Journal:

**Day 1: Foundational Flask Development and Local Database Integration**The primary goal of Day 1 was to build the core local functionality of the Flask web application, including user-facing forms and backend API endpoints capable of interacting with a PostgreSQL database running locally in Docker.

* **Task 1: Enhance Flask Frontend with HTML/CSS**
* **Actions:**

1. **Created templates/ and static/ directories:** Standard Flask convention for organizing HTML templates and static assets (CSS, JavaScript, images) respectively. This promotes a clean project structure.
2. **Developed templates/book.html:** This HTML file provides the user interface for booking an appointment. It includes a form with fields for patient name, doctor selection, and appointment time.

* *Reason:* To allow users to visually interact with the system and submit appointment requests.

1. **Created static/style.css:** Basic CSS rules were added to improve the visual appearance and usability of the book.html form.

* *Reason:* To make the user interface more presentable and user-friendly.

1. **Updated app.py to render book.html:** The /book route in app.py was modified to use Flask's render\_template('book.html') function.

* *Reason:* To connect the URL endpoint (/book) to the actual HTML page, so users visiting that URL see the booking form.

1. **Initial Testing and Troubleshooting:**

* Addressed a ModuleNotFoundError: No module named 'flask'. This occurred because the Flask library was not installed in the Python environment where app.py was being run.
* *Reason & Solution:* Python couldn't locate the Flask package. Resolved by installing Flask using pip install Flask.
* Added print statements to app.py to display accessible URLs in the terminal upon startup.
* *Reason:* User request for clearer feedback during development, making it easier to know which URLs to access for testing.
* **Task 2: Develop Backend API Endpoints & Local Database Setup**
* **Actions:**

1. **Installed psycopg2-binary:** This package is the PostgreSQL database adapter for Python.

* *Reason:* To enable the Flask application to communicate with (send queries to and receive results from) the PostgreSQL database.

1. **Configured Database Connection in app.py:**

* Connection parameters (DB\_NAME, DB\_USER, DB\_PASS, DB\_HOST, DB\_PORT) were defined. The password was initially read from db\_password.txt to avoid hardcoding it directly in the script.
* *Reason:* To provide the necessary credentials and location details for psycopg2 to connect to the database.
* A get\_db\_connection() helper function was created to encapsulate the logic for establishing a database connection.
* *Reason:* To promote reusable code and centralize connection management.

1. **Developed API Endpoints in app.py:**

* **POST /book-appointment:** This endpoint handles the submission of the booking form. It receives form data, connects to the PostgreSQL database, and executes an SQL INSERT statement to store the new appointment.
* *Reason:* To provide the core functionality of creating and saving new appointments in the system.
* **GET /appointments:** This endpoint retrieves all existing appointments from the database using an SQL SELECT statement and returns them, typically as JSON.
* *Reason:* To allow for viewing of booked appointments, useful for administration or patient history features.

1. **PostgreSQL Database Setup and Troubleshooting:**

* A postgres Docker container (named postgres-db) was used to host the database. A db\_setup.sql file was created to define the healthcare database and the appointments table structure.
* **Major Challenge: FATAL: database "healthcare" does not exist:** This was a persistent issue. The Flask application (running directly on the host at this stage) could not connect to the healthcare database within the postgres-db Docker container, even though the container itself was running.
* *Root Cause Investigation:* Extensive troubleshooting involved checking container logs, app.py connection strings, and the database initialization process. The key issue was related to **Docker volume persistence and initialization**. The PostgreSQL Docker image has a mechanism to run initialization scripts (like creating a database specified by POSTGRES\_DB or running scripts in /docker-entrypoint-initdb.d/) *only when a data volume is initialized for the first time*. If the named volume (pgdata\_healthcarebooking) existed from previous (potentially failed) attempts and wasn't correctly initialized with the healthcare database, subsequent container starts wouldn't recreate it if the image logic deemed the volume already "initialized."
* *Resolution Steps:*

1. The postgres-db container was stopped and removed.
2. Crucially, the Docker named volume (pgdata\_healthcarebooking) was explicitly removed (docker volume rm pgdata\_healthcarebooking) to ensure a completely fresh start.
3. The postgres-db container was restarted with the -e POSTGRES\_DB=healthcare environment variable (which instructs the official image to create this database if it doesn't exist on fresh volume initialization) and the volume mount.
4. After confirming the healthcare database was created, the db\_setup.sql script (which creates the appointments table) was manually executed within the container using psql.
5. The DB\_HOST in app.py was confirmed to be 127.0.0.1 (or localhost) because, at this point, the Flask app was running directly on the host, and the postgres-db container had its port 5432 mapped to the host.

* Successful booking and retrieval of appointments via Postman confirmed the fix.

**Day 2: Containerization and Kubernetes Deployment for Flask**The focus of Day 2 was to containerize the now-functional Flask application, deploy it to a local Kubernetes cluster (via Docker Desktop), and implement autoscaling.

* **Task 1: Create Dockerfile and Docker Image for Flask App**
* **Actions:**

1. **Created deployment/Dockerfile:** This file contains instructions to build a Docker image for the Flask application. It specifies the base Python image, sets up the working directory, copies application code and dependencies, installs Python packages from requirements.txt, exposes port 5000, and defines the command to run the Flask app.

* *Reason:* To package the Flask application and all its dependencies into a standardized, portable, and reproducible unit (a Docker image).

1. **Created deployment/requirements.txt:** This file lists the Python dependencies (Flask, psycopg2-binary).

* *Reason:* To explicitly declare the application's Python package dependencies, ensuring they are installed consistently when the Docker image is built.

1. **Built the Docker Image:** The command docker build -t healthcarebooking\_app:v1 -f deployment/Dockerfile . was used.

* *Reason:* To create the runnable Docker image named healthcarebooking\_app with tag v1 from the Dockerfile specifications.

1. **Ran the Containerized Flask App and Tested DB Connection:** The Flask app container was run.

* **Networking Challenge: DB\_HOST for Container-to-Container:** When the Flask app ran *inside its own container*, DB\_HOST = "127.0.0.1" in app.py failed because 127.0.0.1 now referred to the Flask container itself.
* *Reason & Solution:* Containers on the same Docker network can resolve each other by their container names. DB\_HOST in app.py was changed to "postgres-db" (the name of the running PostgreSQL container). The Flask app Docker image was rebuilt with this change.
* Successful booking confirmed the Flask container could connect to the postgres-db container.
* **Task 2: Create Kubernetes YAML Files and Deploy Flask App**
* **Actions:**

1. **Pushed Docker Image to Docker Hub:** docker push hayden2310/healthcarebooking\_app:v1.

* *Reason:* To make the application image accessible to the Kubernetes cluster. Kubernetes pulls images from a registry (like Docker Hub) to run them.

1. **Created deployment/flask-deployment.yaml:** This Kubernetes manifest defines a Deployment object.

* *Reason:* To declare the desired state of the Flask application in Kubernetes, including which Docker image to use, the number of replicas (pods), port configurations, labels for organization, environment variables, resource requests/limits for scheduling, and readiness/liveness probes for health checking.

1. **Created deployment/flask-service.yaml:** This manifest defines a Service object of type ClusterIP.

* *Reason:* To provide a stable internal network endpoint (a consistent IP address and DNS name within the cluster) to access the Flask application pods. The Service load balances traffic across the pods managed by the Deployment.

1. **Applied YAMLs to Kubernetes:** kubectl apply -f ....

* *Reason:* To instruct Kubernetes to create or update resources as defined in the YAML files.

1. **Verified and Accessed:** Used kubectl get pods,svc,deployment to check status and kubectl port-forward service/flask-app-service 5000:5000 to access the application running in Kubernetes from the local machine.

* **Task 3: Configure Horizontal Pod Autoscaling (HPA)**
* **Actions:**

1. **Created deployment/flask-hpa.yaml:** This manifest defines a HorizontalPodAutoscaler object.

* *Reason:* To enable Kubernetes to automatically increase or decrease the number of Flask application pods based on observed CPU utilization (or other metrics), allowing the application to handle variable loads efficiently.

1. **Applied HPA YAML.**
2. **Troubleshooting HPA:**

* **Issue 1: HPA Metrics <unknown>/50%:** The HPA couldn't fetch CPU utilization metrics.
* *Reason & Solution:* The pods targeted by the HPA (defined in flask-deployment.yaml) must have CPU resource requests specified in their container definition. Without this, Kubernetes cannot calculate CPU utilization as a percentage of the request. Added resources: {requests: {cpu: "..."}} to the deployment.
* **Issue 2: Pods Pending due to Insufficient cpu during HPA Scale-Up:** When HPA tried to create new pods, they couldn't be scheduled.
* *Reason & Solution:* The sum of CPU requests from all pods (existing + new ones HPA wanted to create) exceeded the allocatable CPU of the single node in Docker Desktop. Resolved by:

1. Reducing the requests.cpu per pod in flask-deployment.yaml (e.g., to 50m).
2. Adjusting the replicas in flask-deployment.yaml to 1 (initial state before HPA scales).
3. Setting minReplicas: 1 and maxReplicas: 3 (a realistic maximum for the local node) in flask-hpa.yaml.
4. **Load Testing and Observation:** Generated HTTP traffic to the Flask service and used kubectl get hpa -w and kubectl get pods -w to observe the HPA scaling the deployment up to 3 pods under load, and then scaling back down to 1 pod when the load was removed. This confirmed the HPA was working correctly.

“flash\_booking\_form.png”

A screenshot of a computer

AI-generated content may be incorrect.

“api\_post.png”

A screenshot of a computer

AI-generated content may be incorrect.

“api\_get.png”

A screenshot of a computer

AI-generated content may be incorrect.

“flask\_k8s.png”

A screenshot of a computer

AI-generated content may be incorrect.

“hpa\_scaling.png”

A screenshot of a computer

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

Encountered errors:

A/ 2 Postgresql servers are running at the same time:

have **TWO** PostgreSQL servers running:

1. **The Docker Container (postgres-db):** This is the one running Debian, where you've been using docker exec to create the healthcare database. Its logs showed version 17.4 earlier.
2. **A Native Windows PostgreSQL Installation:** This is running directly on your Windows host, was compiled with Visual C++, and is version 16.2. This is the server that your test\_db.py and Flask app are actually connecting to on 127.0.0.1:5432. This native Windows PostgreSQL instance *does not* have the healthcare database.

When you run python app.py or python test\_db.py on your Windows machine and connect to 127.0.0.1:5432 (or localhost:5432), it's connecting to the PostgreSQL server that "answers" on that port *first* from the perspective of your Windows host. It seems your native Windows PostgreSQL installation is taking precedence or is the one configured to listen on that specific IP/port combination in a way that psycopg2 resolves to it.

Solutions:

* Open "Services" in Windows (search for services.msc).
* Look for a service related to PostgreSQL (e.g., postgresql-x64-16)
* Stop the Native Windows PostgreSQL Service.
  + In the Services window, right-click the PostgreSQL service and select "Stop".
  + To prevent it from starting on boot, right-click -> Properties -> Startup type: "Manual" or "Disabled".

**B/ Inter-Container Network Communication in Docker Environment**

* **Issue Encountered:** An initial challenge arose when containerizing the Flask application and its PostgreSQL database. The Flask application, running in its Docker container, was unable to establish a connection with the PostgreSQL container. The application's database host configuration was initially set to localhost or 127.0.0.1, leading to connection errors such as psycopg2.OperationalError: connection refused.
* **Root Cause Analysis:** This issue stemmed from the network isolation inherent in Docker containers. Within a container's isolated network namespace, localhost refers to the container itself, not the host machine or other sibling containers. Consequently, the Flask application was attempting to connect to a PostgreSQL instance it presumed to be within its own network boundaries, which was incorrect.
* **Resolution Implemented:** The problem was rectified by leveraging Docker's internal DNS resolution capabilities for containers sharing a common Docker network. The database host (DB\_HOST) parameter within the Flask application's configuration (app.py) was modified from localhost to the service name of the PostgreSQL container (i.e., postgres-db). This change enabled the Flask container to correctly resolve and connect to the PostgreSQL container using its designated service name, thereby establishing successful database communication. In addition, a network called healthcare-network through command: docker network create healthcare-network.

**C/ Kubernetes Pod Scheduling and Resource Allocation for Horizontal Pod Autoscaling (HPA)**

* **Issue Encountered:** Upon deploying the Flask application to the Kubernetes cluster (via Docker Desktop), difficulties were observed with pod scheduling and the functionality of the Horizontal Pod Autoscaler (HPA). New pods, whether created by the initial deployment or by HPA scaling events, frequently entered a Pending state. Examination of pod events using kubectl describe pod revealed Warning FailedScheduling messages, citing "Insufficient CPU" as the cause, even when the node appeared to have available capacity. Furthermore, the HPA initially displayed <unknown> for CPU utilization targets, indicating an inability to retrieve necessary metrics.
* **Root Cause Analysis:** These issues were traced to aspects of Kubernetes' resource management and HPA configuration:
* **Resource Requests and Limits:** Kubernetes schedules pods based on their declared resource requests (e.g., CPU, memory). If the sum of the requests of all pods on a node (or attempting to be scheduled on it) exceeded the node's allocatable capacity, new pods could not be scheduled. The initial CPU request per pod might have been set too high for the single-node Docker Desktop Kubernetes environment, particularly when HPA attempted to scale out.
* **HPA Metrics Dependency:** For the HPA to effectively monitor and scale based on CPU utilization, the target pods' container specifications must include defined requests for CPU. An absence or misconfiguration of resources.requests.cpu in the Deployment YAML prevented the Metrics Server from reporting, and thus the HPA from acting upon, the pods' CPU usage.
* **Resolution Implemented:** A multi-faceted approach was adopted to resolve these Kubernetes-specific challenges:
* **Resource Request Optimization:** The cpu and memory requests within the flask-deployment.yaml for the application container were carefully reviewed and adjusted to more modest values (e.g., requests.cpu reduced to 50m). This ensured that the cumulative resource requests, even during HPA scale-out events, remained within the allocatable capacity of the local Kubernetes node.
* **HPA Configuration Tuning:** The flask-hpa.yaml was modified to set a minReplicas of 1 and a maxReplicas of 3, aligning the scaling behavior with the resource capacity of the test environment. The initial replicas in the Deployment was also set to 1. These changes provided a more stable baseline and prevented premature resource exhaustion during scaling.
* **Ensuring Metric Availability:** The explicit definition of resources.requests.cpu in the Deployment manifest was confirmed, enabling the Kubernetes Metrics Server to collect and provide the necessary data for the HPA to function correctly, which was subsequently reflected in the HPA status showing actual CPU utilization percentages.