

**IMPORTANT** Please remember to destroy all the resources after each work session. You can recreate infrastructure by creating new PR and merging it to master.

The screenshot shows the GitHub Actions interface for the repository 'tbd-workshop-1'. The 'Actions' tab is selected. On the left, under 'Actions', the 'Destroy' workflow is chosen. The main area displays '22 workflow runs' for the 'destroy' workflow. One run is highlighted, showing it was triggered by a 'workflow\_dispatch' event and manually run by 'mwilewicz'. A modal window is open over the list, titled 'Run workflow', with the branch set to 'master'. A green 'Run workflow' button is visible at the bottom of the modal.

## 1. Authors:

14

<https://github.com/Tuvalu120/tbd-workshop-1>

## 2. Follow all steps in README.md.

## 3. From available Github Actions select and run destroy on main branch.

## 4. Create new git branch and:

### 1. Modify tasks-phase1.md file.

### 2. Create PR from this branch to **YOUR** master and merge it to make new release.

The screenshot shows the GitHub Actions interface for the repository 'tbd-workshop-1'. The 'Actions' tab is selected. On the left, under 'Actions', the 'All workflows' option is chosen. The main area displays '5 workflow runs' across different branches. The runs are: 
 

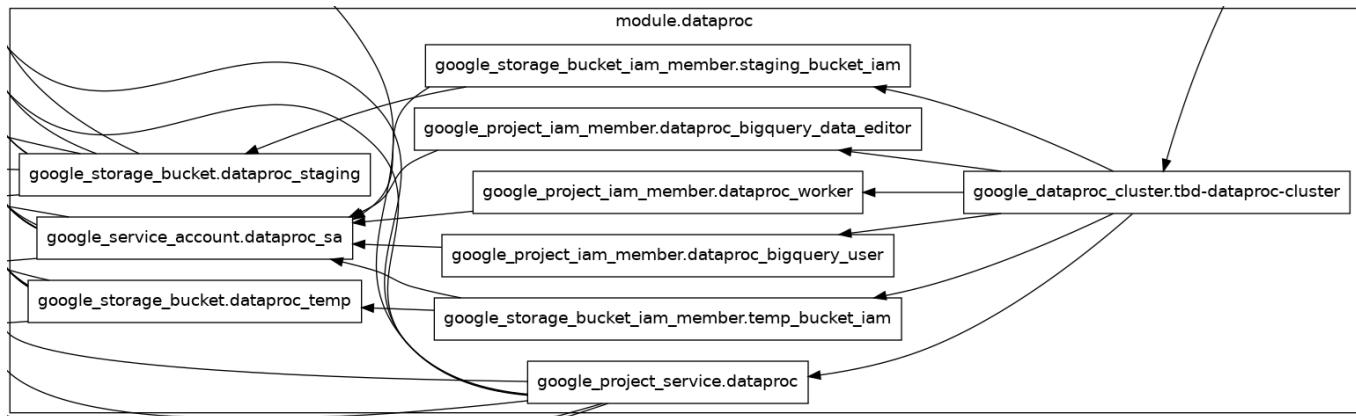
- Merge pull request #1 from Tuvalu120/task-4 (master, 34 minutes ago)
- Change tasks-phase1.md (task-4, 36 minutes ago)
- Destroy (master, Nov 8, 1:52 PM GMT+1)
- Initialize repo (master, Nov 8, 11:15 AM GMT+1)
- Destroy (master, Oct 30, 10:05 PM GMT+1)

 A modal window is open over the list, titled 'Run workflow', with the branch set to 'master'. A green 'Run workflow' button is visible at the bottom of the modal.

## 5. Analyze terraform code. Play with terraform plan, terraform graph to investigate different modules.

Moduł dataproc służy do stworzenia i konfiguracji klastra Apache Spark w wykorzystywanej w tym ćwiczeniu usłudze Google Cloud Dataproc. W architekturze naszego projektu pełni rolę środowiska obliczeniowego dla zadań przetwarzania danych oraz środowiska interaktywnego. Moduł ten jest ściśle powiązany z modułem sieciowym vpc (depends\_on = [module.vpc]), czyli wymaga on uprzednio utworzonej sieci vpc. Zadaniem modułu jest zapewnienie, że sieć vpc jest gotowa, stworzenie klastra

w określonym projekcie (project\_name = var.project\_name) i regionie (var.region), umieszczenie węzłów klastra w odpowiedniej podsieci vpc (subnet = module.vpc.subnets[local.notebook\_subnet\_id].id). Dodatkowo instaluje on oprogramowanie oparte na obrazie Dataproc w wersji 2.2 z Ubuntu 22 (image\_version = "2.2.69-ubuntu22") i używa maszyn wirtualnych typu e2-standard-2 (machine\_type = "e2-standard-2").



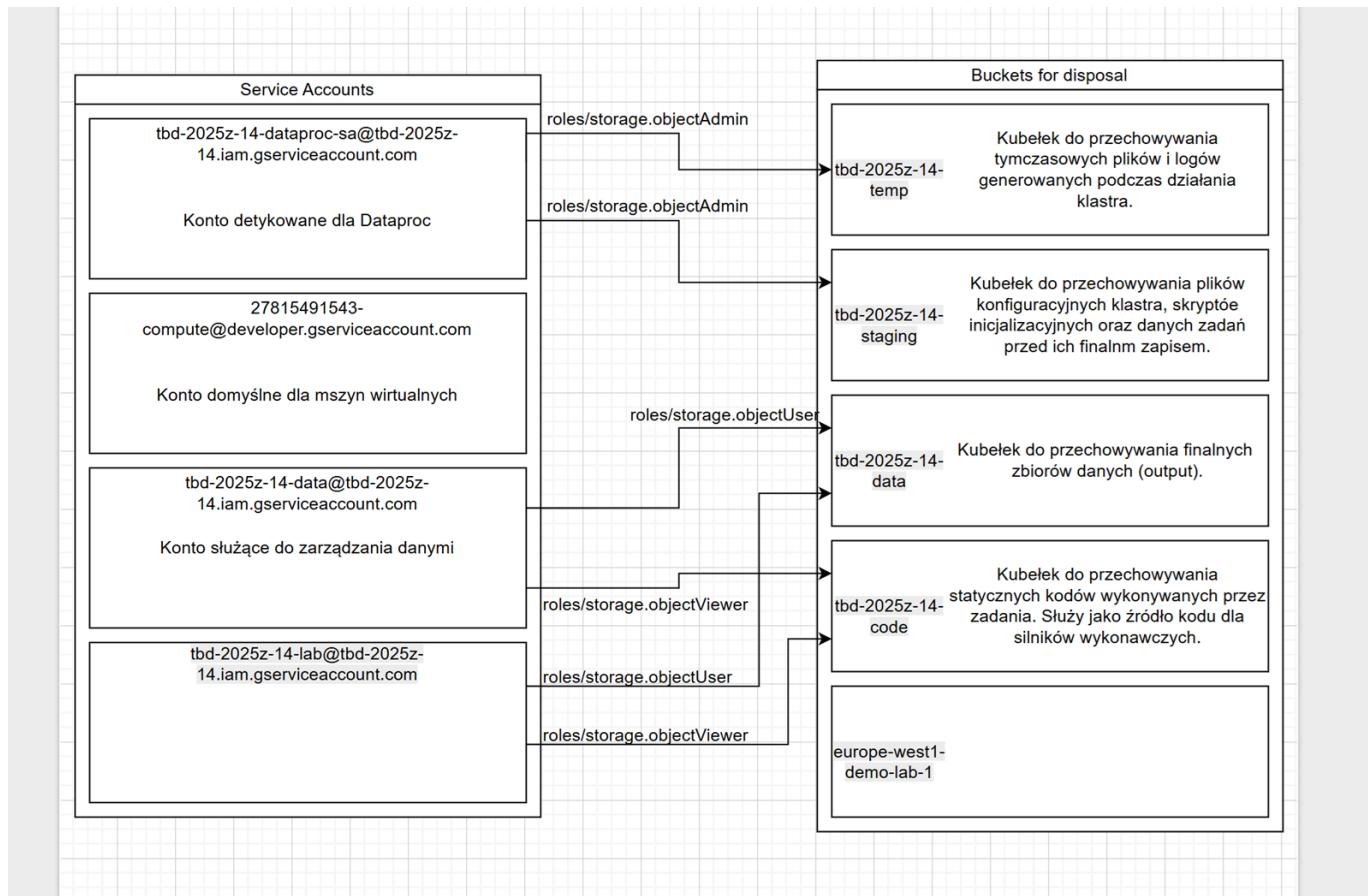
## 6. Reach YARN UI

```
gcloud compute ssh --zone "europe-west1-b" "tbd-cluster-m" --tunnel-through-iap --project "tbd-2025z-14" -- -L 8088:localhost:8088
```

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime
application_1763904207222_0003	root	Shakespeare WordCount	SPARK	dataproj_hash_93637cc0-fa69-3c36-91ad-33ad8961c209.dataproj_job_70b0bacb-cd1c-4141-93b8-c4c03d5b524.dataproj_job_attempt_timestamp_1763906572136.dataproj_master_index_0.dataproj_uuid_70b0bacb-cd1c-4141-93b8-8c4c03d5b524	default	0	Sun Nov 23 15:03:09 +0100 2025	N/A
application_1763904207222_0002	root	Shakespeare WordCount	SPARK	dataproj_hash_149ee1fb-eb34-374c-ba74-6881a36d5f587.dataproj_job_attempt_timestamp_1763906185033.dataproj_job_fd97cf02-1ebb-488f-9c50-d1f36325a15.dataproj_master_index_0.dataproj_uuid_fd97cf02-1ebb-488f-9c50-d1f36325a15	default	0	Sun Nov 23 14:56:44 +0100 2025	N/A
application_1763904207222_0001	root	Shakespeare WordCount	SPARK	dataproj_hash_ba920a93-2a54-3386-add1-d88148921eb.dataproj_job_21f907a2-d67f-420f-9ea3-18af9e4cef89.dataproj_job_attempt_timestamp_1763905786911.dataproj_master_index_0.dataproj_uuid_21f907a2-d67f-420f-9ea3-18af9e4cef89	default	0	Sun Nov 23 14:50:08 +0100 2025	N/A

## 7. Draw an architecture diagram (e.g. in draw.io) that includes:

1. Description of the components of service accounts
2. List of buckets for disposal



8. Create a new PR and add costs by entering the expected consumption into Infracost For all the resources of type: `google_artifact_registry`, `google_storage_bucket`, `google_service_networking_connection` create a sample usage profiles and add it to the Infracost task in CI/CD pipeline. Usage file [example](#)

github-actions bot commented 33 minutes ago • edited

Infracost report

Monthly estimate increased by \$92 🚨

Changed project	Baseline cost	Usage cost*	Total change	New monthly cost
Tuvalu120/tbd-workshop-1	+\$0	+\$65	+\$65	\$65
Tuvalu120/tbd-workshop-1/bootstrap	+\$0	+\$13	+\$13	\$13
Tuvalu120/tbd-workshop-1/mlops	+\$0.30	+\$14	+\$14 (+53%)	\$42

\*Usage costs were estimated using `infracost-usage.yml`, see [docs](#) for other options.

▼ Estimate details (includes details of unsupported resources)

Key: \* usage cost, ~ changed, + added, - removed

9. Create a BigQuery dataset and an external table using SQL

```
CREATE SCHEMA IF NOT EXISTS dataset;

CREATE OR REPLACE EXTERNAL TABLE dataset.shakespeare
OPTIONS (
```

```

format = 'ORC',
uris = ['gs://tbd-2025z-14-data/data/shakespeare/.orc']);

SELECT * FROM dataset.shakespeare
ORDER BY sum_word_count;

```

The screenshot shows the Google Cloud BigQuery interface. On the left, there's a sidebar with project navigation and a search bar. The main area has a toolbar with various icons. A tab labeled "Untitled query" is active, showing the SQL code run earlier:

```

1 CREATE SCHEMA IF NOT EXISTS dataset;
2
3 CREATE OR REPLACE EXTERNAL TABLE dataset.shakespeare
4   OPTIONS (
5     format = 'ORC',
6     uris = ['gs://tbd-2025z-14-data/data/shakespeare/*.orc']);
7
8
9 SELECT * FROM dataset.shakespeare
10 ORDER BY sum_word_count DESC;
11

```

Below the code, a message says "Query completed". A summary table shows the job details:

Elapsed time	Statements processed	Job status
6 sec	3	SUCCESS

Under "Job history", three rows are listed:

Status	End time	SQL	Stages completed	Bytes processed	Action
✓	16:08 [1:1]	CREATE SCHEMA IF NOT EXISTS dataset	0	0 B	<a href="#">View results</a>
✓	16:08 [3:1]	CREATE OR REPLACE EXTERNAL TABLE dataset.shakespeare	0	0 B	<a href="#">View results</a>
✓	16:08 [9:3]	SELECT * FROM dataset.shakespeare	2	469.62 KB	<a href="#">View results</a>

At the bottom, a table displays the top 5 words from the Shakespeare dataset:

Row	word	sum_word_count
1	the	25568
2	I	21028
3	and	19649
4	to	17361
5	of	16438

ORC to format plików używany do przechowywania danych w systemach Big Data. Jedną z jego głównych zalet jest to, iż pliki ORC przechowują informacje o swoim schemacie (nazwy kolumn i typy danych) wewnątrz struktury. Dzięki temu systemy takie jak BigQuery mogą odczytać plik i automatycznie odtworzyć strukturę danych, bez konieczności ręcznego definiowania schematu podczas tworzenia tabeli zewnętrznej.

#### 10. Find and correct the error in spark-job.py

Problem był spowodowany tym, iż w pliku spark-job.py zawarty był niewłaściwy Bucket. Informacje o złym Buckecie można było odczytać z logów nieudanego joba w konsoli Goggle Cloud i dataproc. Po zamianie nazwy Bucketa na:

```
DATA_BUCKET = "gs://tbd-2025z-14-data/data/shakespeare/"
```

job został wykonany pomyślnie.

```
The top words in shakespeare are
25/11/24 14:35:38 INFO GhfsGlobalStorageStatistics: periodic connector metrics: {action_http_delete_request=2, action_http_delete_request_duration=76,
[CONTEXT ratelimit_period="5 MINUTES" ]
+-----+
|word|sum_word_count|
+-----+
| the| 25568|
| I| 21028|
| and| 19649|
| to| 17361|
| of| 16438|
| a| 13409|
| you| 12527|
| my| 11291|
| in| 10589|
| is| 8735|
|that| 8561|
| not| 8395|
| me| 8030|
| And| 7780|
|with| 7224|
| it| 7137|
| his| 6811|
| be| 6724|
|your| 6244|
| for| 6154|
+-----+
only showing top 20 rows
```

<input type="checkbox"/> Job ID	Status	Region	Type	Cluster	Start time
<a href="#">d53d2e1f-b099-41d4-a9dd-805aa2eb27c9</a>	<span style="color: green;">✓ Succeeded</span>	europe-west1	PySpark	<a href="#">tbd-cluster</a>	24 Nov 2025,

## 11. Add support for preemptible/spot instances in a Dataproc cluster

```
[modules/dataproc/main.tf] (modules/dataproc/main.tf)

~~~
preemptible_worker_config {
    num_instances = 2
    preemptibility = "SPOT"

    disk_config {
        boot_disk_type      = "pd-standard"
        boot_disk_size_gb = 100
    }
}
~~~
```

## 12. Triggered Terraform Destroy on Schedule or After PR Merge. Goal: make sure we never forget to clean up resources and burn money.

Add a new GitHub Actions workflow that:

1. runs `terraform destroy -auto-approve`
2. triggers automatically:
  - a) on a fixed schedule (e.g. every day at 20:00 UTC)
  - b) when a PR is merged to main containing [CLEANUP] tag in title

Steps:

1. Create file `.github/workflows/auto-destroy.yml`
2. Configure it to authenticate and destroy Terraform resources
3. Test the trigger (schedule or cleanup-tagged PR)

<.github/workflows/auto-destroy.yml>

The screenshot shows the GitHub Actions interface for the 'Auto Terraform Destroy' workflow. At the top, there's a 'Help us improve GitHub Actions' survey. Below it, a summary table shows two workflow runs:

Run	Event	Status	Branch	Actor	Created At	Last Updated	Duration
Auto Terraform Destroy #2: Scheduled	Scheduled	Completed	master	report-doc	Nov 23, 9:23 PM GMT+1	Nov 23, 9:23 PM GMT+1	20s
[CLEANUP] Add several info into tasks-phase1.md	Pull request #4 closed	Completed	report-doc	Tuvalu120	Nov 23, 6:20 PM GMT+1	Nov 23, 6:20 PM GMT+1	8m 12s

Below the table, a detailed view of the second run is shown. It includes a 'Summary' section with links to 'Run details', 'Usage', and 'Workflow file'. The main area displays the steps of the 'terraform-destroy' job:

- > Set up job (0s)
- > Checkout repo (1s)
- > Setup Terraform (2s)
- > Authenticate to Google Cloud (0s)
- > Terraform Init (7s)
- > Terraform Destroy (3s)
- > Post Authenticate to Google Cloud (0s)
- > Post Checkout repo (0s)
- > Complete job (0s)

There are also 'Search logs' and 'Re-run all jobs' buttons at the bottom.

Automatyczne czyszczenie środowiska o zadanej godzinie pomaga zminimalizować zużycie zasobów, kiedy nie są one używane (np. kiedy ktoś zapomniał o jego czyszczeniu po skończonej pracy), a czyszczenie środowiska z użyciem taga [CLEANUP] znaczco ułatwia osobom zajmującym się projektem uporządkowanie zasobów po skończonej pracy nad projektem.