Log type root cause identification

June 23, 2023

1 Context

The high availability of information systems requires efficient monitoring tools. The rise of connected objects and the increasing complexity of digital environments are questioning the traditional tools of CIOs and technical departments. To meet these challenges, EasyVista is targeting a new generation of AIOps ”Artificial Intelligence Operations Systems” software, formalized by Gartner. AIOps combines AI and Bigdata to automate problem detection and resolution. Considering the volumes of objects and events, the technical and economic challenges are colossal for all the actors involved in high availability. EasyVista seeks to structurally improve problem-solving activities through the assistance of Artificial Intelligence based on causal inference.

2 Problem

In a recent development, EasyVista has introduced a significant advancement called EasyRCA (*Easy* *R*oot *C*ause *A*nalysis) [AEZZ23], which enables the automation of identifying the root causes of collective anomalies [1]. This achievement relies on a summary causal graph [ADG22] [2], representing the normal state of the monitored system, along with observational time series from both the normal and anomalous states [AEZZ23]. However, it is important to note that the detected root cause is limited to the level of the summary causal graph, which serves as a simplified representation of the complex IT system.

Simultaneously, the Syslog (System Log) [Ger09] has gained widespread usage as a means to track the status of IT systems. The detailed log messages contained within the Syslog provide deeper insights into system failures and their underlying causes. This leads us to the following question:

Question: Can the access to the Syslog of a specific subsystem facilitate the identification of significant system events that can be traced back along the causal chain, thereby pinpointing a more specific root cause that empowers experts to make informed decisions in resolving the anomaly? Put simply, our objective is to discover the fine-grained root cause from the Syslog that corresponds to the root cause identified in the summary graph.

3 Data

In this rakaton, we consider a system built upon the Storm ingestion topology (referred to as ”Strom” hereafter). The time series obtained from this system are sampled at a frequency of one minute, and they are labeled as follows:

[1] In time series, a collective anomaly is a sequence of data instances that is anomalous with respect to the entire time series [CBK09].

[2] A summary causal graph establishes direct relationships between variables without considering time [ADG22].

- *PMDB*: represents the extraction of some information about the messages received by the Storm system;
- *MDB*: refers to an activity of a process that orients messages to another process with respect to different types of messages;
- *CMB*: represents the activity of extracting metrics from messages;
- *MB*: represents the activity of inserting data into a database;
- *LMB*: reflects the updates of the latest metric values in Cassandra;
- *RTMB*: represents the activity of searching and merging data with information from the check message bolt;
- *GSIB*: represents the activity of inserting historical status into the database;
- *ESB*: represents the activity of writing data into Elasticsearch.

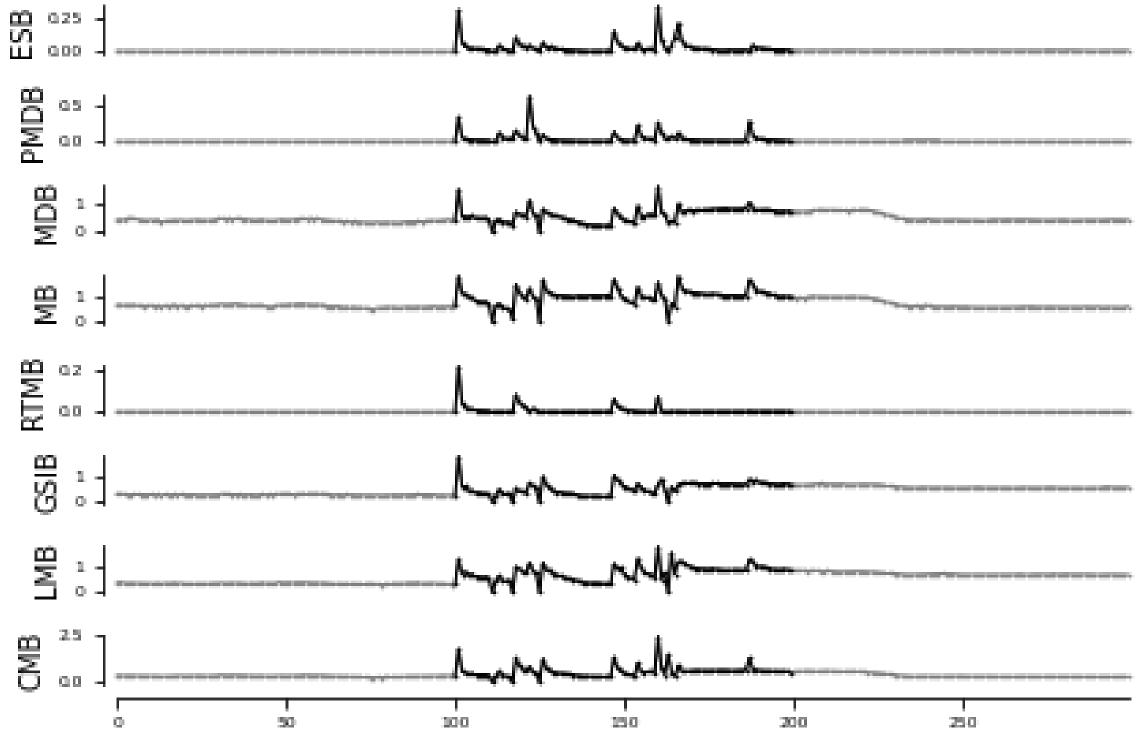


Figure 1: A section of the Storm data is displayed in a plot, where the abnormal period is emphasized with dark black lines. Furthermore, gray lines indicate 100 sampling points preceding and following this abnormal period.

For each time series, anomalies are assumed to occur simultaneously and have a size of 100. Figure 1 displays a plot of these time series, where the abnormal period is represented by dark black lines, and 100 points preceding and following this period are shown as grey lines.

In Figure 2, the summary causal graph associated with the Storm system is depicted, utilizing expert knowledge. In the graph, a self-loop indicates a self-dependency within a time series, while an arrow signifies a causal relationship from the cause to the effect. Upon applying *EasyRCA* to both the abnormal and normal segments of the time series, it is determined that *ESB* emerges as one of the root causes.

The Syslog of ElasticSearch plays a crucial role in capturing and recording various events, errors, and activities occurring within the ElasticSearch cluster. It contains a chronological record of significant

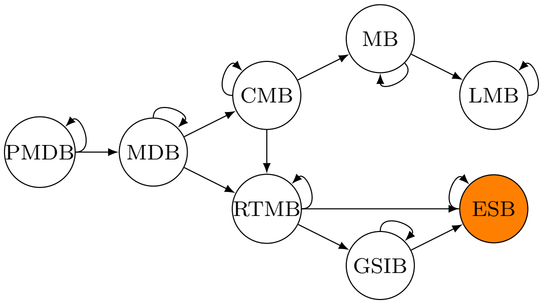
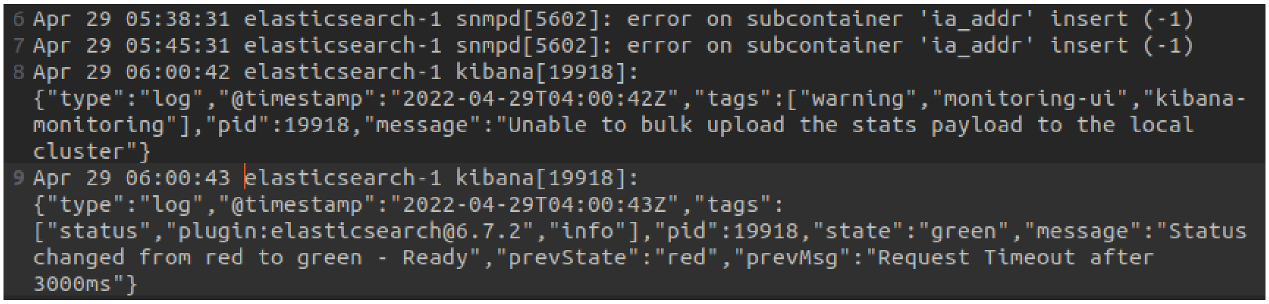


Figure 2: IT monitoring system.

events, including process failures and recoveries, index operations, search queries, and various other diagnostic information. The log entries offer details about exceptions, warnings, and informational messages, aiding in the analysis and resolution of errors.

In this case, we focus on the Syslog of the server named ’elasticsearch-1’ within the ElasticSearch cluster. The Syslog provided begins recording events from 2022-04-28 05:38:28 and continues until 2022-04-29 16:40:09. During this timeframe, the server encountered an incident, and the details of the anomaly period can be cross-referenced using the data provided in Figure 1. Figure 3 presents an example of the Syslog from this server, where each entry can be divided into six parts: **Month**, **Day**, **Time**, **Host**, **ProcessName**, and **Content**, as illustrated in Table 1. Ultimately, the expected form of the root cause(s) to be found from the Syslog is a combination of the **ProcessName** and the relevant **Content**.



```
Apr 29 05:38:31 elasticsearch-1 snmpd[5602]: error on subcontainer 'ia_addr' insert (-1)
Apr 29 05:45:31 elasticsearch-1 snmpd[5602]: error on subcontainer 'ia_addr' insert (-1)
Apr 29 06:00:42 elasticsearch-1 kibana[19918]:
{"type":"log","@timestamp":"2022-04-29T04:00:42Z","tags":["warning","monitoring-ui","kibana-
monitoring"],"pid":19918,"message":"Unable to bulk upload the stats payload to the local
cluster"}
Apr 29 06:00:43 elasticsearch-1 kibana[19918]:
{"type":"log","@timestamp":"2022-04-29T04:00:43Z","tags":
["status","plugin:elasticsearch@6.7.2","info"],"pid":19918,"state":"green","message":"Status
changed from red to green - Ready","prevState":"red","prevMsg":"Request Timeout after
3000ms"}
```

Figure 3: A snapshot of Syslog of Elasticsearch.

Month	Day	Time	Host	ProcessName	Content
Apr	29	05:38:31	elasticsearch-1	snmpd[5602]	error on subcontainer ’ia_addr’ insert (-1)
Apr	29	06:00:42	elasticsearch-1	kibana[19918]	”type”:”log”,”@timestamp”:”2022-04-29”. . .
Apr	29	06:00:43	elasticsearch-1	kibana[19918]	”type”:”log”,”@timestamp”:”2022-04-29. . .

Table 1: Each entry of the log messages is composed of six parts.

4 Expected results

A comprehensive presentation should be prepared, covering the questions outlined in Section 2 and showcasing the developed code. The presentation should include literature-based investigations, a detailed explanation of the proposed method, and the results obtained from the proposed data.

The evaluation of this work will primarily concentrate on its research merits, emphasizing the significance of scientific reasoning, the choices made in selecting solutions, and the interpretation of the results.

Contact:
Lei Zan, lzan@easyvista.com

References

[ADG22] Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

[AEZZ23] Charles K Assaad, Imad Ez-Zejjari, and Lei Zan. Root cause identification for collective anomalies in time series given an acyclic summary causal graph with loops. In *International Conference on Artificial Intelligence and Statistics*, pages 8395–8404. PMLR, 2023.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), jul 2009.

[Ger09] Rainer Gerhards. The syslog protocol. Technical report, 2009.