

Load Testing & Bottlenecks

Software Architecture

April 26, 2023

Brae Webb



1 This Week

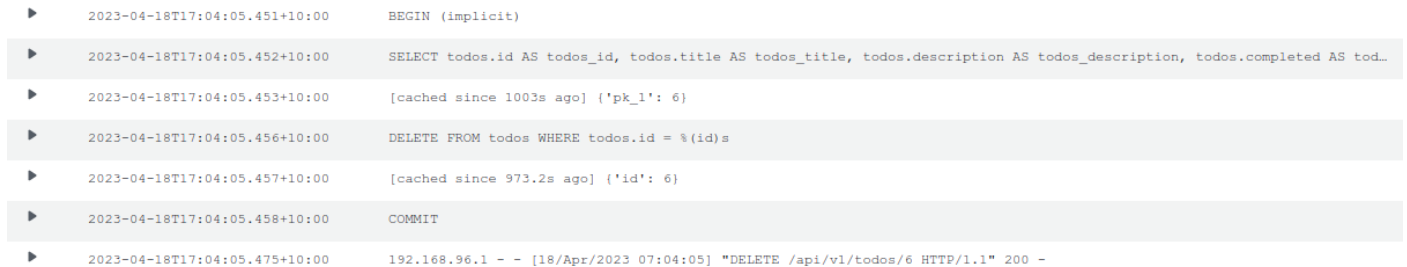
Our goal is to:

- Introduce structured logging to TaskOverflow.
- Deploy TaskOverflow to AWS.
- Write k6 tests to ensure TaskOverflow can handle given scenarios.
- Uncover and fix bottlenecks in the TaskOverflow application using the logs.

2 Watchtower

In this course we have repeatedly claimed that there is value in monitoring and logging. This week, we aim to prove it by using logging to help diagnose issues with a deployment of TaskOverflow. For this task, we have integrated [Watchtower](https://pypi.org/project/watchtower/)¹ into the project. Watchtower is a Python library that allows us to send logs to AWS CloudWatch, allowing us to monitor the applications performance in real time.

Currently the project is configured to log API accesses made with Flask and to log database queries created by SQL Alchemy. This results in an unstructured log stream as seen in Figure 1.



```
▶ 2023-04-18T17:04:05.451+10:00 BEGIN (implicit)
▶ 2023-04-18T17:04:05.452+10:00 SELECT todos.id AS todos_id, todos.title AS todos_title, todos.description AS todos_description, todos.completed AS tod...
▶ 2023-04-18T17:04:05.453+10:00 [cached since 1003s ago] {'pk_1': 6}
▶ 2023-04-18T17:04:05.456+10:00 DELETE FROM todos WHERE todos.id = %(id)s
▶ 2023-04-18T17:04:05.457+10:00 [cached since 973.2s ago] {'id': 6}
▶ 2023-04-18T17:04:05.458+10:00 COMMIT
▶ 2023-04-18T17:04:05.475+10:00 192.168.96.1 - - [18/Apr/2023 07:04:05] "DELETE /api/v1/todos/6 HTTP/1.1" 200 -
```

Figure 1: An example of logs made to AWS CloudWatch for a DELETE request in the TaskOverflow API.

Getting Started

1. Using the GitHub Classroom link for this practical provided by your tutor in Slack, create a repository to work within.
2. Install Terraform if not already installed, as it will be required again this week.
3. Start your learner lab and copy the AWS Learner Lab credentials into a credentials file in the root of the repository.

What's New We are returning to TaskOverflow roughly from the state at the end of the last practical. The following notable changes have been made:

- Watchtower has been installed as a dependency.
- In `docker-compose.yml`, we mount the credentials file to `/root/.aws/credentials`. This allows local testing of watchtower to log to AWS CloudWatch.
- Logging has been introduced for Flask and SQL Alchemy.

Our first task will be to replicate the above logs in Figure 1. Once you have copied `credentials` into the project root, start docker compose with:

```
$ docker-compose up
```

¹<https://pypi.org/project/watchtower/>

```

app_1 | * Serving Flask app 'todo'
app_1 | * Debug mode: on
app_1 | INFO:werkzeug:WARNING: This is a development server. Do not use it in a
      | production deployment. Use a production WSGI server instead.
app_1 | * Running on all addresses (0.0.0.0)
app_1 | * Running on http://127.0.0.1:6400
app_1 | * Running on http://192.168.96.3:6400
app_1 | INFO:werkzeug:Press CTRL+C to quit
app_1 | INFO:werkzeug: * Restarting with stat
app_1 | INFO:botocore.credentials:Found credentials in shared credentials file: ~/.
      | aws/credentials
app_1 | INFO:sqlalchemy.engine.Engine:select pg_catalog.version()
app_1 | INFO:sqlalchemy.engine.Engine:[raw sql] {}
app_1 | INFO:sqlalchemy.engine.Engine:select current_schema()
app_1 | INFO:sqlalchemy.engine.Engine:[raw sql] {}
app_1 | INFO:sqlalchemy.engine.Engine:show standard_conforming_strings
app_1 | INFO:sqlalchemy.engine.Engine:[raw sql] {}
app_1 | INFO:sqlalchemy.engine.Engine:BEGIN (implicit)
app_1 | INFO:sqlalchemy.engine.Engine:SELECT pg_catalog.pg_class.relname
app_1 | FROM pg_catalog.pg_class JOIN pg_catalog.pg_namespace ON pg_catalog.
      | pg_namespace.oid = pg_catalog.pg_class.relnamespace
app_1 | WHERE pg_catalog.pg_class.relname = %(table_name)s AND pg_catalog.pg_class.
      | relkind = ANY (ARRAY[%(param_1)s, %(param_2)s, %(param_3)s, %(param_4)s, %(
      | param_5)s]) AND pg_catalog.pg_table_is_visible(pg_catalog.pg_class.oid) AND
      | pg_catalog.pg_namespace.nspname != %(nspname_1)s
app_1 | INFO:sqlalchemy.engine.Engine:[generated in 0.00011s] {'table_name': 'todos',
      | 'param_1': 'r', 'param_2': 'p', 'param_3': 'f', 'param_4': 'v', 'param_5': 'm',
      | 'nspname_1': 'pg_catalog'}
app_1 | INFO:sqlalchemy.engine.Engine:COMMIT
app_1 | WARNING:werkzeug: * Debugger is active!

```

You should see logs similar to the above. Notice that information about Flask is prefixed with `INFO:werkzeug` and information about SQL Alchemy is prefixed with `INFO:sqlalchemy.engine.Engine`. This prefix indicates which log stream they are put into.

Open the AWS CloudWatch console and go to Log groups on the side panel. You should see a log group called `taskoverflow`, if you click on that group you can see the two log streams.



2.1 Structured Logging

Our first task will be to convert the current logging into a structured logging format. As we saw in last weeks tutorial, structured logging can be as simple as logging a JSON object. This allows logging services to quickly filter through logs based on criteria on the objects fields.

In `todo/__init__.py` we have the following code within the `create_app` function. This code configures watchtower to log to AWS CloudWatch.

```
» cat todo/__init__.py

def create_app(config_overrides=None):
    ...
    handler = watchtower.CloudWatchLogHandler(
        log_group_name="taskoverflow",
        boto3_client=boto3.client("logs", region_name="us-east-1")
    )
    app.logger.addHandler(handler)
    ...
```

We want to wrap all our logs in JSON objects and inject metadata to these objects. This allows us to search the logs with greater ease. To do this, we will create a custom log formatter that is a subclass of `watchtower.CloudWatchLogFormatter`.

```
» cat todo/__init__.py

from flask import has_request_context, request
...

class StructuredFormatter(watchtower.CloudWatchLogFormatter):
    def format(self, record):
        record.msg = {
            'timestamp': record.created,
            'location': record.name,
            'message': record.msg,
        }
        if has_request_context():
            record.msg['request_id'] = request.environ.get('REQUEST_ID')
            record.msg['url'] = request.environ.get('PATH_INFO')
            record.msg['method'] = request.environ.get('REQUEST_METHOD')
        return super().format(record)

def create_app(config_overrides=None):
    ...
    handler = ...
    handler.setFormatter(StructuredFormatter())
    ...
```

These are just some metadata fields we might want to add to our logging. We may also consider `socket.gethostname()` to identify the instance handling a request or other helpful information.

2.2 Correlation IDs

Correlation IDs are a mechanism to help understand the path of a request, event, message, etc. through a system. When logging it can be helpful to be able to trace the execution of a particular request.

For our system we will generate a new random identifier for each incoming request. This identifier will be included as part of the logging metadata. For a system with multiple API endpoints requests will often come with an established `REQUEST_ID` header.

```
» cat todo/__init__.py

import uuid
...

def create_app(config_overrides=None):
    ...
    requests = logging.getLogger("requests")
    requests.addHandler(handler)

    @app.before_request
    def before_request():
        request.environ['REQUEST_ID'] = str(uuid.uuid4())
        requests.info("Request started")

    @app.after_request
    def after_request(response):
        requests.info("Request finished")
        return response
    ...
```

The above in `create_app` will generate a unique identifier to each incoming request. It will also log when a request starts being processed and when it finishes processing.

2.3 Verify and Refine

Ensure that the above modifications are working as expected by returning to the AWS CloudWatch logs. Launch the service locally and make some API requests.

You should now be able to go to the *Log Insights* interface from the sidebar. As we saw in the tutorial, this interface allows you to make queries about structured logging data. Give the following query a try to ensure that your logs seem sensible.

```
fields request_id, url, method, @timestamp, message.message, @message, @method
| filter not isempty(request_id)
| limit 400
```

Take the time to refine your logging now to produce logs that are easy to search through and understand the trace of your API requests.

3 Load Testing

We will now generate some load on our API using k6 and see how it performs. We briefly saw k6 at the end of the week 7 practical. To refresh, k6 is a load testing tool that is written in Go but provides an interface using a JavaScript subset.

It will help us to generate a large number of concurrent API requests. Go ahead and install k6: <https://k6.io/docs/get-started/installation/>

3.1 Scenario

We want to simulate the following scenario using our testing framework.

Indecisive planners and studious reviewers Our TaskOverflow application is being used at a point in semester where most students have already setup their tasks. These students are routinely visiting the website and listing the tasks they have yet to complete. At the same time, there are 40 very indecisive students who, quite late in semester, are trying to setup their tasks for the rest of semester. They continually create a task, realise that they're mis-typed, delete the task, and start over.

As the tasks are shared globally amongst all students, the indecisive planners are altering the list of task seen by the organised students.

3.2 Setup

To get started, we will create a file called `planners-and-studiers.js`. Most k6 tests start with the following imports.

```
» cat planners-and-studiers.js

import http from "k6/http";
import { check, sleep } from "k6";

const ENDPOINT = __ENV.ENDPOINT;
```

- `http` holds the methods used to make HTTP requests,
- `check` allows us to assert the state of HTTP responses, and
- `sleep` gives us the ability to put a simulated user to sleep rather than continuously spamming the service with requests.

The `ENDPOINT` line retrieves the endpoint URL from the `ENDPOINT` environment variable.

3.3 User Simulation

The behaviour of our users will be defined by an exported JavaScript function. Our studying student will be listing out all of the tasks they have left to complete by using the `/api/v1/todos` endpoint.

```
» cat planners-and-studiers.js
```

```

export function studyingStudent() {
  let url = ENDPOINT + `/api/v1/todos`;

  // What tasks do I have left to work on?
  let request = http.get(url);

  check(request, {
    'is status 200': (r) => r.status === 200,
  });

  // Alright I'll go work on my next task for around 2 minutes
  sleep(120);
}

```

Of course, this test is very basic as it only ensures the response code is 200. There is no guarantee that the returned data is sensible.

Info

If you would like a challenge, you can use the `randomIntBetween` function to have the student occasionally tick off tasks.

<https://k6.io/docs/javascript-api/jslib/utils/randomintbetween/>

Next we need to handle our indecisive users. For them, we will need to make a POST request to the `/api/v1/todos` endpoint to create a task. Then we will need to use the ID given by the POST response to DELETE the mis-typed task.

```

» cat planners-and-studiers.js

export function indecisivePlanner() {
  let url = ENDPOINT + `/api/v1/todos`;

  // I need to work on the CSSE6400 Cloud Assignment!
  const payload = JSON.stringify({
    "title": "CSSE6400 Cloud Assignment",
    "completed": false,
    "description": "",
    "deadline_at": "2023-12-04T14:00:00",
  });

  const params = {
    headers: {
      'Content-Type': 'application/json',
    },
  };

  let request = http.post(url, payload, params);
  check(request, {
    'is status 200': (r) => r.status === 200,
  });
}

```



```

});

sleep(10);

// Oh no! Not the Clout assignment, the Cloud assignment!
const wrongId = request.id;

request = http.delete(`${url}/${wrongId}`);

check(request, {
  'is status 200': (r) => r.status === 200,
});

// I'll come back to it later :(
sleep(10);
}

```

3.4 Configure Behaviour

We have now outlined how our user agents will interact with our API. We can use the `options` to configure how these interactions occur over time. To prevent overloading the API immediately, and allow any auto-scaling behaviour to be triggers, we will slowly increase the amount of studiers. We will have a consistent amount of planners, 20, who will perform a total of 200 corrections over the test.

```

export const options = {
  scenarios: {
    studier: {
      exec: 'studyingStudent',
      executor: 'ramping-vus',
      stages: [
        { duration: '2m', target: 1000 },
        { duration: '2m', target: 2500 },
        { duration: '2m', target: 0 },
      ],
    },
    planner: {
      exec: 'indecisivePlanner',
      executor: 'shared-iterations',
      vus: 20,
      iterations: 400,
    },
  },
};

```

3.5 Running the tests

Now we have a completed load test. Deploy the service to AWS:


```
$ terraform apply
```

Once the deployment is finished, you should be given a URL where the endpoint is deployed. Use that endpoint to set the ENDPOINT environment variable and run the tests.

```
> export ENDPOINT=...
> k6 run planners-and-studiers.js
```

These tests should take about 10 minutes to run. You should see that most of the tests pass but that there will be a few failures throughout, as seen below.

```
WARN[0152] Request Failed error="Get \"http://54.166.206.229:6400//api/v1/todos\":
request timeout"
WARN[0183] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
WARN[0239] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
WARN[0241] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
WARN[0272] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
WARN[0330] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
WARN[0339] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
WARN[0371] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
WARN[0412] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
WARN[0412] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
WARN[0467] Request Failed error="Get \"http://54.166.206.229:6400/api/v1/todos\":
request timeout"
```

```
running (09m30.0s), 0000/2520 VUs, 4914 complete and 1947 interrupted iterations
planner [ 100% ] 20 VUs 07m09.0s/10m0s 400/400 shared iters
studier [ 100% ] 0000/2500 VUs 9m0s
```

```
is status 200
94% 6646 / 411
```

```
checks.....: 94.17% 6646 411
data_received.....: 166 MB 292 kB/s
data_sent.....: 727 kB 1.3 kB/s
iterations.....: 4914 8.620734/s
vus.....: 1 min=1 max=2520
vus_max.....: 2520 min=2520 max=2520
```

4 Debugging

Now we can start to debug our service. We desire to not only increase our 94% pass rate to 100% but also to increase the amount of studies and planners.

4.1 Tips

Open TaskOverflow If you open TaskOverflow, you should find it has been left in an unusual state after the tests. See if you can find out from the logs how this occurred.

Follow an Interaction Use the correlation IDs that we have now included to see where potential bottlenecks may be occurring. You can introduce more fine-grained logs to find where time is being spent.

4.2 Back-of-the-envelope

Back-of-the-envelope calculations are normally quite helpful when trying to scale services. You need some rough estimates of how many requests your infrastructure can handle to help identify bottlenecks.

Our production service is deployed using Gunicorn, which is a Python WSGI HTTP Server for UNIX.² It is a pre-fork worker model, which means that it will create a number of worker processes when it starts. Each worker process can handle the amount of available threads concurrently. This means that the number of concurrent requests that can be handled by the service is the number of workers multiplied by the number of threads.

The number of sync gunicorn workers is determined by the number of CPUs:

$$workers = (2 * cpu) + 1$$

You can use the logs that you have produced to help measure the time it takes each request to process. We can then use this to calculate how many requests each worker can handle per second.

You can use the sleeps in the load tests and the number of virtual users to calculate how many requests are made per second.

The service also makes use of RDS. The formula for calculating the amount of concurrent connections RDS can handle is roughly:

$$LEAST(DBInstanceClassMemory/9531392, 5000)$$

Use these calculations to work out the amount of resources required to scale your service to handle the traffic produced by the tests.

²<https://docs.gunicorn.org/en/stable/design.html>