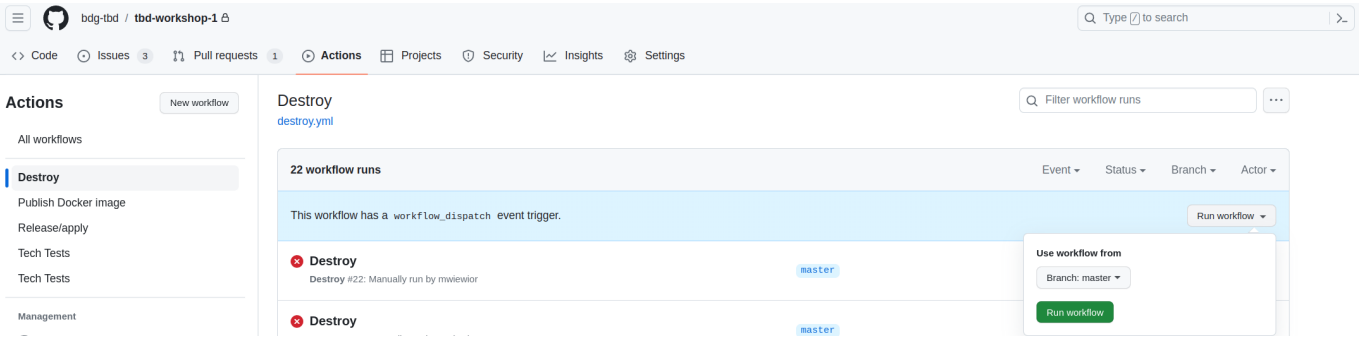
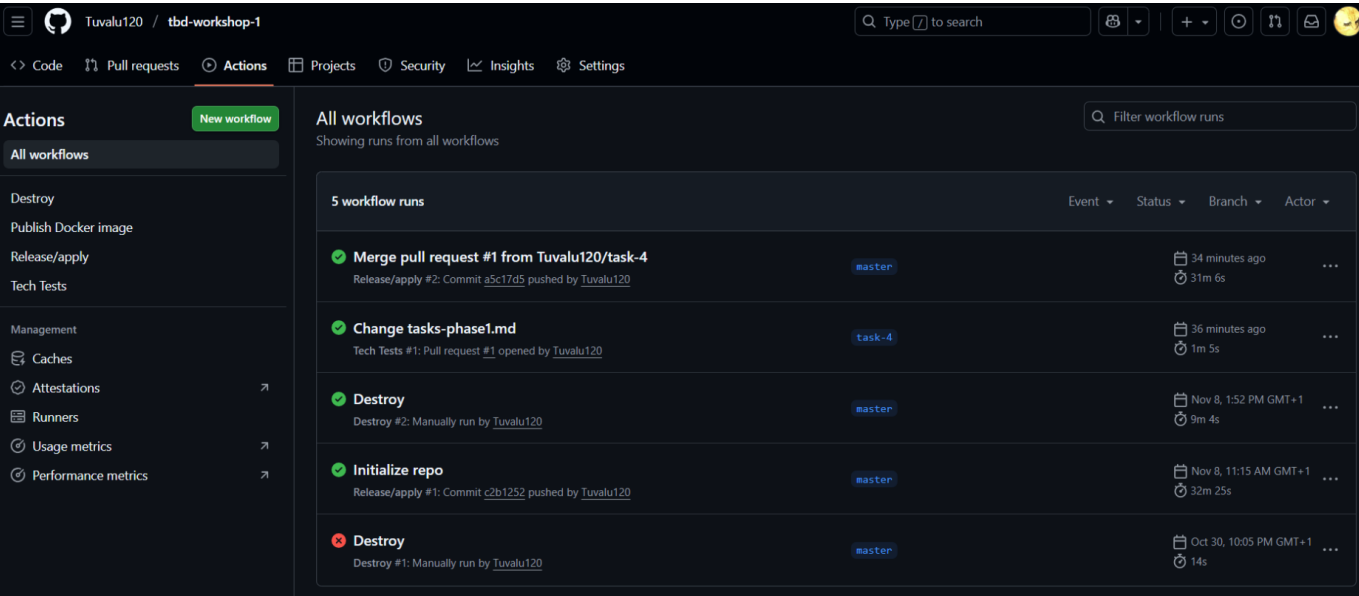


IMPORTANT Please remember to destroy all the resources after each work session. You can recreate infrastructure by creating new PR and merging it to master.



1. Authors:
- 14
- <https://github.com/Tuvalu120/tbd-workshop-1>
2. Follow all steps in README.md.
3. From avaialble Github Actions select and run destroy on main branch.
4. Create new git branch and:
1. Modify tasks-phase1.md file.

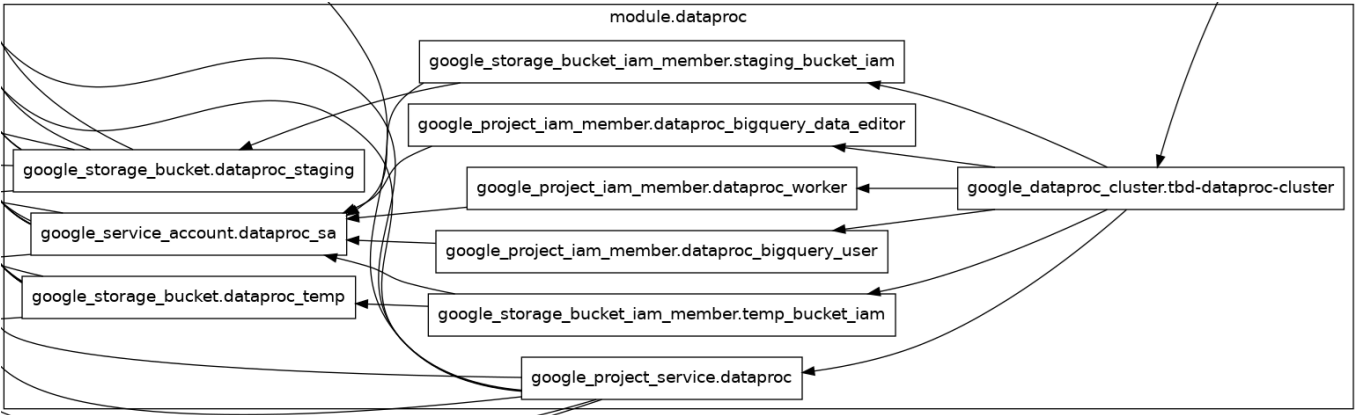
2. Create PR from this branch to **YOUR** master and merge it to make new release.



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
```



All Applicati

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

Cluster Metrics

Apps Submitted	0	Apps Pending	0	Apps Running	3	Apps Completed	0	Containers Running		Used Resources	
										<memory:0 B, vCores 0>	<m

Cluster Nodes Metrics

Active Nodes	0	Decommissioning Nodes	0	Decommissioned Nodes	0
--------------	---	-----------------------	---	----------------------	---

Scheduler Metrics

Scheduler Type	Capacity Scheduler	Scheduling Resource Type	[memory-mb (unit=Mi), vcores]	Minimum Allocation	<memory:1, vCores:1>
----------------	--------------------	--------------------------	-------------------------------	--------------------	----------------------

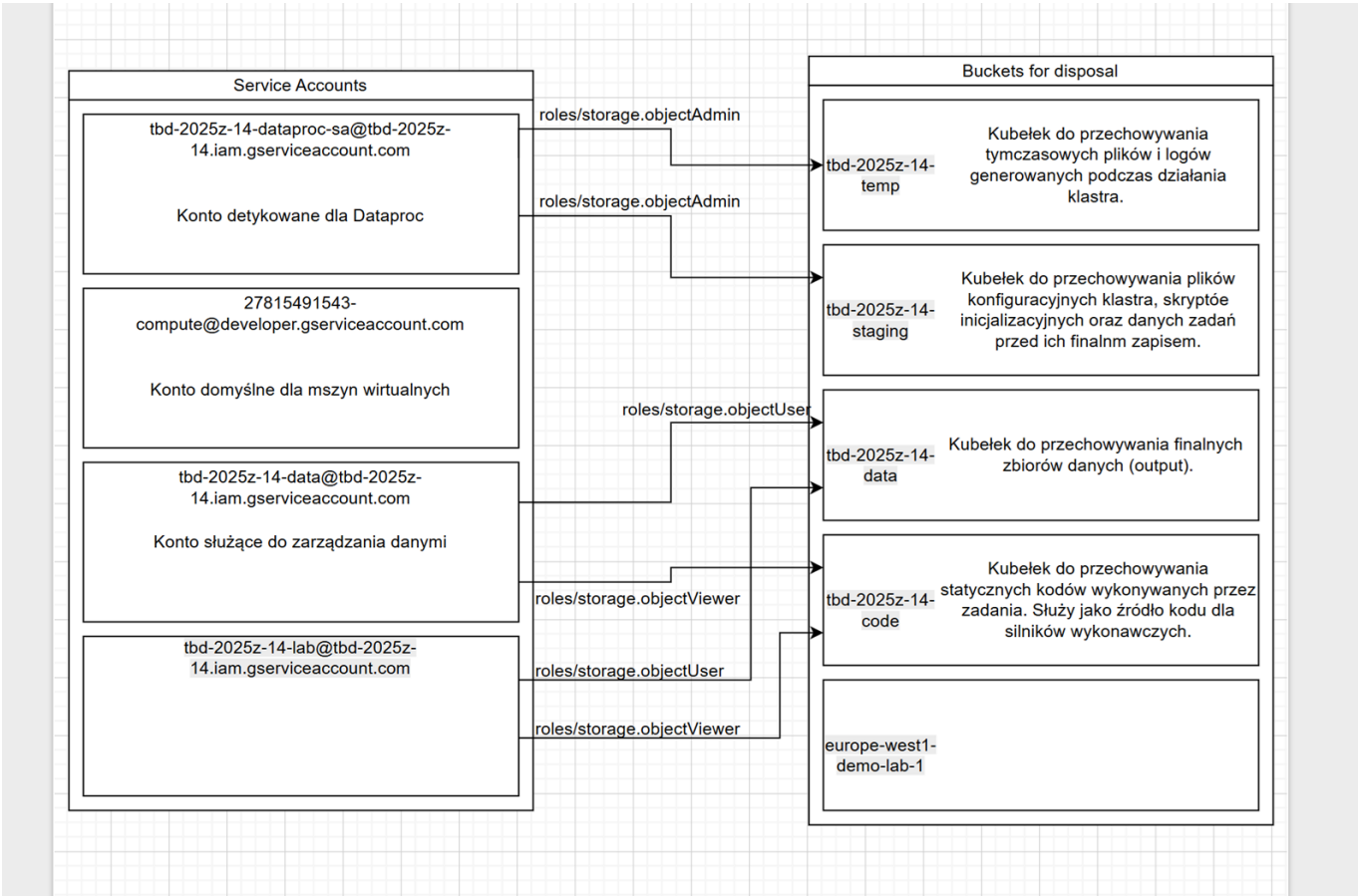
Show: 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTi
application_1763904207222_0003	root	Shakespeare WordCount	SPARK	dataproc_hash_93637cc0-fa69-3c36-91ad-33ad8961c209.dataproc_job_70b0bacb-cd1c-4141-933b-8c4c03d58524.dataproc_job_attempt_timestamp_1763906572136.dataproc_master_index_0.dataproc_uuid_70b0bacb-cd1c-4141-933b-8c4c03d58524	default	0	Sun Nov 23 15:03:09 +0100 2025	N/A
application_1763904207222_0002	root	Shakespeare WordCount	SPARK	dataproc_hash_149eeef5b-eb34-374c-ba74-6881a36df587.dataproc_job_attempt_timestamp_1763906185033.dataproc_job_f997cf02-1ebb-488f-9c50-d1fe36325a15.dataproc_master_index_0.dataproc_uuid_f997cf02-1ebb-488f-9c50-d1fe36325a15	default	0	Sun Nov 23 14:56:44 +0100 2025	N/A
application_1763904207222_0001	root	Shakespeare WordCount	SPARK	dataproc_hash_ba920a93-2a54-3386-add1-d88148921eab.dataproc_job_21f907a2-d87f-420f-9ea3-18a9e4cef89.dataproc_job_attempt_timestamp_1763905786911.dataproc_master_index_0.dataproc_uuid_21f907a2-d87f-420f-9ea3-18a9e4cef89	default	0	Sun Nov 23 14:50:08 +0100 2025	N/A


Showing 1 to 3 of 3 entries


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;
```

Google Cloud

TBD tbd-2025z-14 project

Search (/) for resources, docs, products and more

Q Search

◆ 📦 📄 🔔 ? ⋮

B

🔍 Search BigQuery resources

Show starred only

tbd-2025z-14

Repositories

Queries

Notebooks

Data canvases

Data preparations

Pipelines

Connections

dataset

shakespeare

Untitled query

Run Save Download Share Schedule Open in More

```
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 SELECT * FROM dataset.shakespeare
9 ORDER BY sum_word_count DESC;
10
11
```

Query completed

All results

Elapsed time

Statements processed

Job status

6 sec

3

✓ SUCCESS

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

Job history

Show

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
<input type="checkbox"/>	d53d2e1f-b099-41d4-a9dd-805aa2eb27c9	✔ Succeeded	europa-west1	PySpark	tbd-cluster	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](#)

Auto Terraform Destroy

auto-destroy.yml

Filter workflow runs

...

Help us improve GitHub Actions

Tell us how to make GitHub Actions work better for you with three quick questions.

Give feedback

×

2 workflow runs

Event ▾Status ▾Branch ▾Actor ▾

Auto Terraform Destroy

Auto Terraform Destroy #2: Scheduled

master

Nov 23, 9:23 PM GMT+1

20s

...

[CLEANUP] Add several info into tasks-phase1.md

Auto Terraform Destroy #1: Pull request #4 closed by Tuvalu120

report-doc

Nov 23, 6:20 PM GMT+1

8m 12s

...

Auto Terraform Destroy

Auto Terraform Destroy #2

Re-run all jobs

...

Summary

Jobs

Run details

Usage

Workflow file

terraform-destroy

succeeded 18 hours ago in 15s

Search logs

↺

⚙

Set up job

Checkout repo

Setup Terraform

Authenticate to Google Cloud

Terraform Init

Terraform Destroy

Post Authenticate to Google Cloud

Post Checkout repo

Complete job

0s

1s

2s

0s

7s

3s

0s

0s

0s

Automatycznie 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] znacząco ułatwia osobom zajmującym się projektem uporządkowanie zasobów po skończonej pracy nad projektem.