**Detailed Troubleshooting Report: GKE Deployment for Student Task Manager**This report provides an in-depth analysis of the technical challenges encountered during the deployment of the Student Task Manager application to Google Kubernetes Engine (GKE). It details the symptoms observed, the diagnostic steps taken, the root causes identified, and the comprehensive solutions implemented to achieve a successful deployment.

**1. Pod Stuck in Pending State Due to Insufficient CPU Resources**

**Symptoms Observed:**

* Upon applying the deployment.yaml using kubectl apply -f k8s/deployment.yaml, the command kubectl get pods consistently showed the application pod in a Pending state.
* Further investigation using kubectl describe pod <pod-name> revealed critical information in the Events section. Events with Type: Warning, Reason: FailedScheduling indicated that the Kubernetes scheduler was unable to assign the pod to any available node. The message associated with this event was specific: 0/1 nodes are available: 1 Insufficient cpu.
* An additional event, Type: Normal, Reason: NotTriggerScaleUp, with the message pod didn't trigger scale-up: 1 Insufficient cpu, suggested that even the cluster autoscaler did not deem it appropriate to add more nodes or larger nodes based on this single pod's request.

**Diagnostic Steps:**

1. Reviewed the resources.requests.cpu value in k8s/deployment.yaml. It was initially set to 250m (0.25 vCPU).
2. Examined the GKE node's capacity using kubectl get nodes and then kubectl describe node <node-name>. This showed that the node (an e2-small instance with 2 vCPUs total) had an Allocatable CPU capacity of 940m. While 250m is less than 940m, the Insufficient cpu error implied that either other system pods were consuming a portion of the allocatable CPU, or the requested amount, combined with potential CPU fragmentation, prevented scheduling.

**Root Cause Analysis:**

The primary cause was that the CPU resources requested by the pod (250m) exceeded the *effectively available and schedulable* CPU on the sole GKE node at the time of scheduling. GKE nodes reserve a portion of their total capacity for the Kubernetes control plane components (like kubelet, kube-proxy) and the underlying operating system. The remaining "allocatable" resources are available for user pods. The scheduler could not guarantee the 250m CPU requested amidst the existing load and reservations.

**Solution Implemented:**

The cpu resource request in the k8s/deployment.yaml file, under spec.template.spec.containers.resources.requests, was adjusted downwards.

* Initial change: Modified from 250m to 100m.
* Verification: The updated deployment.yaml was applied using kubectl apply -f k8s/deployment.yaml.
* Outcome: The Kubernetes scheduler successfully found sufficient resources for the 100m CPU request, and the pod transitioned from Pending to ContainerCreating and then eventually to Running (though it subsequently hit another issue). This confirmed that the resource request was the primary blocker for scheduling.

**2. Pod Crashing with CrashLoopBackOff Due to Missing Service Account Key File**

**Symptoms Observed:**

* Once the scheduling issue was resolved, kubectl get pods showed the pod transitioning to Running briefly, but then quickly changing to CrashLoopBackOff or Error. The RESTARTS count for the pod incremented repeatedly.
* The command kubectl logs <pod-name> was used to retrieve the application logs from the failing container. The logs clearly displayed a startup error: Error: Cannot find module './sit323-studenttaskmanager-43b726fc0e9a.json'. The stack trace indicated this error originated from the line in index.js where the Firebase Admin SDK was being initialized with this JSON key file.

**Diagnostic Steps:**

1. Reviewed index.js to confirm the require('./sit323-studenttaskmanager-43b726fc0e9a.json') statement.
2. Checked the Dockerfile to see how the application source code was copied (COPY . .).
3. Checked the .dockerignore file, which correctly listed sit323-studenttaskmanager-43b726fc0e9a.json. This confirmed the key file was intentionally and correctly excluded from the Docker image build process.

**Root Cause Analysis:**

The application was designed for local development to use a downloaded JSON service account key for Firebase authentication. However, for security and best practices in containerized environments (especially GKE), service account keys should not be bundled into Docker images. The .dockerignore ensured this, but the application code was not yet adapted to run in an environment (like GKE with Workload Identity) where credentials would be provided ambiently. The application crashed because it couldn't find the expected key file at runtime inside the container.

**Solution Implemented:**

A multi-step approach was taken to address this, preparing for Workload Identity:

1. **Code Modification (index.js):** The explicit loading of the JSON key file was removed. The Firebase Admin SDK initialization was changed to admin.initializeApp({ projectId: 'sit323-studenttaskmanager' });. This allows the SDK to automatically detect and use ambient credentials when available (e.g., from Workload Identity).
2. **New Docker Image (:v2):** A new version of the Docker image (student-task-manager:v2) was built using docker build -t student-task-manager:v2 . to include the updated index.js.
3. **Push to Artifact Registry:** The new :v2 image was tagged and pushed to the designated Google Artifact Registry repository.
4. **Deployment Update:** The k8s/deployment.yaml was modified to reference the new image tag (.../student-task-manager:v2).

* Outcome: After applying these changes, the pod no longer crashed due to the missing file. It reached a Running state, although subsequent Firestore operations would initially fail due to lack of authenticated identity, which was the next problem to solve.

**3. API Calls Failing with 7 PERMISSION\_DENIED: Request had insufficient authentication scopes**

**Symptoms Observed:**

* With the pod running (after fixing the missing module error), attempts to access API endpoints that interacted with Firestore (e.g., GET /tasks via Postman) resulted in a 500 Internal Server Error from the application.
* The application logs, retrieved using kubectl logs <pod-name>, showed a clear error from the Firebase Admin SDK: Error: 7 PERMISSION\_DENIED: Request had insufficient authentication scopes. The error details specified reason: 'ACCESS\_TOKEN\_SCOPE\_INSUFFICIENT' and service: 'firestore.googleapis.com'.

**Diagnostic Steps:**

1. **Workload Identity Configuration Verification:**

* Confirmed a Kubernetes Service Account (KSA), student-task-manager-ksa, was created.
* Verified the Google Service Account (GSA), student-task-manager-dev@..., had the Cloud Datastore User role in IAM.
* Ensured the gcloud iam service-accounts add-iam-policy-binding command was correctly run to link the KSA to the GSA with the roles/iam.workloadIdentityUser role.
* Confirmed the KSA was annotated using kubectl annotate serviceaccount ... iam.gke.io/gcp-service-account=YOUR\_GSA\_EMAIL.
* Checked that k8s/deployment.yaml specified serviceAccountName: student-task-manager-ksa for the pod template.

All these steps for Workload Identity setup appeared correct.

1. **IAM Permission vs. OAuth Scopes:** The error message specifically mentioned "scopes," not just general permissions. This indicated that while the GSA *had permission* to access Firestore, the access token being used by the application (obtained via Workload Identity acting as the GSA) was not minted with the *scope* required for the firestore.googleapis.com service.

**Root Cause Analysis:**

GKE nodes are provisioned with a set of default OAuth 2.0 scopes that their node service account can use when generating access tokens for Google Cloud APIs. If these default scopes are too restrictive and do not include https://www.googleapis.com/auth/datastore (or the broader https://www.googleapis.com/auth/cloud-platform), then even if Workload Identity correctly impersonates a GSA with the right IAM roles, the access tokens generated might still lack the required scope for certain services. This was the case: the original node pool was created without explicitly providing sufficient scopes.

**Solution Implemented:**The GKE node pool (default-pool) was recreated with the necessary OAuth scopes:

1. **Set gcloud context:** Project, zone, and cluster were set as defaults for subsequent gcloud commands.
2. **Node Pool Deletion:** The existing default-pool was deleted using gcloud container node-pools delete default-pool --quiet. This temporarily removed the compute capacity for the application.
3. **Node Pool Recreation with Scopes:** A new default-pool was created with the same name but now including the --scopes "https://www.googleapis.com/auth/cloud-platform" flag. This flag ensures the node's service account can request tokens with all necessary scopes for Google Cloud services, including Firestore. The machine type was also upgraded from e2-small to e2-medium during this recreation for better resource availability.

A blue screen with text

AI-generated content may be incorrect.

Verification: After the new node pool became active, the studenttaskmanager-deployment automatically scheduled its pod onto the new node.

Outcome: API calls made to the application's external IP address via Postman (e.g., GET /tasks) now returned a 200 OK status with data successfully retrieved from Firestore. The "insufficient authentication scopes" error was resolved, confirming that the broader scopes on the node pool enabled Workload Identity to function correctly for Firestore access.