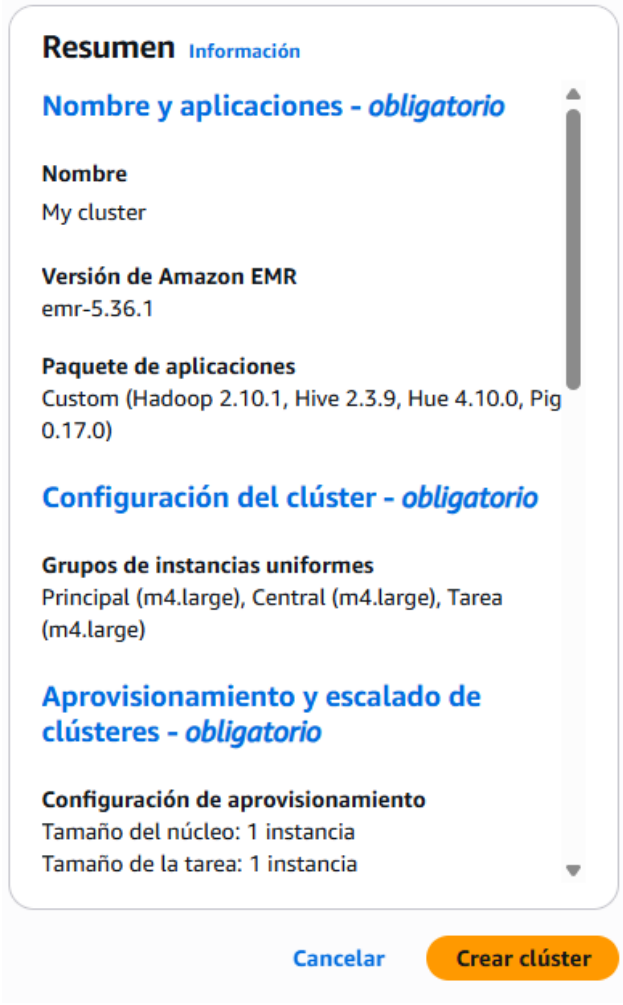
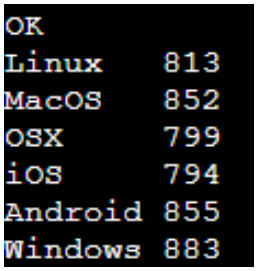
AWS Skillbuilder

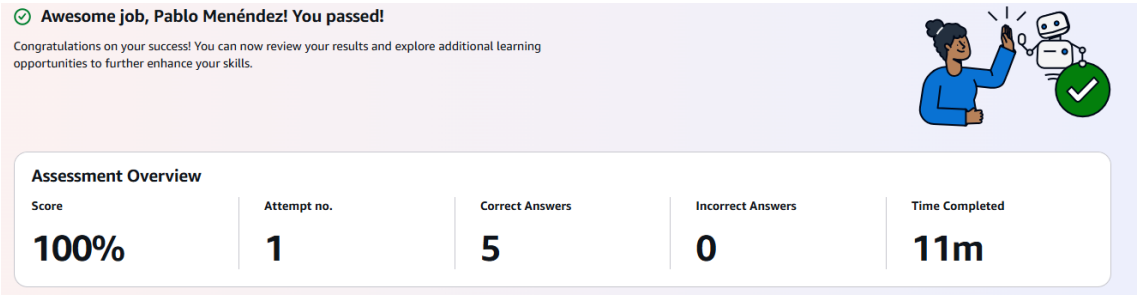
1) APARTADO A

1) Realiza el laboratorio “Analyze Big Data with Hadoop (Español de España)”.



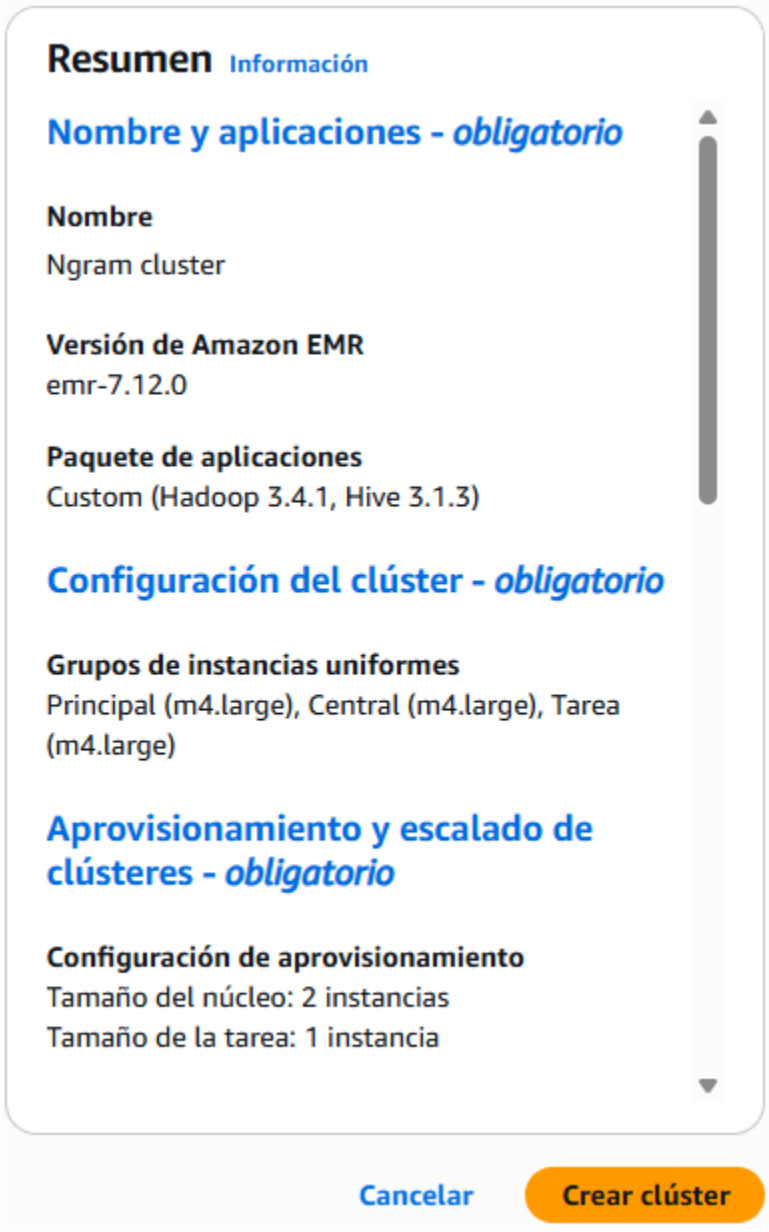
Consulta de Hive:





2) APARTADO B

1) Realiza el laboratorio “Exploring Google Ngrams with Amazon EMR and Hive (Español de España)”.



```
hive> CREATE EXTERNAL TABLE ngrams
    > (gram string, year int, occurrences bigint, pages bigint, books bigint)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
    > STORED AS SEQUENCEFILE
    > LOCATION 's3://datasets.elasticmapreduce/ngrams/books/20090715/eng-1M/1gram/';
OK
Time taken: 5.943 seconds
hive> DESCRIBE ngrams;
OK
gram                    string
year                    int
occurrences             bigint
pages                   bigint
books                   bigint
Time taken: 0.337 seconds, Fetched: 5 row(s)
```

```
hive> SELECT * FROM ngrams LIMIT 10;
OK
#       1574    1       1       1
#       1584    6       6       1
#       1614    1       1       1
#       1631    115     100     1
#       1632    3       3       1
#       1635    1       1       1
#       1640    1       1       1
#       1641    1       1       1
#       1642    5       5       1
#       1644    234     193     1
Time taken: 4.375 seconds, Fetched: 10 row(s)
```

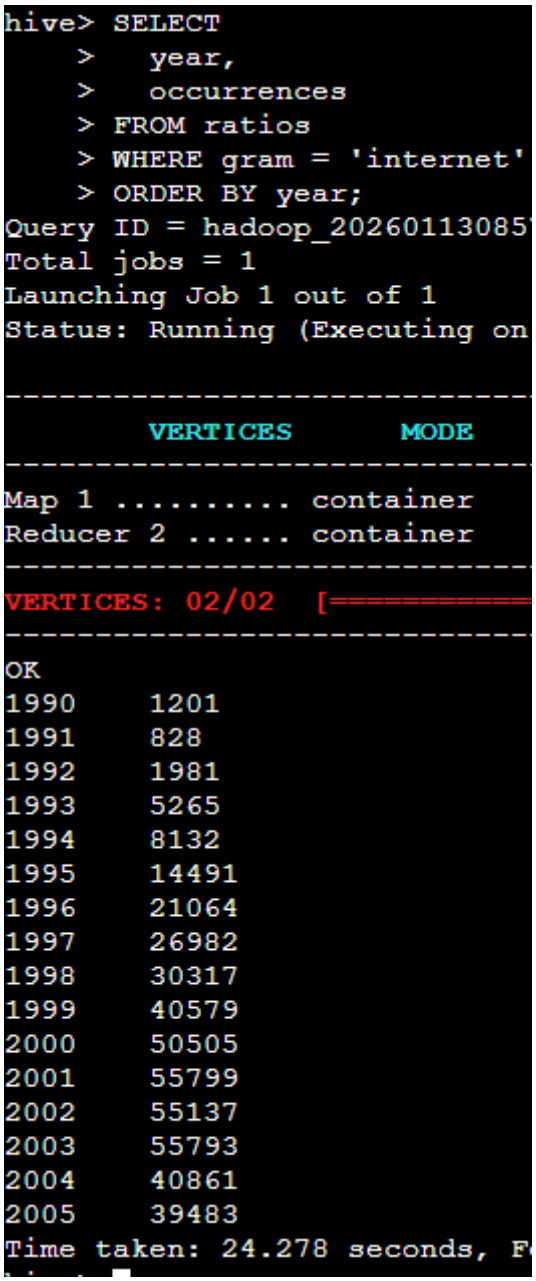
```
hive> INSERT OVERWRITE TABLE normalized
    > SELECT lower(gram), year, occurrences
    > FROM ngrams
    > WHERE year BETWEEN 1990 AND 2005
    > AND gram REGEXP "^[A-Za-z+\'-]{3,}$";
Query ID = hadoop_20260113082945_f9f77029-f481-4712-a892-240a5933b1cc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1768292604981_0001)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 187.26 s
--------------------------------------------------------------------------------
Loading data to table default.normalized
OK
Time taken: 192.345 seconds
```

```
hive> SELECT * FROM normalized LIMIT 20;
OK
ingermany       1991    1
ingermany       1993    1
ingermany       1994    3
ingermany       1996    1
ingermany       2001    1
ingermany       2004    1
ingermany       2005    1
ingreece        1990    1
ingreece        2001    1
ingreece        2004    1
injuly  1990    7
injuly  1991    3
injuly  1992    6
injuly  1993    4
injuly  1994    1
injuly  1995    5
injuly  1996    4
injuly  1998    4
injuly  1999    3
injuly  2000    6
Time taken: 0.159 seconds, Fetched: 20 row(s)
```

```
hive> SELECT
    >   gram,
    >   sum(occurrences) as total_occurrences
    > FROM normalized
    > GROUP BY gram
    > ORDER BY total_occurrences DESC
    > LIMIT 50;
Query ID = hadoop_20260113083711_a1cc6acf-2317-4305-b2b2-f5
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id appl

----------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED
----------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     21         21
Reducer 2 ...... container     SUCCEEDED      1          1
Reducer 3 ...... container     SUCCEEDED      1          1
----------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPS
----------------------------------------------------------------
OK
the     600731810
and     269591500
that    94084329
for     80649257
with    61620362
was     57843905
this    45202579
are     44749547
from    40039900
not     38905683
his     33689806
have    31100386
but     29014171
which   28732316
you     27930990
they    27538747
had     26712182
were    24825903
their   24729315
one     23646682
all     21382534
can     19855720
her     19278458
has     18670586
more    18489067
there   17394454
when    17001334
been    16865057
she     16766345
```

```
hive> SELECT
    >   gram,
    >   sum(occurrences) as total_occurrences
    > FROM normalized
    > WHERE length(gram) > 10
    > GROUP BY gram
    > ORDER BY total_occurrences DESC
    > LIMIT 50;
Query ID = hadoop_20260113084059_a5189451-8d0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster wi

----------------------------------------------
        VERTICES      MODE        STATUS  TOT
----------------------------------------------
Map 1 .......... container     SUCCEEDED
Reducer 2 ...... container     SUCCEEDED
Reducer 3 ...... container     SUCCEEDED
----------------------------------------------
VERTICES: 03/03  [==========================>
----------------------------------------------
OK
development     4584319
information     4419750
international   2731441
relationship    2013252
significant     1762598
particularly    1709008
performance     1669887
understanding   1631334
environment     1511561
organization    1491293
established     1441713
association     1385892
individuals     1376707
differences     1353913
traditional     1314358
appropriate     1309280
application     1289144
distribution    1278733
environmental   1170366
temperature     1140588
independent     1120174
communication   1114437
introduction    1111958
administration  1107008
relationships   1022809
institutions    1015093
construction    1005391
professional    975041
```

```
hive> SELECT year, gram, occurrences, CONCAT(CAST
    > (
    >   SELECT
    >     y2.gram,
    >     y2.year,
    >     y2.occurrences,
    >     y2.ratio / y1.ratio as increase,
    >     rank() OVER (PARTITION BY y2.year ORDER
    >   FROM ratios y2
    >   JOIN ratios y1 ON y1.gram = y2.gram and y
    >   WHERE
    >       y2.year BETWEEN 1991 and 2005
    >   AND y1.occurrences > 1000
    >   AND y2.occurrences > 1000
    > ) grams
    > WHERE rank = 1
    > ORDER BY year;
Query ID = hadoop_20260113085448_cc3b8c52-3832-4f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with A

----------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL
----------------------------------------------------
Map 4 .......... container     SUCCEEDED     18
Map 1 .......... container     SUCCEEDED     18
Reducer 2 ...... container     SUCCEEDED      1
Reducer 3 ...... container     SUCCEEDED      1
----------------------------------------------------
VERTICES: 04/04  [==========================>>] 1
----------------------------------------------------
OK
1991    amyloid 6405    5x increase
1992    comm    18841   8x increase
1993    abstr   7033    6x increase
1994    carole  8358    7x increase
1995    mansfield       4570    3x increase
1996    polymerization  14442   8x increase
1997    tho     19259   8x increase
1998    oswald  8774    6x increase
1999    sql     12516   6x increase
2000    dlb     12369   10x increase
2001    dcs     6031    5x increase
2002    proust  6231    5x increase
2003    olfactory       8538    6x increase
2004    eeg     8873    5x increase
2005    rectum  6981    6x increase
Time taken: 45.659 seconds, Fetched: 15 row(s)
```


hive> SELECT
> year,
> occurrences
> FROM ratios
> WHERE gram = 'internet'
> ORDER BY year;
Query ID = hadoop_20260113085
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on
VERTICES MODE
Map 1 ..... container
Reducer 2 ..... container
VERTICES: 02/02
OK
1990 1201
1991 828
1992 1981
1993 5265
1994 8132
1995 14491
1996 21064
1997 26982
1998 30317
1999 40579
2000 50505
2001 55799
2002 55137
2003 55793
2004 40861
2005 39483
Time taken: 24.278 seconds, F

```
----------------------------------------------
        VERTICES      MODE        STATUS  TOT
----------------------------------------------
Map 1 .......... container     SUCCEEDED
Reducer 2 ...... container     SUCCEEDED
Reducer 3 ...... container     SUCCEEDED
Reducer 4 ...... container     SUCCEEDED
----------------------------------------------
VERTICES: 04/04  [==========================>
----------------------------------------------
OK
3       the
4       that
5       which
6       people
7       between
8       american
9       different
10      university
11      development
12      relationship
13      international
14      administration
15      characteristics
16      responsibilities
17      industrialization
18      telecommunications
19      hyperparathyroidism
20      institutionalization
21      psychopharmacological
22      electroencephalography
23      electroencephalographic
24      cholangiopancreatography
25      methylenetetrahydrofolate
26      abcdefghijklmnopqrstuvwxyz
27      ooooooooooooooooooooooooooo
28      trimethoprimsulfamethoxazole
29      methylenedioxymethamphetamine
30      dipalmitoylphosphatidylcholine
31      dichlorodiphenyltrichloroethane
32      oooooooooooooooooooooooooooooo
33      ooooooooooooooooooooooooooooooo
34      oooooooooooooooooooooooooooooooo
35      ooooooooooooooooooooooooooooooooo
36      oooooooooooooooooooooooooooooooooo
```

2) Contesta a las siguientes respuestas:

   i. ¿Qué contiene el bucket s3://datasets.elasticmapreduce/ngrams/books/20090715/eng-1M/1gram/? ¿Cuánto ocupa el archivo que contiene?

El bucket contiene todos los datos importados desde Google Ngrams y outputs de acciones realizadas. Ocupa 50 MB.

ii. ¿Cuántos registros contiene la tabla ngrams que creaste en HIVE? ¿Desde qué año hasta qué año abarca la información que contiene?

Contiene varios millones. Abarca entre los años 1990 y 2005.

iii. En la creación de la tabla normalized ¿qué significa la expresión REGEXP "^[A-Za-z+\'-]{3,}$"? ¿Cuántos registros contiene la tabla normalized?

Significa que puede tener letras mayúsculas, minúsculas o ciertos símbolos, un mínimo de tres. Contiene varios cientos de miles.