

# Using Transformers for Log Supervision

Supervised by Aristeidis Fkiaras

A quick peek over the project I was assigned from mid-June to mid-July 2023

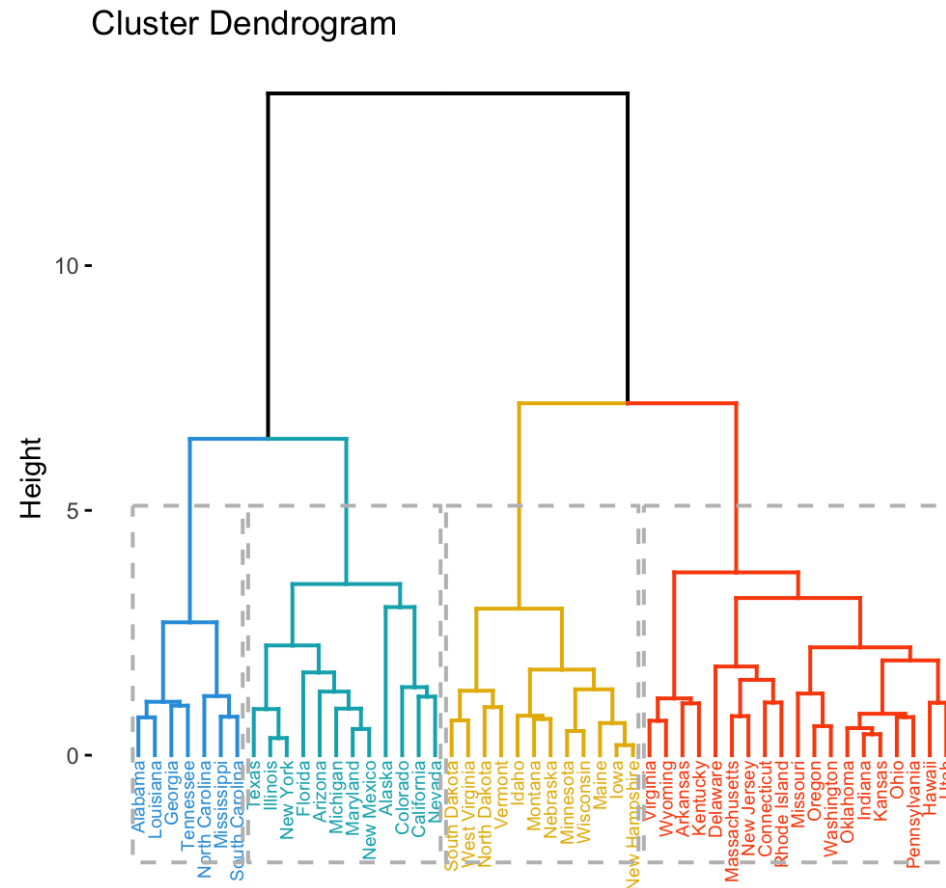
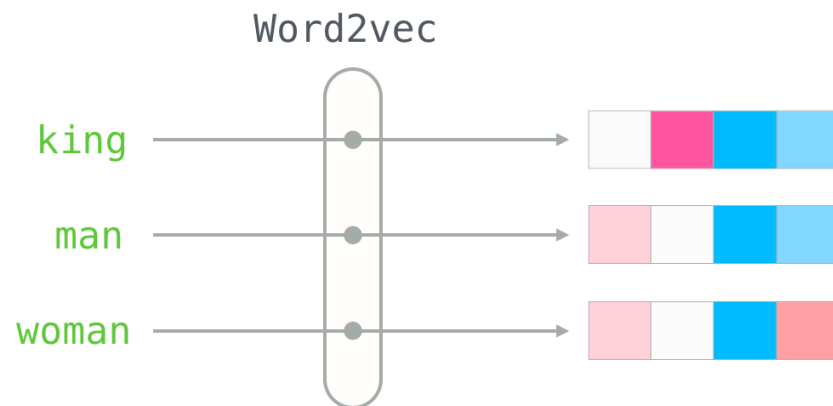
Mathis Lindner

# Problems at hand

- Supervising system logs can be tedious
  - Even with good filters
  - Similar messages but not exactly the same
  - Similar messages but on different hosts
- Cannot store logs forever
  - Adding an entry to the logbook instead of storing all the logs.

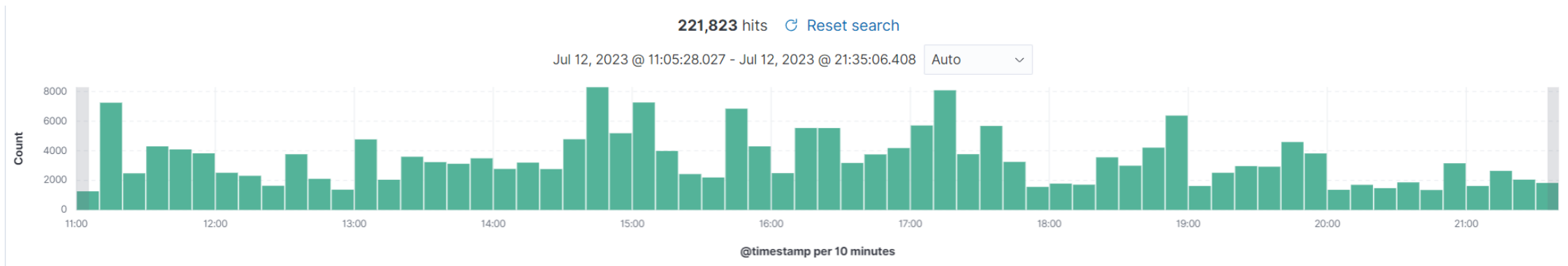
# Log Reduction

- Sentence transformer to vectorise the logs
- Hierarchical clustering to group similar logs together.
- Send them back to opensearch as a new index each hour.

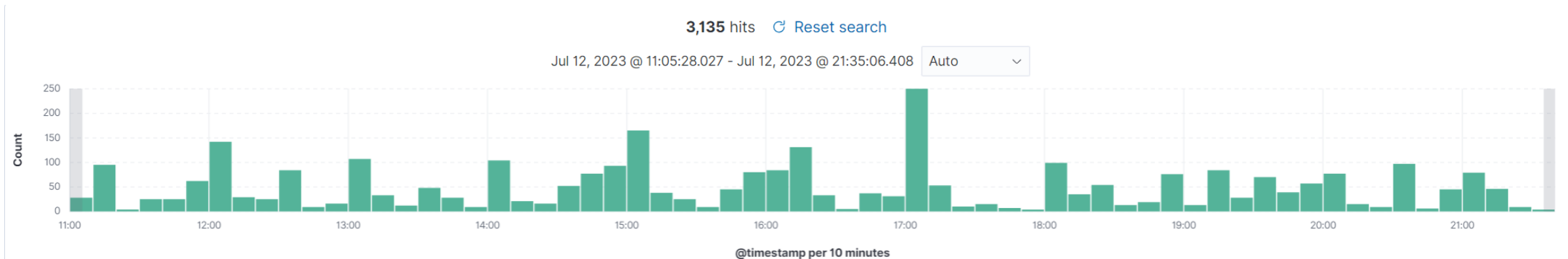


# Results on sytem logs errors and warnings

## Before



## After



host	message	n_unique_hosts	n_similar_messages	@timestamp	syslog_severity
n2173301	igb 0000:06:00.0 eno1: Reset adapter	1	2	2023-07-12 20:51:09.170000+00:00	error
n2173301	rcu: 4-....: (1 GPs behind) idle=479/1/0x4000000000000000 softirq=15122273/15122274 fqs=455	1	3	2023-07-12 20:42:40.183000+00:00	error
n2173301	rcu: INFO: rcu_preempt self-detected stall on CPU	1	7	2023-07-12 20:42:40.178000+00:00	error
n2081302	Error receiving audit netlink packet (No buffer space available)	1	2	2023-07-12 16:20:17.934000+00:00	error
crguest14	Failed to enqueue OnFailure= job: Resource deadlock avoided	1	2	2023-07-12 16:03:24.044000+00:00	error
megtestq1n1	usb 2-1-port1: over-current condition	1	2	2023-07-12 15:50:40.459000+00:00	error
diri01	hid-generic 0003:03F0:034A.006C: usb_submit_urb(ctrl) failed: -19	1	2	2023-07-12 15:24:31.224000+00:00	error
n2060301	Exiting the SSSD. Could not restart critical service [nss].	1	7	2023-07-12 15:04:01.941000+00:00	error

# Logbook Reports

- Vectorise the hourly logs (some of them are repeated each hour)
- Cluster them as well
- Ask to get a summary from a Large Language Model to send it to the LHCb logbook.
- We tried a few open-source models (Falcon 40b, Llama 65b) and ran them locally for privacy/independency reasons, but the results were not satisfying.
  - Only a matter of time until open-source models can do that too.
- So for demonstration purposes we used the OpenAI's GPT-3.5 for the following...

## System Log Report - July 12, 2023

### Morning (00:00 - 08:59):

The day began with several errors logged across different hosts. At midnight, the server n2201703 experienced a failure in the systemd-coredump service, resulting in a failed step and a read-only file system error. Another server, cronl03, reported a BIOS error related to the iwlwifi device.

In the early hours, the server lab14 encountered a forbidden error while checking for jobs, possibly due to access restrictions for the runner 'mv72UWrp.' Additionally, the server uceb13 reported a device detection error, suggesting the need for bus re-enumeration.

On cradm01, multiple issues occurred simultaneously. The process python2.7 was killed by an unsupported signal, the bluez5-util.c failed to retrieve ManagedObjects, and vdagent virtio channel access was denied...

... Around 8:40 AM, uceb10 experienced an error related to network interface renaming, while crot02 failed to start the LSB service responsible for networking. A problem with the specified group 'plugdev' was also recorded on crot02.

### **Late Morning (09:00 - 11:59):**

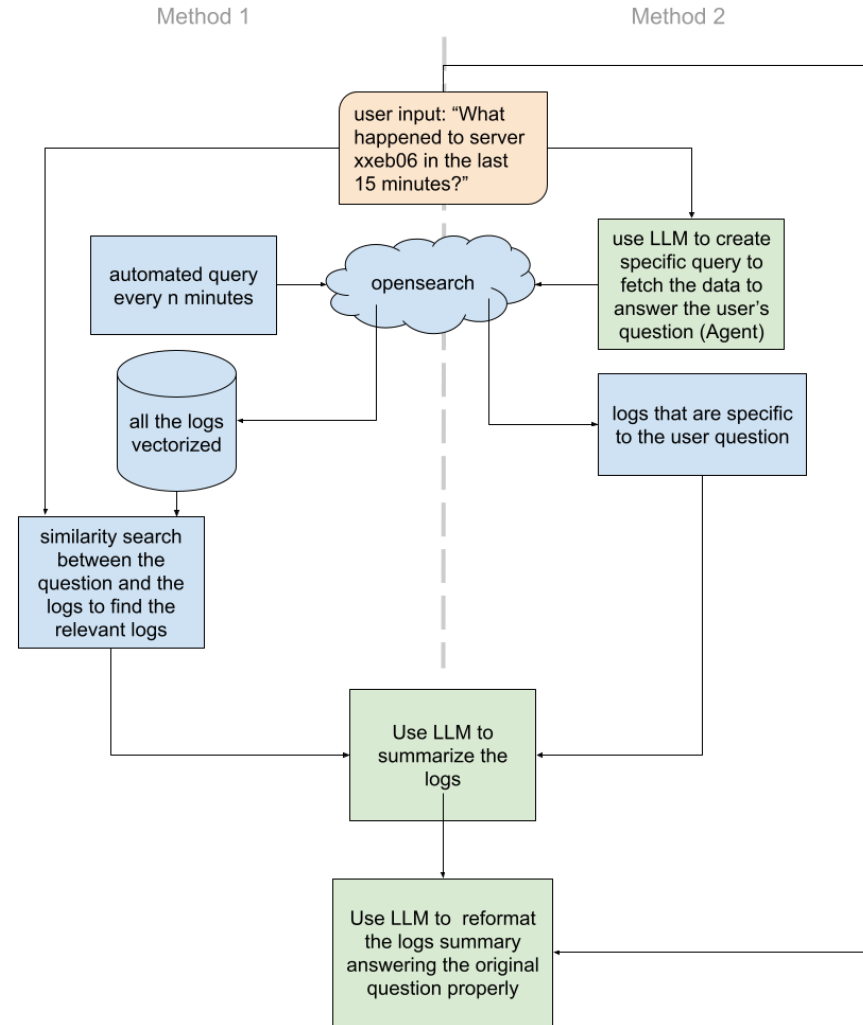
The server n6093103 encountered a connection refusal while attempting to send a WATCHDOG=1 notification message. Daqv0207 experienced a timeout while waiting for the primary device, and n8190704 reported a crash that was not saved in the specified file.

Meanwhile, loadbalancer-pluscc faced issues with the mailer process when attempting to mail the output of the `cron.daily` job. These errors might have affected the regular operation of the system.



# Future Work

- Idea of chatbot answering questions specific to the logs
- **Issue:** Most Large Language Models do not support long inputs
- **Idea:** Only give it access the logs relevant to the question using cosimilarity search on their embeddings.
- **Need:** To train a good sentence transformer adapted for logs. (Most models are trained on forums news and other text sources and do not perform well on logs)
- **Need:** To find a way to deal with time related questions with method 1.



side notes:

issue with a similarity search for logs: if you want to query time related, trouble finding the range, compared to an agent that looks through structured data.

method 3 would be using langchain LLM to query from the downloaded logs in a dataframe

# Quick Demo of method 2

- Using [Langchain module](#)
  - Prompt LLM on how it can access information.
  - Tell the LLM that it needs to use a specific thought process.
  - Add the question after this prompt.

## Sidenote

- Unfortunately there are limitations on complexity of the question.
- Models we tried from [huggingface](#) are not performing well on other tasks than conversation yet.
- The most complicated things you can ask is a simple:
  - "what happened to host xyz"
  - can find logs specific to the question and summarize them a bit.

**Question:** What is the last log entry?

**Thought:** I should use the `sort_values` function to sort the dataframe by timestamp.

**Action:** `python_repl_ast`

**Action Input:** `df.sort_values(by='@timestamp')`

Observation:

host		message	@timestamp
0	vedcsvss01	Failed to start backup_pvssInst.service.	2023-06-28 00:00:34.726000+00:00
1	crot03	iwlfwif 0000:00:14.3: BIOS contains WGDS but n...	2023-06-28 00:00:39.239000+00:00
2	n4061401	ERROR: Checking for jobs... forbidden ...	2023-06-28 00:00:55.171000+00:00
3	n4061401	ERROR: Runner https://gitlab.cern.ch/zs...	2023-06-28 00:00:55.923000+00:00
4	n4061401	ERROR: Checking for jobs... forbidden ...	2023-06-28 00:01:55.200000+00:00
...	...	...	...
9995	n2232702	libceph: auth protocol 'cephx' mauth authentic...	2023-06-28 12:14:22.365000+00:00
9996	n2232702	libceph: auth protocol 'cephx' mauth authentic...	2023-06-28 12:14:22.388000+00:00
9997	n2232702	libceph: auth protocol 'cephx' mauth authentic...	2023-06-28 12:14:25.460000+00:00
9998	n2232702	libceph: auth protocol 'cephx' mauth authentic...	2023-06-28 12:14:32.605000+00:00
9999	n2232702	libceph: auth protocol 'cephx' mauth authentic...	2023-06-28 12:14:32.605000+00:00

[10000 rows x 3 columns]

**Thought:** I should use the `iloc` function to get the last row

**Action:** `python_repl_ast`

**Action Input:** `df.sort_values(by='@timestamp').iloc[-1]`

**Observation:**

```
host                                     n2232702
message      libceph: auth protocol 'cephx' mauth authentic...
@timestamp      2023-06-28 12:14:32.605000+00:00
Name: 9999, dtype: object
```

**Thought:** I now know the final answer

**Final Answer:** `n2232702`

**Thank you for your attention**

More slides

