

Dokumentacja

Etap 1a

README -- GKE i Dataproc

Udało się poprawnie odpalić joby na GKE i Dataproc opisane w README

```
Apply complete! Resources: 19 added, 0 changed, 0 destroyed.

Outputs:

data_generator_filepath = "data/input.csv"
data_generator_lines_num = 56
bluealien99@bluebuntu-vm:~/elka9/project-team-tbd-3-2022Z$ gcloud dataproc workflow-templates instantiate ${TF_VAR_tbd_semester}-${TF_VAR_group_id}-workflow --region europe-central2
Waiting on operation [projects/tbd-2022z-3/regions/europe-central2/operations/c93be835-b6fb-34b4-88b8-8b7cede4b4f7].
WorkflowTemplate [tbd-2022z-3-workflow] RUNNING
Creating cluster: Operation ID [projects/tbd-2022z-3/regions/europe-central2/operations/de823017-b37a-4d54-8c3f-00a907bde7b5].
Created cluster: tbd-2022z-3-cluster-aillmq2f6ag4m.
Job ID tbd-2022z-3-job-aillmq2f6ag4m RUNNING
Job ID tbd-2022z-3-job-aillmq2f6ag4m COMPLETED
Deleting cluster: Operation ID [projects/tbd-2022z-3/regions/europe-central2/operations/76fc891e-2197-4656-8bc2-02c046026bab].
WorkflowTemplate [tbd-2022z-3-workflow] DONE
Deleted cluster: tbd-2022z-3-cluster-aillmq2f6ag4m.

To take a quick anonymous survey, run:
$ gcloud survey

bluealien99@bluebuntu-vm:~/elka9/project-team-tbd-3-2022Z$ gsutil cat "gs://tbd-2022z-3-staging/data/output-dataproc.csv/*"
count(1)
56
bluealien99@bluebuntu-vm:~/elka9/project-team-tbd-3-2022Z$ gsutil cat "gs://tbd-2022z-3-staging/data/output-k8s.csv/*"
count(1)
56
bluealien99@bluebuntu-vm:~/elka9/project-team-tbd-3-2022Z$
```







tbd-2022z-3-staging

Location	Storage class	Public access	Protection
europe-central2 (Warsaw)	Standard	Subject to object ACLs	None

[OBJECTS](#) [CONFIGURATION](#) [PERMISSIONS](#) [PROTECTION](#) [LIFECYCLE](#) [OBSERVABILITY](#) [NEW](#)

Buckets > tbd-2022z-3-staging > data

[UPLOAD FILES](#) [UPLOAD FOLDER](#) [CREATE FOLDER](#) [TRANSFER DATA](#) [MANAGE HOLDS](#) [DOWNLOAD](#) [DELETE](#)

Filter by name prefix only		Filter objects and folders		Show deleted data			
<input type="checkbox"/>	Name	Size	Type	Created	Storage class	Last modified	
<input type="checkbox"/>	 input.csv	1,009 B	text/plain; charset=utf-8	Dec 3, 2022, 8:06:55 PM	Standard	Dec 3, 2022, 8:06:55 PM	 
<input type="checkbox"/>	 output-dataproc.csv/	—	Folder	—	—	—	
<input type="checkbox"/>	 output-k8s.csv/	—	Folder	—	—	—	

Sposoby odwoływania się do modułów Terraform wykorzystane w projekcie

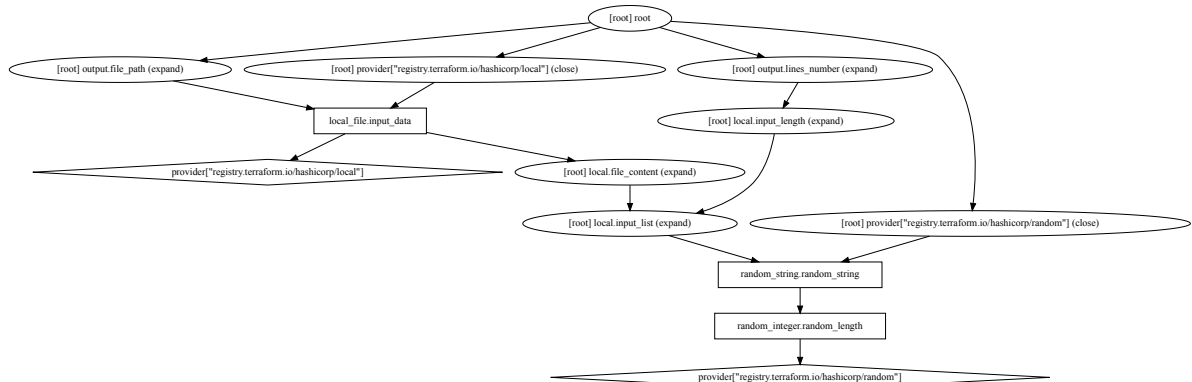
Root Module -- główny moduł (projekt) wykorzystujący child modules.

Child Modules:

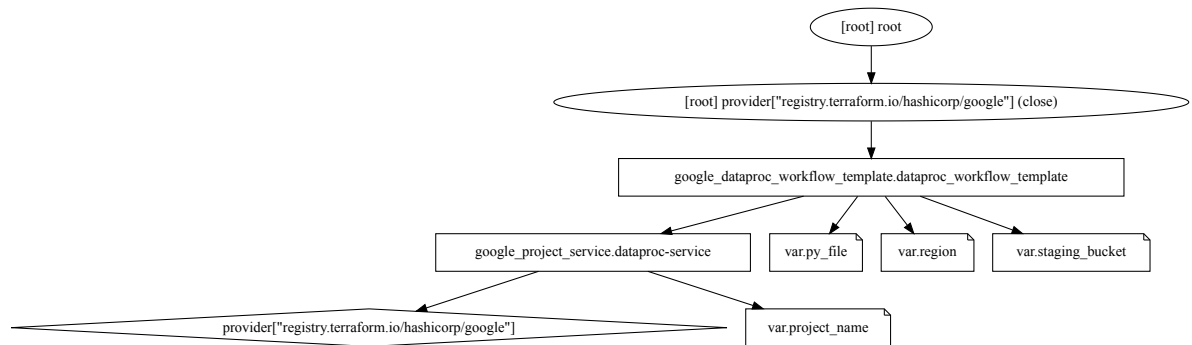
- Local modules -- znajdują się w `./modules`:
 - `data-generator` -- generuje plik zawierający losowe stringi.
 - `dataproc-pyspark-job` -- tworzy i konfiguruje szablon workflowu Dataprocowego zawierającego job PySparkowy.

- `gke` -- tworzy i konfiguruje klaster Kubernetesowy oraz jego node'y. Zwraca endpoint, pod którym się znajduje, oraz jego certyfikat CA.
- Git repository:
 - `k8s-spark-operator` -- pozwala na uruchamianie aplikacji Sparkowych na Kubernetesie w łatwy i idiomatyczny sposób.

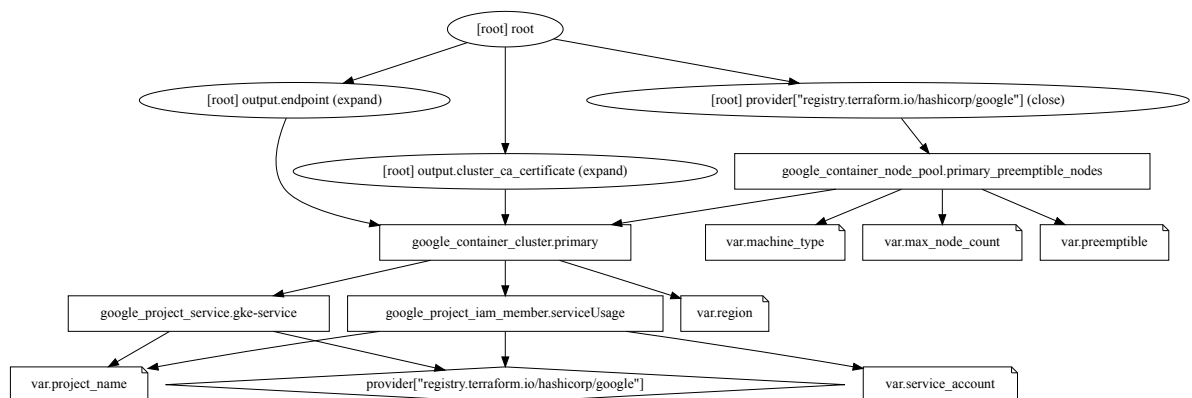
Graf dla modułu `data-generator`



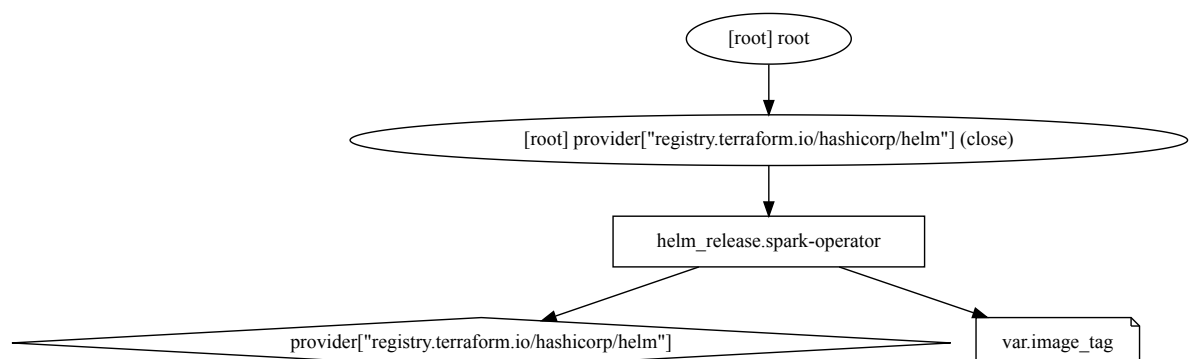
Graf dla modułu `dataproc-pyspark-job`



Graf dla modułu `gke`



Graf dla modułu `k8s-spark-operator`



Moduł monitorujący budżet

Moduł znajduje się w folderze `./modules/budget-monitoring`.

Wszelkie czynności związane z `google_billing_budget` muszą być dokonywane przy użyciu *service account*. W przypadku *end user credentials* dostaniemy `403` z informacją `"reason": "SERVICE_DISABLED"` dla `"service": "billingbudgets.googleapis.com"`, który w rzeczywistości jest włączony.

`google_billing_budget` można skonfigurować podając wiele `threshold_rules` dla jednego resource'a, jednak nie jest możliwe użycie w takim wypadku `for_each`, a co za tym idzie, nie można poziomów dla alertów trzymać w liście.

Postanowiliśmy wykorzystać `for_each`, a także umożliwić konfigurowanie poziomów dla alertów poprzez zmienną (z domyślnymi wartościami). Skutek jest jednak taki, że w konsoli widnieją 3 różne budżety zamiast jednego z kilkoma poziomami, jednak nie powinien być to raczej problem, ponieważ cała konfiguracja jest zarządzana i modyfikowana przez Terraforma.

Lista budżetów

Google Cloud

Search Products, resources, docs (/)

1

?

B

Billing

Billing account
Billing Account for Education

Overview

Reports

Cost table

Cost breakdown

Budgets & alerts

Billing export

Cost optimization

Committed use discounts (C...

Budgets & alerts

CREATE BUDGET

DELETE

LEARN

Budgets track expenses within a Google Cloud Platform project or billing account. Your budget can be a specified amount or based on previous spend. You can set alerts to notify billing admins and users when a budget goes over a specified amount.

Setting a budget does not cap resource or API consumption. [Learn more.](#)

Filter

Enter property name or value

Budget name	Budget period	Budget type	Applies to	Trigger alerts at	Spend and budget amount
<input type="checkbox"/> TBD Billing Budget	Monthly	Specified amount	This billing accoi	100%	\$1.57 / \$50.00 Excludes -\$1.57 credit
<input type="checkbox"/> TBD Billing Budget	Monthly	Specified amount	This billing accoi	80%	\$1.57 / \$50.00 Excludes -\$1.57 credit
<input type="checkbox"/> TBD Billing Budget	Monthly	Specified amount	This billing accoi	50%	\$1.57 / \$50.00 Excludes -\$1.57 credit

Szczegóły budżetu

Google Cloud

Search Products, resources, docs (/)

Billing

Billing account

Billing Account for Education

Overview

Reports

Cost table

Cost breakdown

Budgets & alerts

Billing export

Cost optimization

Committed use discounts (C...

CUD analysis

Pricing

Billing management

Account management

Release Notes

Edit Budget

LEARN

TBD Billing Budget

Name *

TBD Billing Budget

Scope

A budget enables you to track your actual spend against your planned spend.

Time range

Monthly

The month starts on the first of the month and reset at the beginning of each month.

A budget can be scoped to focus on a specific set of resources.

Projects

All projects (1)

Services

All services (1958)

Labels

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

Amount

Set a monthly budget amount

Budget type

Specified amount

A fixed amount that your spend will be compared against.

Target amount *

\$ 50

Actions

Set alert threshold rules

Send email alert notifications after the actual or forecasted spend exceeds a percent of the budget or a specified amount. [Learn more.](#)

Percent of budge...

100 %

Amount 1 *

\$ 50

Trigger on 1

Actual

+ ADD THRESHOLD

SAVE

CANCEL

Cost trend

November 1, 2021 – November 30, 2022

\$60

\$50

\$40

\$30

\$20

\$10

\$0

Dec

Jan

Mar

May

Jun

Jul

Aug

Oct

Nov

Actual cost

View report

Wykorzystanie budżetu na koniec etapu 1a

November 1 – 10, 2022 (total cost)

\$2.96

includes \$0.00 in credits

↑ -

\$2.96 over October 22 – 31, 2022

November 2022 (forecasted total cost)

Not enough historical data to project cost

Daily

Service	Cost	Discounts	Promotions and others	Subtotal
Compute Engine	\$1.89	—	—	\$1.89
Kubernetes Engine	\$0.96	—	—	\$0.96
Networking	\$0.11	—	—	\$0.11
Cloud Storage	\$0.01	—	—	\$0.01
Cloud Logging	\$0.00	—	—	\$0.00
Subtotal				\$2.96
Tax				—
Filtered total				\$2.96

Klaster dataproc

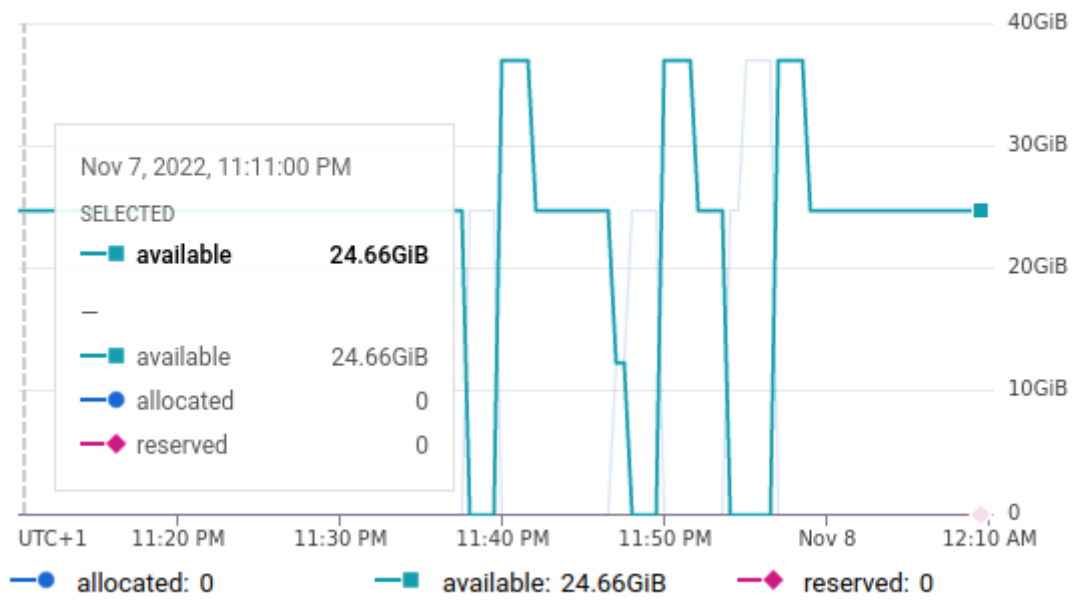
Projekt znajduje się w folderze `./dataproc`.

W poleceniu jest mowa na zmianę o jobie sparkowym i pysparkowym, jednak przykładowy job pysparkowy jest typowym *hello worldem*, który praktycznie nic nie robi, a zatem w celu przetestowania polityki autoscalowania wybrany został przykładowy job sparkowy, który oblicza wartość liczby pi. Na zrzucie ekranu widoczne są 3 scale upy, ponieważ job sparkowy został uruchomiony 3 razy z kolejno argumentami: 100000, 200000 i 400000, które oznaczają liczbę równoległych tasków zwanych *slice / partition*.

Można zauważyć, że gdy zaczynało brakować dostępnej pamięci, dodawany był kolejny (trzeci) node, a po zakończeniu obliczeń był on usuwany.

Działanie polityki autoscalowania

YARN memory



YARN NodeManagers

