Aditya Rao - 017432454

Project Topic: A virtual assistant for developers to help them assess system/infrastructure health.

Problem Description: In a tech company developers are tasked with monitoring and maintaining certain services. Their task is to ensure that these services continue functioning without any impediments. Impediments can arise out of new deployments or changes in infrastructure. These changes via deployments occur very frequently in a rapidly growing organization. Developers usually meet every week to discuss their system health by evaluating key tech metrics like API throughput, API latency and system error rates etc. Any issues in terms of API failure rates or irregular data needs to be studied and analysed and presented before the team with a root cause analysis. This task usually takes up a lot of the developer's time. Developers usually use tools like Amazon CloudWatch, Instana or Kibana to monitor their services.

Project Proposal: A virtual assistant which uses the data of tools like Instana or Amazon CloudWatch to suggest helpful pointers for the developers to analyse the failure root cause, give system health data based on prompts, suggest possible enhancements and best practises.

Queries:

1. There was a downtime in system "X" yesterday night, generate a throughput vs time graph for all APIs under service 'X'.
2. There was a downtime in system "X" yesterday night, generate a latency vs time graph for all APIs under service 'X'.
3. There was a downtime in system "X" yesterday night, generate an error rate vs time graph for all APIs under service 'X'.
4. Print all log messages between 9:00PM to 10:PM yesterday for service "X"
5. Print all log messages between 9:00PM to 10:PM yesterday for service "X" with "500 Response code"
6. What is the throughput difference for service "X" between this and last week.
7. What is the error rate for service "X" between this and last week.
8. What is the latency difference for service "X" between this and last week.
9. What is the deployment status of service "X" now.
10. Return all services which have a very high error rate.
11. Return all services which are causing failures due to the recent deployment.
12. What is the database input output difference between this week and last week for service "X".
13. What is the cost incurred on scaling service "X" from m2.Micro system to m2.Medium, compare data from last two weeks.
14. What could be the reason for high error rates on service "X"?
15. What could be the reason for high latency issues on service "X"?
16. What could be the reason for high throughput issues on service "X"?
17. There was a production failure caused due to a recent deployment, can you find the service and root cause for the failure.
18. My system is experiencing major slowdowns what could be the possible solutions to solve it?
19. I'm expecting a major increase in user growth in the next month, what could be some of the best practises to implement.
20. I'm expecting a decline in user base and wish to deprecate some of the services, outline the steps to enable that.

Data to be used: For proof of concept, I'll be storing log data of API throughput, latency and error rates in a SQL database which the virtual assistant will be able to access.

Models to be used: Llama 3 8B (large) and Mistral 7B (small)