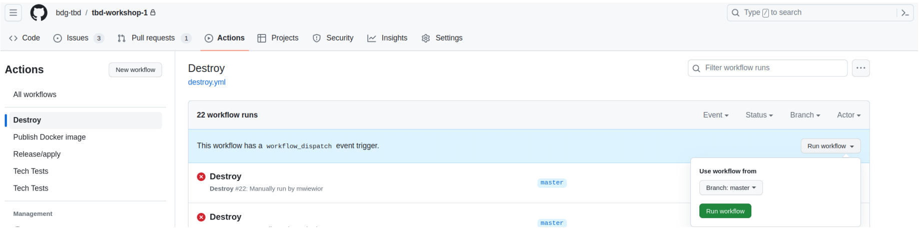
IMPORTANT Please remember to destroy all the resources after each work session. You can recreate infrastructure by creating new PR and merging it to master.



This workflow has a `workflow_dispatch` event trigger.

Destroy #22: Manually run by mwiewior

Figure 1: img.png

1. Authors:

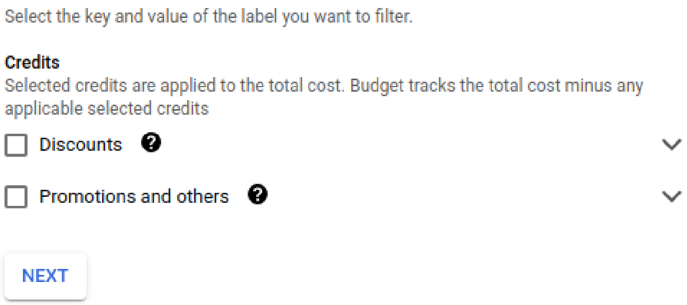
   Grupa nr 7

   - Adam Górski
   - Zuzanna Górecka
   - Michał Oracki

   Link to repo

2. Follow all steps in README.md.

3. Select your project and set budget alerts on 5%, 25%, 50%, 80% of 50$ (in cloud console -> billing -> budget & alerts -> create buget; unclick discounts and promotions&others while creating budget).



Select the key and value of the label you want to filter.

Credits

Selected credits are applied to the total cost. Budget tracks the total cost minus any applicable selected credits

☐ Discounts

☐ Promotions and others

Figure 2: img.png

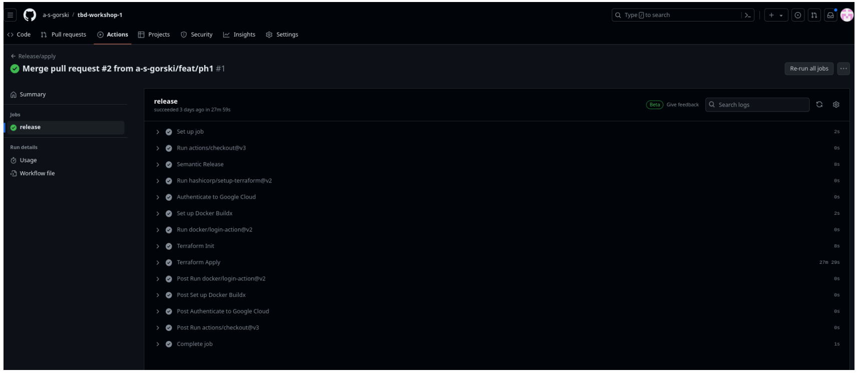
5. From avaialble Github Actions select and run destroy on main branch.

6. 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.

place the screenshot from GA after succesfull application of release



a-s-gorski / tbd-workshop-1

Code | Pull requests | Actions | Projects | Security | Insights | Settings

Release/apply

Merge pull request #2 from a-s-gorski/feat/ph1 #1

Re-run all jobs

Summary

Jobs

release

Run details

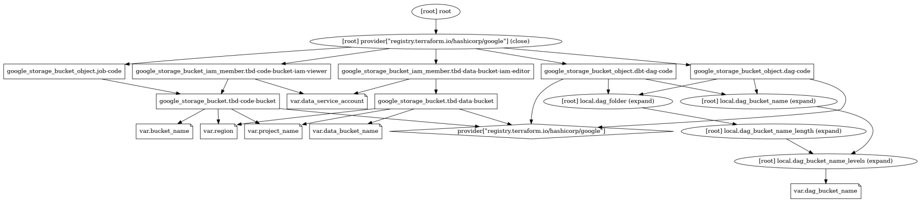
Usage

Workflow file

release
succeeded 3 days ago in 27m 59s

```
Set up job                          2s
Run actions/checkout@v3             0s
Semantic Release                    8s
Run hashicorp/setup-terraform@v2    0s
Authenticate to Google Cloud        0s
Set up Docker Buildx                2s
Run docker/login-action@v2          0s
Terraform Init                      8s
Terraform Apply                27m 29s
Post Run docker/login-action@v2     0s
Post Set up Docker Buildx           0s
Post Authenticate to Google Cloud   0s
Post Run actions/checkout@v3        0s
Complete job                        1s
```

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



[root] root
[root] provider["registry.terraform.io/hashicorp/google"] (close)
google_storage_bucket_object.job-code
google_storage_bucket_iam_member.tbd-code-bucket-iam-viewer
google_storage_bucket_iam_member.tbd-data-bucket-iam-editor
google_storage_bucket_object.dbt-dag-code
google_storage_bucket_object.dag-code
google_storage_bucket.tbd-code-bucket
var.data_service_account
google_storage_bucket.tbd-data-bucket
[root] local.dag_folder (expand)
[root] local.dag_bucket_name (expand)
var.bucket_name
var.region
var.project_name
var.data_bucket_name
provider["registry.terraform.io/hashicorp/google"]
[root] local.dag_bucket_name_length (expand)
[root] local.dag_bucket_name_levels (expand)
var.dag_bucket_name

Moduł, który przeanalizowalśmy to data-pipeline. Tworzy on joby dla dgt i dag oraz trzyma dane wyjściowe z nich w bucketach.

8. Check if pyspark kernel exists - if not then in your Jupyterlab enviroment add Python3.8 kernel

Pyspark kernel nie istniał i został utworzony przy pomocy polecenia:

```
python3.8 -m ipykernel install --user --name pyspark
```

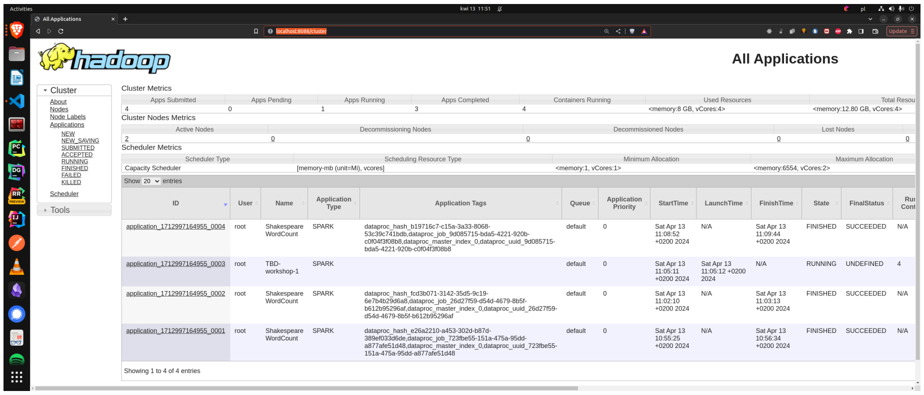
9. Reach YARN UI

```
gcloud compute ssh --zone "europe-west1-d" "tbd-cluster-m" --tunnel-through-iap --proj
```

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

    1. VPC topology with service assignment to subnets
    2. Description of the components of service accounts
    3. List of buckets for disposal
    4. Description of network communication (ports, why it is necessary to specify the host for the driver) of Apache Spark running from Vertex AI Workbech

W Apache Spark węzeł główny zarządza i dystrybuuje zadania do odpowiednich wykonawców. Dlatego potrzebne jest podanie hosta, aby wykonawcy wiedzieli



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resou
4	0	1	3	4	<memory:8 GB, vCores:4>	<memory:12.80 GB, vCores:4>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	
2	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1, vCores:1>	<memory:6554, vCores:2>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Ru Cont
application_1712997164955_0004	root	Shakespeare WordCount	SPARK	dataproc_hash_b19716c7-c15a-3a33-8068-53c39c741bdb,dataproc_job_9d085715-bda5-4221-920b-c0f04f3f08b8,dataproc_master_index_0,dataproc_uuid_9d085715-bda5-4221-920b-c0f04f3f08b8	default	0	Sat Apr 13 11:08:52 +0200 2024	N/A	Sat Apr 13 11:09:44 +0200 2024	FINISHED	SUCCEEDED	N/A
application_1712997164955_0003	root	TBD-workshop-1	SPARK		default	0	Sat Apr 13 11:05:11 +0200 2024	Sat Apr 13 11:05:12 +0200 2024	N/A	RUNNING	UNDEFINED	4
application_1712997164955_0002	root	Shakespeare WordCount	SPARK	dataproc_hash_fcd3b071-3142-35d5-9c19-6e7b4b29d6a8,dataproc_job_26d27f59-d54d-4679-8b5f-b612b95296af,dataproc_master_index_0,dataproc_uuid_26d27f59-d54d-4679-8b5f-b612b95296af	default	0	Sat Apr 13 11:02:10 +0200 2024	N/A	Sat Apr 13 11:03:13 +0200 2024	FINISHED	SUCCEEDED	N/A
application_1712997164955_0001	root	Shakespeare WordCount	SPARK	dataproc_hash_e26a2210-a453-302d-b87d-389ef033d6de,dataproc_job_723fbe55-151a-475a-95dd-a877afe51d48,dataproc_master_index_0,dataproc_uuid_723fbe55-151a-475a-95dd-a877afe51d48	default	0	Sat Apr 13 10:55:25 +0200 2024	N/A	Sat Apr 13 10:56:34 +0200 2024	FINISHED	SUCCEEDED	N/A

Showing 1 to 4 of 4 entries

Figure 3: yarn_ui

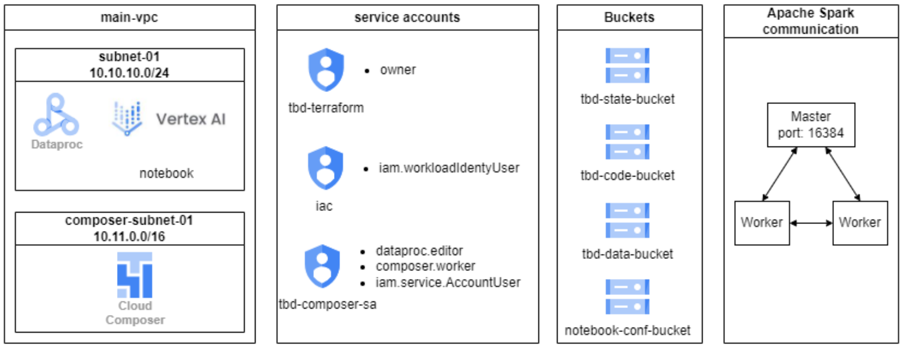


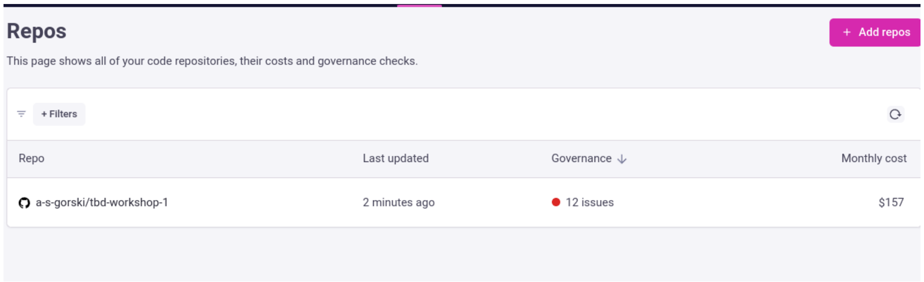
Figure 4: architecture_diagram.drawio

skąd nasłuchiwać, z kim się komunikować i gdzie wysyłać wyniki swojej pracy.

11. 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

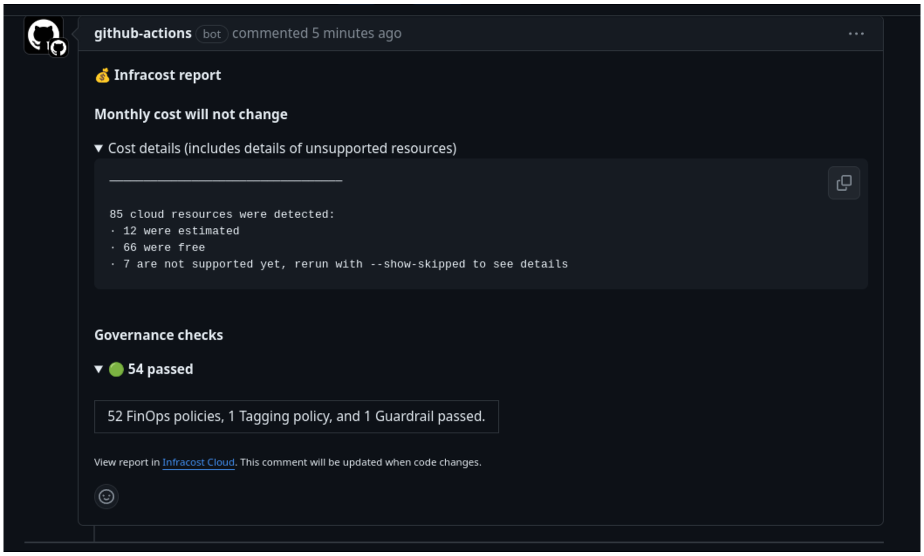
place the expected consumption you entered here

```
version: 0.1

google_artifact_registry_repository.registry:
    storage_gb: 15 # Total data stored in the repository in GB
    monthly_egress_data_transfer_gb: # Monthly data delivered from the artifact registry r
    europe_west1: 10 # GB of data delivered from the artifact registry to europe-north1.


google_storage_bucket.my_bucket:
storage_gb: 150                   # Total size of bucket in GB.
monthly_class_a_operations: 40 # Monthly number of class A operations (object adds, bucket/
monthly_class_b_operations: 20 # Monthly number of class B operations (object gets, retrie
monthly_data_retrieval_gb: 5    # Monthly amount of data retrieved in GB.
monthly_egress_data_transfer_gb:  # Monthly data transfer from Cloud Storage to the follow
    same_continent: 55

google_storage_bucket.mlflow_artifacts_bucket:
storage_gb: 150                   # Total size of bucket in GB.
monthly_class_a_operations: 40 # Monthly number of class A operations (object adds, bucket/
monthly_class_b_operations: 20 # Monthly number of class B operations (object gets, retrie
monthly_data_retrieval_gb: 5    # Monthly amount of data retrieved in GB.
monthly_egress_data_transfer_gb:  # Monthly data transfer from Cloud Storage to the follow
    same_continent: 55


google_storage_bucket.tbd-state-bucket:
storage_gb: 150                   # Total size of bucket in GB.
monthly_class_a_operations: 40 # Monthly number of class A operations (object adds, bucket/
monthly_class_b_operations: 20 # Monthly number of class B operations (object gets, retrie
monthly_data_retrieval_gb: 5    # Monthly amount of data retrieved in GB.
monthly_egress_data_transfer_gb:  # Monthly data transfer from Cloud Storage to the follow
    same_continent: 55

google_storage_bucket.tbd-code-bucket:
storage_gb: 150                   # Total size of bucket in GB.
monthly_class_a_operations: 40 # Monthly number of class A operations (object adds, bucket/
monthly_class_b_operations: 20 # Monthly number of class B operations (object gets, retrie
monthly_data_retrieval_gb: 5    # Monthly amount of data retrieved in GB.
monthly_egress_data_transfer_gb:  # Monthly data transfer from Cloud Storage to the follow
    same_continent: 55

google_storage_bucket.tbd-data-bucket:
storage_gb: 150                   # Total size of bucket in GB.
monthly_class_a_operations: 40 # Monthly number of class A operations (object adds, bucket/
monthly_class_b_operations: 20 # Monthly number of class B operations (object gets, retrie
monthly_data_retrieval_gb: 5    # Monthly amount of data retrieved in GB.
monthly_egress_data_transfer_gb:  # Monthly data transfer from Cloud Storage to the follow
    same_continent: 55

google_storage_bucket.notebook-conf-bucket:
storage_gb: 150                   # Total size of bucket in GB.
monthly_class_a_operations: 40 # Monthly number of class A operations (object adds, bucket/
monthly_class_b_operations: 20 # Monthly number of class B operations (object gets, retrie
monthly_data_retrieval_gb: 5    # Monthly amount of data retrieved in GB.
monthly_egress_data_transfer_gb:  # Monthly data transfer from Cloud Storage to the follow
    same_continent: 55


google_service_networking_connection.private_vpc_connection:
    monthly_egress_data_transfer_gb: # Monthly VM-VM data transfer from VPN gateway to the
    same_region: 250                # VMs in the same Google Cloud region.
    europe: 70                      # Between Google Cloud regions within Europe.
    worldwide: 200                  # to a Google Cloud region on another continent.
```



This page shows all of your code repositories, their costs and governance checks.

Repo	Last updated	Governance ↓	Monthly cost
a-s-gorski/tbd-workshop-1	2 minutes ago	● 12 issues	$157



github-actions bot commented 5 minutes ago

💰 **Infracost report**

Monthly cost will not change

▼ Cost details (includes details of unsupported resources)

```
──────────────────────────────────

85 cloud resources were detected:
∙ 12 were estimated
∙ 66 were free
∙ 7 are not supported yet, rerun with --show-skipped to see details
```

Governance checks

▼ 🟢 **54 passed**

52 FinOps policies, 1 Tagging policy, and 1 Guardrail passed.

View report in Infracost Cloud. This comment will be updated when code changes.

Dodatkowo poza dodaniem jako część CI/CD wykonaliśmy też instrukcję lokalnie:

```
infracost breakdown --sync-usage-file --usage-file infracost-usage.yml --path .
```

i dostaliśmy takie wyniki

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

ORC nie wymaga schematu, ponieważ konektor jest w stanie go samemu inferować.

12. Start an interactive session from Vertex AI workbench:

13. Find and correct the error in spark-job.py

Gdy weszliśmy w szczegóły DAG uzyskaliśmy link do logów dla jobu, w którym pojawił się błąd. Następnie po wczytaniu się w szczegóły, okazało się, że kubełek nie istnieje (gs://tbd-2024l-9910-data/data/shakespeare/), żeby to naprawić podmieniliśmy nazwę kubełka na nasz własny (gs://tbd-2024l-3040540-data/data/shakespeare).

```
DATA_BUCKET = "gs://tbd-2024l-3040540-data/data/shakespeare/"
```

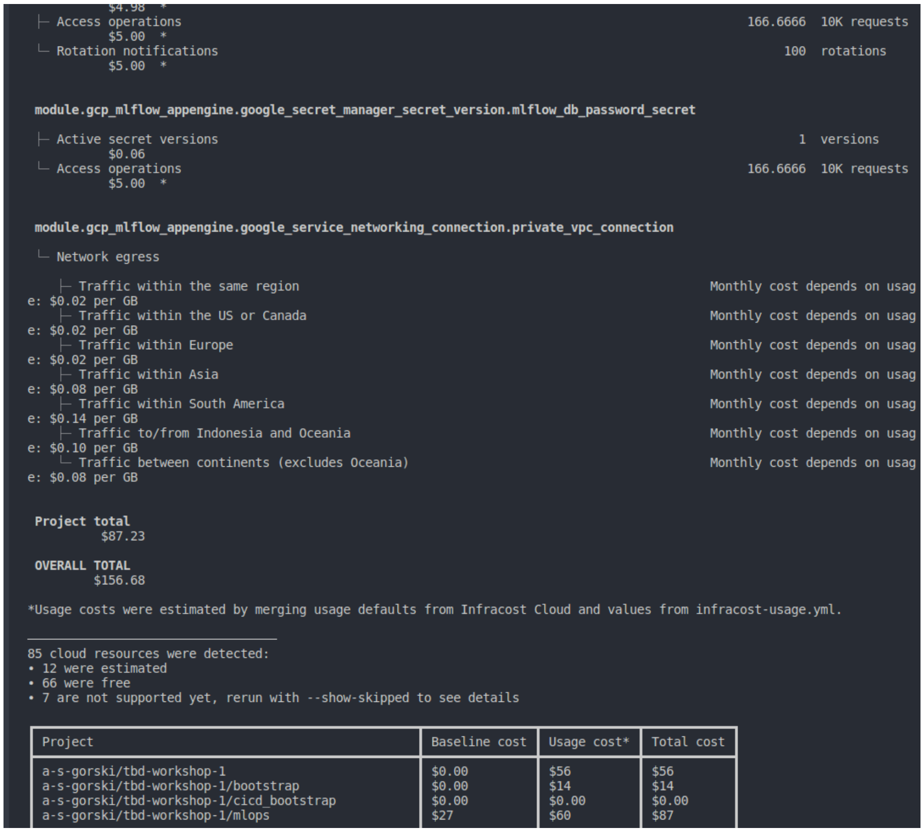
14. Additional tasks using Terraform:

    1. Add support for arbitrary machine types and worker nodes for a Dataproc cluster and JupyterLab instance

    Main dataproc Variables dataproc Main vertex_ai Variables vertex_ai

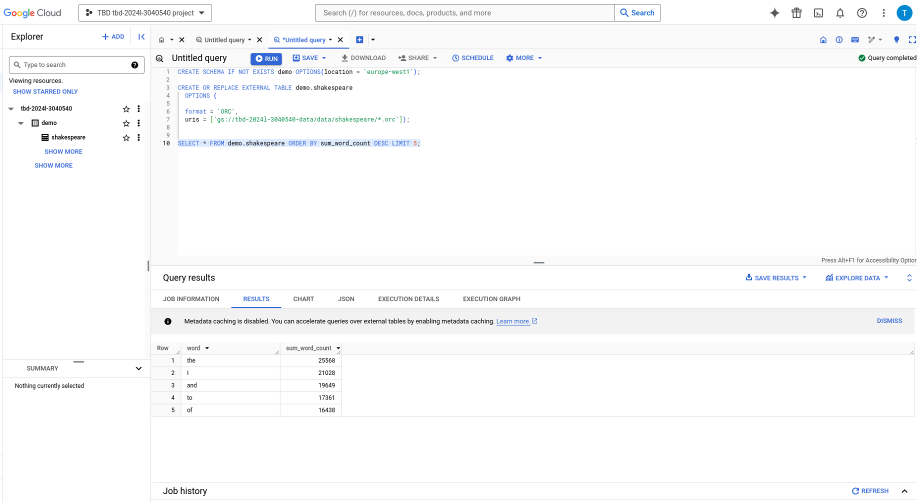
    3. Add support for preemptible/spot instances in a Dataproc cluster

    Link Text



```
          $4.98  *
 ├─ Access operations                                                         166.6666  10K requests
          $5.00  *
 └─ Rotation notifications                                                         100  rotations
          $5.00  *


module.gcp_mlflow_appengine.google_secret_manager_secret_version.mlflow_db_password_secret

 ├─ Active secret versions                                                           1  versions
          $0.06
 └─ Access operations                                                         166.6666  10K requests
          $5.00  *


module.gcp_mlflow_appengine.google_service_networking_connection.private_vpc_connection

 └─ Network egress

    ├─ Traffic within the same region                                    Monthly cost depends on usag
e: $0.02 per GB
    ├─ Traffic within the US or Canada                                   Monthly cost depends on usag
e: $0.02 per GB
    ├─ Traffic within Europe                                             Monthly cost depends on usag
e: $0.02 per GB
    ├─ Traffic within Asia                                               Monthly cost depends on usag
e: $0.08 per GB
    ├─ Traffic within South America                                      Monthly cost depends on usag
e: $0.14 per GB
    ├─ Traffic to/from Indonesia and Oceania                             Monthly cost depends on usag
e: $0.10 per GB
    └─ Traffic between continents (excludes Oceania)                     Monthly cost depends on usag
e: $0.08 per GB


 Project total
         $87.23

 OVERALL TOTAL
        $156.68

*Usage costs were estimated by merging usage defaults from Infracost Cloud and values from infracost-usage.yml.

──────────────────────────────────
85 cloud resources were detected:
• 12 were estimated
• 66 were free
• 7 are not supported yet, rerun with --show-skipped to see details

┌────────────────────────────────────────────────────┬───────────────┬─────────────┬────────────┐
│ Project                                            │ Baseline cost │ Usage cost* │ Total cost │
├────────────────────────────────────────────────────┼───────────────┼─────────────┼────────────┤
│ a-s-gorski/tbd-workshop-1                          │ $0.00         │ $56         │ $56        │
│ a-s-gorski/tbd-workshop-1/bootstrap                │ $0.00         │ $14         │ $14        │
│ a-s-gorski/tbd-workshop-1/cicd_bootstrap           │ $0.00         │ $0.00       │ $0.00      │
│ a-s-gorski/tbd-workshop-1/mlops                    │ $27           │ $60         │ $87        │
```

Figure 5: infracost_local.png



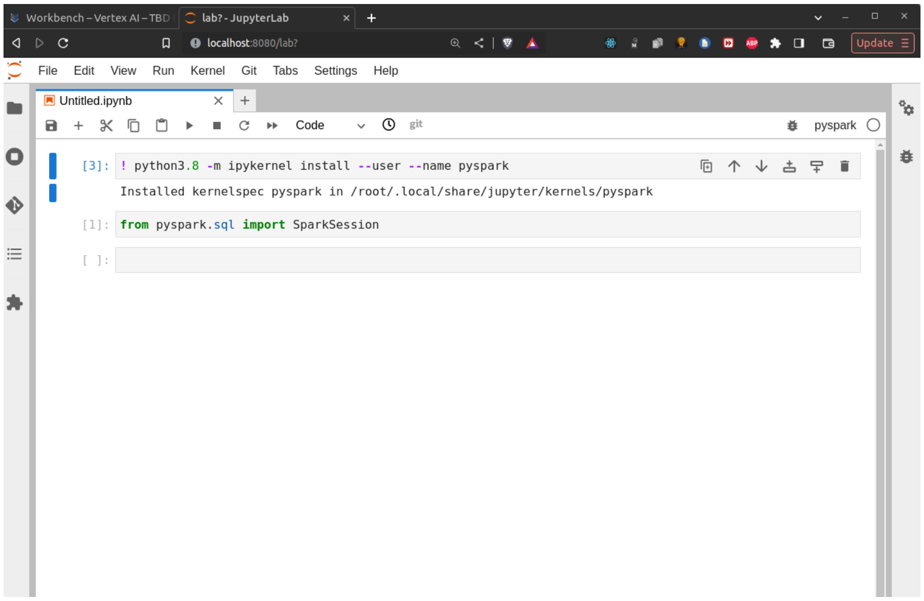
```
CREATE SCHEMA IF NOT EXISTS demo OPTIONS(location = 'europe-west1');

CREATE OR REPLACE EXTERNAL TABLE demo.shakespeare
  OPTIONS (

  format = 'ORC',
  uris = ['gs://tbd-2024l-3040540-data/data/shakespeare/*.orc']);


SELECT * FROM demo.shakespeare ORDER BY sum_word_count DESC LIMIT 5;
```

Query results

Metadata caching is disabled. You can accelerate queries over external tables by enabling metadata caching. Learn more.

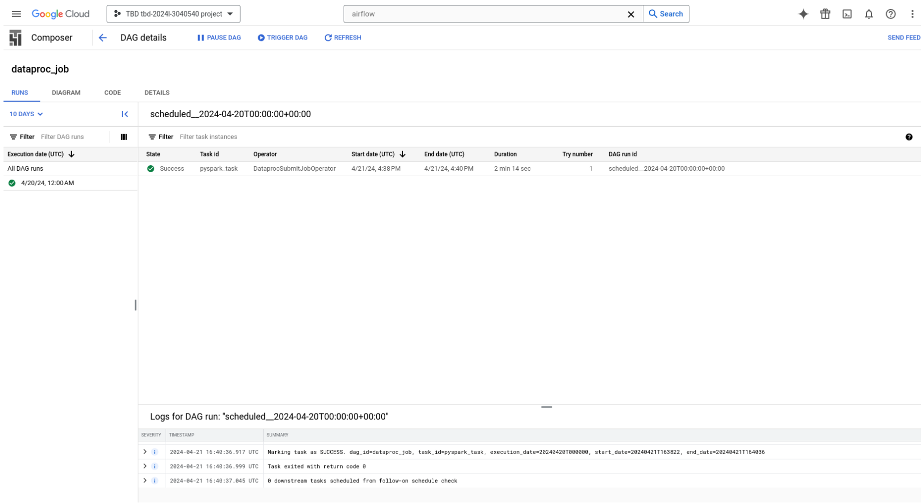
Row	word	sum_word_count
1	the	25568
2	I	21028
3	and	19649
4	to	17361
5	of	16438

Figure 6: bigquery



```
[3]: ! python3.8 -m ipykernel install --user --name pyspark
     Installed kernelspec pyspark in /root/.local/share/jupyter/kernels/pyspark

[1]: from pyspark.sql import SparkSession

[ ]:
```

Figure 7: vertex_ai.png



dataproc_job

Execution date (UTC) ↓
All DAG runs
✓ 4/20/24, 12:00 AM

scheduled__2024-04-20T00:00:00+00:00

State	Task id	Operator	Start date (UTC) ↓	End date (UTC)	Duration	Try number	DAG run id
✓ Success	pyspark_task	DataprocSubmitJobOperator	4/21/24, 4:38 PM	4/21/24, 4:40 PM	2 min 14 sec	1	scheduled__2024-04-20T00:00:00+00:00

Logs for DAG run: "scheduled__2024-04-20T00:00:00+00:00"

```
SEVERITY  TIMESTAMP                    SUMMARY
> i       2024-04-21 16:40:36.917 UTC  Marking task as SUCCESS. dag_id=dataproc_job, task_id=pyspark_task, execution_date=20240420T000000, start_date=20240421T163822, end_date=20240421T164036
> i       2024-04-21 16:40:36.999 UTC  Task exited with return code 0
> i       2024-04-21 16:40:37.045 UTC  0 downstream tasks scheduled from follow-on schedule check
```

Figure 8: airflow

```
preemptible_worker_config {
  num_instances = 2
  preemptibility = "SPOT"
}
```

3. Perform additional hardening of Jupyterlab environment, i.e. disable sudo access and enable secure boot Link Text

   ```
   metadata = {
       vmDnsSetting = "GlobalDefault"
       network-disable-root = true
   }
   post_startup_script = "gs://${google_storage_bucket_object.post-startup.bucket}/${g

   shielded_instance_config {
       enable_secure_boot = true
   }
   ```

4. (Optional) Get access to Apache Spark WebUI