W255 Final Project

Sudhrity Mondal

Class: W255-2 Saturday 8:00 AM

Semester: Spring 2022

Professor: James Winegar

Deploy and conduct performance tests on a Pytorch API application in Azure Kubernetes Service

1 Objective

The goal of `project` is to deploy the API application built throughout the semester on `Azure Kubernetes Service (AKS)` with the following enhancements

- Package up an NLP model (DistilBERT) for running efficient CPU-based inferencing for POSITIVE/NEGATIVE sentiment
- Have results be cached to protect endpoint from abuse
- Use `grafana` to understand the dynamics of the system.
- Leverage `k6` to load test the `/predict` endpoint
- Leverage `pytest` to ensure that the application works correctly prior to deployment on `AKS`.
- Leverage `poetry` to manage the runtime environment in a portable way.
- Leverage `Docker` to package applications in a reuseable fashion

2 Environment information

The environment used to develop and run initial load tests is a **VMware Workstation VM** running **Ubuntu 20.04** with 8 cores, 200GB disk and 32GB RAM. The final load test was executed on a **Ubuntu VM on AWS**.

Project file directory structure

The folder structure and files for the project are shown below under **project** directory

```
ubuntu@ubuntu:~/w255$ tree spring22-sudhrity/project/
spring22-sudhrity/project/
├── image0-1.png
├── image0-2.png
├── image1.png
├── image2.png
├── image3.png
├── image4.png
├── load.js
├── mlapi
│   ├── distilbert-base-uncased-finetuned-sst2
│   │   ├── config.json
│   │   ├── pytorch_model.bin
│   │   ├── README.md
│   │   ├── special_tokens_map.json
│   │   ├── tokenizer_config.json
│   │   ├── tokenizer.json
│   │   ├── training_args.bin
│   │   └── vocab.txt
│   ├── docker-compose.yml
│   ├── Dockerfile
│   ├── mlapi
│   │   ├── example.py
│   │   ├── __init__.py
│   │   ├── main.py
│   │   └── __pycache__
│   │       ├── __init__.cpython-310.pyc
│   │       └── main.cpython-310.pyc
│   ├── model_pipeline.pkl
│   ├── poetry.lock
│   ├── pyproject.toml
│   ├── README.rst
│   ├── run_prod.sh
│   ├── run.sh
│   ├── tests
│   │   ├── __init__.py
│   │   ├── __pycache__
│   │   │   ├── __init__.cpython-310.pyc
│   │   │   └── test_mlapi.cpython-310-pytest-7.1.1.pyc
│   │   └── test_mlapi.py
│   └── train.py
├── README.md
├── run_dev.sh
├── run_k6.sh
└── run_prod.sh
```

git-lfs installation

```
sudo apt-get update -y
sudo apt-get install -y git-lfs
```

Project files

- **README.md** - This file containing description and tasks of the lab and results from the performance tests
- **run_prod.sh** - Execute build and deploy to AKS
- **run_dev.sh** - Execute build and deploy to local minikube environment
- **run_k6.sh** - Bash script to run load test using k6 script.
- ***.png** - Screenshot files.
- **load.js** - K6 load test script

3 Build, deploy in Minikube and execute tests in Dev environment

Below is a copy of the *run_dev.sh*

```
#!/bin/bash
APP_NAME=mlapi
IMAGE_NAME=project
NAMESPACE=sudhrity

kubectl config use-context minikube

# minikube
minikube start --kubernetes-version=v1.21.7 --extra-config=apiserver.service-node-port-range=1-65535
APP_HOST=`minikube ip`

kubectl config use-context minikube -n sudhrity
kubectl delete -k ${APP_NAME}/.k8s/overlays/dev

kubectl delete service project-service -n sudhrity

eval $(minikube -p minikube docker-env)

docker rmi -f ${NAMESPACE}/${IMAGE_NAME}

docker build --no-cache -t ${NAMESPACE}/${IMAGE_NAME} ./${APP_NAME}/

kubectl kustomize ${APP_NAME}/.k8s/overlays/dev
kubectl apply -k ${APP_NAME}/.k8s/overlays/dev

sleep 60

kubectl get all -n sudhrity

kubectl expose deployment project --type=LoadBalancer --name=project-service -n sudhrity

#APP_URL=`minikube service list | grep http | awk -F'|' '{print $5}'`

APP_PORT=$(kubectl get service project-service -n sudhrity --output json | jq '.spec.ports[0].nodePort')
```

```
sleep 20

curl -X 'POST' \
 "${APP_HOST}:${APP_PORT}/predict" \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{"text": ["I hate you.", "I love you."]}'

echo

curl -X 'GET' \
 "${APP_HOST}:${APP_PORT}/predict" \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{"text": ["I love you."]}'

echo
```

The execution results of *run_dev.sh* on the dev environment is shown below:

```
...
apiVersion: autoscaling/v1
kind: HorizontalPodAutoscaler
metadata:
  name: project
  namespace: sudhrity
spec:
  maxReplicas: 10
  minReplicas: 1
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: project
  targetCPUUtilizationPercentage: 50
configmap/redis created
service/project created
service/redis created
Warning: Detected changes to resource local-pv1 which is currently being
deleted.
persistentvolume/local-pv1 unchanged
deployment.apps/project created
statefulset.apps/redis created
horizontalpodautoscaler.autoscaling/project created
NAME                           READY   STATUS    RESTARTS   AGE
pod/project-766c757bd6-hrtgz   1/1     Running   0          60s
pod/redis-0                    1/1     Running   0          60s

NAME              TYPE        CLUSTER-IP    EXTERNAL-IP   PORT(S)    AGE
service/project   ClusterIP   10.97.99.44   <none>        8000/TCP   60s
service/redis     ClusterIP   None          <none>        6379/TCP   60s

NAME                      READY   UP-TO-DATE   AVAILABLE   AGE
```

```
deployment.apps/project   1/1     1            1           60s

NAME                                 DESIRED   CURRENT   READY   AGE
replicaset.apps/project-766c757bd6   1         1         1       60s

NAME                     READY   AGE
statefulset.apps/redis   1/1     60s

NAME                                          REFERENCE            TARGETS
  MINPODS   MAXPODS   REPLICAS   AGE
horizontalpodautoscaler.autoscaling/project   Deployment/project   <unknown>/50%
  1         10        1          60s
service/project-service exposed
{"predictions": [[{"label": "NEGATIVE", "score": 0.8838168382644653}, {"label":
"POSITIVE", "score": 0.11618312448263168}], [{"label": "NEGATIVE", "score":
0.003921119030565023}, {"label": "POSITIVE", "score": 0.9960789084434509}]]}
{"predictions": [[{"label": "NEGATIVE", "score": 0.003921119030565023},
{"label": "POSITIVE", "score": 0.9960789084434509}]]}
```

3.1 Pytest execution

The test script (*test_mlapi.py*) used to execute pytest and the test results are shown below:

```
from fastapi.testclient import TestClient
from numpy.testing import assert_almost_equal

from mlapi import __version__
from mlapi.main import app

client = TestClient(app)


def test_predict():
    data = {"text": ["I hate you.", "I love you."]}
    response = client.post(
        "/predict",
        json=data,
    )

    assert response.status_code == 200
    assert type(response.json()["predictions"]) is list
    assert type(response.json()["predictions"][0]) is list
    assert type(response.json()["predictions"][0][0]) is dict
    assert type(response.json()["predictions"][1][0]) is dict
    assert set(response.json()["predictions"][0][0].keys()) == {"label",
"score"}
    assert set(response.json()["predictions"][0][1].keys()) == {"label",
"score"}
    assert set(response.json()["predictions"][1][0].keys()) == {"label",
"score"}
    assert set(response.json()["predictions"][1][1].keys()) == {"label",
"score"}
    assert response.json()["predictions"][0][0]["label"] == "NEGATIVE"
    assert response.json()["predictions"][0][1]["label"] == "POSITIVE"
    assert response.json()["predictions"][1][0]["label"] == "NEGATIVE"
    assert response.json()["predictions"][1][1]["label"] == "POSITIVE"
    assert (
        assert_almost_equal(
            response.json()["predictions"][0][0]["score"], 0.883, decimal=3
        )
        is None
    )
    assert (
        assert_almost_equal(
            response.json()["predictions"][0][1]["score"], 0.116, decimal=3
        )
        is None
    )
    assert (
        assert_almost_equal(
            response.json()["predictions"][1][0]["score"], 0.004, decimal=3
        )
        is None
    )
    assert (
        assert_almost_equal(
            response.json()["predictions"][1][1]["score"], 0.996, decimal=3
        )
        is None
    )
```

The test execution results are shown below:

```
ubuntu@ubuntu:~/w255/spring22-sudhrity/project/mlapi$ poetry run pytest -v
================================================= test session starts
=================================================
platform linux -- Python 3.10.4, pytest-7.1.1, pluggy-1.0.0 --
/home/ubuntu/.cache/pypoetry/virtualenvs/mlapi-ta4nsihM-py3.10/bin/python
cachedir: .pytest_cache
rootdir: /home/ubuntu/w255/spring22-sudhrity/project/mlapi
plugins: anyio-3.5.0
collected 1 item


tests/test_mlapi.py::test_predict PASSED
                               [100%]

================================================== warnings summary
===================================================
../../../../.cache/pypoetry/virtualenvs/mlapi-ta4nsihM-
py3.10/lib/python3.10/site-packages/torch/nn/modules/module.py:1402
  /home/ubuntu/.cache/pypoetry/virtualenvs/mlapi-ta4nsihM-
py3.10/lib/python3.10/site-packages/torch/nn/modules/module.py:1402: UserWarning:
positional arguments and argument "destination" are deprecated.
nn.Module.state_dict will not accept them in the future. Refer to
https://pytorch.org/docs/master/generated/torch.nn.Module.html#torch.nn.Module.s
tate_dict for details.
    warnings.warn(
-- Docs: https://docs.pytest.org/en/stable/how-to/capture-warnings.html
============================================== 1 passed, 1 warning in 1.71s ===============================================
```

4 Build, deploy in AKS and execute tests in Production environment

Below is a copy of the *run_prod.sh*

```
#!/bin/bash
APP_NAME=mlapi
IMAGE_NAME=project
NAMESPACE=sudhrity
kubectl config use-context w255-aks

# minikube
minikube start --kubernetes-version=v1.21.7 --extra-config=apiserver.service-node-port-range=1-65535
#APP_HOST=`minikube ip`

kubectl config use-context w255-aks
az acr login --name w255mids

kubectl delete -k ${APP_NAME}/.k8s/overlays/prod

docker rmi -f ${IMAGE_NAME}
docker rmi -f ${IMAGE_FQDN}

docker build --no-cache -t ${IMAGE_NAME} ./${APP_NAME}/

IMAGE_PREFIX=$(az account list --all | jq '.[].user.name' | grep -i berkeley.edu | awk -F@ '{print $1}' | tr -d '"' | uniq)
ACR_DOMAIN=w255mids.azurecr.io
IMAGE_FQDN="${ACR_DOMAIN}/${IMAGE_PREFIX}/${IMAGE_NAME}"
az acr login --name w255mids


#TAG=$(echo $RANDOM | md5sum | head -c 8; echo;)
#sed "s/\[TAG\]/${TAG}/g" ${APP_NAME}/.k8s/overlays/prod/patch-deployment-lab4_copy.yaml > ${APP_NAME}/.k8s/overlays/prod/patch-deployment-lab4.yaml

TAG=latest
docker tag ${IMAGE_NAME} ${IMAGE_FQDN}:${TAG}
docker push ${IMAGE_FQDN}:${TAG}
docker pull ${IMAGE_FQDN}:${TAG}

kubectl kustomize ${APP_NAME}/.k8s/overlays/prod
kubectl apply -k ${APP_NAME}/.k8s/overlays/prod

sleep 60
```

```
kubectl get all -n sudhrity

APP_HOST=${NAMESPACE}.mids-w255.com
APP_PORT=443

# wait for the /health endpoint to return a 200 and then move on
finished=false
while ! $finished; do
    health_status=$(curl -o /dev/null -s -w "%{http_code}\n" -X GET "https://${APP_HOST}:${APP_PORT}/health")
    if [ $health_status == "200" ]; then
        finished=true
        echo "API is ready"
    else
        echo "API not responding yet https://${APP_HOST}:${APP_PORT}/health"
        sleep 5
        # set this to avoid github action infinite loop. run.sh works locally but health check fails
        # when executed as a github action
        finished=true

    fi
done

sleep 30

# check a few endpoints and their http response
curl -o /dev/null -s -w "%{http_code}\n" -X GET "https://${APP_HOST}:${APP_PORT}/docs"

curl -iLX 'GET' \
 'https://sudhrity.mids-w255.com/predict' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{"text": ["I love you."]}'

echo

# output and tail the logs for the container
kubectl logs -f -n ${NAMESPACE} -l app=${IMAGE_NAME}
```

The execution results of *run_prod.sh* on the prod environment is shown below:

```
...
configmap/redis created
service/project created
service/redis created
persistentvolume/local-pv1 unchanged
deployment.apps/project created
statefulset.apps/redis created
horizontalpodautoscaler.autoscaling/project created
virtualservice.networking.istio.io/project created
NAME                              READY   STATUS            RESTARTS   AGE
pod/project-9d6cbbb4b-xp7vc       0/2     PodInitializing   0          61s
```

```
pod/redis-0                   2/2     Running            0          61s

NAME                 TYPE        CLUSTER-IP      EXTERNAL-IP   PORT(S)    AGE
service/project      ClusterIP   10.0.129.229    <none>        8000/TCP   62s
service/redis        ClusterIP   None            <none>        6379/TCP   61s

NAME                      READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/project   0/1     1            0           62s

NAME                                 DESIRED   CURRENT   READY   AGE
replicaset.apps/project-9d6cbbb4b    1         1         0       62s

NAME                     READY   AGE
statefulset.apps/redis   1/1     62s

NAME                                              REFERENCE            TARGETS
  MINPODS   MAXPODS   REPLICAS   AGE
horizontalpodautoscaler.autoscaling/lab4          Deployment/lab4      1%/50%
   1         10        1          28h
horizontalpodautoscaler.autoscaling/project       Deployment/project   <unknown>/50%
  1         10        1          62s
API not responding yet https://sudhrity.mids-w255.com:443/health
200
HTTP/2 200
date: Thu, 14 Apr 2022 20:26:32 GMT
server: istio-envoy
x-mlapi-cache: Miss
expires: Thu, 14 Apr 2022 20:27:33 GMT
cache-control: max-age=60
etag: W/3828612506244308857
content-length: 125
content-type: application/json
x-envoy-upstream-service-time: 31

{"predictions": [[{"label": "NEGATIVE", "score": 0.003921117167919874},
{"label": "POSITIVE", "score": 0.9960789084434509}]]}
...
INFO:     127.0.0.6:45123 - "GET /health HTTP/1.1" 200 OK
INFO:     127.0.0.6:45621 - "GET /docs HTTP/1.1" 200 OK
INFO:fastapi_redis_cache.client: 04/14/2022 08:26:33 PM | KEY_ADDED_TO_CACHE:
key=mlapi-cache:mlapi.main.predict(sentiments=text=['I love you.'])
INFO:     127.0.0.6:46595 - "GET /predict HTTP/1.1" 200 OK
INFO:     127.0.0.6:47075 - "GET /health HTTP/1.1" 200 OK
...
INFO:     127.0.0.6:35533 - "GET /health HTTP/1.1" 200 OK
INFO:fastapi_redis_cache.client: 04/14/2022 08:27:31 PM | KEY_FOUND_IN_CACHE:
key=mlapi-cache:mlapi.main.predict(sentiments=text=['I love you.'])
INFO:     127.0.0.6:37799 - "GET /predict HTTP/1.1" 200 OK
```

5 Instructions to conduct Performance tests

A **run_k6.sh** script is provided in the project root directory to execute performance test on application API application deployed on AKS.

```
#!/bin/bash
k6 run  --summary-trend-stats="min,med,avg,max,p(90),p(95),p(99)" load.js
```

The K6 script used for the performance test is `load.js` and is shown below:

```
import http from 'k6/http';
import { check, group, sleep } from 'k6';

export const options = {
  stages: [
    { duration: '30s', target: 10 }, // simulate ramp-up of traffic from 1 to 10 users over 30 seconds.
    { duration: '7m', target: 10 }, // stay at 10 users for 7 minutes
    { duration: '3m', target: 0 }, // ramp-down to 0 users
  ],
  thresholds: {
    'http_req_duration': ['p(99)<2000'] // 99% of requests must complete below 2s
  },
};

const fixed = ["I love you!", "I hate you!", "I am a Kubernetes Cluster!"]
var random_shuffler = [
  "I love you!",
  "I hate you!",
  "I am a Kubernetes Cluster!",
  "I ran to the store",
  "The students are very good in this class",
  "Working on Saturday morning is brutal",
  "How much wood could a wood chuck chuck if a wood chuck could chuck wood?",
  "A Wood chuck would chuck as much wood as a wood chuck could chuck if a wood chuck could chuck wood",
  "Food is very tasty",
  "Welcome to the thunderdome"
];

const generator = (cacheRate) => {
  const rand = Math.random()
  const text = rand > cacheRate
    ? random_shuffler.map(value => ({ value, sort: Math.random() }))
      .sort((a, b) => a.sort - b.sort)
      .map(({ value }) => value)
    : fixed
  return {
    text
  }
}

const NAMESPACE = 'sudhrity'
const BASE_URL = `https://${NAMESPACE}.mids-w255.com`;
const CACHE_RATE = .95
```

```
export default () => {
  const healthRes = http.get(`${BASE_URL}/health`)
  check(healthRes, {
    'is 200': (r) => r.status === 200
  })

  const payload = JSON.stringify(generator(CACHE_RATE))
  const predictionRes = http.request('POST', `${BASE_URL}/predict`, payload)
  check(predictionRes, {
    'is 200': (r) => r.status === 200
  })
};
```

The script used to setup performance dashboard using Grafana is as follows:

```
kubectl port-forward -n prometheus svc/grafana 3000:3000
```

Grafana is accessible using the following URL:

```
http://localhost:3000/?orgId=1
```

6 Performance Test Run

The performance test is conducted with a ramp-up period, load test period and ramp-down period. In these tests the duration used for the test runs are as below:

- Ramp-up - 3 minutes
- Load-test - 7 minutes
- Ram-down - 3 minutes

Performance test scenarios

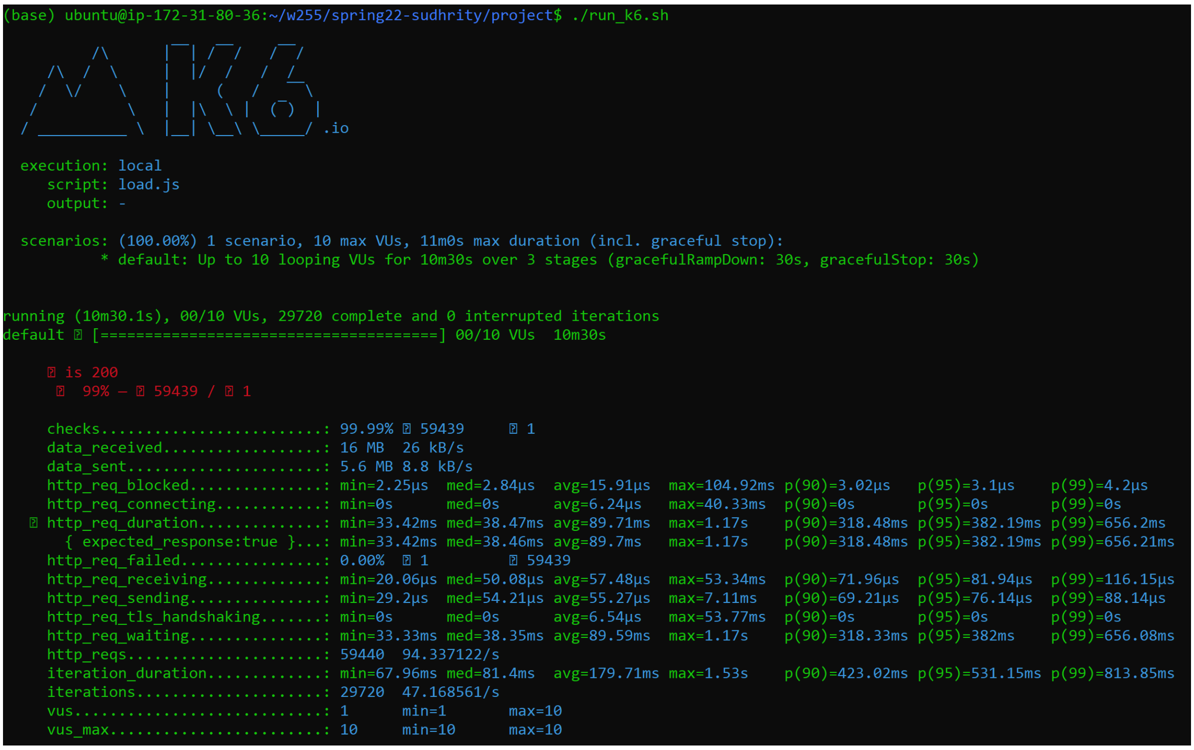
The performance tests were conducted with various cache rates, to evaluate how the application performance varies with varying cache rates. The test scenarios are shown below:

Test #	Cache Rates	Test Script Execution Environment	Cluster State
Test 1	0.95	AWS VM with Ubuntu 20.04 on local machine - t2.xlarge	REPLICAS=1
Test 2	0.95	AWS VM with Ubuntu 20.04 on local machine - t2.xlarge	REPLICAS=10

7 Test 1 - CACHE_RATE=0.95 after cluster started. 10+ minute sustained load

The results are shown in the screenshots below:

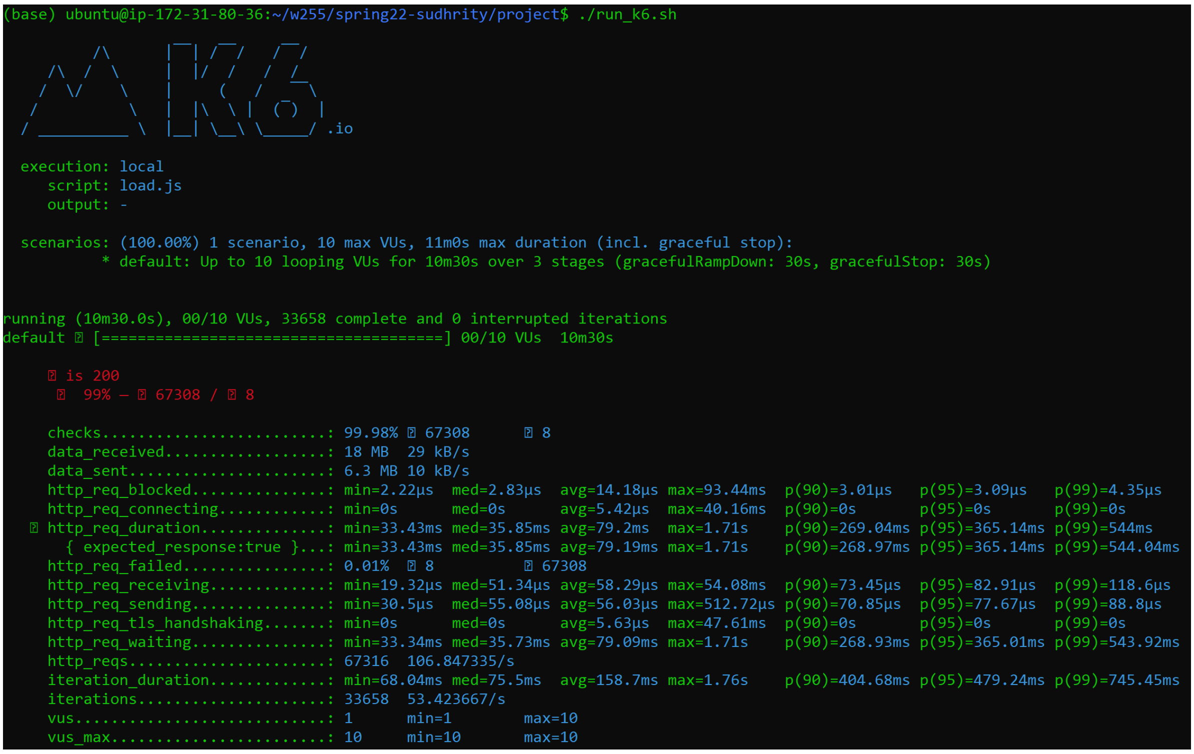
Execution Results



```
(base) ubuntu@ip-172-31-80-36:~/w255/spring22-sudhrity/project$ ./run_k6.sh

          /\      |‾‾| /‾‾/   /‾‾/
     /\  /  \     |  |/  /   /  /
    /  \/    \    |     (   /   ‾‾\
   /          \   |  |\  \ |  (‾)  |
  / __________ \  |__| \__\ \_____/ .io

  execution: local
     script: load.js
     output: -

  scenarios: (100.00%) 1 scenario, 10 max VUs, 11m0s max duration (incl. graceful stop):
           * default: Up to 10 looping VUs for 10m30s over 3 stages (gracefulRampDown: 30s, gracefulStop: 30s)


running (10m30.1s), 00/10 VUs, 29720 complete and 0 interrupted iterations
default ⍰ [======================================] 00/10 VUs  10m30s

     ⍰ is 200
      ⍰  99% — ⍰ 59439 / ⍰ 1

     checks.........................: 99.99% ⍰ 59439      ⍰ 1
     data_received..................: 16 MB  26 kB/s
     data_sent......................: 5.6 MB 8.8 kB/s
     http_req_blocked...............: min=2.25µs  med=2.84µs  avg=15.91µs  max=104.92ms p(90)=3.02µs   p(95)=3.1µs    p(99)=4.2µs
     http_req_connecting............: min=0s      med=0s      avg=6.24µs   max=40.33ms  p(90)=0s       p(95)=0s       p(99)=0s
   ⍰ http_req_duration..............: min=33.42ms med=38.47ms avg=89.71ms  max=1.17s    p(90)=318.48ms p(95)=382.19ms p(99)=656.2ms
       { expected_response:true }...: min=33.42ms med=38.46ms avg=89.7ms   max=1.17s    p(90)=318.48ms p(95)=382.19ms p(99)=656.21ms
     http_req_failed................: 0.00%  ⍰ 1          ⍰ 59439
     http_req_receiving.............: min=20.06µs med=50.08µs avg=57.48µs  max=53.34ms  p(90)=71.96µs  p(95)=81.94µs  p(99)=116.15µs
     http_req_sending...............: min=29.2µs  med=54.21µs avg=55.27µs  max=7.11ms   p(90)=69.21µs  p(95)=76.14µs  p(99)=88.14µs
     http_req_tls_handshaking.......: min=0s      med=0s      avg=6.54µs   max=53.77ms  p(90)=0s       p(95)=0s       p(99)=0s
     http_req_waiting...............: min=33.33ms med=38.35ms avg=89.59ms  max=1.17s    p(90)=318.33ms p(95)=382ms    p(99)=656.08ms
     http_reqs......................: 59440  94.337122/s
     iteration_duration.............: min=67.96ms med=81.4ms  avg=179.71ms max=1.53s    p(90)=423.02ms p(95)=531.15ms p(99)=813.85ms
     iterations.....................: 29720  47.168561/s
     vus............................: 1      min=1        max=10
     vus_max........................: 10     min=10       max=10
```

8 Test 2 - CACHE_RATE=0.95 after previous test run with 0 delay. 10+ minute sustained load

Execution Results

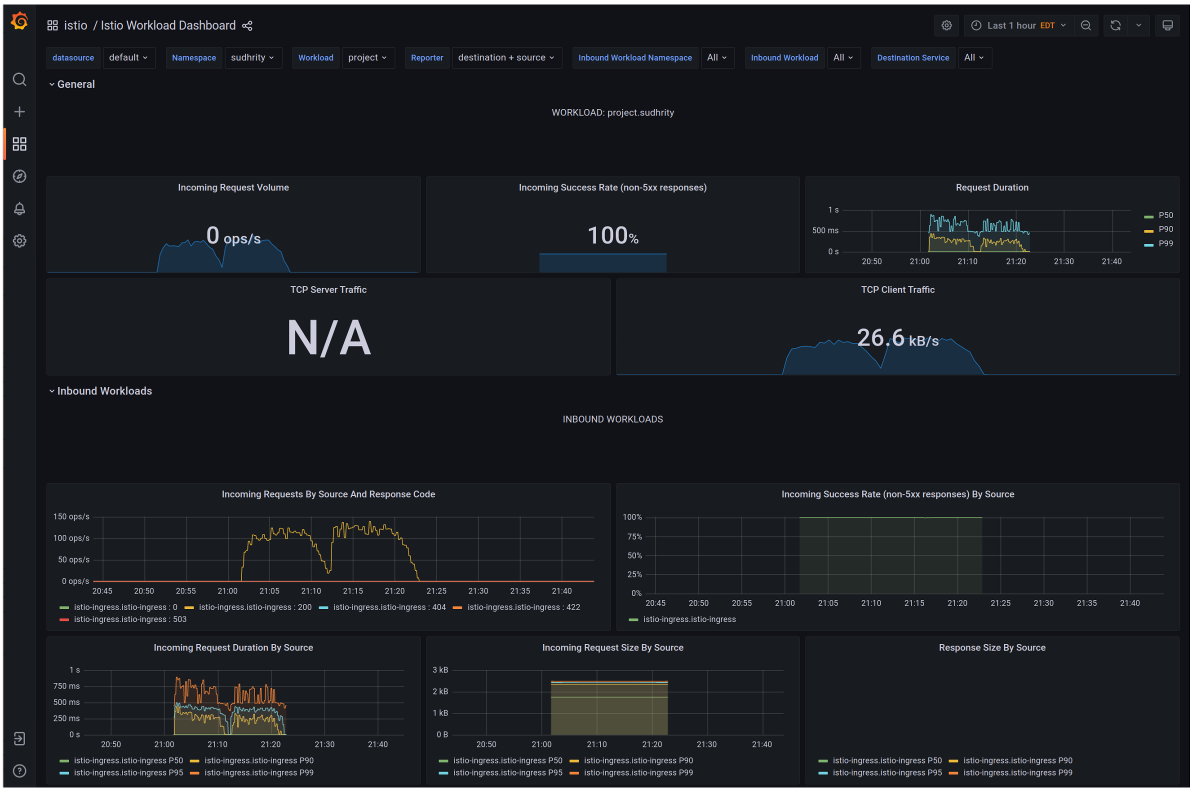


```
(base) ubuntu@ip-172-31-80-36:~/w255/spring22-sudhrity/project$ ./run_k6.sh

          /\      |‾‾| /‾‾/   /‾‾/
     /\  /  \     |  |/  /   /  /
    /  \/    \    |     (   /   ‾‾\
   /          \   |  |\  \ |  (‾)  |
  / __________ \  |__| \__\ \_____/ .io

  execution: local
     script: load.js
     output: -

  scenarios: (100.00%) 1 scenario, 10 max VUs, 11m0s max duration (incl. graceful stop):
           * default: Up to 10 looping VUs for 10m30s over 3 stages (gracefulRampDown: 30s, gracefulStop: 30s)


running (10m30.0s), 00/10 VUs, 33658 complete and 0 interrupted iterations
default ⍰ [======================================] 00/10 VUs  10m30s

     ⍰ is 200
      ⍰  99% — ⍰ 67308 / ⍰ 8

     checks.........................: 99.98% ⍰ 67308      ⍰ 8
     data_received..................: 18 MB  29 kB/s
     data_sent......................: 6.3 MB 10 kB/s
     http_req_blocked...............: min=2.22µs  med=2.83µs  avg=14.18µs max=93.44ms  p(90)=3.01µs   p(95)=3.09µs   p(99)=4.35µs
     http_req_connecting............: min=0s      med=0s      avg=5.42µs  max=40.16ms  p(90)=0s       p(95)=0s       p(99)=0s
   ⍰ http_req_duration..............: min=33.43ms med=35.85ms avg=79.2ms  max=1.71s    p(90)=269.04ms p(95)=365.14ms p(99)=544ms
       { expected_response:true }...: min=33.43ms med=35.85ms avg=79.19ms max=1.71s    p(90)=268.97ms p(95)=365.14ms p(99)=544.04ms
     http_req_failed................: 0.01%  ⍰ 8          ⍰ 67308
     http_req_receiving.............: min=19.32µs med=51.34µs avg=58.29µs max=54.08ms  p(90)=73.45µs  p(95)=82.91µs  p(99)=118.6µs
     http_req_sending...............: min=30.5µs  med=55.08µs avg=56.03µs max=512.72µs p(90)=70.85µs  p(95)=77.67µs  p(99)=88.8µs
     http_req_tls_handshaking.......: min=0s      med=0s      avg=5.63µs  max=47.61ms  p(90)=0s       p(95)=0s       p(99)=0s
     http_req_waiting...............: min=33.34ms med=35.73ms avg=79.09ms max=1.71s    p(90)=268.93ms p(95)=365.01ms p(99)=543.92ms
     http_reqs......................: 67316  106.847335/s
     iteration_duration.............: min=68.04ms med=75.5ms  avg=158.7ms max=1.76s    p(90)=404.68ms p(95)=479.24ms p(99)=745.45ms
     iterations.....................: 33658  53.423667/s
     vus............................: 1      min=1        max=10
     vus_max........................: 10     min=10       max=10
```

9 Response Times (ms)

Test #	Cache Rate	Min	Med	Avg	Max	P(90)	P(95)	P(99)	# req ✓	# req ✗	req/s
1	0.95	33.42	38.47	89.71	1170	318.48	382.19	**656.2**	59440	1	**94.33**
2	0.95	33.43	35.85	79.2	1710	269.04	365.14	**544.04**	67316	8	**106.84**

10 Istio Workload Dashboards



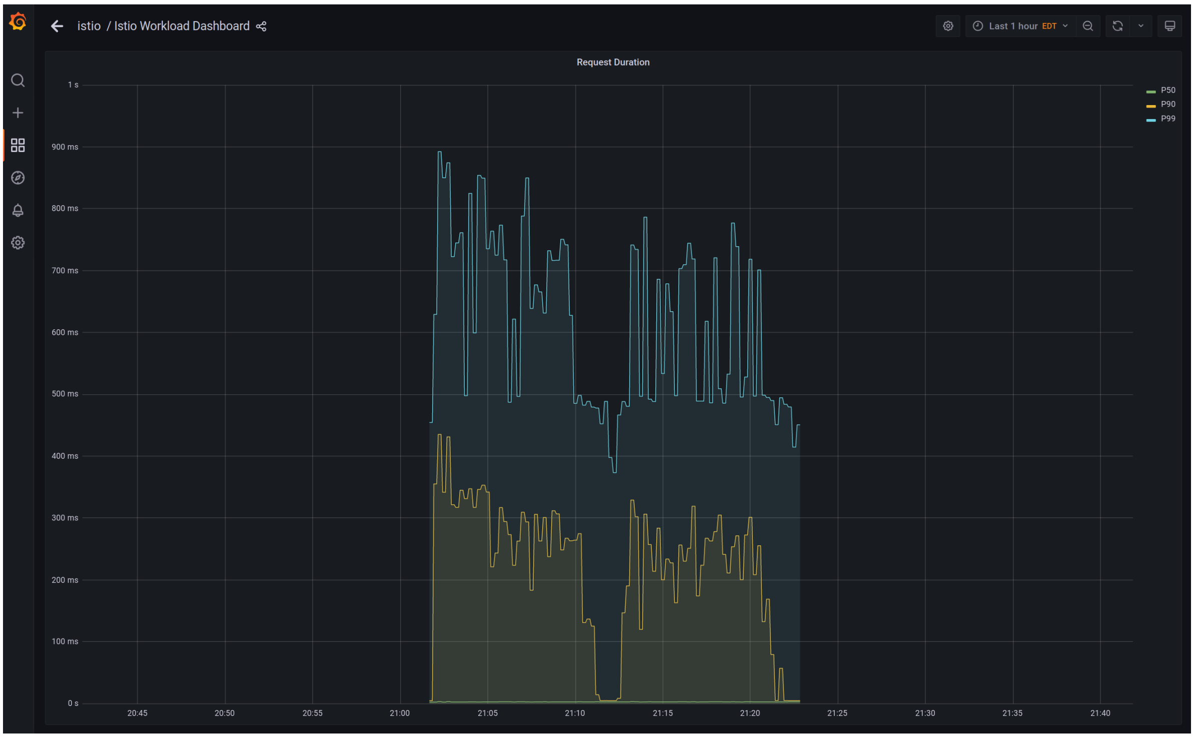
istio / Istio Workload Dashboard

datasource: default · Namespace: sudhrity · Workload: project · Reporter: destination + source · Inbound Workload Namespace: All · Inbound Workload: All · Destination Service: All

Last 1 hour EDT

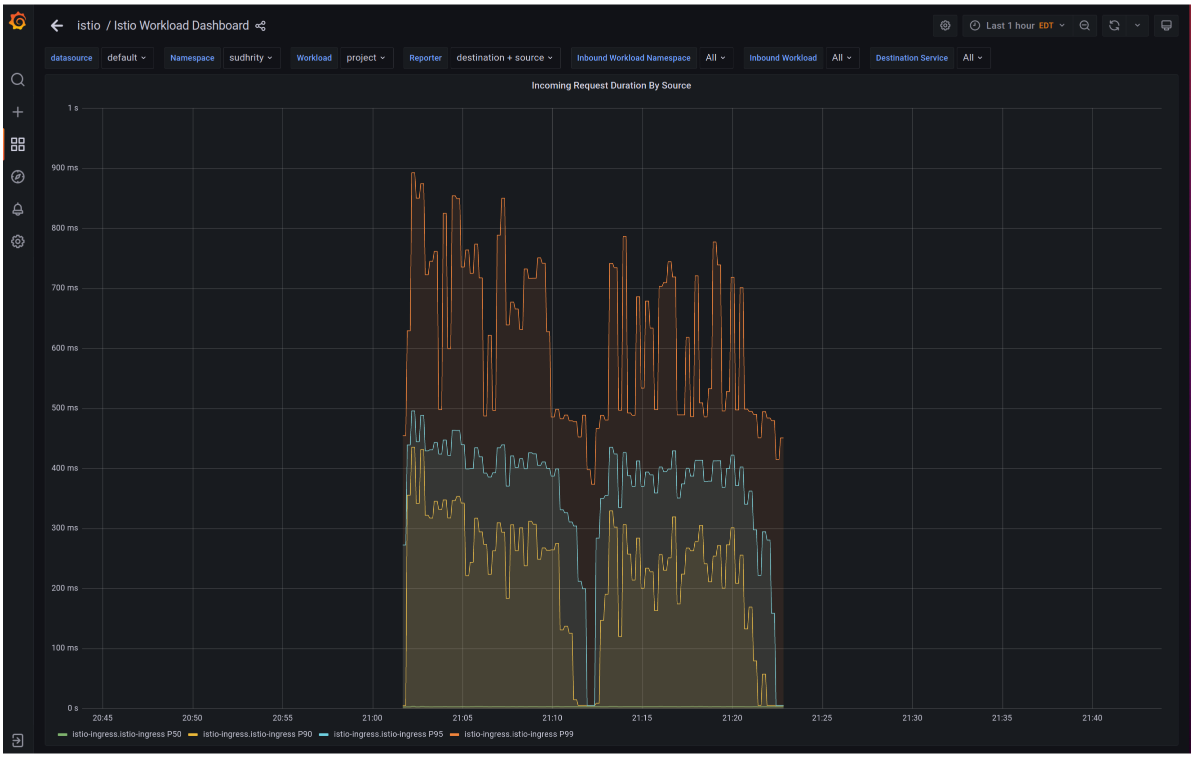
WORKLOAD: project.sudhrity

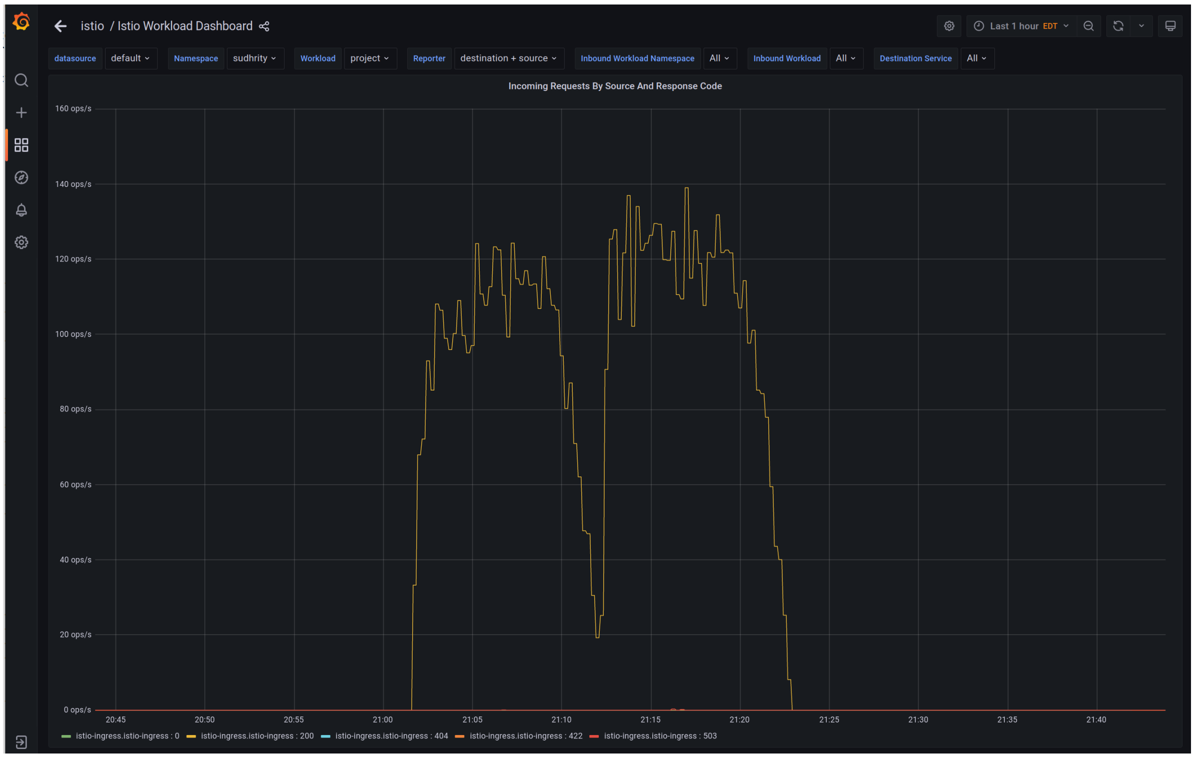
General
Incoming Request Volume
0 ops/s
Incoming Success Rate (non-5xx responses)
100%
Request Duration
1 s
500 ms
0 s
20:50
21:00
21:10
21:20
21:30
21:40
P50
P90
P99
TCP Server Traffic
N/A
TCP Client Traffic
26.6 kB/s
Inbound Workloads
INBOUND WORKLOADS
Incoming Requests By Source And Response Code
150 ops/s
100 ops/s
50 ops/s
0 ops/s
20:45
20:50
20:55
21:00
21:05
21:10
21:15
21:20
21:25
21:30
21:35
21:40
istio-ingress.istio-ingress : 0
istio-ingress.istio-ingress : 200
istio-ingress.istio-ingress : 404
istio-ingress.istio-ingress : 422
istio-ingress.istio-ingress : 503
Incoming Success Rate (non-5xx responses) By Source
100%
75%
50%
25%
0%
istio-ingress.istio-ingress
Incoming Request Duration By Source
1 s
750 ms
500 ms
250 ms
0 s
istio-ingress.istio-ingress P50
istio-ingress.istio-ingress P90
istio-ingress.istio-ingress P95
istio-ingress.istio-ingress P99
Incoming Request Size By Source
3 kB
2 kB
1 kB
0 B
istio-ingress.istio-ingress P50
istio-ingress.istio-ingress P90
istio-ingress.istio-ingress P95
istio-ingress.istio-ingress P99
Response Size By Source
istio-ingress.istio-ingress P50
istio-ingress.istio-ingress P90
istio-ingress.istio-ingress P95
istio-ingress.istio-ingress P99



istio / Istio Workload Dashboard

Last 1 hour EDT





11.1 Requirements

1. Write pydantic models to meet the specified input model: **- Done**

```
{"text": ["example 1", "example 2"]}
```

2. Pull the following model winegarj/distilbert-base-uncased-finetuned-sst2 locally to allow for loading into your application. Put this at the root of your project directory for an easier time. **- Done**
3. Run pytest to ensure your application is working as intended. **- Done**
    - Update your application as neccessary
4. Build and test your docker container locally. **- Done**
    - Minikube or docker-compose are fine. - *I used Minikube*
    - `kustomize` overlays and `docker-compose.yml` files are provided to minimize effort - *I used kustomize*
5. Push your image to ACR use a prefix based on your namespace, and call the image project. **- Done**
6. Deploy your application to AKS leveraging Istio similarly to `lab 4/5` **- Done**
7. Test your endpoint works with a simple example **- Done**
8. If your endpoint is unresponsive, make sure you review the pods and logs and see if there are any issues. **- Done** - *There were issues with resources and memory was increased for the pod in patch-deployment files.*
9. Run k6 against your endpoint with the provided `load.js` **- Done**
10. Feel extremely proud about all the learning you went through over the semester and how this will help you develop professionally and enable you to deploy an API effectively during capstone. There is much to learn, but getting the fundamentals are key. **- Done**

11.2 Rubric

- `pytest` **(provided) pass for your project: 2 point**

Yes. Results provided in section 3.1

- `Model` **is loaded into the container as part of the build process instead of being dynamically pulled down from** `HuggingFace` **on model instantiation:**

Yes. The main.py code is provided below. This also shows that pydantic models were used to meet the specified input model. The model winegarj/distilbert-base-uncased-finetuned-sst2 is copied locally into the root of the project directory to allow for loading into the API application.

```
import logging
import os
from typing import Dict

from fastapi import FastAPI, Request, Response
from fastapi_redis_cache import FastApiRedisCache, cache_one_minute
from pydantic import BaseModel
from transformers import pipeline, AutoModelForSequenceClassification,
AutoTokenizer
```

```
model_path = "./distilbert-base-uncased-finetuned-sst2"
model = AutoModelForSequenceClassification.from_pretrained(model_path)
tokenizer = AutoTokenizer.from_pretrained(model_path)
classifier = pipeline(
    task="text-classification",
    model=model,
    tokenizer=tokenizer,
    device=-1,
    return_all_scores=True,
)

logger = logging.getLogger(__name__)
LOCAL_REDIS_URL = "redis://redis:6379/0"
app = FastAPI()

@app.on_event("startup")
def startup():
    redis_cache = FastApiRedisCache()
    redis_cache.init(
        host_url=os.environ.get("REDIS_URL", LOCAL_REDIS_URL),
        prefix="mlapi-cache",
        response_header="X-MLAPI-Cache",
        ignore_arg_types=[Request, Response],
    )

class SentimentRequest(BaseModel):
    text: list[str]

class Sentiment(BaseModel):
    label: str
    score: float

class SentimentResponse(BaseModel):
    predictions: list[list[Sentiment]]

@app.get("/predict", response_model=SentimentResponse)
@cache_one_minute()
def predict(sentiments: SentimentRequest):
    return {"predictions": classifier(sentiments.text)}

@app.post("/predict", response_model=SentimentResponse)
@cache_one_minute()
def predict(sentiments: SentimentRequest):
    return {"predictions": classifier(sentiments.text)}

@app.get("/health")
async def health():
    return {"status": "healthy"}
```

- **Ability to hit `/predict` endpoint and get sentiment responses: 2 points**

Yes. The results are shown for dev environment in **Section 3** and for the production environment in **Section 4**

- **Ability to hit `/predict` endpoint 10/s: 2 points**

Yes. Results from the test on production for Tests 1 and 2 are **94.33/s** and **106.84/s**. This is shown in **Section 9**. This is also shown in the **Istio Dashboard - Incoming Requests by Source and Response Code** in **Section 10**

- **p(99) < 2 second for `/predict` endpoint under 10 Virtual User (`k6` VU) load: 2 points**

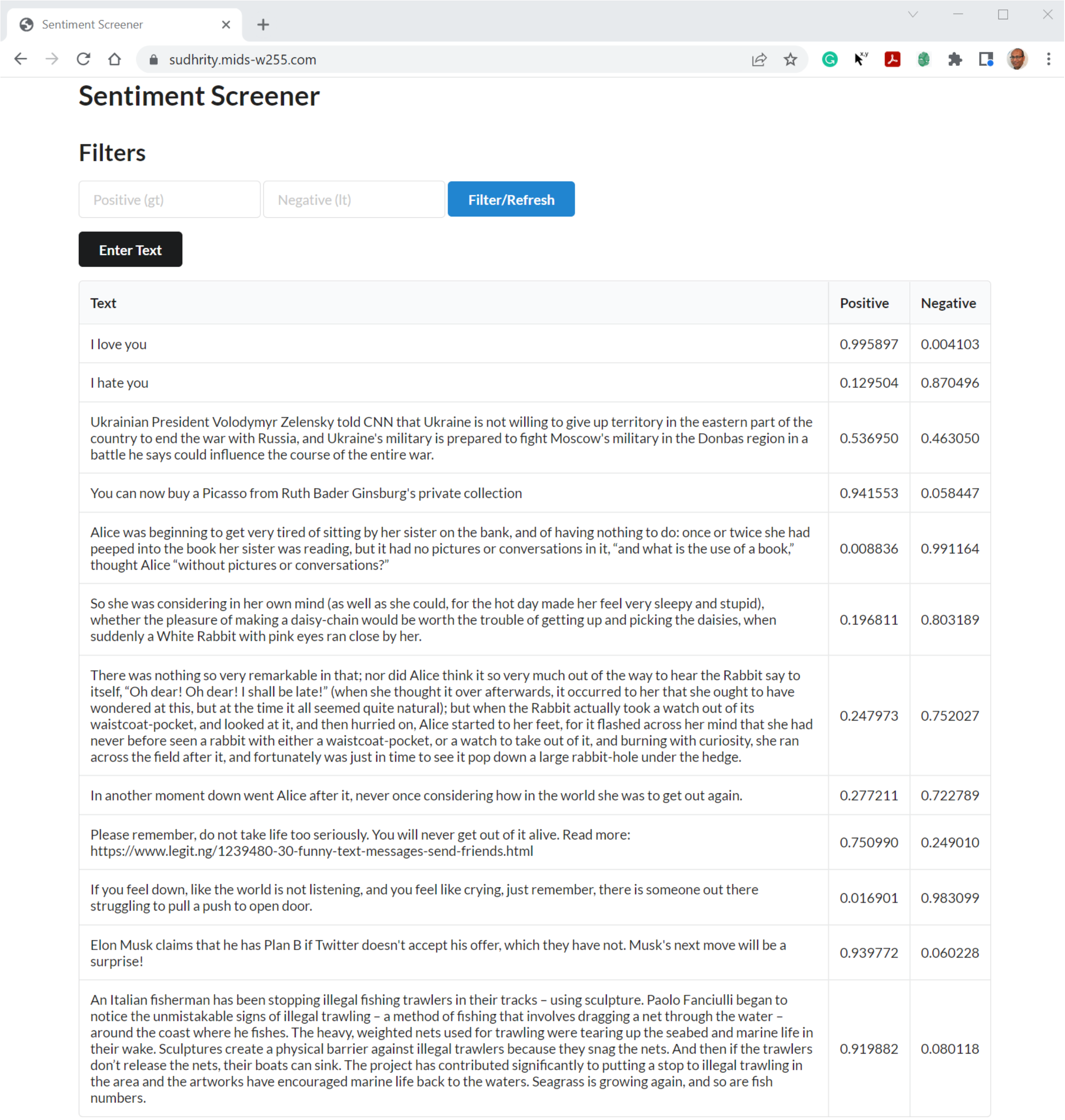
Yes. Results from the test on production for Tests 1 and 2 are **0.6562 s** and **0.5440 s**. This is shown in **Section 9**. This is also shown in the **Istio Dashboard - Request Duration** and **Incoming Request Duration by Source** in **Section 10**

12 UI for the Sentiment API application

The UI for the API application can be accessed at: https://sudhrity.mids-w255.com/ This is a very basic UI, not fully tested and is shown below.

Instructions to use the UI is as follows:

- To enter a text for sentiment screening, press the **Enter Text** button. Multiple sentences can be added. To get the sentiment scores, click on **Filter/Refresh**.
- To filter texts, you may enter **Positive** and/or **Negative** scores in the *Positive (gt)* and *Negative (lt)* text boxes and click on Filter/Refresh. This will return tests with Positive scores greater than the score entered and/or Negative score less than the score entered.



Sentiment Screener

Filters

Text	Positive	Negative
I love you	0.995897	0.004103
I hate you	0.129504	0.870496
Ukrainian President Volodymyr Zelensky told CNN that Ukraine is not willing to give up territory in the eastern part of the country to end the war with Russia, and Ukraine's military is prepared to fight Moscow's military in the Donbas region in a battle he says could influence the course of the entire war.	0.536950	0.463050
You can now buy a Picasso from Ruth Bader Ginsburg's private collection	0.941553	0.058447
Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, "and what is the use of a book," thought Alice "without pictures or conversations?"	0.008836	0.991164
So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.	0.196811	0.803189
There was nothing so very remarkable in that; nor did Alice think it so very much out of the way to hear the Rabbit say to itself, "Oh dear! Oh dear! I shall be late!" (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually took a watch out of its waistcoat-pocket, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.	0.247973	0.752027
In another moment down went Alice after it, never once considering how in the world she was to get out again.	0.277211	0.722789
Please remember, do not take life too seriously. You will never get out of it alive. Read more: https://www.legit.ng/1239480-30-funny-text-messages-send-friends.html	0.750990	0.249010
If you feel down, like the world is not listening, and you feel like crying, just remember, there is someone out there struggling to pull a push to open door.	0.016901	0.983099
Elon Musk claims that he has Plan B if Twitter doesn't accept his offer, which they have not. Musk's next move will be a surprise!	0.939772	0.060228
An Italian fisherman has been stopping illegal fishing trawlers in their tracks – using sculpture. Paolo Fanciulli began to notice the unmistakable signs of illegal trawling – a method of fishing that involves dragging a net through the water – around the coast where he fishes. The heavy, weighted nets used for trawling were tearing up the seabed and marine life in their wake. Sculptures create a physical barrier against illegal trawlers because they snag the nets. And then if the trawlers don't release the nets, their boats can sink. The project has contributed significantly to putting a stop to illegal trawling in the area and the artworks have encouraged marine life back to the waters. Seagrass is growing again, and so are fish numbers.	0.919882	0.080118