Máster en Big Data

Tecnologías de Almacenamiento

11. Hands-On: Hive

Àlex Balló Vergés

2019

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 Desarrollo_Hadoop y todo será ejecutado vía shell

3. Creación de tablas

a) Ejecutar el Hive Shell

   hive

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

```
hive> CREATE  EXTERNAL TABLE IF NOT EXISTS tb_movie
    > (
    > id int,
    > name string,
    > anio int
    > )
    > row format delimited
    > fields terminated by ','
    > LOCATION '/user/training/movie';
OK
Time taken: 0.361 seconds
```

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

```
hive> CREATE  EXTERNAL TABLE IF NOT EXISTS tb_movieratings
    > (
    > userid int,
    > movieid int,
    > rating tinyint
    > )
    > row format delimited
    > fields terminated by ','
    > LOCATION '/user/training/movierating';
OK
Time taken: 0.116 seconds
```

d) Listar todas las tablas de Hive

```
hive> show tables;
OK
customers
order_details
orders
products
tb_movie
tb_movieratings
Time taken: 0.136 seconds
```

e) Ver la metainformación de las tablas movie y movierating

```
hive> describe tb_movie;
OK
id      int
name    string
anio    int
Time taken: 0.204 seconds
hive> describe tb_movieratings;
OK
userid  int
movieid int
rating  tinyint
Time taken: 0.221 seconds
```

4. Consultas con Hive

f) Listar todas las películas lanzadas antes de 1930

```
hive> select * from tb_movie where anio<'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_201905151032_0001, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=jo
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_201905151032_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2019-05-15 14:43:47,836 Stage-1 map = 0%,  reduce = 0%
2019-05-15 14:43:53,957 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.45 sec
2019-05-15 14:43:54,984 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.45 sec
2019-05-15 14:43:56,001 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.45 sec
2019-05-15 14:43:57,021 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.45 sec
2019-05-15 14:43:58,047 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.45 sec
2019-05-15 14:43:59,063 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.45 sec
2019-05-15 14:44:00,083 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.45 sec
MapReduce Total cumulative CPU time: 1 seconds 450 msec
Ended Job = job_201905151032_0001
MapReduce Jobs Launched:
Job 0: Map: 1   Cumulative CPU: 1.45 sec   HDFS Read: 102389 HDFS Write: 3981 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 450 msec
OK
30      Shanghai Triad  0
47      Seven   0
68      French Twist    0
82      Antonia's Line  0
97      Hate    0
106     Nobody Loves Me 0
```

training@localhost:~

g) Descarta todas aquellas que no tengan año conocido (year=0) y ordenalas por nombre

```
hive> insert overwrite table tb_movie select*from tb_movie where anio<>'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_201905151032_0002, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=jo
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_201905151032_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-05-15 14:50:39,811 Stage-1 map = 0%,  reduce = 0%
2019-05-15 14:50:47,871 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.18 sec
2019-05-15 14:50:48,893 Stage-1 map = 100%,  training@localhost:~ ive CPU 2.18 sec
```

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 tb_movieratings 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_201905151032_0004, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=j
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_201905151032_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2019-05-15 14:57:35,551 Stage-1 map = 0%,  reduce = 0%
2019-05-15 14:57:43,601 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.21 sec
2019-05-15 14:57:44,617 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.21 sec
2019-05-15 14:57:45,629 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.21 sec
2019-05-15 14:57:46,643 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.21 sec
2019-05-15 14:57:47,663 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.21 sec
2019-05-15 14:57:48,673 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.21 sec
MapReduce Total cumulative CPU time: 2 seconds 210 msec
Ended Job = job_201905151032_0004
MapReduce Jobs Launched:
Job 0: Map: 1   Cumulative CPU: 2.21 sec   HDFS Read: 11553769 HDFS Write: 3966 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 210 msec
OK
1249    4
1177    4                                                    training@localhost:~
```

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

```
select movieid,rating,name from tb_movieratings join tb_movie on tb_movie.id= tb_movieratings.mov
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_201905151032_0005, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=jo
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_201905151032_0005
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2019-05-15 15:03:47,355 Stage-1 map = 0%,  reduce = 0%
2019-05-15 15:04:04,625 Stage-1 map = 50%,  reduce = 0%
2019-05-15 15:04:07,671 Stage-1 map = 74%,  reduce = 0%, Cumulative CPU 3.03 sec
2019-05-15 15:04:08,737 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.17 sec
2019-05-15 15:04:09,749 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.17 sec
2019-05-15 15:04:10,762 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.17 sec
2019-05-15 15:04:11,775 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.17 sec
2019-05-15 15:04:12,790 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.17 sec
2019-05-15 15:04:13,804 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.17 sec
2019-05-15 15:04:14,823 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.17 sec
2019-05-15 15:04:15,844 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.17 sec
```

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

```
hive> select avg (rating) from tb_movieratings where userid='149';
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_201905151032_0006, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=jo
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_201905151032_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-05-15 15:07:55,235 Stage-1 map = 0%,  reduce = 0%
2019-05-15 15:08:04,297 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.89 sec
2019-05-15 15:08:05,310 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.89 sec
2019-05-15 15:08:06,323 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.89 sec
2019-05-15 15:08:07,337 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.89 sec
2019-05-15 15:08:08,357 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.89 sec
2019-05-15 15:08:09,369 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.89 sec
2019-05-15 15:08:10,385 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.89 sec
2019-05-15 15:08:11,401 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.89 sec
2019-05-15 15:08:12,416 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.58 sec
2019-05-15 15:08:13,440 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.58 sec
2019-05-15 15:08:14,455 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.58 sec
2019-05-15 15:08:15,471 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.58 sec
2019-05-15 15:08:16,480 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.58 sec
```

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

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