Big Data Analytics

Lab Program-3

Aditya Raj

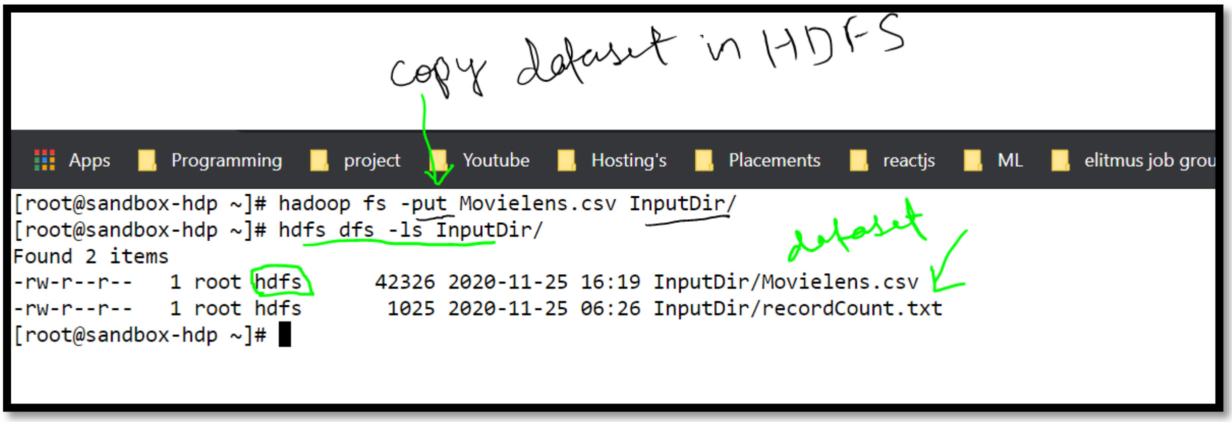
1RD18MCA01

Program on Pig Script Using movie lens dataset

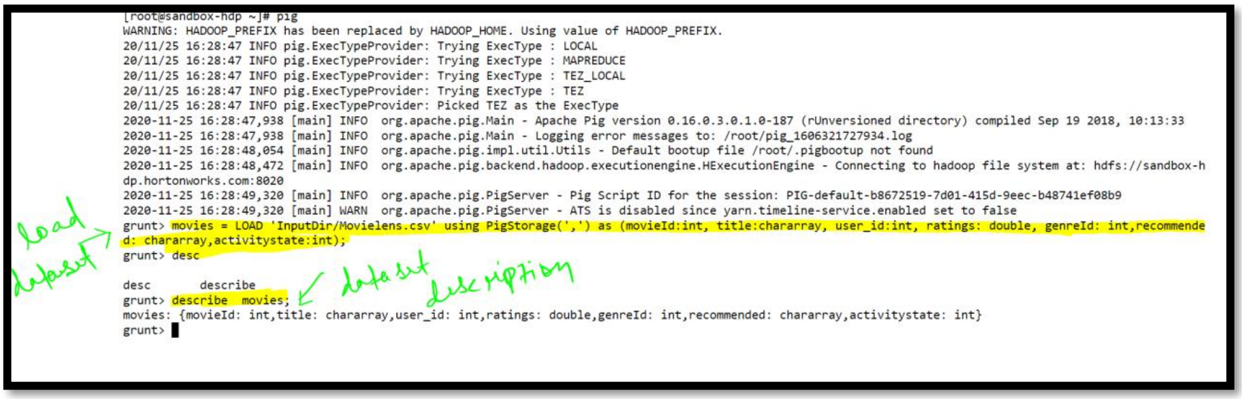
Movielens Dataset

It consists of 987records with the field’s movie id, title of the movie, user id, rating, genre id, recommended and activity state.

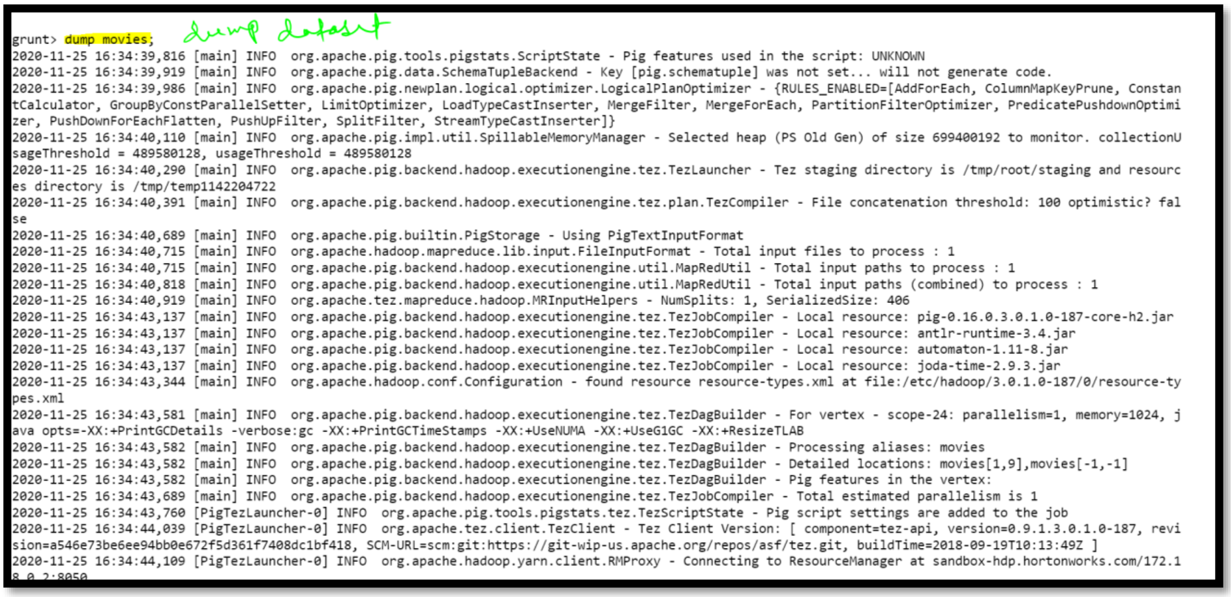
Pig Script on Movielens Dataset



```
[root@sandbox-hdp ~]# hadoop fs -put Movielens.csv InputDir/
[root@sandbox-hdp ~]# hdfs dfs -ls InputDir/
Found 2 items
-rw-r--r--   1 root hdfs      42326 2020-11-25 16:19 InputDir/Movielens.csv
-rw-r--r--   1 root hdfs       1025 2020-11-25 06:26 InputDir/recordCount.txt
[root@sandbox-hdp ~]#
```

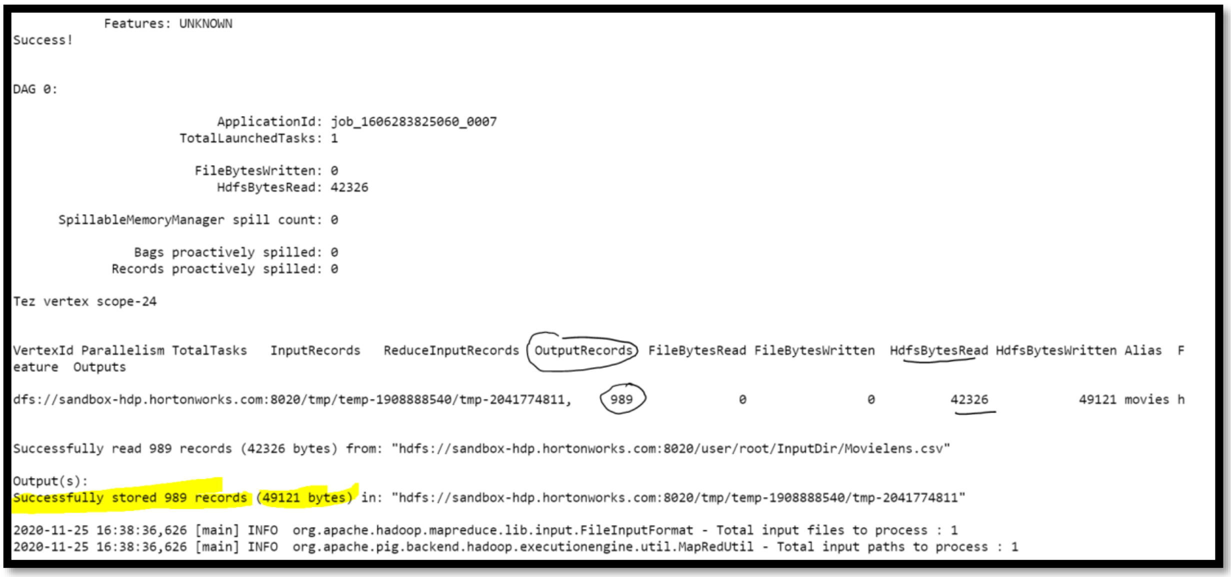


```
[root@sandbox-hdp ~]# pig
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
20/11/25 16:28:47 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
20/11/25 16:28:47 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
20/11/25 16:28:47 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
20/11/25 16:28:47 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
20/11/25 16:28:47 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
2020-11-25 16:28:47,938 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0.3.0.1.0-187 (rUnversioned directory) compiled Sep 19 2018, 10:13:33
2020-11-25 16:28:47,938 [main] INFO  org.apache.pig.Main - Logging error messages to: /root/pig_1606321727934.log
2020-11-25 16:28:48,054 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /root/.pigbootup not found
2020-11-25 16:28:48,472 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://sandbox-h
dp.hortonworks.com:8020
2020-11-25 16:28:49,320 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-b8672519-7d01-415d-9eec-b48741ef08b9
2020-11-25 16:28:49,320 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> movies = LOAD 'InputDir/Movielens.csv' using PigStorage(',') as (movieId:int, title:chararray, user_id:int, ratings: double, genreId: int,recommende
d: chararray,activitystate:int);
grunt> desc

desc       describe
grunt> describe  movies;
movies: {movieId: int,title: chararray,user_id: int,ratings: double,genreId: int,recommended: chararray,activitystate: int}
grunt>
```

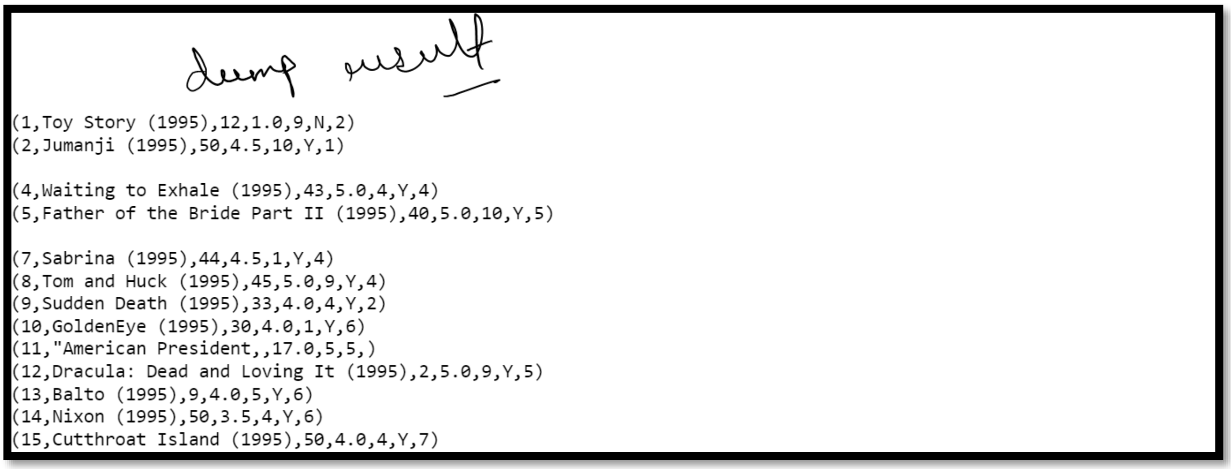


dump dataset

```
grunt> dump movies;
2020-11-25 16:34:39,816 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2020-11-25 16:34:39,919 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-11-25 16:34:39,986 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, Constan
tCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimi
zer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2020-11-25 16:34:40,110 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionU
sageThreshold = 489580128, usageThreshold = 489580128
2020-11-25 16:34:40,290 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Tez staging directory is /tmp/root/staging and resourc
es directory is /tmp/temp1142204722
2020-11-25 16:34:40,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.plan.TezCompiler - File concatenation threshold: 100 optimistic? fal
se
2020-11-25 16:34:40,689 [main] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2020-11-25 16:34:40,715 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 16:34:40,715 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2020-11-25 16:34:40,818 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2020-11-25 16:34:40,919 [main] INFO  org.apache.tez.mapreduce.hadoop.MRInputHelpers - NumSplits: 1, SerializedSize: 406
2020-11-25 16:34:43,137 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: pig-0.16.0.3.0.1.0-187-core-h2.jar
2020-11-25 16:34:43,137 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: antlr-runtime-3.4.jar
2020-11-25 16:34:43,137 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: automaton-1.11-8.jar
2020-11-25 16:34:43,137 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: joda-time-2.9.3.jar
2020-11-25 16:34:43,344 [main] INFO  org.apache.hadoop.conf.Configuration - found resource resource-types.xml at file:/etc/hadoop/3.0.1.0-187/0/resource-ty
pes.xml
2020-11-25 16:34:43,581 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - For vertex - scope-24: parallelism=1, memory=1024, j
ava opts=-XX:+PrintGCDetails -verbose:gc -XX:+PrintGCTimeStamps -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB
2020-11-25 16:34:43,582 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Processing aliases: movies
2020-11-25 16:34:43,582 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Detailed locations: movies[1,9],movies[-1,-1]
2020-11-25 16:34:43,582 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Pig features in the vertex:
2020-11-25 16:34:43,689 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Total estimated parallelism is 1
2020-11-25 16:34:43,760 [PigTezLauncher-0] INFO  org.apache.pig.tools.pigstats.tez.TezScriptState - Pig script settings are added to the job
2020-11-25 16:34:44,039 [PigTezLauncher-0] INFO  org.apache.tez.client.TezClient - Tez Client Version: [ component=tez-api, version=0.9.1.3.0.1.0-187, revi
sion=a546e73be6ee94bb0e672f5d361f7408dc1bf418, SCM-URL=scm:git:https://git-wip-us.apache.org/repos/asf/tez.git, buildTime=2018-09-19T10:13:49Z ]
2020-11-25 16:34:44,109 [PigTezLauncher-0] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at sandbox-hdp.hortonworks.com/172.1
8.0.2:8050
```



```
            Features: UNKNOWN
Success!


DAG 0:

                           ApplicationId: job_1606283825060_0007
                      TotalLaunchedTasks: 1

                        FileBytesWritten: 0
                           HdfsBytesRead: 42326

      SpillableMemoryManager spill count: 0

                 Bags proactively spilled: 0
             Records proactively spilled: 0

Tez vertex scope-24


VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias  F
eature  Outputs

dfs://sandbox-hdp.hortonworks.com:8020/tmp/temp-1908888540/tmp-2041774811,    989            0              0         42326          49121 movies h


Successfully read 989 records (42326 bytes) from: "hdfs://sandbox-hdp.hortonworks.com:8020/user/root/InputDir/Movielens.csv"

Output(s):
Successfully stored 989 records (49121 bytes) in: "hdfs://sandbox-hdp.hortonworks.com:8020/tmp/temp-1908888540/tmp-2041774811"

2020-11-25 16:38:36,626 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 16:38:36,626 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
```

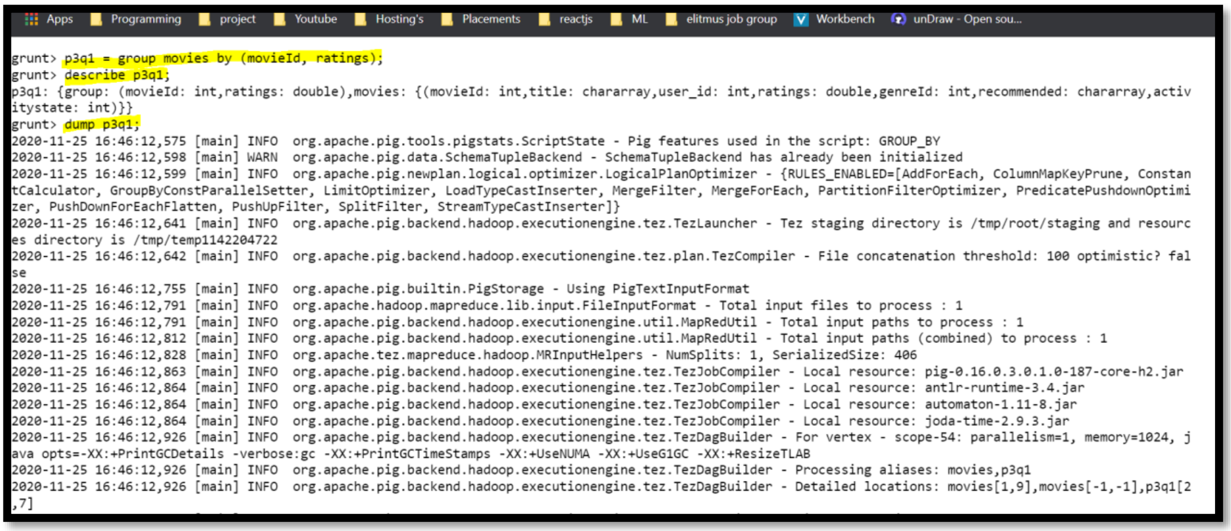


dump result

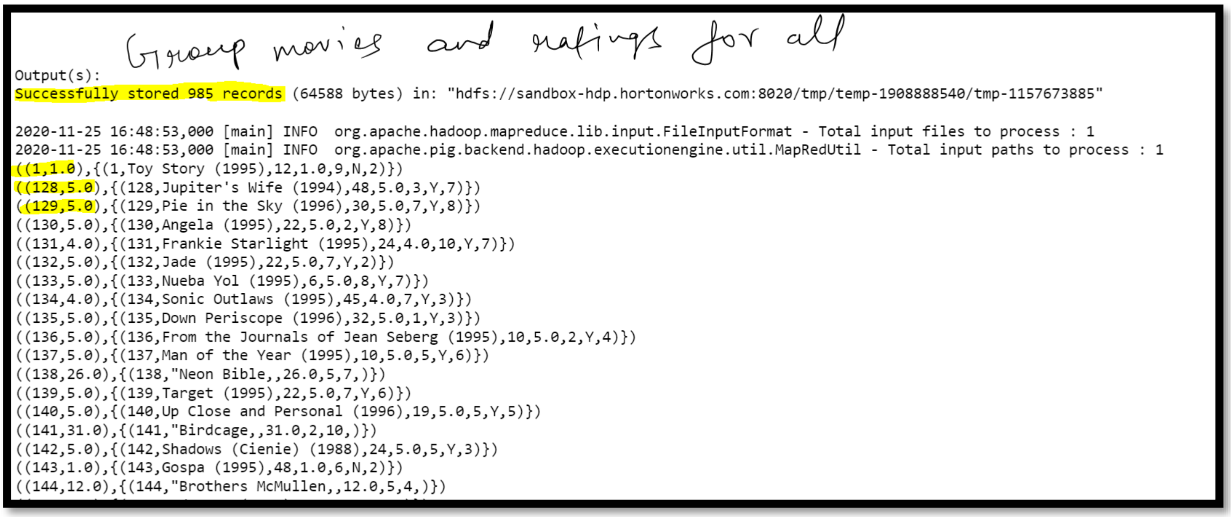
```
(1,Toy Story (1995),12,1.0,9,N,2)
(2,Jumanji (1995),50,4.5,10,Y,1)

(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)

(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)
```

1. List all the movies and number of ratings.

   >> p3q1 = group movie by (movieId, ratings);
   >> dump p3q1;



```
grunt> p3q1 = group movies by (movieId, ratings);
grunt> describe p3q1;
p3q1: {group: (movieId: int,ratings: double),movies: {(movieId: int,title: chararray,user_id: int,ratings: double,genreId: int,recommended: chararray,activ
itystate: int)}}
grunt> dump p3q1;
2020-11-25 16:46:12,575 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2020-11-25 16:46:12,598 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2020-11-25 16:46:12,599 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, Constan
tCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimi
zer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2020-11-25 16:46:12,641 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Tez staging directory is /tmp/root/staging and resourc
es directory is /tmp/temp1142204722
2020-11-25 16:46:12,642 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.plan.TezCompiler - File concatenation threshold: 100 optimistic? fal
se
2020-11-25 16:46:12,755 [main] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2020-11-25 16:46:12,791 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 16:46:12,791 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2020-11-25 16:46:12,812 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2020-11-25 16:46:12,828 [main] INFO  org.apache.tez.mapreduce.hadoop.MRInputHelpers - NumSplits: 1, SerializedSize: 406
2020-11-25 16:46:12,863 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: pig-0.16.0.3.0.1.0-187-core-h2.jar
2020-11-25 16:46:12,864 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: antlr-runtime-3.4.jar
2020-11-25 16:46:12,864 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: automaton-1.11-8.jar
2020-11-25 16:46:12,864 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: joda-time-2.9.3.jar
2020-11-25 16:46:12,926 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - For vertex - scope-54: parallelism=1, memory=1024, j
ava opts=-XX:+PrintGCDetails -verbose:gc -XX:+PrintGCTimeStamps -XX:+UseNUMA -XX:+UseG1GC -XX:+ResizeTLAB
2020-11-25 16:46:12,926 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Processing aliases: movies,p3q1
2020-11-25 16:46:12,926 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezDagBuilder - Detailed locations: movies[1,9],movies[-1,-1],p3q1[2
,7]
```



Group movies and ratings for all

```
Output(s):
Successfully stored 985 records (64588 bytes) in: "hdfs://sandbox-hdp.hortonworks.com:8020/tmp/temp-1908888540/tmp-1157673885"

2020-11-25 16:48:53,000 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 16:48:53,000 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((1,1.0),{(1,Toy Story (1995),12,1.0,9,N,2)})
((128,5.0),{(128,Jupiter's Wife (1994),48,5.0,3,Y,7)})
((129,5.0),{(129,Pie in the Sky (1996),30,5.0,7,Y,8)})
((130,5.0),{(130,Angela (1995),22,5.0,2,Y,8)})
((131,4.0),{(131,Frankie Starlight (1995),24,4.0,10,Y,7)})
((132,5.0),{(132,Jade (1995),22,5.0,7,Y,2)})
((133,5.0),{(133,Nueba Yol (1995),6,5.0,8,Y,7)})
((134,4.0),{(134,Sonic Outlaws (1995),45,4.0,7,Y,3)})
((135,5.0),{(135,Down Periscope (1996),32,5.0,1,Y,3)})
((136,5.0),{(136,From the Journals of Jean Seberg (1995),10,5.0,2,Y,4)})
((137,5.0),{(137,Man of the Year (1995),10,5.0,5,Y,6)})
((138,26.0),{(138,"Neon Bible,,26.0,5,7,)})
((139,5.0),{(139,Target (1995),22,5.0,7,Y,6)})
((140,5.0),{(140,Up Close and Personal (1996),19,5.0,5,Y,5)})
((141,31.0),{(141,"Birdcage,,31.0,2,10,)})
((142,5.0),{(142,Shadows (Cienie) (1988),24,5.0,5,Y,3)})
((143,1.0),{(143,Gospa (1995),48,1.0,6,N,2)})
((144,12.0),{(144,"Brothers McMullen,,12.0,5,4,)})
```

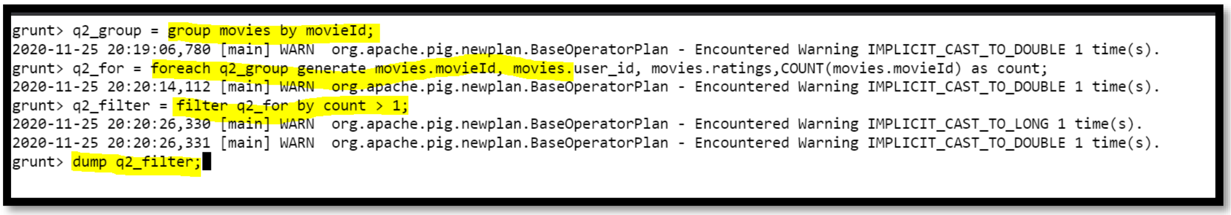
2. List all the users who have rated the same movie and find the number of ratings

   >> q2_group = group movie by movieId;

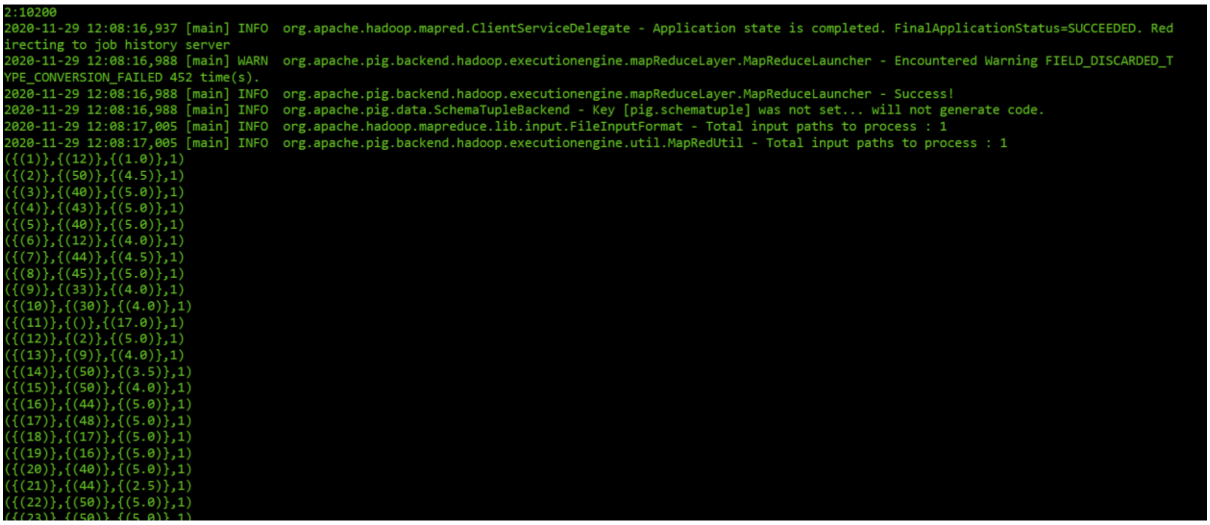
   >> q2_for = foreach q2_group generate movies.movieId, movies.user_id, movies.ratings, COUNT(movies.movieId) as count;

   >> q2_filter = filter q2_for by count > 1;

   >> dump q2_filter;



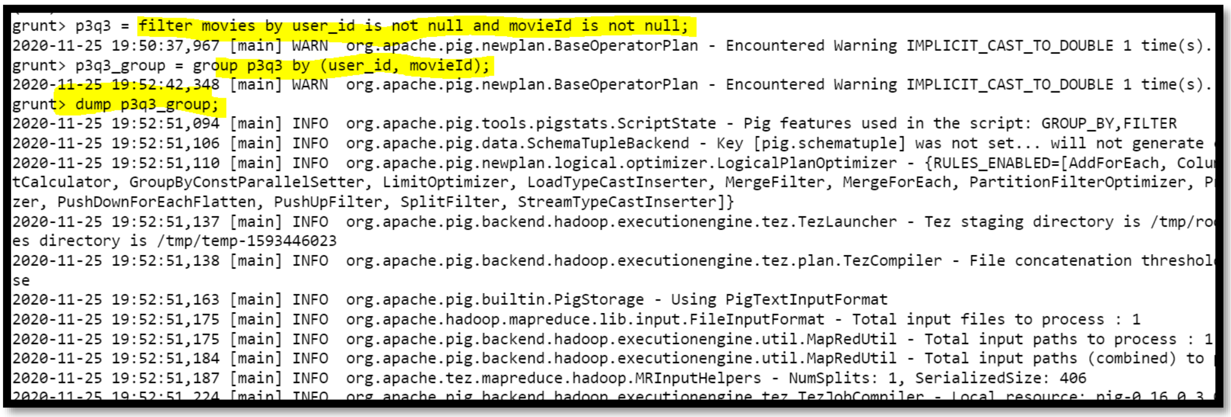
```
grunt> q2_group = group movies by movieId;
2020-11-25 20:19:06,780 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> q2_for = foreach q2_group generate movies.movieId, movies.user_id, movies.ratings,COUNT(movies.movieId) as count;
2020-11-25 20:20:14,112 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> q2_filter = filter q2_for by count > 1;
2020-11-25 20:20:26,330 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
2020-11-25 20:20:26,331 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> dump q2_filter;
```



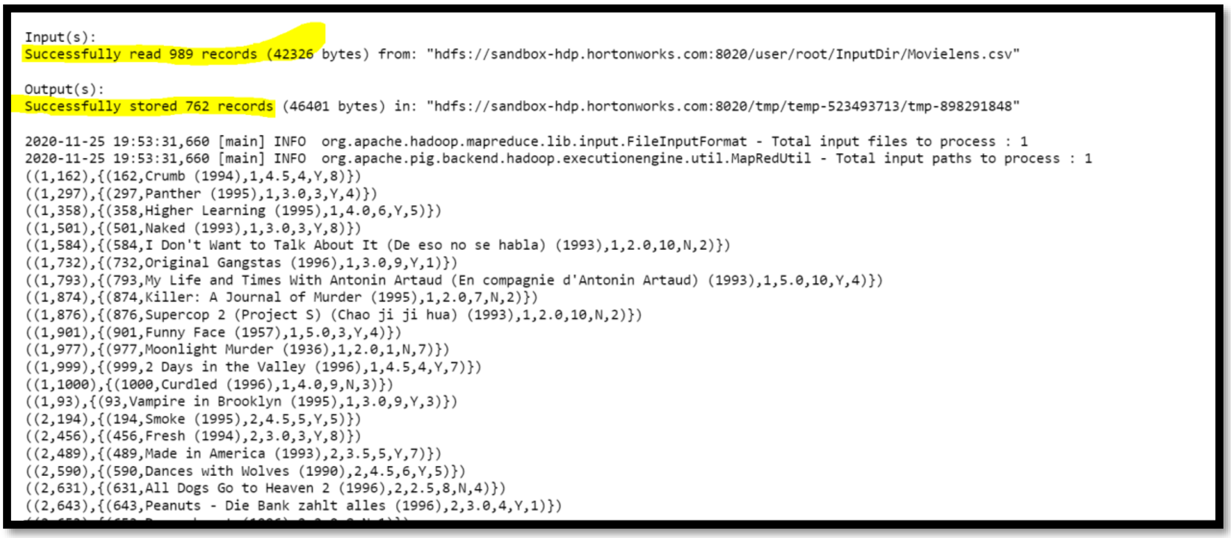
```
2:10200
2020-11-29 12:08:16,937 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Red
irecting to job history server
2020-11-29 12:08:16,988 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_T
YPE_CONVERSION_FAILED 452 time(s).
2020-11-29 12:08:16,988 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-29 12:08:16,988 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-11-29 12:08:17,005 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2020-11-29 12:08:17,005 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
({(1)},{(12)},{(1.0)},1)
({(2)},{(50)},{(4.5)},1)
({(3)},{(40)},{(5.0)},1)
({(4)},{(43)},{(5.0)},1)
({(5)},{(40)},{(5.0)},1)
({(6)},{(12)},{(4.0)},1)
({(7)},{(44)},{(4.5)},1)
({(8)},{(45)},{(5.0)},1)
({(9)},{(33)},{(4.0)},1)
({(10)},{(30)},{(4.0)},1)
({(11)},{()},{(17.0)},1)
({(12)},{(2)},{(5.0)},1)
({(13)},{(9)},{(4.0)},1)
({(14)},{(50)},{(3.5)},1)
({(15)},{(50)},{(4.0)},1)
({(16)},{(44)},{(5.0)},1)
({(17)},{(48)},{(5.0)},1)
({(18)},{(17)},{(5.0)},1)
({(19)},{(16)},{(5.0)},1)
({(20)},{(40)},{(5.0)},1)
({(21)},{(44)},{(2.5)},1)
({(22)},{(50)},{(5.0)},1)
({(23)},{(50)},{(5.0)},1)
```

3. List all the users who have rated the movies (Users who have rated at least one movie)

   >> p3q3 = filter movie by user_id is not null and movieId is not null;
   >> p3q3_group = group p3q3 by (user_id,movieId);
   >> dump p3q3;



```
grunt> p3q3 = filter movies by user_id is not null and movieId is not null;
2020-11-25 19:50:37,967 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> p3q3_group = group p3q3 by (user_id, movieId);
2020-11-25 19:52:42,348 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt> dump p3q3_group;
2020-11-25 19:52:51,094 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,FILTER
2020-11-25 19:52:51,106 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate
2020-11-25 19:52:51,110 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, Colu
tCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, P
zer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2020-11-25 19:52:51,137 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher - Tez staging directory is /tmp/ro
es directory is /tmp/temp-1593446023
2020-11-25 19:52:51,138 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.plan.TezCompiler - File concatenation threshol
se
2020-11-25 19:52:51,163 [main] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2020-11-25 19:52:51,175 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 19:52:51,175 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2020-11-25 19:52:51,184 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to
2020-11-25 19:52:51,187 [main] INFO  org.apache.tez.mapreduce.hadoop.MRInputHelpers - NumSplits: 1, SerializedSize: 406
2020-11-25 19:52:51,224 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezJobCompiler - Local resource: pig-0.16.0.3
```



```
Input(s):
Successfully read 989 records (42326 bytes) from: "hdfs://sandbox-hdp.hortonworks.com:8020/user/root/InputDir/Movielens.csv"

Output(s):
Successfully stored 762 records (46401 bytes) in: "hdfs://sandbox-hdp.hortonworks.com:8020/tmp/temp-523493713/tmp-898291848"

2020-11-25 19:53:31,660 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 19:53:31,660 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((1,162),{(162,Crumb (1994),1,4.5,4,Y,8)})
((1,297),{(297,Panther (1995),1,3.0,3,Y,4)})
((1,358),{(358,Higher Learning (1995),1,4.0,6,Y,5)})
((1,501),{(501,Naked (1993),1,3.0,3,Y,8)})
((1,584),{(584,I Don't Want to Talk About It (De eso no se habla) (1993),1,2.0,10,N,2)})
((1,732),{(732,Original Gangstas (1996),1,3.0,9,Y,1)})
((1,793),{(793,My Life and Times With Antonin Artaud (En compagnie d'Antonin Artaud) (1993),1,5.0,10,Y,4)})
((1,874),{(874,Killer: A Journal of Murder (1995),1,2.0,7,N,2)})
((1,876),{(876,Supercop 2 (Project S) (Chao ji ji hua) (1993),1,2.0,10,N,2)})
((1,901),{(901,Funny Face (1957),1,5.0,3,Y,4)})
((1,977),{(977,Moonlight Murder (1936),1,2.0,1,N,7)})
((1,999),{(999,2 Days in the Valley (1996),1,4.5,4,Y,7)})
((1,1000),{(1000,Curdled (1996),1,4.0,9,N,3)})
((1,93),{(93,Vampire in Brooklyn (1995),1,3.0,9,Y,3)})
((2,194),{(194,Smoke (1995),2,4.5,5,Y,5)})
((2,456),{(456,Fresh (1994),2,3.0,3,Y,8)})
((2,489),{(489,Made in America (1993),2,3.5,5,Y,7)})
((2,590),{(590,Dances with Wolves (1990),2,4.5,6,Y,5)})
((2,631),{(631,All Dogs Go to Heaven 2 (1996),2,2.5,8,N,4)})
((2,643),{(643,Peanuts - Die Bank zahlt alles (1996),2,3.0,4,Y,1)})
```

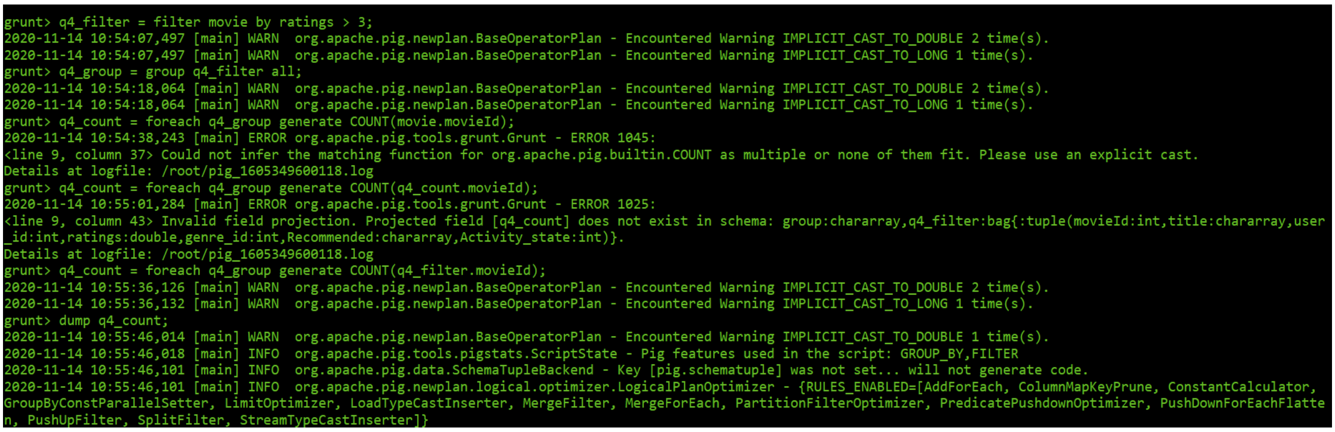
4. Find the count of the movies which have ratings more than 3

   >> q4_filter = filter movie by ratings > 3;

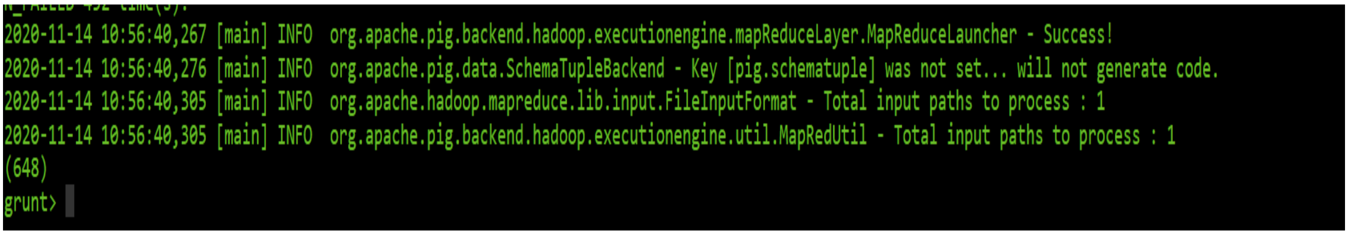
   >> q4_group = group q4_filter all;

   >> q4_count = foreach q4_group generate COUNT(q4_filter.movieId);

   >> dump q4_count;



```
grunt> q4_filter = filter movie by ratings > 3;
2020-11-14 10:54:07,497 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2 time(s).
2020-11-14 10:54:07,497 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
grunt> q4_group = group q4_filter all;
2020-11-14 10:54:18,064 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2 time(s).
2020-11-14 10:54:18,064 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
grunt> q4_count = foreach q4_group generate COUNT(movie.movieId);
2020-11-14 10:54:38,243 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1045:
<line 9, column 37> Could not infer the matching function for org.apache.pig.builtin.COUNT as multiple or none of them fit. Please use an explicit cast.
Details at logfile: /root/pig_1605349600118.log
grunt> q4_count = foreach q4_group generate COUNT(q4_count.movieId);
2020-11-14 10:55:01,284 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1025:
<line 9, column 43> Invalid field projection. Projected field [q4_count] does not exist in schema: group:chararray,q4_filter:bag{:tuple(movieId:int,title:chararray,user
_id:int,ratings:double,genre_id:int,Recommended:chararray,Activity_state:int)}.
Details at logfile: /root/pig_1605349600118.log
grunt> q4_count = foreach q4_group generate COUNT(q4_filter.movieId);
2020-11-14 10:55:36,126 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2 time(s).
2020-11-14 10:55:36,132 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 1 time(s).
grunt> dump q4_count;
2020-11-14 10:55:46,014 [main] WARN  org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
2020-11-14 10:55:46,018 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,FILTER
2020-11-14 10:55:46,101 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-11-14 10:55:46,101 [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]}
```



```
N_FAILED 492 time(s).
2020-11-14 10:56:40,267 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-14 10:56:40,276 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2020-11-14 10:56:40,305 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2020-11-14 10:56:40,305 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(648)
grunt>
```

5. Find the max, min and average ratings for all the movie

   >> q5_group = group movie all;

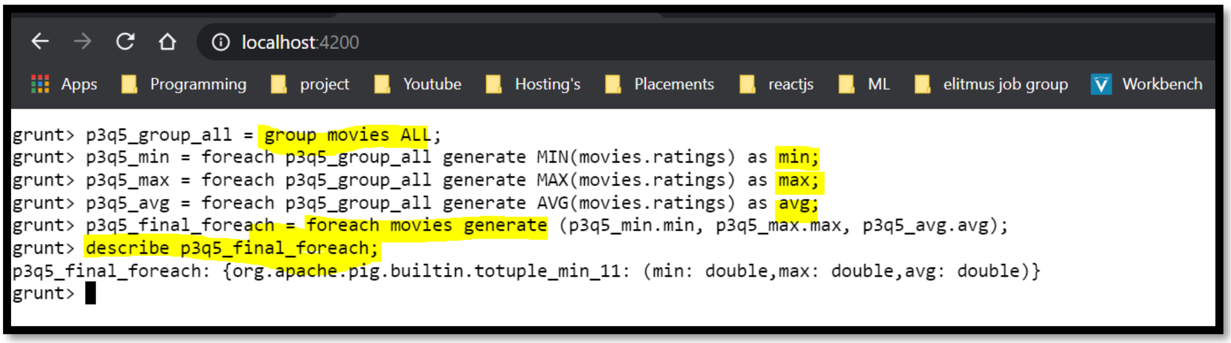
   >> q5_max = foreach q5_group generate MAX(movie.ratings) as max;

   >> q5_avg = foreach q5_group generate AVG(movie.ratings) as avg;

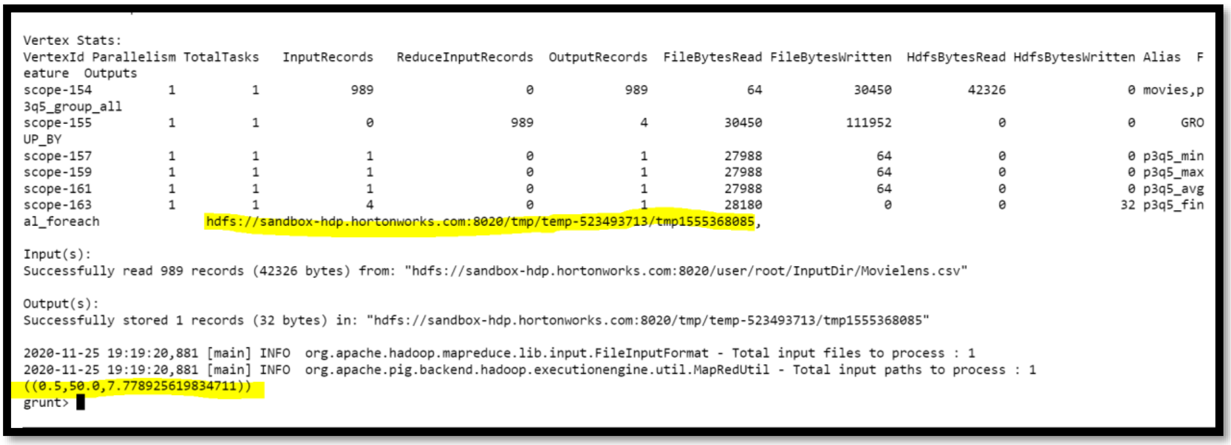
   >> q5_min = foreach q5_group generate MIN(movie.ratings) as min;

   >> q5_all = foreach movies generate (q5_max.max, q5_avg.avg, q5_min.min);

   >> dump q5_all;



```
grunt> p3q5_group_all = group movies ALL;
grunt> p3q5_min = foreach p3q5_group_all generate MIN(movies.ratings) as min;
grunt> p3q5_max = foreach p3q5_group_all generate MAX(movies.ratings) as max;
grunt> p3q5_avg = foreach p3q5_group_all generate AVG(movies.ratings) as avg;
grunt> p3q5_final_foreach = foreach movies generate (p3q5_min.min, p3q5_max.max, p3q5_avg.avg);
grunt> describe p3q5_final_foreach;
p3q5_final_foreach: {org.apache.pig.builtin.totuple_min_11: (min: double,max: double,avg: double)}
grunt>
```



```
Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias  F
eature  Outputs
scope-154          1          1            989                    0            989             64            30450          42326                0 movies,p
3q5_group_all
scope-155          1          1              0                  989              4          30450           111952              0                0      GRO
UP_BY
scope-157          1          1              1                    0              1          27988               64              0                0 p3q5_min
scope-159          1          1              1                    0              1          27988               64              0                0 p3q5_max
scope-161          1          1              1                    0              1          27988               64              0                0 p3q5_avg
scope-163          1          1              4                    0              1          28180                0              0               32 p3q5_fin
al_foreach             hdfs://sandbox-hdp.hortonworks.com:8020/tmp/temp-523493713/tmp1555368085,

Input(s):
Successfully read 989 records (42326 bytes) from: "hdfs://sandbox-hdp.hortonworks.com:8020/user/root/InputDir/Movielens.csv"

Output(s):
Successfully stored 1 records (32 bytes) in: "hdfs://sandbox-hdp.hortonworks.com:8020/tmp/temp-523493713/tmp1555368085"

2020-11-25 19:19:20,881 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2020-11-25 19:19:20,881 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
((0.5,50.0,7.778925619834711))
grunt>
```