AIOps Challenge

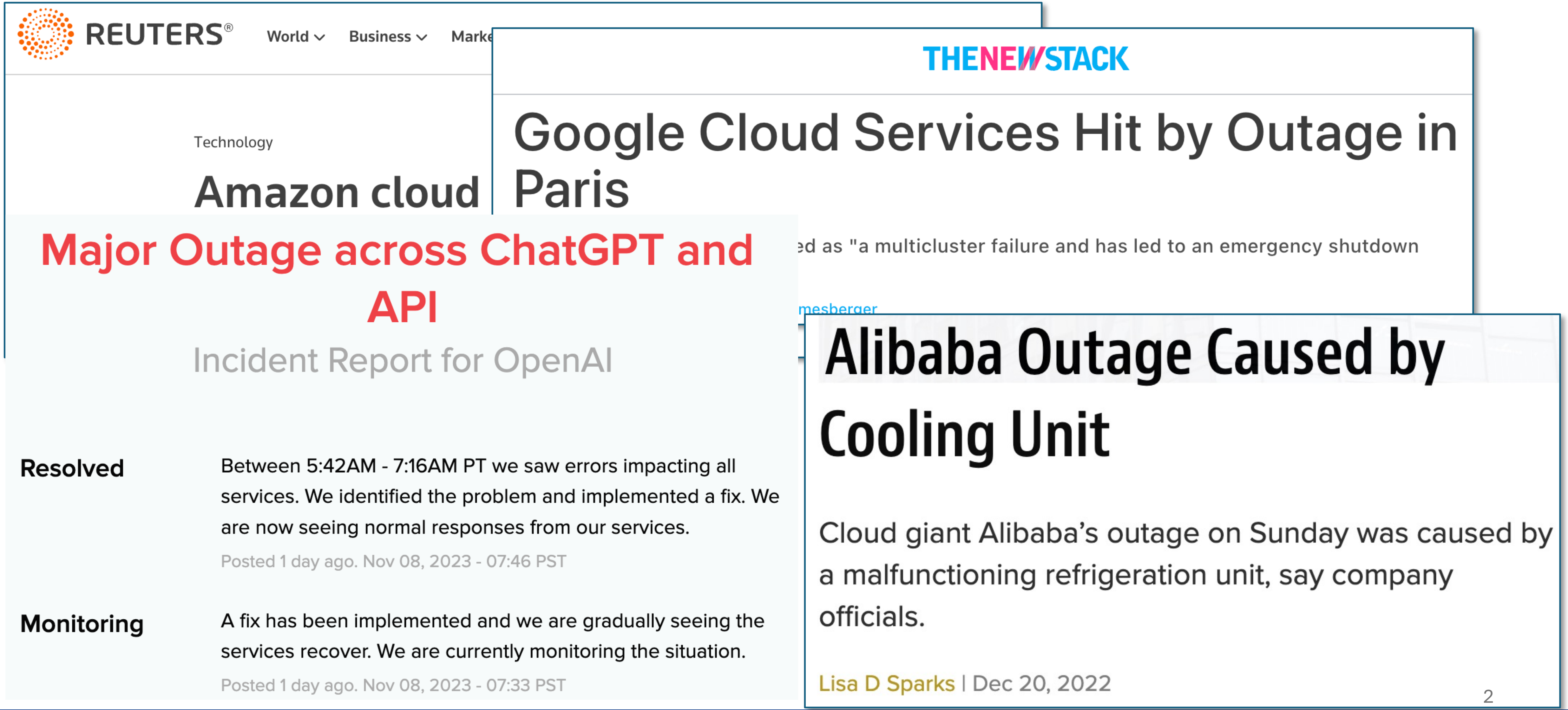
2023 CCF国际AIOps挑战赛决赛
暨“大模型时代的AIOps”研讨会

Automatic Root Cause Analysis via Large Language Models for Cloud Incidents

马明华 微软主管研究员

主办单位：中国计算机学会（CCF)、清华大学、中国建设银行股份有限公司、南开大学
承办单位：中国计算机学会互联网专委会、清华大学计算机科学与技术系、中国建设银行股份有限公司运营数据中心、南开大学软件学院、北京必示科技有限公司
赞助单位：华为技术有限公司、国网宁夏电力有限公司电力科学研究院、软通动力信息技术（集团）股份有限公司

Cloud and Its Incidents (!) are on the Rise



REUTERS®

Technology

Amazon cloud

THENEWSTACK

Google Cloud Services Hit by Outage in Paris

ed as "a multicluster failure and has led to an emergency shutdown

mesberger

Major Outage across ChatGPT and API

Incident Report for OpenAI

Resolved Between 5:42AM - 7:16AM PT we saw errors impacting all services. We identified the problem and implemented a fix. We are now seeing normal responses from our services.

Posted 1 day ago. Nov 08, 2023 - 07:46 PST

Monitoring A fix has been implemented and we are gradually seeing the services recover. We are currently monitoring the situation.

Posted 1 day ago. Nov 08, 2023 - 07:33 PST

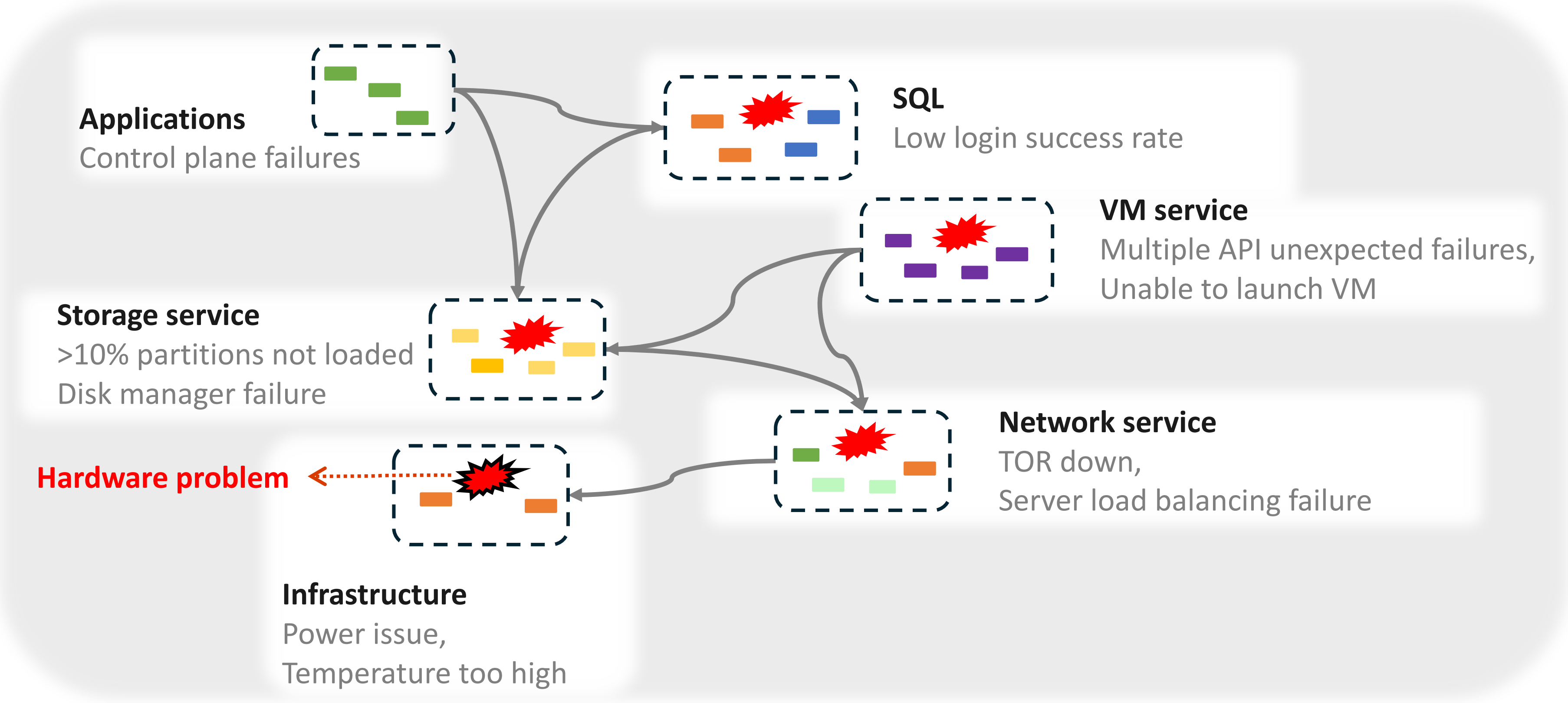
Alibaba Outage Caused by Cooling Unit

Cloud giant Alibaba’s outage on Sunday was caused by a malfunctioning refrigeration unit, say company officials.

Lisa D Sparks | Dec 20, 2022

Incident Root Cause Analysis (RCA)

- Triage the incident to the corresponding service team.
- Solve the incidents fundamentally and improve the service reliability.
- Prevent the similar incidents happen again in the future.



Applications
Control plane failures

SQL
Low login success rate

VM service
Multiple API unexpected failures,
Unable to launch VM

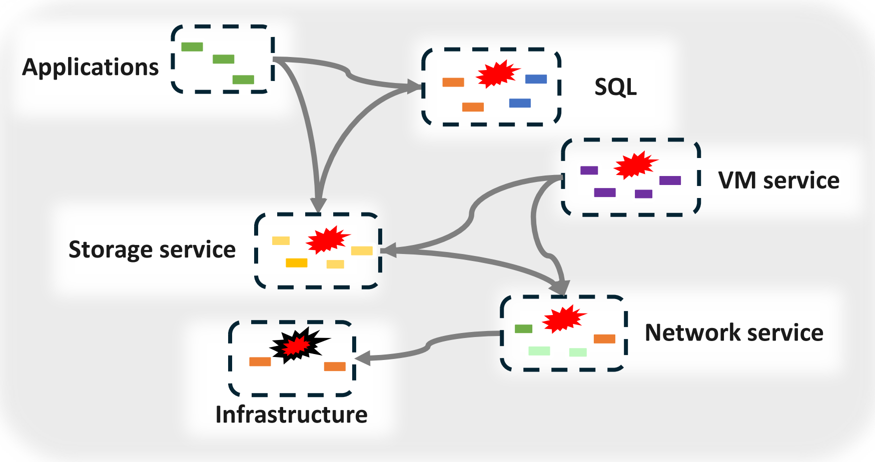
Storage service
>10% partitions not loaded
Disk manager failure

Network service
TOR down,
Server load balancing failure

Hardware problem

Infrastructure
Power issue,
Temperature too high

Challenges for Incident Root Cause Analysis

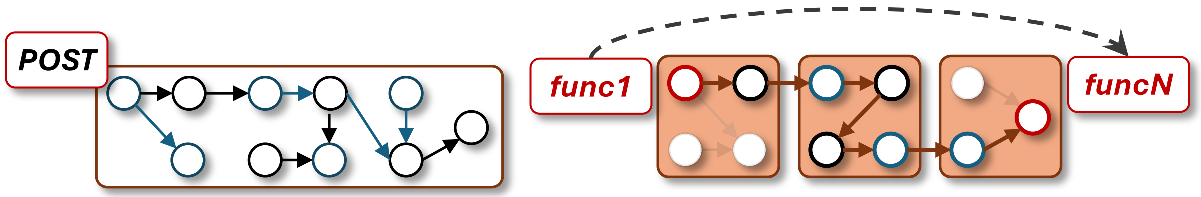


To win this war in fog, we have ...

Logs

```
07-29 19:17:57,939 – INFO [/10.10.10.01:2222] – Received connection request /11.11.11.01:5555
07-29 19:17:57,956 – WARN [Worker: 188979561024] – Interrupting SendWorker
07-29 19:18:01,926 – WARN [Worker: 188979561024] – Interrupting while waiting for msg on queue
07-29 19:18:07,944 – WARN [Worker: 188979561024] – Interrupting SendWorker
07-29 19:18:07,958 – WARN [Worker: 188979561024] – Interrupting SendWorker
```

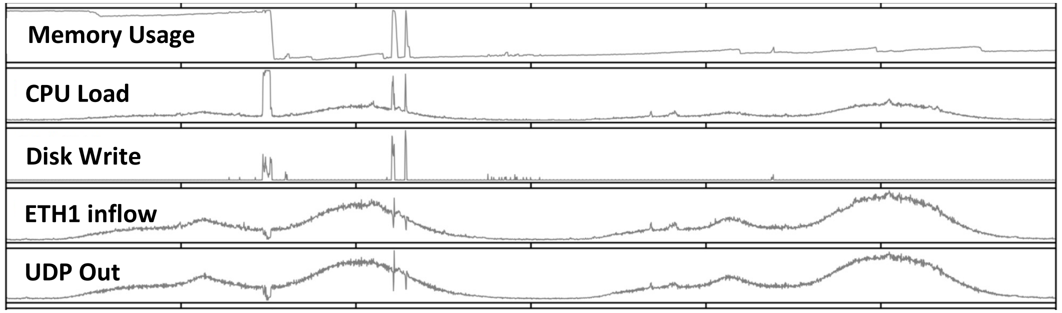
Traces



Request traces

Exception traces

Metrics



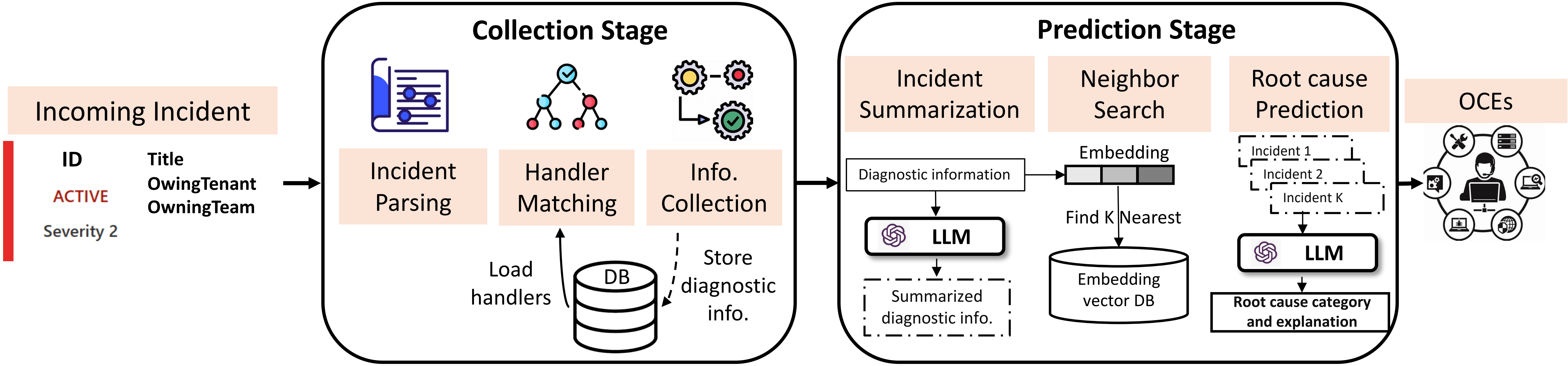
The debug information for **on-call engineers (OCE)** may be either **diffcult to collect** or **overwhelming**.

Goals of RCAssistant

When an incident happens, RCAssistant is able to:

- automatically collect incident-related information from multiple data sources, e.g., logs, metrics and traces
- automatically interpret and analyze the collected incident-related information and predict the root cause

RCAssistant - Architechture



Incoming Incident

ID

ACTIVE

Severity 2

Title

OwingTenant

OwningTeam

Collection Stage

Incident Parsing

Handler Matching

Info. Collection

Load handlers

DB

Store diagnostic info.

Prediction Stage

Incident Summarization

Diagnostic information

LLM

Summarized diagnostic info.

Neighbor Search

Embedding

Find K Nearest

Embedding vector DB

Root cause Prediction

Incident 1

Incident 2

Incident K

LLM

Root cause category and explanation

OCEs

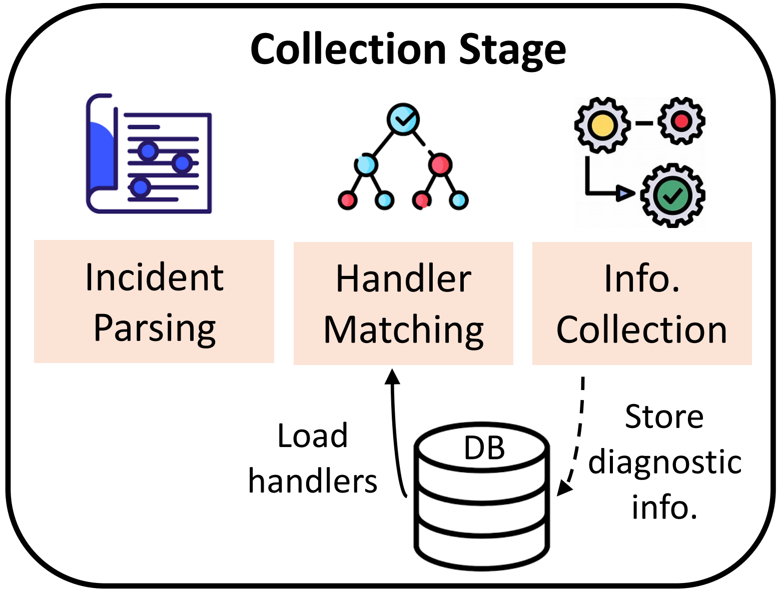
Diagnostic Information Collection Stage

Collecting diagnostic info -> Decision tree

RCAssistant will execute the predefined incident handler when an incident comes. Each incident handler is composed of multiple actions.

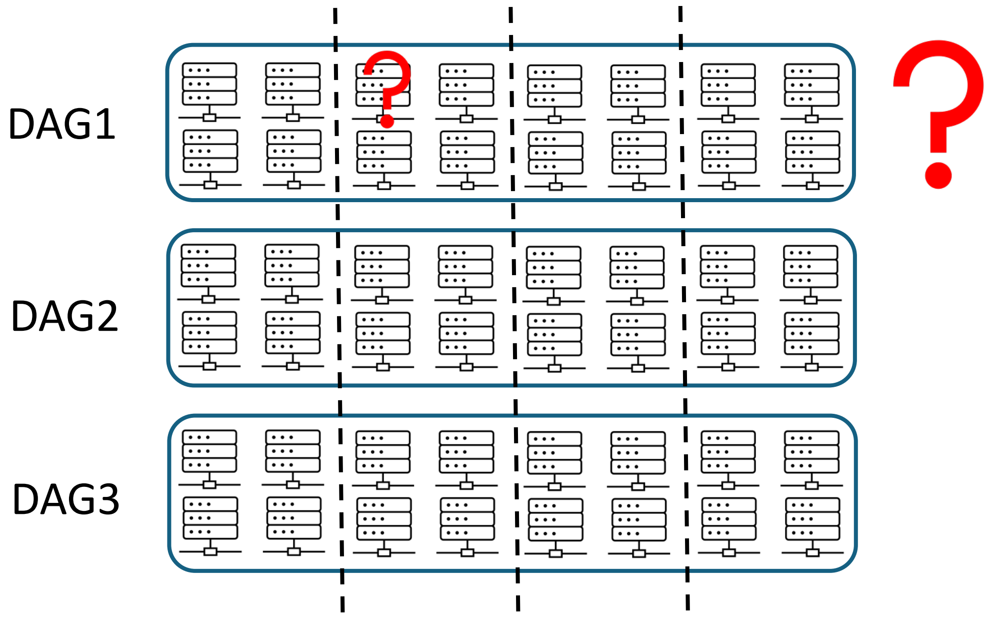
RCAssistant supports three types of actions:

- Scope switching action
- Query action
- Mitigation action



Collection Stage – Incident Handler and Actions

- Scope switching action
- Query action
- Mitigation action

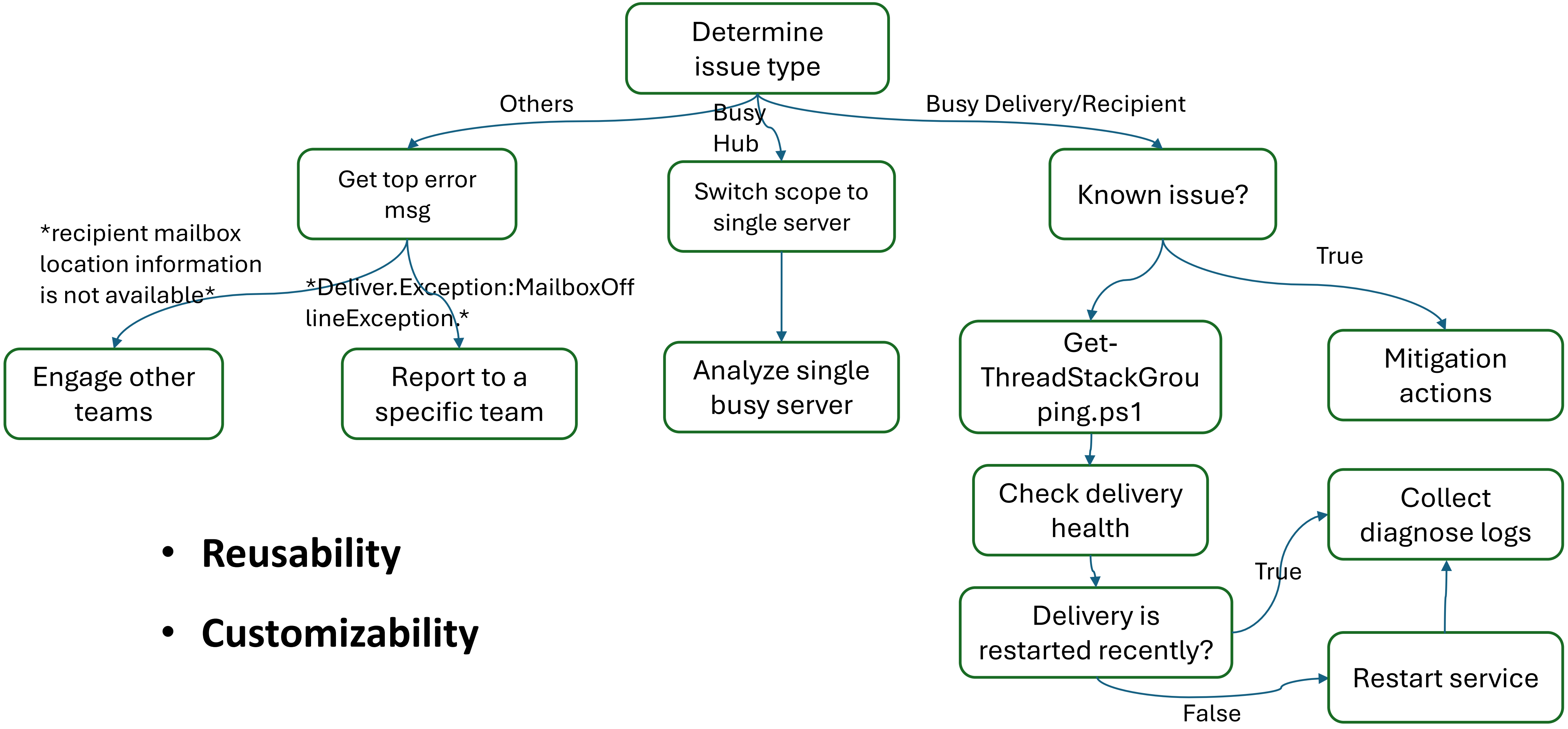


DAG: Database Available Group

Query action can query data from different sources and output the result as a key-value pair table.

Mitigation action suggests steps to fix, alleviate or triage an incident, such as "restart", or "engage other team".

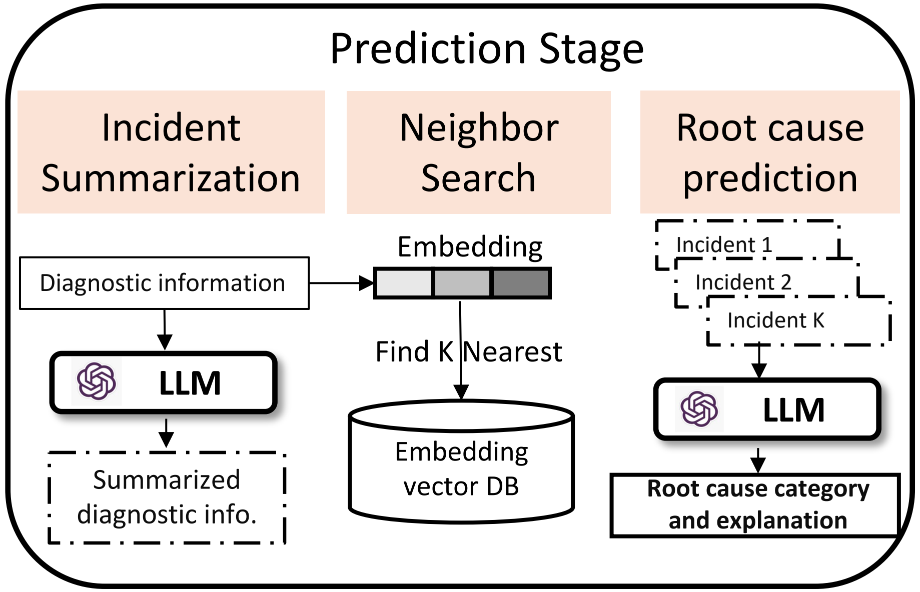
Collection Stage – Incident Handler and Actions

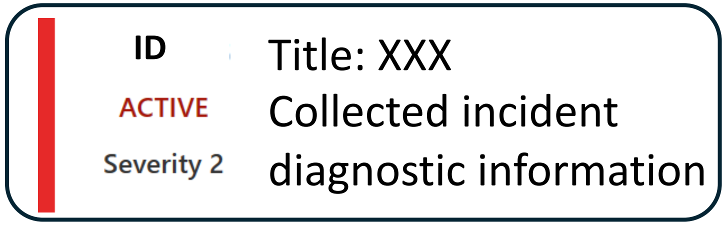

Determine issue type
Others
Busy Hub
Busy Delivery/Recipient
Get top error msg
Switch scope to single server
Known issue?
recipient mailbox location information is not available
Deliver.Exception:MailboxOfflineException.
True
Engage other teams
Report to a specific team
Analyze single busy server
Get-ThreadStackGrouping.ps1
Mitigation actions
Check delivery health
Collect diagnose logs
True
Delivery is restarted recently?
Restart service
False

- **Reusability**
- **Customizability**

Root Cause Prediction Stage

- Automatic few-shots chain-of-thoughts prompt construction
- Root cause categroy prediction and explanation





ID

ACTIVE

Severity 2

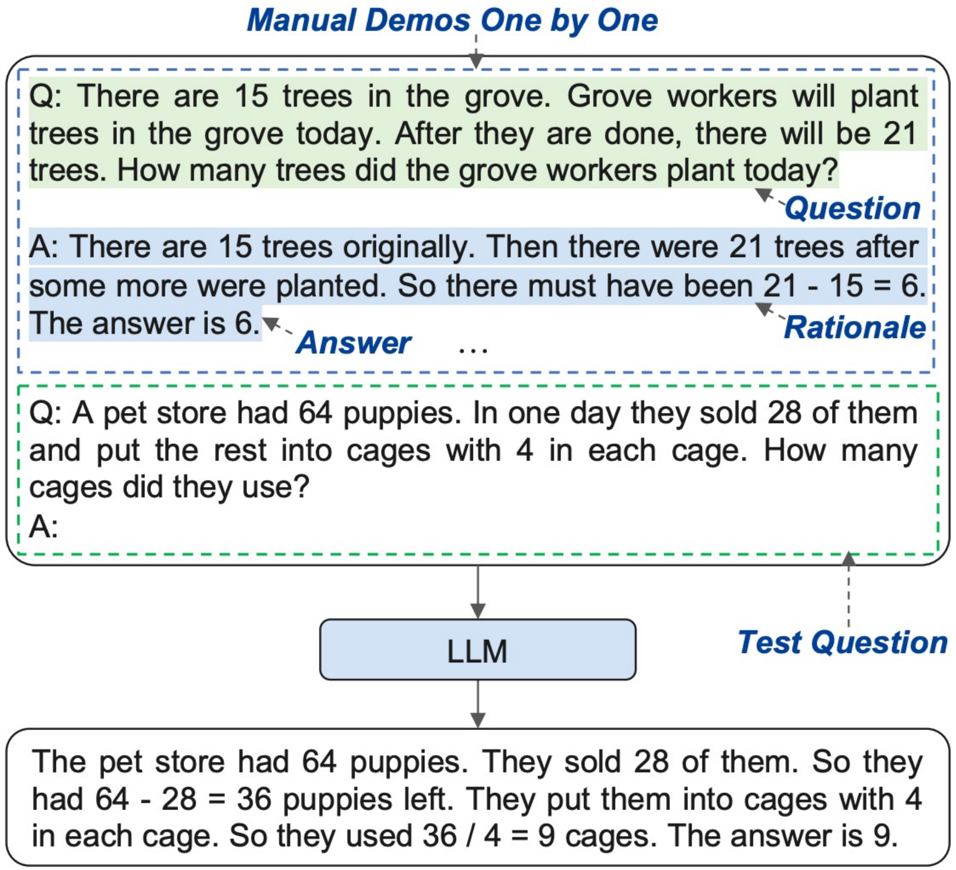
Title: XXX
Collected incident diagnostic information

Root Cause Category: HubPortExhaustion
Root cause details: The UDP hub ports on the machine [machine-XXX] had been run out ...

Prediction Stage – Chain-of-Thoughts

In few-shots CoT prompting, a few **demonstrations** that are composed of a question and a reasoning chain that leads to an answer for each of them.

- Demonstrations: historical incidents
    - Reasoning: diagnostic information
- Answer: root cause category label



Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.

Q: A pet store had 64 puppies. In one day they sold 28 of them and put the rest into cages with 4 in each cage. How many cages did they use?
A:

The pet store had 64 puppies. They sold 28 of them. So they had 64 - 28 = 36 puppies left. They put them into cages with 4 in each cage. So they used 36 / 4 = 9 cages. The answer is 9.

Root Cause Prediction Stage

Automatic few-shots chain-of-thoughts prompt construction

💡 **Solution:**

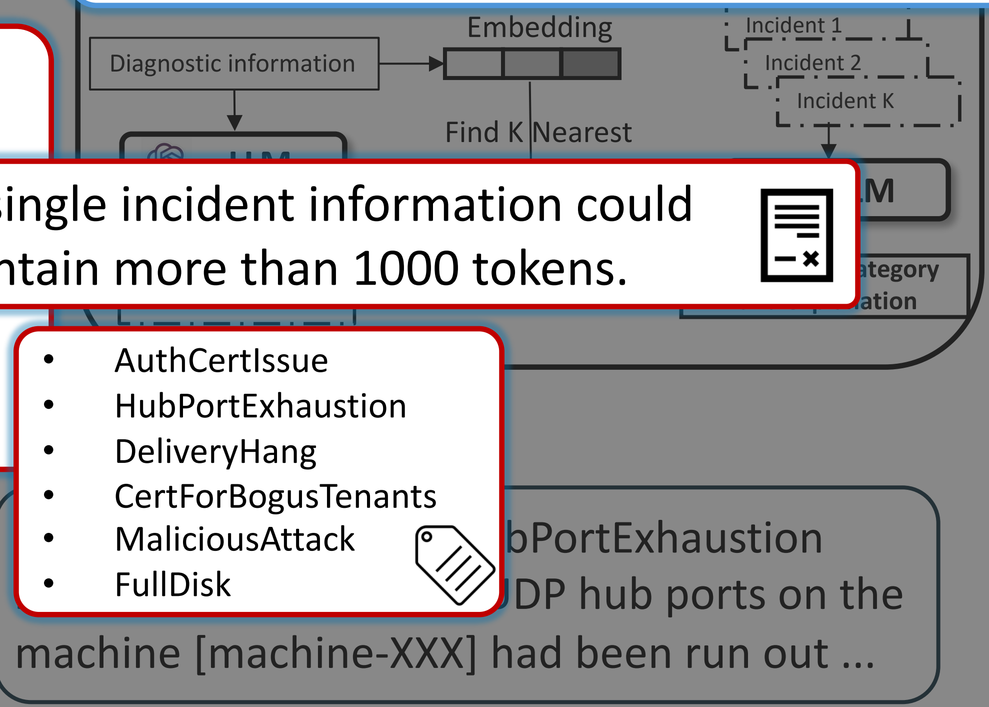
- **Similar incident retrieval**
- **Incident summarization**

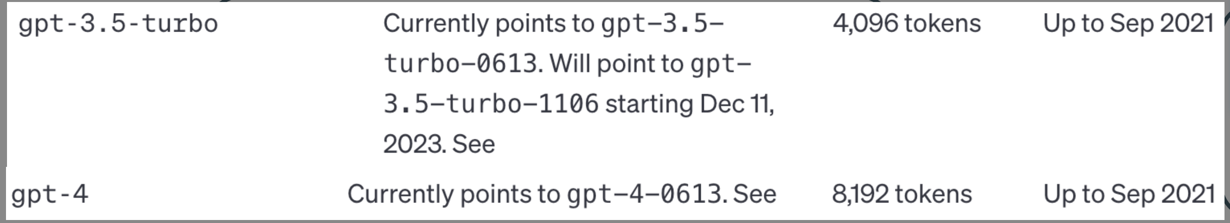
The collected incident information cannot fit into the prompt directly:

- Diagnostic information itself is lengthy
- Hundreds of root cause categories
- Token limit of large language models

A single incident information could contain more than 1000 tokens.

- AuthCertIssue
- HubPortExhaustion
- DeliveryHang
- CertForBogusTenants
- MaliciousAttack
- FullDisk





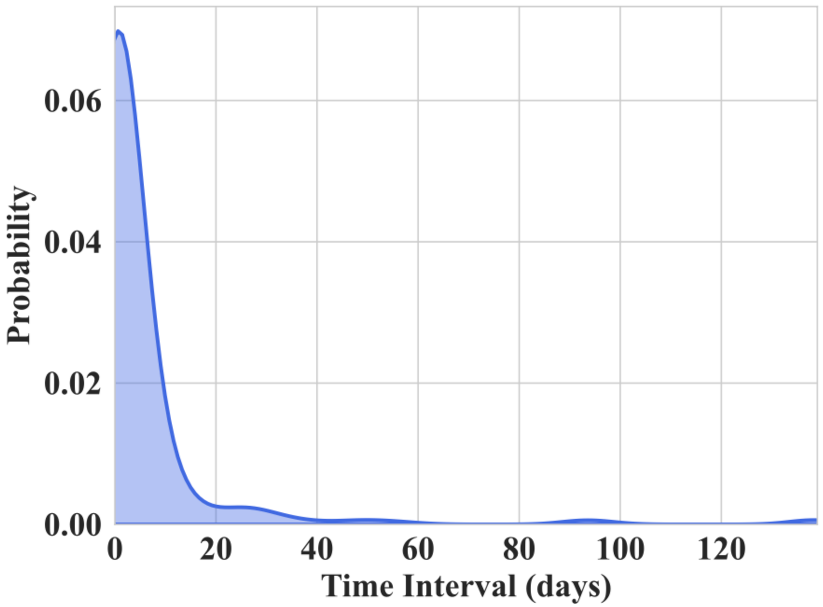
gpt-3.5-turbo	Currently points to gpt-3.5-turbo-0613. Will point to gpt-3.5-turbo-1106 starting Dec 11, 2023. See	4,096 tokens	Up to Sep 2021
gpt-4	Currently points to gpt-4-0613. See	8,192 tokens	Up to Sep 2021

Similar Incident Retrieval

- On-call engineers refer to historical incidents – Provide examples for LLM

How to measure the similarity?

- Study insight: incidents stemming from similar or identical root causes often recur within a short period – Time locality



Most recurring incidents (93.8%) tend to reappear within 20 days.

Similar Incident Retrieval

- On-call engineers refer to historical incidents – Provide examples for LLM
- Study insight: incidents stemming from similar or identical root causes often recur within a short period – Time locality

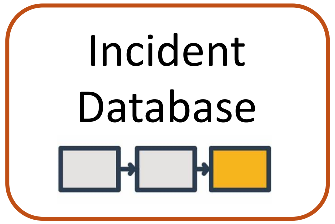
Embed incident diagnosis information and store in the database.

Measure similarity: $Distance(a, b) = ||a - b||_2$

fastText

$$Similarity(a, b) = \frac{1}{1 + Distance(a, b)} * e^{-\alpha|T(a) - T(b)|}$$

T(x) denotes the date of the incident.

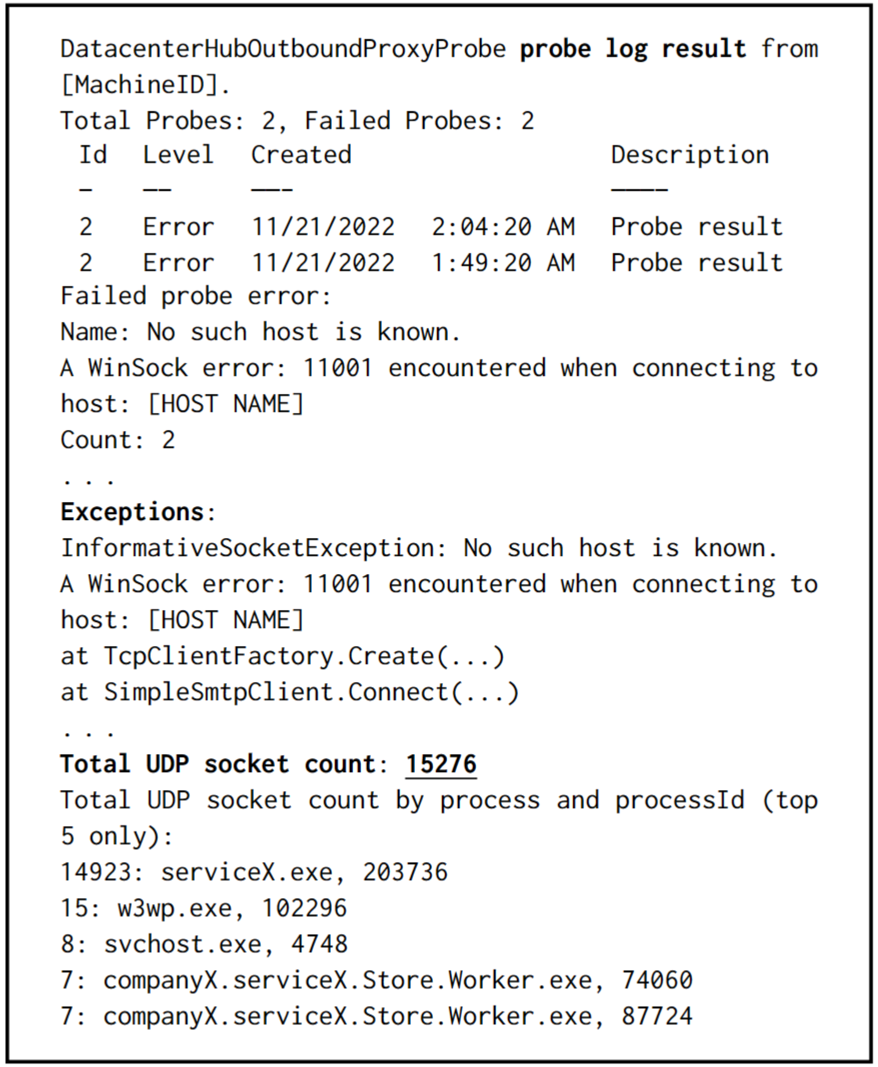


Incident Database

Incident Summarization

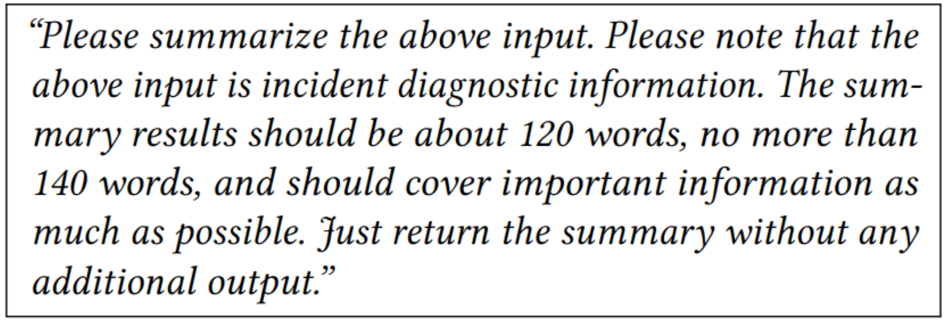
Original information collected by RCASSISTANT handler:

1000+ tokens



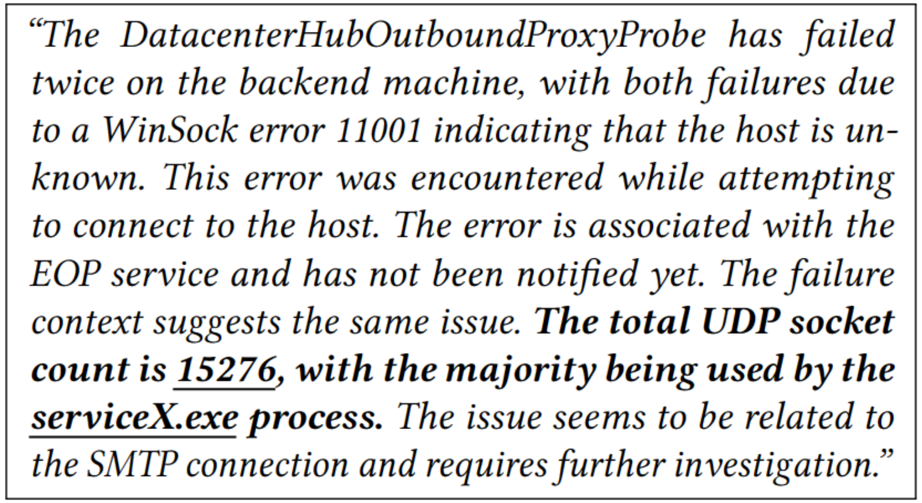
```
DatacenterHubOutboundProxyProbe probe log result from
[MachineID].
Total Probes: 2, Failed Probes: 2
 Id  Level  Created                 Description
 –   ––     –––                     ––––
 2   Error  11/21/2022  2:04:20 AM  Probe result
 2   Error  11/21/2022  1:49:20 AM  Probe result
Failed probe error:
Name: No such host is known.
A WinSock error: 11001 encountered when connecting to
host: [HOST NAME]
Count: 2
. . .
Exceptions:
InformativeSocketException: No such host is known.
A WinSock error: 11001 encountered when connecting to
host: [HOST NAME]
at TcpClientFactory.Create(...)
at SimpleSmtpClient.Connect(...)
. . .
Total UDP socket count: 15276
Total UDP socket count by process and processId (top
5 only):
14923: serviceX.exe, 203736
15: w3wp.exe, 102296
8: svchost.exe, 4748
7: companyX.serviceX.Store.Worker.exe, 74060
7: companyX.serviceX.Store.Worker.exe, 87724
```

Prompt used to summarize the incident information:



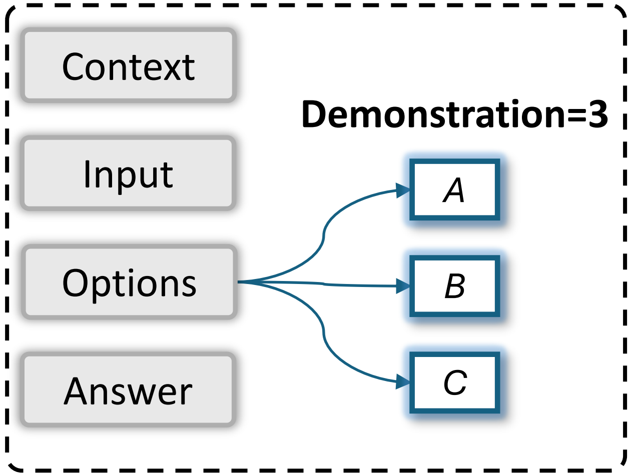
"Please summarize the above input. Please note that the above input is incident diagnostic information. The summary results should be about 120 words, no more than 140 words, and should cover important information as much as possible. Just return the summary without any additional output."

RCASSISTANT summary result:



"The DatacenterHubOutboundProxyProbe has failed twice on the backend machine, with both failures due to a WinSock error 11001 indicating that the host is unknown. This error was encountered while attempting to connect to the host. The error is associated with the EOP service and has not been notified yet. The failure context suggests the same issue. ***The total UDP socket count is 15276, with the majority being used by the serviceX.exe process.*** *The issue seems to be related to the SMTP connection and requires further investigation."*

Automatic Chain-of-Thoughts Prompting



Context: The following description shows ... Please select the incident information that is most likely to have the same root cause and give your explanation ... If not, please select the option ‘None’.

Input: The DatacenterHubOutboundProxyProbe probe result from [BackEndMachine] is a failure...

Options:

- A: ***Keyword***: Delivery hang. ***Summary***: There are 62 managed threads in process [MSExchangeDelivery]. The most common thread stacks are .IL_STUB_PInvoke, System.Threading.WaitHandle ...
- B: ***Keyword***: Code regression. ***Summary***: The DatacenterHubOutboundProxyProbe probe from the [BackEndMachine] failed with ...
- C: ***Keyword***: None

Answer: ... ...

Evaluation Results

RCAssistant achieves 0.766 F1-score when predicting the root causes.

Method	F1-score		Average Time (seconds)	
	Micro	**Macro**	**Train.**	**Infer.**
FastText	0.076	0.004	10.592	0.524
XGBoost	0.022	0.009	11.581	1.211
Fine-tune GPT	0.103	0.144	3192	4.262
GPT-4 Prompt	0.026	0.004	-	3.251
GPT-4 Embed.	0.257	0.122	1925	3.522
RCAssistant (GPT-3.5)	0.761	0.505	10.562	4.221
RCAssistant (GPT-4)	**0.766**	**0.533**	10.562	4.205

Conclusion

- We propose RCAssistant, an automated end-to-end solution for cloud incident root cause analysis:
  - **Efficient** incident-related diagnostic data collection
  - Integration of a **large language model** to predict incident root cause categories with explanations
  - Successfully deployed in the real-world cloud systems

AIOps
Challenge

2023 CCF国际AIOps挑战赛决赛暨“大模型时代的AIOps”研讨会

THANKS