

Big Data Analytics

Lab Program-3

Arpita Manoj Chikkodi

1RV18MCA07

Program on Pig Script Using movie lens dataset

Movielens Dataset

It consists of 1000 records with the fields movie id,title of the movie, user id, rating, genre id, recommended and activity state.

Procedure

Step1: Add Movielens.csv through Ambari

Step 2: Copy the file to root folder

>mkdir lab3

>hdfs dfs -copyToLocal /MCA_LAB/lab3/Movielens.csv lab3/Movielens.csv

Step 3: Start grunt shell

> pig -x local **or** **>pig**

Step 4: Load the dataset and display

>movie = LOAD 'lab3/Movielens.csv' USING PigStorage(',') as (movie_id:int, title:chararray, user_id:int, ratings:double, genre_id:int, Recommended:chararray, Activity_State: int);

>dump movie;

```
root@sandbox: ~ - Shell in A Box x +
192.168.56.101:4200
Apps Gmail BDA CC Seminar1 MP-II Big Data Analytics V SEM - SPM An Efficient Machin... Accurate classificati... Deep Dive into Mac... root@sandbox: ~ - ... PDBe home < EMB...

root@sandbox: ~# ls
anaconda-ks.cfg build.out install.log lab2 lc_input.txt pig_1601080245976.log sandbox.info start_hbase.sh
blueprint.json hdp install.log.syslog lcfinal.jar lcl.jar pig_1601081645193.log start_ambari.sh
root@sandbox: ~# mkdir lab3
root@sandbox: ~# ls
anaconda-ks.cfg build.out install.log lab2 lcfinal.jar lcl.jar pig_1601081645193.log start_ambari.sh
blueprint.json hdp install.log.syslog lab3 lc_input.txt pig_1601080245976.log sandbox.info start_hbase.sh
root@sandbox: ~# hdfs -copyToLocal /MCA_LAB/lab3/MovieLens.csv lab3/MovieLens.csv
root@sandbox: ~# ls lab3/
MovieLens.csv
root@sandbox: ~# pig -x local
2020-09-26 01:03:42 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2020-09-26 01:03:42 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2020-09-26 01:03:42,355 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0.2.5.0-1245 (reexported) compiled Aug 26 2016, 02:07:35
2020-09-26 01:03:42,355 [main] INFO org.apache.pig.Main - Logging error messages to: /root/pig_1601082222353.log
2020-09-26 01:03:42,441 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /root/.pigbootstrap not found
2020-09-26 01:03:42,774 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2020-09-26 01:03:43,354 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-4446fe9b-48ca-4989-b842-c33534b79ef9
2020-09-26 01:03:43,354 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> movie = LOAD 'lab3/MovieLens.csv' USING PigStorage(',') as (movie_id:int,title:chararray, user_id:int, ratings:double, genre_id:int, Recommended:chararray, Activ
ity_State:int);
grunt> dump movie;
2020-09-26 01:04:23,841 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2020-09-26 01:04:24,175 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator,
GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatte
n, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2020-09-26 01:04:24,324 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThreshold
= 489580128, usageThreshold = 489580128
2020-09-26 01:04:24,706 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCCompiler - File concatenation threshold: 100 optimistic? false
2020-09-26 01:04:24,742 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2020-09-26 01:04:24,742 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2020-09-26 01:04:24,819 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Initializing JVM Metrics with processName=JobTracker, sessionId=
2020-09-26 01:04:24,866 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScripState - Pig script settings are added to the job
2020-09-26 01:04:24,875 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not
set, set to default: 0.3
2020-09-26 01:04:24,899 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2020-09-26 01:04:24,927 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2020-09-26 01:04:24,927 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
```

Job Status

```
root@sandbox: ~ - Shell in A Box x +
192.168.56.101:4200
Apps Gmail BDA CC Seminar1 MP-II Big Data Analytics V SEM - SPM An Efficient Machin... Accurate classificati... Deep Dive into Mac... root@sandbox: ~ - ... PDBe home < EMB...

2020-09-26 01:04:42,596 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2020-09-26 01:04:42,601 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.7.3.2.5.0-1245 0.16.0.2.5.0-1245 root 2020-09-26 01:04:24 2020-09-26 01:04:42 UNKNOWN

Success!

Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias F
eature Outputs
job_local1674071291_0001 1 0 n/a n/a n/a n/a 0 0 0 0 movie MAP_ONLY file:/tmp/temp-36940135/tmp-4646
93799,

Input(s):
Successfully read 983 records from: "file:///root/lab3/MovieLens.csv"

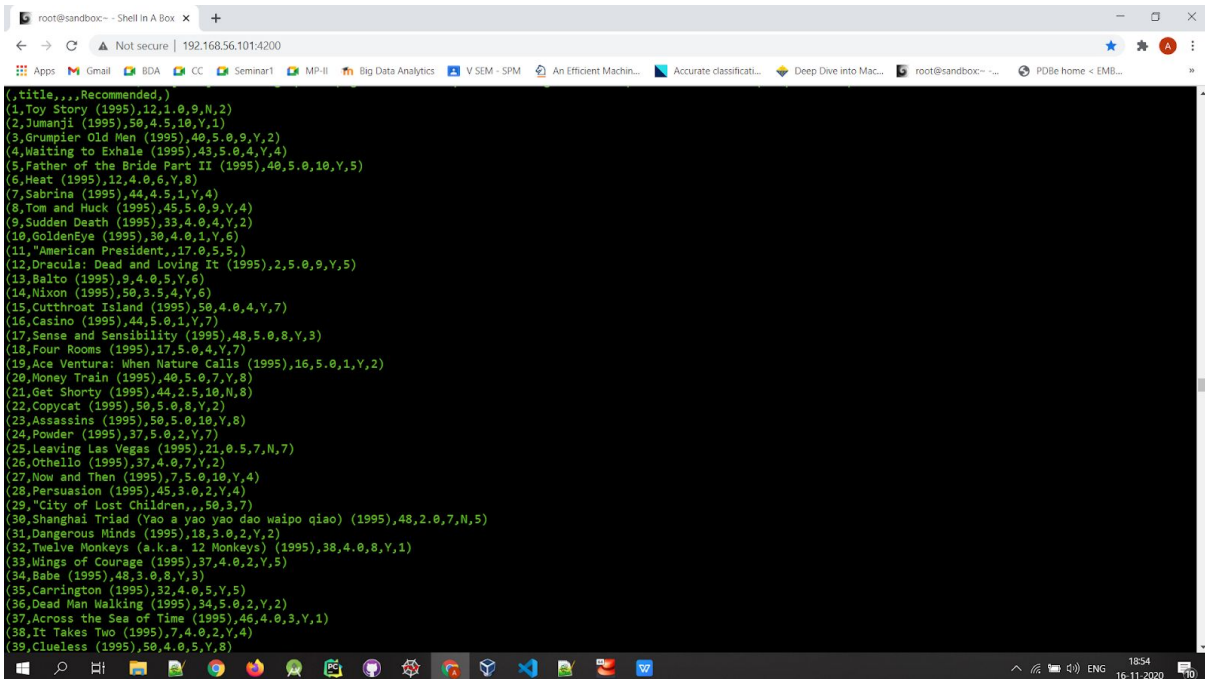
Output(s):
Successfully stored 983 records in: "file:/tmp/temp-36940135/tmp-464693799"

Counters:
Total records written : 983
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1674071291_0001

2020-09-26 01:04:42,606 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 01:04:42,609 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 01:04:42,612 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 01:04:42,625 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSI
ON_FAILED 452 time(s).
2020-09-26 01:04:42,626 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

After dumping the dataset that is loaded to pig

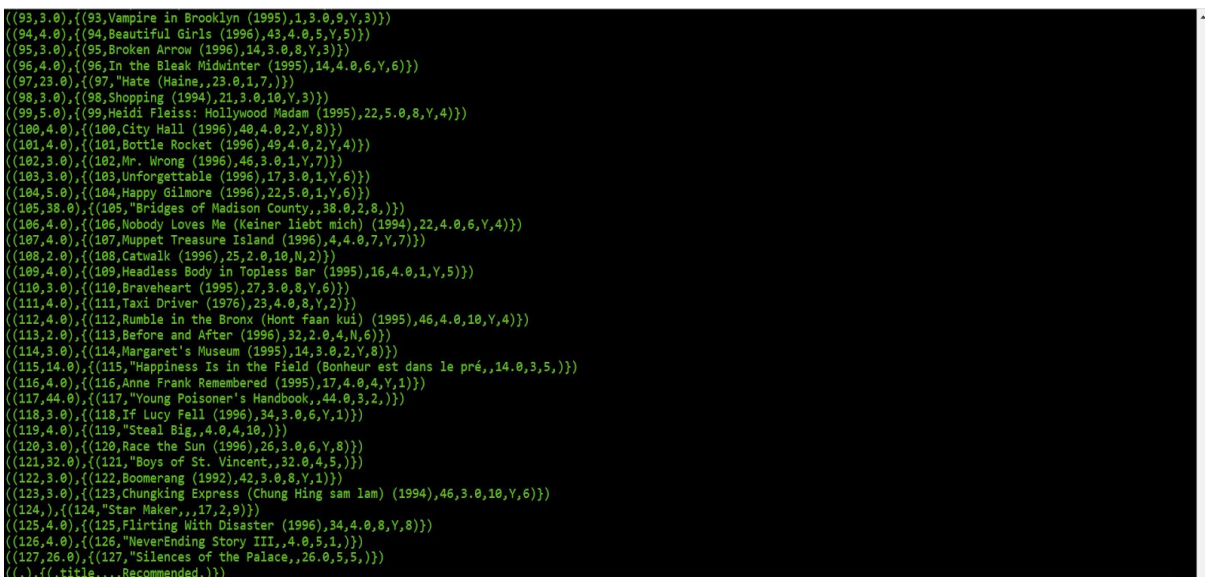


```
{title,,,Recommended,}
(1,Toy Story (1995),12,1.0,9,N,2)
(2,Jumanji (1995),50,4.5,10,Y,1)
(3,Grumpier Old Men (1995),40,5.0,9,Y,2)
(4,Waiting to Exhale (1995),43,5.0,4,Y,4)
(5,Father of the Bride Part II (1995),40,5.0,10,Y,5)
(6,Heat (1995),12,4.0,6,Y,8)
(7,Sabrina (1995),44,4.5,1,Y,4)
(8,Tom and Huck (1995),45,5.0,9,Y,4)
(9,Sudden Death (1995),33,4.0,4,Y,2)
(10,GoldenEye (1995),30,4.0,1,Y,6)
(11,"American President,,17.0,5,5)
(12,Dracula: Dead and Loving It (1995),2,5.0,9,Y,5)
(13,Balto (1995),9,4.0,5,Y,6)
(14,Nixon (1995),50,3.5,4,Y,6)
(15,Cutthroat Island (1995),50,4.0,4,Y,7)
(16,Casino (1995),44,5.0,1,Y,7)
(17,Sense and Sensibility (1995),48,5.0,8,Y,3)
(18,Four Rooms (1995),17,5.0,4,Y,7)
(19,Ace Ventura: When Nature Calls (1995),16,5.0,1,Y,2)
(20,Money Train (1995),40,5.0,7,Y,8)
(21,Get Shorty (1995),44,2.5,10,N,8)
(22,Copycat (1995),50,5.0,8,Y,2)
(23,Assassins (1995),50,5.0,10,Y,8)
(24,Powder (1995),37,5.0,2,Y,7)
(25,Leaving Las Vegas (1995),21,0.5,7,N,7)
(26,Othello (1995),37,4.0,7,Y,2)
(27,Now and Then (1995),7,5.0,10,Y,4)
(28,Persuasion (1995),45,3.0,2,Y,4)
(29,"City of Lost Children,,50,3,7)
(30,Shanghai Triad (Yao a yao yao dao waipo qiao) (1995),48,2.0,7,N,5)
(31,Dangerous Minds (1995),18,3.0,2,Y,2)
(32,Twelve Monkeys (a.k.a. 12 Monkeys) (1995),38,4.0,8,Y,1)
(33,Wings of Courage (1995),37,4.0,2,Y,5)
(34,Babe (1995),48,3.0,8,Y,3)
(35,Carrington (1995),32,4.0,5,Y,5)
(36,Dead Man Walking (1995),34,5.0,2,Y,2)
(37,Across the Sea of Time (1995),46,4.0,3,Y,1)
(38,It Takes Two (1995),7,4.0,2,Y,4)
(39,Clueless (1995),50,4.0,5,Y,8)
```

a) List all the movies and the number of ratings

```
>query1 = group movie by (movie_id,ratings);
```

```
>dump query1;
```



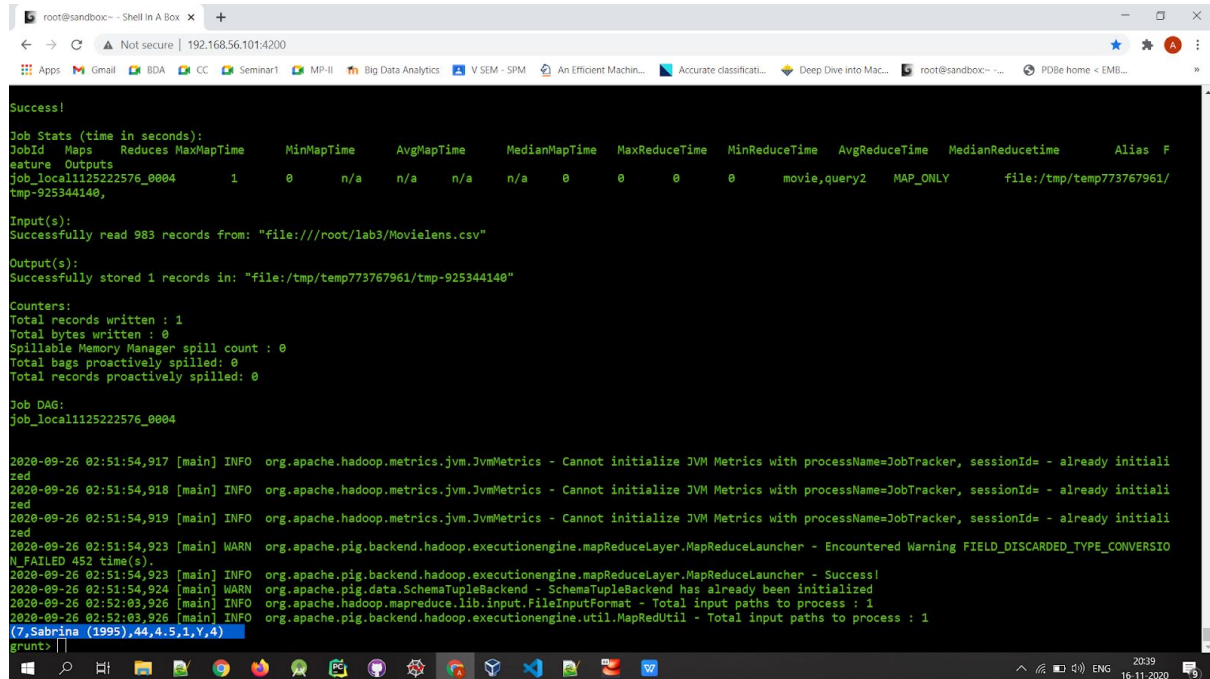
```
((93,3.0),{(93,Vampire in Brooklyn (1995),1,3.0,9,Y,3)})
((94,4.0),{(94,Beautiful Girls (1996),43,4.0,5,Y,5)})
((95,3.0),{(95,Broken Arrow (1996),14,3.0,8,Y,3)})
((96,4.0),{(96,In the Bleak Midwinter (1995),14,4.0,6,Y,6)})
((97,23.0),{(97,"Hate (Haine,,23.0,1,7,))}
((98,3.0),{(98,Shopping (1994),21,3.0,10,Y,3)})
((99,5.0),{(99,Heidi Fleiss: Hollywood Madam (1995),22,5.0,8,Y,4)})
((100,4.0),{(100,City Hall (1996),40,4.0,2,Y,8)})
((101,4.0),{(101,Bottle Rocket (1996),49,4.0,2,Y,4)})
((102,3.0),{(102,Mr. Wrong (1996),46,3.0,1,Y,7)})
((103,3.0),{(103,Unforgettable (1996),17,3.0,1,Y,6)})
((104,5.0),{(104,Happy Gilmore (1996),22,5.0,1,Y,6)})
((105,38.0),{(105,"Bridges of Madison County,38.0,2,8,))}
((106,4.0),{(106,Nobody Loves Me (Keiner liebt mich) (1994),22,4.0,6,Y,4)})
((107,4.0),{(107,Muppet Treasure Island (1996),4,4.0,7,Y,7)})
((108,2.0),{(108,Catwalk (1996),25,2.0,10,N,2)})
((109,4.0),{(109,Headless Body in Topless Bar (1995),16,4.0,1,Y,5)})
((110,3.0),{(110,Braveheart (1995),27,3.0,8,Y,6)})
((111,4.0),{(111,Taxi Driver (1976),23,4.0,8,Y,2)})
((112,4.0),{(112,Rumble in the Bronx (Hont faan kui) (1995),46,4.0,10,Y,4)})
((113,2.0),{(113,Before and After (1996),32,2.0,4,N,6)})
((114,3.0),{(114,Margaret's Museum (1995),14,3.0,2,Y,8)})
((115,14.0),{(115,"Happiness Is in the Field (Bonheur est dans le pré,,14.0,3,5,))}
((116,4.0),{(116,Anne Frank Remembered (1995),17,4.0,4,Y,1)})
((117,44.0),{(117,"Young Poisoner's Handbook,,44.0,3,2,))}
((118,3.0),{(118,If Lucy Fell (1996),34,3.0,6,Y,1)})
((119,4.0),{(119,"Steal Big,,4.0,4,10,))}
((120,3.0),{(120,Race the Sun (1996),26,3.0,6,Y,8)})
((121,32.0),{(121,"Boys of St. Vincent,,32.0,4,5,))}
((122,3.0),{(122,Boomerang (1992),42,3.0,8,Y,1)})
((123,3.0),{(123,Chungking Express (Chung Hing sam lam) (1994),46,3.0,10,Y,6)})
((124,,{(124,"Star Maker,,17,2,9,))}
((125,4.0),{(125,Flirting With Disaster (1996),34,4.0,8,Y,8)})
((126,4.0),{(126,"NeverEnding Story III,,4.0,5,1,))}
((127,26.0),{(127,"Silences of the Palace,,26.0,5,5,))}
(,),(title,,,Recommended,))
```

b) List all the users who have rated the same movie and find the number of ratings

For dataset considered, each movie is rated by one user only

>query2 = filter movie by movie_id==7;

>dump query2;



```
Success!
Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReduceTime  Alias  F
feature  Outputs
Job_local112522576_0004  1  0  n/a  n/a  n/a  n/a  0  0  0  0  movie,query2  MAP_ONLY  file:/tmp/temp773767961/
tmp-925344140,

Input(s):
Successfully read 983 records from: "file:///root/lab3/MovieLens.csv"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp773767961/tmp-925344140"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
Job_local112522576_0004

2020-09-26 02:51:54,917 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:51:54,918 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:51:54,919 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:51:54,923 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSIO
N FAILED 452 time(s).
2020-09-26 02:51:54,924 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-09-26 02:51:54,924 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-09-26 02:52:03,926 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2020-09-26 02:52:03,926 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(7,Sabrina (1995),44,4.5,1,Y,4)
grunt> |
```

For different dataset if multiple users have rated same movie

>query2 = filter movie by movie_id==7;

>dump query2;

>query21 = group query2 all;

>query22 = foreach query21 generate COUNT(query2.ratings)

c) List all the Users who have rated the movies (Users who have rated at least one movie)

>query3 = group movie by (user_id,movie_id);

>dump query3;

Or

>query3 = filter movie by user_id is not null and movie_id is not null;

>dump query3;

d) Find the count of the Movie which has the ratings more than 3

>query4 = filter movie by ratings>3;

>query41 = group query4 all;

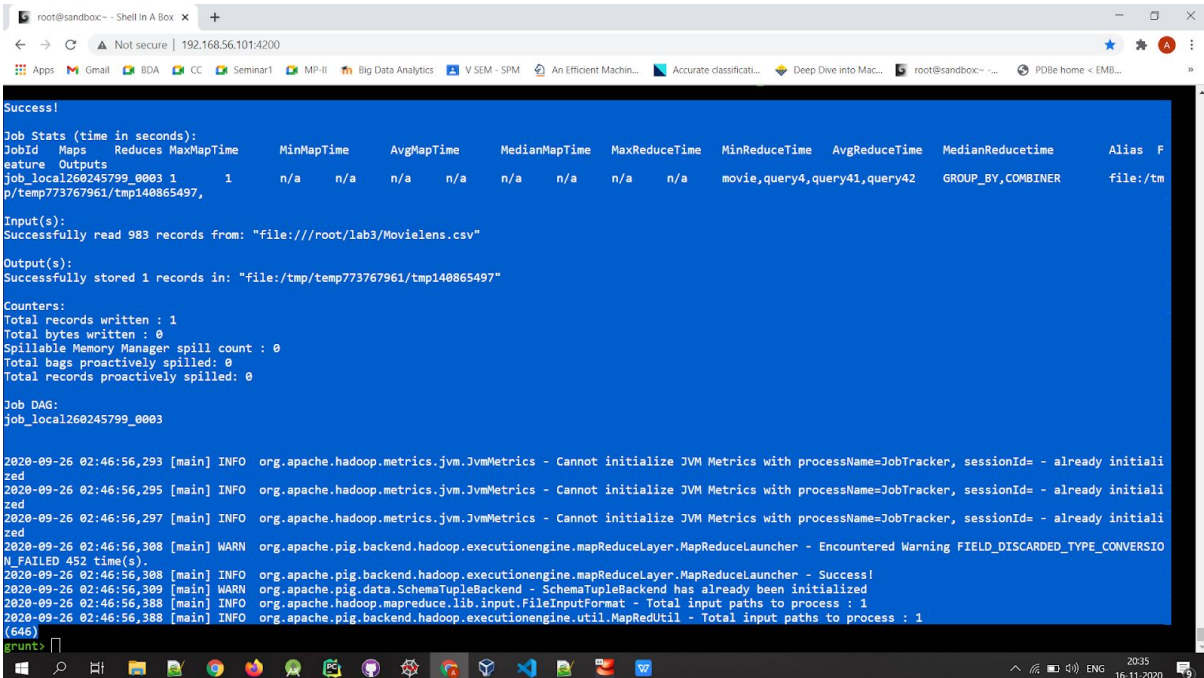
(dump query41 to see all movies with rating greater than 3)

>query42 = foreach query41 generate COUNT(query4.movie_id);

>dump query42;

Output

646



```
Success!
Job Stats (time in seconds):
JobId  Maps  Reduces MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReducetime  Alias  F
eature  Outputs
job_local260245799_0003  1      1      n/a      n/a      n/a      n/a      n/a      n/a      n/a      n/a      movie,query4,query41,query42  GROUP_BY,COMBINER  file:/tm
p/temp773767961/tmp140865497,

Input(s):
Successfully read 983 records from: "file:///root/lab3/Movielens.csv"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp773767961/tmp140865497"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local260245799_0003

2020-09-26 02:46:56,293 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:46:56,295 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:46:56,297 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:46:56,308 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSIO
N_FAILED 452 time(s).
2020-09-26 02:46:56,308 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-09-26 02:46:56,309 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-09-26 02:46:56,388 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2020-09-26 02:46:56,388 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(646)
```

e) Find the max, min, average ratings for all the movie

>query5 = group movie all;

>queryavg = foreach query5 generate AVG(movie.ratings) as avg;

>querymax = foreach query5 generate MAX(movie.ratings) as max;

>querymin = foreach query5 generate MIN(movie.ratings) as min;

>queryres = foreach query5 generate (queryavg.avg, querymax.max, querymin.min);

>dump queryres;

```
tmp-265947234,
Input(s):
Successfully read 983 records from: "file:///root/lab3/MovieLens.csv"

Output(s):
Successfully stored 1 records in: "file:/tmp/temp-36940135/tmp-265947234"

Counters:
Total records written : 1
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1553860959_0005    ->    job_local1945587538_0006,
job_local1945587538_0006

2020-09-26 02:12:53,969 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:12:53,970 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:12:53,971 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:12:53,979 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:12:53,981 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:12:53,982 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2020-09-26 02:12:53,986 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSIO
N FAILED 452 time(s).
2020-09-26 02:12:53,986 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-09-26 02:12:53,987 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-09-26 02:12:54,010 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2020-09-26 02:12:54,010 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((7.797717842323651,50.0,0.5))
```