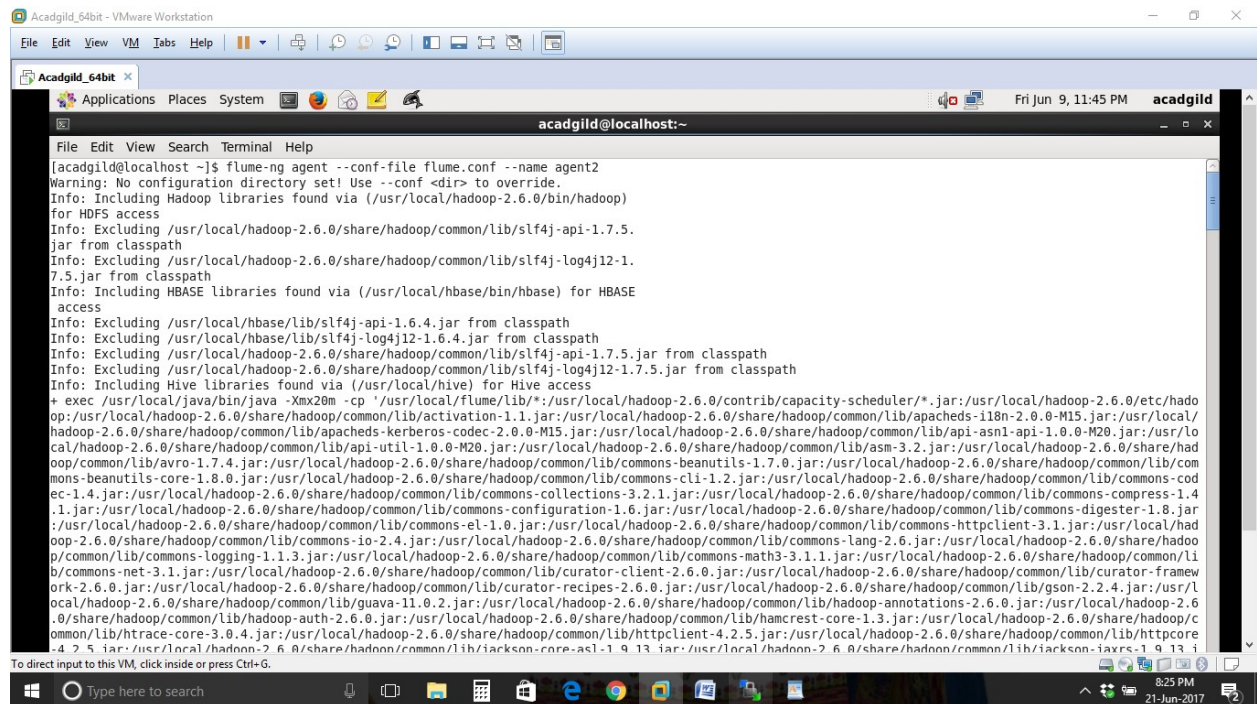


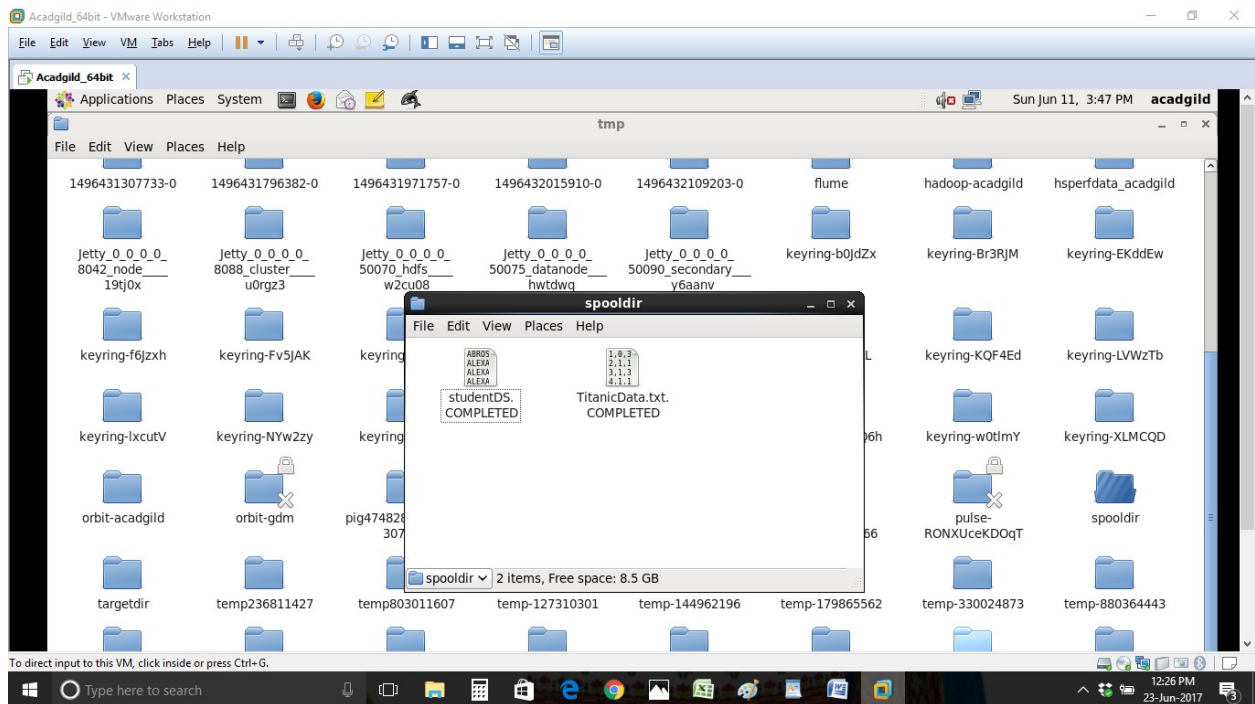
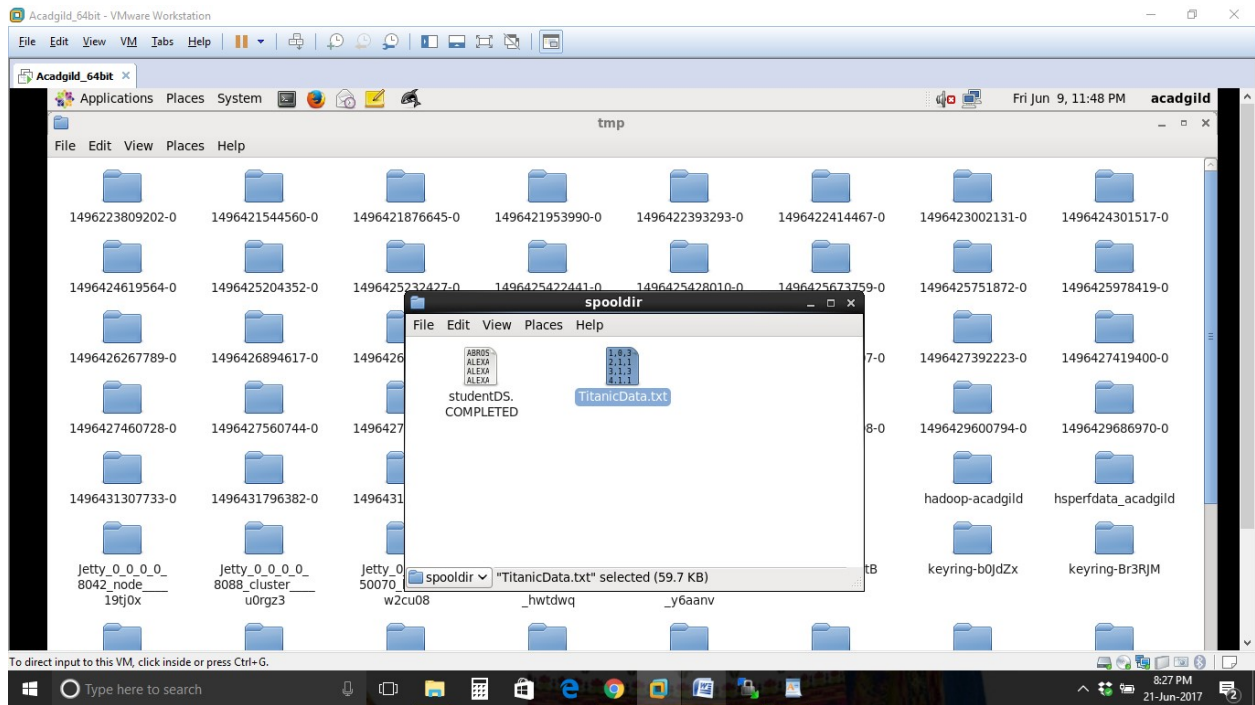
# Project 1.2

## Titanic Data Analysis

### Copy the data set into HDFS using Flume



```
[acadgild@localhost ~]$ flume-ng agent --conf-file flume.conf --name agent2
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/usr/local/hadoop-2.6.0/bin/hadoop)
for HDFS access
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.
jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.
7.5.jar from classpath
Info: Including HBase libraries found via (/usr/local/hbase/bin/hbase) for HBASE
access
Info: Excluding /usr/local/hbase/lib/slf4j-api-1.6.4.jar from classpath
Info: Excluding /usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Including Hive libraries found via (/usr/local/hive) for Hive access
+ exec /usr/local/java/bin/java -Xmx20m -cp '/usr/local/flume/lib/*:/usr/local/hadoop-2.6.0/contrib/capacity-scheduler/*.jar:/usr/local/hadoop-2.6.0/etc/hado
op:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/activation-1.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/apacheds-118n-2.0.0-M15.jar:/usr/local/
hadoop-2.6.0/share/hadoop/common/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/api-asn1-api-1.0.0-M20.jar:/usr/lo
cal/hadoop-2.6.0/share/hadoop/common/lib/api-util-1.0.0-M20.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/asm-3.2.jar:/usr/local/hadoop-2.6.0/share/had
oop/common/lib/avro-1.7.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/com
mons-beanutils-core-1.8.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-cli-1.2.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-cod
ec-1.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-collections-3.2.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-compress-1.4
.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-configuration-1.6.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-digester-1.8.jar
:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-el-1.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-httpclient-3.1.jar:/usr/local/had
oop-2.6.0/share/hadoop/common/lib/commons-io-2.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-lang-2.6.jar:/usr/local/hadoop-2.6.0/share/hadoo
p/common/lib/commons-logging-1.1.3.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/li
b/commons-net-3.1.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/curator-client-2.6.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/curator-framew
ork-2.6.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/curator-recipes-2.6.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/gson-2.2.4.jar:/usr/l
ocal/hadoop-2.6.0/share/hadoop/common/lib/guava-11.0.2.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/hadoop-annotations-2.6.0.jar:/usr/local/hadoop-2.6
.0/share/hadoop/common/lib/hadoop-auth-2.6.0.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/hamcrest-core-1.3.jar:/usr/local/hadoop-2.6.0/share/hadoop/c
ommon/lib/htrace-core-3.0.4.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/httpclient-4.2.5.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/httpcore
-4.2.5.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/jackson-axrs-1.9.13.i
```



## Pig :Latin Code to load the Titanic Passenger Data

```
titanic_data = load 'TitanicData.txt' Using PigStorage(',') as
(pass_id:int,survived:int,class:int,name:chararray,gender:chararray,age:int,sibsp:int,parch:int,ticket:chararray,fare:float,cabin:chararray,embarked:chararray);
```

## Problem Statement

**In this problem statement, we will find the average fare of each class.**

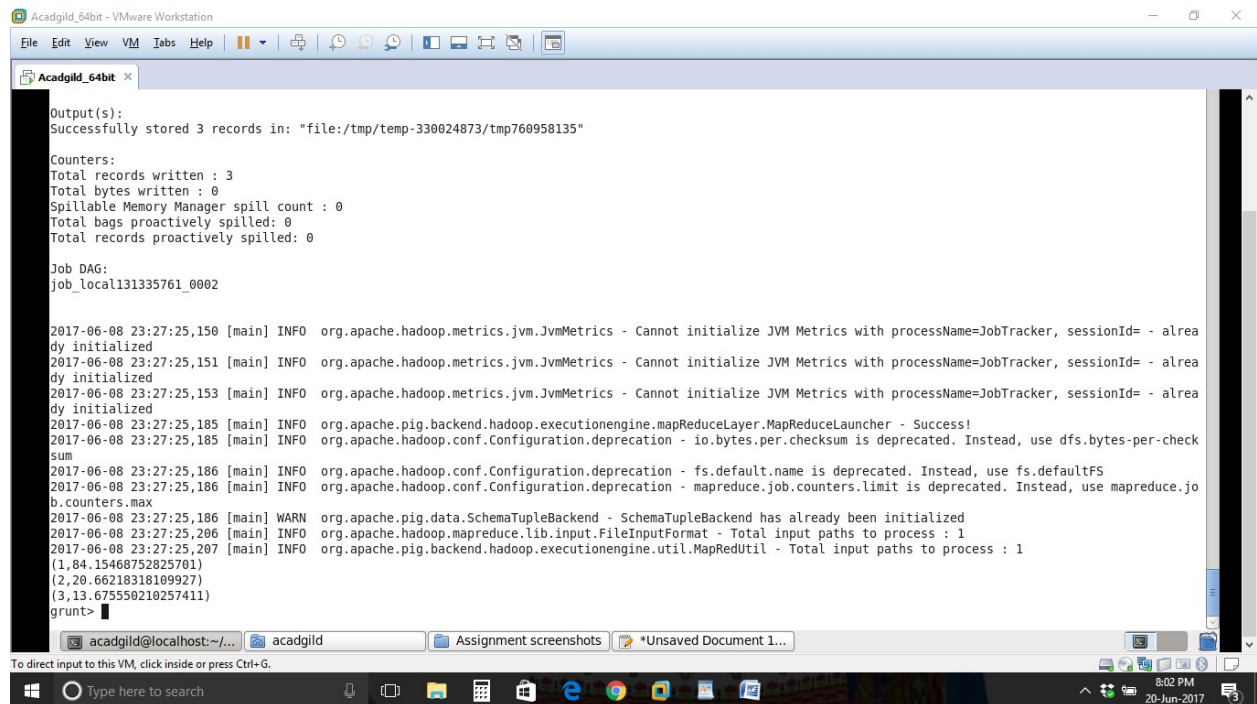
```
fare_class = foreach titanic_data generate class,fare;( generate only class & fare details)
```

```
fare_class_group = group fare_class by class;(group data by classwise)
```

```
avg_class_fare = foreach fare_class_group generate group as class, AVG(fare_class.fare) as
average_fare; ( calculating the average fare for each class based on grouped data)
```

```
dump avg_class_fare;
```

## **Average Fare for each Class**



```
Acadgild_64bit - VMware Workstation
File Edit View VM Tabs Help
Acadgild_64bit x
Output(s):
Successfully stored 3 records in: "file:/tmp/temp-330024873/tmp760958135"

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local131335761_0002

2017-06-08 23:27:25,150 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-08 23:27:25,151 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-08 23:27:25,153 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-08 23:27:25,185 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-06-08 23:27:25,185 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2017-06-08 23:27:25,186 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-06-08 23:27:25,186 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.jo
b.counters.max
2017-06-08 23:27:25,186 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-06-08 23:27:25,206 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-06-08 23:27:25,207 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,84.15468752825701)
(2,20.66218318189927)
(3,13.675550210257411)
grunt>
```

## Problem Statement

In this problem statement, we will find the number of people alive in each class and embarked at Southampton.

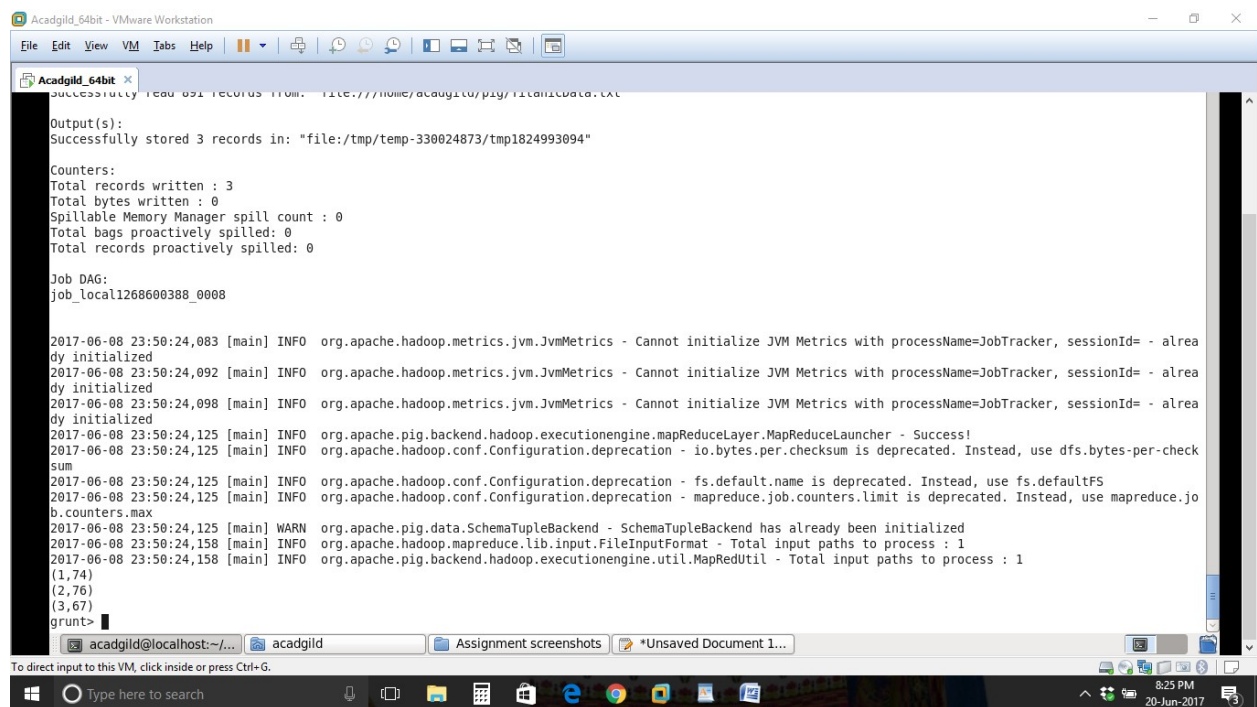
alive\_southampton = filter titanic\_data by embarked matches 'S' and survived==1; (**filter data of passengers who embarked in Southampton and are alive**)

alive\_southampton\_class = group alive\_southampton by class; (**Group the data class wise**)

alive\_southampton\_count = foreach alive\_southampton\_class generate group, COUNT(alive\_southampton.name); (**generating the count of alive passenger in each class who embarked from Southampton**)

dump alive\_southampton\_count;

**Number of passenger alive in each class who embarked in Southampton**



```
Acadgild_64bit - VMware Workstation
File Edit View VM Tabs Help
Acadgild_64bit
Successfully read 031 records from: file:///home/acadgild/pig/titanicdata.txt
Output(s):
Successfully stored 3 records in: "file:///tmp/temp-330024873/tmp1824993094"
Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
Job DAG:
job_local1268600388_0008
2017-06-08 23:50:24,083 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-06-08 23:50:24,092 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-06-08 23:50:24,098 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2017-06-08 23:50:24,125 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-06-08 23:50:24,125 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2017-06-08 23:50:24,125 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-06-08 23:50:24,125 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max
2017-06-08 23:50:24,125 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-06-08 23:50:24,158 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-06-08 23:50:24,158 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,74)
(2,76)
(3,67)
grunt>
```

## Problem Statement

In this problem statement, we will find out number of males and females who died in each class.

pass\_died = filter titanic\_data by survived==0;  
(filtering details of passenger who died)

pass\_died\_data = foreach pass\_died generate class,gender;  
(generating only class and gender details of passenger died)

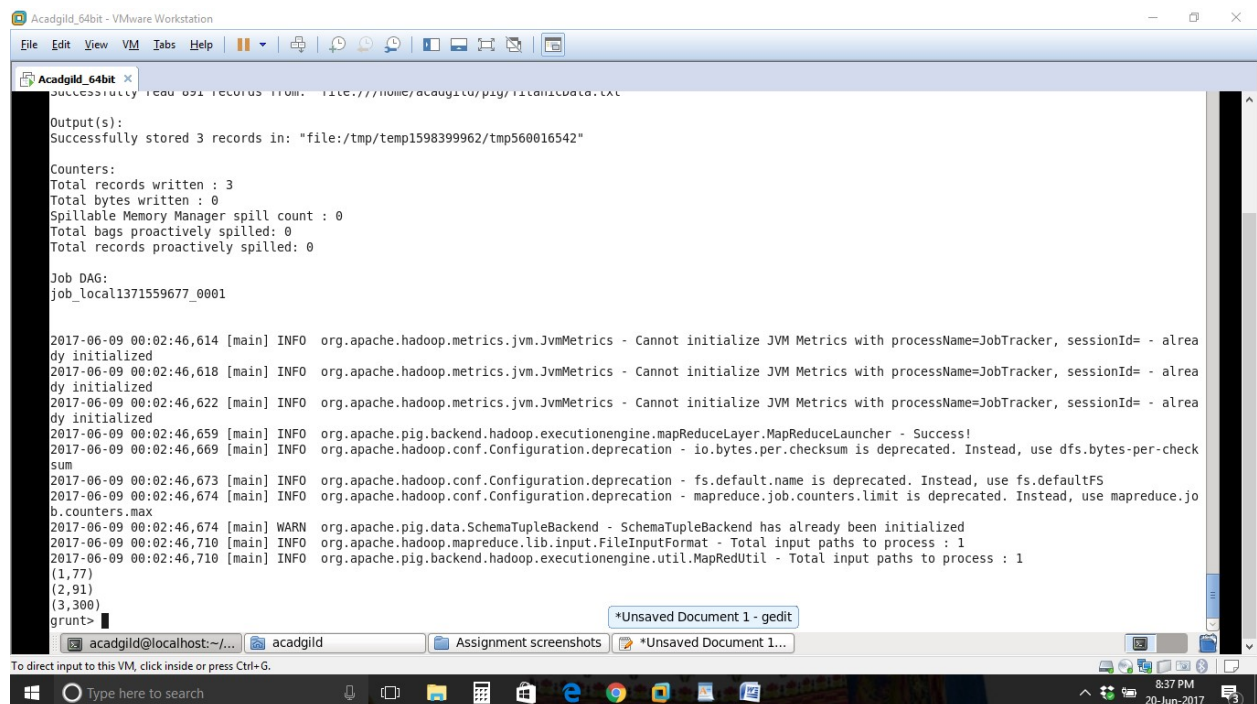
male\_pass\_died = filter pass\_died\_data by gender matches'male';  
(filtering details of only male passenger)

male\_died\_group = group male\_pass\_died by class;  
( group the male passenger died by class)

male\_died\_count = foreach male\_died\_group generate group,  
COUNT(male\_pass\_died.gender);  
(generating count of male passenger died class wise)

Dump male\_died\_count;

**Number of male passengers died in each class**



The screenshot shows a terminal window titled "Acadgild\_64bit - VMware Workstation". The terminal displays the output of a Hadoop MapReduce job. The output includes a summary of the job's progress, such as "Successfully stored 3 records in: 'file:/tmp/tmp1598399962/tmp560016542'", and a list of counters. The counters show that 3 records were written, with 0 bytes and 0 bags spilled. The output also includes a list of log messages from the Hadoop framework, including warnings about deprecated configuration options and information about the job's progress. The terminal window has a standard Linux-style interface with a prompt "grunt>" and a status bar at the bottom showing the time as 8:37 PM on 20-Jun-2017.

```
Acadgild_64bit x
Successfully read 031 records from: 'file:///home/acadgild/pig/titanicdata.txt'

Output(s):
Successfully stored 3 records in: "file:/tmp/tmp1598399962/tmp560016542"

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1371559677_0001

2017-06-09 00:02:46,614 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-09 00:02:46,618 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-09 00:02:46,622 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-09 00:02:46,659 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-06-09 00:02:46,669 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
sum
2017-06-09 00:02:46,673 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-06-09 00:02:46,674 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.jo
b.counters.max
2017-06-09 00:02:46,674 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-06-09 00:02:46,710 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-06-09 00:02:46,710 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,77)
(2,91)
(3,300)
grunt>
```



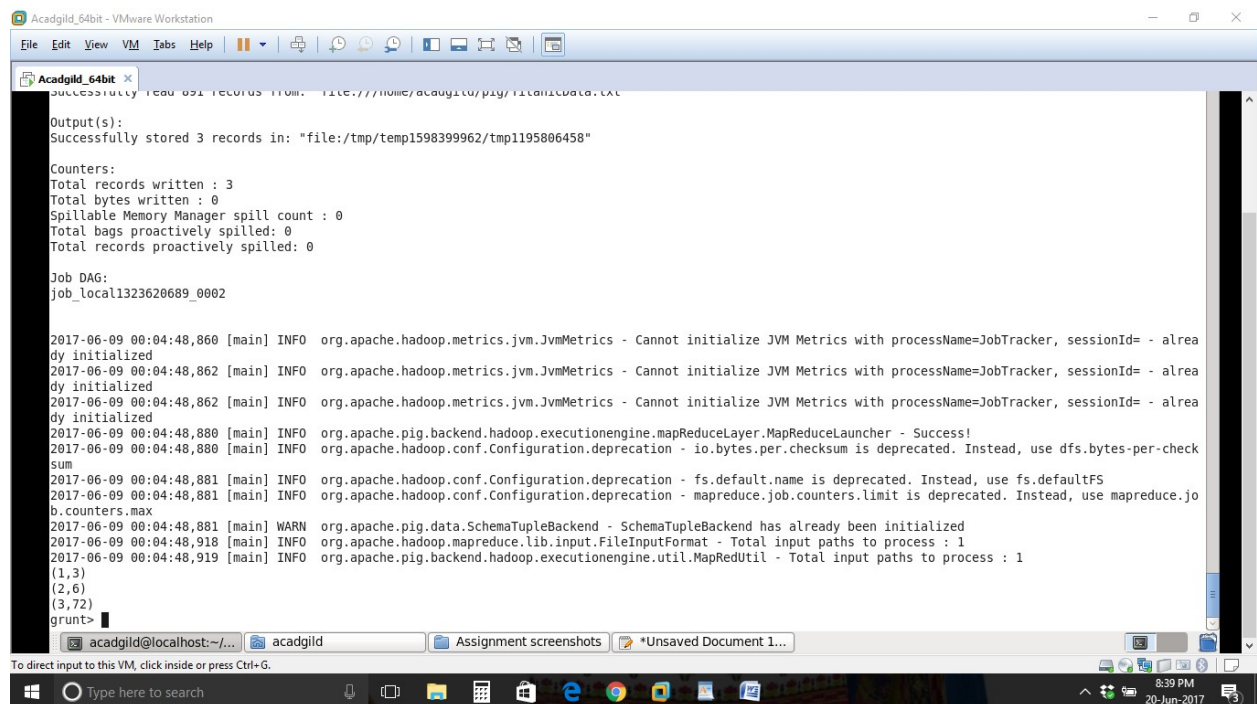
female\_pass\_died = filter pass\_died\_data by gender matches'female';  
(filtering details of only female passenger)

female\_died\_group = group female\_pass\_died by class;  
(group the female passenger died by class

female\_died\_count = foreach female\_died\_group generate group,  
COUNT(female\_pass\_died.gender);  
(generating count of male passenger died class wise)

dump female\_died\_count;

## Number of female passengers died in each class



```
Acadgild_64bit - VMware Workstation
File Edit View VM Tabs Help
Acadgild_64bit x
Successfully read 631 records from: file:///home/acadgild/pig/flightdata.txt
Output(s):
Successfully stored 3 records in: "file:///tmp/temp1598399962/tmp1195806458"

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1323620689_0002

2017-06-09 00:04:48,860 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-09 00:04:48,862 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-09 00:04:48,862 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alrea
dy initialized
2017-06-09 00:04:48,880 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-06-09 00:04:48,880 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-check
Sum
2017-06-09 00:04:48,881 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-06-09 00:04:48,881 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.jo
b.counters.max
2017-06-09 00:04:48,881 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-06-09 00:04:48,918 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-06-09 00:04:48,919 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,3)
(2,6)
(3,72)
grunt>
```