

# CA4022 - Hive and Pig: Cloud Deployment

Name: Lara Murphy  
Email: [lara.murphy239@mail.dcu.ie](mailto:lara.murphy239@mail.dcu.ie)  
Student ID: 18390323

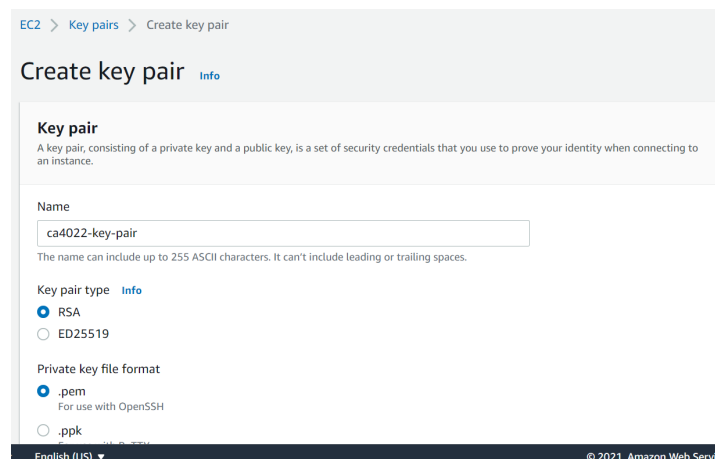
## 1 Setting up EMR

### 1.1 Signing up to AWS.

To sign up to AWS, I was required to make an account, enter billing information, verify my identity through phone, and choose a support plan. I chose the free, basic support plan.

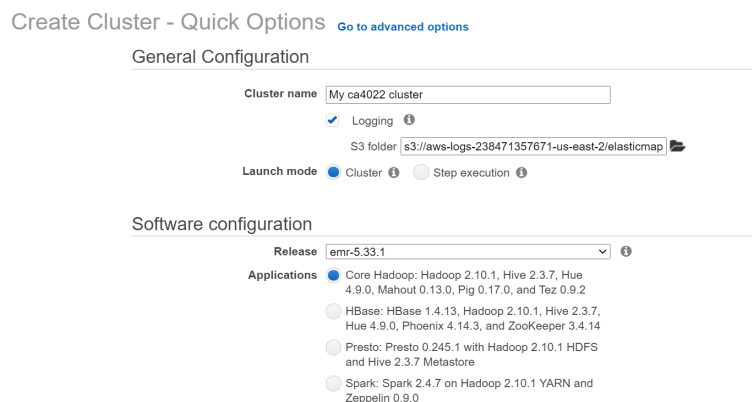
### 1.2 Create an Amazon EC2 key pair for SSH (following [these instructions](#))

From the EC2 console, using the Network & Security pane, I created an EC2 key pair and named it 'ca4022-key-pair'. I set the key pair type to RSA and set the private key file format to .pem. Using Ubuntu, I set the permissions for this private key file to read-only (chmod 400).



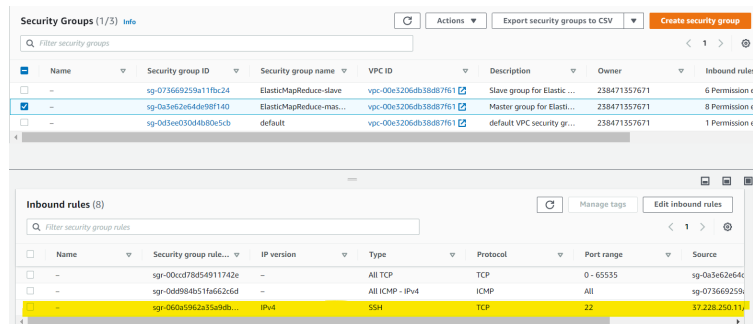
### 1.3 Create a cluster (following [these steps](#))

In the Elastic Map Reduce console, I created a cluster. I left the default values for Release, Instance type, Number of instances, and Permissions, and named the cluster 'My ca4022 cluster'.



## 1.4 Update the SSH Rule

To solve errors regarding a timeout when connecting to the cluster, I updated the SSH rules. From the EC2 dashboard, I added an inbound rule to 'ElasticMapReduce-master' to allow for SSH connections and changed the default source to my IP.



The screenshot shows the AWS Management Console interface for Security Groups. The 'ElasticMapReduce-master' security group is selected, and its inbound rules are displayed. A new rule has been added for SSH access (TCP port 22) from a specific IP address (57.228.250.11).

Name	Security group rule...	IP version	Type	Protocol	Port range	Source
-	sg-00c4f78d54911742e	-	All TCP	TCP	0 - 65535	sg-0a3e62e64c...
-	sg-0d8984b51fa662c6d	-	All ICMP - IPv4	ICMP	All	sg-0736692593...
-	sg-066a5962a35a9db...	IPv4	SSH	TCP	22	57.228.250.11

## 2 Running PIG Scripts

### 2.1 Running PIG Cleaning Scripts

I ran a single cleaning script for each file (movies, ratings, tags and links) and one extra to merge the movies and ratings datasets.

### 2.2 Running PIG Analysis Scripts

I performed three analyses using Pig, namely:

1. Finding the movie with the highest number of ratings

```
2021-10-25 18:16:52,361 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-10-25 18:16:52,361 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(356,Forrest Gump,329)
2021-10-25 18:16:52,406 [main] INFO org.apache.pig.Main - Pig script completed in 8 seconds and 154 milliseconds (8154 ms)
laramurphyx@LAPTOP-T2T1G8JL:~/hadoop-3.3.0$
```

2. Movies with the highest proportion of five stars

```
2021-10-25 18:22:03,813 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-10-25 18:22:03,813 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(318,Shawshank Redemption, The,153,317,0.48264983)
(858,Godfather, The,88,192,0.45833334)
(1208,Apocalypse Now,45,107,0.42056075)
(527,Schindler's List,92,220,0.4181818)
(260,Star Wars: Episode IV - A New Hope,104,251,0.41434264)
2021-10-25 18:22:03,865 [main] INFO org.apache.pig.Main - Pig script completed in 8 seconds and 166 milliseconds (8166 ms)
laramurphyx@LAPTOP-T2T1G8JL:~/hadoop-3.3.0$
```

3. Users with the highest average rating

```
2021-10-26 18:48:21,471 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(53,20,5.0)
(251,23,4.826086956521739)
(515,26,4.8076923076923075)
(25,26,4.769230769230769)
(30,34,4.705882352941177)
2021-10-26 18:48:21,503 [main] INFO org.apache.pig.Main - Pig script completed in 4 seconds and 450 milliseconds (4450 ms)
laramurphyx@LAPTOP-T2T1G8JL:~/hadoop-3.3.0$
```

### 3 Running Hive Analysis Scripts

I performed six analyses using Hive, namely:

#### 1. Finding the Movie with the Highest Number of Ratings

```
-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    1         1         0         0         0         0
Reducer 2 ..... container  SUCCEEDED    2         2         0         0         0         0
Reducer 3 ..... container  SUCCEEDED    1         1         0         0         0         0
-----
VERTICES: 03/03 [=====>>] 100% ELAPSED TIME: 4.97 s
-----
OK
296    Pulp Fiction    325
Time taken: 12.476 seconds, Fetched: 1 row(s)
[hadoop@ip-172-31-42-93 assignment1]$
```

#### 2. Finding the Movie with the Highest Proportion of Five Star Ratings

```
-----
VERTICES: 04/04 [=====>>] 100% ELAPSED TIME: 0.59 s
-----
Moving data to directory hdfs://ip-172-31-42-93.us-east-2.compute.internal:8020/user/hive/warehouse
OK
Time taken: 1.525 seconds
OK
858    Godfather, The    112    210    0.5333333333333333
318    Shawshank Redemption, The  157    308    0.5097402597402597
527    Schindler's List      114    248    0.4596774193548387
912    Casablanca           55     125    0.44
50     Usual Suspects, The   100    228    0.43859649122807015
1136   Monty Python and the Holy Grail 65     154    0.42207792207792205
1221   Godfather: Part II, The 59     140    0.42142857142857143
2959   Fight Club           79     207    0.38164251207729466
1196   Star Wars: Episode V - The Empire Strikes Back 87     228    0.3815789473684211
260    Star Wars: Episode IV - A New Hope 104    273    0.38095238095238093
Time taken: 0.132 seconds, Fetched: 10 row(s)
[hadoop@ip-172-31-42-93 assignment1]$
```

#### 3. Finding the Users with the Highest Average Rating

```
-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED    1         1         0         0         0         0
Reducer 2 ..... container  SUCCEEDED    2         2         0         0         0         0
Reducer 3 ..... container  SUCCEEDED    1         1         0         0         0         0
-----
VERTICES: 03/03 [=====>>] 100% ELAPSED TIME: 5.11 s
-----
OK
78     35     5.0
15     73     4.835616438356165
550    21     4.761904761904762
432    200    4.6
520    37     4.594594594594595
Time taken: 12.216 seconds, Fetched: 5 row(s)
[hadoop@ip-172-31-42-93 assignment1]$
```

#### 4. Finding the Count of Each Rating per Movie

```
-----
VERTICES: 09/09 [=====>>>] 100% ELAPSED TIME: 1.13 s
-----
OK
1027 Robin Hood: Prince of Thieves 39 78 234 702 390 78
1035 Sound of Music, The 58 174 232 580 928 1392
1036 Die Hard 165 495 1320 6765 12210 6270
105 Bridges of Madison County, The 35 140 105 315 455 175
1088 Dirty Dancing 112 112 280 1344 784 504
1093 Doors, The 32 32 224 352 320 64
1097 E.T. the Extra-Terrestrial 156 780 1560 7176 8268 6396
110 Braveheart 496 496 1736 12896 24304 21576
1100 Days of Thunder 116 29 145 406 87 58
110102 Captain America: The Winter Soldier 34 17 17 85 119 17
Time taken: 2.41 seconds, Fetched: 10 row(s)
[hadoop@ip-172-31-42-93 assignment1]$
```

#### 5. Finding the Most Popular Rating for All Movies

```
-----
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 2 2 0 0 0 0
Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0
-----
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 0.74 s
-----
OK
4 37067
Time taken: 1.289 seconds, Fetched: 1 row(s)
[hadoop@ip-172-31-42-93 assignment1]$
```

#### 6. Exploring how the Ratings are Distributed by Genre

```
-----
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
-----
VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 5.32 s
-----
OK
31205 23076 5966 8098 38055 18291 1206 46960 10889 1210 7983 4287 8320 19094 16795 2
9288 5828 2314
Time taken: 12.387 seconds, Fetched: 1 row(s)
[hadoop@ip-172-31-42-93 assignment1]$
```

## 4 Conclusion

These images are uploaded to GitHub in the ['Cloud Deployment' directory](#).

There are further explanations for the analyses performed in the same GitHub repository, specifically in the markdown document, [Documentation.md](#).