Konfiguracja środowiska GCP

W założeniu następujące elementy są gotowe:

- Konto Google
- Założony projekt przeznaczony na nasz kurs (`https://console.cloud.google.com/`)
- Utworzone konto rozliczeniowe

Wprowadzenie

Przetwarzane dane i ich pochodzenie

Dane które będziemy wykorzystywali w ramach zajęć pochodzą z różnych miejsc. Z reguły są one wskazane np. w zestawach zadań, które z nich korzystają. Warto jednak mieć świadomość, że miejsca te mogą już nie istnieć, dane, które tam się znajdują mogły ulec zmianie, a także może mieć miejsce sytuacja, w której wykorzystywane przez nas dane były celowo odpowiednio zmodyfikowane przez prowadzącego.

W związku z powyższym każdorazowo będziemy korzystali z kopii tych danych udostępnianych przez prowadzącego. W wielu miejscach zestawów zadań pojawia się tajemniczy `<host>`. Naszym hostem jest `jankiewicz.pl`.
Dla przykładu:

```
wget http://<host>/bigdata/cycle-share-dataset.zip
```

oznacza:

```
wget http://jankiewicz.pl/bigdata/cycle-share-dataset.zip
```

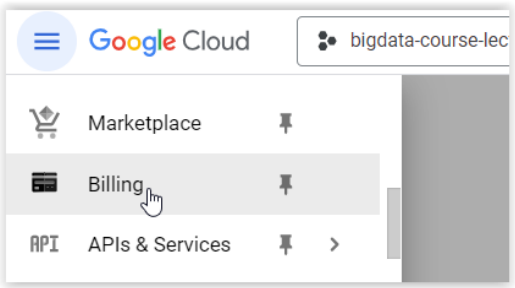
Środowisko GCP

Środowisko GCP oprócz dziesiątków różnego typu narzędzi odpowiadających za przechowywanie danych o różnorodnej charakterystyce (*Cloud Storage*, *Bigtable*, *SQL*, *Datastore*, …), przetwarzanie danych i akwizycję danych (*Cloud Functions*, *Compute Engine*, *Pub/Sub*, …) posiada także klaster *Hadoop*, o nazwie *Dataproc*, z całym szeregiem preinstalowanych (*Pig, Hive, Spark*) i opcjonalnych narzędzi (*Zookeeper, HBase, Apache Kafka*), które wykorzystamy w ramach naszego kursu.

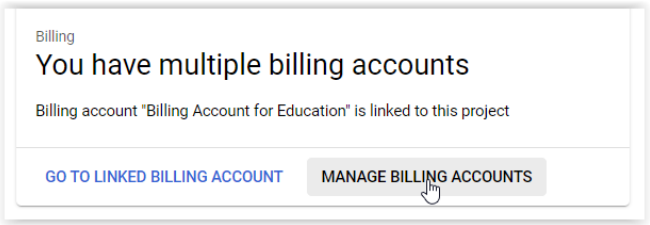
1. Przejdź do konsoli platformy GCP https://console.cloud.google.com

Sprawdzenie/podłączenie konta rozliczeniowego do projektu

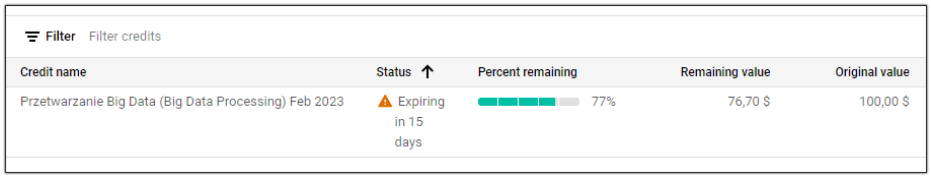
1. Z menu nawigacyjnego wybierz pozycję *Billing.* Aby mieć łatwiejszy dostęp do tej pozycji można ją „przyszpilić”.



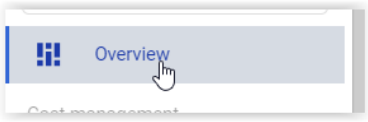
2. Przejdź do zarządzania kontem rozliczeniowym podpiętym do Twojego konta. Jeśli masz wiele kont rozliczeniowych wybierz opcję *Manage Billing Accounts*, a następnie wybierz konto rozliczeniowe którym chcesz zarządzać

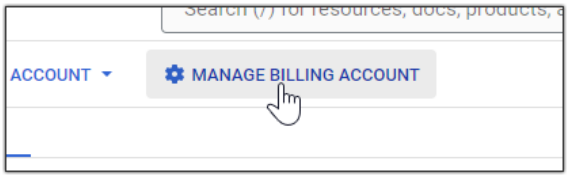


3. Zatrzymaj się na chwilę w tym miejscu. Możesz tu sprawdzić ile środków Ci jeszcze zostało lub wygenerować raport zużycia. Zaglądaj tu od czasu do czasu aby kontrolować Twoje środki

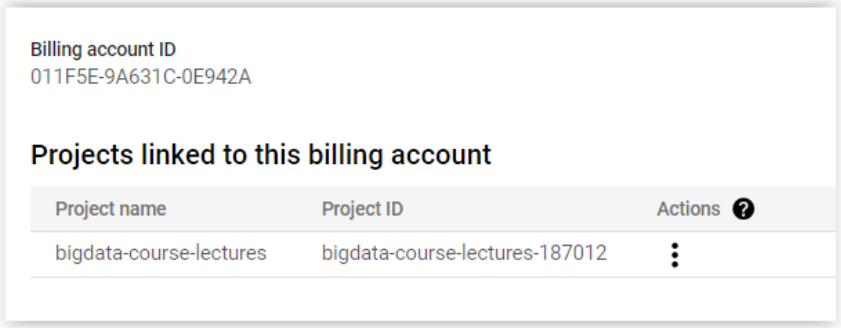


4. Wybierz Overview, a następnie, przejdź do zarządzania tym kontem rozliczeniowym

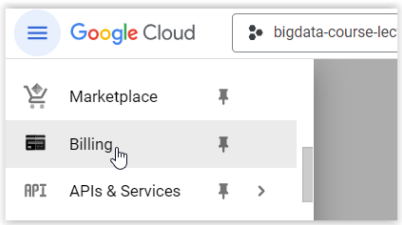


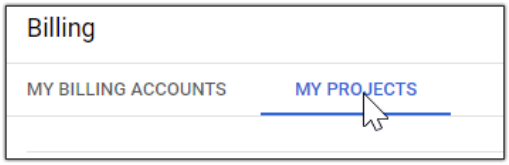


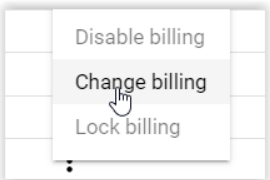
5. Upewnij się, że jedynie Twój projekt przeznaczony na potrzeby naszego kursu jest podpięty pod konto rozliczeniowe



6. Jeśli żadnego projektu nie ma podczepionego do Twojego konta rozliczeniowego, to wybierz ponownie, *Billing*, *Manage Billing Accounts*, a następnie wybierz zakładkę *My Projects*. Przy swoim projekcie wybierz w menu *Actions* opcję *Change Billing* i przypisz utworzone konto rozliczeniowe.



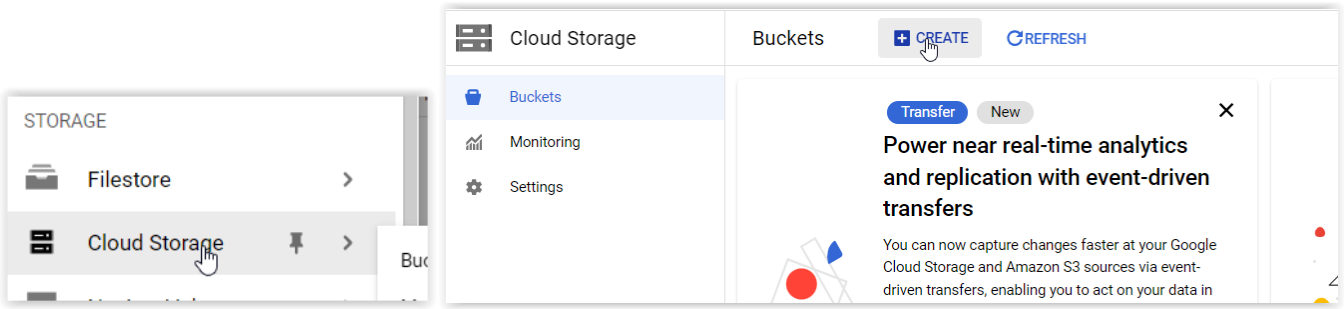




Utworzenie zasobnika (*bucket*)

Jedną z ważniejszych usług, które będziemy wykorzystywali jest *Cloud Storage*. Jest to nic innego jak stary dobry Google File System, pod chwytliwą marketingową nazwą. Google File System czyli rozproszony system plików Google. Tworząc w ramach tej usługi zasobnik utworzymy nasz własny "katalog" w tej usłudze, w którym będziemy mogli trwale i bezpiecznie przechowywać nasze dane. Zarówno te które będziemy przetwarzali, jak i te, które będą np. wynikami naszego przetwarzania, lub efektami naszej pracy (notatniki platform Zeppelin czy Jupyter).

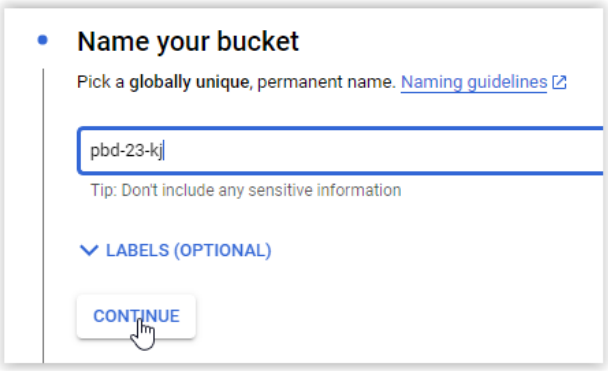
7. Za pomocą menu nawigacyjnego *Storage/Cloud Storage* otwórz stronę z utworzonymi zasobnikami



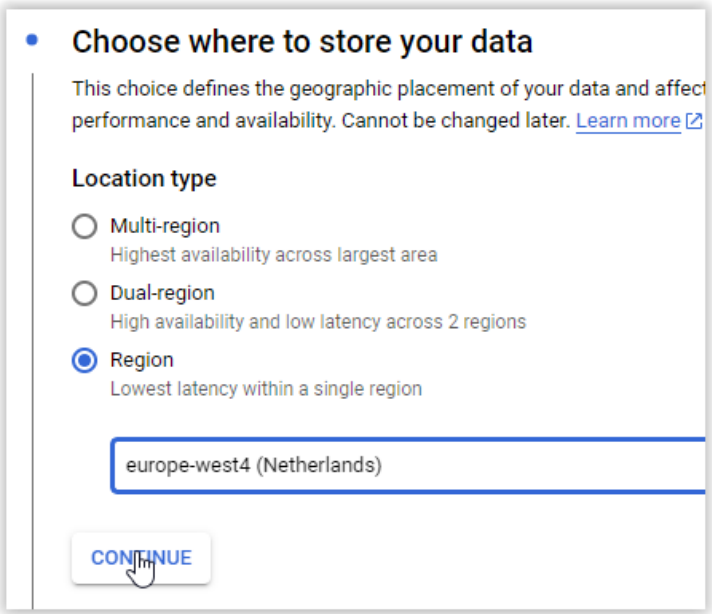
8. Aby mieć łatwiejszy dostęp do tej pozycji można ją „przyszpilić".

9. Korzystając z przycisku *Create* utwórz nowy zasobnik (*bucket*)

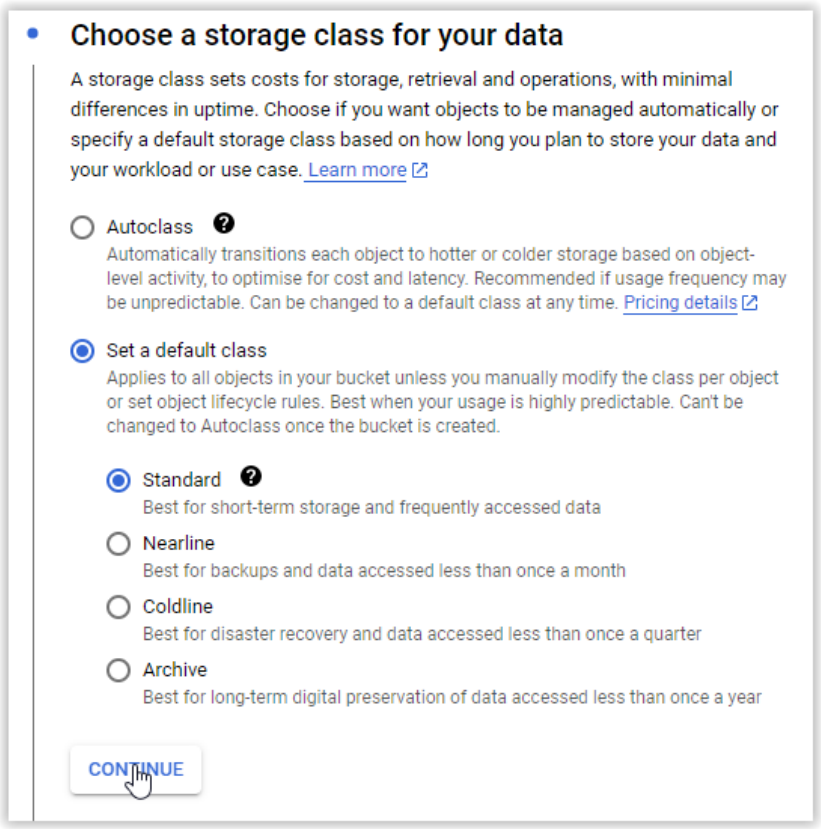
   a. Określ globalnie unikalną nazwę (używaj tylko małych liter, cyfr oraz myślników), wybierz *Continue*



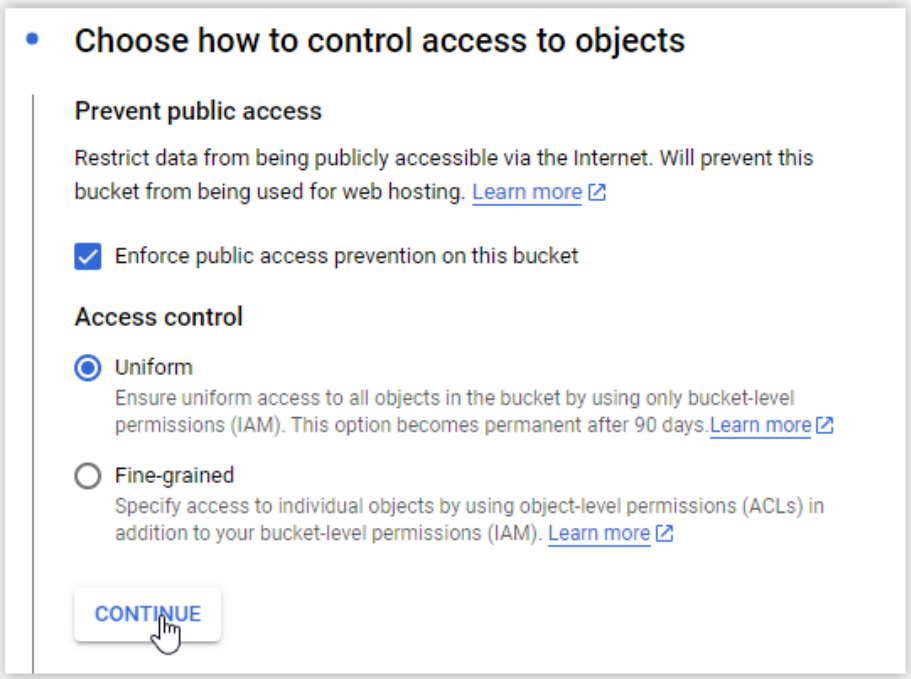
   b. Określ region, w którym zostanie utworzony Twój zasobnik i w którym będą w przyszłości tworzone będą Twoje klastry (dzięki temu, że będą blisko danych, będzie można uzyskać lepszą wydajność), wybierz *Continue*



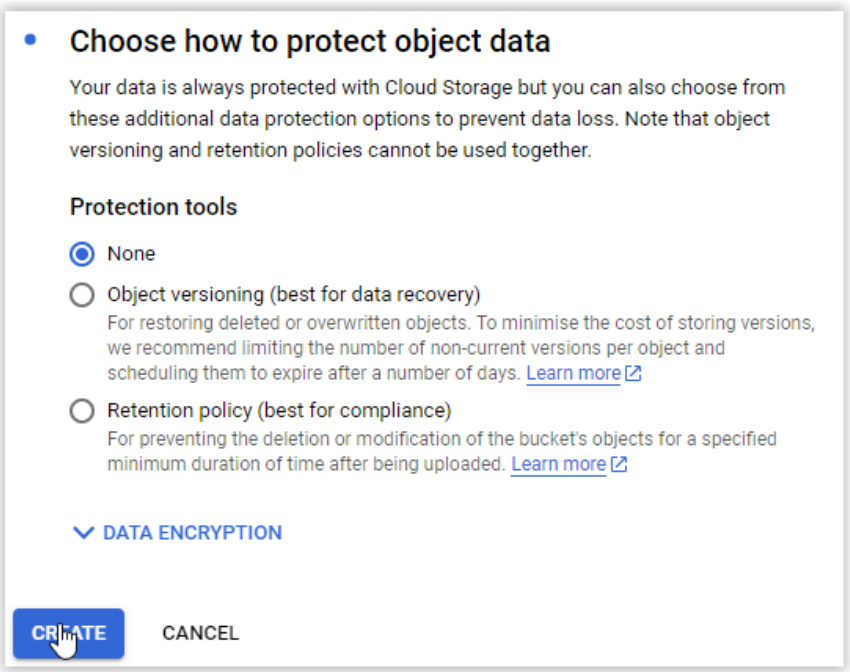
c. Wybierz klasę składowania danych (pozostaw Standard, gdyż nasze dane będą intensywnie wykorzystywane)
Wybierz *Continue*

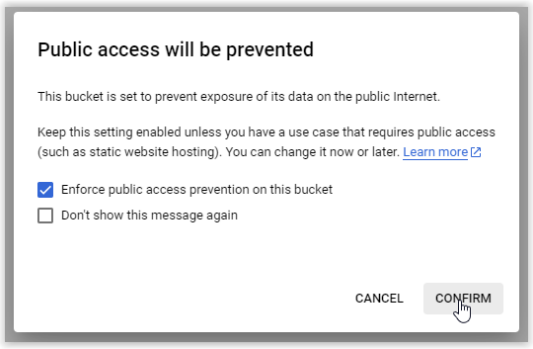


d. Pozostaw poziom sterowania uprawnieniami na *Uniform*
Wybierz *Continue*



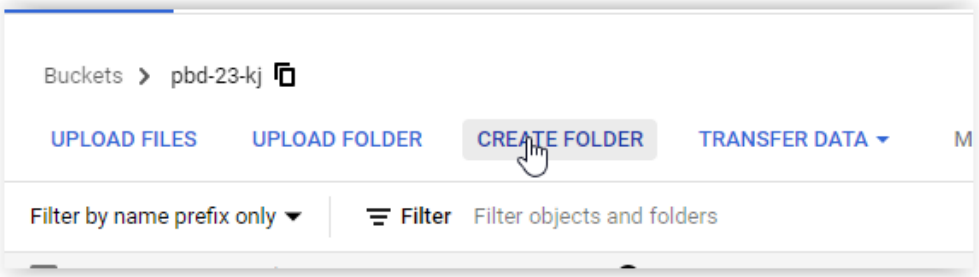
e. Nie włączaj żadnych dodatkowych funkcji chroniących dane przed utratą
Wybierz *Create*, a następnie potwierdź chęć utworzenia prywatnego zasobnika

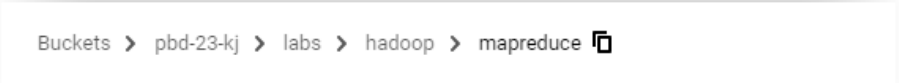




Utworzenie katalogów i plików

10. Korzystając z przycisków *Create folder*, utwórz zestaw katalogów `/labs/hadoop/mapreduce`

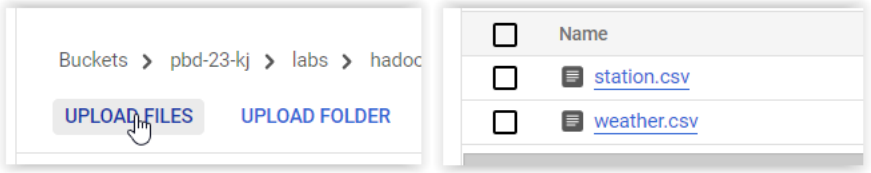




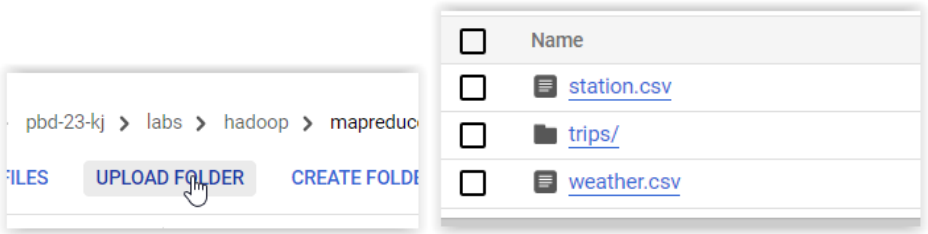
11. Pobierz do swojego lokalnego systemu plików poniższy plik, a następnie go rozpakuj
    `https://jankiewicz.pl/bigdata/cycle-share-dataset.zip`

```
PS C:\tmp\20221114\cycle-share-dataset> ls -r


    Directory: C:\tmp\20221114\cycle-share-dataset


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
d-----        27.09.2023     18:16                trips
-a----        07.11.2016     03:36           5316 station.csv
-a----        07.11.2016     03:36          56516 weather.csv


    Directory: C:\tmp\20221114\cycle-share-dataset\trips


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----        01.07.2022     13:28       16585055 trip1.csv
-a----        01.07.2022     13:29       16612028 trip2.csv
-a----        01.07.2022     13:29        6076971 trip3.csv
```

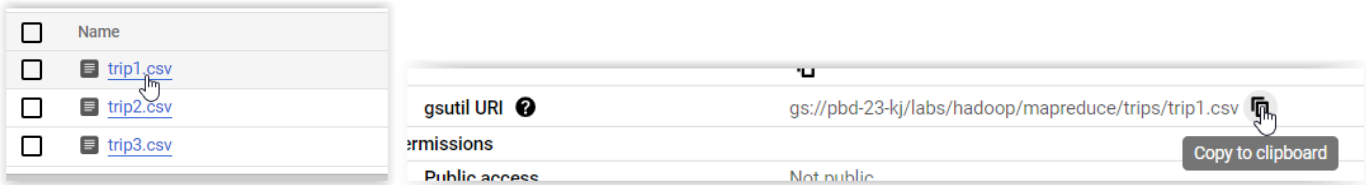
12. Załaduj pliki `station.csv` oraz `weather.csv` do Twojego zasobnika do katalogu `/labs/hadoop/mapreduce` korzystając z przycisku *Upload Files.* Możesz załadować oba na raz o ile je zaznaczysz razem



13. Załaduj cały katalog `trips` wykorzystując *Upload Folder*



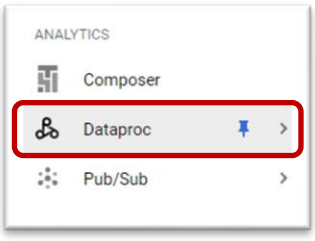
14. Wybierz link dotyczący pliku `trip1.csv` znajdujący się w katalogu `trips`.
    Zwróć uwagę na własność *gsutil URI* – za pomocą tego URI będziemy mogli odwoływać się do tego pliku np. z poziomu klastra Hadoop (*Dataproc*)

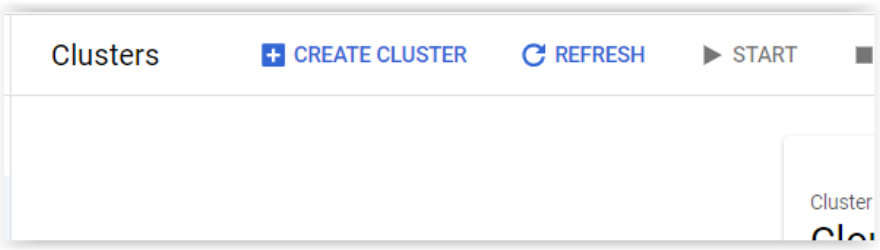


Uruchomienie klastra

Przygotowania

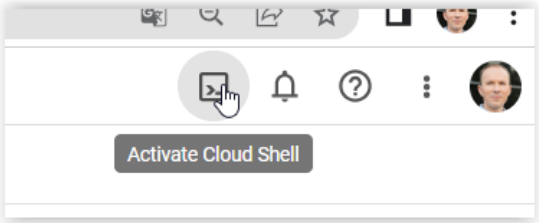
15. Za pomocą menu nawigacyjnego *Analytics/Dataproc* otwórz stronę z uruchomionymi klastrami *Dataproc*





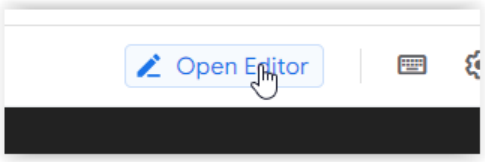
16. Aby mieć łatwiejszy dostęp do tej pozycji można ją „przyszpilić”.

17. Korzystając z paska nawigacji (lewy górny róg konsoli), uruchom terminal *Cloud Shell*.

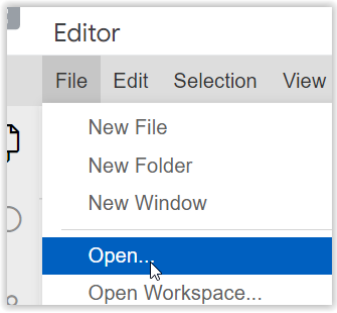


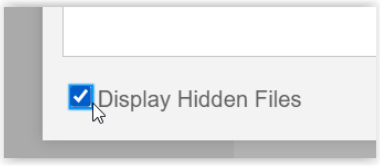
Cloud Shell to mała darmowa wirtualna maszyna z zainstalowanym API, za pomocą którego można wykonywać cały szereg różnych czynności administracyjnych w środowisku GCP. Przykładowo, można za pomocą zainstalowanych tam narzędzi uruchamiać klastry *Dataproc*.

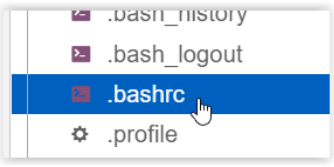
18. Otwórz edytor, który dostępny jest na poziomie tego narzędzia

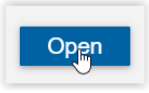


19. Wybierz z menu pozycje *File->Open*. Następnie włącz widoczność plików ukrytych i otwórz plik `.bashrc`





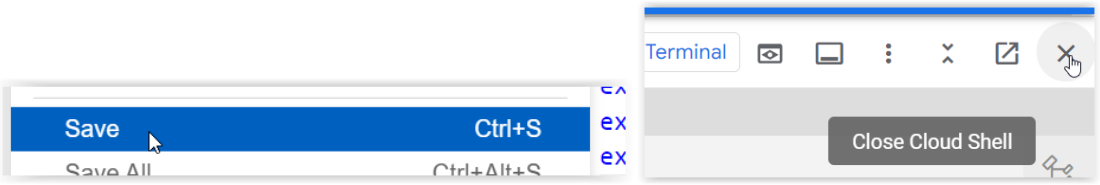




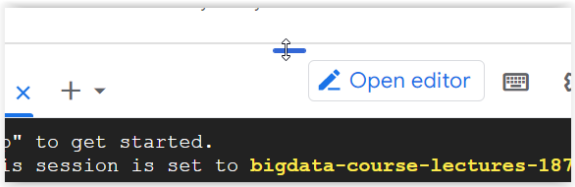
20. Na końcu pliku dodaj następujące polecenia, które zainicjują wartości kilku zmiennych za każdym razem kiedy będziemy otwierali sesję w za pomocą terminala *Cloud Shell*.
    Koniecznie popraw nazwę regionu oraz nazwę zasobnika

```
export REGION=TU_WPISZ_NAZWĘ_REGIONU_W_KTORYM_ZOSTAL_UMIESZCZONY_ZASOBNIK
export CLUSTER_NAME=pbd-cluster
export PROJECT_ID=$(gcloud config get-value project)
export BUCKET_NAME=TU_WPISZ_NAZWĘ_SWOJEGO_ZASOBNIKA
```

21. Zapisz zmiany dokonane w pliku, a następnie zamknij *Cloud Shell.*



22. Otwórz ponownie *Cloud Shell*. Jeśli wypełnił on całe okno możesz go zmniejszyć do jego dolnej części.

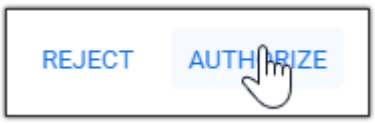


Uruchomienie klastra

23. Korzystając z terminala *Cloud Shell* uruchom polecenie, które utworzy klaster *Dataproc.*

```
gcloud dataproc clusters create ${CLUSTER_NAME} \
--enable-component-gateway --region ${REGION} --subnet default \
--master-machine-type n2-standard-4 --master-boot-disk-size 50 \
--num-workers 2 --worker-machine-type n2-standard-2 --worker-boot-disk-size 50 \
--image-version 2.2-debian12 --optional-components ZEPPELIN \
--project ${PROJECT_ID} --bucket ${BUCKET_NAME} --max-age=2h
```

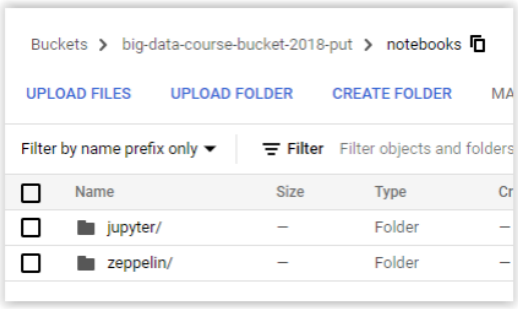
24. Jeśli pojawi się pytanie dotyczące tego czy włączyć API obsługujące klastry *Dataproc*, potwierdź swoją decyzję. Polecenie zakończy swoje działanie w ciągu kilku minut. Ty jednak możesz kontynuować.



Kilka uwag

Warto zwrócić uwagę na polecenie wykorzystane do utworzenia klastra.

Dzięki przypisaniu zasobnika do klastra (parametr `--bucket ${BUCKET_NAME}`), pewne elementy pochodzące z działań na klastra będą automatycznie zapisywane trwale w zasobniku. Dla przykładu będą tam automatycznie zapisywane notatniki platform Zeppelin czy Jupyter



Określając region (parametr `--region ${REGION}`) sprawiliśmy, że nasz klaster został utworzony w ogólnie określonej lokalizacji. Patrz na: `https://cloud.google.com/compute/docs/regions-zones`

Nasz klaster będzie się składał z trzech maszyn, jednej master i dwóch roboczych (parametr `--num-workers 2`).

Maszyna master będzie miała 16GB RAM i 4 wirtualne procesory (`n2-standard-4`) oraz dysk o rozmiarze 50GB (parametr `--master-boot-disk-size 50`). Każda maszyna robocza będzie miała 8GB RAM i 2 wirtualne procesory (`n2-standard-2`) oraz dysk o rozmiarze 50GB. Patrz na:
`https://cloud.google.com/compute/docs/general-purpose-machines#n2_series`

Warto zwrócić także uwagę na obraz maszyny wykorzystywany do utworzenia węzłów klastra (parametr `--image-version 2.2-debian12`).
Patrz na: `https://cloud.google.com/dataproc/docs/concepts/versioning/dataproc-versions`

Każdy z tych obrazów różni się wersjami narzędzi jakie są w nim dostępne.

Version	Last Updated	Released On
2.2-debian12	2024/02/01	2023/12/08
2.1-debian11	2024/02/01	2022/12/12
2.0-debian10	2024/02/01	2021/01/22

Apache Flink optional component	1.17.0
Apache Hadoop installed	3.3.6
Apache Hive installed	3.1.3

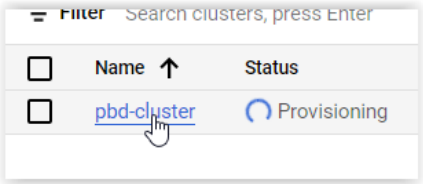
Apache Kafka initialization action	3.1.0
Apache Pig installed	0.18.0-SNAPSHOT
Apache Spark installed	3.5.0

Wiele z tych narzędzi jest już zainstalowanych w obrazie maszyny i podczas tworzenia klastra wymagają one jedynie konfiguracji. Część z nich może zostać dodatkowo zainstalowane jako dodatkowe komponenty np.: `--optional-components JUPYTER`, lub za pomocą akcji inicjalizacyjnych.

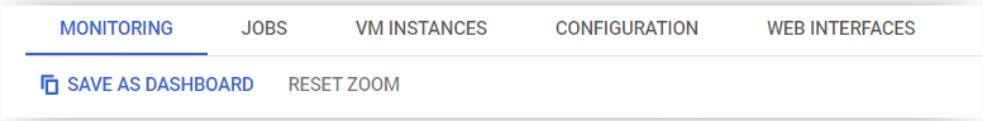
Podczas tworzenia klastra zażyczyliśmy sobie ponadto, aby najważniejsze interfejsy sieciowe komponentów zainstalowanych na klastrze zostały udostępnione (parametr `--enable-component-gateway`). Podczas zapoznawania się ze środowiskiem zaglądnij na zakładkę *Web Interfaces* aby się o tym przekonać.

Zapoznanie się ze środowiskiem

25. Korzystając z konsoli GCP wybierz nazwę klastra, przechodząc w ten sposób do jego szczegółowych informacji.



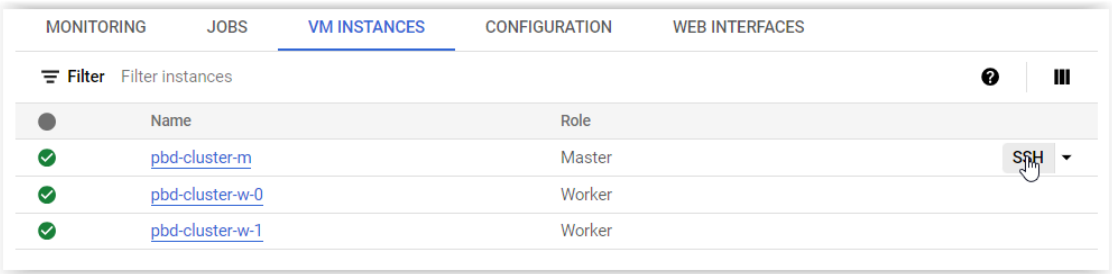
26. Na otwartej przed chwilą stronie znajdziesz pięć zakładek
    a. *Monitoring* – wykresy pozwalające na monitowanie zasobów klastra
    b. *Jobs* – zadania, które zostały uruchomione na klastrze
    c. *VM Instances* – instancje wirtualnych maszyn wchodzących w skład klastra
    d. *Configuration* – parametry konfiguracyjne klastra
    e. *Web Interfaces* – interfejsu sieciowe komponentów wchodzących w skład klastra



Maszyny wirtualne

27. Otwórz zakładkę dotyczącą maszyn wirtualnych. Znajdziesz tam trzy maszyny.
    a. Przyglądnij się poleceniu, które zostało wykorzystane do utworzenia klastra. Znajdź w nim parametry, które odpowiadają za taką, a nie inną liczbę maszyn.
    b. Czy wszystkie maszyny są takiego samego typu?

28. Pierwsza z maszyn jest tzw. maszyną master. Uruchomione są na niej nieco inne komponenty (lub w innej roli) niż na tzw. węzłach roboczych. Wybierz przycisk SSH aby uruchomić terminal SSH maszyny master.



29. System operacyjny maszyn wchodzących w skład klastra *Dataproc* został wybrany podczas jego tworzenia. Zaglądnij na stronę https://cloud.google.com/dataproc/docs/concepts/versioning/dataproc-versions, aby sprawdzić możliwości. Wszystkie dostępne systemy operacyjne oparte są oczywiście na systemie Linux.

30. Sprawdź za pomocą poniższych poleceń:
    c. zawartość Twojego domowego katalogu

```
ls
```

    d. wersję platformy Hadoop

```
hadoop version
```

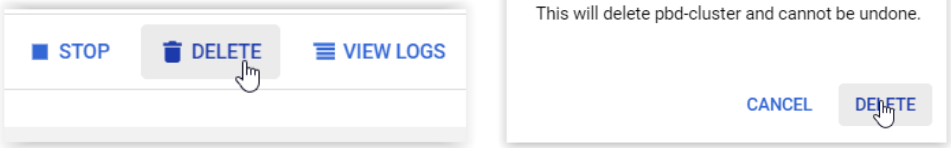
```
jankiewicz_krzysztof@hadoop-intro-m:~$ hadoop version
Hadoop 3.3.6
Source code repository https://bigdataoss-internal.googlesource.com
0d4882c72
Compiled by bigtop on 2024-01-30T20:45Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum a07f708b719648bc7616ed771d9b4173
This command was run using /usr/lib/hadoop/hadoop-common-3.3.6.jar
```

    e. wersję Sparka

```
spark-submit --version
```

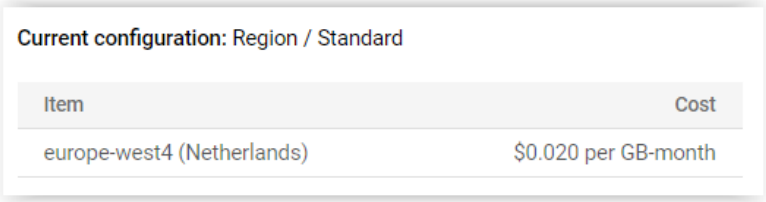
```
jankiewicz_krzysztof@hadoop-intro-m:~$ spark-submit --version
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Scala version 2.12.18, OpenJDK 64-Bit Server VM, 11.0.20.1
Branch dataproc-branch-3.5.0
Compiled by user  on 2024-01-30T21:30:10Z
Revision 8c74211db15b74df03f0cfb2fd46a60ebe95f5c8
Url https://bigdataoss-internal.googlesource.com/third_party/apache/spark
Type --help for more information.
```

31. Jeśli wszystko zadziałało poprawnie, usuń klaster. W tym celu przełącz się na konsolę GCP na stronę dotyczącą szczegółów klastra. A następnie wybierz przycisk *Delete*.



Pamiętaj – to ważne

Zasobnika nie będziemy usuwali aż do końca naszego kursu. Jego koszt jest w praktyce pomijalny.



Jednak uruchomiony, a nawet zatrzymany klaster generuje znaczące koszty. Działający klaster o parametrach, które będziemy z reguły wykorzystywali, **w ciągu kilku (2-4) dni może wykorzystać całe zasoby jakie mamy na naszych kontach rozliczeniowych**.

Dlatego:

- klastry **koniecznie usuwamy** po każdym warsztacie (po zakończonej pracy)
- zawsze sprawdzamy, czy polecenie uruchamiające klaster ma ustawiony parametr `--max-age`
- regularnie kontrolujemy środki jakie nam pozostały, mając świadomość, że realizacja projektów może wymagać ich większej ilości