GKE Node Management and Infrastructure Provisioning Documentation

Overview

This documentation outlines the process and issues encountered while managing Google Kubernetes Engine (GKE) nodes to optimize billing, as well as the challenges faced during the infrastructure provisioning pipeline using Terraform. The primary focus is on resolving the cluster recreation issue to prevent potential data loss.

I. Billing Optimization with GKE Nodes

A. Initial Approach

To control costs, the initial strategy involved stopping GKE nodes during periods of inactivity to avoid extra billing.

B. Revised Approach

Upon discovering that stopping GKE nodes did not significantly impact billing, the strategy was revised. Instead of stopping nodes, the decision was made to delete and recreate GKE nodes using the following gcloud commands:

```
gcloud container node-pools delete general --cluster=disearch-cluster --zone us-central1-c --quiet
```

```
gcloud container node-pools create general --cluster disearch-cluster --num-nodes=1 --machine-type=e2-highmem-4 --labels=role=general --enable-autorepair --region us-central1-c
```

This approach effectively reduced billing without affecting the cluster's functionality.

II. Infrastructure Provisioning Pipeline

A. Issue Description

Running the infrastructure provisioning pipeline, which utilizes Terraform, resulted in unexpected behaviors:

1. **Cluster Recreation: Instead of recreating nodes, the entire GKE cluster was deleted and recreated.**
2. **Service Disruption: Services within the GKE cluster experienced sudden disruptions without any error indications. Manually changing the service's IP address temporarily**

resolved the issue.

B. Investigation Steps

1. **Cluster Recreation Issue:**
   - **The gcloud container node-pools create command was expected to recreate nodes without affecting the entire cluster.**
   - **Investigation focused on the Terraform script & Gcloud command to identify any misconfigurations or unexpected behavior triggering full cluster recreation.**
2. **Service Disruption Issue:**
   - **Analyzing service logs, network configurations, and GKE cluster settings to understand the cause of sudden disruptions.**
   - **Manual IP address changes as a workaround indicated potential issues with automatic configurations or dependencies.**

C. Current Priority

Given the severity of the potential data loss resulting from the cluster recreation issue, resolving this problem is currently the top priority.

III. Next Steps

1. **Cluster Recreation Issue:**
   - **Identify and rectify any misconfigurations causing full cluster recreation.**
   - **Conduct thorough testing of Terraform scripts to ensure proper node recreation without impacting the entire cluster.**
2. **Service Disruption Issue:**
   - **Investigate and implement a permanent solution to prevent sudden service disruptions.**
   - **Ensure that automatic configurations within the GKE cluster do not interfere with service functionality.**

IV. Conclusion

This documentation provides insights into the billing optimization strategy for GKE nodes and the challenges faced during the infrastructure provisioning pipeline. Addressing the cluster recreation issue is of utmost importance to prevent any potential data loss. Ongoing collaboration with the team and continuous testing will contribute to the successful resolution of these issues.