UNIVERSIDAD DE GRANADA

Big Data I

máster ciencia de datos e ingeniería de computadores

Cloud Computing y Big Data

Práctica sobre Contenedores

Autor
Ignacio Vellido Expósito
ignaciove@correo.ugr.es

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Curso 2020-2021

Índice

1. Contenedor con SGDB MySQL

1.1. Descripción

Contendor docker partiendo de una instalación base de MariaDB

1.2. Archivo Dockerfile

1.3. Proceso de construcción

1.3.1. En hadoop.ugr.es

1.3.2. En Azure

1.4. Evaluación

Para la evaluación del contenedor se añade una pequeña base de datos sobre la que se realizan las siguientes pruebas.

2. Contenedor para actividades de ciencia de datos basado en Python

2.1. Descripción

Contenedor partiendo de una imagen base de Ubuntu al que se le añade Python con distintos paquetes de ciencia de datos, concretamente:

- pandas
- scikit-learn
- seaborn
- scipy
- numpy
- matplotlib
- xlrd

2.2. Archivo Dockerfile

```
1  FROM ubuntu
2  LABEL author="Ignacio Vellido Expósito"
3  ENV http_proxy http://stargate.ugr.es:3128
4
5  # To don't get asked about geographic location (disable dpkg interactivity)
6  ENV DEBIAN_FRONTEND=noninteractive
7
8  RUN apt-get update && \
9      apt-get install -y python3 python3-pip
10
11 WORKDIR /usr/src/app
12
13 COPY requirements.txt ./
14
15 RUN pip3 install --upgrade pip && \
16     pip3 install --no-cache-dir -r requirements.txt && \
17     rm requirements.txt
18
19 # Launch test script
20 COPY data /home/data
21 COPY testDocker.py /home/testDocker.py
22 RUN cd /home && \
23     python3 /home/testDocker.py > /home/testOutput.txt
24
25 # Launch Python
26 CMD [ "python3" ]
```

Figura 1

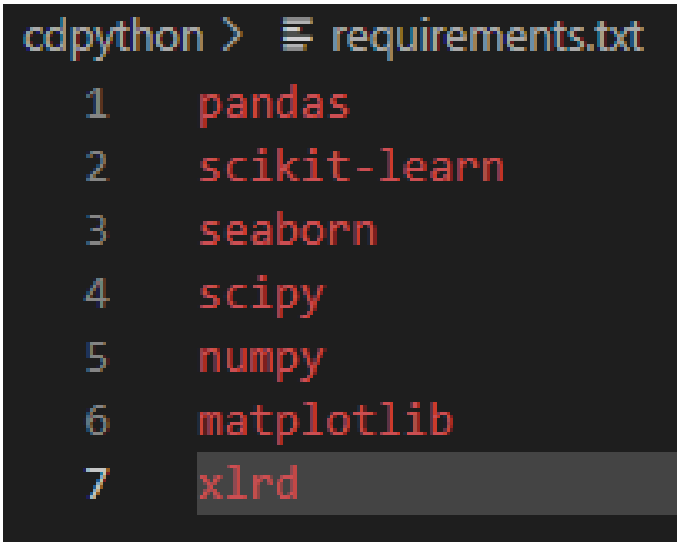

Figura 2: Archivo con los paquetes a instalar

Para la construcción del archivo Dockerfile se parte de las recomendaciones de `https://hub.docker.com/_/python` y se adapta para una instalación base de Ubuntu.

2.3. Proceso de construcción

2.3.1. En hadoop.ugr.es

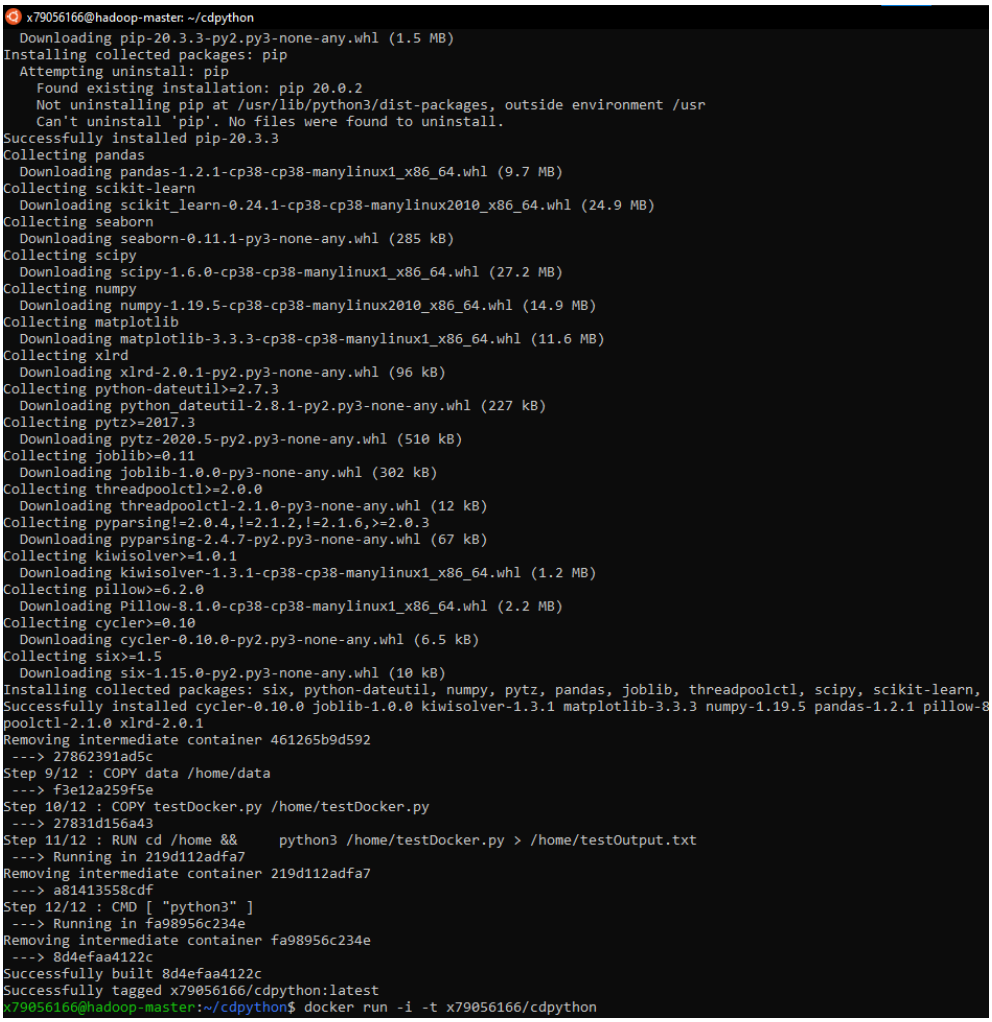

Figura 3: Construcción de la imagen

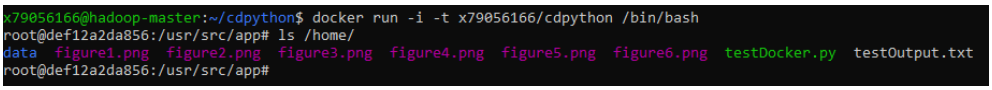

Figura 4: Lanzando la imagen

```
root@def12a2da856:/usr/src/app# cat /home/testOutput.txt
Datos sin normalizar:
        LBE     LB   AC   FM   UC  ...  Median  Variance  Tendency    E  NSP
1     120.0  120.0  0.0  0.0  0.0  ...   121.0      73.0       1.0  0.0  2.0
2     132.0  132.0  4.0  0.0  4.0  ...   140.0      12.0       0.0  0.0  1.0
3     133.0  133.0  2.0  0.0  5.0  ...   138.0      13.0       0.0  0.0  1.0
4     134.0  134.0  2.0  0.0  6.0  ...   137.0      13.0       1.0  0.0  1.0
5     132.0  132.0  4.0  0.0  5.0  ...   138.0      11.0       1.0  0.0  1.0
...     ...    ...  ...  ...  ...  ...     ...       ...       ...  ...  ...
2122  140.0  140.0  0.0  0.0  6.0  ...   152.0       2.0       0.0  1.0  2.0
2123  140.0  140.0  1.0  0.0  9.0  ...   151.0       3.0       1.0  1.0  2.0
2124  140.0  140.0  1.0  0.0  7.0  ...   152.0       4.0       1.0  1.0  2.0
2125  140.0  140.0  1.0  0.0  9.0  ...   151.0       4.0       1.0  1.0  2.0
2126  142.0  142.0  1.0  1.0  5.0  ...   145.0       1.0       0.0  0.0  1.0

[2126 rows x 25 columns]
----------------------------------------------------------------------
Datos normalizados:
        LBE     LB   AC   FM   UC  ...    Mean  Median  Variance  Tendency    E
1     120.0  120.0  0.0  0.0  0.0  ...   137.0   121.0      73.0       1.0  0.0
2     132.0  132.0  4.0  0.0  4.0  ...   136.0   140.0      12.0       0.0  0.0
3     133.0  133.0  2.0  0.0  5.0  ...   135.0   138.0      13.0       0.0  0.0
4     134.0  134.0  2.0  0.0  6.0  ...   134.0   137.0      13.0       1.0  0.0
5     132.0  132.0  4.0  0.0  5.0  ...   136.0   138.0      11.0       1.0  0.0
...     ...    ...  ...  ...  ...  ...     ...     ...       ...       ...  ...
2122  140.0  140.0  0.0  0.0  6.0  ...   150.0   152.0       2.0       0.0  1.0
2123  140.0  140.0  1.0  0.0  9.0  ...   148.0   151.0       3.0       1.0  1.0
2124  140.0  140.0  1.0  0.0  7.0  ...   148.0   152.0       4.0       1.0  1.0
2125  140.0  140.0  1.0  0.0  9.0  ...   147.0   151.0       4.0       1.0  1.0
2126  142.0  142.0  1.0  1.0  5.0  ...   143.0   145.0       1.0       0.0  0.0

[2126 rows x 24 columns]
              LBE           LB  ...     Tendency            E
count  2126.000000  2126.000000  ...  2126.000000  2126.000000
mean    133.303857   133.303857  ...     0.320320     0.033866
std       9.840844     9.840844  ...     0.610829     0.180928
min     106.000000   106.000000  ...    -1.000000     0.000000
25%     126.000000   126.000000  ...     0.000000     0.000000
50%     133.000000   133.000000  ...     0.000000     0.000000
75%     140.000000   140.000000  ...     1.000000     0.000000
max     160.000000   160.000000  ...     1.000000     1.000000

[8 rows x 24 columns]
            counts     freqs
categories
N             1655  0.778457
S              295  0.138758
P              176  0.082785
----------------------------------------------------------------------
Clasificando con SVM
----------------------------------------------------------------------
Mejores hiperparámetros del modelo:
{'C': 100000.0, 'gamma': 0.001, 'kernel': 'poly'}

Mejor score obtenido:
0.8505882352941176

Resultados de la predicción sobre test:
              precision    recall  f1-score   support

      Normal       0.87      0.96      0.91       325
     Suspect       0.86      0.71      0.78        42
  Pathologic       0.53      0.29      0.37        59
```

Figura 5: Contenido de la imagen

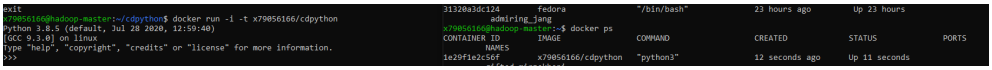

Figura 6: Comprobando ejecución

2.3.2. En Azure

2.4. Evaluación

Para evaluar el correcto funcionamiento se lanza el siguiente script, que carga los paquetes instaladas y realiza un aprendizaje sobre un conjunto de datos con SVM.

```python
#! /usr/bin/env python3
# -*- coding: utf-8 -*-
################################################################################
# Librerías
################################################################################

import random

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Preprocesamiento
from sklearn.preprocessing import Normalizer
from sklearn.model_selection import train_test_split
from sklearn.decomposition import PCA

# Algoritmos
from sklearn.svm import SVC

from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import KFold

# Evaluación
from sklearn.metrics import classification_report, \
                            confusion_matrix, \
                            plot_confusion_matrix

################################################################################
# Lectura
################################################################################

# Semilla con la que se han analizado los resultados
random.seed(9999)

# Cargamos los datos (sheet Raw Data nos es más cómodo que Data)
data = pd.read_excel("data/CTG.xls", "Raw Data")

# Eliminamos las 3 últimas filas que solo contienen valores de máximos y mínimos
data = data[:-3]

# Eliminamos la primera fila que está vacía
data = data[1:]

# Eliminamos las columnas que no contienen información relevante para la
# clasificación
removed_columns = ["FileName", "Date", "SegFile", "b", "e", "A", "B", "C", "D",
                   "AD", "DE", "LD", "FS", "SUSP", "CLASS"]
```

Figura 7: Script de prueba

3. Contenedor para actividades de ciencia de datos basado en R

3.1. Descripción

Contenedor partiendo de una imagen base de Ubuntu al que se le añade R con distintos paquetes de ciencia de datos, concretamente:

- tidyverse
- caret
- RSNNS
- frbs
- FSinR
- forecast

3.2. Archivo Dockerfile

```
FROM ubuntu
LABEL author="Ignacio Vellido Expósito"
ENV http_proxy http://stargate.ugr.es:3128

# To don't get asked about geographic location (disable dpkg interactivity)
ENV DEBIAN_FRONTEND=noninteractive

# Install R and tidyverse
RUN apt-get update && \
    apt-get install -y r-base \
                       r-cran-tidyverse

RUN R -e "install.packages(c('caret','RSNNS','frbs','FSinR','forecast'), \
                           dependencies=TRUE, repos='http://cran.rstudio.com/')"

# Launch test script
COPY testDocker.R /home/testDocker.R
RUN cd /home && \
    Rscript /home/testDocker.R > /home/testOutput.txt

# Launch R
CMD [ "R" ]
```

Figura 8

El paquete "tidyverse" es necesario instalarlo a través de apt-get para evitar errores. Se incluye el proceso de testeo dentro del dockerfile para agilizar las pruebas, y se concluye indicando el comando por defecto de ejecución del script.

3.3. Proceso de construcción

3.3.1. En hadoop.ugr.es

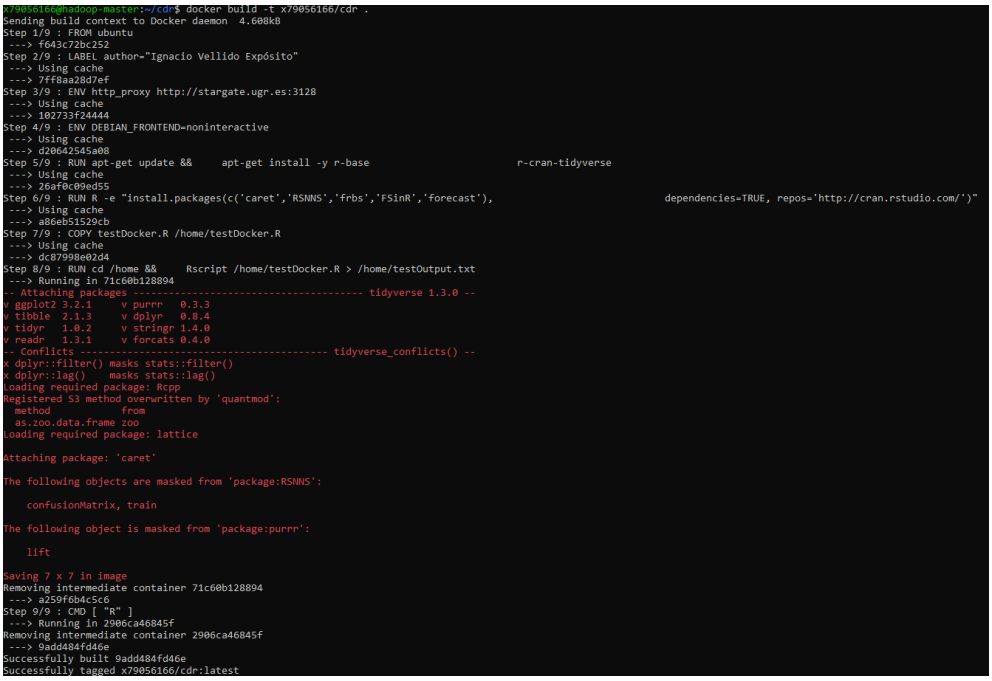

Figura 9: Construcción de la imagen

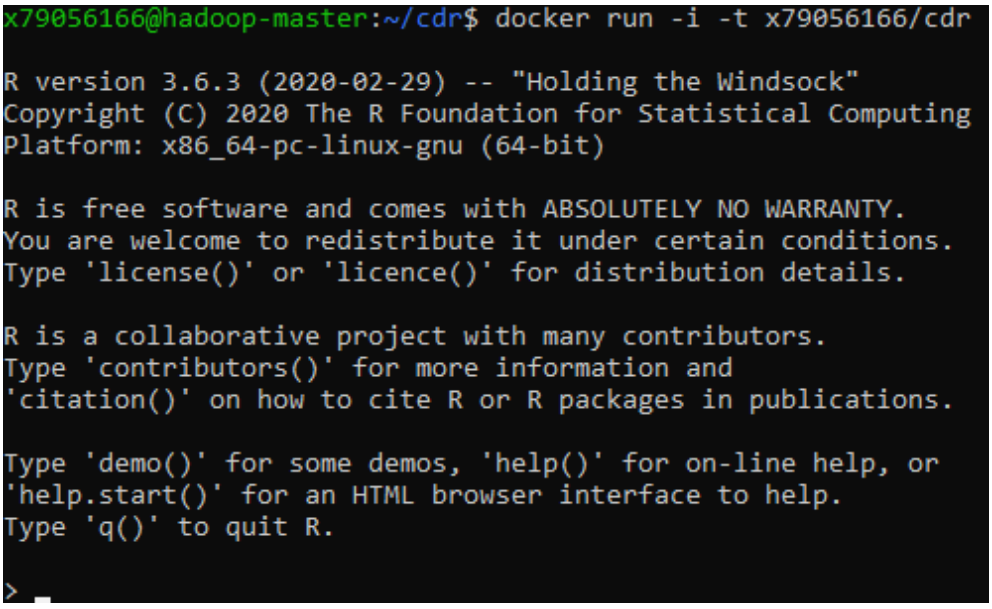

Figura 10: Lanzando la imagen

```
x79056166@hadoop-master:~/cdr$ docker run -i -t x79056166/cdr /bin/bash
root@5ace4be08114:/# ls /home/
Rplots.pdf  testDocker.R  testGgplot.png  testOutput.txt
root@5ace4be08114:/# cat /home/testOutput.txt
k-Nearest Neighbors

150 samples
  4 predictor
  2 classes: 'negative', 'positive'

Pre-processing: centered (4), scaled (4)
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 120, 120, 120, 120, 120, 120, ...
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
   1  0.9400000  0.9600000  0.9200000
   3  0.9701667  0.9666667  0.9266667
   5  0.9895000  0.9766667  0.9400000
   7  0.9923333  0.9766667  0.9066667
   9  0.9918333  0.9800000  0.9066667
  11  0.9908333  0.9733333  0.9133333

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 7.
```

Figura 11: Contenido de la imagen

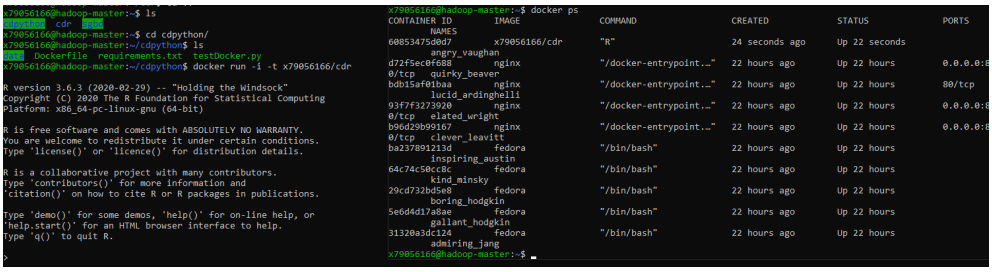

Figura 12: Comprobando ejecución

3.3.2. En Azure

Primeramente se crea un repositorio privado en Docker Hub, y se le cambia el nombre a la imagen de hadoop para adaptarla al repositorio

```
x79056166@hadoop-master:~$ docker login
Login with your Docker ID to push and pull images from Docker Hub. If you don't have a Docker ID,
ub.docker.com to create one.
Username: ignaciove
Password:
WARNING! Your password will be stored unencrypted in /home/x79056166/.docker/config.json.
Configure a credential helper to remove this warning. See
https://docs.docker.com/engine/reference/commandline/login/#credentials-store

Login Succeeded
x79056166@hadoop-master:~$ docker push ignaciove/big_data_i:cdr
The push refers to repository [docker.io/ignaciove/big_data_i]
19c81e3a4cbd: Pushed
6cdcaaf11e6c: Pushed
468c259c5bf8: Pushed
0a63c95bf6c7: Pushed
f6253634dc78: Mounted from danijorq/sgbd
9069f84dbbe9: Mounted from danijorq/sgbd
bacd3af13903: Mounted from danijorq/sgbd
cdr: digest: sha256:193718fa7fd7520e7ce6ba2fa8c79c09283a9693948dea4b293be9ac79d9821a size: 1785
```

Figura 13: Subiendo imagen al repositorio

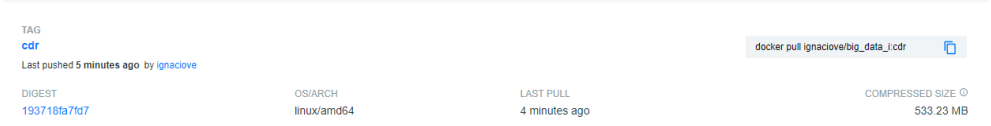

Figura 14: Imagen en Docker Hub

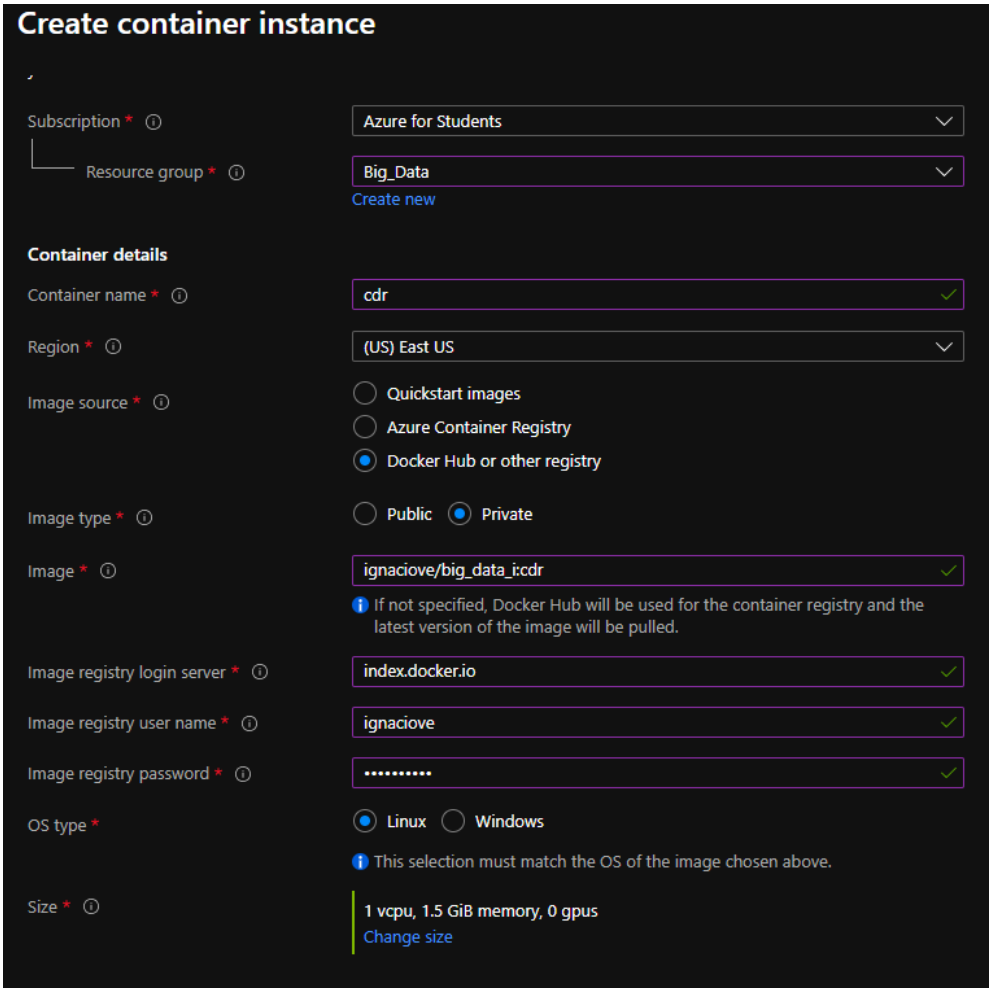

Figura 15: Desplegando el contenedor en Azure

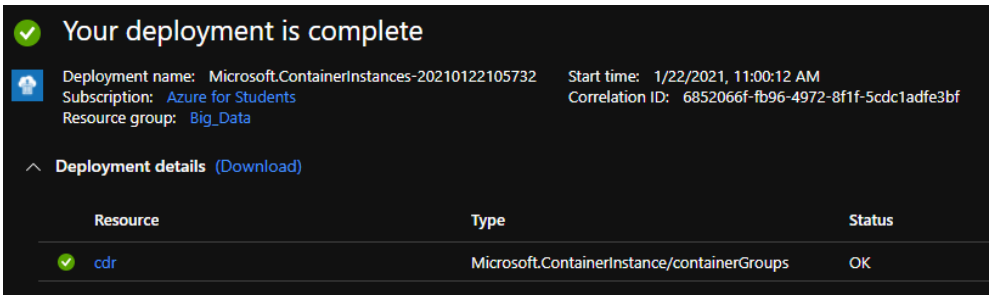

Figura 16: Desplegando el contenedor en Azure

3.4. Evaluación

Para evaluar se ha desplegado el contenedor en las diferentes plataformas y comprobado la salida del scripts. En este se cargan todas las bibliotecas adicionales instaladas y se aplican operaciones con algunas de ellas.

```r
# Test libraries
library(tidyverse)
library(RSNNS)
library(frbs)
library(FSinR)
library(forecast)
library(caret)

# Test ggplot (and tidyverse)
ggplot(iris, aes(x=Sepal.Length, y=Petal.Length))
ggsave("testGgplot.png")

# Test caret
learn_model <-function(dataset, ctrl, message){
  model.fit <- caret::train(Class ~ ., data = dataset, method = "knn",
                            trControl = ctrl, preProcess = c("center","scale"), metric="ROC",
                            tuneGrid = expand.grid(k = c(1,3,5,7,9,11)))
  model.pred <- predict(model.fit,newdata = dataset)
  model.cm <- caret::confusionMatrix(model.pred, dataset$Class,positive = "positive")
  model.probs <- predict(model.fit,newdata = dataset, type="prob")

  return(model.fit)
}

df <- iris
df$Class <- ifelse(df$Species == "virginica", "positive", "negative") %>% as.factor()
df$Species <- NULL

ctrl <- trainControl(method="repeatedcv",number=5,repeats = 3,
                     classProbs=TRUE,summaryFunction = twoClassSummary)
model.raw <- learn_model(df, ctrl, "RAW ")

print(model.raw)
```

Figura 17: Script de prueba

Referencias