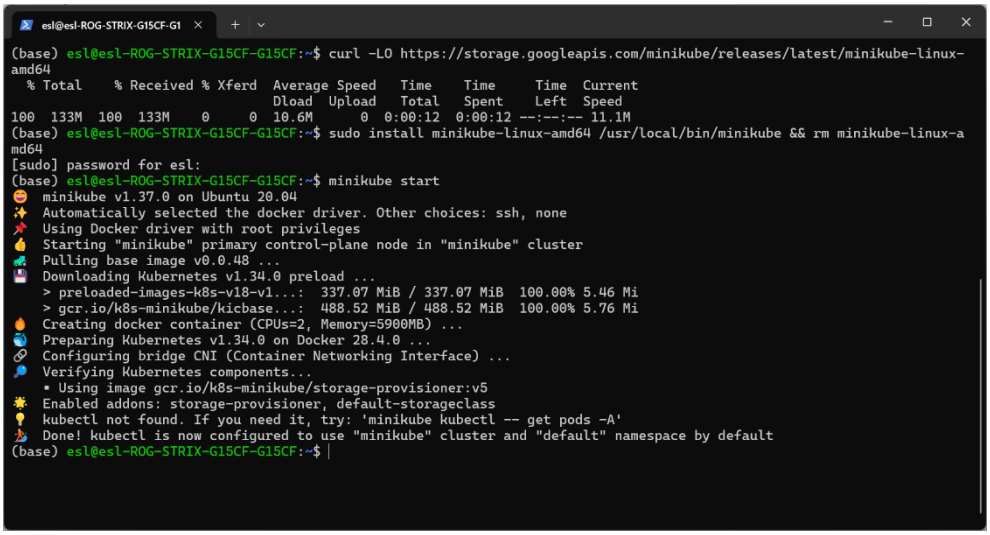
開源軟體開發　MLOps 作業

姓名：張健勳　　學號：614410073

1. 基礎環境建置與檢測

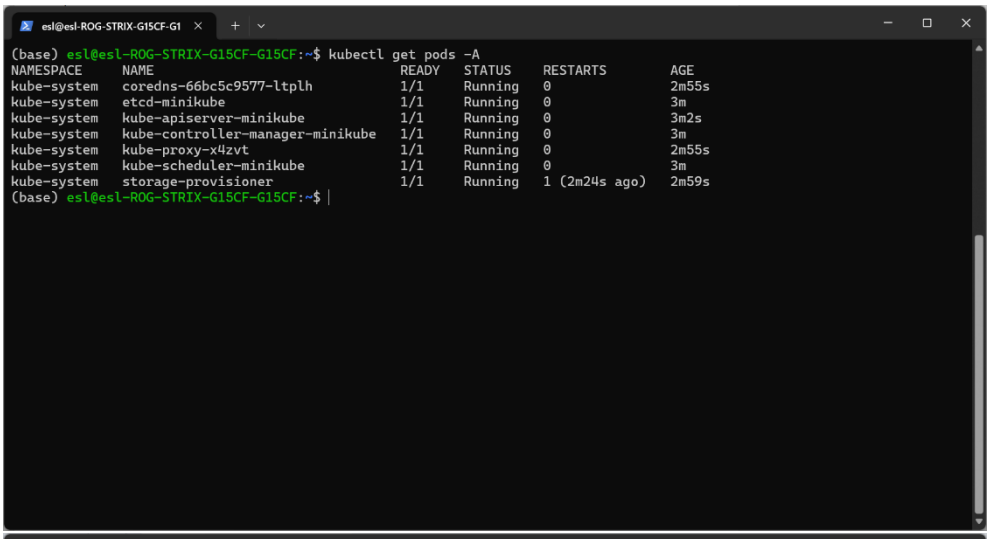
首先啟動本地 Kubernetes 叢集。使用 minikube start 成功初始化控制平面（Control Plane），並下載基礎映像檔。隨後透過 kubectl get pods -A 檢查系統元件，確認 etcd、kube-apiserver、coredns 等核心服務皆處於 Running 狀態，叢集運作正常。

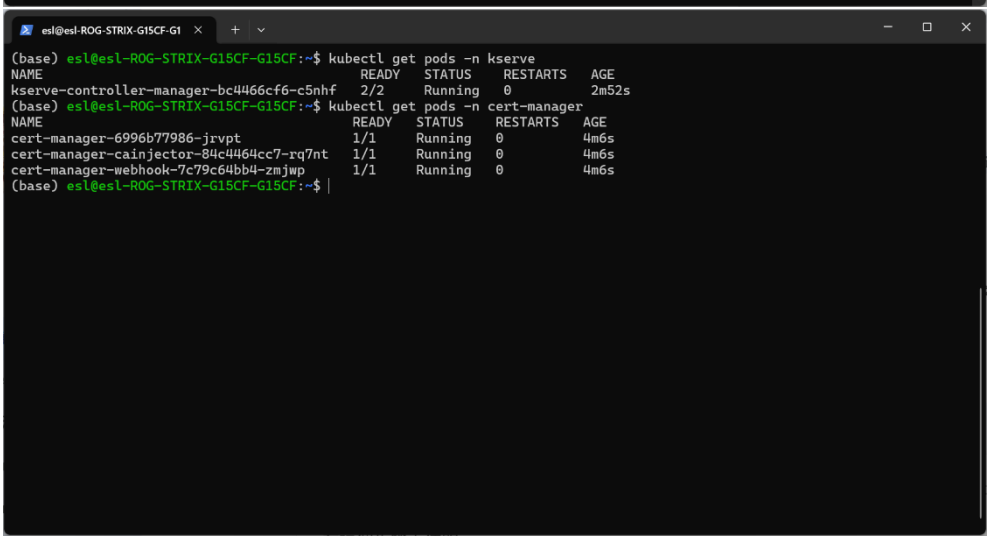


2. KServe 服務元件檢查

在部署模型前，驗證 KServe 及其依賴元件的狀態。

- KServe Controller: 確認 kserve-controller-manager 處於 Running 且狀態為 2/2，表示控制器已就緒。
- Cert-Manager: 檢查 cert-manager、cainjector 與 webhook 等 Pod 皆正常運作，確保 Webhook 憑證管理功能無誤。

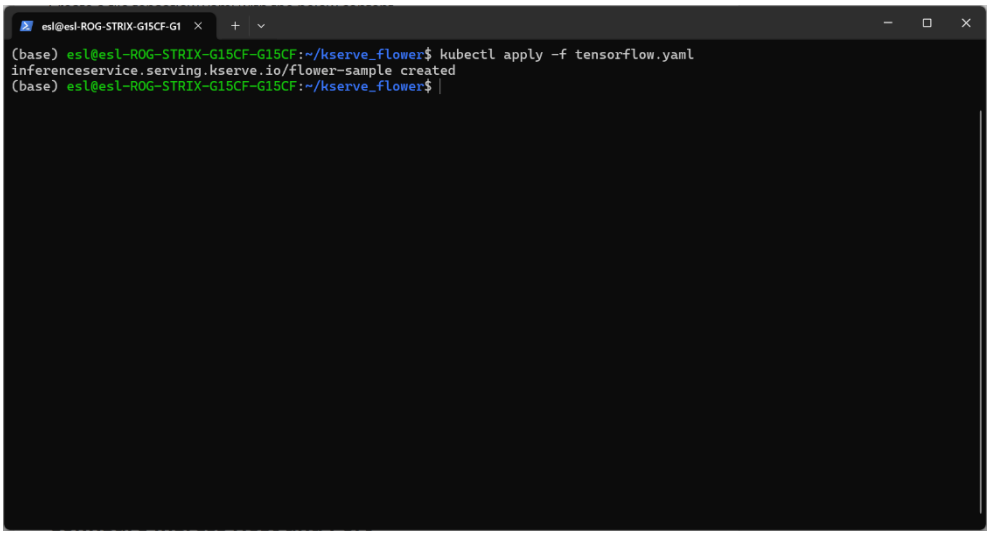




3. 模型部署 (InferenceService)

使用定義好的 YAML 設定檔部署模型：

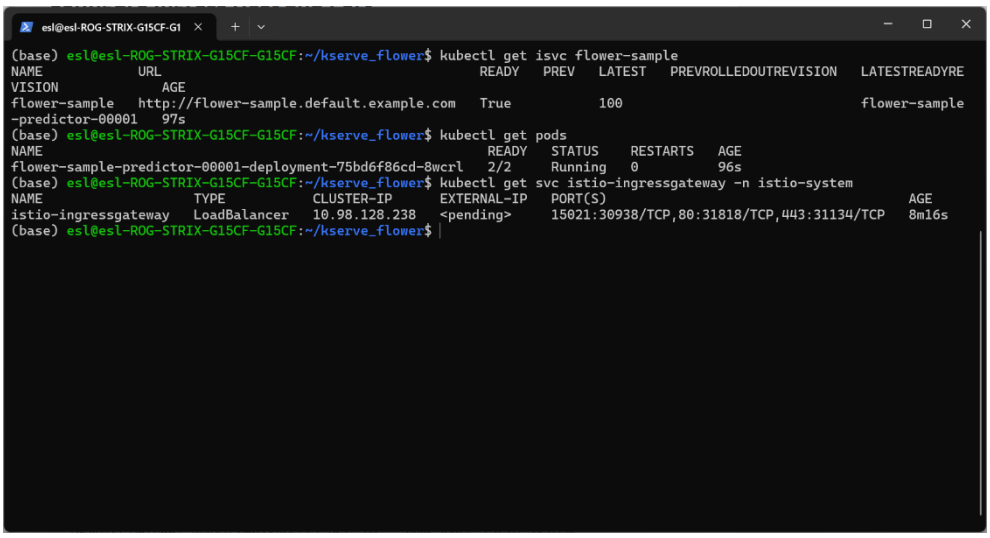
- 執行 kubectl apply -f tensorflow.yaml 建立資源，系統回傳 inferenceservice.serving.kserve.io/flower-sample created。
- 透過 kubectl get isvc flower-sample 監控部署進度。初期狀態可能顯示 Unknown，待模型容器下載並初始化完成後，READY 欄位轉為 True，且 URL 已生成 (http://flower-sample.default.example.com)。
- 同步確認 Pod 狀態，flower-sample-predictor 顯示 2/2 Running，代表模型伺服器已上線。



4. 網路配置與端口轉發 (Port-Forwarding)

為了從本地端存取叢集內的模型服務，進行以下網路設定：

- 獲取 Istio Ingress Gateway 的服務名稱與端口資訊。
- 設定環境變數 INGRESS_HOST 與 INGRESS_PORT。
- 執行 kubectl port-forward 將本地的 8080 端口映射至 Istio Ingress Gateway 的 80 端口，建立連線通道。



5. 推論測試 (Inference Testing)

最後進行實際的 API 請求測試：

- 準備包含圖片數據的 input.json。
- 使用 curl 發送 POST 請求至模型預測端點 (/v1/models/flower-sample:predict)。

- 測試結果： 伺服器回傳 HTTP/1.1 200 OK，表示請求成功。回傳的 JSON 數據中包含 predictions 物件，顯示分類預測值為 0，信心分數（Score）高達 0.9991。