Universidad del Desarrollo
Facultad de Ingeniería

Hands-on I Big Data & Cloud Computing

ABRIL - 2021

SR. Gustavo Adolfo Lobos Curamil

ABRIL 2021

Universidad de Desarrollo
Facultad de Ingeniería
BIG DATA & CLOUD COMPUTING

Introducción:

En el presente documento se encuentran las respuestas al primer Hands On del ramo BIG DATA incluidas en el PDF 'Hands-on_I_2021_step-by-step'.
El logro de los objetivos fue realizado utilizando una máquina virtual con Ubuntu en su versión 20.10.

Como el dataset de GDELT se conforma de 03 archivos de extensión .csv y que de por si estaban separados y subidos cronológicamente, descargar los archivos uno por uno no era práctico.
Es por esta razón por la que se confeccionó un pequeño script diseñado en Bash para esta tarea:

```
nash@vm:~/gdelt$ ls -l |grep mdd
-rwxr-xr-x 1 root root     1464 abr 18 15:24 mdd.sh
nash@vm:~/gdelt$
```

La función de este script consiste en descargar el contenido de la página '[Master CSV Data File List](#)' que se actualiza cada 15 minutos.
El contenido de esta web es importante para el script dado que de ahí se extrae la información con los enlaces de descarga de cada uno de los archivos, y eso es lo que busca el script.

```
#!/bin/bash
wget https://www.dropbox.com/s/c9o2ouce2t1zfah/Webs.txt
cat Webs.txt |grep /2020 > 2020
cat 2020 |awk '{print $3}' > links
cat links |grep /2020 |grep -e 202009 -e 202010 -e 202011 -e 202012 > trimestr4
cat trimestr4 |grep mentions > mentions
cat trimestr4 |grep export > export
cat trimestr4 |grep gkg > gkg
```

Posteriormente lo único que se hace es generar la estructura de directorios sugerida en el desafío para que los archivos sean descargados directamente.

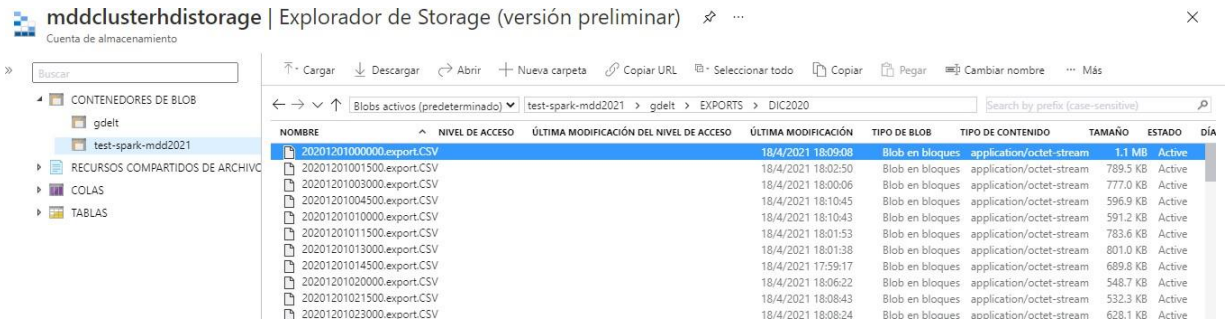
```
mkdir MENTIONS
mkdir EXPORTS
mkdir GKG
mkdir MENTIONS/SEP2020
mkdir MENTIONS/OCT2020
mkdir MENTIONS/NOV2020
mkdir MENTIONS/DIC2020
mkdir EXPORTS/SEP2020
mkdir EXPORTS/OCT2020
mkdir EXPORTS/NOV2020
mkdir EXPORTS/DIC2020
mkdir GKG/SEP2020
mkdir GKG/OCT2020
mkdir GKG/NOV2020
mkdir GKG/DIC2020
```

```
nash@vm:~/gdelt/EXPORTS$ ls -l
total 980
drwxr-xr-x 2 root root 319488 abr 18 15:47 DIC2020
drwxr-xr-x 2 root root 196608 abr 18 15:58 NOV2020
drwxr-xr-x 2 root root 184320 abr 18 16:11 OCT2020
drwxr-xr-x 2 root root 290816 abr 18 16:12 SEP2020
nash@vm:~/gdelt/EXPORTS$ cd
```

Por último, se descomprimen los archivos, y se eliminan todos los archivos que tengan la extensión “.zip” para sólo dejar los “.csv” ordenados dentro de los directorios.

```
nash@vm:~/gdelt/EXPORTS/SEP2020$ ls |head
20200901000000.export.CSV
20200901001500.export.CSV
20200901003000.export.CSV
20200901004500.export.CSV
20200901010000.export.CSV
20200901011500.export.CSV
20200901013000.export.CSV
20200901014500.export.CSV
20200901020000.export.CSV
20200901021500.export.CSV
nash@vm:~/gdelt/EXPORTS/SEP2020$
```

Una vez que se validó que el script funcionaba correctamente se armó el cluster Hadoop en Azure y se ejecutó allí. Cuando terminó se envió los directorios al Storage y permanecieron ahí para las pruebas.



PRIMERA ETAPA

La primera etapa de las preguntas enumeradas de 1 a 6 son las siguientes:

1. ¿Qué información contiene el dataset GDELT? (1 punto)
    a. El proyecto de la base de datos global de eventos, lenguaje y tono (GDELT por sus siglas en inglés) es una base datos abierta global en tiempo real de la sociedad humana según los medios informativos del mundo, que profundiza en los acontecimientos, las reacciones y emociones de cada parte del mundo en tiempo casi real. Toda esta información está disponible de forma gratuita para investigar, analizar, visualizar e incluso predecir la sociedad humana de acuerdo con la cobertura de noticias globales. También incluye un catálogo completo y de alta resolución de los eventos sociopolíticos geo-referenciados desde 1979 hasta la actualidad. El proyecto GDELT hace un seguimiento de cada boletín de noticias impreso y digital accesible en todo el mundo cada 15 minutos en más de 100 idiomas. La información se procesa utilizando una gran variedad de algoritmos para identificar cientos de categorías de eventos (desde protestas hasta llamamientos por la paz), miles de emociones (desde la ansiedad hasta la felicidad), millones de temas narrativos (desde los derechos de las mujeres hasta el acceso al agua potable), así como ubicaciones, personas, organizaciones y otros indicadores.

2. ¿Cuál es el número total de archivos? (1 punto)

   Dentro de este dataset separado por año, mes y día vienen 3 tipos de archivos y la cantidad de archivos por cada dataset es la siguiente:
   - Export.csv = 9134
   - Gkg.csv = 9134
   - Mentions.csv = 9134

```
sshuser@hn0-mddspa:~$ cat archivos_trimestre4 |grep gkg |wc -l
9134
sshuser@hn0-mddspa:~$ cat archivos_trimestre4 |wc -l
27402
sshuser@hn0-mddspa:~$ cat archivos_trimestre4 |grep gkg |wc -l
9134
sshuser@hn0-mddspa:~$ cat archivos_trimestre4 |grep export |wc -l
9134
sshuser@hn0-mddspa:~$ cat archivos_trimestre4 |grep mention |wc -l
9134
sshuser@hn0-mddspa:~$
```

3. ¿Cuántas columnas tiene cada tipo de archivo (export, mentions o gkg)? (1 punto)
   a. Como podemos apreciar en la siguiente figura la cantidad de columnas están definidas como sigue:

```
[ ] export = pd.read_csv('/content/20150218230000.export.CSV', sep = '\t' , header=None)
    gkg = pd.read_csv('/content/20150218230000.gkg.csv', sep = '\t', header=None)
    mentions = pd.read_csv('/content/20150218230000.mentions.CSV', sep = '\t', header=None)
```

```
[ ] print('Cantidad de columnas por tipo de archivo:')
    print('export   : {}\ngkg      : {}\nmentions : {}'.format(export.shape[1], gkg.shape[1],mentions.shape[1]))

    Cantidad de columnas por tipo de archivo:
    export   : 61
    gkg      : 27
    mentions : 16
```

Utilizando bash en el cluster, se llega a la misma conclusión

```
sshuser@hn0-mddspa:~/archivos$ awk -F"\t" '{print NF}' 20201231234500.export.CSV |sort |uniq -c
    961 61
sshuser@hn0-mddspa:~/archivos$ awk -F"\t" '{print NF}' 20201231234500.mentions.CSV |sort |uniq -c
   3548 16
sshuser@hn0-mddspa:~/archivos$ awk -F"\t" '{print NF}' 20201231234500.gkg.csv |sort |uniq -c
      1 1
   1296 27
sshuser@hn0-mddspa:~/archivos$
```

4. ¿Cuál es el tamaño promedio de cada tipo de archivo? (2 puntos)

```
globos@linuxmdd:~$ ll |grep export |gawk '{sum += $5; n++;} END {print sum/n}'
554094
globos@linuxmdd:~$ ll |grep mention |gawk '{sum += $5; n++;} END {print sum/n}'
815958
globos@linuxmdd:~$ ll |grep gkg |gawk '{sum += $5; n++;} END {print sum/n}'
1.94033e+07
globos@linuxmdd:~$
```

- Mentions = 815958 bytes
- Export = 554094 bytes
- GKG = 19403300 bytes (18.5 megas)

5. ¿Cuántas filas en promedio tiene cada tipo de archivo? (2 puntos)
    - Mentions = 3825.38
    - Export = 1356.42
    - GKG =

```
globos@linuxmdd:~$ awk 'END{FILAS=ARGC-1; print (NR-FILAS)/FILAS}' *.mentions.CSV
3825.38
globos@linuxmdd:~$ awk 'END{FILAS=ARGC-1; print (NR-FILAS)/FILAS}' *.export.CSV
1356.42
```

6. ¿Existe algún archivo cuyo peso en bytes esté fuera de rango? Debe definir qué significa fuera de rango. (3 puntos)
   a. Primero se extrae el tamaño de cada archivo para posteriormente procesarlo con Python y calcular los outliers.

```
globos@linuxmdd:~$ ll |grep mentions |awk '{print $5}' > df_mentions
globos@linuxmdd:~$ ll |grep export |awk '{print $5}' > df_export
globos@linuxmdd:~$ ll |grep gkg |awk '{print $5}' > df_gkg
```

   b. Se considerará como valor fuera de rango o atípico, aquel valor que se encuentre a 1.5 veces el rango intercuartil (Q3 - Q1) de acuerdo con Tukey. Conforme a la siguiente figura, y a la definición anterior si existen outliers.

```
[8] def IQR(data):
      '''Calcula el Rango intercuartil de acuerdo a los datos entregados'''
      q1 = int(np.quantile(data,[.25]))
      q3 = int(np.quantile(data,[.75]))
      IQR = q3-q1
      return IQR
```

```
[9] iqr_export =IQR(df_export)
    iqr_gkg =IQR(df_gkg)
    iqr_mentions =IQR(df_mentions)
```

```
[10] cant_out_1 = df_export[df_export>iqr_export*1.5].count()
     cant_out_2 = df_gkg[df_gkg>iqr_gkg*1.5].count()
     cant_out_3 = df_mentions[df_mentions>iqr_mentions*1.5].count()
     print('Cant. de outlier por tipo de archivos')
     print('export   = {:,}\ngkg      = {:,}\nmentions = {:,}'.format(cant_out_1,
                                                                      cant_out_2,
                                                                      cant_out_3))
```

```
Cant. de outlier por tipo de archivos
export   = 5,698
gkg      = 5,065
mentions = 5,527
```

SEGUNDA ETAPA

1. ¿Cuál fue el sitio web con mayor cantidad de eventos registrados el 6 de octubre del 2020?

   a. Utilizando PIG, la respuesta a esta pregunta es Reuters.com con 7993 menciones. El script confeccionado fue:

```
sshuser@hn0-mdd202:~$ cat query1.pig
set tez.runtime.io.sort.mb 1089;
DATA = LOAD 'wasb://test-spark-mdd2021@mddclusterhdistorage.blob.core.windows.net/gdelt/EXPORTS/OCT2020/20201006*' using PigStorage('\t');
PAGINA = GROUP DATA BY $4;
CONTADOR = FOREACH PAGINA GENERATE group,COUNT(DATA);
ORDENADOR = order CONTADOR by $1 desc;
TOP_10 = limit ORDENADOR 10;
dump TOP_10;
sshuser@hn0-mdd202:~$
sshuser@hn0-mdd202:~$
```

Y la salida indica:

```
sshuser@hn0-mdd202:~$ cat respuesta11
(reuters.com,7993)
(msn.com,7692)
(iheart.com,5390)
```

   b. Con Hive, se llega a la misma conclusión, con los mismos contadores

```
INFO  : Completed executing command(queryId:
59df48d0); Time taken: 48.111 seconds
INFO  : OK
+----------------------+-------+
|  mentionsourcename   |  _c1  |
+----------------------+-------+
| reuters.com          | 7993  |
| msn.com              | 7692  |
| iheart.com           | 5390  |
| yahoo.com            | 3575  |
| dailymail.co.uk      | 2190  |
| allafrica.com        | 2008  |
| washingtontimes.com  | 1763  |
| indiatimes.com       | 1632  |
| wickedlocal.com      | 1558  |
| chron.com            | 1321  |
+----------------------+-------+
10 rows selected (48.644 seconds)
0: jdbc:hive2://hn0-mdd202:10001/>
```

Y la sentencia ejecutada fue:

```
select mentionsourcename, count(*) from default.mentions6 where
mentiontimedate like '20201006%' group by mentionsourcename order
by 2 desc limit 10;
```

2. ¿Cuál es la ciudad con mayor número de apariciones en los eventos registrados el 6 de octubre del 2020?
   a. Utilizando PIG, el primer lugar lo ocupa ninguna ciudad, ya que en la mayoría de los eventos no se realiza la relación con alguna ciudad, por lo cual ya en los siguientes puestos comienzan a aparecer locaciones según el 'Geo_FeatureID' deGDELT. En ese sentido la ciudad con mayor número de apariciones es Washington DC.

      El script utilizado fue:

```
sshuser@hn0-mdd202:~$ cat query2.pig
set tez.runtime.io.sort.mb 1089;
DATA = LOAD 'wasb://test-spark-mdd2021@mddclusterhdistorage.blob.core.windows.net/gdelt/EXPORTS/OCT2020/20201006*' using PigStorage('\t');
PAGINA = GROUP DATA BY $52;
CONTADOR = FOREACH PAGINA GENERATE group,COUNT(DATA);
ORDENADOR = order CONTADOR by $1 desc;
TOP_10 = limit ORDENADOR 10;
dump TOP_10;
sshuser@hn0-mdd202:~$
```

      Y el resultado fue:

```
sshuser@hn0-mdd202:~$ cat respuesta2
(,38031)
(White House, District of Columbia, United States,2320)
(United States,1802)
```

   b. Hive señala que, en este caso, se llega a la misma conclusión con los mismos contadores. Aunque el número de contadores depende de la columna seleccionada, todas dan el mismo resultado de ciudad:

```
INFO  : OK
+-----------------------------------------------------+--------+
|                actor2geo_fullname                   |  _c1   |
+-----------------------------------------------------+--------+
|                                                     | 38031  |
| White House, District of Columbia, United States    | 2320   |
| United States                                       | 1802   |
| Washington, District of Columbia, United States     | 1490   |
| New York, United States                             | 1468   |
| United Kingdom                                      | 1439   |
| London, London, City of, United Kingdom             | 1018   |
| China                                               | 1005   |
| Beijing, Beijing, China                             | 857    |
| Texas, United States                                | 841    |
+-----------------------------------------------------+--------+
10 rows selected (43.216 seconds)
```

```
select Actor2Geo_FullName, count(*) from default.exports2 where DATEADDED like '20201006%' group by Actor2Geo_FullName order by 2 desc limit 10;
```

3. ¿Cuál fue el evento generado durante noviembre del 2020 con mayor cantidad de menciones en los medios?

a. Utilizando PIG el evento encontrado fue 957018602 con 1079 menciones.

```
sshuser@hn0-mdd202:~$ cat t3
(957018602,1079)
(953950018,894)
(957018542,834)
```

Y este corresponde a un evento en donde un senador fue covid positivo tras tener contacto con Donald Trump

```
sshuser@hn0-mdd202:~$ hadoop fs -cat 'wasb://test-spark-mdd2021@mddclusterhdistorage.blob.core.windows.net/gdelt/EXPORTS/NOV2020/' |grep 957018602
cat: `wasb://test-spark-mdd2021@mddclusterhdistorage.blob.core.windows.net/gdelt/EXPORTS/NOV2020': Is a directory
sshuser@hn0-mdd202:~$ hadoop fs -cat 'wasb://test-spark-mdd2021@mddclusterhdistorage.blob.core.windows.net/gdelt/EXPORTS/NOV2020/*' |grep 957018602
957018602       20201130        202011  2020    2020.9041       USA     UNITED STATES   USA                                                     1
040     040     04      1       1.0     26      5       24      -0.632135533183327      3       White House, District of Columbia, United States        US      U
SDC             38.8951 -77.0364        531871  0                                                       3       White House, District of Columbia, Unite
d States        US      USDC            38.8951 -77.0364        531871  20201130000000  https://www.gjsentinel.com/news/us/source-pa-lawmaker-gets-a-positive-te
st-at-trump-meeting/article_abea9f45-6066-5504-89e2-f29a95f1a8ba.html
```

Para responder esta pregunta se realizó la siguiente query:

```
sshuser@hn0-mdd202:~$ cat query3.pig
set tez.runtime.io.sort.mb 1091;
DATA = LOAD 'wasb://test-spark-mdd2021@mddclusterhdistorage.blob.core.windows.net/gdelt/MENTIONS/NOV2020/' using PigStorage ('\t')  AS
        (GlobalEventID:int,
        EventTimeDate:int,
        MentionTimeDate:int,
        MentionType:int,
        MentionSourceName:chararray,
        MentionIdentifier:chararray,
        SentenceID:int,
        Actor1CharOffset:int,
        Actor2CharOffset:int,
        ActionCharOffset:int,
        InRawText:int,
        Confidence:int,
        MentionDocLen:int,
        MentionDocTone:int,
        MentionDocTranslationInfo:chararray,
        Extras:chararray);
grp = group DATA by GlobalEventID;
FILE_GROUP_COUNT = FOREACH grp GENERATE group,COUNT(DATA);
ORDERBYCOUNT = order FILE_GROUP_COUNT by $1 desc;
TOPCOUNT = limit ORDERBYCOUNT 40;
dump TOPCOUNT;

sshuser@hn0-mdd202:~$
```

b. Utilizando Hive se llega a la misma conclusión con exactamente los mismos contadores:

```
INFO  : Completed executing command(
INFO  : OK
+----------------+-------+
| globaleventid  |  _c1  |
+----------------+-------+
| 957018602      | 1079  |
| 953950018      | 894   |
| 957018542      | 834   |
| 956034393      | 727   |
| 955735471      | 719   |
| 953924747      | 680   |
| 953924802      | 660   |
| 955732092      | 632   |
| 956614494      | 627   |
| 953830927      | 620   |
+----------------+-------+
10 rows selected (52.081 seconds)
0: jdbc:hive2://hn0-mdd202:10001/>
```

El evento 957018602 se repite 1079 veces, y si se busca en la carpeta exports arroja la misma salida que cuando se utilizó PIG

```
+----------------+-------+
| globaleventid  |  _c1  |
+----------------+-------+
| 957018602      | 1079  |
| 953950018      | 894   |
| 957018542      | 834   |
| 956034393      | 727   |
| 955735471      | 719   |
| 953924747      | 680   |
| 953924802      | 660   |
| 955732092      | 632   |
| 956614494      | 627   |
| 953830927      | 620   |
+----------------+-------+
10 rows selected (52.081 seconds)
0: jdbc:hive2://hn0-mdd202:10001/> Closing: 0: jdbc:hive2://hn0-mdd202:10001/;transportMode=http
sshuser@hn0-mdd202:~$ hadoop fs -cat 'wasb://test-spark-mdd2021@mddclusterhdistorage.blob.core.windows.net/gdelt/EXPORTS/NOV2020/*' |grep 957018602
957018602       20201130        202011  2020    2020.9041       USA     UNITED STATES   USA                                                                                                     1
040     040     04      1       1.0     26      5       24      -0.632135533183327      3       White House, District of Columbia, United States        US      USDC            38.8951
-77.0364        531871  0                                                               3       White House, District of Columbia, United States        US      USDC            38.8951
-77.0364        531871  20201130000000  https://www.gjsentinel.com/news/us/source-pa-lawmaker-gets-a-positive-test-at-trump-meeting/article_abea9f45-6066-5504-89e2-f29a95f1a8ba.html
sshuser@hn0-mdd202:~$
```

Sentencia HQL:

```
select globaleventid, count(*) from default.mentions7
where eventtimedate like '202011%' group by
globaleventid order by 2 desc limit 10;
```

4. ¿Cuáles son los 10 sitios web que generaron en promedio la mayor cantidad de menciones poco confiables (<40% de confianza) durante noviembre del 2020?

    a. Implemente la consulta usando Pig (preguntar si los archivos yafueron transformados a un formato legible por otro alumno) (6 puntos)

Los sitios con confiabilidad menor al 40% en noviembre 2020 son:

```
sshuser@hn0-mddspa:~$ cat query5b
(timesofisrael.com,23573)
(sputniknews.com,11480)
(breitbart.com,11336)
(palestinemonitor.org,10222)
(nbcnews.com,8455)
(voanews.com,8031)
(haaretz.com,7763)
(usatoday.com,6498)
(politico.com,6430)
(northwestgeorgianews.com,6271)
sshuser@hn0-mddspa:~$
```

El script utilizado para realizar esta tarea fue:

```
sshuser@hn0-mdd202:~$ cat query4.pig
set tez.runtime.io.sort.mb 1091;
DATA = LOAD 'wasb://test-spark-mdd2021@mddclusterhdistorage.blob.core.windows.net/gdelt/MENTIONS/NOV2020/' using PigStorage ('\t')  AS
        (GlobalEventID:int,
        EventTimeDate:int,
        MentionTimeDate:int,
        MentionType:int,
        MentionSourceName:chararray,
        MentionIdentifier:chararray,
        SentenceID:int,
        Actor1CharOffset:int,
        Actor2CharOffset:int,
        ActionCharOffset:int,
        InRawText:int,
        Confidence:int,
        MentionDocLen:int,
        MentionDocTone:int,
        MentionDocTranslationInfo:chararray,
        Extras:chararray);
grp = group DATA by MentionSourceName;
mean = foreach grp {
    sum = SUM(DATA.Confidence);
    count = COUNT(DATA);
    generate group as id, sum/count as mean;};
HIGHCONFID = filter mean by $1 < 40;
ORDERBYCOUNT = order HIGHCONFID by $1 desc;
TOPCOUNT = limit ORDERBYCOUNT 40;
dump TOPCOUNT;
```

b. Al parecer Hive es más exacto que PIG para realizar cálculos. En PIG la salida para el promedio fue un entero, sin embargo, en Hive muestra flotantes, lo que distorsiona la lista.

contador_sitios	mentionsourcename	_c2
23573	timesofisrael.com	35.97463199423069
11480	sputniknews.com	39.150696864111495
11336	breitbart.com	39.7635850388144
10222	palestinemonitor.org	30.845235765994914
8455	nbcnews.com	38.502661147250144
8031	voanews.com	39.76715228489603
7763	haaretz.com	38.383356949632876
6498	usatoday.com	38.72883964296707
6430	politico.com	38.10108864696734
6271	northwestgeorgianews.com	39.09743262637538

```
10 rows selected (154.469 seconds)
0: jdbc:hive2://hn0-mddspa:10001/>
```

```
select count(mentionsourcename) as contador_sitios,
mentionsourcename, avg(confidence) from default.mentions1
group by mentionsourcename having avg(confidence) < 40
order by 1 desc limit 10;
```

5. ¿Cuáles son los 10 sitios web que generaron en promedio la mayor cantidad de menciones muy confiables (>80% de confianza) durante septiembre, octubre, noviembre y diciembre del 2020? (promedio de todas las apariciones durante los 4 meses)
    1. Los 10 sitios con mayor cantidad de menciones con un nivel de confianza superior al 80% son:

```
sshuser@hn0-mddspa:~$ cat query5
(meed.com,3128)
(live.com,1923)
(business2community.com,1271)
(lewistownnews.com,1136)
(coolsocial.net,1025)
(alleywatch.com,860)
(androidheadlines.com,816)
(justjaredjr.com,759)
(energyfm.net,732)
(androidpolice.com,730)
sshuser@hn0-mddspa:~$
```

Aquí el script utilizado

```
sshuser@hn0-mdd202:~$ cat query5.pig
set tez.runtime.io.sort.mb 1091;
DATA = LOAD 'wasb://test-spark-mdd2021@mddclusterhdistorage.blob.core.windows.net/gdelt/MENTIONS/' using PigStorage ('\t')  AS
        (GlobalEventID:int,
        EventTimeDate:int,
        MentionTimeDate:int,
        MentionType:int,
        MentionSourceName:chararray,
        MentionIdentifier:chararray,
        SentenceID:int,
        Actor1CharOffset:int,
        Actor2CharOffset:int,
        ActionCharOffset:int,
        InRawText:int,
        Confidence:int,
        MentionDocLen:int,
        MentionDocTone:int,
        MentionDocTranslationInfo:chararray,
        Extras:chararray);
grp = group DATA by MentionSourceName;
mean = foreach grp {
    sum = SUM(DATA.Confidence);
    count = COUNT(DATA);
    generate group as id, sum/count as mean;};
HIGHCONFID = filter mean by $1 > 80;
ORDERBYCOUNT = order HIGHCONFID by $1 desc;
TOPCOUNT = limit ORDERBYCOUNT 40;
dump TOPCOUNT;
```

Implemente la consulta usando Hive (preguntar si la tabla yafue creada por otro alumno) (12 puntos).

- Mismo caso que el anterior, se detecta que Hive realiza cálculos de manera distinta con la función incorporada AVG. Se listan los sitios con confidence superior a 80%.

contador_sitios	mentionsourcename	_c2
3128	meed.com	88.49744245524296
1923	live.com	99.97919916796671
1271	business2community.com	84.02045633359559
1136	lewistownnews.com	89.87676056338029
1025	coolsocial.net	97.59024390243903
860	alleywatch.com	92.25581395348837
816	androidheadlines.com	80.08578431372548
759	justjaredjr.com	80.14492753623189
732	energyfm.net	96.93989071038251
730	androidpolice.com	82.3013698630137

```
10 rows selected (779.953 seconds)
0: jdbc:hive2://hn0-mddspa:10001/>
```

Query:

```
select count(mentionsourcename) as contador_sitios,
mentionsourcename, avg(confidence) from default.mentions2
group by mentionsourcename having avg(confidence) > 80
order by 1 desc limit 10;
```

6. Diseñe una consulta que le parezca relevante y que involucre cruzar los datasets export y mentions usando la llave GlobalEventID (primeros campos en ambos).