Contents

Query1.pig

```
-- Format:
--{follower_twitter_account_id: int,subject_twitter_account_id: int}
follows_account = LOAD '/oracle/stuhomes/yu1357/hk4/HW4-follows_account.txt'
using PigStorage(',') AS (follower_twitter_account_id:
int,subject_twitter_account_id: int);


-- Format:
--{twitter_account_id: int,email_address: chararray,phone_number:
chararray,user_location: chararray,num_tweets: int}
twitter_account = LOAD '/oracle/stuhomes/yu1357/hk4/HW4-twitter_account.txt'
USING PigStorage(',') AS (twitter_account_id: int, email_address: chararray ,
phone_number: chararray , user_location: chararray , num_tweets: int);

-- Format:
--{twitter_account_id: int,twitter_account_rank: float}
old_twitter_account_rank = LOAD '/oracle/stuhomes/yu1357/hk4/HW4-
old_twitter_account_rank.txt' using PigStorage(',') AS (twitter_account_id:int,
twitter_account_rank:float);

-- Format:
--{email_address: chararray,reputation: int,num_questions: int}
stack_overflow_account = LOAD '/oracle/stuhomes/yu1357/hk4/HW4-
stack_overflow_account.txt' using PigStorage(',') AS (email_address:chararray,
reputation: int, num_questions: int);



--query_1a 1st test
---------------------------------------------------------------------------------
-
-- Format
-- {twitter_account_id: int,email_address: chararray,phone_number:
chararray,user_location: chararray,num_tweets: int}
filter_phone = FILTER twitter_account by STARTSWITH(phone_number, '430');
DESCRIBE filter_phone;

-- Format
-- {group: chararray,filter_phone: {(twitter_account_id: int,email_address:
chararray,phone_number: chararray,user_location: chararray,num_tweets: int)}}
grouped_phone = GROUP filter_phone BY SUBSTRING(phone_number, 0, 3);
DESCRIBE grouped;
```

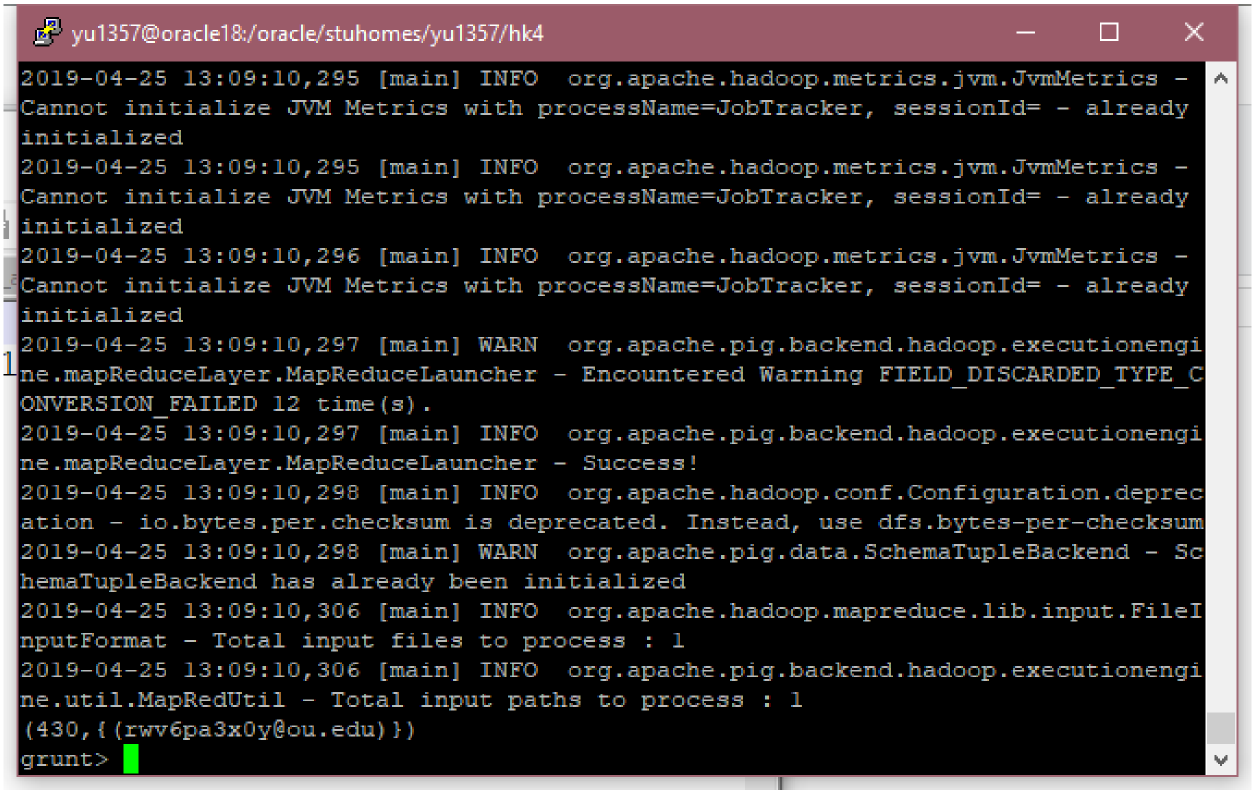
```
-- Format
-- {group: chararray,{(email_address: chararray)}}
result_a1 = FOREACH grouped_phone GENERATE group, $1.email_address;
DESCRIBE result_a1;

STORE result_a1 INTO query_1a1;
dump result_a1;


--query_1a 2rd test
----------------------------------------------------------------------------------
-
-- Format
-- {twitter_account_id: int,email_address: chararray,phone_number:
chararray,user_location: chararray,num_tweets: int}
filter_phone = FILTER twitter_account by STARTSWITH(phone_number, '405');
DESCRIBE filter_phone;

-- Format
-- {group: chararray,filter_phone: {(twitter_account_id: int,email_address:
chararray,phone_number: chararray,user_location: chararray,num_tweets: int)}}
grouped_phone = GROUP filter_phone BY SUBSTRING(phone_number, 0, 3);
DESCRIBE grouped;


-- Format
-- {group: chararray,{(email_address: chararray)}}
result_a2 = FOREACH grouped_phone GENERATE group, $1.email_address;
DESCRIBE result_a2;

STORE result_a2 INTO query_1a2;
dump result_a2;


--query_1a 3rd test
----------------------------------------------------------------------------------
-
-- Format
-- {twitter_account_id: int,email_address: chararray,phone_number:
chararray,user_location: chararray,num_tweets: int}
filter_phone = FILTER twitter_account by STARTSWITH(phone_number, '555');
DESCRIBE filter_phone;

-- Format
-- {group: chararray,filter_phone: {(twitter_account_id: int,email_address:
chararray,phone_number: chararray,user_location: chararray,num_tweets: int)}}
grouped_phone = GROUP filter_phone BY SUBSTRING(phone_number, 0, 3);
DESCRIBE grouped;


-- Format
-- {group: chararray,{(email_address: chararray)}}
result_a3 = FOREACH grouped_phone GENERATE group, $1.email_address;
DESCRIBE result_a3;

STORE result_a3 INTO query_1a3;
dump result_a3;


--query_1b
---------------------------------------------------------------------------------
-
--Format
--{group: chararray,twitter_account: {(twitter_account_id: int,email_address:
chararray,phone_number: chararray,user_location: chararray,num_tweets: int)}}
group_city = GROUP twitter_account BY user_location;


--Format
--result_b = FOREACH group_city GENERATE group AS user_location, COUNT($1) AS
num_acc;
result_b = FOREACH group_city GENERATE group AS user_location, COUNT($1) AS
num_acc;

DESCRIBE result_b;

STORE result_b INTO query_1b;
dump result_b;

--query_1c  1st test
---------------------------------------------------------------------------------
-
--Format
--{twitter_account::twitter_account_id: int,twitter_account::email_address:
chararray,twitter_account::phone_number:
chararray,twitter_account::user_location: chararray,twitter_account::num_tweets:
int,stack_overflow_account::email_address:
chararray,stack_overflow_account::reputation:
int,stack_overflow_account::num_questions: int}
```

```
joined = JOIN twitter_account BY email_address, stack_overflow_account by
email_address;

--Format
--{twitter_account::twitter_account_id: int,twitter_account::email_address:
chararray,twitter_account::phone_number:
chararray,twitter_account::user_location: chararray,twitter_account::num_tweets:
int,stack_overflow_account::email_address:
chararray,stack_overflow_account::reputation:
int,stack_overflow_account::num_questions: int}
filtered_num_tweets = FILTER joined BY num_tweets > 666;

--Format
--{twitter_account::num_tweets: int,stack_overflow_account::reputation: int}
email_reputations = FOREACH filtered_num_tweets GENERATE
twitter_account::num_tweets, stack_overflow_account::reputation;

--Format
--{group: chararray,email_reputations: {(twitter_account::num_tweets:
int,stack_overflow_account::reputation: int)}}
email_reputations_group = GROUP email_reputations ALL;

--Format
--{num_tweets: long,avg_stack_overflow_reputation: double}
result_c1 = FOREACH email_reputations_group GENERATE SUM($1.$0) AS num_tweets,
AVG($1.$1) AS avg_stack_overflow_reputation;


DESCRIBE result_c1;
STORE result_c1 INTO query_1c;
dump result_c;


--query_1c  2rd test
---------------------------------------------------------------------------------
-
--Format
--{twitter_account::twitter_account_id: int,twitter_account::email_address:
chararray,twitter_account::phone_number:
chararray,twitter_account::user_location: chararray,twitter_account::num_tweets:
int,stack_overflow_account::email_address:
chararray,stack_overflow_account::reputation:
int,stack_overflow_account::num_questions: int}
joined = JOIN twitter_account BY email_address, stack_overflow_account by
email_address;
```

```
--Format
--{twitter_account::twitter_account_id: int,twitter_account::email_address:
chararray,twitter_account::phone_number:
chararray,twitter_account::user_location: chararray,twitter_account::num_tweets:
int,stack_overflow_account::email_address:
chararray,stack_overflow_account::reputation:
int,stack_overflow_account::num_questions: int}
filtered_num_tweets = FILTER joined BY num_tweets > 666;

--Format
--{twitter_account::num_tweets: int,stack_overflow_account::reputation: int}
email_reputations = FOREACH filtered_num_tweets GENERATE
twitter_account::num_tweets, stack_overflow_account::reputation;

--Format
--{group: chararray,email_reputations: {(twitter_account::num_tweets:
int,stack_overflow_account::reputation: int)}}
email_reputations_group = GROUP email_reputations ALL;

--Format
--{num_tweets: long,avg_stack_overflow_reputation: double}
result_c2 = FOREACH email_reputations_group GENERATE SUM($1.$0) AS num_tweets,
AVG($1.$1) AS avg_stack_overflow_reputation;


DESCRIBE result_c2;
STORE result_c2 INTO query_1c2;
dump result_c2;


--query_1c  2rd test
---------------------------------------------------------------------------------
-
--Format
--{twitter_account::twitter_account_id: int,twitter_account::email_address:
chararray,twitter_account::phone_number:
chararray,twitter_account::user_location: chararray,twitter_account::num_tweets:
int,stack_overflow_account::email_address:
chararray,stack_overflow_account::reputation:
int,stack_overflow_account::num_questions: int}
joined = JOIN twitter_account BY email_address, stack_overflow_account by
email_address;

--Format
```

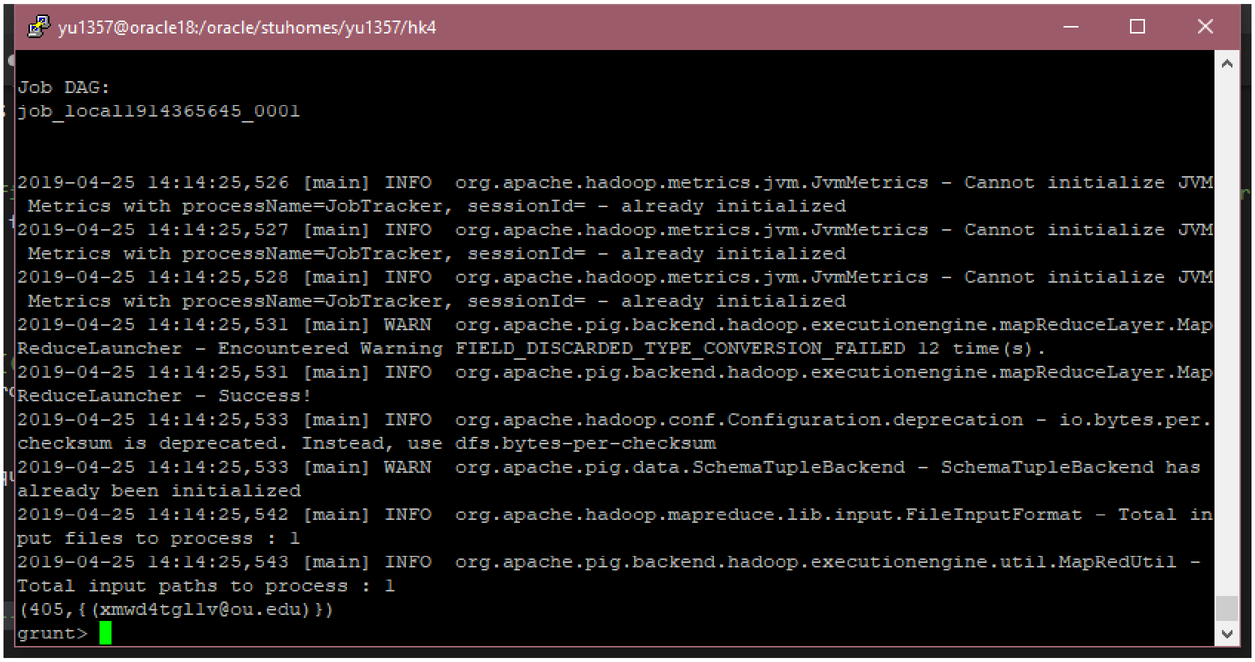
```
--{twitter_account::twitter_account_id: int,twitter_account::email_address:
chararray,twitter_account::phone_number:
chararray,twitter_account::user_location: chararray,twitter_account::num_tweets:
int,stack_overflow_account::email_address:
chararray,stack_overflow_account::reputation:
int,stack_overflow_account::num_questions: int}
filtered_num_tweets = FILTER joined BY num_tweets > 666;

--Format
--{twitter_account::num_tweets: int,stack_overflow_account::reputation: int}
email_reputations = FOREACH filtered_num_tweets GENERATE
twitter_account::num_tweets, stack_overflow_account::reputation;

--Format
--{group: chararray,email_reputations: {(twitter_account::num_tweets:
int,stack_overflow_account::reputation: int)}}
email_reputations_group = GROUP email_reputations ALL;

--Format
--{num_tweets: long,avg_stack_overflow_reputation: double}
result_c3 = FOREACH email_reputations_group GENERATE SUM($1.$0) AS num_tweets,
AVG($1.$1) AS avg_stack_overflow_reputation;


DESCRIBE result_c3;
STORE result_c3 INTO query_1c3;
dump result_c3;
```

Query_1a output test 1 Input: 430



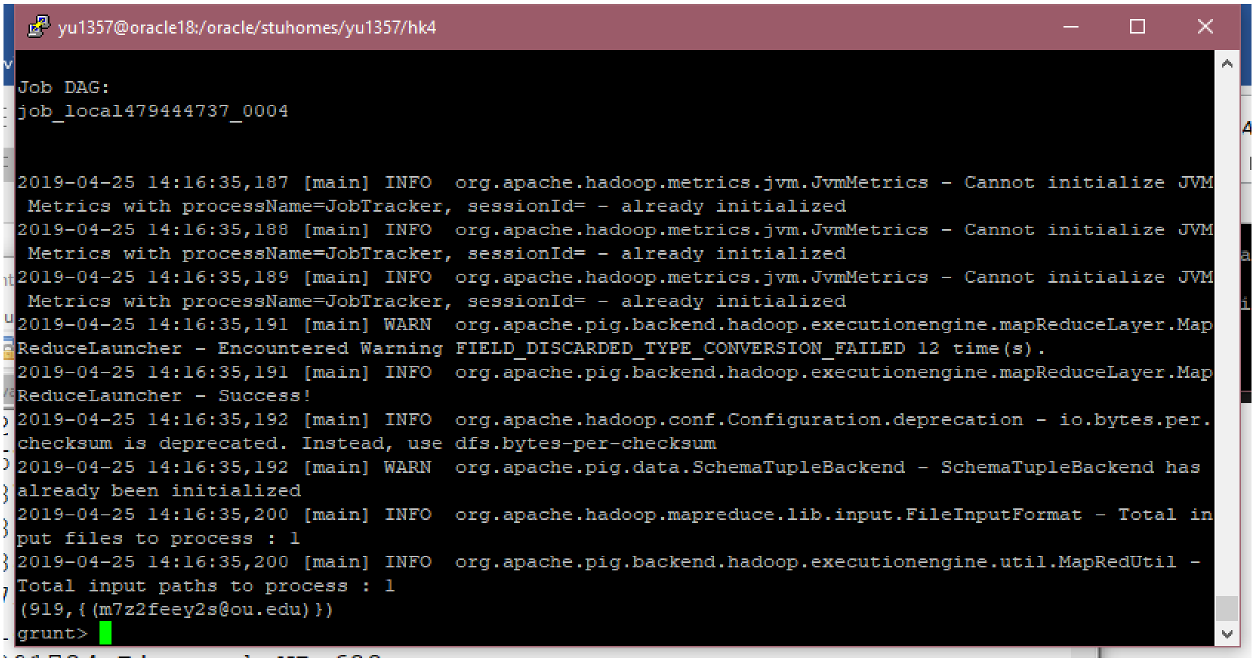
```
2019-04-25 13:09:10,295 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-04-25 13:09:10,295 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-04-25 13:09:10,296 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-04-25 13:09:10,297 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_C
ONVERSION_FAILED 12 time(s).
2019-04-25 13:09:10,297 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2019-04-25 13:09:10,298 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-25 13:09:10,298 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2019-04-25 13:09:10,306 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input files to process : 1
2019-04-25 13:09:10,306 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(430,{(rwv6pa3x0y@ou.edu)})
grunt>
```

Query_1a output test 2 input: 405



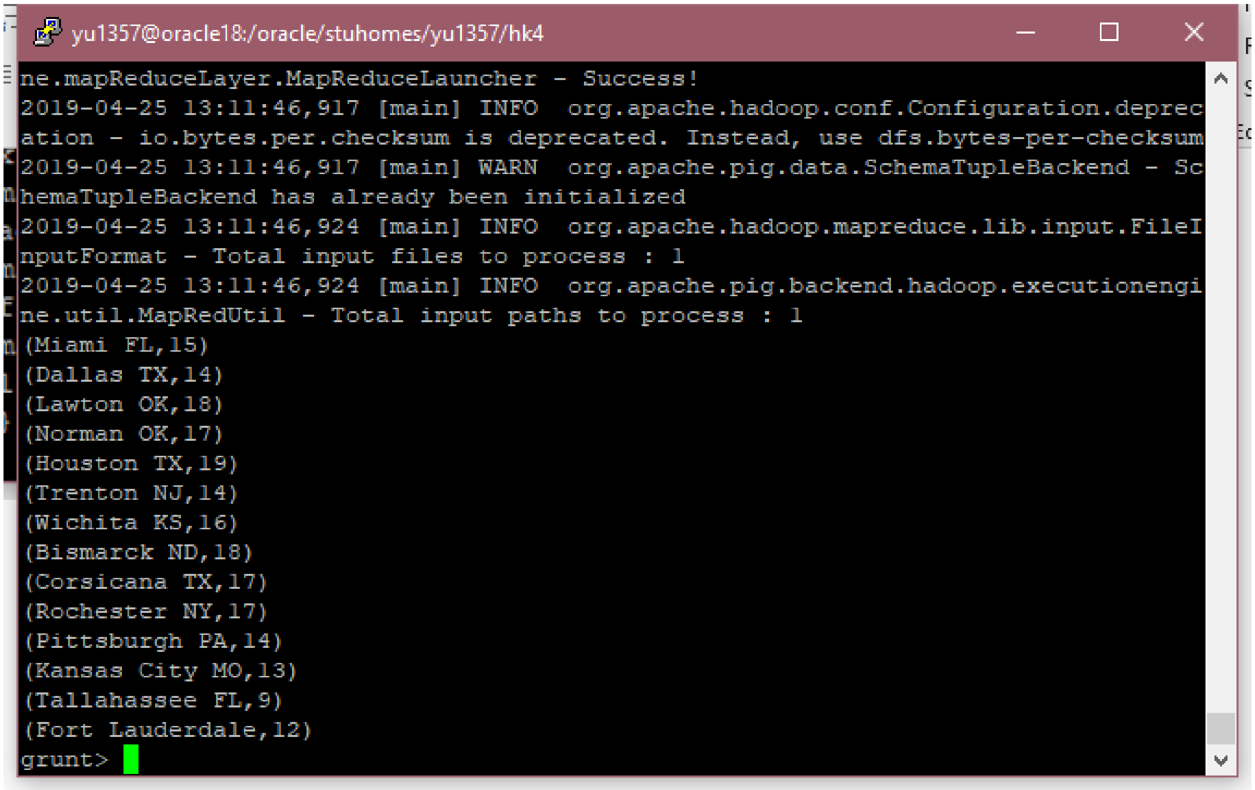
```
Job DAG:
job_local1914365645_0001


2019-04-25 14:14:25,526 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:14:25,527 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:14:25,528 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:14:25,531 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 12 time(s).
2019-04-25 14:14:25,531 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Success!
2019-04-25 14:14:25,533 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.
checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-25 14:14:25,533 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has
already been initialized
2019-04-25 14:14:25,542 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
put files to process : 1
2019-04-25 14:14:25,543 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
(405,{(xmwd4tgllv@ou.edu)})
grunt>
```

Query_1a output test 3 input: 919



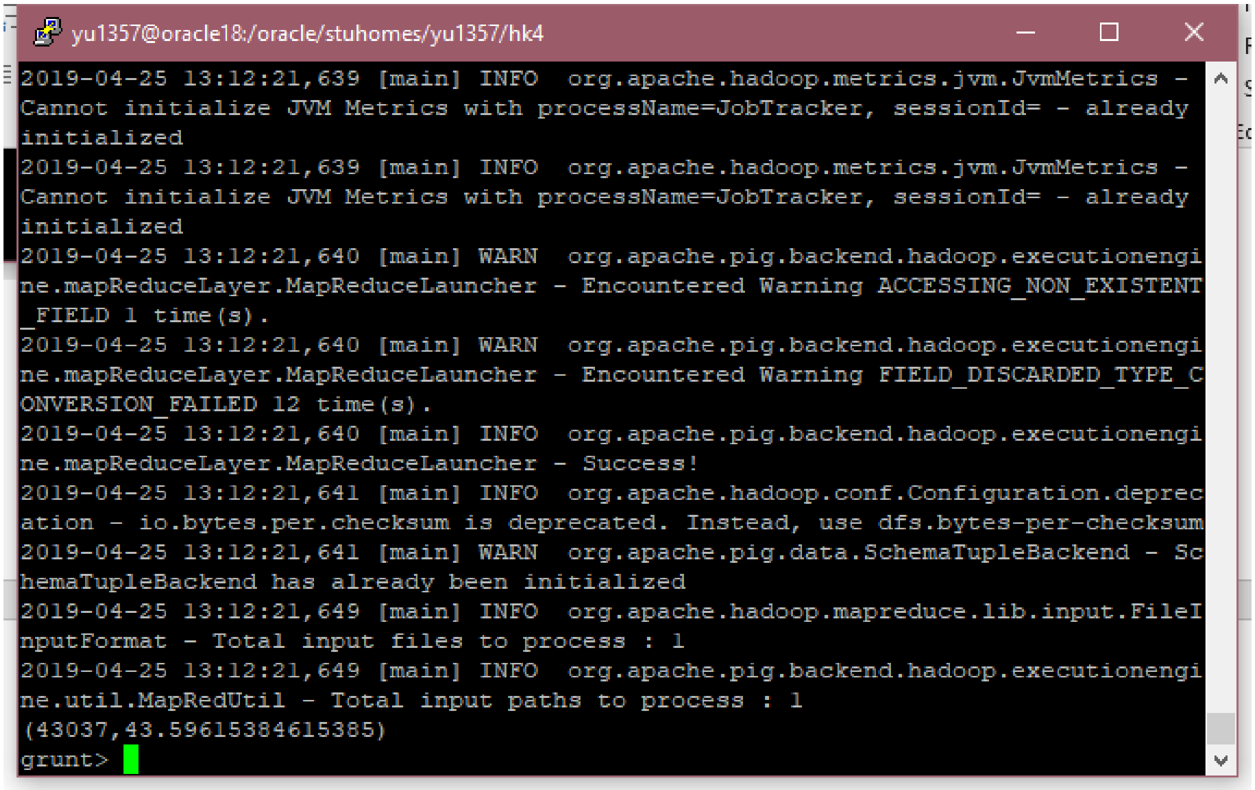
```
Job DAG:
job_local479444737_0004


2019-04-25 14:16:35,187 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:16:35,188 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:16:35,189 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:16:35,191 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 12 time(s).
2019-04-25 14:16:35,191 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Success!
2019-04-25 14:16:35,192 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.
checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-25 14:16:35,192 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has
already been initialized
2019-04-25 14:16:35,200 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
put files to process : 1
2019-04-25 14:16:35,200 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
(919,{(m7z2feey2s@ou.edu)})
grunt>
```

Query_1b output:



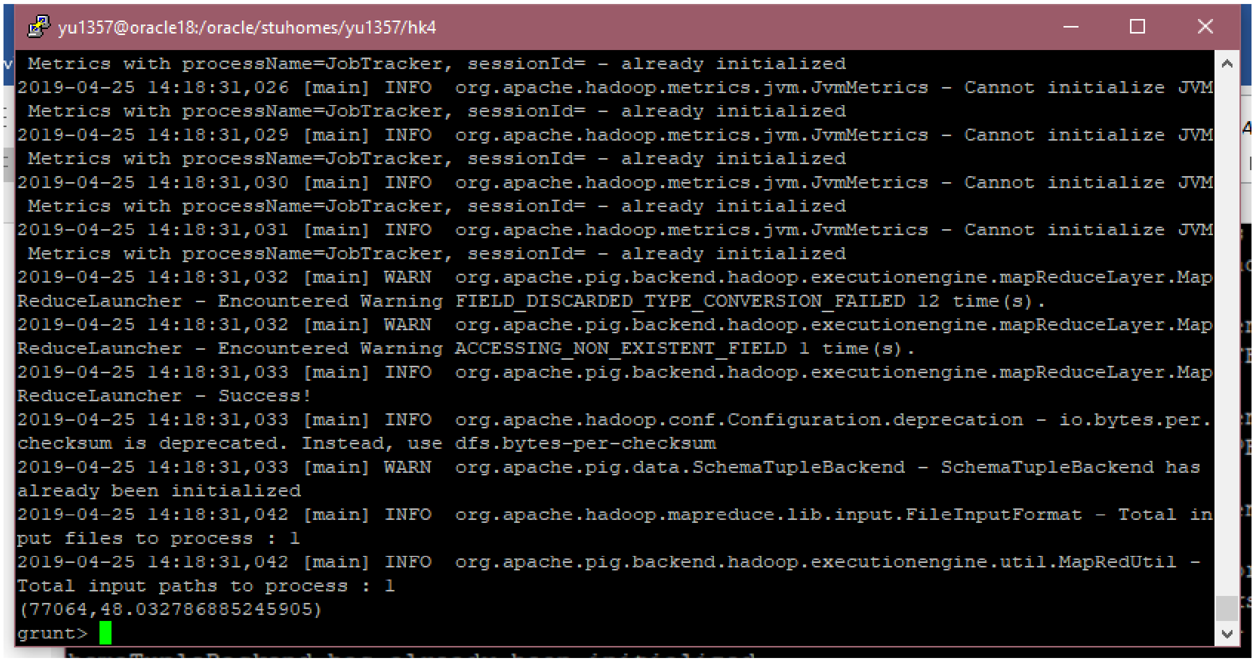
```
ne.mapReduceLayer.MapReduceLauncher - Success!
2019-04-25 13:11:46,917 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-25 13:11:46,917 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2019-04-25 13:11:46,924 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input files to process : 1
2019-04-25 13:11:46,924 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Miami FL,15)
(Dallas TX,14)
(Lawton OK,18)
(Norman OK,17)
(Houston TX,19)
(Trenton NJ,14)
(Wichita KS,16)
(Bismarck ND,18)
(Corsicana TX,17)
(Rochester NY,17)
(Pittsburgh PA,14)
(Kansas City MO,13)
(Tallahassee FL,9)
(Fort Lauderdale,12)
grunt>
```

Query_1c test 1 input: 666



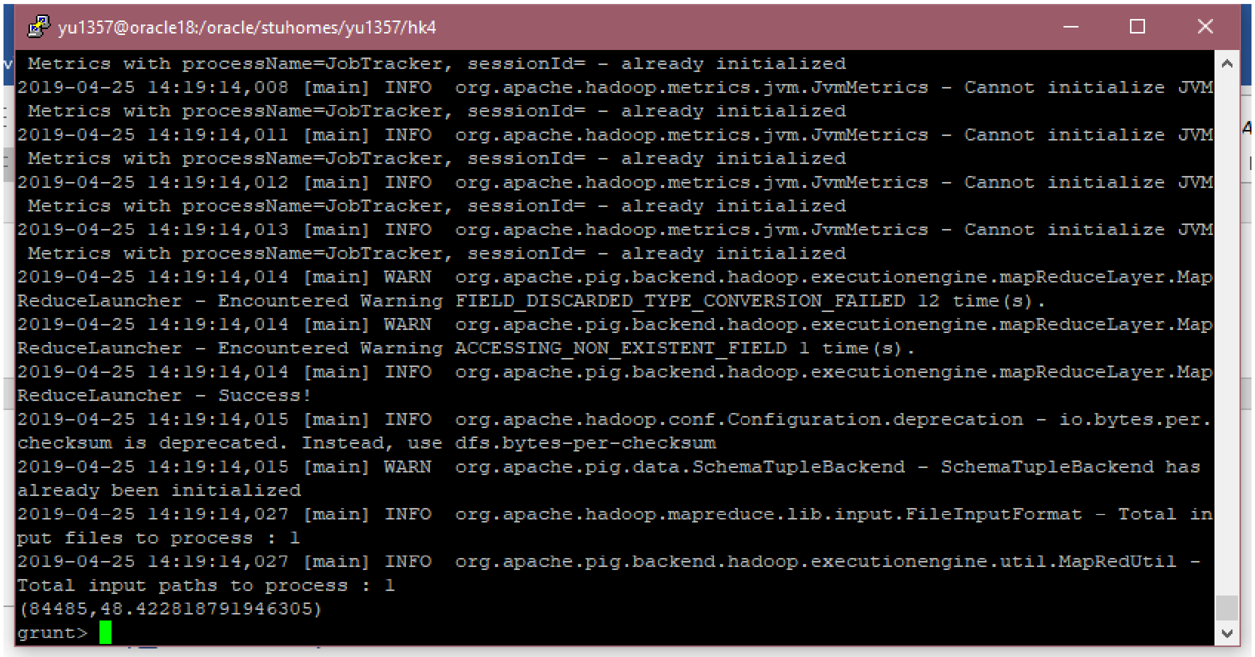
```
2019-04-25 13:12:21,639 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-04-25 13:12:21,639 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics -
Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2019-04-25 13:12:21,640 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT
_FIELD 1 time(s).
2019-04-25 13:12:21,640 [main] WARN  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_C
ONVERSION_FAILED 12 time(s).
2019-04-25 13:12:21,640 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2019-04-25 13:12:21,641 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-25 13:12:21,641 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2019-04-25 13:12:21,649 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input files to process : 1
2019-04-25 13:12:21,649 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(43037,43.59615384615385)
grunt>
```

Query_1c test 2 input: 333



```
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:18:31,026 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:18:31,029 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:18:31,030 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:18:31,031 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:18:31,032 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 12 time(s).
2019-04-25 14:18:31,032 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 1 time(s).
2019-04-25 14:18:31,033 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Success!
2019-04-25 14:18:31,033 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.
checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-25 14:18:31,033 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has
already been initialized
2019-04-25 14:18:31,042 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
put files to process : 1
2019-04-25 14:18:31,042 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
(77064,48.032786885245905)
grunt>
```

Query_1c test 3 input: 233

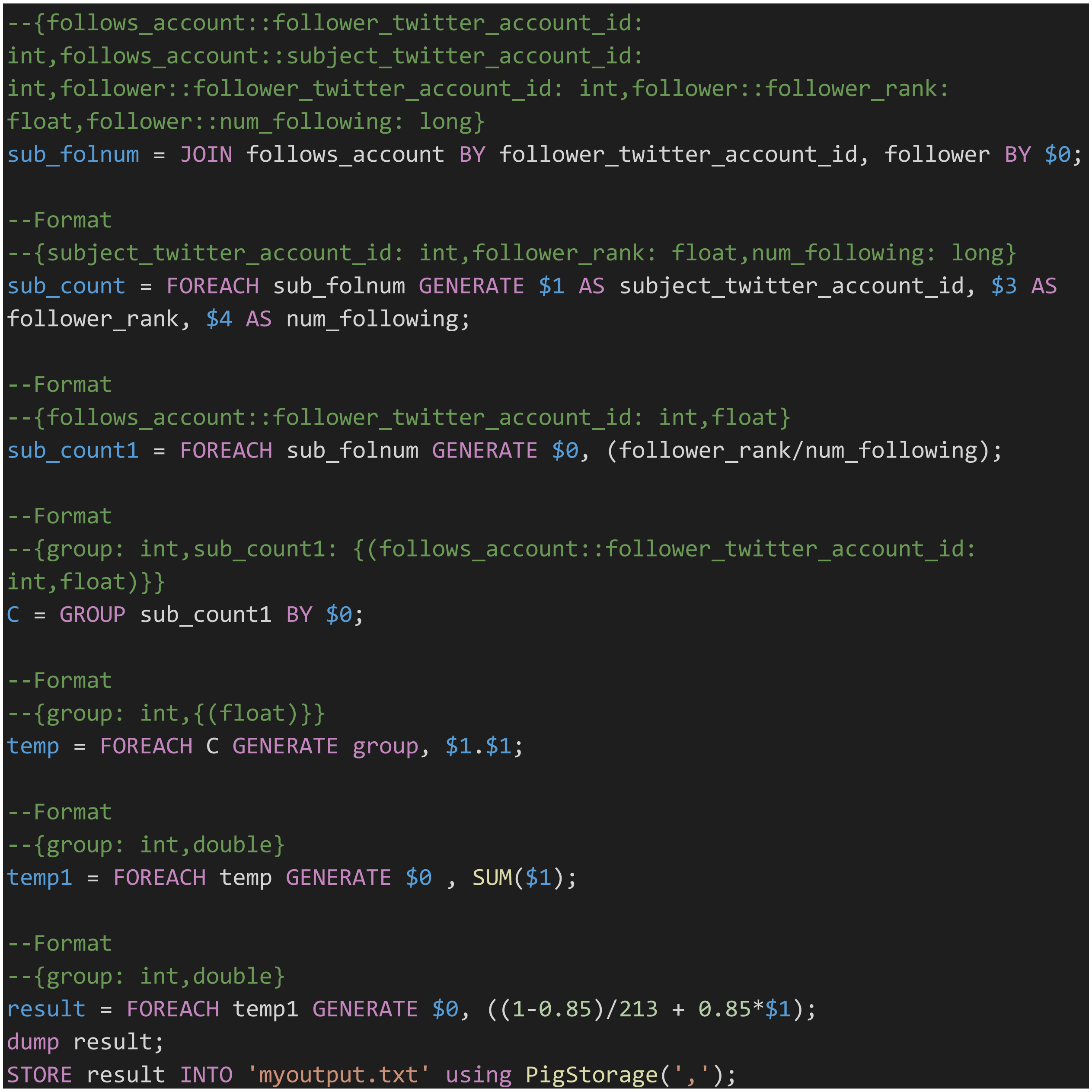


```
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:19:14,008 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:19:14,011 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:19:14,012 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:19:14,013 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:19:14,014 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 12 time(s).
2019-04-25 14:19:14,014 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 1 time(s).
2019-04-25 14:19:14,014 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Success!
2019-04-25 14:19:14,015 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.
checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-25 14:19:14,015 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has
already been initialized
2019-04-25 14:19:14,027 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
put files to process : 1
2019-04-25 14:19:14,027 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
(84485,48.422818791946305)
grunt>
```

Task 2

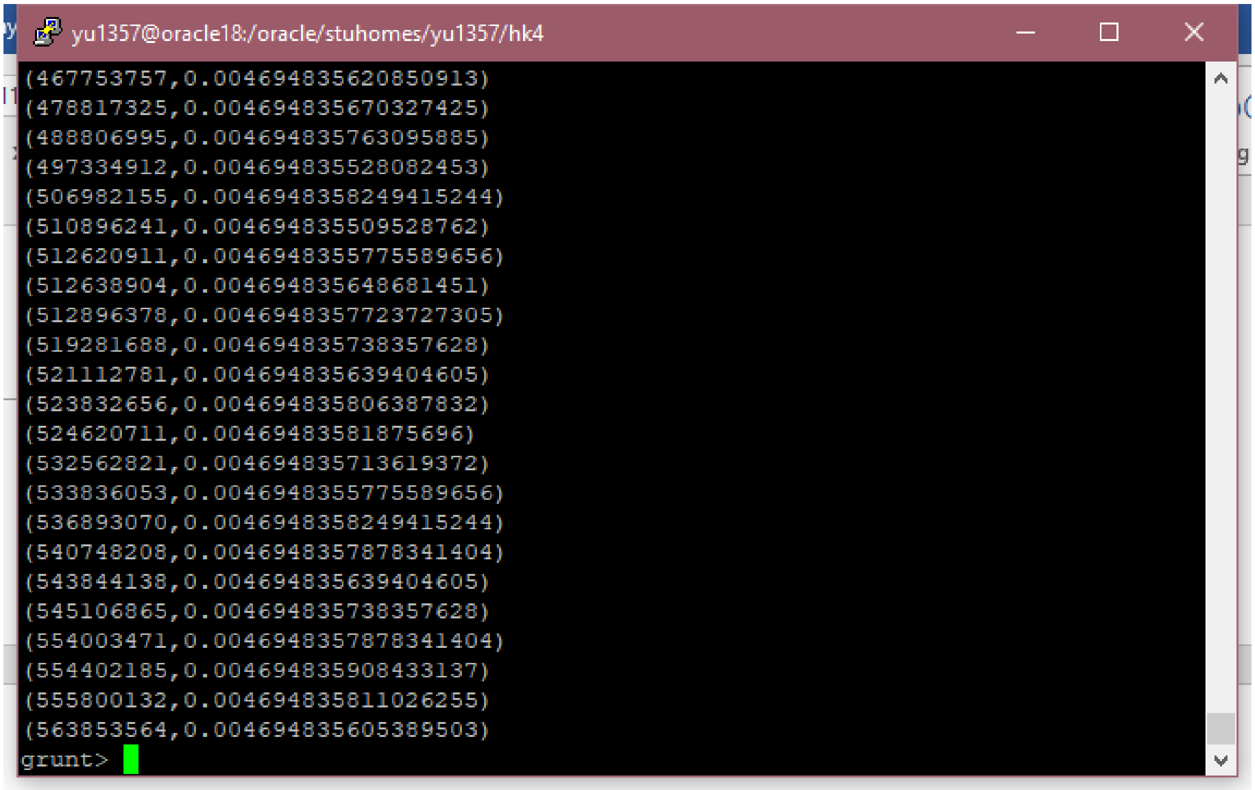
query2.PIG:

```
-- Format:
--{follower_twitter_account_id: int,subject_twitter_account_id: int}
follows_account = LOAD '/oracle/stuhomes/yu1357/hk4/HW4-follows_account.txt'
using PigStorage(',') AS (follower_twitter_account_id:
int,subject_twitter_account_id: int);


-- Format:
--{twitter_account_id: int,email_address: chararray,phone_number:
chararray,user_location: chararray,num_tweets: int}
twitter_account = LOAD '/oracle/stuhomes/yu1357/hk4/HW4-twitter_account.txt'
USING PigStorage(',') AS (twitter_account_id: int, email_address: chararray ,
phone_number: chararray , user_location: chararray , num_tweets: int);

-- Format:
--{twitter_account_id: int,twitter_account_rank: float}
old_twitter_account_rank = LOAD '/oracle/stuhomes/yu1357/hk4/HW4-
old_twitter_account_rank.txt' using PigStorage(',') AS (twitter_account_id:int,
twitter_account_rank:float);


--Format
--{group: int,follows_account: {(follower_twitter_account_id:
int,subject_twitter_account_id: int)}}
B = GROUP follows_account BY follower_twitter_account_id;

--Format
--{group: int,long}
follower = FOREACH B GENERATE group, COUNT($1);

--Format
--{old_twitter_account_rank::twitter_account_id:
int,old_twitter_account_rank::twitter_account_rank: float,follower::group:
int,long}
follower = JOIN old_twitter_account_rank BY twitter_account_id, follower BY $0;

--Format
--{follower_twitter_account_id: int,follower_rank: float,num_following: long}
follower = FOREACH follower GENERATE $0 AS follower_twitter_account_id, $1 AS
follower_rank, $3 AS num_following;

--Format
```



```
--{follows_account::follower_twitter_account_id:
int,follows_account::subject_twitter_account_id:
int,follower::follower_twitter_account_id: int,follower::follower_rank:
float,follower::num_following: long}
sub_folnum = JOIN follows_account BY follower_twitter_account_id, follower BY $0;

--Format
--{subject_twitter_account_id: int,follower_rank: float,num_following: long}
sub_count = FOREACH sub_folnum GENERATE $1 AS subject_twitter_account_id, $3 AS
follower_rank, $4 AS num_following;

--Format
--{follows_account::follower_twitter_account_id: int,float}
sub_count1 = FOREACH sub_folnum GENERATE $0, (follower_rank/num_following);

--Format
--{group: int,sub_count1: {(follows_account::follower_twitter_account_id:
int,float)}}
C = GROUP sub_count1 BY $0;

--Format
--{group: int,{(float)}}
temp = FOREACH C GENERATE group, $1.$1;

--Format
--{group: int,double}
temp1 = FOREACH temp GENERATE $0 , SUM($1);

--Format
--{group: int,double}
result = FOREACH temp1 GENERATE $0, ((1-0.85)/213 + 0.85*$1);
dump result;
STORE result INTO 'myoutput.txt' using PigStorage(',');
```

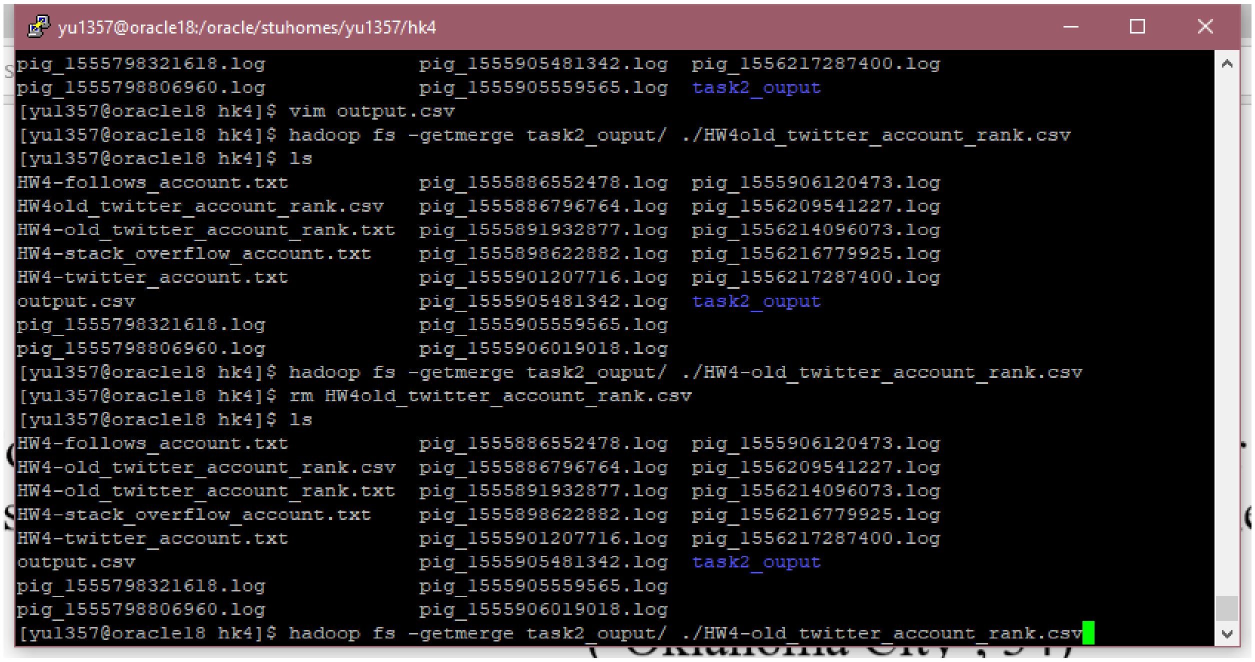
New Rank Twitter Account



yu1357@oracle18:/oracle/stuhomes/yu1357/hk4

```
(467753757,0.004694835620850913)
(478817325,0.004694835670327425)
(488806995,0.004694835763095885)
(497334912,0.004694835528082453)
(506982155,0.0046948358249415244)
(510896241,0.004694835509528762)
(512620911,0.0046948355775589656)
(512638904,0.004694835648681451)
(512896378,0.0046948357723727305)
(519281688,0.004694835738357628)
(521112781,0.004694835639404605)
(523832656,0.004694835806387832)
(524620711,0.00469483581875696)
(532562821,0.004694835713619372)
(533836053,0.0046948355775589656)
(536893070,0.0046948358249415244)
(540748208,0.0046948357878341404)
(543844138,0.004694835639404605)
(545106865,0.004694835738357628)
(554003471,0.0046948357878341404)
(554402185,0.004694835908433137)
(555800132,0.004694835811026255)
(563853564,0.004694835605389503)
grunt>
```

Copy output data into file: HW4old_twitter_account_rank.csv



yu1357@oracle18:/oracle/stuhomes/yu1357/hk4

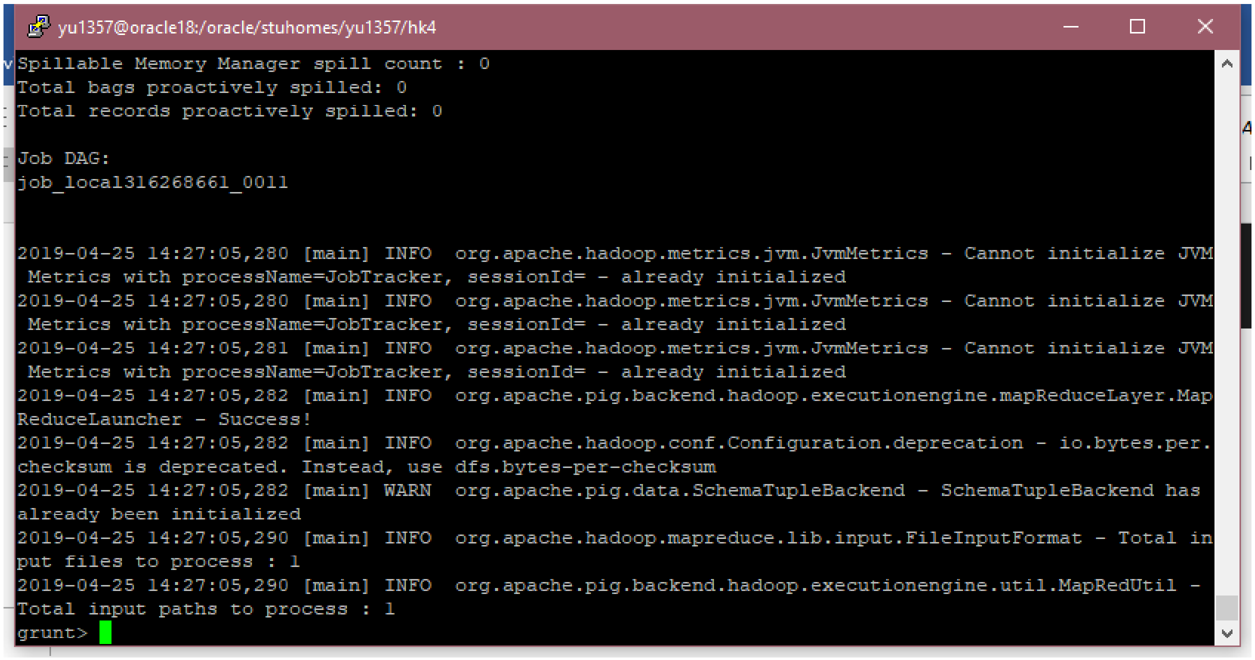
```
pig_1555798321618.log            pig_1555905481342.log  pig_1556217287400.log
pig_1555798806960.log            pig_1555905559565.log  task2_ouput
[yu1357@oracle18 hk4]$ vim output.csv
[yu1357@oracle18 hk4]$ hadoop fs -getmerge task2_ouput/ ./HW4old_twitter_account_rank.csv
[yu1357@oracle18 hk4]$ ls
HW4-follows_account.txt          pig_1555886552478.log  pig_1555906120473.log
HW4old_twitter_account_rank.csv  pig_1555886796764.log  pig_1556209541227.log
HW4-old_twitter_account_rank.txt pig_1555891932877.log  pig_1556214096073.log
HW4-stack_overflow_account.txt   pig_1555898622882.log  pig_1556216779925.log
HW4-twitter_account.txt          pig_1555901207716.log  pig_1556217287400.log
output.csv                       pig_1555905481342.log  task2_ouput
pig_1555798321618.log            pig_1555905559565.log
pig_1555798806960.log            pig_1555906019018.log
[yu1357@oracle18 hk4]$ hadoop fs -getmerge task2_ouput/ ./HW4-old_twitter_account_rank.csv
[yu1357@oracle18 hk4]$ rm HW4old_twitter_account_rank.csv
[yu1357@oracle18 hk4]$ ls
HW4-follows_account.txt          pig_1555886552478.log  pig_1555906120473.log
HW4-old_twitter_account_rank.csv pig_1555886796764.log  pig_1556209541227.log
HW4-old_twitter_account_rank.txt pig_1555891932877.log  pig_1556214096073.log
HW4-stack_overflow_account.txt   pig_1555898622882.log  pig_1556216779925.log
HW4-twitter_account.txt          pig_1555901207716.log  pig_1556217287400.log
output.csv                       pig_1555905481342.log  task2_ouput
pig_1555798321618.log            pig_1555905559565.log
pig_1555798806960.log            pig_1555906019018.log
[yu1357@oracle18 hk4]$ hadoop fs -getmerge task2_ouput/ ./HW4-old_twitter_account_rank.csv
```

Find k_percentile_accounts.pig

```
-- Format:
--{twitter_account_id: int,email_address: chararray,phone_number:
chararray,user_location: chararray,num_tweets: int}
twitter_account = LOAD '/oracle/stuhomes/yu1357/hk4/HW4-twitter_account.txt'
USING PigStorage(',') AS (twitter_account_id: int, email_address: chararray ,
phone_number: chararray , user_location: chararray , num_tweets: int);


-- Format:
--{twitter_account_id: int,twitter_account_rank: float}
old_twitter_account_rank = LOAD '/oracle/stuhomes/yu1357/hk4/HW4-
old_twitter_account_rank.csv' using PigStorage(',') AS (twitter_account_id:int,
twitter_account_rank:float);

--------------------------------first time testing----------------------------
-- Format
--{twitter_account::twitter_account_id: int,twitter_account::email_address:
chararray,twitter_account::phone_number:
chararray,twitter_account::user_location: chararray,twitter_account::num_tweets:
int,old_twitter_account_rank::twitter_account_id:
int,old_twitter_account_rank::twitter_account_rank: float}
join_twitter = JOIN twitter_account BY twitter_account_id,
old_twitter_account_rank BY twitter_account_id;


--Format
--{twitter_account_id: int,email_address: chararray,twitter_rank: float}
percentile_accounts = FOREACH join_twitter GENERATE $0 AS twitter_account_id, $1
AS email_address, $6 AS twitter_rank;


result1 = FILTER  percentile_accounts BY twitter_rank > 0.3;


DESCRIBE result1;
STORE result1 INTO task3_1;
dump result1;


--------------------------------Second time testing----------------------------
-- Format
--{twitter_account::twitter_account_id: int,twitter_account::email_address:
chararray,twitter_account::phone_number:
chararray,twitter_account::user_location: chararray,twitter_account::num_tweets:
```

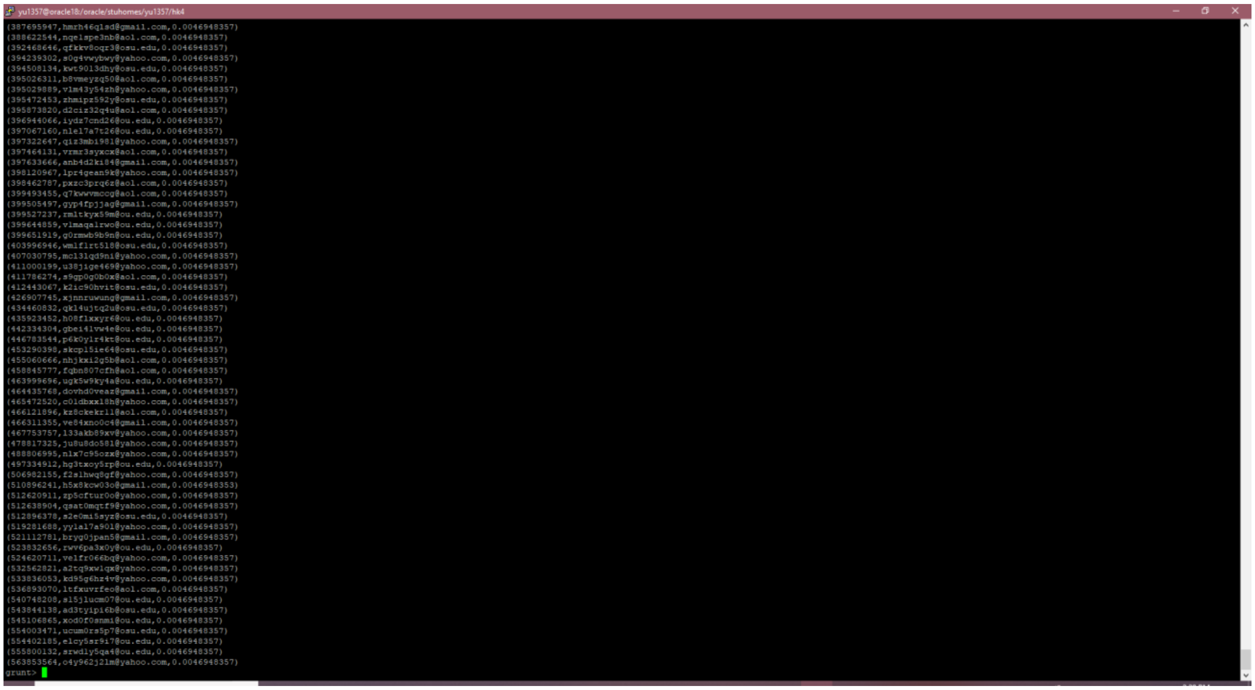
```
int,old_twitter_account_rank::twitter_account_id:
int,old_twitter_account_rank::twitter_account_rank: float}
join_twitter = JOIN twitter_account BY twitter_account_id,
old_twitter_account_rank BY twitter_account_id;


--Format
--{twitter_account_id: int,email_address: chararray,twitter_rank: float}
percentile_accounts = FOREACH join_twitter GENERATE $0 AS twitter_account_id, $1
AS email_address, $6 AS twitter_rank;

result2 = FILTER  percentile_accounts BY twitter_rank > 0.0042;


DESCRIBE result2;
STORE result2 INTO task3_2;
dump result2;
--------------------------------thrid time testing----------------------------
-- Format
--{twitter_account::twitter_account_id: int,twitter_account::email_address:
chararray,twitter_account::phone_number:
chararray,twitter_account::user_location: chararray,twitter_account::num_tweets:
int,old_twitter_account_rank::twitter_account_id:
int,old_twitter_account_rank::twitter_account_rank: float}
join_twitter = JOIN twitter_account BY twitter_account_id,
old_twitter_account_rank BY twitter_account_id;


--Format
--{twitter_account_id: int,email_address: chararray,twitter_rank: float}
percentile_accounts = FOREACH join_twitter GENERATE $0 AS twitter_account_id, $1
AS email_address, $6 AS twitter_rank;

result3 = FILTER  percentile_accounts BY twitter_rank > 0.005;


DESCRIBE result3;
STORE result3 INTO task3_3;
dump result3;
```

Task3 output test 1 Input: 0.3



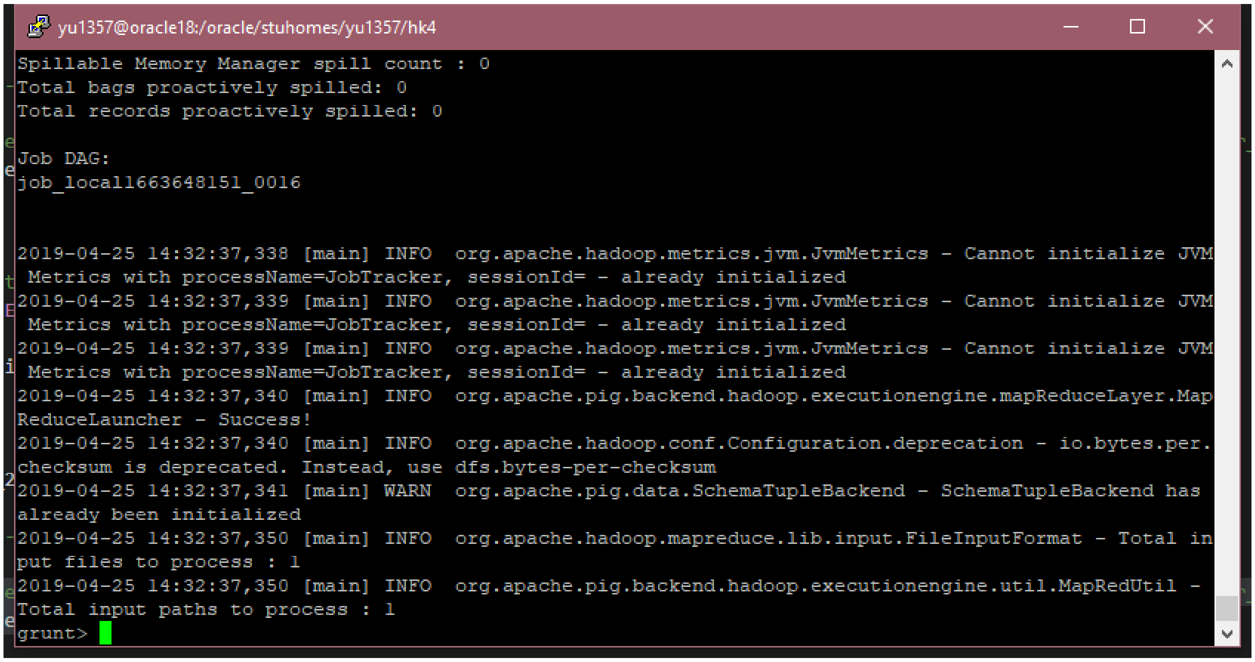
```
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local316268661_0011


2019-04-25 14:27:05,280 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:27:05,280 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:27:05,281 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:27:05,282 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Success!
2019-04-25 14:27:05,282 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.
checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-25 14:27:05,282 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has
already been initialized
2019-04-25 14:27:05,290 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
put files to process : 1
2019-04-25 14:27:05,290 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
grunt>
```

Task3 output test 2 Input: 0.0042



```
(387695947,hmrh46qlsd@gmail.com,0.0046948357)
(388622544,nqelspe3nb@aol.com,0.0046948357)
(392468646,qfkkv8oqr3@osu.edu,0.0046948357)
(394239302,s0g4vwybwy@yahoo.com,0.0046948357)
(394508134,kwt9013dhy@osu.edu,0.0046948357)
(395026311,b8vmeyzq50@aol.com,0.0046948357)
(395029889,vlm43y54zh@yahoo.com,0.0046948357)
(395472453,zhmipz592y@osu.edu,0.0046948357)
(395873820,d2ciz32q4u@aol.com,0.0046948357)
(396944066,iydz7cnd26@ou.edu,0.0046948357)
(397067160,nlel7a7t26@ou.edu,0.0046948357)
(397322647,qiz3mbi981@yahoo.com,0.0046948357)
(397464131,vrmr3syxcx@aol.com,0.0046948357)
(397633666,anb4d2ki84@gmail.com,0.0046948357)
(398120967,lpr4gean9k@yahoo.com,0.0046948357)
(398462787,pxzc3prq6z@aol.com,0.0046948357)
(399493455,q7kwwvmccg@aol.com,0.0046948357)
(399505497,gyp4fpjjag@gmail.com,0.0046948357)
(399527237,rmltkyx59m@ou.edu,0.0046948357)
(399644859,vlmaqalrwo@ou.edu,0.0046948357)
(399651919,g0rmwb9b9n@ou.edu,0.0046948357)
(403996946,wmlflrt518@osu.edu,0.0046948357)
(407030795,mcl31qd9ni@yahoo.com,0.0046948357)
(411000199,u38jige469@yahoo.com,0.0046948357)
(411786274,s9gp0g0b0x@aol.com,0.0046948357)
(412443067,k2ic90hvit@osu.edu,0.0046948357)
(426907745,xjnnruwung@gmail.com,0.0046948357)
(434460832,qkl4ujtq2u@osu.edu,0.0046948357)
(435923452,h08flxxyr6@ou.edu,0.0046948357)
(442334304,gbei4lvw4e@ou.edu,0.0046948357)
(446783544,p6k0ylr4kt@ou.edu,0.0046948357)
(453290398,skcpl5ie64@osu.edu,0.0046948357)
(455060666,nhjkxi2g5b@aol.com,0.0046948357)
(458845777,fqbn807cfh@aol.com,0.0046948357)
(463999696,ugk5w9ky4a@ou.edu,0.0046948357)
(464435768,dovhd0veaz@gmail.com,0.0046948357)
(465472520,c0ldbxxl8h@yahoo.com,0.0046948357)
(466121896,kz8ckekrll@aol.com,0.0046948357)
(466311355,ve84xno0c4@gmail.com,0.0046948357)
(467753757,l33akb89xv@yahoo.com,0.0046948357)
(478817325,ju8u8do581@yahoo.com,0.0046948357)
(488806995,nlx7c95ozx@yahoo.com,0.0046948357)
(497334912,hg3txoy5rp@ou.edu,0.0046948357)
(506982155,f2slhwq8gf@yahoo.com,0.0046948357)
(510896241,h5x8kcw03o@gmail.com,0.0046948353)
(512620911,zp5cftur0o@yahoo.com,0.0046948357)
(512638904,qsat0mqtf9@yahoo.com,0.0046948357)
(512896378,s2e0mi5syz@osu.edu,0.0046948357)
(519281688,yylal7a901@yahoo.com,0.0046948357)
(521112781,bryg0jpan5@gmail.com,0.0046948357)
(523832656,rwv6pa3x0y@ou.edu,0.0046948357)
(524620711,velfr066bq@yahoo.com,0.0046948357)
(532562821,a2tq9xwlqx@yahoo.com,0.0046948357)
(533836053,kd95g6hz4v@yahoo.com,0.0046948357)
(536893070,ltfxuvrfeo@aol.com,0.0046948357)
(540748208,sl5jlucm07@ou.edu,0.0046948357)
(543844138,ad3tyipi6b@osu.edu,0.0046948357)
(545106865,xod0f0snmi@ou.edu,0.0046948357)
(554003471,ucum0rs5p7@osu.edu,0.0046948357)
(554402185,elcy5sr9i7@ou.edu,0.0046948357)
(555800132,srwdly5qa4@ou.edu,0.0046948357)
(563853564,o4y962j2lm@yahoo.com,0.0046948357)
grunt>
```

Task3 output test 2 Input: 0.005



```
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1663648151_0016


2019-04-25 14:32:37,338 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:32:37,339 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:32:37,339 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM
 Metrics with processName=JobTracker, sessionId= - already initialized
2019-04-25 14:32:37,340 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Success!
2019-04-25 14:32:37,340 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.
checksum is deprecated. Instead, use dfs.bytes-per-checksum
2019-04-25 14:32:37,341 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has
already been initialized
2019-04-25 14:32:37,350 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total in
put files to process : 1
2019-04-25 14:32:37,350 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
Total input paths to process : 1
grunt>
```