

SESSION 5 ASSIGNMENT 1

Task_1:

```
grunt>emp_group = GROUP emp_details by rating;

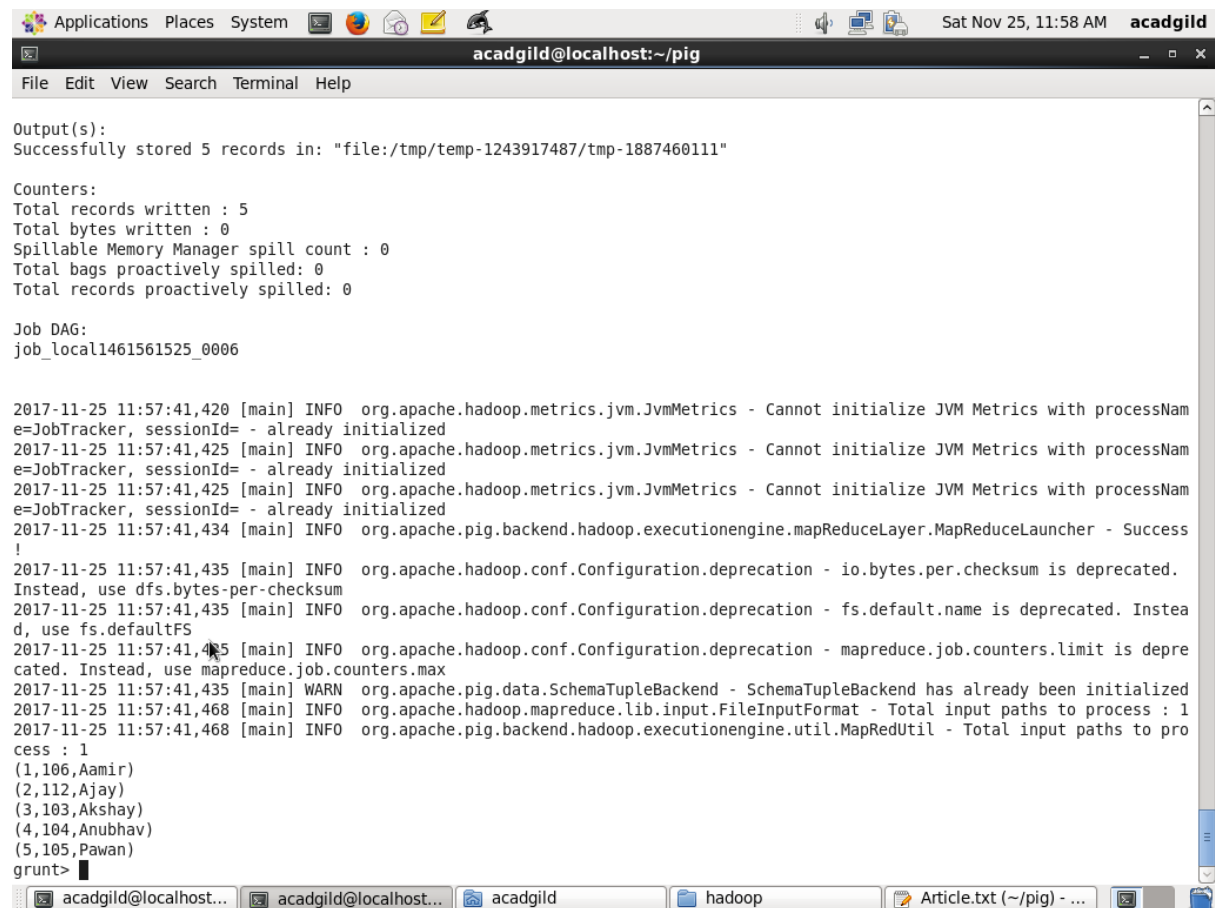
grunt> emp_5 = FOREACH emp_group {

>> da = ORDER emp_details BY rating DESC, name;

>> db = LIMIT da 1;

>> GENERATE FLATTEN(group), FLATTEN(db.id), FLATTEN(db.name);

>> }
```



The screenshot shows a terminal window titled 'acadgild@localhost:~/pig' with a menu bar (File, Edit, View, Search, Terminal, Help). The output of a Pig job is displayed, showing successful storage of 5 records and various counters. It also includes a Job DAG and a series of log messages from the Hadoop ecosystem, including warnings about deprecated configuration parameters and information about the SchemaTupleBackend and MapReduceUtil.

```
Output(s):
Successfully stored 5 records in: "file:/tmp/temp-1243917487/tmp-1887460111"

Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1461561525_0006

2017-11-25 11:57:41,420 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-25 11:57:41,425 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-25 11:57:41,425 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-25 11:57:41,434 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-11-25 11:57:41,435 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-25 11:57:41,435 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-25 11:57:41,435 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max
2017-11-25 11:57:41,435 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-25 11:57:41,468 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-25 11:57:41,468 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,106,Aamir)
(2,112,Ajay)
(3,103,Akshay)
(4,104,Anubhav)
(5,105,Pawan)
grunt>
```

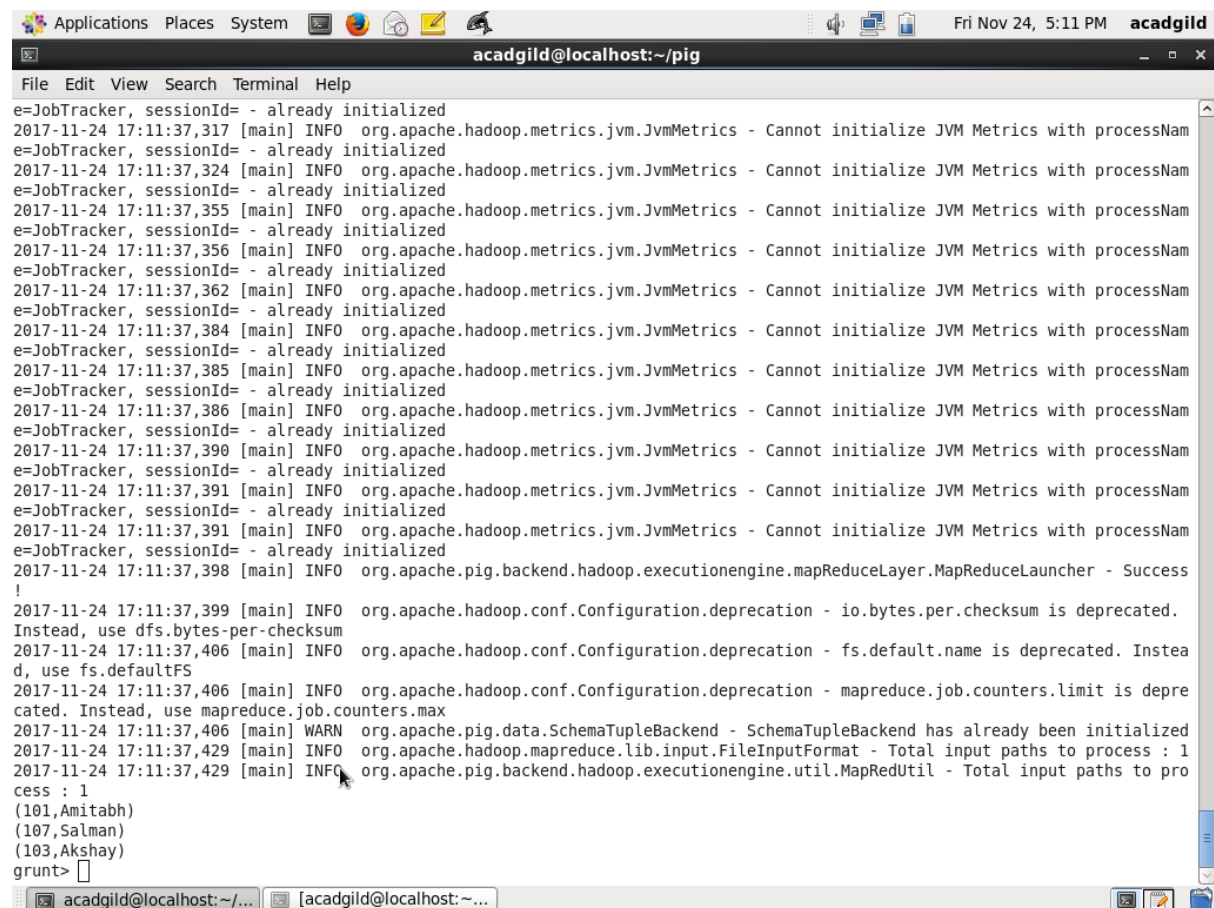
Task_2:

```
grunt> emp_odd = FILTER emp_details BY (id%2 != 0);

grunt> emp_odd_order = ORDER emp_odd BY salary desc, name;

grunt> emp = LIMIT emp_odd_order 3;

grunt> emp_result = FOREACH emp GENERATE id,name;
```



The screenshot shows a terminal window titled 'acadgild@localhost:~/pig'. The window displays a series of log messages from Hadoop and Pig. The logs include timestamps, log levels (INFO, WARN), and messages from various components like 'org.apache.hadoop.metrics.jvm.JvmMetrics', 'org.apache.hadoop.conf.Configuration.deprecation', 'org.apache.pig.data.SchemaTupleBackend', 'org.apache.hadoop.mapreduce.lib.input.FileInputFormat', and 'org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil'. The final output of the Pig script is displayed at the bottom: (101,Amitabh), (107,Salman), and (103,Akshay).

```
File Edit View Search Terminal Help
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,317 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,324 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,355 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,356 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,362 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,384 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,385 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,386 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,390 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,391 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,391 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2017-11-24 17:11:37,398 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2017-11-24 17:11:37,399 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2017-11-24 17:11:37,406 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2017-11-24 17:11:37,406 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2017-11-24 17:11:37,406 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-24 17:11:37,429 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-24 17:11:37,429 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(101,Amitabh)
(107,Salman)
(103,Akshay)
grunt>
```

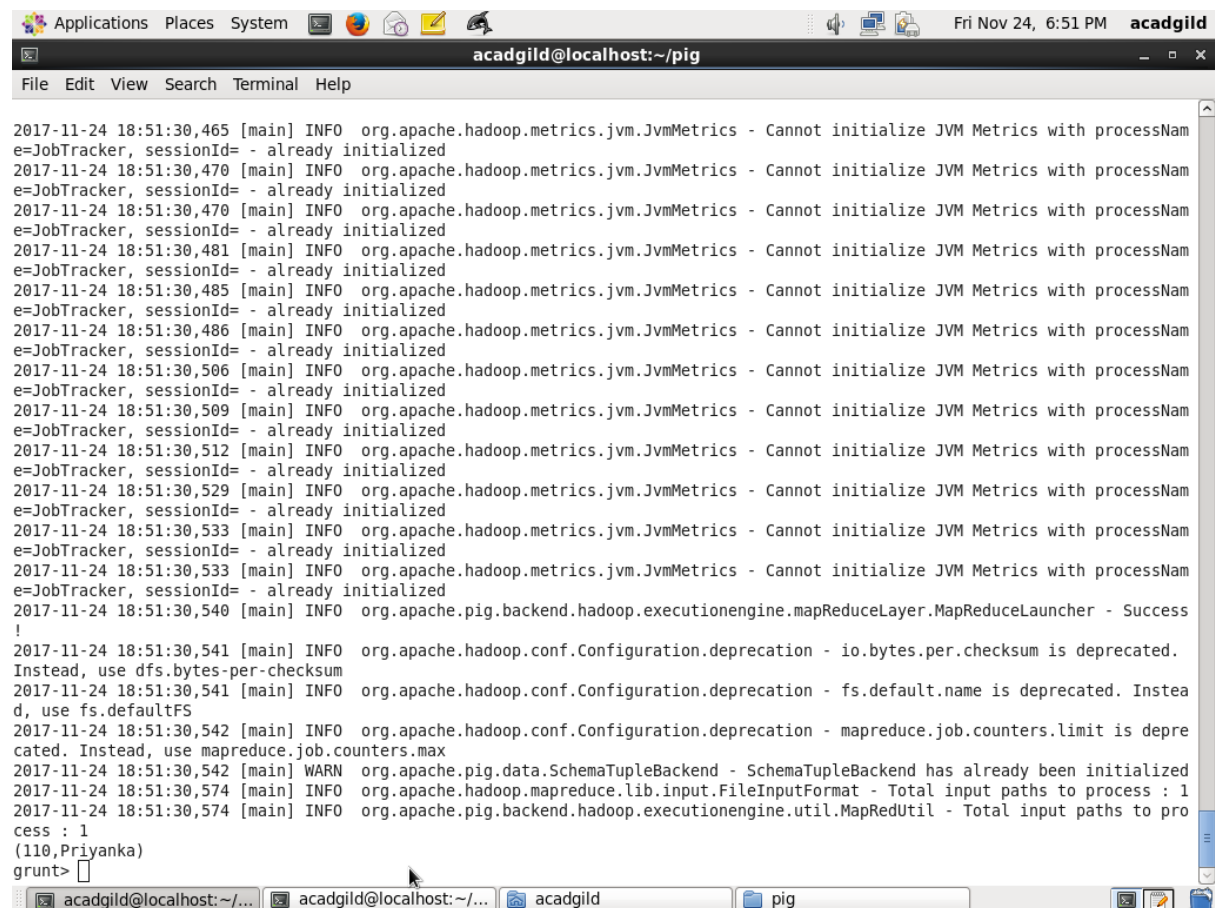
Task_3:

```
grunt> A = JOIN emp_details BY id, emp_expenses BY id;

grunt> B = ORDER A BY expense DESC,name;

grunt> C = LIMIT B 1;

grunt> D= FOREACH C GENERATE emp_details::id,name;
```



The screenshot shows a terminal window titled 'acadgild@localhost:~/pig'. The window displays the output of a Pig script execution. The output consists of multiple log messages from the Hadoop and Pig frameworks. Key messages include:

- Repeated warnings: 'org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized'.
- Configuration deprecation warnings: 'org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum', 'org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS', and 'org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max'.
- Success message: 'org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!'.
- Warnings about already initialized components: 'org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized' and 'org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1'.
- Final summary: 'org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1 (110, Priyanka)'.

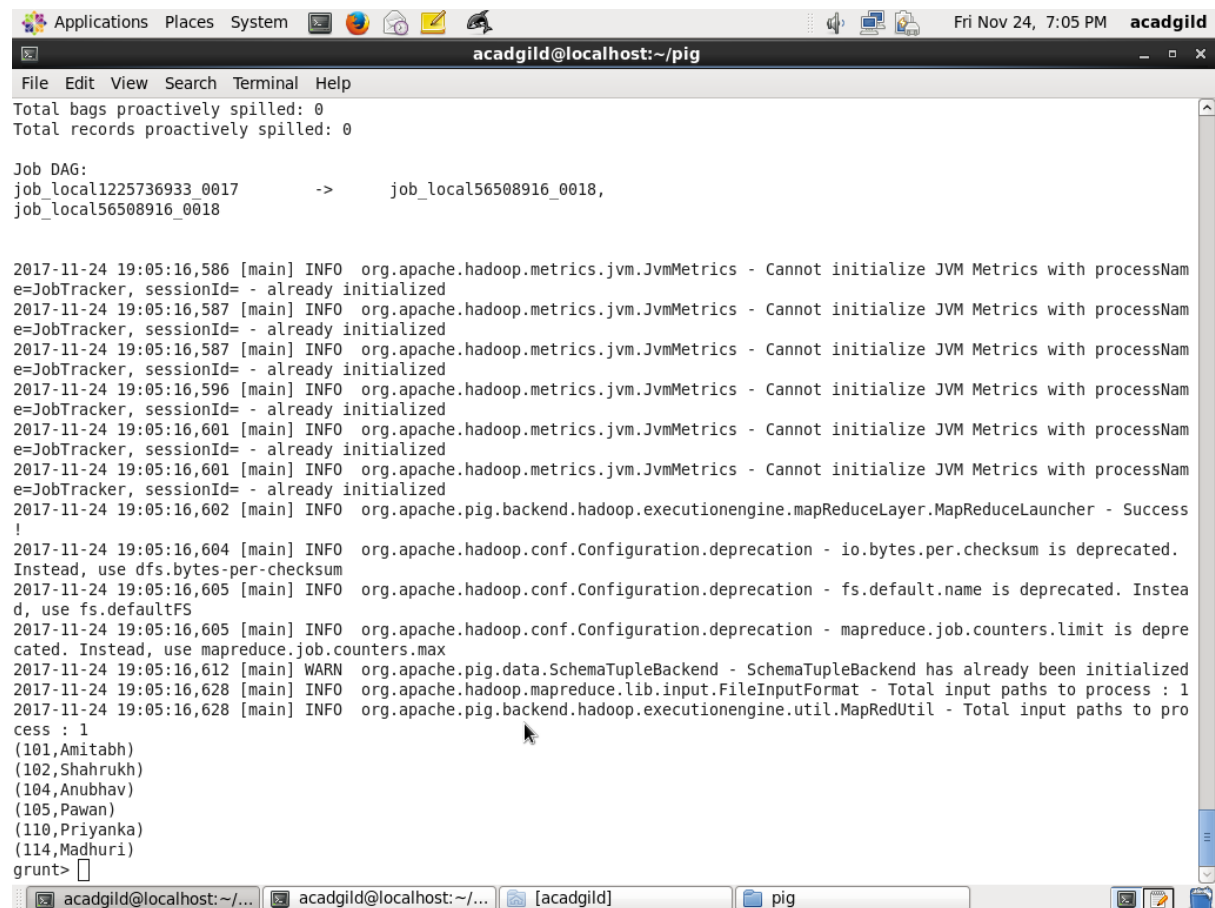
The terminal prompt 'grunt>' is visible at the bottom of the output.

Task_4:

```
grunt> joined = JOIN emp_details BY id, emp_expenses BY id;

grunt> emp = FOREACH joined GENERATE emp_details::id, name;

grunt> emp_data = DISTINCT emp;
```



The screenshot shows a terminal window titled 'acadgild@localhost:~/pig'. The window displays the output of a Pig script execution. At the top, it shows 'Total bags proactively spilled: 0' and 'Total records proactively spilled: 0'. Below this, the 'Job DAG' is shown with two jobs: 'job_local1225736933_0017' and 'job_local56508916_0018'. The main part of the output consists of log messages from the Hadoop and Pig frameworks. These include warnings about deprecated configuration properties like 'io.bytes.per.checksum', 'fs.default.name', and 'mapreduce.job.counters.limit'. There are also several 'INFO' messages from 'org.apache.hadoop.metrics.jvm.JvmMetrics' stating 'Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized'. A 'WARN' message from 'org.apache.pig.data.SchemaTupleBackend' indicates it has already been initialized. Finally, 'INFO' messages from 'org.apache.hadoop.mapreduce.lib.input.FileInputFormat' and 'org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil' show the total input paths to process as 1. The output concludes with a list of names: (101,Amitabh), (102,Shahrukh), (104,Anubhav), (105,Pawan), (110,Priyanka), and (114,Madhuri). The prompt 'grunt>' is visible at the bottom of the terminal.

```
File Edit View Search Terminal Help
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1225736933_0017    ->    job_local56508916_0018,
job_local56508916_0018

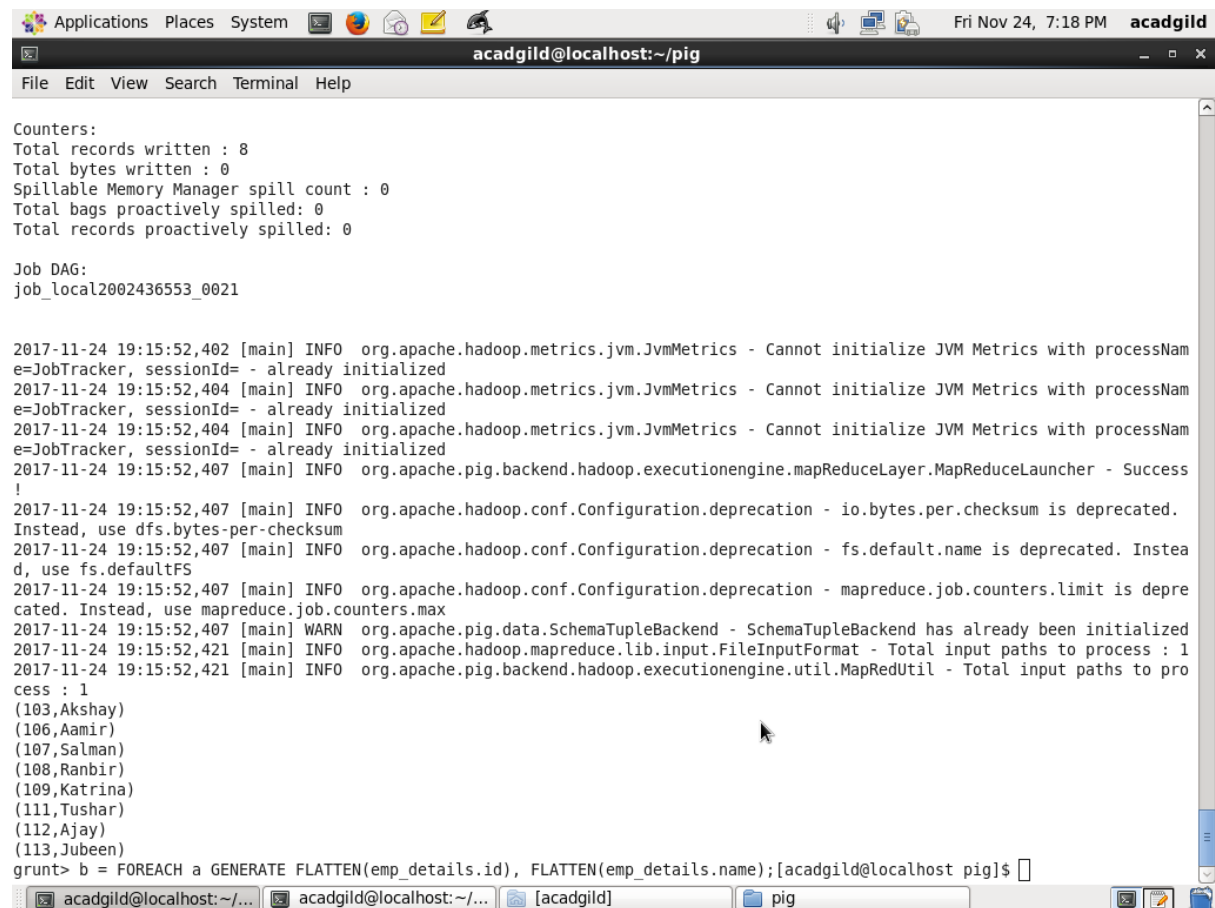
2017-11-24 19:05:16,586 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-24 19:05:16,587 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-24 19:05:16,587 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-24 19:05:16,596 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-24 19:05:16,601 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-24 19:05:16,601 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-24 19:05:16,602 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-11-24 19:05:16,604 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-24 19:05:16,605 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-24 19:05:16,605 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max
2017-11-24 19:05:16,612 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-24 19:05:16,628 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-24 19:05:16,628 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,Amitabh)
(102,Shahrukh)
(104,Anubhav)
(105,Pawan)
(110,Priyanka)
(114,Madhuri)
grunt>
```

Task_5:

```
grunt> joined = GROUP emp_details BY id, emp_expenses BY id;

grunt> a= FILTER joined BY IsEmpty(emp_expenses);

grunt> b = FOREACH a GENERATE FLATTEN(emp_details.id), FLATTEN(emp_details.name);
```



The screenshot shows a terminal window titled 'acadgild@localhost:~/pig'. The window displays the output of a Pig job. At the top, it shows 'Counters:' with values: Total records written: 8, Total bytes written: 0, Spillable Memory Manager spill count: 0, Total bags proactively spilled: 0, and Total records proactively spilled: 0. Below this, it shows 'Job DAG:' with the job ID 'job_local2002436553_0021'. The main part of the output consists of log messages from the Hadoop and Pig frameworks, including warnings about deprecated configuration options and information about the job's progress. The log messages are timestamped '2017-11-24 19:15:52'. The terminal window also shows a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The status bar at the bottom shows the current directory as '~/pig' and the file 'pig'.

```
Applications Places System Fri Nov 24, 7:18 PM acadgild
acadgild@localhost:~/pig
File Edit View Search Terminal Help

Counters:
Total records written : 8
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local2002436553_0021

2017-11-24 19:15:52,402 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-24 19:15:52,404 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-24 19:15:52,404 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-11-24 19:15:52,407 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-11-24 19:15:52,407 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-11-24 19:15:52,407 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-24 19:15:52,407 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max
2017-11-24 19:15:52,407 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-24 19:15:52,421 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-24 19:15:52,421 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(103,Akshay)
(106,Aamir)
(107,Salman)
(108,Ranbir)
(109,Katrina)
(111,Tushar)
(112,Ajay)
(113,Jubeen)
grunt> b = FOREACH a GENERATE FLATTEN(emp_details.id), FLATTEN(emp_details.name);[acadgild@localhost pig]$
```