Máster en Big Data

Tecnologías de Almacenamiento

11. Hands-On: Hive

Presentado por: Jose David Angulo y Albert Ripoll

Índice

1. Introducción

El objetivo de este Hands-On es familiarizarse con la utilización de Hive, tanto en la creación de tablas como en la realización de consultas.

2. Entorno

Para este Handos On, utilizaremos la máquina virtual desplegada en Hands-On anteriores llamada Developer_Hadoop y todo será ejecutado vía Shell.

3. Creación de tablas

a) Ejecutar el Hive Shell

Para iniciar el Shell se hace con el comando hive

```
[training@localhost ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
Hive history file=/tmp/training/hive_job_log_training_202405081506_237650753.txt
hive>
```

b) Crear la tabla movie basada en el archivo “movie” importado anteriormente

```
CREATE TABLE movie
   (id INT, name STRING, year INT)
   ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ';'
   STORED AS TEXTFILE
```

```
hive> CREATE TABLE movie
    > (id INT, name STRING, year INT)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE;
OK
Time taken: 0.57 seconds
```

```
DESCRIBE movie;
```

```
hive> DESCRIBE movie;
OK
id      int
name    string
year    int
Time taken: 0.117 seconds
hive>
```

hadoop fs -rm -r /user/hadoop/movielens/movie/_logs

```
[training@localhost ~]$ hadoop fs -ls /user/hadoop/movielens/movie
Found 3 items
-rw-r--r--   1 training supergroup          0 2024-05-02 16:06 /user/hadoop/movi
elens/movie/_SUCCESS
drwxrwxrwx   - training supergroup          0 2024-05-02 16:06 /user/hadoop/movi
elens/movie/_logs
-rw-r--r--   1 training supergroup     102052 2024-05-02 16:06 /user/hadoop/movi
elens/movie/part-m-00000
[training@localhost ~]$ hadoop fs -rm -r /user/hadoop/movielens/movie/_logs
Deleted /user/hadoop/movielens/movie/_logs
[training@localhost ~]$ hadoop fs -ls /user/hadoop/movielens/movie
```

LOAD DATA INPATH "/user/hadoop/movielens/movie1/part-m-*"

> INTO TABLE movie;

```
hive> LOAD DATA INPATH "/user/hadoop/movielens/movie1/part-m-*"
    > INTO TABLE movie;
Loading data to table default.movie
OK
Time taken: 0.115 seconds
```

c) Crear la tabla movierating basada en el archivo “movierating” importado anteriormente

```
hive> CREATE EXTERNAL TABLE movierating (userid INT, movieid INT, rating INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY "\;" LOCATION "/user/training/movierating";
OK
Time taken: 0.054 seconds
hive> SELECT * from  movierating LIMIT 5;
OK
1       1193    5
1       661     3
1       914     3
1       3408    4
1       2355    5
```

d) Listar todas las tablas de Hive

```
hive> SHOW TABLES;
OK
customers
movie
movie2
movie3
movierating
order_details
orders
products
```

e) Ver la metainformación de las tablas movie y movierating

```
hive> DESCRIBE movie3;
OK
id      int
name    string
year    smallint
```

```
hive> DESCRIBE movierating;
OK
userid  int
movieid int
rating  int
```

4. Consultas con Hive

f) Listar todas las películas lanzadas antes de 1930

```
hive> SELECT * FROM movie3
    > WHERE year < 1930;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_202404221529_0049, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202404221529_0049
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_202404221529_0049
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-05-07 03:17:14,122 Stage-1 map = 0%,  reduce = 0%
2024-05-07 03:17:28,298 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.9 sec
2024-05-07 03:17:29,310 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.9 sec
2024-05-07 03:17:30,319 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 0.9 sec
MapReduce Total cumulative CPU time: 900 msec
Ended Job = job_202404221529_0049
MapReduce Jobs Launched:
Job 0: Map: 1   Cumulative CPU: 0.9 sec   HDFS Read: 102389 HDFS Write: 7181 SUCCESS
Total MapReduce CPU Time Spent: 900 msec
OK
30      Shanghai Triad  0
47      Seven   0
58      Postino, Il     0
59      Confessional, The       0
68      French Twist    0
80      White Balloon, The      0
```

g) Listar todas las películas lanzadas antes de 1930. Descarta todas aquellas que no tengan año conocido (year=0) y ordenalas por nombre

```
hive> SELECT * FROM movie3
    > WHERE year < 1930 AND year <> 0
    > ORDER BY name ASC;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202404221529_0050, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202404221529
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_202404221529_0050
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-05-07 03:31:53,790 Stage-1 map = 0%,  reduce = 0%
2024-05-07 03:32:01,827 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.14 sec
2024-05-07 03:32:02,836 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.14 sec
2024-05-07 03:32:03,844 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.14 sec
2024-05-07 03:32:04,851 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.14 sec
2024-05-07 03:32:05,865 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.44 sec
2024-05-07 03:32:06,872 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.44 sec
2024-05-07 03:32:07,879 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.44 sec
2024-05-07 03:32:08,891 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.44 sec
MapReduce Total cumulative CPU time: 2 seconds 440 msec
Ended Job = job_202404221529_0050
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 2.44 sec   HDFS Read: 102389 HDFS Write: 796 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 440 msec
OK
2230    Always Tell Your Wife   1923
3012    Battling Butler 1926
3517    Bells, The      1926
2221    Blackmail       1929
1926    Broadway Melody, The    1929
2222    Champagne       1928
3306    Circus, The     1928
2777    Cobra   1925
3132    Daddy Long Legs 1919
3309    Dog's Life, A   1920
2224    Downhill        1927
2225    Easy Virtue     1927
2223    Farmer's Wife, The      1928
```

h) Selecciona todas las películas valoradas por el usuario con id = 149 (Muestra solamente los campos relativos al id de película y su valoración)

```
hive> SELECT movieid, rating FROM movierating
    > WHERE userid = 149;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_202404221529_0051, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202404221529_0051
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_202404221529_0051
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2024-05-07 03:43:08,734 Stage-1 map = 0%,  reduce = 0%
2024-05-07 03:43:17,790 Stage-1 map = 24%,  reduce = 0%
2024-05-07 03:43:20,807 Stage-1 map = 48%,  reduce = 0%
2024-05-07 03:43:23,822 Stage-1 map = 76%,  reduce = 0%
2024-05-07 03:43:28,871 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 15.8 sec
2024-05-07 03:43:29,888 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 15.8 sec
2024-05-07 03:43:30,976 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 15.8 sec
2024-05-07 03:43:32,070 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 15.8 sec
2024-05-07 03:43:33,130 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 15.8 sec
2024-05-07 03:43:34,170 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 15.8 sec
2024-05-07 03:43:35,228 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 15.8 sec
2024-05-07 03:43:36,282 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 15.8 sec
```

```
MapReduce Total cumulative CPU time: 15 seconds 800 msec
Ended Job = job_202404221529_0051
MapReduce Jobs Launched:
Job 0: Map: 1   Cumulative CPU: 15.8 sec   HDFS Read: 11553769 HDFS Write: 3966 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 800 msec
OK
1249    4
1177    4
1179    4
647     4
648     4
1321    4
3863    2
1250    4
3865    2
2990    4
2126    3
3793    2
2991    4
1252    4
3794    2
720     4
2993    4
```

i) Utiliza información de las dos tablas, por ejemplo, incluyendo el nombre de la película en la lista generada en el apartado anterior

```
hive> SELECT m.id, m.name, m.year , mr.rating
    > FROM movie3 m
    > LEFT OUTER JOIN movierating mr ON m.id =mr.movieid
    > WHERE mr.userid=149;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202404221529_0058, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202404221529_0058
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_202404221529_0058
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2024-05-08 13:47:56,724 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 40.27 sec
MapReduce Total cumulative CPU time: 40 seconds 270 msec
Ended Job = job_202404221529_0058
MapReduce Jobs Launched:
Job 0: Map: 2  Reduce: 1   Cumulative CPU: 40.27 sec   HDFS Read: 11656158 HDFS Write: 16597 SUCCESS
Total MapReduce CPU Time Spent: 40 seconds 270 msec
OK
1       Toy Story       1995    3
2       Jumanji 1995    3
6       Heat    1995    4
9       Sudden Death    1995    3
10      GoldenEye       1995    4
16      Casino  1995    4
17      Sense and Sensibility   1995    4
21      Get Shorty      1995    5
```

j) Calcula el promedio con el que el usuario 149 califica las pel·lícules

```
hive> SELECT m.id,m.name, AVG(mr.rating) AS promedio_calificacioXmovie
    > FROM movie3 m
    > LEFT OUTER JOIN movierating mr ON m.id=mr.movieid
    > WHERE mr.userid=149
    > GROUP BY m.id, m.name;
Total MapReduce jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202404221529_0059, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202404221529_0059
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_202404221529_0059
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
```

```
2024-05-08 14:04:10,578 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 41.5 sec
2024-05-08 14:04:11,583 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 41.5 sec
2024-05-08 14:04:12,590 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 41.5 sec
MapReduce Total cumulative CPU time: 41 seconds 500 msec
Ended Job = job_202404221529_0059
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202404221529_0060, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202404221529_0060
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_202404221529_0060
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2024-05-08 14:04:16,931 Stage-2 map = 0%,  reduce = 0%
2024-05-08 14:04:20,947 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.82 sec
2024-05-08 14:04:21,954 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.82 sec
2024-05-08 14:04:22,961 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.82 sec
2024-05-08 14:04:23,967 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.82 sec
2024-05-08 14:04:24,973 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.82 sec
2024-05-08 14:04:25,978 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.82 sec
2024-05-08 14:04:26,984 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.44 sec
2024-05-08 14:04:27,993 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.44 sec
2024-05-08 14:04:28,998 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.44 sec
2024-05-08 14:04:30,002 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.44 sec
2024-05-08 14:04:31,008 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.44 sec
MapReduce Total cumulative CPU time: 2 seconds 440 msec
```

```
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 2.44 sec   HDFS Read: 30526 HDFS Write: 14911 SUCCESS
Total MapReduce CPU Time Spent: 43 seconds 940 msec
OK
1       Toy Story       3.0
2       Jumanji 3.0
6       Heat    4.0
9       Sudden Death    3.0
10      GoldenEye       4.0
16      Casino  4.0
17      Sense and Sensibility   4.0
21      Get Shorty      5.0
22      Copycat 2.0
24      Powder  4.0
25      Leaving Las Vegas       2.0
29      City of Lost Children, The      4.0
```

k) Lista cada usuario que ha valorado películas, el número de películas que ha valorado y el promedio de valoración que ha proporcionado

```
hive> SELECT mr.userid, COUNT(mr.movieid),AVG(mr.rating)
    > FROM movierating mr
    > GROUP BY mr.userid;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202404221529_0061, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202404221529_0061
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_202404221529_0061
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-05-08 15:10:35,967 Stage-1 map = 0%,  reduce = 0%
2024-05-08 15:10:44,022 Stage-1 map = 24%,  reduce = 0%
2024-05-08 15:10:50,058 Stage-1 map = 48%,  reduce = 0%
2024-05-08 15:10:56,090 Stage-1 map = 76%,  reduce = 0%
2024-05-08 15:11:10,229 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 24.88 sec
MapReduce Total cumulative CPU time: 24 seconds 880 msec
Ended Job = job_202404221529_0061
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 24.88 sec   HDFS Read: 11553769 HDFS Write: 150900 SUCCESS
Total MapReduce CPU Time Spent: 24 seconds 880 msec
OK
1       53      4.188679245283019
2       129     3.7131782945736433
3       51      3.9019607843137254
4       21      4.190476190476191
5       198     3.1464646464646466
6       71      3.9014084507042255
7       31      4.32258064516129
8       139     3.884892086330935
9       106     3.7358490566037736
10      401     4.114713216957606
11      137     3.2773722627737225
12      23      3.8260869565217392
13      108     3.388888888888889
14      25      3.32
```

l) Inserta toda la información del apartado anterior en una nueva tabla llamada “userrating”.

```
hive> INSERT INTO TABLE userrating
    > SELECT mr.userid, COUNT(*), AVG(mr.rating)
    > FROM movierating mr
    > GROUP BY mr.userid;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202404221529_0062, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_202404221529_0062
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_202404221529_0062
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-05-08 15:37:37,659 Stage-1 map = 0%,  reduce = 0%
2024-05-08 15:37:49,741 Stage-1 map = 24%,  reduce = 0%
2024-05-08 15:37:55,771 Stage-1 map = 48%,  reduce = 0%
2024-05-08 15:38:01,801 Stage-1 map = 76%,  reduce = 0%
2024-05-08 15:38:05,830 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.8 sec
2024-05-08 15:38:06,845 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.8 sec
2024-05-08 15:38:07,859 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.8 sec
2024-05-08 15:38:08,866 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.8 sec
2024-05-08 15:38:09,880 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.8 sec
2024-05-08 15:38:10,889 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.8 sec
2024-05-08 15:38:11,900 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 24.8 sec
2024-05-08 15:38:12,916 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 27.84 sec
2024-05-08 15:38:13,929 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 27.84 sec
2024-05-08 15:38:14,940 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 27.84 sec
2024-05-08 15:38:15,949 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 27.84 sec
2024-05-08 15:38:16,956 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 27.84 sec
```

```
2024-05-08 15:38:16,956 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 27.84 sec
MapReduce Total cumulative CPU time: 27 seconds 840 msec
Ended Job = job_202404221529_0062
Loading data to table default.userrating
6040 Rows loaded to userrating
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 27.84 sec   HDFS Read: 11553769 HDFS Write: 62279 SUCCESS
Total MapReduce CPU Time Spent: 27 seconds 840 msec
OK
Time taken: 45.994 seconds
hive> SELECT * FROM userrating LIMIT 20;
OK
1       53      4
2       129     3
3       51      3
4       21      4
5       198     3
```

m) Exporta la tabla a MySQL con Sqoop.

```
sqoop export \
--connect jdbc:mysql://localhost:3306/Albert \
--username training \
--password training \
--table mobie_table \
--export-dir /user/hive/movielens/movie_table \
--input-fields-terminated-by '\t' \
-m 1
```