *{1,1} prox3 (squid-1): *{2,2} 172.26.15.36 TCP_MISS/200 *{1,1} CONNECT *{1,1} - *{1,1} -"
            },
            {
                "cluster_id":"1",
                "support":"333",
                "pattern":"May 1 *{1,1} prox3 (squid-1): *{2,2} 172.26.15.36 TCP_CLIENT_REFRESH_MISS/200 *{1,1} GET *{1,1} - *{2,2}"
            }
        ],
        "num_outliers":361,
        "outliers":"lc3_201705-01-prox3_1.out"
    }
}
```
Anomaly Detection

1. Anomaly Detection in Metrics
   • Discord Discovery

2. Anomaly Detection in Logs
   • LogScore

Both algorithms

• Learn *normality* from past data
• Unsupervised operation
Large Scale Anomaly Detection in Data Center Logs and Metrics

Functional architecture
Large Scale Anomaly Detection in Data Center Logs and Metrics

Deployment

GET
POST
DELETE

LOAD BALANCER

GET Running Jobs
Restart Job

WATCHDOG

MAGDA-ENGINE 1

REST API

TASKS
Task1, Task2, ...

... 

MAGDA-ENGINE N

REST API

TASKS
Task1, Task2, ...

Global Status

kubernetes
Large Scale Anomaly Detection in Data Center Logs and Metrics

Tests and results

• 4 months of data (52 GB, 121 machines, 306M log events)

• Not only detect strange events
  • But also provide a summary suitable for human supervision

• Good perceived performance
Large Scale Anomaly Detection in Data Center Logs and Metrics

Conclusions

• High-value solution with hard requirements
• Fully scalable in large Big Data environments
• One company already pushing these technologies into the market
Thank you!

Rafael P. Martínez-Álvarez, PhD
rmartinez@gradiant.org

(+34) 986 120 430 | gradiant@gradiant.org | www.gradiant.org