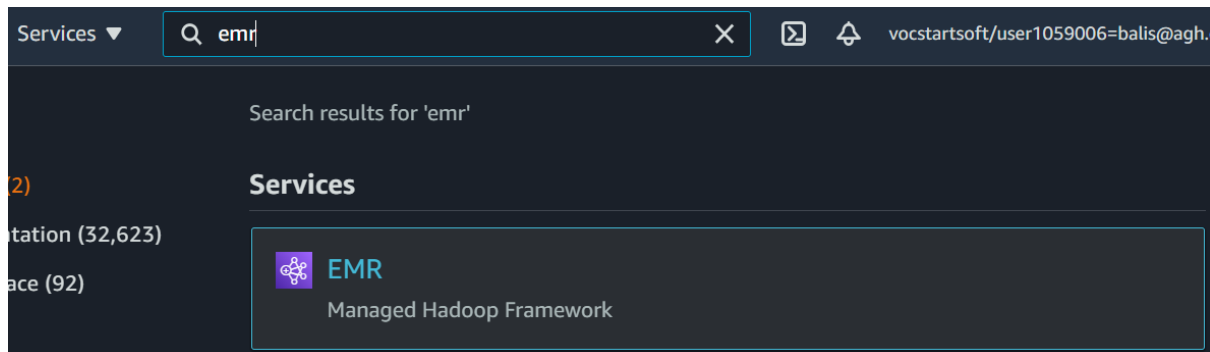


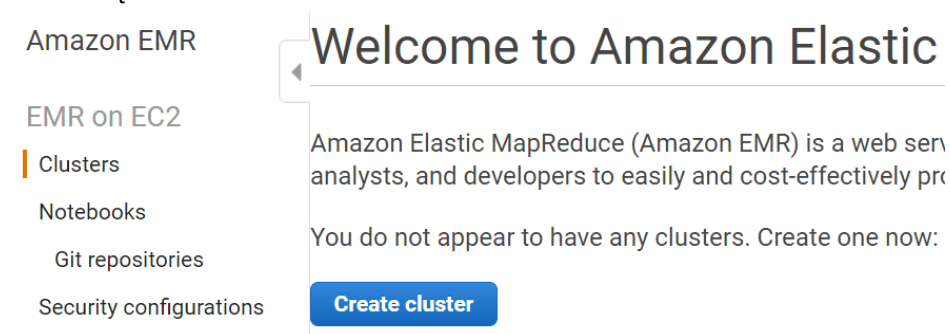
Konfiguracja klastra Elastic Map Reduce w chmurze AWS

1. Wyszukujemy i przechodzimy do usługi **EMR**:



2. Tworzymy klaster

Po kliknięciu w Create cluster:



wpisujemy nazwę dla klastra, reszta konfiguracji może zostać domyślna:

General Configuration

Cluster name

☒ Logging ⓘ

S3 folder ⓘ

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

Software configuration

Release ⓘ

Applications

- ☒ Core Hadoop: Hadoop 2.10.1, Hive 2.3.7, Hue 4.8.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- ☐ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Hue 4.8.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
- ☐ Presto: Presto 0.240.1 with Hadoop 2.10.1 HDFS and Hive 2.3.7 Metastore
- ☐ Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.8.2

☐ Use AWS Glue Data Catalog for table metadata ⓘ

Po przewinięciu strony w dół widzimy konfigurację instancji i klucza SSH. Dla celów testowych można wybrać tańszą instancję, np. m4.large. Koniecznie należy również ustawić “EC2 key pair”. Jeśli nie mamy utworzonej pary kluczy, należy to zrobić przez konsolę EC2. Tworząc parę kluczy można importować swój istniejący klucz publiczny, aby uzyskać dostęp do instancji przez ssh.

Hardware configuration

Instance type

Number of instances (1 master and 2 core nodes)

Cluster scaling ☐ scale cluster nodes based on workload

Security and access

EC2 key pair

Permissions ☒ Default ☐ Custom
Use default IAM roles. If roles are not present, they will be created for you with managed policies for automatic policy updates.

EMR role [EMR_DefaultRole](#) ⓘ

EC2 instance profile [EMR_EC2_DefaultRole](#) ⓘ

3. Finalizujemy klikając **Create cluster**. Przejdziemy do strony podsumowującej, na której będzie widać, że maszyny klastra są w trakcie tworzenia (“Provisioning”):

Network and hardware

Availability zone: --

Subnet ID: [subnet-1dbbe813](#)

Master: Provisioning 1 m4.large

Core: Provisioning 2 m4.large

Task: --

Cluster scaling: Not enabled

Utworzony klaster widać na liście w konsoli usługi EMR:

[Create cluster](#) [View details](#) [Clone](#) [Terminate](#)

Filter: 1 cluster (all loaded)

	Name	ID	Status
<input type="checkbox"/> ▶	MPR_cluster	j-2HDL4U4C6VVY7	Starting

Po wejściu na widok klastra widać nazwę DNS węzła master.

Cluster: MPR_cluster **Starting** Configuring cluster software

Summary

Application user interfaces

Monitoring

Hardware

Con

Summary

ID: j-2HDL4U4C6VVY7


Creation date: 2021-03-23 12:47 (UTC+1)

Elapsed time: 8 minutes

After last step completes: Cluster waits

Termination protection: Off [Change](#)

Tags: -- [View All / Edit](#)

Master public DNS: ec2-3-235-63-94.compute-1.amazonaws.com 

[Connect to the Master Node Using SSH](#)

Po kliknięciu na “Connect to the Master Node Using SSH” można skopiować komendę by połączyć się z węzłem master przez SSH.

Windows

Mac / Linux

1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace ~/bb_key_pair.pem with the location and filename of the private key file (.pem) used to launch the cluster.

```
ssh -i ~/bb_key_pair.pem hadoop@ec2-3-235-63-94.compute-1.amazonaws.com
```

3. Type yes to dismiss the security warning.

Uwaga: może istnieć konieczność dodania reguły dla ruchu SSH w ustawieniach grupy bezpieczeństwa Mastera (Security groups for Master):


Security and access

Key name: bb_key_pair

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for Master: [sg-0279fa4733efb92ab](#)  (ElasticMapReduce-master)

--

Maszyny klastra można również zobaczyć w widoku konsoli EC2:

Instances (1/3) [Info](#)

Instance state ▼

Instance state: running ✕

Clear filters

<input type="checkbox"/>	Name ▼	Instance ID	Instance state ▼	Instance type
<input type="checkbox"/>	–	i-036d70c11db896581	Running	m4.large
<input type="checkbox"/>	–	i-0e9d2741724d10fb4	Running	m4.large
<input type="checkbox"/>	–	i-0844e4963e6fb2aa5	Running	m4.large

Logujemy się na węzeł Master:

[illegible]

Uruchomienie przykładu

1. Tworzymy katalog na dane wejściowe

```
$ mkdir books
```

2. Tworzymy skrypt `download.sh`:

```
$ cat >> download.sh
#!/bin/bash
for i in {1300..1400}
do
    wget "http://www.gutenberg.org/files/$i/$i.txt"
    wget "http://www.gutenberg.org/files/$i/$i-0.txt"
done
<ctrl+d>
$ chmod u+x download.sh
```

Zmieniając zakres 1300..1400 można pobrać więcej danych.

3. Pobieramy dane testowe i kopiujemy je do HDFS:

```
$ cd books
$ ../download.sh
$ hdfs dfs -mkdir books-input
$ hdfs dfs -put *.txt books-input
```

4. Następnie kopiujemy [mapper.py](#) i [reducer.py](#). Alternatywne (bardziej zoptymalizowane) implementacje korzystają z generatorów: [mapper-adv.py](#) i [reducer-adv.py](#).

Przetestować ich działanie można komendą:

```
$ printf 'aaa bbb ccc\naaa' | ./mapper.py | sort | ./reducer.py
aaa      2
bbb      1
ccc      1
```

5. Uruchamiamy przetwarzania Map Reduce (z pomiarem czasu):

```
$ time hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -files
mapper.py,reducer.py -mapper mapper.py -reducer reducer.py -input
books-input -output books-output
```

Inne przydatne komendy:

hdfs dfs -rm -r books-output	Usuwanie katalogu z danymi wyjściowymi
hdfs dfs -get books-output	Pobranie danych wyjściowych