2024 TSMC IT CareerHack
Digital Excellence × Generative AI

以生成式 AI 技術實現

IT Infra 自動化監控與管理

C5 我想吃滷豆乾跟sudo

國立成功大學資訊工程學系 大二

李緒成 劉哲佑 李逵安 張百鴻

OUTLINE

User Story

System Architecture

Application Implement

Demo

DevOps

Testing

Future

User Story (1/2)

Senior SA

Jason

Jason:

We must closely monitor the consumer, as any errors could result in significant financial losses. We can only assign this task to our Junior SA Henry.

Junior SA

Henry

Henry:

It feels like such a waste of my skills, and monitoring around the clock is impossible. Handling these trivial tasks daily is incredibly boring. I'm seriously considering quitting : (

AI

Jerry

Jerry:

It’s my show time!

User Story (2/2)

Junior SA

AI Monitor System

- Replace the role of a junior SA
- Analyze Consumer’s log
- Resolve the issue automatically

System Error Detected

Auto-Scale Consumer resources

Discord Notification

- By Discord bot
- Senior SA can send feedback to help generating better notification mess

OUTLINE

User Story

System Architecture

Application Implement

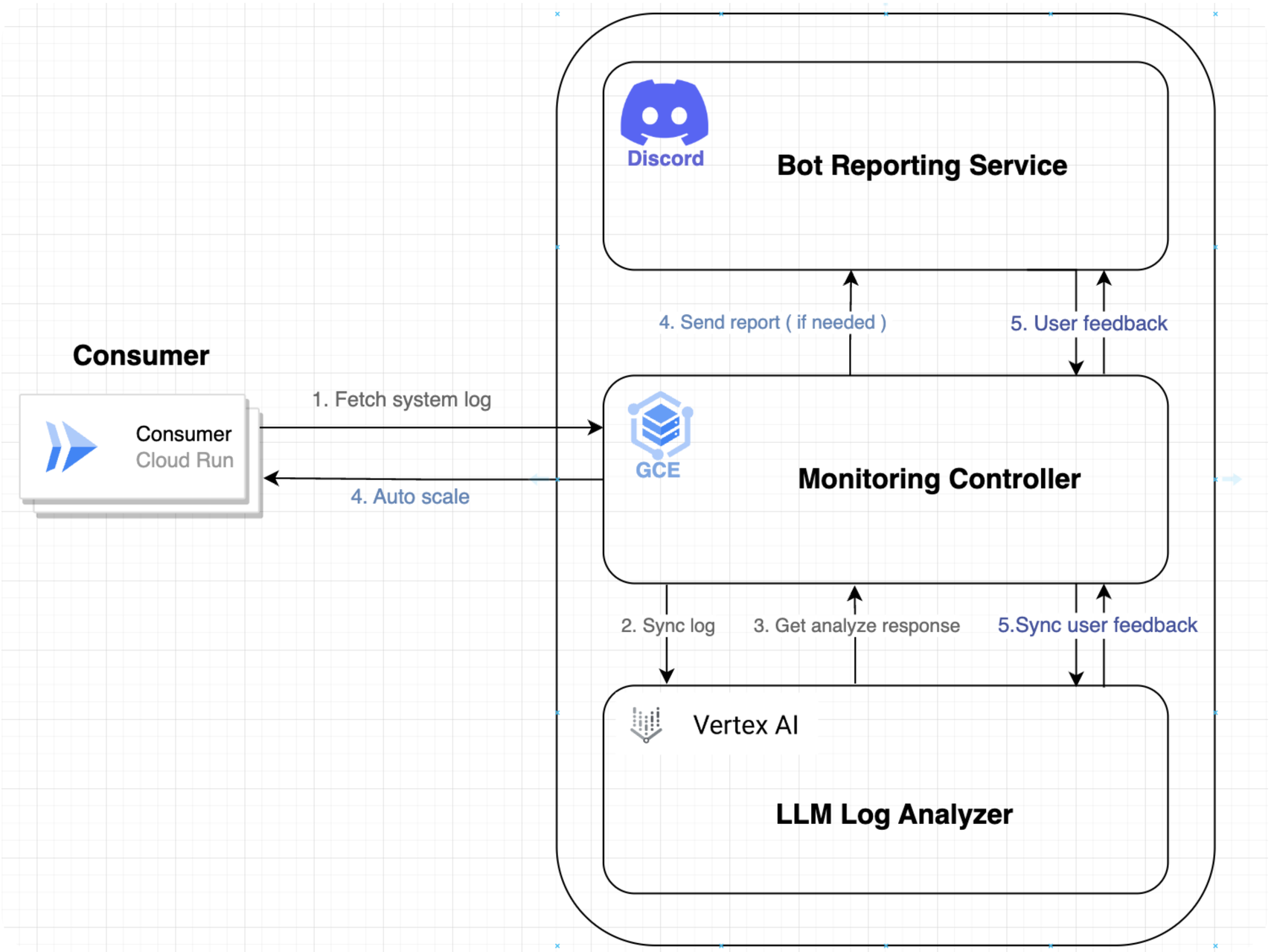
Demo

DevOps

Testing

Future

System Architecture



OUTLINE

User Story

System Architecture

Application Implement

Demo

DevOps

Testing

Future

Consumer

- stimulate async task
- stimulate error status
- feature metrics :
  - remain_count
  - avg_exe_time

Consumer

Health

GET / Health

Job

GET /api/job/start Start Normal Behavior

GET /api/job/stop Stop Normal Behavior

GET /api/job/status Status Normal Behavior

GET /api/job/sleep/{seconds} Sleep

Full

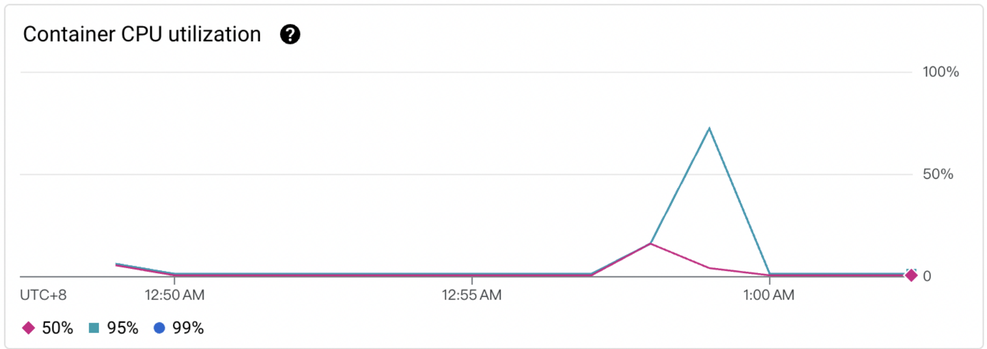
GET /api/full/cpu/duration/{duration} Full Cpu

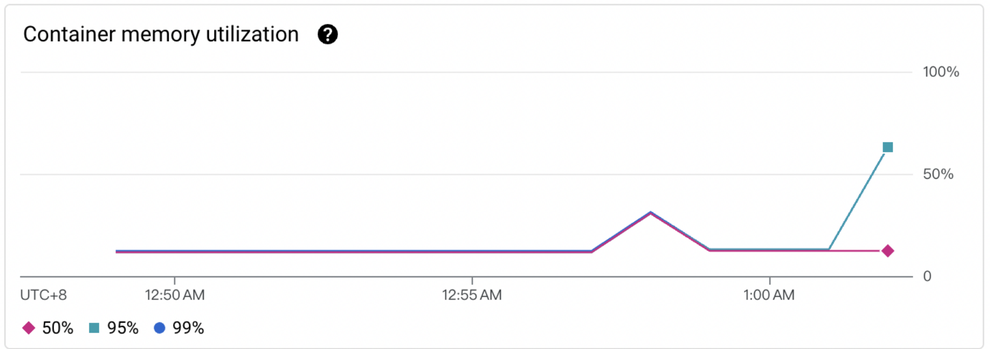
GET /api/full/ram/duration/{duration} Full Ram

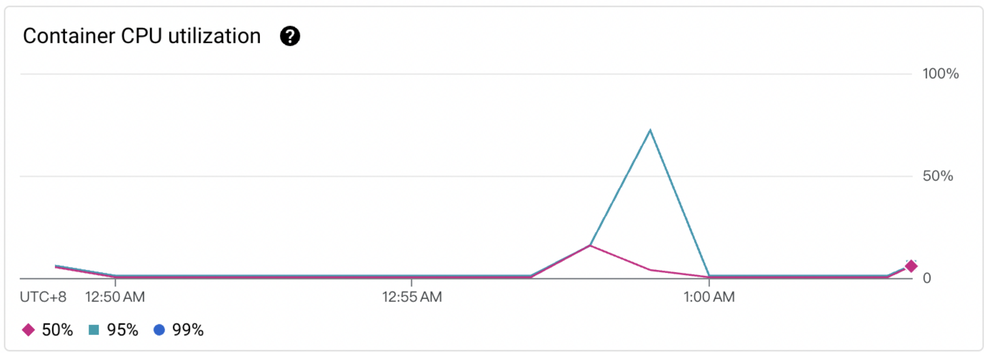
GET /api/full/enque/{num} Enque

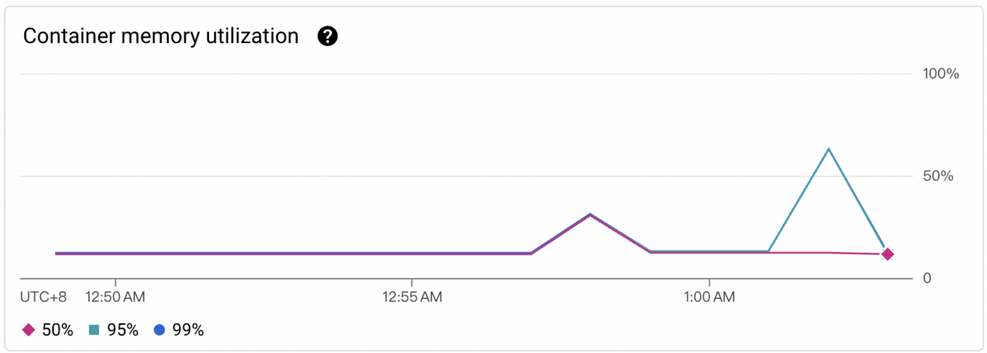
GET /api/full/error Error

Consumer


Container CPU utilization
100%
50%
0
UTC+8
12:50 AM
12:55 AM
1:00 AM
50%
95%
99%

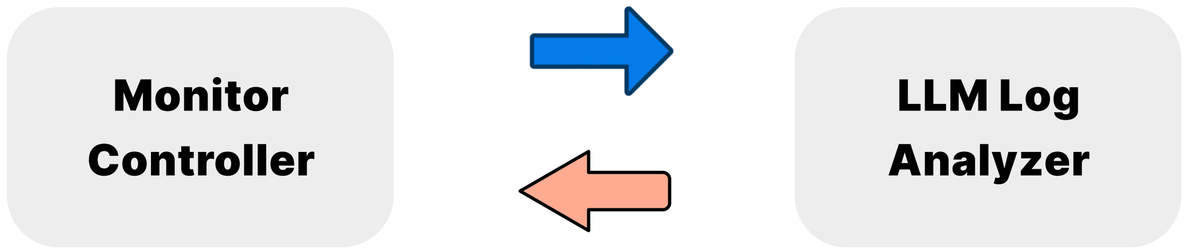

Container memory utilization
100%
50%
0
UTC+8
12:50 AM
12:55 AM
1:00 AM
50%
95%
99%


Container CPU utilization
100%
50%
0
UTC+8
12:50 AM
12:55 AM
1:00 AM
50%
95%
99%


Container memory utilization
100%
50%
0
UTC+8
12:50 AM
12:55 AM
1:00 AM
50%
95%
99%

GenAI - Analyze Flow

1. send log data
   to analyzer



2. send feedback
   to report and
   scale

Problem:

- **AI could give the wrong suggestion**
- **No ways for it to improve**
- **We need a way for AI to learn from experience**

GenAI - Analyze Flow with memory

1. send log data to analyzer

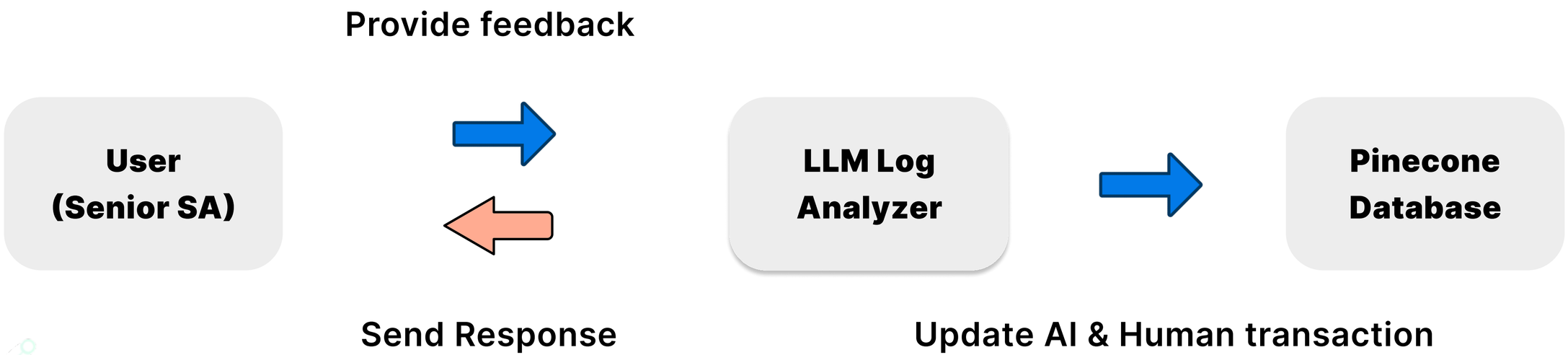
2. retrieve similar log from past and its analysis feedback

4. send feedback to report and scale

3. Store log data and newly generated analysis feedback

GenAI - Feedback Response

After monitor controller send report to user(senior SA) ...



GenAI - Prompt Engineering

1. Detect error/warning thresholds by rule. (heuristic_analysis)
2. Retrieve similar log and analysis from Pinecone. (memory)
3. Formulate prompt with 1. and 2. and ask GenAI for analysis
4. Parse the analysis result into JSON format for scaling and reporting

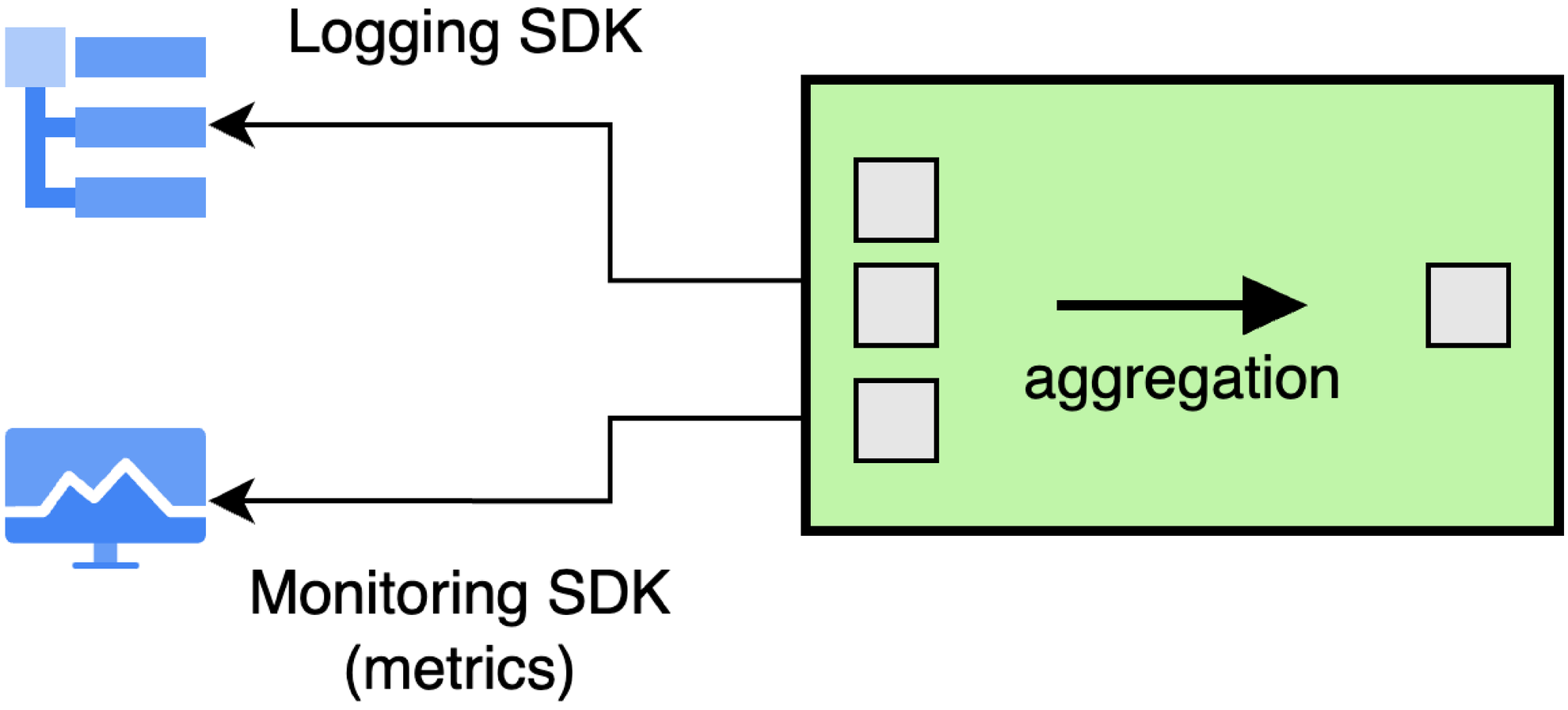
```
The following text contains log data for a Google Cloud Run application. \
This data is presented in CSV format and encompasses the most recent {time_span} minutes:
{log_data}

{memory}

Some heuristic analysis has been performed beforehand, the following text is a summary of the analysis:
{heuristic_analysis}

Your task is to provide an in-depth analysis based on the provided log data and heuristic analysis feedback, \
and scale the application by providing the number of CPUs, amount of memory in MB, and amount of instances to add(positive) or remove(negative).
The system will automatically scale the application based on your feedback.

{format_instruction}
```

Monitor Controller - Logs & Metrics

Application Log and Metrics Real-Time Retrieval

- Application Log Structure
  - Remaining Task Count in Queue
  - Average Task Execution Time
  - ...
- What metrics we retrieved?
  - Container CPU Utilization (%)
  - Container Memory Utilization (%)
  - Instance Count (active)
  - Instance Count (idle)

Monitor Controller - Logs & Metrics

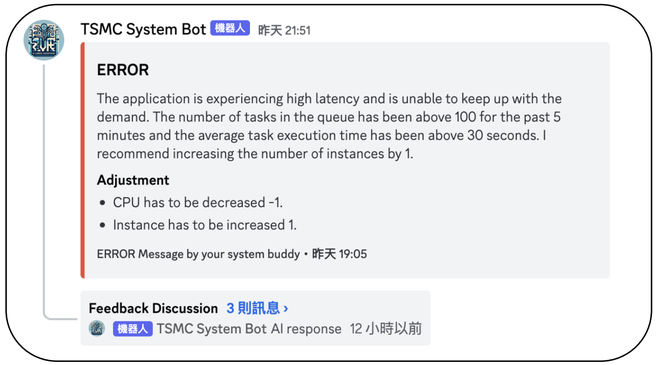


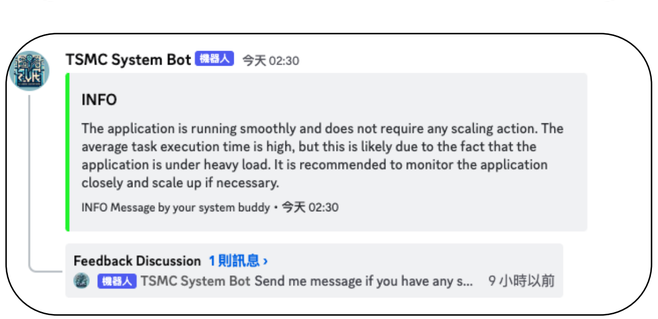
Discord bot

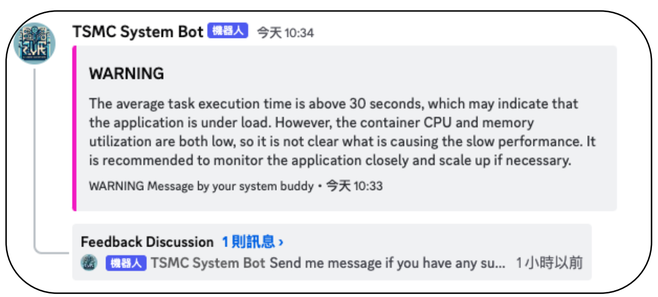
FLOW

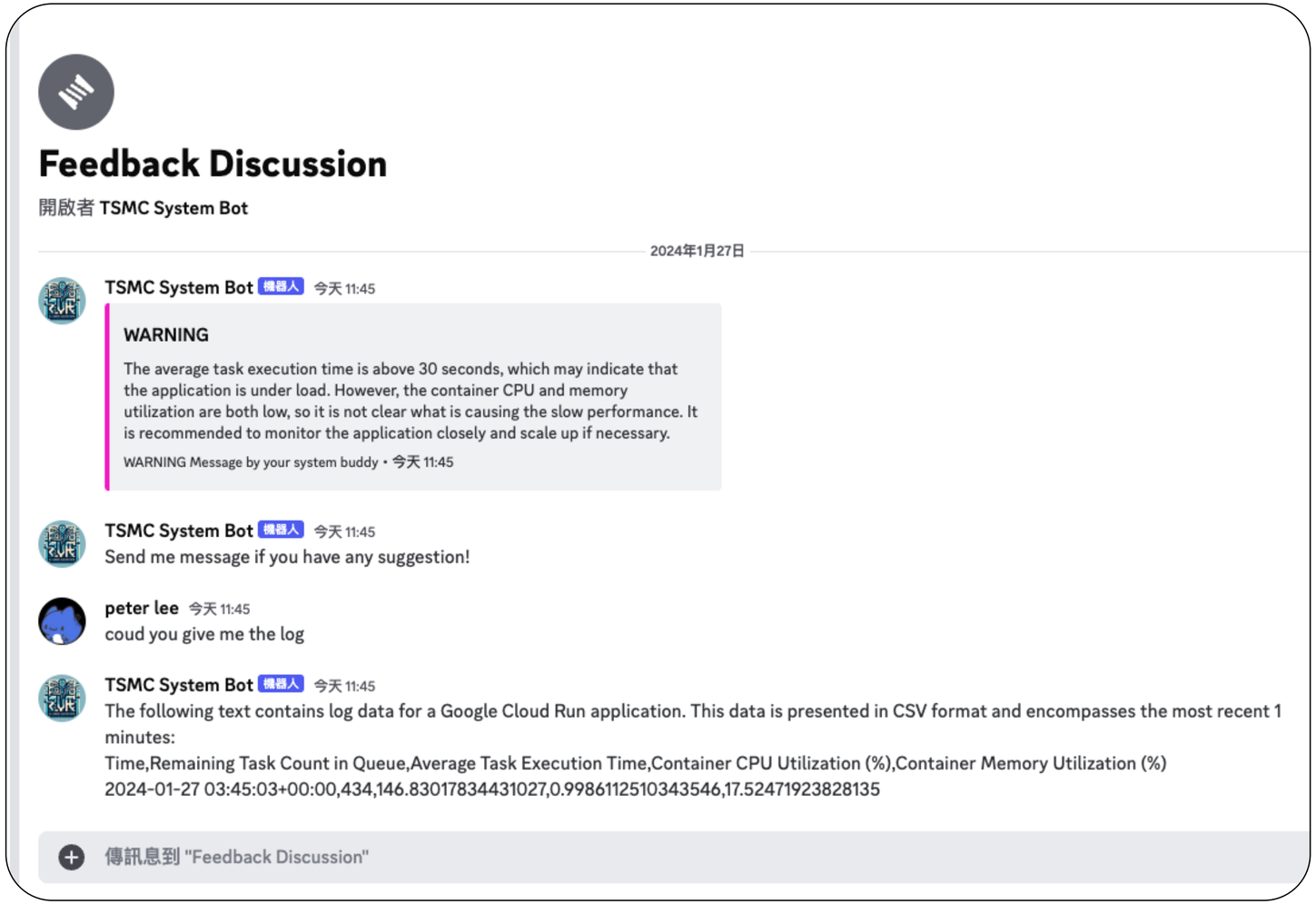
1. Monitor send **Report Request**
2. Analyze notification data and transfer it into **message**. Open a **thread** at the same time for feedback.
3. If there are any **feedback**. Send it to **GenAI** and wait for **response**.

Discord bot


TSMC System Bot 機器人 昨天 21:51
ERROR
The application is experiencing high latency and is unable to keep up with the demand. The number of tasks in the queue has been above 100 for the past 5 minutes and the average task execution time has been above 30 seconds. I recommend increasing the number of instances by 1.
Adjustment
• CPU has to be decreased -1.
• Instance has to be increased 1.
ERROR Message by your system buddy • 昨天 19:05
Feedback Discussion 3 則訊息 ›
機器人 TSMC System Bot AI response 12 小時以前


TSMC System Bot 機器人 今天 02:30
INFO
The application is running smoothly and does not require any scaling action. The average task execution time is high, but this is likely due to the fact that the application is under heavy load. It is recommended to monitor the application closely and scale up if necessary.
INFO Message by your system buddy • 今天 02:30
Feedback Discussion 1 則訊息 ›
機器人 TSMC System Bot Send me message if you have any s... 9 小時以前

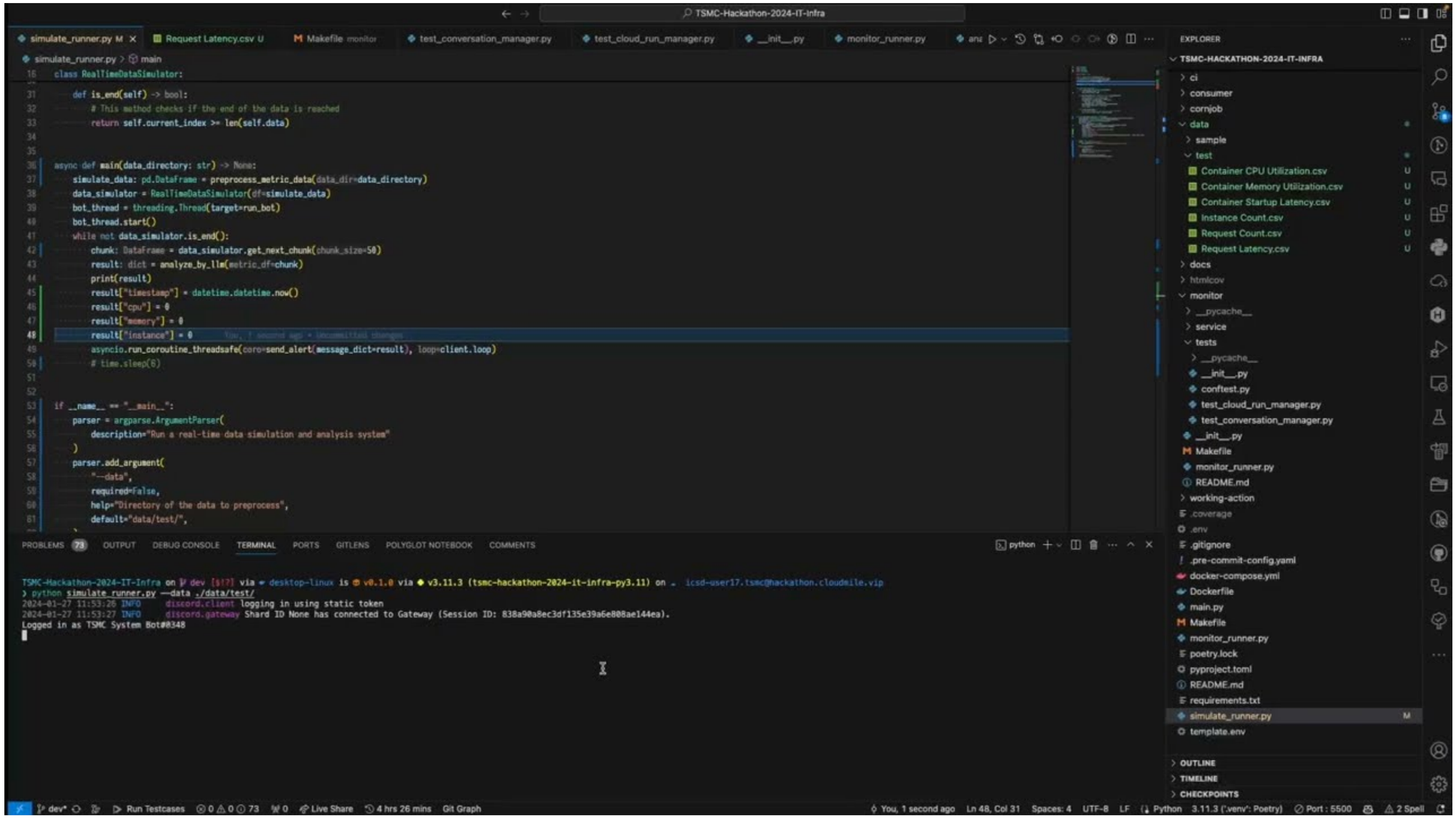

TSMC System Bot 機器人 今天 10:34
WARNING
The average task execution time is above 30 seconds, which may indicate that the application is under load. However, the container CPU and memory utilization are both low, so it is not clear what is causing the slow performance. It is recommended to monitor the application closely and scale up if necessary.
WARNING Message by your system buddy • 今天 10:33
Feedback Discussion 1 則訊息 ›
機器人 TSMC System Bot Send me message if you have any su... 1 小時以前


Feedback Discussion
開啟者 TSMC System Bot
2024年1月27日
TSMC System Bot 機器人 今天 11:45
WARNING
The average task execution time is above 30 seconds, which may indicate that the application is under load. However, the container CPU and memory utilization are both low, so it is not clear what is causing the slow performance. It is recommended to monitor the application closely and scale up if necessary.
WARNING Message by your system buddy • 今天 11:45
TSMC System Bot 機器人 今天 11:45
Send me message if you have any suggestion!
peter lee 今天 11:45
coud you give me the log
TSMC System Bot 機器人 今天 11:45
The following text contains log data for a Google Cloud Run application. This data is presented in CSV format and encompasses the most recent 1 minutes:
Time,Remaining Task Count in Queue,Average Task Execution Time,Container CPU Utilization (%),Container Memory Utilization (%)
2024-01-27 03:45:03+00:00,434,146.83017834431027,0.9986112510343546,17.52471923828135
傳訊息到 "Feedback Discussion"

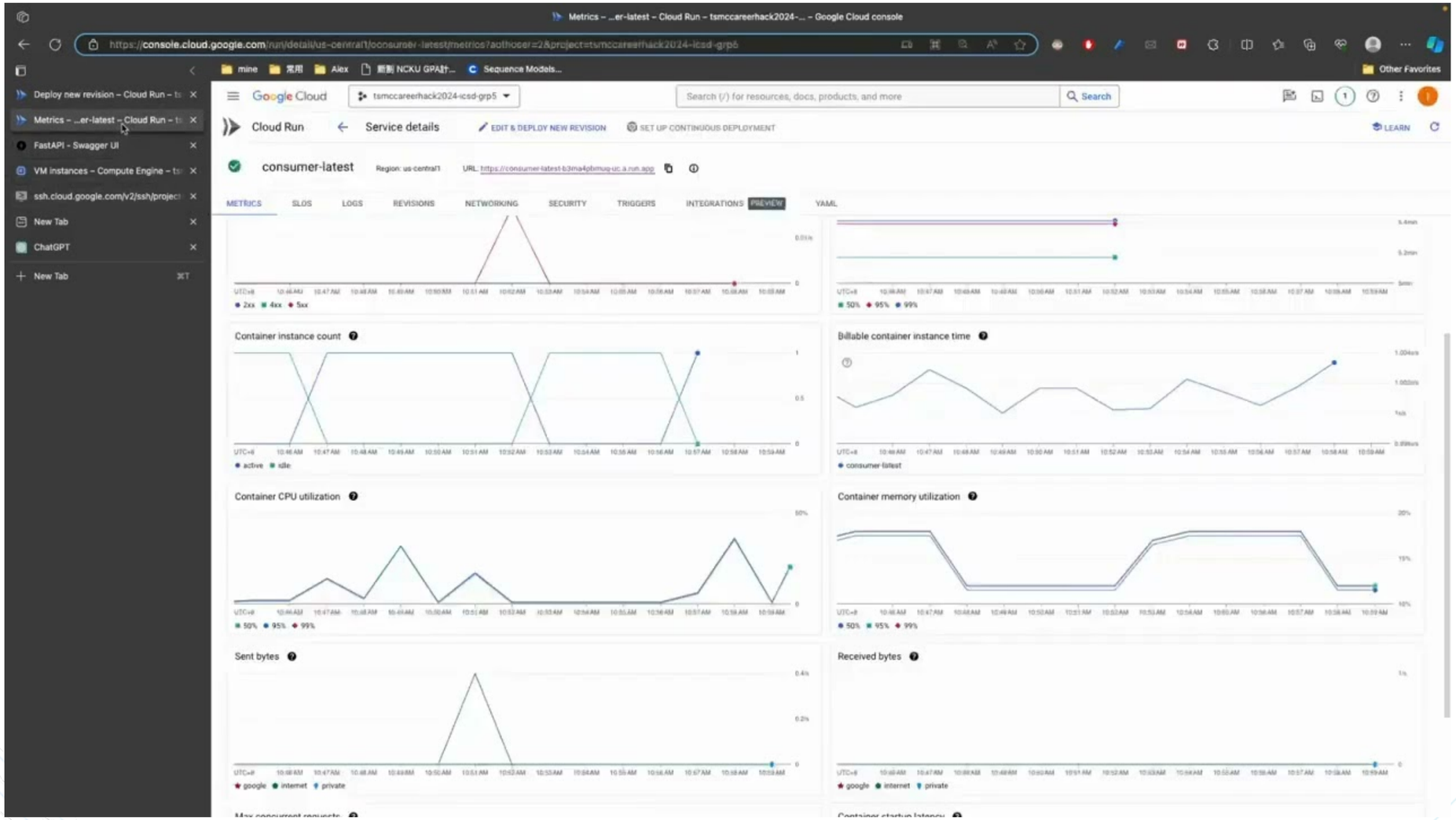
OUTLINE

User Story
System Architecture
Application Implement
Demo
DevOps
Testing
Future

Demo - Testing Dataset



Demo - Real time stimulation



OUTLINE

User Story
System Architecture
Application Implement
Demo
DevOps
Testing
Future

DevOps Outline

- Linting
- CI/CD
- Demo

Linting

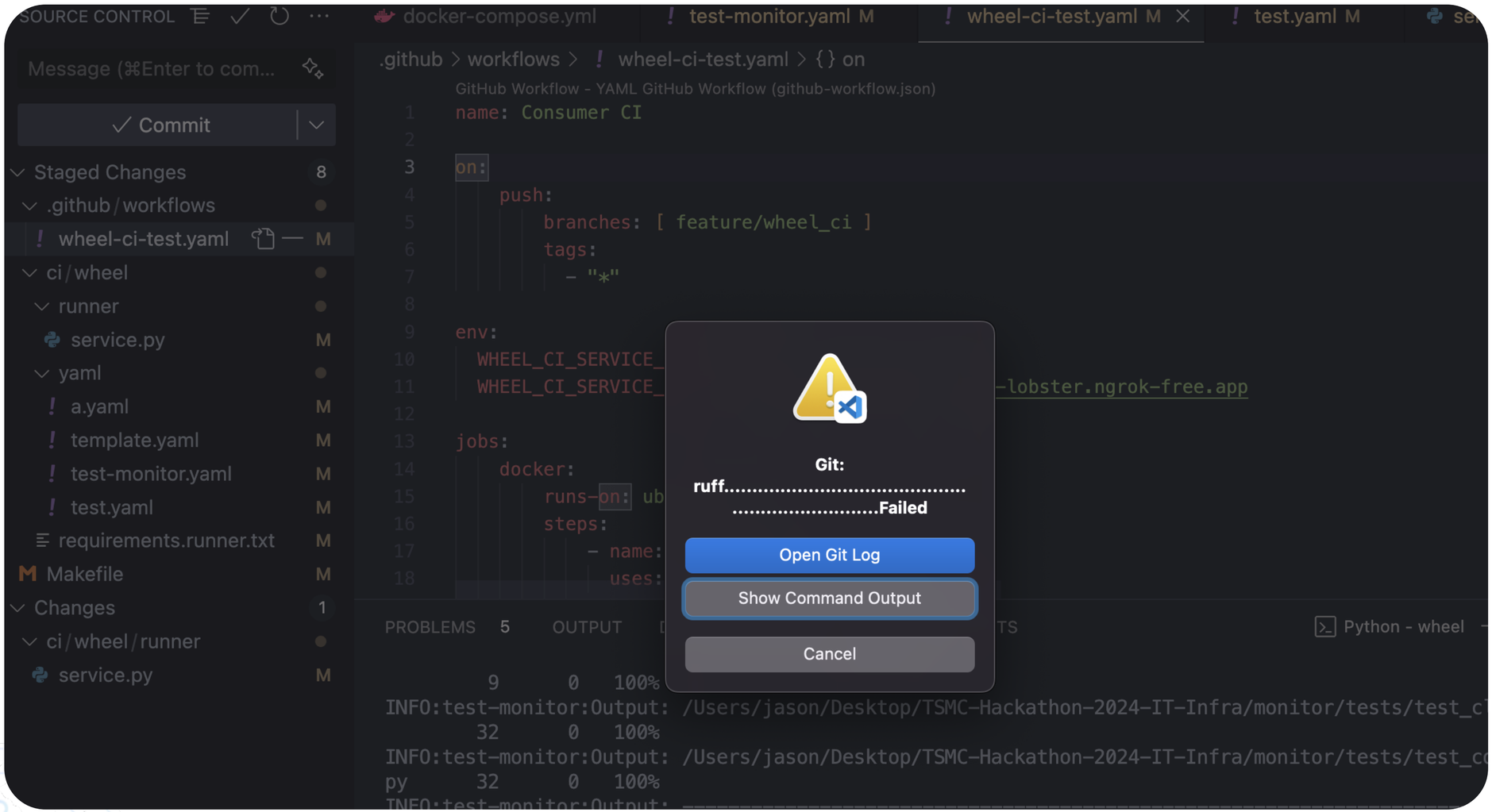
Pre Commit

Auto trigger on commit

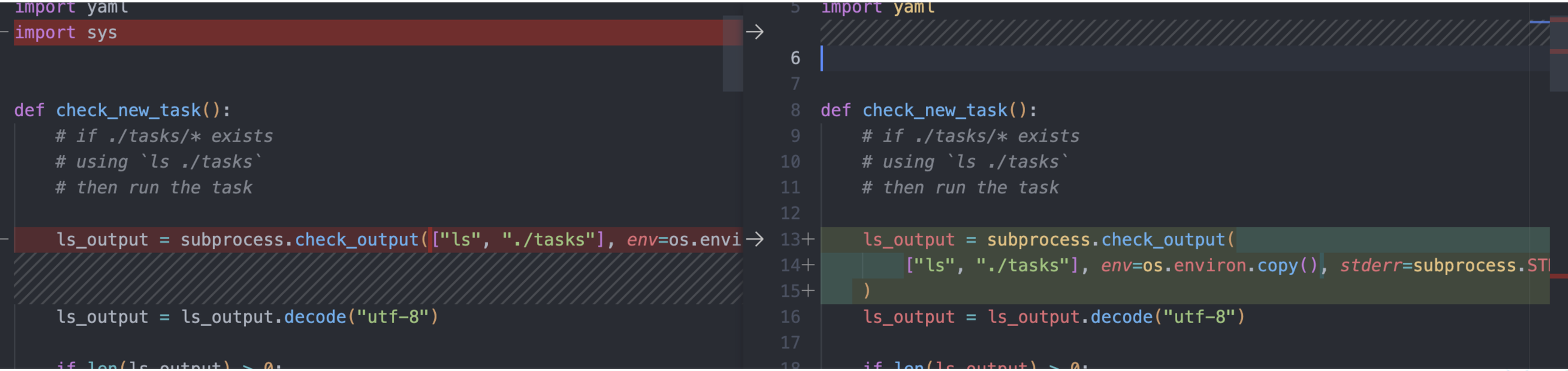
Ruff

Python linter and code formatter, written in Rust.

Linting - commit



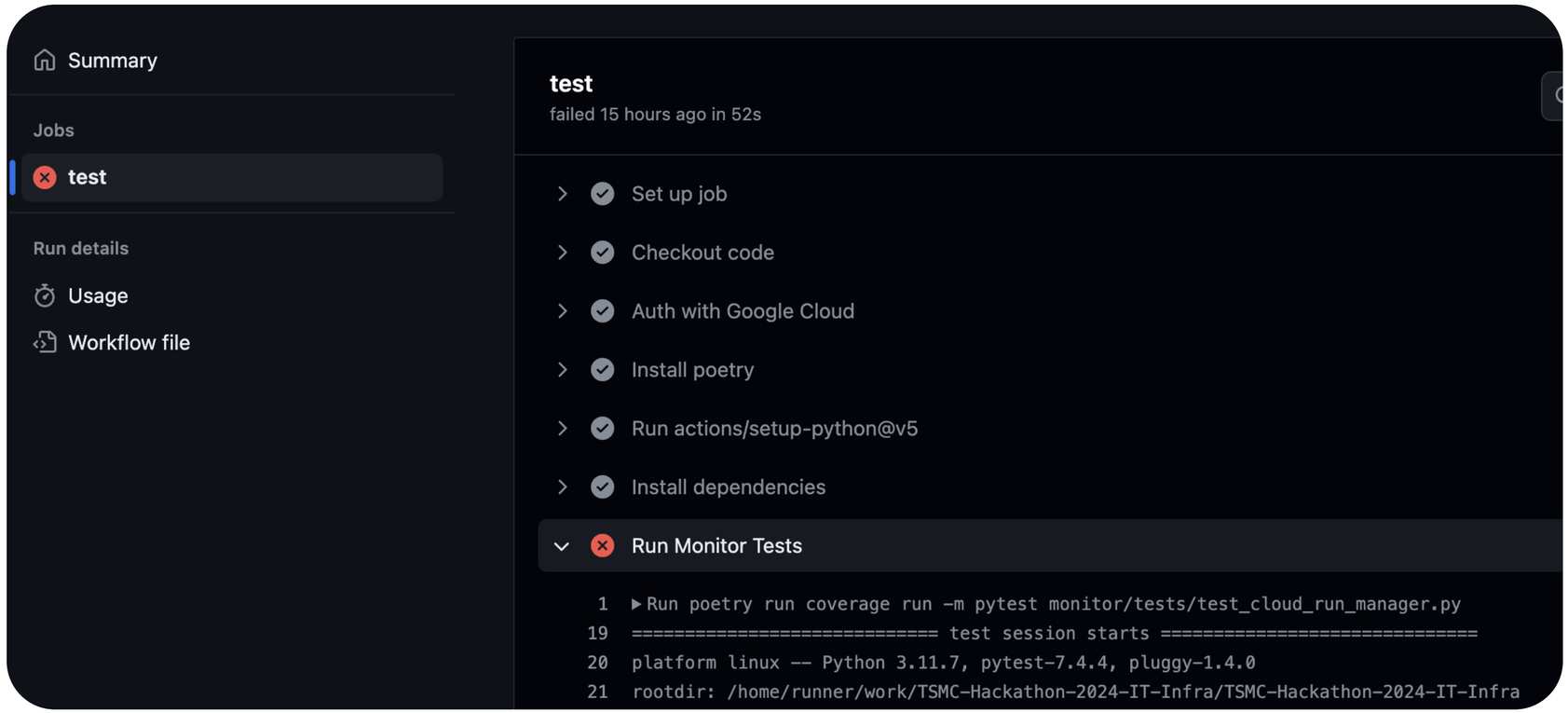
Linting - Ruff



CI/CD - Attempt

- Github Action with GCP
- Github Action with SSH
- Drone CI

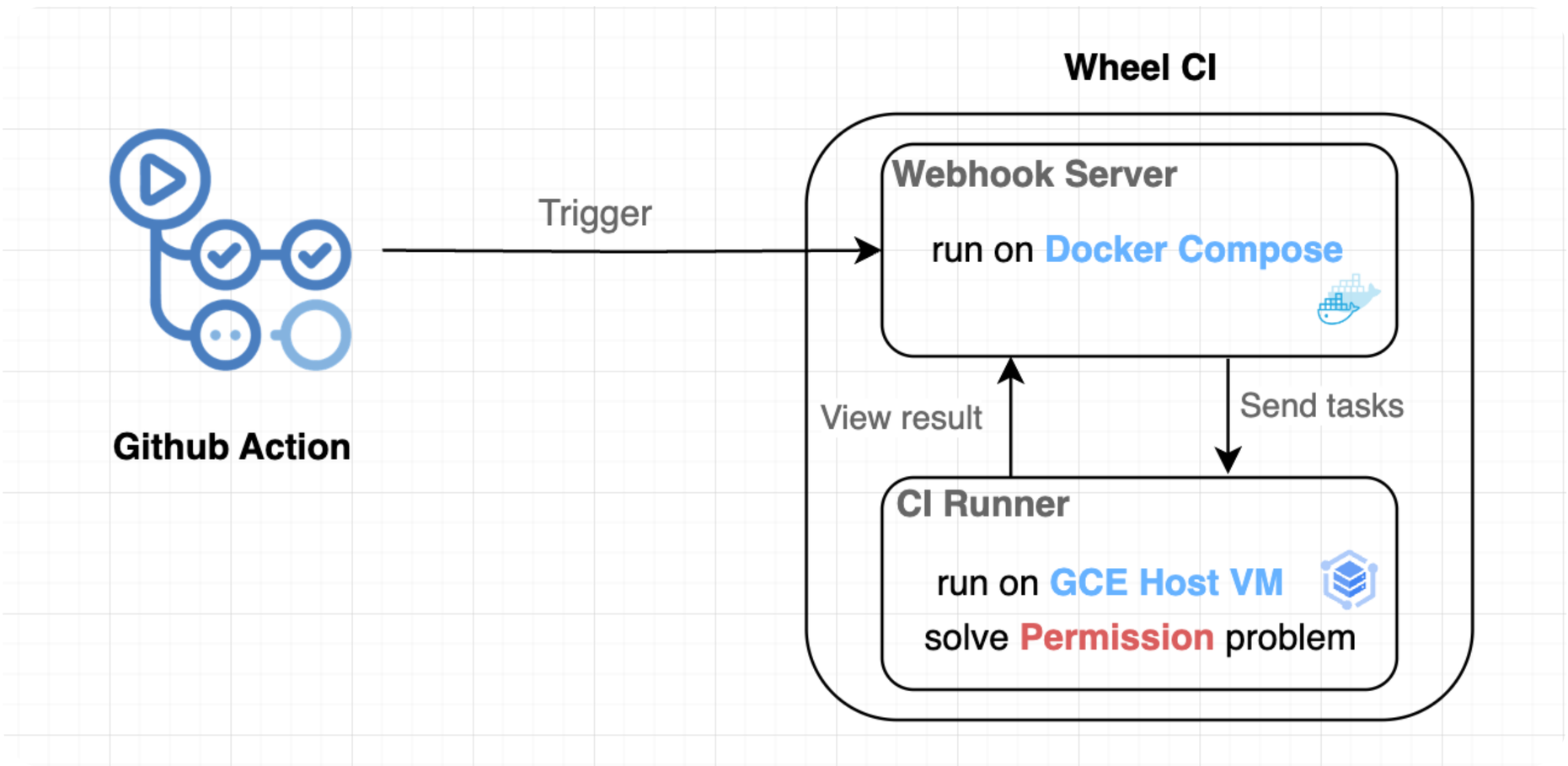
Fail: Due to GCP Permission



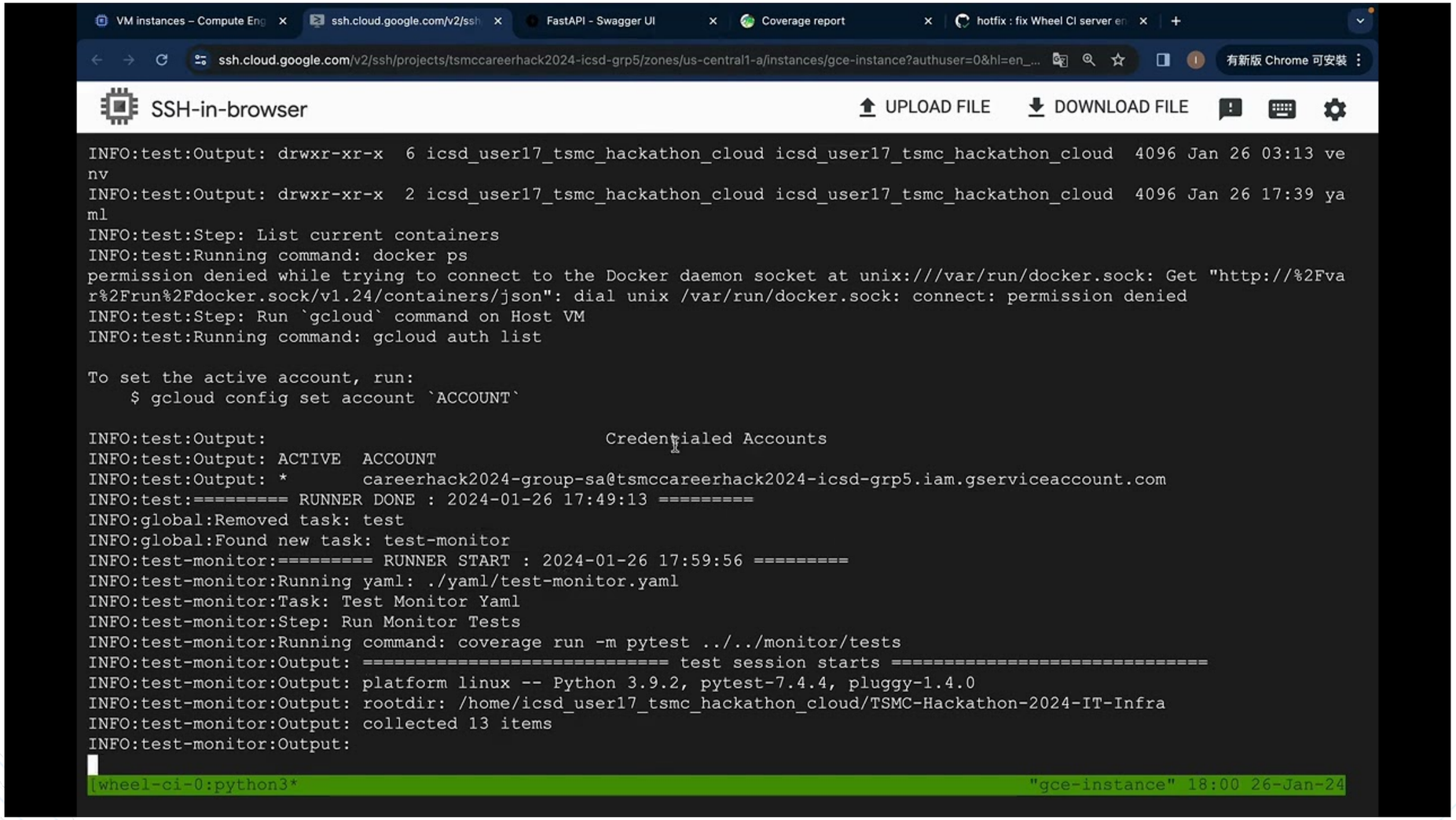

```
    grp5/locations/us-central1/services/consumer-test' (or resource may not exist).
510 FAILED monitor/tests/test_cloud_run_manager.py::test_cloud_run_manager_adjust_instance_count -
    google.api_core.exceptions.PermissionDenied: 403 Permission 'run.services.get' denied on resource 'projects/tsmccareerhack2024-icsd-
    grp5/locations/us-central1/services/consumer-test' (or resource may not exist).
511 FAILED monitor/tests/test_cloud_run_manager.py::test_cloud_run_manager_increase_instance_count -
    google.api_core.exceptions.PermissionDenied: 403 Permission 'run.services.get' denied on resource 'projects/tsmccareerhack2024-icsd-
    grp5/locations/us-central1/services/consumer-test' (or resource may not exist).
512 FAILED monitor/tests/test_cloud_run_manager.py::test_cloud_run_manager_get_metrics - google.api_core.exceptions.PermissionDenied:
    403 Permission 'run.services.get' denied on resource 'projects/tsmccareerhack2024-icsd-grp5/locations/us-central1/services/consumer-
    test' (or resource may not exist).
513 ========================= 6 failed, 1 passed in 3.11s ==========================
514 Error: Process completed with exit code 1.
```

CI/CD - Self-made CI

> build our own Wheel



CI/CD Workflow Demo



OUTLINE

User Story
System Architecture
Application Implement
Demo
DevOps
Testing
Future

Testing

- Log Analyzer
- Monitor Controller

Testing - Log Analyzer

Coverage for **ai2/log_analyzer.py**: 97%

66 statements | 64 run | 2 missing | 0 excluded

« prev ^ index » next coverage.py v7.4.0, created at 2024-01-27 11:19 +0800

```python
import pandas as pd
from langchain.output_parsers import ResponseSchema, StructuredOutputParser
from langchain.prompts import PromptTemplate
from langchain_google_vertexai import VertexAI
from pinecone import Pinecone
from vertexai.language_models import TextEmbeddingModel


class LLMLogAnalyzer:
    """
    LLMLogAnalyzer is a class that performs log analysis and provides scaling recommendations for a Google Cloud Run application.

    Args:
        pinecone_api_key (str): The API key for accessing the Pinecone service.
        index_name (str): The name of the Pinecone index.
        llm_args (dict): Additional arguments for initializing the VertexAI model.

    Attributes:
        llm (VertexAI): The VertexAI model for performing log analysis.
        db (Pinecone.Index): The Pinecone index for storing log embeddings.
        embedding_model (TextEmbeddingModel): The text embedding model for generating log embeddings.
        output_parser (StructuredOutputParser): The output parser for parsing the analysis feedback.
        format_instruction (str): The format instruction for providing scaling recommendations.
        prompt_template (PromptTemplate): The prompt template for generating analysis prompts.

    Methods:
        analyze_log: Analyzes the log data and provides scaling recommendations.
        store_memory: Stores the analysis feedback and log data in the Pinecone index.
        chat: Performs a conversation with the AI based on the stored analysis feedback.

    """

    def __init__(self, pinecone_api_key: str, index_name: str, llm_args: dict) -> None:
        """
        Initializes the LLMLogAnalyzer.

        Args:
            pinecone_api_key (str): The API key for accessing the Pinecone service.
            index_name (str): The name of the Pinecone index.
            llm_args (dict): Additional arguments for initializing the VertexAI model.
        """
        self.llm = VertexAI(**llm_args)

        pc = Pinecone(api_key=pinecone_api_key)
        self.db = pc.Index(index_name)

        self.embedding_model = TextEmbeddingModel.from_pretrained(
            "textembedding-gecko@003"
        )

        # Define output parser and generate format instruction
        severity_schema = ResponseSchema(
```

Total Coverage:

- 98%

Coverage report: 98%

coverage.py v7.4.0, created at 2024-01-27 11:19 +0800

Module	*statements*	*missing*	*excluded*	*coverage*
ai2/log_analyzer.py	66	2	0	97%
ai2/tests/__init__.py	0	0	0	100%
ai2/tests/conftest.py	15	0	0	100%
ai2/tests/log_analyzer_test.py	14	0	0	100%
Total	**95**	**2**	**0**	**98%**

coverage.py v7.4.0, created at 2024-01-27 11:19 +0800

Testing - Monitor Controller

Total Coverage:

- **82%**

Coverage report: 82%

filter...

coverage.py v7.4.0, created at 2024-01-26 18:01 +0000

Module	*statements*	*missing*	*excluded*	*coverage*
/home/icsd_user17_tsmc_hackathon_cloud/TSMC-Hackathon-2024-IT-Infra/monitor/__init__.py	0	0	0	100%
/home/icsd_user17_tsmc_hackathon_cloud/TSMC-Hackathon-2024-IT-Infra/monitor/service/__init__.py	0	0	0	100%
/home/icsd_user17_tsmc_hackathon_cloud/TSMC-Hackathon-2024-IT-Infra/monitor/service/cloudrun.py	151	45	0	70%
/home/icsd_user17_tsmc_hackathon_cloud/TSMC-Hackathon-2024-IT-Infra/monitor/service/conversation_manager.py	33	0	0	100%
/home/icsd_user17_tsmc_hackathon_cloud/TSMC-Hackathon-2024-IT-Infra/monitor/tests/__init__.py	0	0	0	100%
/home/icsd_user17_tsmc_hackathon_cloud/TSMC-Hackathon-2024-IT-Infra/monitor/tests/conftest.py	9	0	0	100%
/home/icsd_user17_tsmc_hackathon_cloud/TSMC-Hackathon-2024-IT-Infra/monitor/tests/test_cloud_run_manager.py	32	0	0	100%
/home/icsd_user17_tsmc_hackathon_cloud/TSMC-Hackathon-2024-IT-Infra/monitor/tests/test_conversation_manager.py	32	0	0	100%
Total	**257**	**45**	**0**	**82%**

coverage.py v7.4.0, created at 2024-01-26 18:01 +0000

OUTLINE

User Story

System Architecture

Application Implement

Demo

DevOps

Testing

Future

Future Improvements

GenAI

- Allow user to choose different language model
- Used advanced chain to enable additional functionalities
  - Ex. Generate code, Visualize data
- Incorporate RAG to learn from documents

Thank you for your attention