I. Preparing dataset.

Step 1:

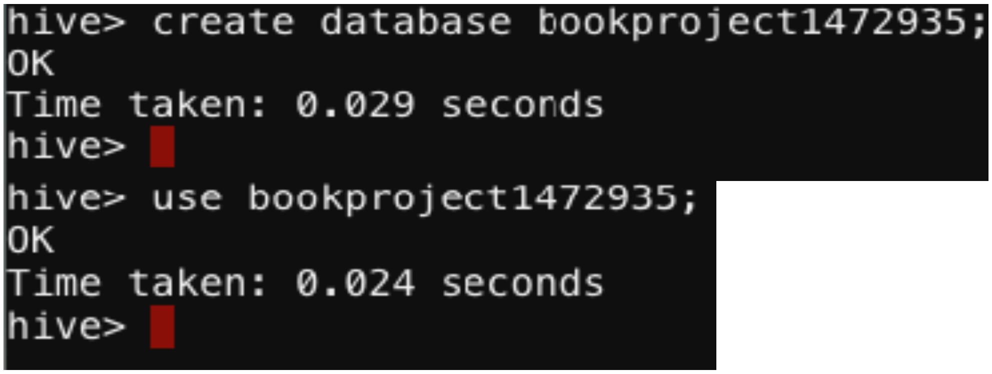
Create a new database bookproject1472935 and use it.

Command:

```
Create database bookproject1472935;

Use bookproject1472935;
```

Screenshot:



```
hive> create database bookproject1472935;
OK
Time taken: 0.029 seconds
hive> 
hive> use bookproject1472935;
OK
Time taken: 0.024 seconds
hive> 
```

Step 2:

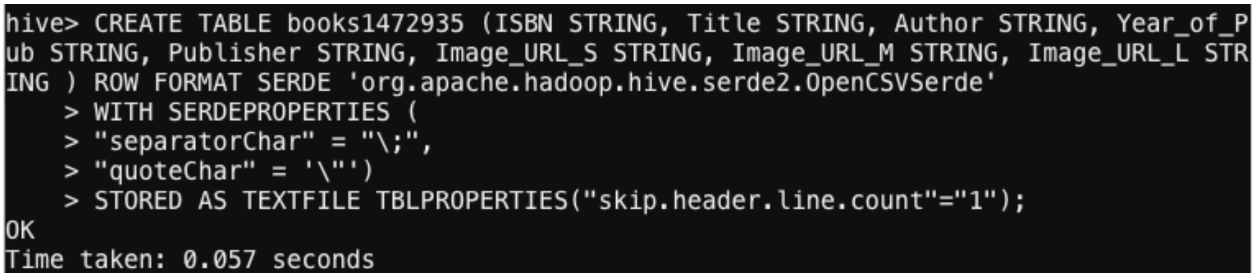
Creating a table and loading to the database just created. (BX-Books.csv)

Command:

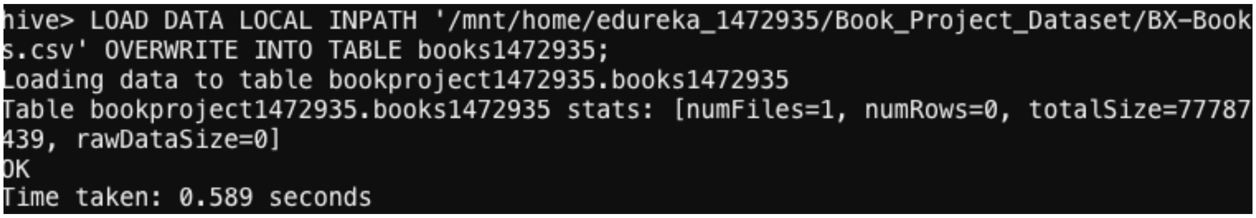
```
CREATE TABLE books1472935 (ISBN STRING, Title STRING, Author STRING,
Year_of_Pub STRING, Publisher STRING, Image_URL_S STRING, Image_URL_M STRING,
Image_URL_L STRING ) ROW FORMAT SERDE
'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
"separatorChar" = "\;",
"quoteChar" = '\"')
STORED AS TEXTFILE TBLPROPERTIES("skip.header.line.count"="1");

LOAD DATA LOCAL INPATH '/mnt/home/edureka_1472935/Book_Project_Dataset/BX-Books.csv' OVERWRITE INTO TABLE books1472935;
```

Screenshot:



```
hive> CREATE TABLE books1472935 (ISBN STRING, Title STRING, Author STRING, Year_of_P
ub STRING, Publisher STRING, Image_URL_S STRING, Image_URL_M STRING, Image_URL_L STR
ING ) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    > "separatorChar" = "\;",
    > "quoteChar" = '\"')
    > STORED AS TEXTFILE TBLPROPERTIES("skip.header.line.count"="1");
OK
Time taken: 0.057 seconds
```



```
hive> LOAD DATA LOCAL INPATH '/mnt/home/edureka_1472935/Book_Project_Dataset/BX-Book
s.csv' OVERWRITE INTO TABLE books1472935;
Loading data to table bookproject1472935.books1472935
Table bookproject1472935.books1472935 stats: [numFiles=1, numRows=0, totalSize=77787
439, rawDataSize=0]
OK
Time taken: 0.589 seconds
```

Step 3:

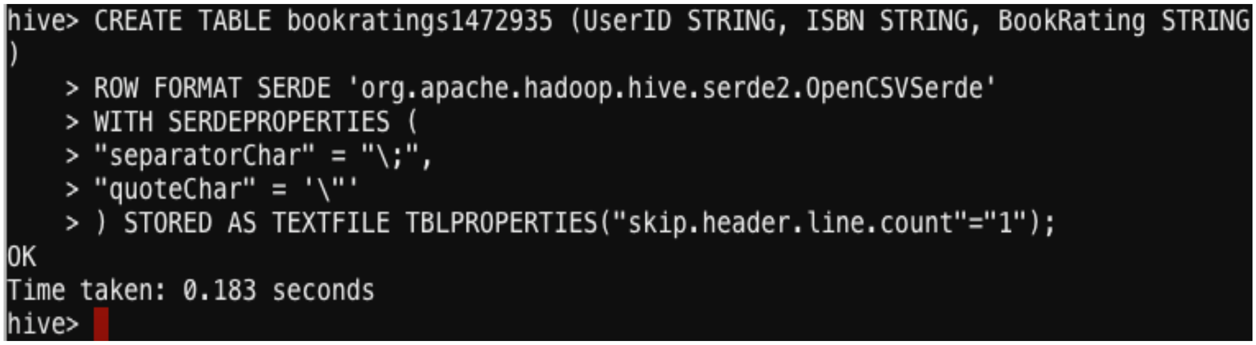
Creating a table and loading to the database just created. (BX-Book-Ratings.csv)

Command:

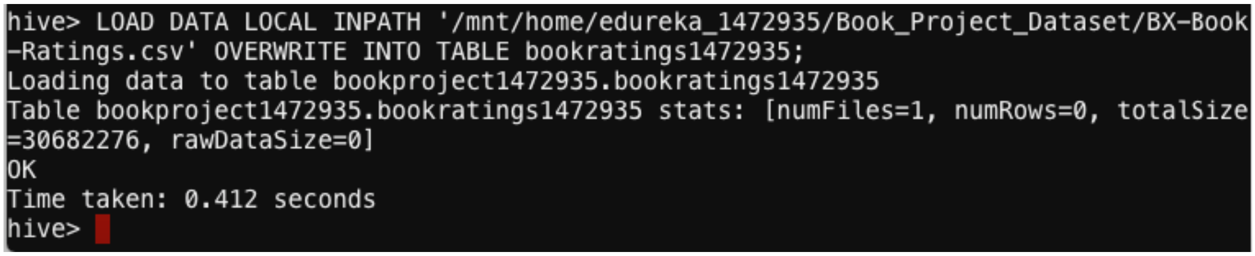
CREATE TABLE bookratings1472935 (UserID STRING, ISBN STRING, BookRating STRING)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
"separatorChar" = "\;",
"quoteChar" = '\"'
) STORED AS TEXTFILE TBLPROPERTIES("skip.header.line.count"="1");

LOAD DATA LOCAL INPATH '/mnt/home/edureka_1472935/Book_Project_Dataset/BX-Book-Ratings.csv' OVERWRITE INTO TABLE bookratings1472935;

Screenshots:



```
hive> CREATE TABLE bookratings1472935 (UserID STRING, ISBN STRING, BookRating STRING
)
    > ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
    > WITH SERDEPROPERTIES (
    > "separatorChar" = "\;",
    > "quoteChar" = '\"'
    > ) STORED AS TEXTFILE TBLPROPERTIES("skip.header.line.count"="1");
OK
Time taken: 0.183 seconds
hive>
```



```
hive> LOAD DATA LOCAL INPATH '/mnt/home/edureka_1472935/Book_Project_Dataset/BX-Book
-Ratings.csv' OVERWRITE INTO TABLE bookratings1472935;
Loading data to table bookproject1472935.bookratings1472935
Table bookproject1472935.bookratings1472935 stats: [numFiles=1, numRows=0, totalSize
=30682276, rawDataSize=0]
OK
Time taken: 0.412 seconds
hive> 
```

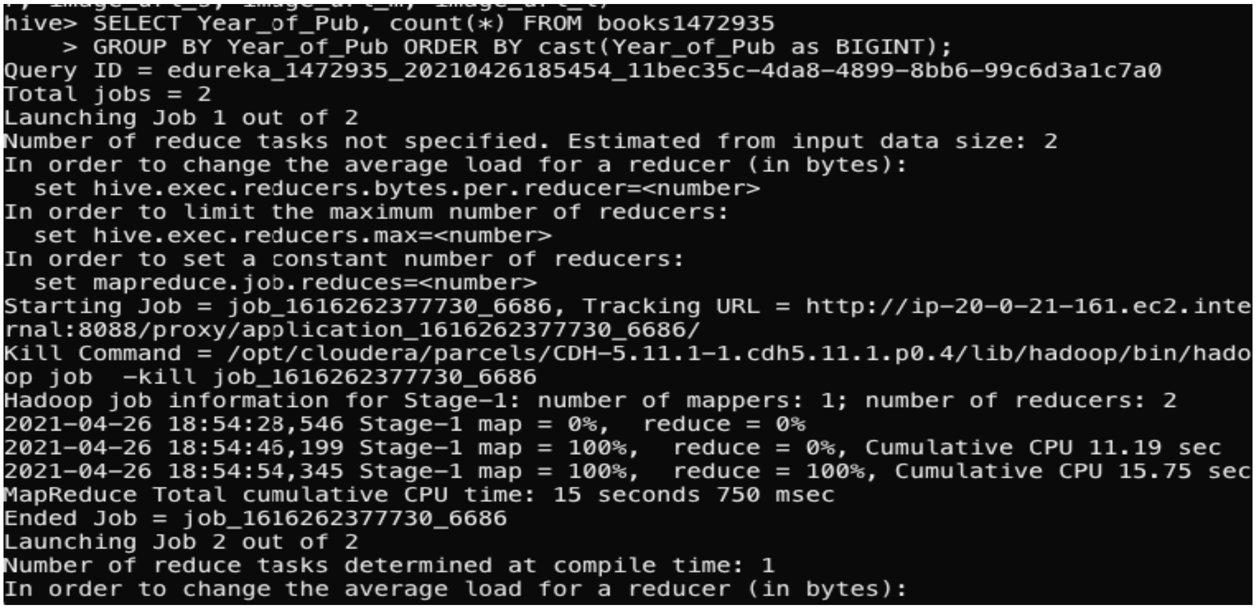
Solutions for the problem statements:

A. Find out the frequency of books published each year. (Hint: Use Boooks.csv file for this)

Command:

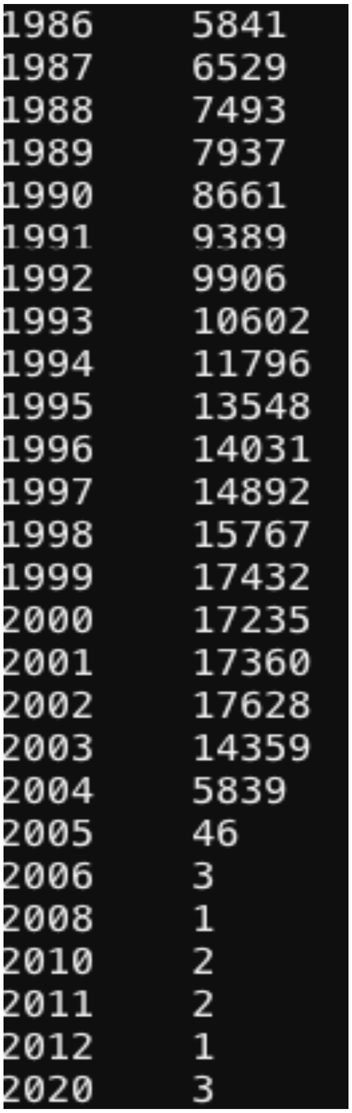
SELECT Year_of_Pub, count(*) FROM books1472935
GROUP BY Year_of_Pub ORDER BY cast(Year_of_Pub as BIGINT);

Screenshot:



```
hive> SELECT Year_of_Pub, count(*) FROM books1472935
    > GROUP BY Year_of_Pub ORDER BY cast(Year_of_Pub as BIGINT);
Query ID = edureka_1472935_20210426185454_11bec35c-4da8-4899-8bb6-99c6d3a1c7a0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1616262377730_6686, Tracking URL = http://ip-20-0-21-161.ec2.inte
rnal:8088/proxy/application_1616262377730_6686/
Kill Command = /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/bin/hado
op job  -kill job_1616262377730_6686
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 2
2021-04-26 18:54:28,546 Stage-1 map = 0%,  reduce = 0%
2021-04-26 18:54:46,199 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 11.19 sec
2021-04-26 18:54:54,345 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 15.75 sec
MapReduce Total cumulative CPU time: 15 seconds 750 msec
Ended Job = job_1616262377730_6686
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
```

Results:



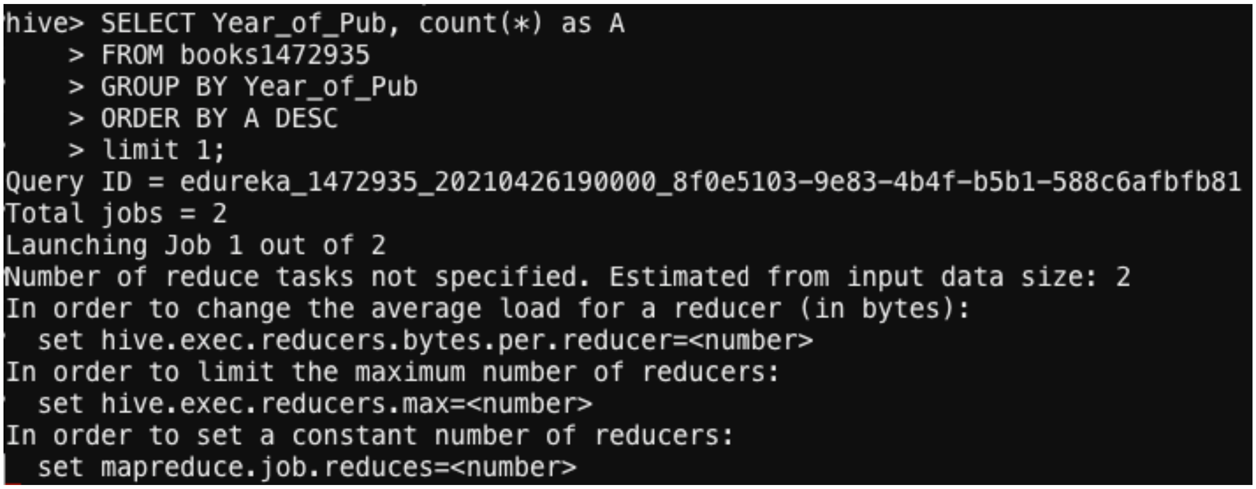
```
1986    5841
1987    6529
1988    7493
1989    7937
1990    8661
1991    9389
1992    9906
1993    10602
1994    11796
1995    13548
1996    14031
1997    14892
1998    15767
1999    17432
2000    17235
2001    17360
2002    17628
2003    14359
2004    5839
2005    46
2006    3
2008    1
2010    2
2011    2
2012    1
2020    3
```

B. Find out in which year maximum number of books were published

Command:

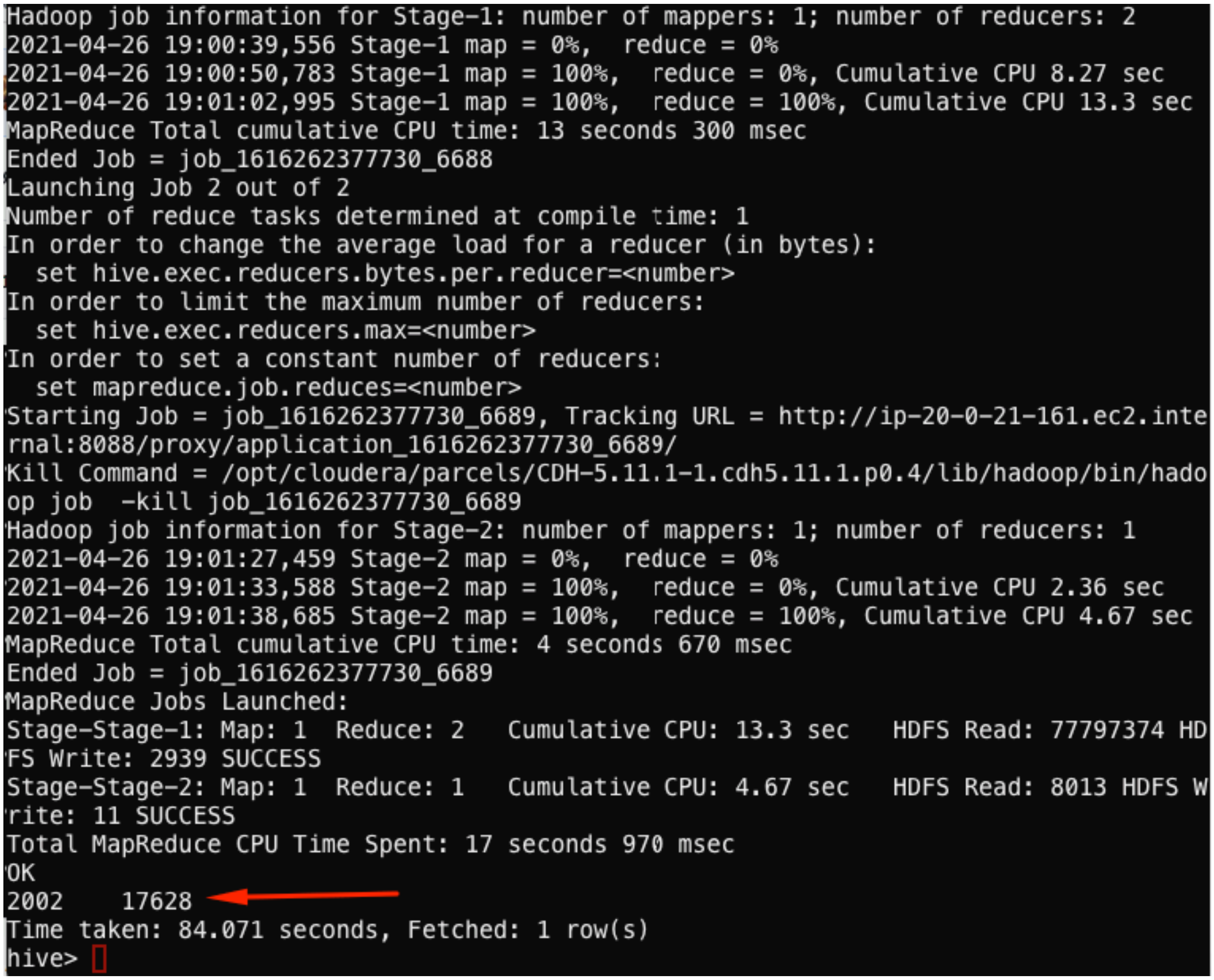
```
SELECT Year_of_Pub, count(*) as A
FROM books1472935
GROUP BY Year_of_Pub
ORDER BY A DESC
limit 1;
```

Screenshot:



```
hive> SELECT Year_of_Pub, count(*) as A
    > FROM books1472935
    > GROUP BY Year_of_Pub
    > ORDER BY A DESC
    > limit 1;
Query ID = edureka_1472935_20210426190000_8f0e5103-9e83-4b4f-b5b1-588c6afbfb81
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
```

Results:



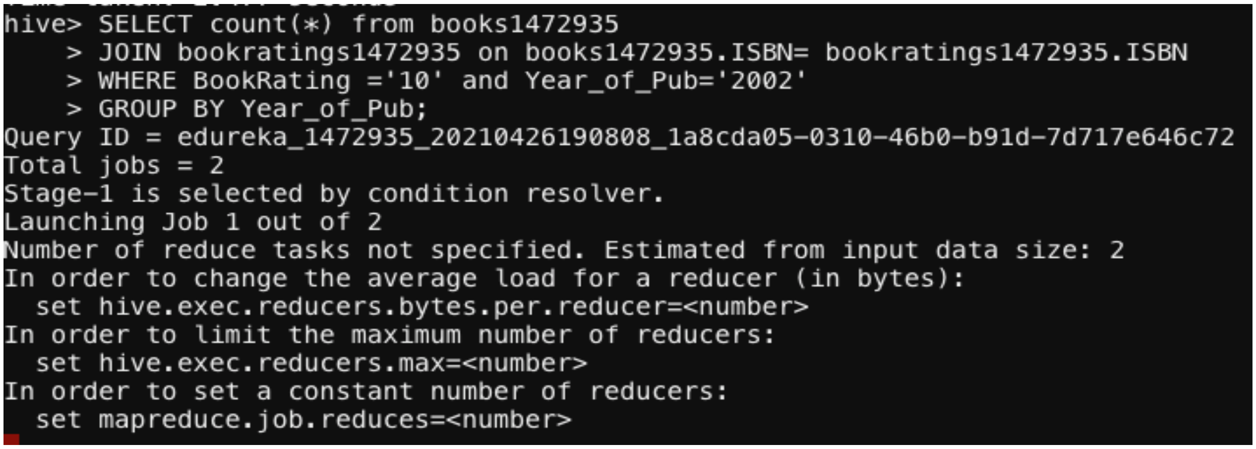
```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 2
2021-04-26 19:00:39,556 Stage-1 map = 0%,  reduce = 0%
2021-04-26 19:00:50,783 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.27 sec
2021-04-26 19:01:02,995 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 13.3 sec
MapReduce Total cumulative CPU time: 13 seconds 300 msec
Ended Job = job_1616262377730_6688
Launching Job 2 out of 2
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 mapreduce.job.reduces=<number>
Starting Job = job_1616262377730_6689, Tracking URL = http://ip-20-0-21-161.ec2.inte
rnal:8088/proxy/application_1616262377730_6689/
Kill Command = /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/bin/hado
op job  -kill job_1616262377730_6689
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-04-26 19:01:27,459 Stage-2 map = 0%,  reduce = 0%
2021-04-26 19:01:33,588 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.36 sec
2021-04-26 19:01:38,685 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.67 sec
MapReduce Total cumulative CPU time: 4 seconds 670 msec
Ended Job = job_1616262377730_6689
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 2   Cumulative CPU: 13.3 sec   HDFS Read: 77797374 HD
FS Write: 2939 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.67 sec   HDFS Read: 8013 HDFS W
rite: 11 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 970 msec
OK
2002    17628
Time taken: 84.071 seconds, Fetched: 1 row(s)
hive>
```

c. **Find out how many books were published based on ranking in the year 2002. (Hint: Use Book.csv and Book-Ratings.csv)**

Command:

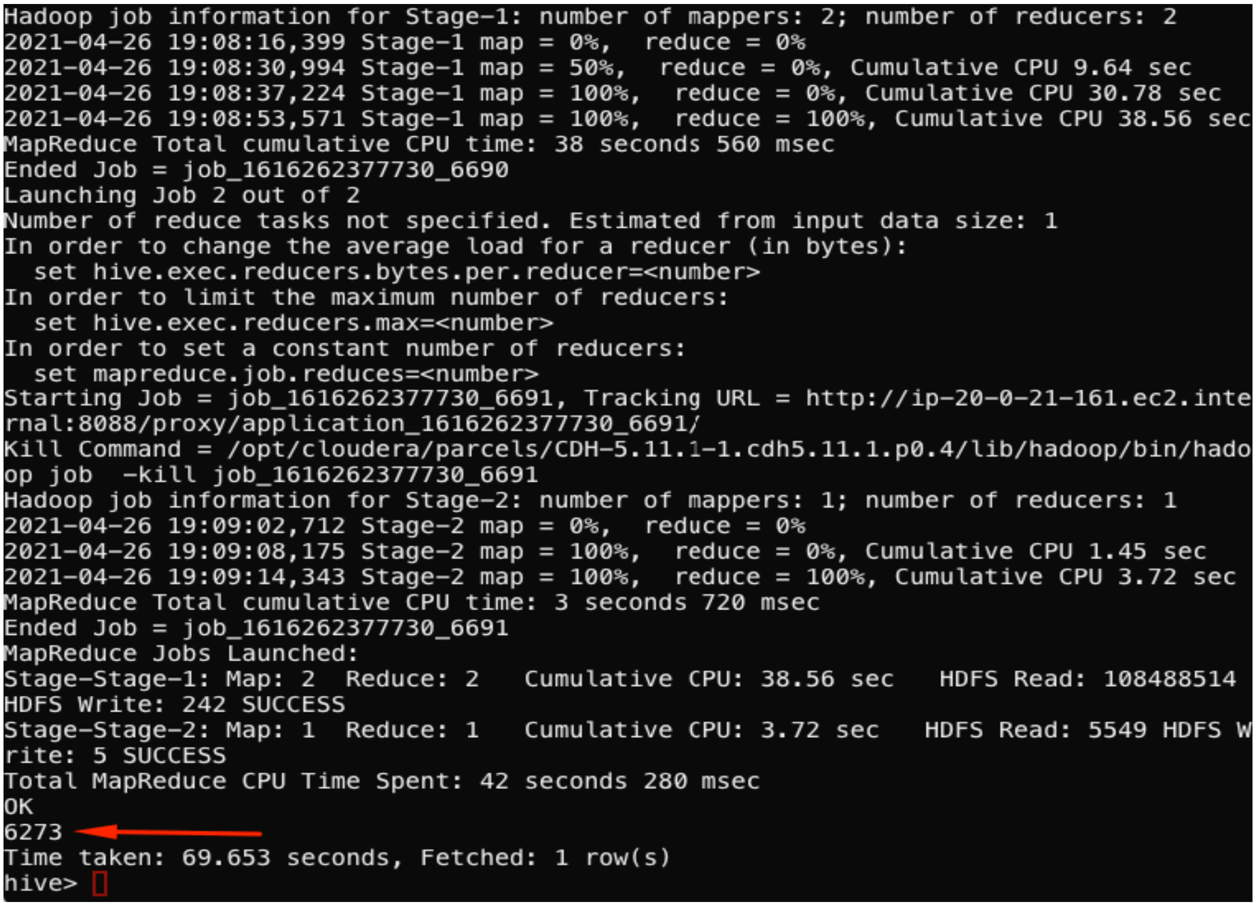
SELECT count(*) from books1472935
JOIN bookratings1472935 on books1472935.ISBN= bookratings1472935.ISBN
WHERE BookRating ='10' and Year_of_Pub='2002'
GROUP BY Year_of_Pub;

Screenshots:



```
hive> SELECT count(*) from books1472935
    > JOIN bookratings1472935 on books1472935.ISBN= bookratings1472935.ISBN
    > WHERE BookRating ='10' and Year_of_Pub='2002'
    > GROUP BY Year_of_Pub;
Query ID = edureka_1472935_20210426190808_1a8cda05-0310-46b0-b91d-7d717e646c72
Total jobs = 2
Stage-1 is selected by condition resolver.
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
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 mapreduce.job.reduces=<number>
```

Results:



```
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2021-04-26 19:08:16,399 Stage-1 map = 0%,  reduce = 0%
2021-04-26 19:08:30,994 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 9.64 sec
2021-04-26 19:08:37,224 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 30.78 sec
2021-04-26 19:08:53,571 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 38.56 sec
MapReduce Total cumulative CPU time: 38 seconds 560 msec
Ended Job = job_1616262377730_6690
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 mapreduce.job.reduces=<number>
Starting Job = job_1616262377730_6691, Tracking URL = http://ip-20-0-21-161.ec2.inte
rnal:8088/proxy/application_1616262377730_6691/
Kill Command = /opt/cloudera/parcels/CDH-5.11.1-1.cdh5.11.1.p0.4/lib/hadoop/bin/hado
op job  -kill job_1616262377730_6691
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2021-04-26 19:09:02,712 Stage-2 map = 0%,  reduce = 0%
2021-04-26 19:09:08,175 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.45 sec
2021-04-26 19:09:14,343 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.72 sec
MapReduce Total cumulative CPU time: 3 seconds 720 msec
Ended Job = job_1616262377730_6691
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 2   Cumulative CPU: 38.56 sec   HDFS Read: 108488514
HDFS Write: 242 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.72 sec   HDFS Read: 5549 HDFS W
rite: 5 SUCCESS
Total MapReduce CPU Time Spent: 42 seconds 280 msec
OK
6273
Time taken: 69.653 seconds, Fetched: 1 row(s)
hive>
```