Solution Extraction Using HIVE

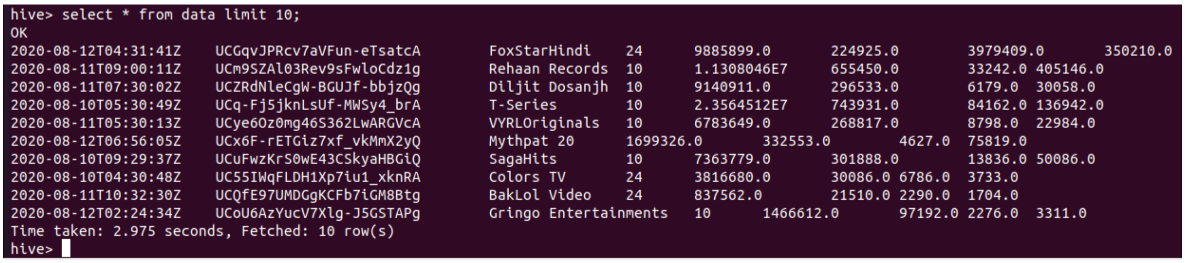
Use the following command to 'Create a Table' in HIVE

```
Hive> create table data (int,month int,day int,order1
int,country int,session_ID double,page_1 int,page_2
string,colour int,location int,model_photography
int,price double,price_2 int,page int
)
ROW FORMAT DELIMITED FIELDS
TERMINATED BY '\t'
tblproperties("skip.header.line.count"="1");
```

This command will create a Hive table named 'YouTube_data_table' in which rows will be delimited and rows fields will be terminated by commas**.**

Selecting the tables

This hive query will select randomly a ten records from the table "data" using the function limit.

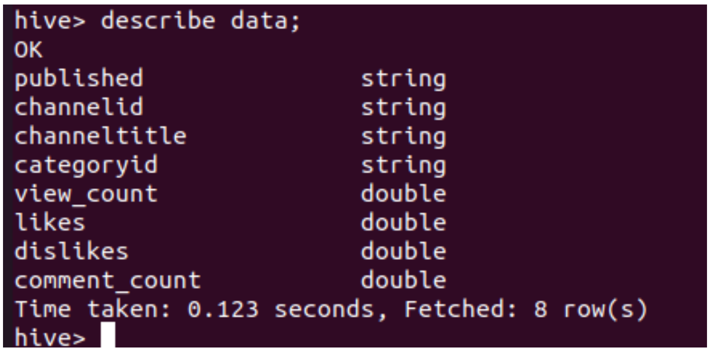


```
hive> select * from data limit 10;
OK
2020-08-12T04:31:41Z          UCGqvJPRcv7aVFun-eTsatcA                FoxStarHindi        24        9885899.0           224925.0            3979409.0           350210.0
2020-08-11T09:00:11Z          UCm9SZAl03Rev9sFwloCdz1g                Rehaan Records      10        1.1308046E7         655450.0            33242.0   405146.0
2020-08-11T07:30:02Z          UCZRdNleCgW-BGUJf-bbjzQg                Diljit Dosanjh      10        9140911.0           296533.0            6179.0    30058.0
2020-08-10T05:30:49Z          UCq-Fj5jknLsUf-MWSy4_brA                T-Series            10        2.3564512E7         743931.0            84162.0   136942.0
2020-08-11T05:30:13Z          UCye6Oz0mg46S362LwARGVcA                VYRLOriginals       10        6783649.0           268817.0            8798.0    22984.0
2020-08-12T06:56:05Z          UCx6F-rETGiz7xf_vkMmX2yQ                Mythpat 20          1699326.0           332553.0            4627.0    75819.0
2020-08-10T09:29:37Z          UCuFwzKrS0wE43CSkyaHBGiQ                SagaHits            10        7363779.0           301888.0            13836.0   50086.0
2020-08-10T04:30:48Z          UC55IWqFLDH1Xp7iu1_xknRA                Colors TV           24        3816680.0           30086.0   6786.0    3733.0
2020-08-11T10:32:30Z          UCQfE97UMDGgKCFb7iGM8Btg                BakLol Video        24        837562.0            21510.0   2290.0    1704.0
2020-08-12T02:24:34Z          UCoU6AzYucV7Xlg-J5GSTAPg                Gringo Entertainments         10        1466612.0           97192.0   2276.0    3311.0
Time taken: 2.975 seconds, Fetched: 10 row(s)
hive> 
```

1. **Describing the dataset loaded:(Structure)**

Hive> describe data;

From this query we can able to see the structure of the dataset and their respective datatypes.



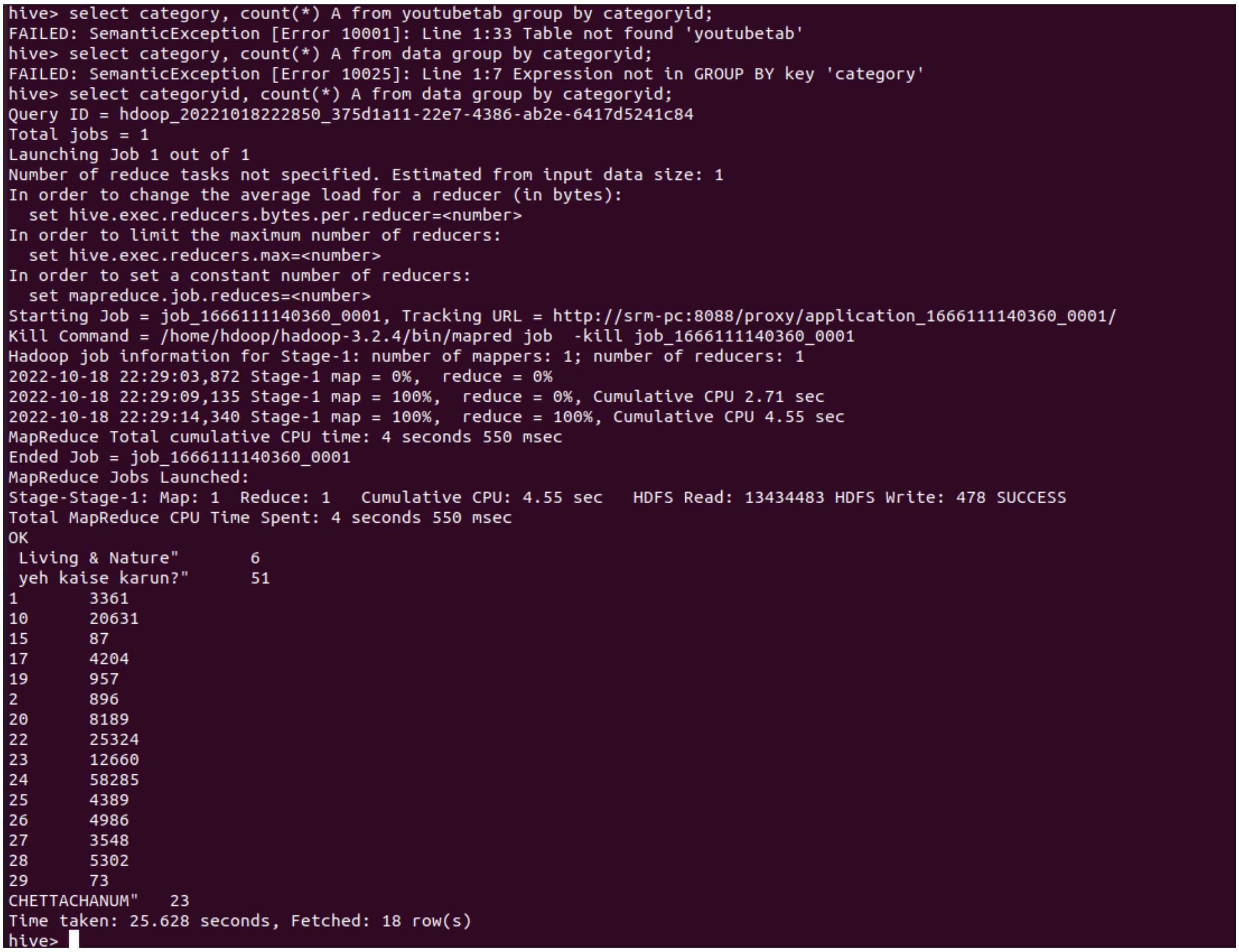
```
hive> describe data;
OK
published               string
channelid               string
channeltitle            string
categoryid              string
view_count              double
likes                   double
dislikes                double
comment_count           double
Time taken: 0.123 seconds, Fetched: 8 row(s)
hive> 
```

2. Group by clause:

The simple group by clause is used to group all the similar rows and increases their count.

```
hive> select category, count(*) A FROM data GROUP BY categoryid;
```

This command will count the total number of category in the row and group it as once.



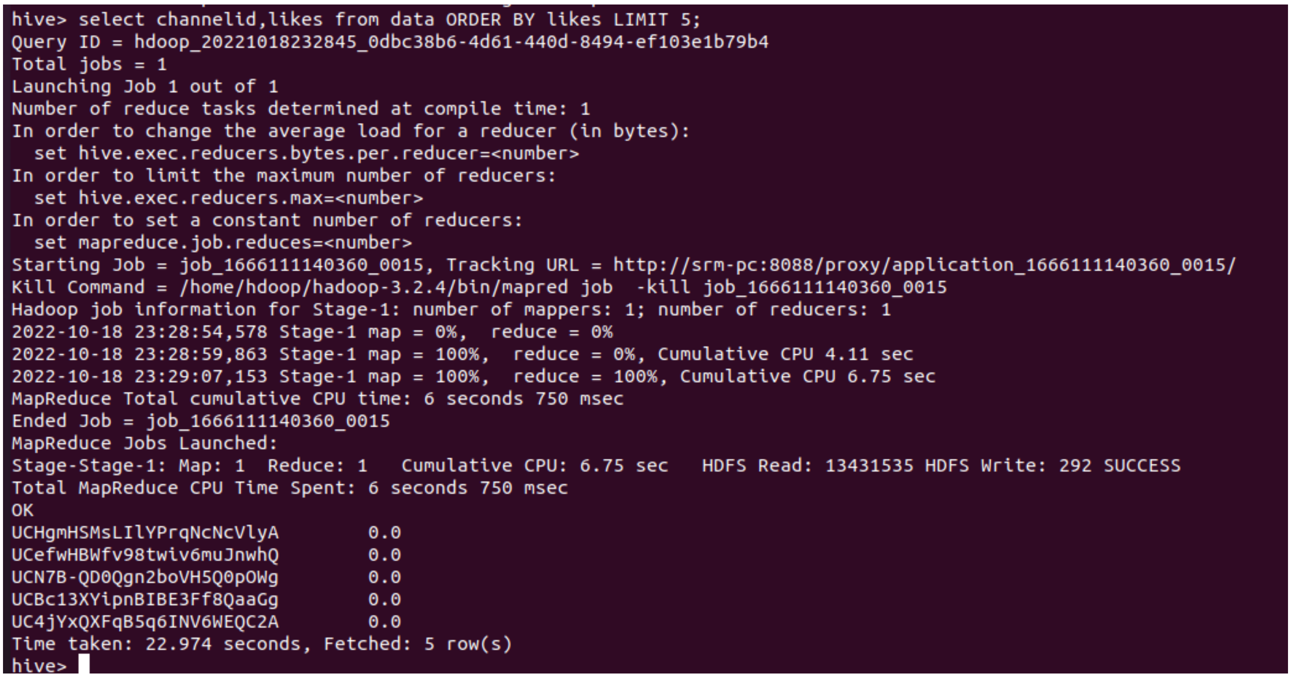
```
hive> select category, count(*) A from youtubetab group by categoryid;
FAILED: SemanticException [Error 10001]: Line 1:33 Table not found 'youtubetab'
hive> select category, count(*) A from data group by categoryid;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'category'
hive> select categoryid, count(*) A from data group by categoryid;
Query ID = hdoop_20221018222850_375d1a11-22e7-4386-ab2e-6417d5241c84
Total 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 mapreduce.job.reduces=<number>
Starting Job = job_1666111140360_0001, Tracking URL = http://srm-pc:8088/proxy/application_1666111140360_0001/
Kill Command = /home/hdoop/hadoop-3.2.4/bin/mapred job  -kill job_1666111140360_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-18 22:29:03,872 Stage-1 map = 0%,  reduce = 0%
2022-10-18 22:29:09,135 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.71 sec
2022-10-18 22:29:14,340 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.55 sec
MapReduce Total cumulative CPU time: 4 seconds 550 msec
Ended Job = job_1666111140360_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.55 sec   HDFS Read: 13434483 HDFS Write: 478 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 550 msec
OK
 Living & Nature"       6
 yeh kaise karun?"      51
1       3361
10      20631
15      87
17      4204
19      957
2       896
20      8189
22      25324
23      12660
24      58285
25      4389
26      4986
27      3548
28      5302
29      73
CHETTACHANUM"   23
Time taken: 25.628 seconds, Fetched: 18 row(s)
hive>
```

3. Order by clause:

The simple order by clause is used to group all the similar rows and display either in ascending or descending order.

```
hive> select channelid, likes FROM data ORDER BY likes LIMIT 5;
```

This command will disply the channelid and likes frm the table data and it by default displays in the ascending order.



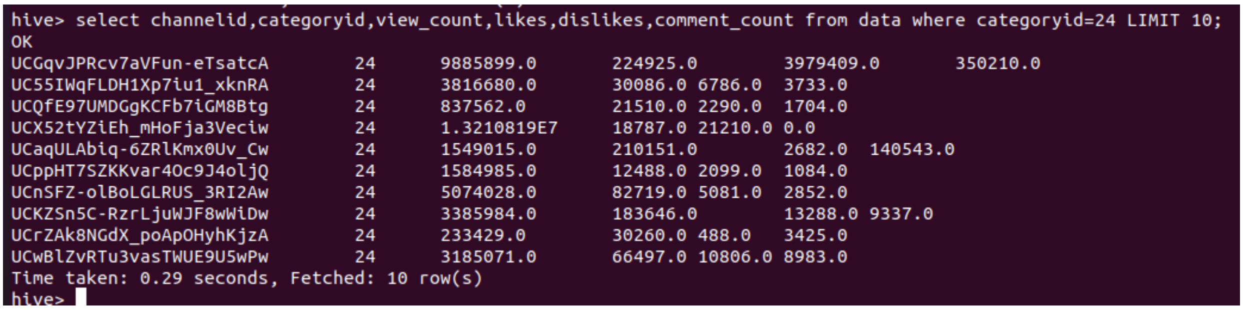
```
hive> select channelid,likes from data ORDER BY likes LIMIT 5;
Query ID = hdoop_20221018232845_0dbc38b6-4d61-440d-8494-ef103e1b79b4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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 mapreduce.job.reduces=<number>
Starting Job = job_1666111140360_0015, Tracking URL = http://srm-pc:8088/proxy/application_1666111140360_0015/
Kill Command = /home/hdoop/hadoop-3.2.4/bin/mapred job  -kill job_1666111140360_0015
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-18 23:28:54,578 Stage-1 map = 0%,  reduce = 0%
2022-10-18 23:28:59,863 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.11 sec
2022-10-18 23:29:07,153 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.75 sec
MapReduce Total cumulative CPU time: 6 seconds 750 msec
Ended Job = job_1666111140360_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.75 sec   HDFS Read: 13431535 HDFS Write: 292 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 750 msec
OK
UCHgmHSMsLIlYPrqNcNcVlyA        0.0
UCefwHBWfv98twiv6muJnwhQ        0.0
UCN7B-QD0Qgn2boVH5Q0pOWg        0.0
UCBc13XYipnBIBE3Ff8QaaGg        0.0
UC4jYxQXFqB5q6INV6WEQC2A        0.0
Time taken: 22.974 seconds, Fetched: 5 row(s)
hive>
```

4. **Conditional statement:**

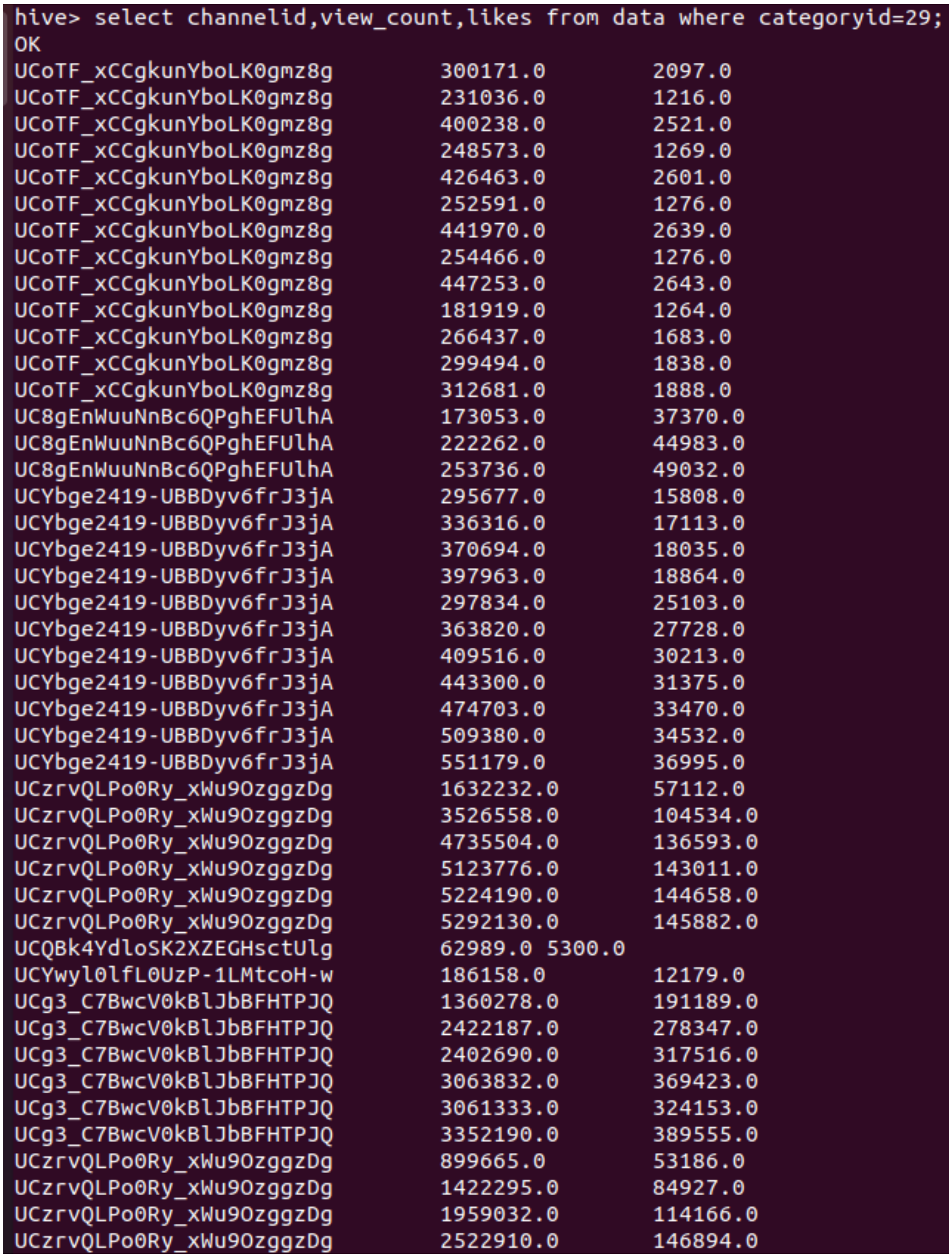
The simple condition is used to display the rows rows some of the conditions used were

```
hive> select channelid, categoryid, view_count,likes,dislikes, comment_count FROM data WHERE categoryid=24 LIMIT 10;
```

- **Equalto(=)** It will gives the exact result of the given conditions.



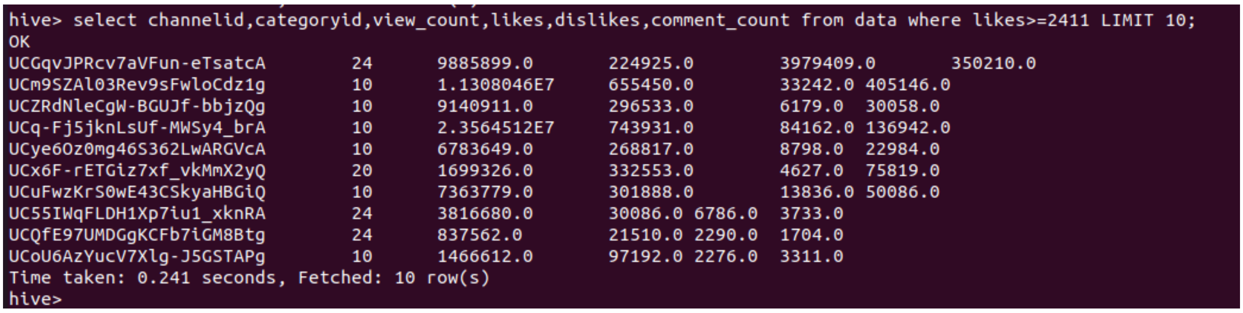
```
hive> select channelid,categoryid,view_count,likes,dislikes,comment_count from data where categoryid=24 LIMIT 10;
OK
UCGqvJPRcv7aVFun-eTsatcA        24      9885899.0       224925.0        3979409.0       350210.0
UC55IWqFLDH1Xp7iu1_xknRA        24      3816680.0       30086.0 6786.0  3733.0
UCQfE97UMDGgKCFb7iGM8Btg        24      837562.0        21510.0 2290.0  1704.0
UCX52tYZiEh_mHoFja3Veciw        24      1.3210819E7     18787.0 21210.0 0.0
UCaqULAbiq-6ZRlKmx0Uv_Cw        24      1549015.0       210151.0        2682.0  140543.0
UCppHT7SZKKvar4Oc9J4oljQ        24      1584985.0       12488.0 2099.0  1084.0
UCnSFZ-olBoLGLRUS_3RI2Aw        24      5074028.0       82719.0 5081.0  2852.0
UCKZSn5C-RzrLjuWJF8wWiDw        24      3385984.0       183646.0        13288.0 9337.0
UCrZAk8NGdX_poApOHyhKjzA        24      233429.0        30260.0 488.0   3425.0
UCwBlZvRTu3vasTWUE9U5wPw        24      3185071.0       66497.0 10806.0 8983.0
Time taken: 0.29 seconds, Fetched: 10 row(s)
hive>
hive> select channelid, view_count,likes FROM data WHERE
categoryid=29;
```



```
hive> select channelid,view_count,likes from data where categoryid=29;
OK
UCoTF_xCCgkunYboLK0gmz8g        300171.0        2097.0
UCoTF_xCCgkunYboLK0gmz8g        231036.0        1216.0
UCoTF_xCCgkunYboLK0gmz8g        400238.0        2521.0
UCoTF_xCCgkunYboLK0gmz8g        248573.0        1269.0
UCoTF_xCCgkunYboLK0gmz8g        426463.0        2601.0
UCoTF_xCCgkunYboLK0gmz8g        252591.0        1276.0
UCoTF_xCCgkunYboLK0gmz8g        441970.0        2639.0
UCoTF_xCCgkunYboLK0gmz8g        254466.0        1276.0
UCoTF_xCCgkunYboLK0gmz8g        447253.0        2643.0
UCoTF_xCCgkunYboLK0gmz8g        181919.0        1264.0
UCoTF_xCCgkunYboLK0gmz8g        266437.0        1683.0
UCoTF_xCCgkunYboLK0gmz8g        299494.0        1838.0
UCoTF_xCCgkunYboLK0gmz8g        312681.0        1888.0
UC8gEnWuuNnBc6QPghEFUlhA        173053.0        37370.0
UC8gEnWuuNnBc6QPghEFUlhA        222262.0        44983.0
UC8gEnWuuNnBc6QPghEFUlhA        253736.0        49032.0
UCYbge2419-UBBDyv6frJ3jA        295677.0        15808.0
UCYbge2419-UBBDyv6frJ3jA        336316.0        17113.0
UCYbge2419-UBBDyv6frJ3jA        370694.0        18035.0
UCYbge2419-UBBDyv6frJ3jA        397963.0        18864.0
UCYbge2419-UBBDyv6frJ3jA        297834.0        25103.0
UCYbge2419-UBBDyv6frJ3jA        363820.0        27728.0
UCYbge2419-UBBDyv6frJ3jA        409516.0        30213.0
UCYbge2419-UBBDyv6frJ3jA        443300.0        31375.0
UCYbge2419-UBBDyv6frJ3jA        474703.0        33470.0
UCYbge2419-UBBDyv6frJ3jA        509380.0        34532.0
UCYbge2419-UBBDyv6frJ3jA        551179.0        36995.0
UCzrvQLPo0Ry_xWu9OzggzDg        1632232.0       57112.0
UCzrvQLPo0Ry_xWu9OzggzDg        3526558.0       104534.0
UCzrvQLPo0Ry_xWu9OzggzDg        4735504.0       136593.0
UCzrvQLPo0Ry_xWu9OzggzDg        5123776.0       143011.0
UCzrvQLPo0Ry_xWu9OzggzDg        5224190.0       144658.0
UCzrvQLPo0Ry_xWu9OzggzDg        5292130.0       145882.0
UCQBk4YdloSK2XZEGHsctUlg        62989.0 5300.0
UCYwyl0lfL0UzP-1LMtcoH-w        186158.0        12179.0
UCg3_C7BwcV0kBlJbBFHTPJQ        1360278.0       191189.0
UCg3_C7BwcV0kBlJbBFHTPJQ        2422187.0       278347.0
UCg3_C7BwcV0kBlJbBFHTPJQ        2402690.0       317516.0
UCg3_C7BwcV0kBlJbBFHTPJQ        3063832.0       369423.0
UCg3_C7BwcV0kBlJbBFHTPJQ        3061333.0       324153.0
UCg3_C7BwcV0kBlJbBFHTPJQ        3352190.0       389555.0
UCzrvQLPo0Ry_xWu9OzggzDg        899665.0        53186.0
UCzrvQLPo0Ry_xWu9OzggzDg        1422295.0       84927.0
UCzrvQLPo0Ry_xWu9OzggzDg        1959032.0       114166.0
UCzrvQLPo0Ry_xWu9OzggzDg        2522910.0       146894.0
```

- **<u>Greaterthan(>=)</u>** This condition will give the values greater than the specified values.

```
7.
hive> select channelid, categoryid, view_count,likes,dislikes,
comment_count FROM data WHERE LIKES>=2411 LIMIT 10;
```

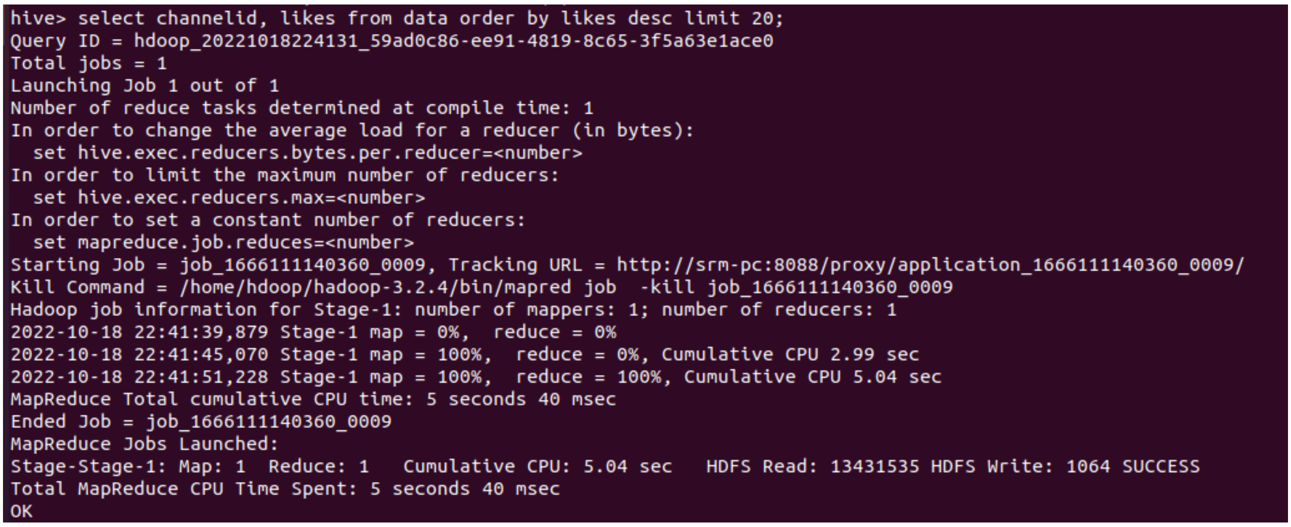


```
hive> select channelid,categoryid,view_count,likes,dislikes,comment_count from data where likes>=2411 LIMIT 10;
OK
UCGqvJPRcv7aVFun-eTsatcA        24      9885899.0       224925.0        3979409.0       350210.0
UCm9SZAl03Rev9sFwloCdz1g        10      1.1308046E7     655450.0        33242.0 405146.0
UCZRdNleCgW-BGUJf-bbjzQg        10      9140911.0       296533.0        6179.0  30058.0
UCq-Fj5jknLsUf-MWSy4_brA        10      2.3564512E7     743931.0        84162.0 136942.0
UCye6Oz0mg46S362LwARGVcA        10      6783649.0       268817.0        8798.0  22984.0
UCx6F-rETGiz7xf_vkMmX2yQ        20      1699326.0       332553.0        4627.0  75819.0
UCuFwzKrS0wE43CSkyaHBGiQ        10      7363779.0       301888.0        13836.0 50086.0
UC55IWqFLDH1Xp7iu1_xknRA        24      3816680.0       30086.0 6786.0  3733.0
UCQfE97UMDGgKCFb7iGM8Btg        24      837562.0        21510.0 2290.0  1704.0
UCoU6AzYucV7Xlg-J5GSTAPg        10      1466612.0       97192.0 2276.0  3311.0
Time taken: 0.241 seconds, Fetched: 10 row(s)
hive>
```

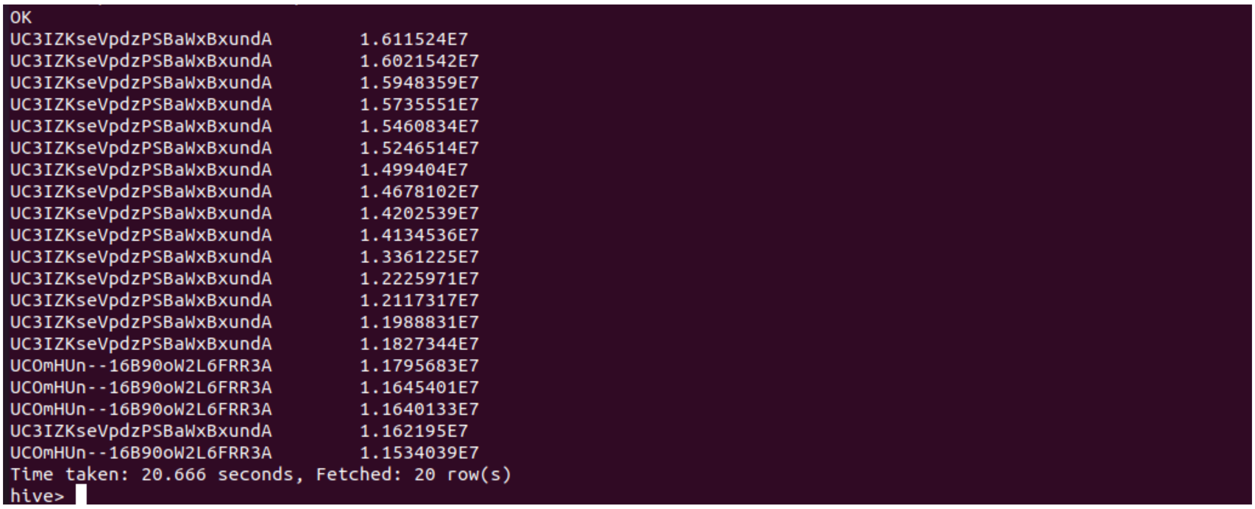
8. Calculate top 10 channels with maximum number of likes

We can extract the top 10 channels with maximum number of likes using the following Hive query. The Hive select query will trigger the following MapReduce job:

```
hive> select channelid,  likes FROM data ORDER BY likes DESC LIMIT 10;
```



```
hive> select channelid, likes from data order by likes desc limit 20;
Query ID = hdoop_20221018224131_59ad0c86-ee91-4819-8c65-3f5a63e1ace0
Total 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 mapreduce.job.reduces=<number>
Starting Job = job_1666111140360_0009, Tracking URL = http://srm-pc:8088/proxy/application_1666111140360_0009/
Kill Command = /home/hdoop/hadoop-3.2.4/bin/mapred job  -kill job_1666111140360_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-18 22:41:39,879 Stage-1 map = 0%,  reduce = 0%
2022-10-18 22:41:45,070 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.99 sec
2022-10-18 22:41:51,228 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.04 sec
MapReduce Total cumulative CPU time: 5 seconds 40 msec
Ended Job = job_1666111140360_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.04 sec   HDFS Read: 13431535 HDFS Write: 1064 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 40 msec
OK
```



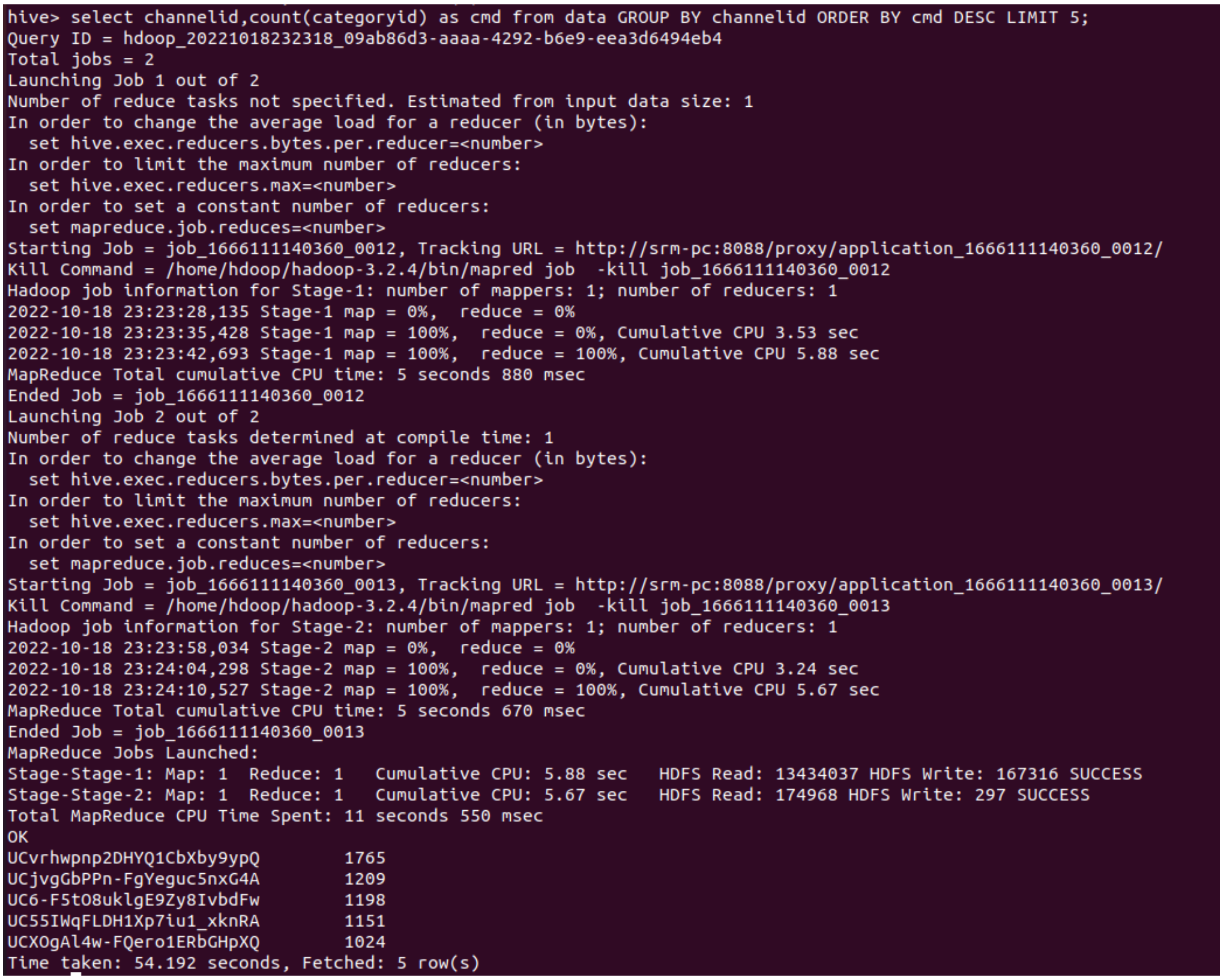
```
OK
UC3IZKseVpdzPSBaWxBxundA        1.611524E7
UC3IZKseVpdzPSBaWxBxundA        1.6021542E7
UC3IZKseVpdzPSBaWxBxundA        1.5948359E7
UC3IZKseVpdzPSBaWxBxundA        1.5735551E7
UC3IZKseVpdzPSBaWxBxundA        1.5460834E7
UC3IZKseVpdzPSBaWxBxundA        1.5246514E7
UC3IZKseVpdzPSBaWxBxundA        1.499404E7
UC3IZKseVpdzPSBaWxBxundA        1.4678102E7
UC3IZKseVpdzPSBaWxBxundA        1.4202539E7
UC3IZKseVpdzPSBaWxBxundA        1.4134536E7
UC3IZKseVpdzPSBaWxBxundA        1.3361225E7
UC3IZKseVpdzPSBaWxBxundA        1.2225971E7
UC3IZKseVpdzPSBaWxBxundA        1.2117317E7
UC3IZKseVpdzPSBaWxBxundA        1.1988831E7
UC3IZKseVpdzPSBaWxBxundA        1.1827344E7
UCOmHUn--16B90oW2L6FRR3A        1.1795683E7
UCOmHUn--16B90oW2L6FRR3A        1.1645401E7
UCOmHUn--16B90oW2L6FRR3A        1.1640133E7
UC3IZKseVpdzPSBaWxBxundA        1.162195E7
UCOmHUn--16B90oW2L6FRR3A        1.1534039E7
Time taken: 20.666 seconds, Fetched: 20 row(s)
hive>
```

The output result describes that for a specific category id, how many likes were received. The number of likes -- or "thumbs-up" -- a video had has a direct significance to the YouTube video's ranking, according to YouTube Analytics. So if a company posts its video on YouTube, then the number of YouTube likes the company has could determine whether the company or its competitors appear more prominently in YouTube search results. The output result shows number of likes for "Disney" channel videos

9. Calculate top 5 channels with maximum number of category_id

```
 hive> select channelid, count(categoryid) as cmd FROM data GROUP BY
channelid ORDER BY cmd DESC LIMIT 10;
```

This command will say top channel in our dataset based in the category so we can analysis easily as this category id has more number of the channels.



```
hive> select channelid,count(categoryid) as cmd from data GROUP BY channelid ORDER BY cmd DESC LIMIT 5;
Query ID = hdoop_20221018232318_09ab86d3-aaaa-4292-b6e9-eea3d6494eb4
Total 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 mapreduce.job.reduces=<number>
Starting Job = job_1666111140360_0012, Tracking URL = http://srm-pc:8088/proxy/application_1666111140360_0012/
Kill Command = /home/hdoop/hadoop-3.2.4/bin/mapred job  -kill job_1666111140360_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-18 23:23:28,135 Stage-1 map = 0%,  reduce = 0%
2022-10-18 23:23:35,428 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.53 sec
2022-10-18 23:23:42,693 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.88 sec
MapReduce Total cumulative CPU time: 5 seconds 880 msec
Ended Job = job_1666111140360_0012
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_1666111140360_0013, Tracking URL = http://srm-pc:8088/proxy/application_1666111140360_0013/
Kill Command = /home/hdoop/hadoop-3.2.4/bin/mapred job  -kill job_1666111140360_0013
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-18 23:23:58,034 Stage-2 map = 0%,  reduce = 0%
2022-10-18 23:24:04,298 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.24 sec
2022-10-18 23:24:10,527 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.67 sec
MapReduce Total cumulative CPU time: 5 seconds 670 msec
Ended Job = job_1666111140360_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.88 sec   HDFS Read: 13434037 HDFS Write: 167316 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.67 sec   HDFS Read: 174968 HDFS Write: 297 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 550 msec
OK
UCvrhwpnp2DHYQ1CbXby9ypQ        1765
UCjvgGbPPn-FgYeguc5nxG4A        1209
UC6-F5tO8uklgE9Zy8IvbdFw        1198
UC55IWqFLDH1Xp7iu1_xknRA        1151
UCXOgAl4w-FQero1ERbGHpXQ        1024
Time taken: 54.192 seconds, Fetched: 5 row(s)
```

10. Calculate top 5 categories with maximum number of comments

hive > select category, max(no_of_comments) as max_no_of_comments from YouTube_data_table GROUP ORDER BY max_no_of_comments DESC LIMIT 5;