ZEBRIUM

Machine Learning for k8s Logs and Metrics

AUTOMATING INCIDENT AND ROOT-CAUSE DETECTION

Larry Lancaster
Founder and CTO
Zebrium

Machine data is my life

- NetApp - *Engineering Informatics*
- EMC / Data Domain - *Product Analytics*
- Glassbeam - *Chief Technology Officer*
- Nimble Storage - *Chief Data Scientist*
- Zebrium - *Founder and CTO*

ZEBRIUM

20 YEARS AGO

Shrink-Wrap:

1 incident 1 user

1 incident 1 monolith

1 incident 10 logfiles

Log use for root-cause:

index and search

20 YEARS AGO

Shrink-Wrap:

1 incident 1 user

1 incident 1 monolith

1 incident 10 logfiles

Log use for root-cause:

index and search

TODAY

SaaS:

1 incident 100K users

1 incident 100 services

1 incident 1K logstreams

Log use for root-cause:

still index and search(!)

Complexity drives MTTR

> "The proportion of medium performers is up. Some are likely improved low performers, while others may be high performers who dropped as **they struggled with increased complexity.**"

Source: State of DevOps (2019)

Automation can't fix it

> "TIME TO RESTORE SERVICE PERFORMANCE STAYED THE SAME
> FOR BOTH ELITE AND LOW PERFORMERS WHEN COMPARED TO THE PREVIOUS YEAR."

Source: State of DevOps (2019)

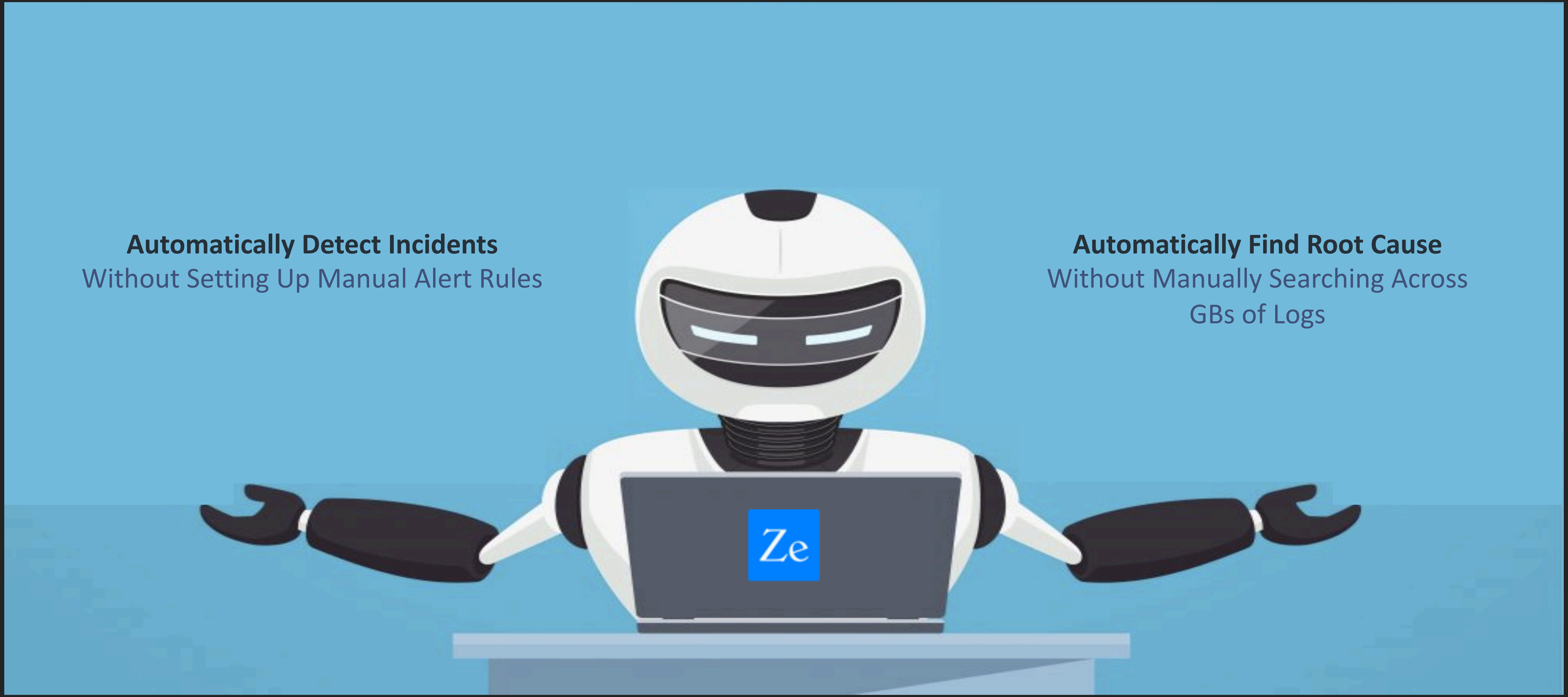
Ze

ZEBRIUM

Our vision

Autonomous RCA will save the world from the cost of complexity.

What I want from a tool



Automatically Detect Incidents
Without Setting Up Manual Alert Rules

Automatically Find Root Cause
Without Manually Searching Across GBs of Logs

My requirements

- Arbitrary application
- Arbitrary runtime
- Arbitrary infrastructure
- Arbitrary environment

- Zero required tracing
- Zero required training
- Zero required alert rules

Is it really too much to ask? :)

Why so harsh?

Because complexity also means:

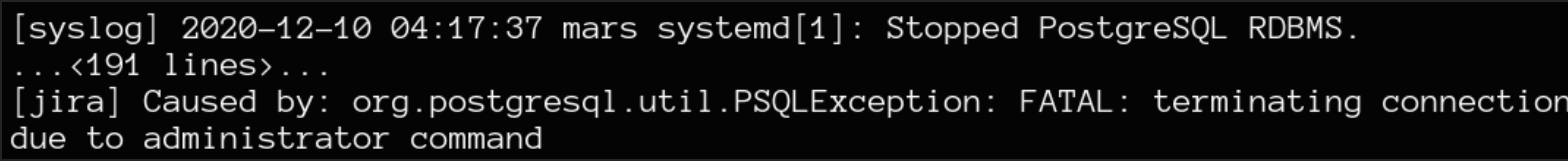
Manual inputs may not scale
Stack assumptions may not hold

Outrageous opinion

Autonomous RCA has to start with logs.

Logs are self-describing

A free-text log tells a story:



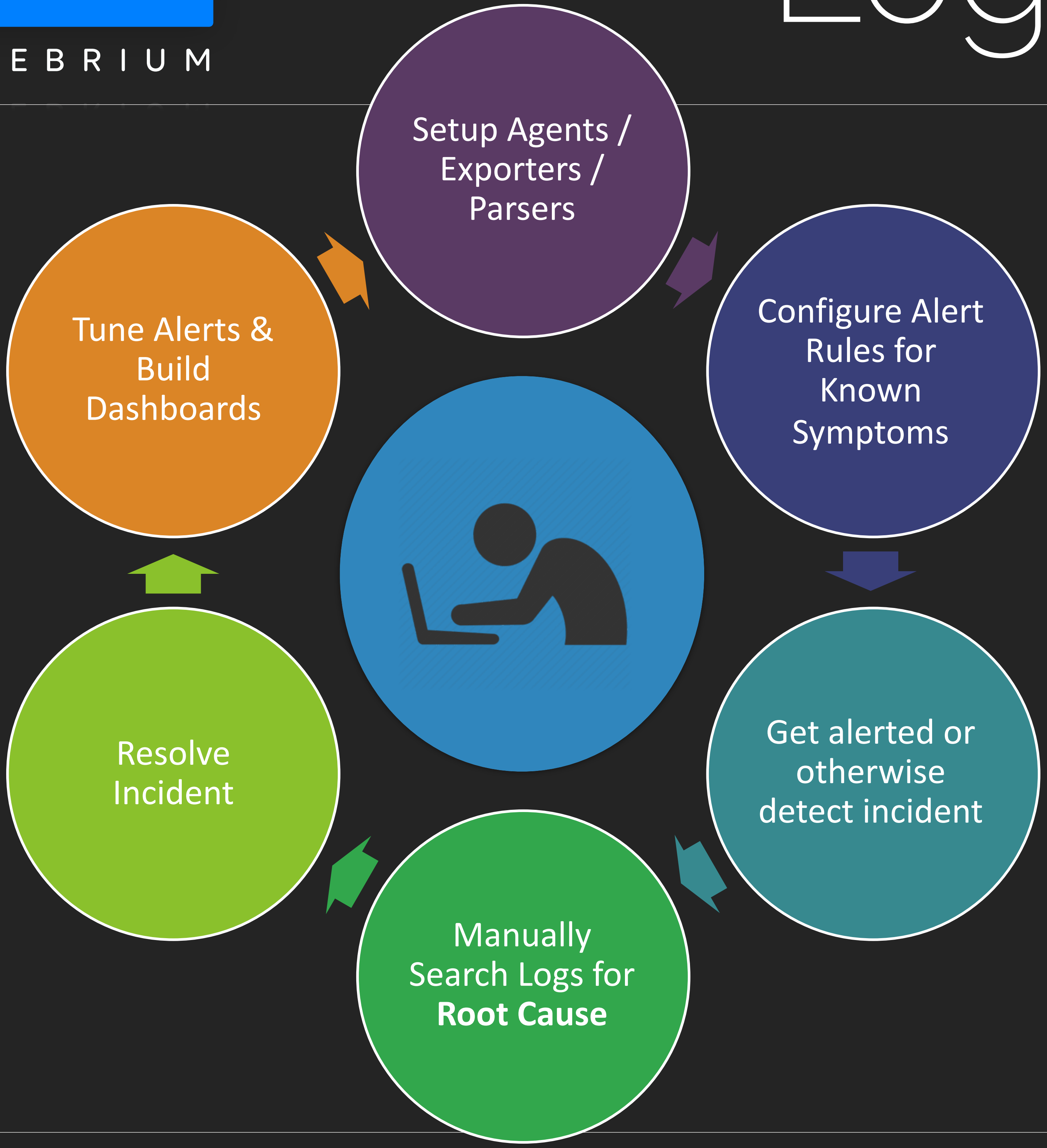
```
[syslog] 2020-12-10 04:17:37 mars systemd[1]: Stopped PostgreSQL RDBMS.
...<191 lines>...
[jira] Caused by: org.postgresql.util.PSQLException: FATAL: terminating connection
due to administrator command
```

People use logs for RCA

...so why aren't they better at helping us monitor?

ZEBRIUM

Log monitoring today


Setup Agents / Exporters / Parsers
Configure Alert Rules for Known Symptoms
Get alerted or otherwise detect incident
Manually Search Logs for **Root Cause**
Resolve Incident
Tune Alerts & Build Dashboards

SLOW (MTTR)

FRAGILE (FORMATS CHG)

ANNOYING (ALERT FATIGUE)

HUMAN-DRIVEN

What keeps logs "dumb"?

Logs are stuck in "index + search"

Why are logs so hard?

Formats change

Parses are ambiguous

Experts are needed to interpret

Apps are bespoke

The junior SRE problem

"Hey, I hadn't seen that happen before...
then everything went sideways!"

--

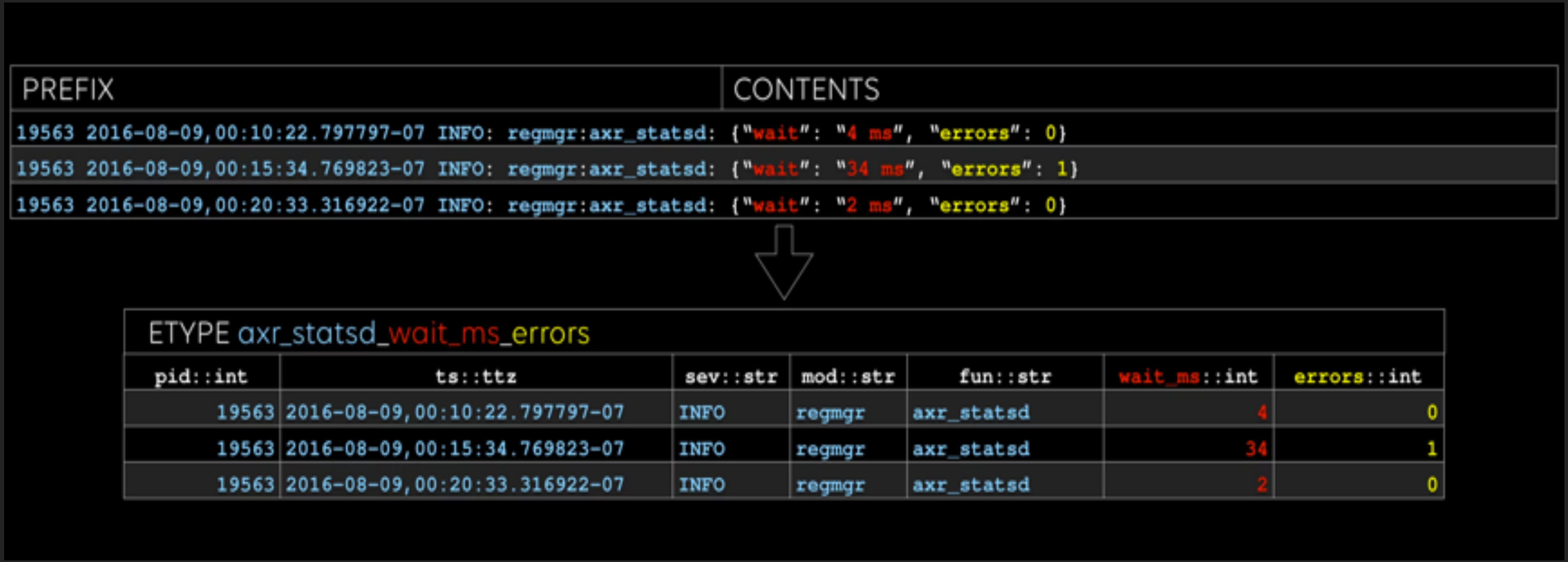
Figure out when rare stuff and bad stuff
are unusually correlated.

Ze: How it works

Complete relational structuring of logs

ZEBRIUM

Ze: How it works



PREFIX	CONTENTS
19563 2016-08-09,00:10:22.797797-07 INFO: regmgr:axr_statsd:	{"wait": "4 ms", "errors": 0}
19563 2016-08-09,00:15:34.769823-07 INFO: regmgr:axr_statsd:	{"wait": "34 ms", "errors": 1}
19563 2016-08-09,00:20:33.316922-07 INFO: regmgr:axr_statsd:	{"wait": "2 ms", "errors": 0}

ETYPE axr_statsd_wait_ms_errors

pid::int	ts::ttz	sev::str	mod::str	fun::str	wait_ms::int	errors::int
19563	2016-08-09,00:10:22.797797-07	INFO	regmgr	axr_statsd	4	0
19563	2016-08-09,00:15:34.769823-07	INFO	regmgr	axr_statsd	34	1
19563	2016-08-09,00:20:33.316922-07	INFO	regmgr	axr_statsd	2	0

Ze: How it works

No information included or required about:

- Known prefix formats
- Specific logtype keywords
- Event grammar / syntax

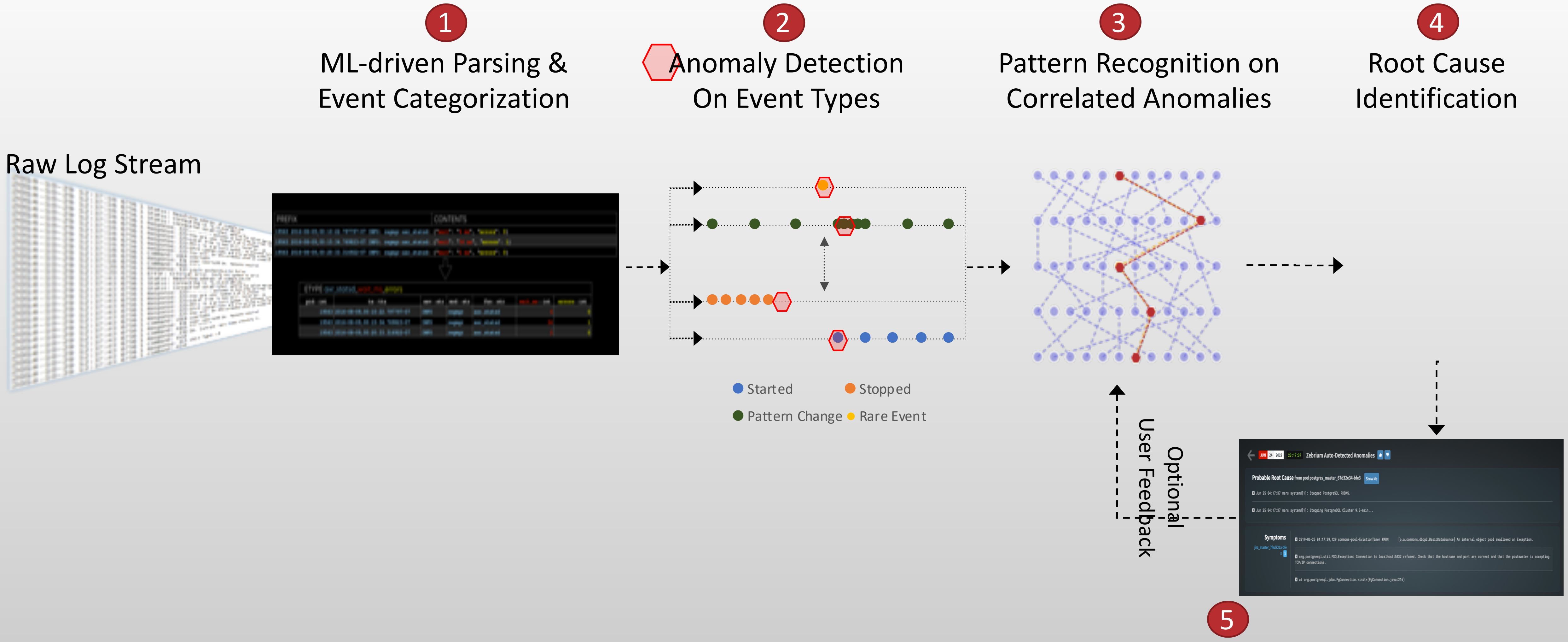
We embrace free-text logs

Ze
ZEBRIUM

Ze: How it works

Anomaly detection on relationally structured data

Ze: How it works


1
ML-driven Parsing & Event Categorization
2
Anomaly Detection On Event Types
3
Pattern Recognition on Correlated Anomalies
4
Root Cause Identification
Raw Log Stream
Started
Stopped
Pattern Change
Rare Event
Optional User Feedback
5

JUN 24 2019 23:17:37 Zebrium Auto-Detected Anomalies

Probable Root Cause from pod postgres_master_67d32e34-bfe3 Show Me

```
Jun 25 04:17:37 mars systemd[1]: Stopped PostgreSQL RDBMS.
Jun 25 04:17:37 mars systemd[1]: Stopping PostgreSQL Cluster 9.5-main...
```

Symptoms

jira_master_7fed321a-bfe3 3

```
2019-06-25 04:17:39,129 commons-pool-EvictionTimer WARN     [o.a.commons.dbcp2.BasicDataSource] An internal object pool swallowed an Exception.
org.postgresql.util.PSQLException: Connection to localhost:5432 refused. Check that the hostname and port are correct and that the postmaster is accepting TCP/IP connections.
at org.postgresql.jdbc.PgConnection.<init>(PgConnection.java:216)
```

Ze: How it works

No information included or required about:

- Connectors, knowledge bases
- Specific application behaviors
- Specific semantic keywords

Works great on bespoke app or stack

Other ML attempts

Use deep learning

Use one algorithm

Work in batch

Use a Swiss army knife

Structure First - Inline

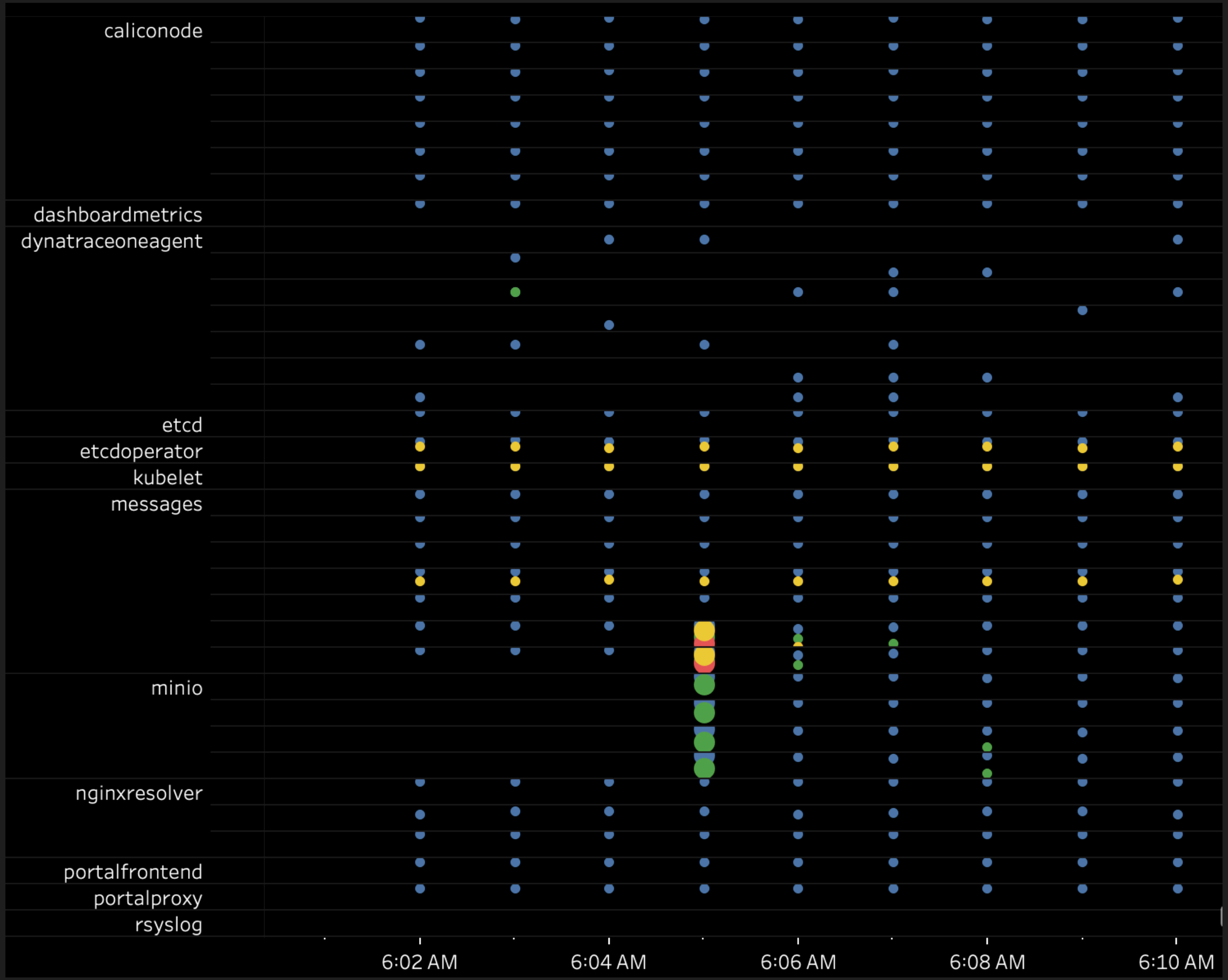
Respect Pareto - multi-stage

AD/RCA gets better w / complexity!!!

GPT3/NLP requires concise RCA

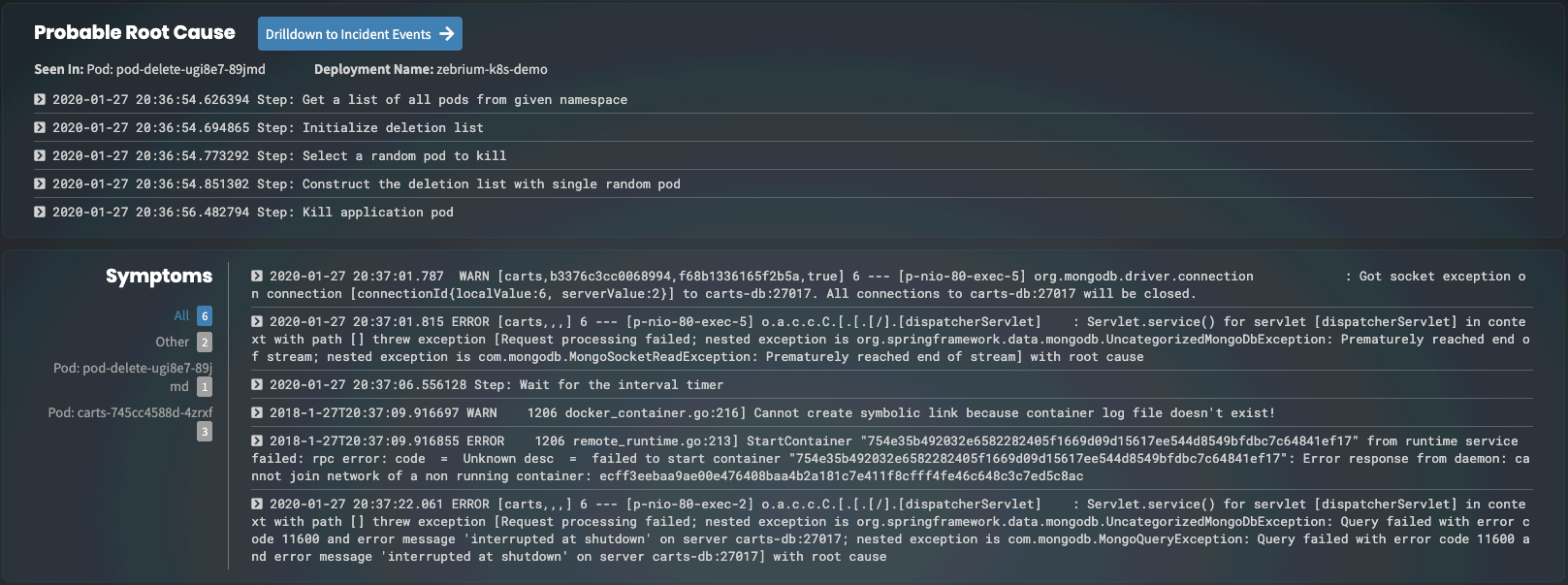
A picture is worth...

An Incident


caliconode
dashboardmetrics
dynatraceoneagent
etcd
etcdoperator
kubelet
messages
minio
nginxresolver
portalfrontend
portalproxy
rsyslog
6:02 AM
6:04 AM
6:06 AM
6:08 AM
6:10 AM

ZEBRIUM

Autonomous Monitoring



Probable Root Cause

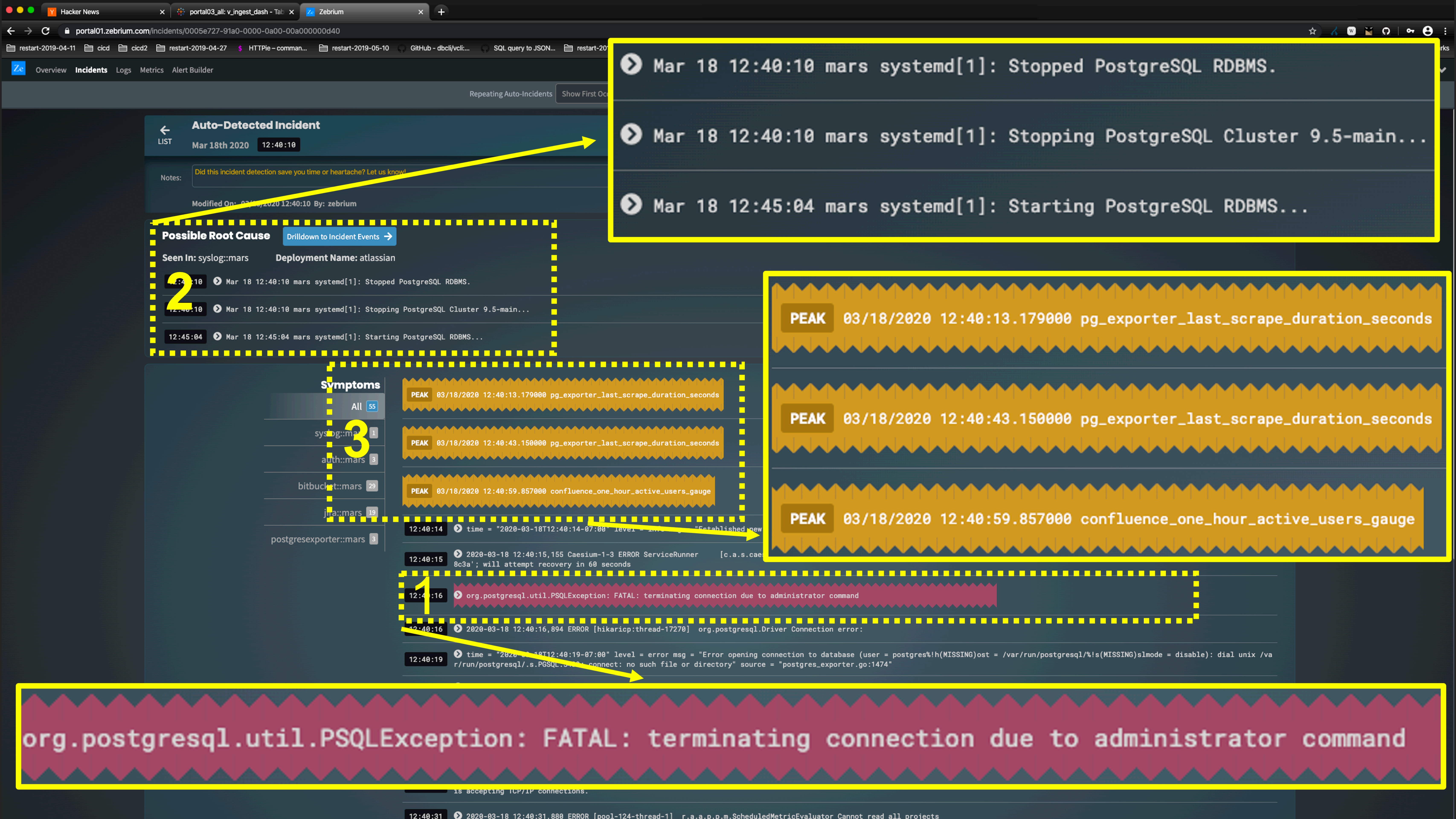
Seen In: Pod: pod-delete-ugi8e7-89jmd **Deployment Name:** zebrium-k8s-demo

```
2020-01-27 20:36:54.626394 Step: Get a list of all pods from given namespace
2020-01-27 20:36:54.694865 Step: Initialize deletion list
2020-01-27 20:36:54.773292 Step: Select a random pod to kill
2020-01-27 20:36:54.851302 Step: Construct the deletion list with single random pod
2020-01-27 20:36:56.482794 Step: Kill application pod
```

Symptoms

All 6
Other 2
Pod: pod-delete-ugi8e7-89jmd 1
Pod: carts-745cc4588d-4zrxf 3

```
2020-01-27 20:37:01.787  WARN [carts,b3376c3cc0068994,f68b1336165f2b5a,true] 6 --- [p-nio-80-exec-5] org.mongodb.driver.connection            : Got socket exception o
n connection [connectionId{localValue:6, serverValue:2}] to carts-db:27017. All connections to carts-db:27017 will be closed.

2020-01-27 20:37:01.815 ERROR [carts,,,] 6 --- [p-nio-80-exec-5] o.a.c.c.C.[.[.[/].[dispatcherServlet]    : Servlet.service() for servlet [dispatcherServlet] in conte
xt with path [] threw exception [Request processing failed; nested exception is org.springframework.data.mongodb.UncategorizedMongoDbException: Prematurely reached end o
f stream; nested exception is com.mongodb.MongoSocketReadException: Prematurely reached end of stream] with root cause

2020-01-27 20:37:06.556128 Step: Wait for the interval timer

2018-1-27T20:37:09.916697 WARN    1206 docker_container.go:216] Cannot create symbolic link because container log file doesn't exist!

2018-1-27T20:37:09.916855 ERROR    1206 remote_runtime.go:213] StartContainer "754e35b492032e6582282405f1669d09d15617ee544d8549bfdbc7c64841ef17" from runtime service
failed: rpc error: code  =  Unknown desc  =  failed to start container "754e35b492032e6582282405f1669d09d15617ee544d8549bfdbc7c64841ef17": Error response from daemon: ca
nnot join network of a non running container: ecff3eebaa9ae00e476408baa4b2a181c7e411f8cfff4fe46c648c3c7ed5c8ac

2020-01-27 20:37:22.061 ERROR [carts,,,] 6 --- [p-nio-80-exec-2] o.a.c.c.C.[.[.[/].[dispatcherServlet]    : Servlet.service() for servlet [dispatcherServlet] in conte
xt with path [] threw exception [Request processing failed; nested exception is org.springframework.data.mongodb.UncategorizedMongoDbException: Query failed with error c
ode 11600 and error message 'interrupted at shutdown' on server carts-db:27017; nested exception is com.mongodb.MongoQueryException: Query failed with error code 11600 a
nd error message 'interrupted at shutdown' on server carts-db:27017] with root cause
```



Auto-Detected Incident

Mar 18th 2020 12:40:10

Notes: Did this incident detection save you time or heartache? Let us know!

Modified On: 03/18/2020 12:40:10 By: zebrium

Possible Root Cause

Seen In: syslog::mars **Deployment Name:** atlassian

```
12:40:10  Mar 18 12:40:10 mars systemd[1]: Stopped PostgreSQL RDBMS.
12:40:10  Mar 18 12:40:10 mars systemd[1]: Stopping PostgreSQL Cluster 9.5-main...
12:45:04  Mar 18 12:45:04 mars systemd[1]: Starting PostgreSQL RDBMS...
```

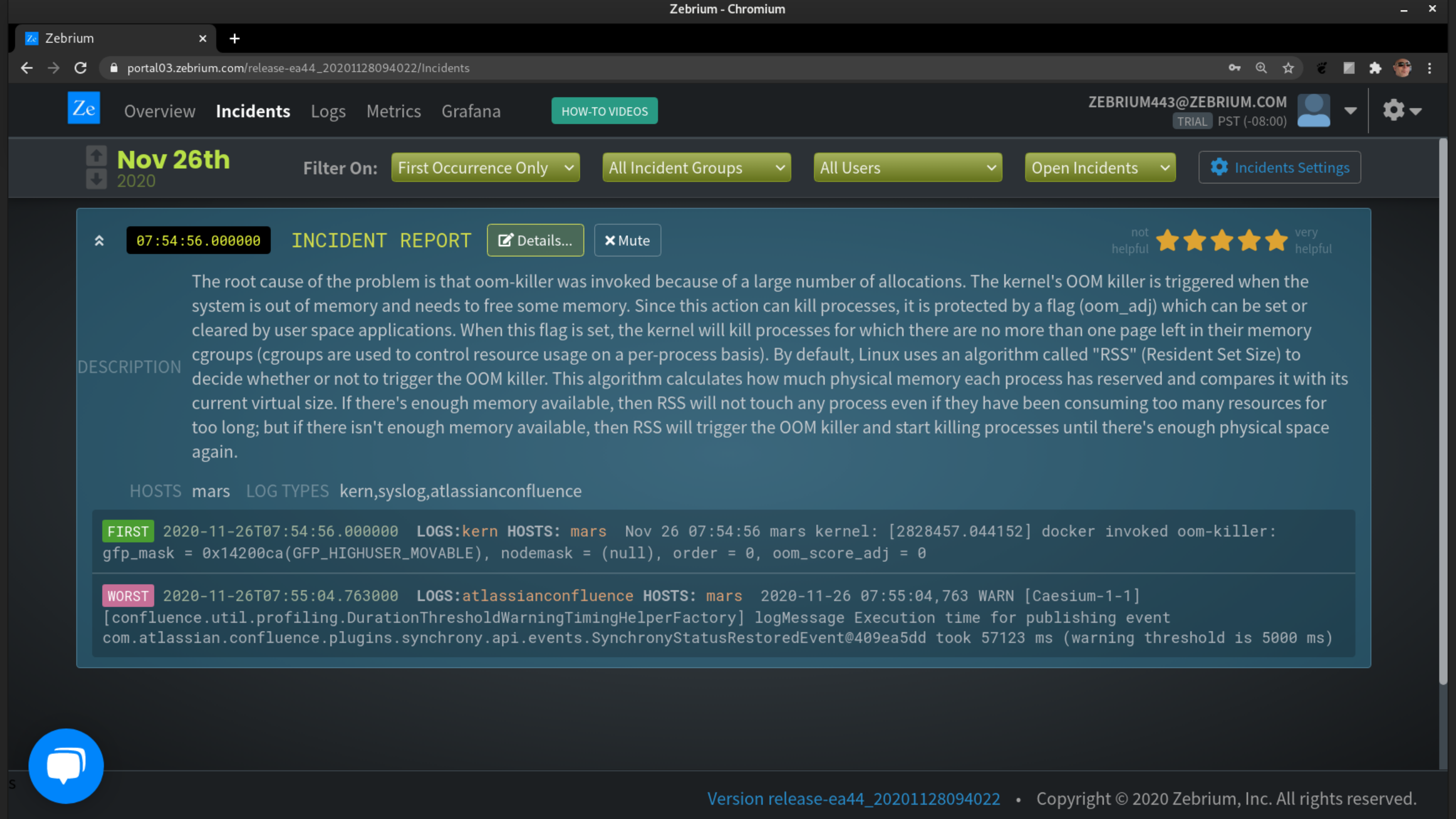
```
PEAK 03/18/2020 12:40:13.179000 pg_exporter_last_scrape_duration_seconds
PEAK 03/18/2020 12:40:43.150000 pg_exporter_last_scrape_duration_seconds
PEAK 03/18/2020 12:40:59.857000 confluence_one_hour_active_users_gauge
```

```
12:40:14  time = "2020-03-18T12:40:14-07:00" level = info ... "Established new
12:40:15  2020-03-18 12:40:15,155 Caesium-1-3 ERROR ServiceRunner     [c.a.s.cae
          8c3a'; will attempt recovery in 60 seconds
12:40:16  org.postgresql.util.PSQLException: FATAL: terminating connection due to administrator command
12:40:16  2020-03-18 12:40:16,894 ERROR [hikaricp:thread-17270]  org.postgresql.Driver Connection error:
12:40:19  time = "2020-03-18T12:40:19-07:00" level = error msg = "Error opening connection to database (user = postgres%!h(MISSING)ost = /var/run/postgresql/%!s(MISSING)slmode = disable): dial unix /va
          r/run/postgresql/.s.PGSQL.5432: connect: no such file or directory" source = "postgres_exporter.go:1474"
          is accepting TCP/IP connections.
12:40:31  2020-03-18 12:40:31,880 ERROR [pool-124-thread-1]  r.a.a.p.p.m.ScheduledMetricEvaluator Cannot read all projects
```

```
Mar 18 12:40:10 mars systemd[1]: Stopped PostgreSQL RDBMS.
Mar 18 12:40:10 mars systemd[1]: Stopping PostgreSQL Cluster 9.5-main...
Mar 18 12:45:04 mars systemd[1]: Starting PostgreSQL RDBMS...
```

```
PEAK 03/18/2020 12:40:13.179000 pg_exporter_last_scrape_duration_seconds
PEAK 03/18/2020 12:40:43.150000 pg_exporter_last_scrape_duration_seconds
PEAK 03/18/2020 12:40:59.857000 confluence_one_hour_active_users_gauge
```

```
org.postgresql.util.PSQLException: FATAL: terminating connection due to administrator command
```



Filter On: First Occurrence Only | All Incident Groups | All Users | Open Incidents

07:54:56.000000 INCIDENT REPORT

★★★★★

DESCRIPTION

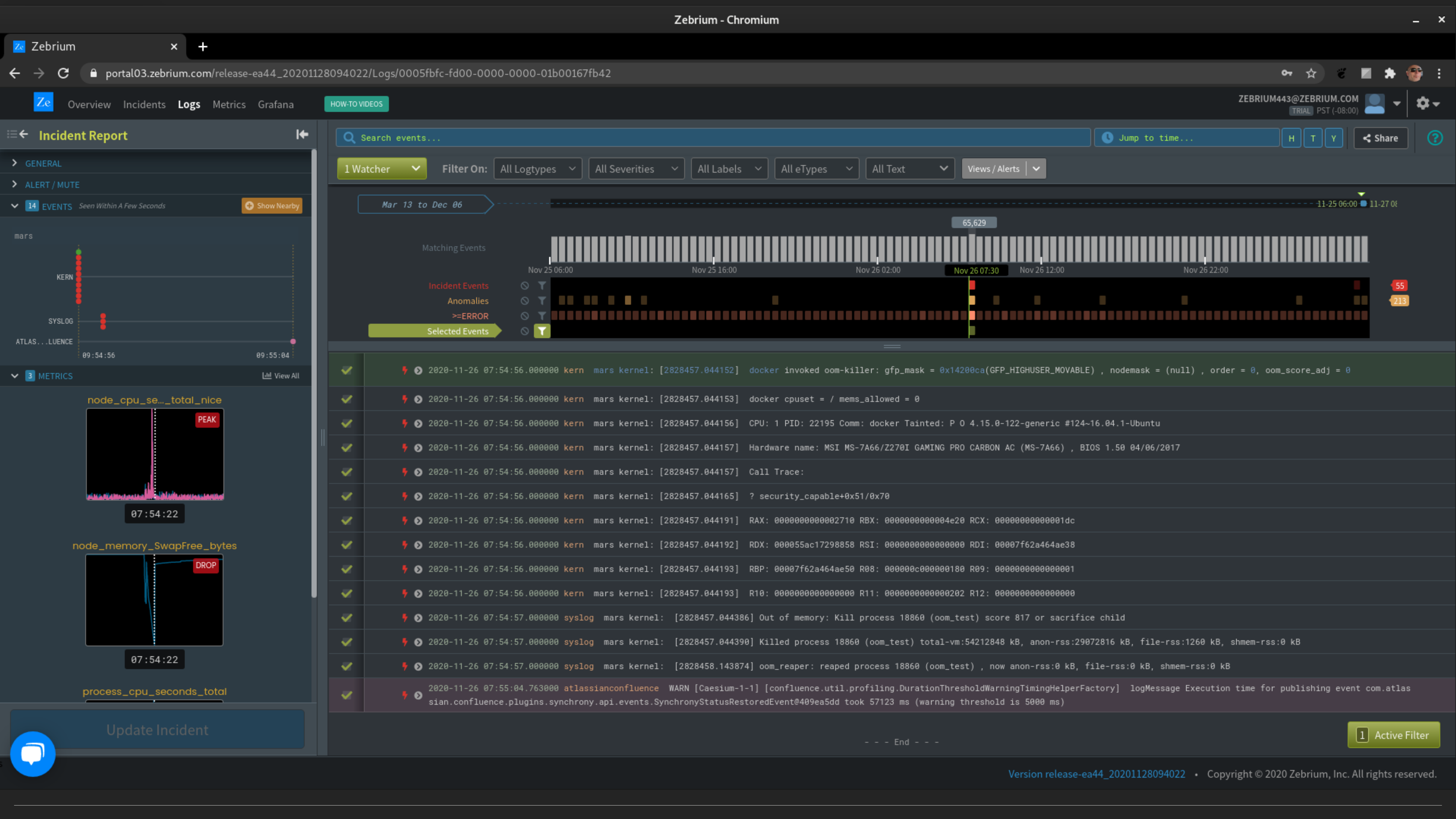
The root cause of the problem is that oom-killer was invoked because of a large number of allocations. The kernel's OOM killer is triggered when the system is out of memory and needs to free some memory. Since this action can kill processes, it is protected by a flag (oom_adj) which can be set or cleared by user space applications. When this flag is set, the kernel will kill processes for which there are no more than one page left in their memory cgroups (cgroups are used to control resource usage on a per-process basis). By default, Linux uses an algorithm called "RSS" (Resident Set Size) to decide whether or not to trigger the OOM killer. This algorithm calculates how much physical memory each process has reserved and compares it with its current virtual size. If there's enough memory available, then RSS will not touch any process even if they have been consuming too many resources for too long; but if there isn't enough memory available, then RSS will trigger the OOM killer and start killing processes until there's enough physical space again.

HOSTS mars LOG TYPES kern,syslog,atlassianconfluence

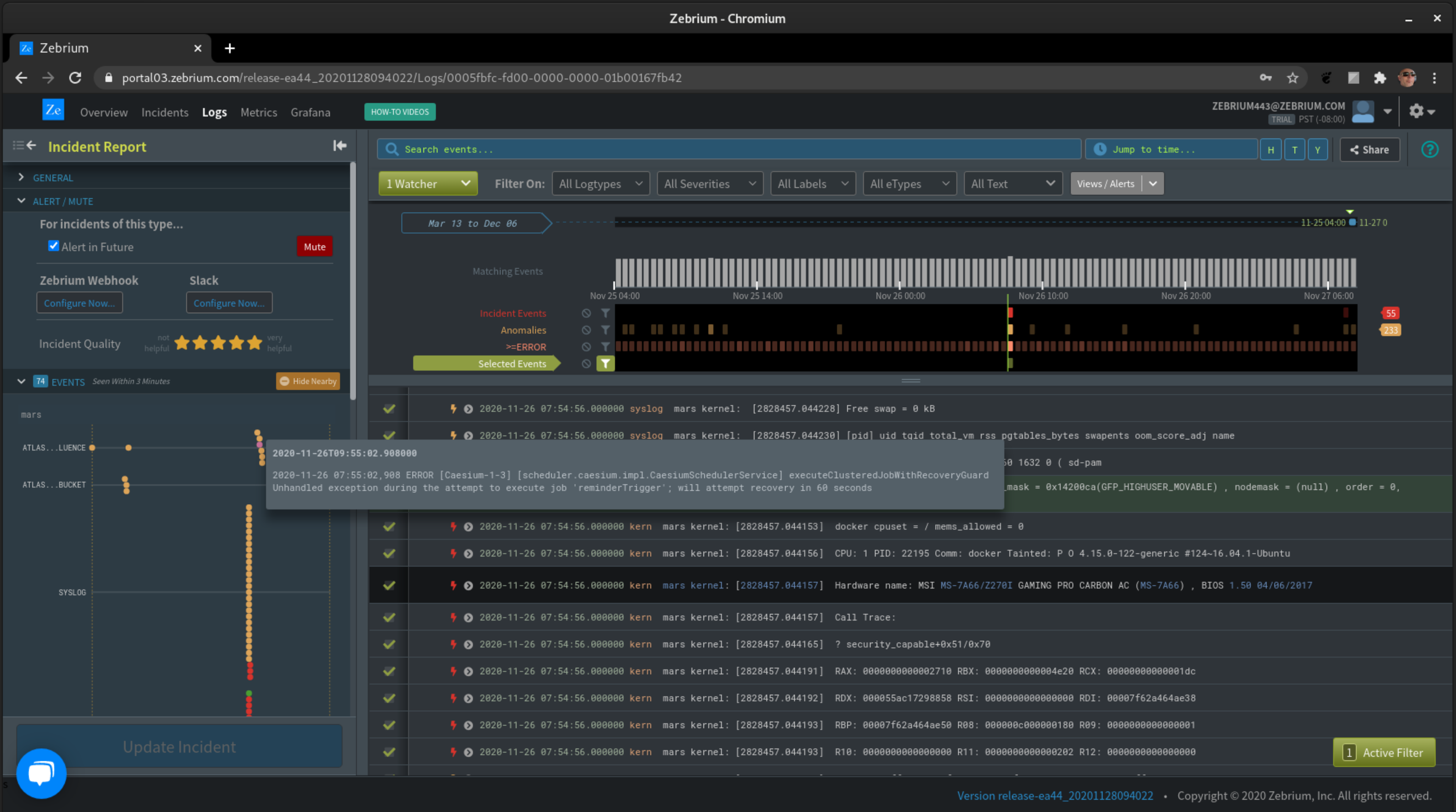
```
FIRST 2020-11-26T07:54:56.000000  LOGS:kern HOSTS: mars  Nov 26 07:54:56 mars kernel: [2828457.044152] docker invoked oom-killer:
gfp_mask = 0x14200ca(GFP_HIGHUSER_MOVABLE), nodemask = (null), order = 0, oom_score_adj = 0
```

```
WORST 2020-11-26T07:55:04.763000  LOGS:atlassianconfluence HOSTS: mars  2020-11-26 07:55:04,763 WARN [Caesium-1-1]
[confluence.util.profiling.DurationThresholdWarningTimingHelperFactory] logMessage Execution time for publishing event
com.atlassian.confluence.plugins.synchrony.api.events.SynchronyStatusRestoredEvent@409ea5dd took 57123 ms (warning threshold is 5000 ms)
```

Version release-ea44_20201128094022 •



```
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044152]  docker invoked oom-killer: gfp_mask = 0x14200ca(GFP_HIGHUSER_MOVABLE) , nodemask = (null) , order = 0, oom_score_adj = 0
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044153]  docker cpuset = / mems_allowed = 0
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044156]  CPU: 1 PID: 22195 Comm: docker Tainted: P O 4.15.0-122-generic #124~16.04.1-Ubuntu
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044157]  Hardware name: MSI MS-7A66/Z270I GAMING PRO CARBON AC (MS-7A66) , BIOS 1.50 04/06/2017
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044157]  Call Trace:
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044165]  ? security_capable+0x51/0x70
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044191]  RAX: 0000000000002710 RBX: 0000000000004e20 RCX: 00000000000001dc
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044192]  RDX: 000055ac17298858 RSI: 0000000000000000 RDI: 00007f62a464ae38
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044193]  RBP: 00007f62a464ae50 R08: 000000c000000180 R09: 0000000000000001
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044193]  R10: 0000000000000000 R11: 0000000000000202 R12: 0000000000000000
2020-11-26 07:54:57.000000 syslog  mars kernel:  [2828457.044386] Out of memory: Kill process 18860 (oom_test) score 817 or sacrifice child
2020-11-26 07:54:57.000000 syslog  mars kernel:  [2828457.044390] Killed process 18860 (oom_test) total-vm:54212848 kB, anon-rss:29072816 kB, file-rss:1260 kB, shmem-rss:0 kB
2020-11-26 07:54:57.000000 syslog  mars kernel:  [2828458.143874] oom_reaper: reaped process 18860 (oom_test) , now anon-rss:0 kB, file-rss:0 kB, shmem-rss:0 kB
2020-11-26 07:55:04.763000 atlassianconfluence  WARN [Caesium-1-1] [confluence.util.profiling.DurationThresholdWarningTimingHelperFactory]  logMessage Execution time for publishing event com.atlas
sian.confluence.plugins.synchrony.api.events.SynchronyStatusRestoredEvent@409ea5dd took 57123 ms (warning threshold is 5000 ms)

- - - End - - -
```



☑ Alert in Future

```
2020-11-26 07:54:56.000000 syslog  mars kernel:  [2828457.044228] Free swap = 0 kB
2020-11-26 07:54:56.000000 syslog  mars kernel:  [2828457.044230] [pid] uid tgid total_vm rss pgtables_bytes swapents oom_score_adj name
                                                                  50 1632 0 ( sd-pam
                                                                  _mask = 0x14200ca(GFP_HIGHUSER_MOVABLE) , nodemask = (null) , order = 0,
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044153]  docker cpuset = / mems_allowed = 0
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044156]  CPU: 1 PID: 22195 Comm: docker Tainted: P O 4.15.0-122-generic #124~16.04.1-Ubuntu
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044157]  Hardware name: MSI MS-7A66/Z270I GAMING PRO CARBON AC (MS-7A66) , BIOS 1.50 04/06/2017
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044157]  Call Trace:
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044165]  ? security_capable+0x51/0x70
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044191]  RAX: 0000000000002710 RBX: 0000000000004e20 RCX: 00000000000001dc
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044192]  RDX: 000055ac17298858 RSI: 0000000000000000 RDI: 00007f62a464ae38
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044193]  RBP: 00007f62a464ae50 R08: 000000c000000180 R09: 0000000000000001
2020-11-26 07:54:56.000000 kern  mars kernel: [2828457.044193]  R10: 0000000000000000 R11: 0000000000000202 R12: 0000000000000000
```

```
2020-11-26T09:55:02.908000

2020-11-26 07:55:02,908 ERROR [Caesium-1-3] [scheduler.caesium.impl.CaesiumSchedulerService] executeClusteredJobWithRecoveryGuard
Unhandled exception during the attempt to execute job 'reminderTrigger'; will attempt recovery in 60 seconds
```

Version release-ea44_20201128094022 • Copyright © 2020 Zebrium, Inc. All rights reserved.

Recent validation

"Zebrium detected every failure and root-cause, with no rules or training."

MayaData

"I didn't build any rules for this. Zebrium just picked up the issue."

DZone
A DEVADA MEDIA PROPERTY

"Zebrium dropped root-cause time from 3 hours to 15 minutes."

Sweetwater®

Ze

ZEBRIUM

Join us on this journey!

URL: zebrium.com/how-to-try

email: larry@zebrium.com

twitter: stochastimus@twitter.com

Gartner

COOL VENDOR 2020

DZone

A DEVADA MEDIA PROPERTY

Best Log Platform for Kubernetes 2020

Forbes 2020

AI 50

MOST PROMISING AI FIRMS

Gartner

Top 25 Enterprise Software Startups to Watch in 2020