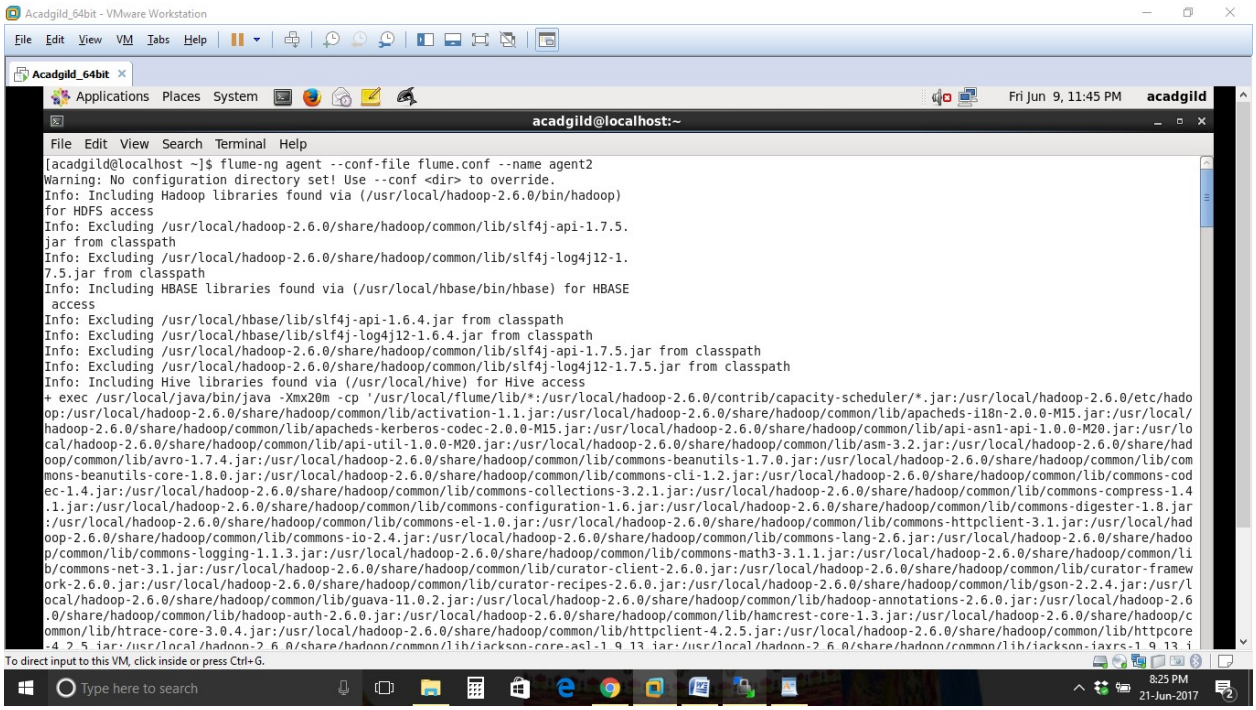
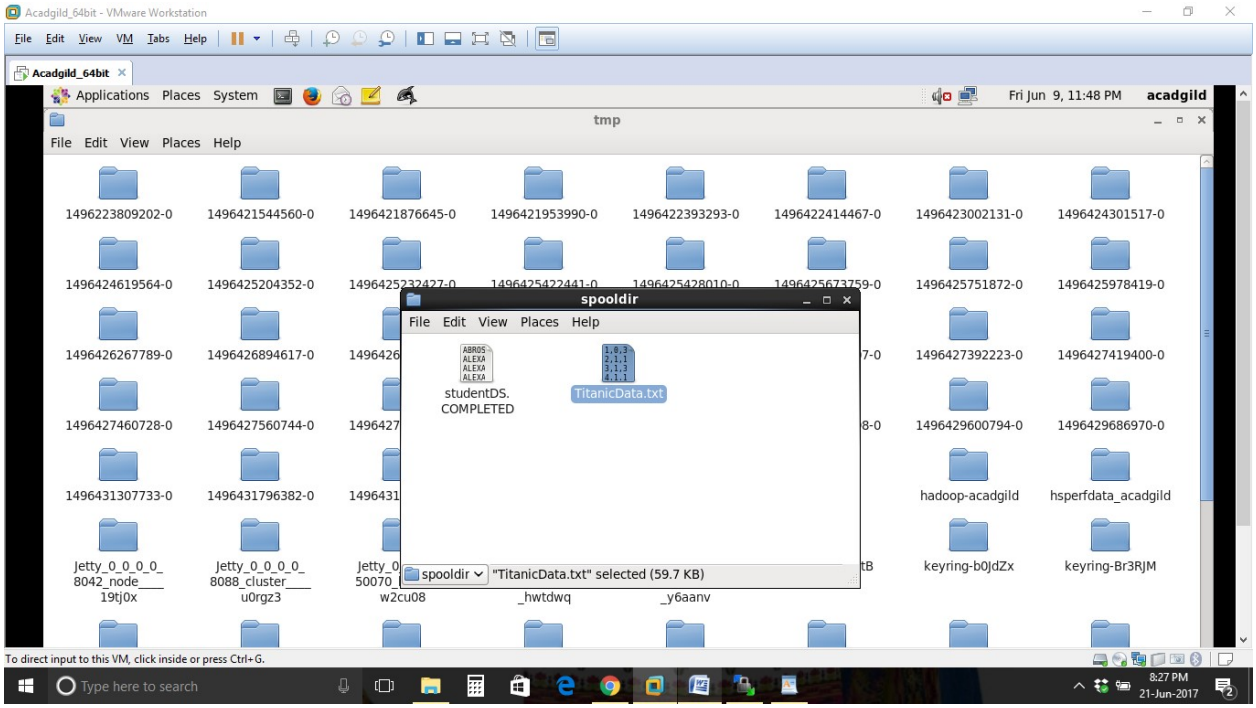
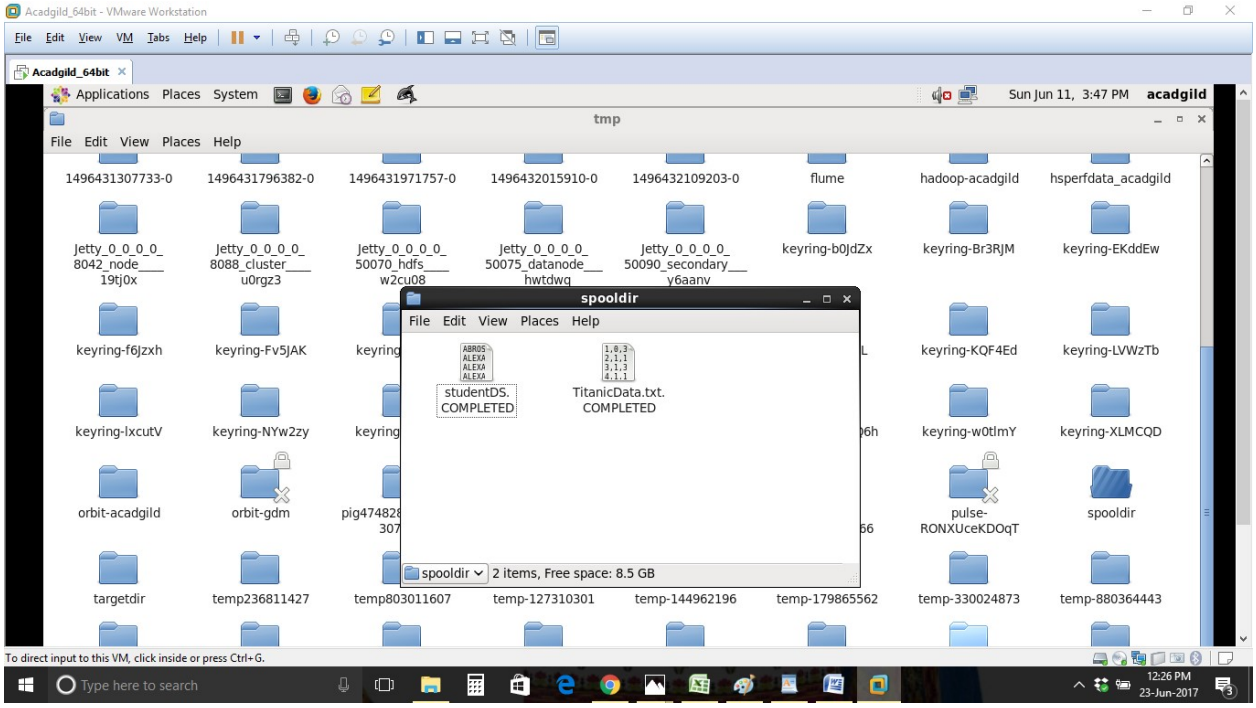
Project 1.2

Titanic Data Analysis

Copy the data set into HDFS using Flume




Acadgild_64bit - VMware Workstation
File Edit View VM Tabs Help
Acadgild_64bit
Applications Places System
Fri Jun 9, 11:48 PM
acadgild
tmp
File Edit View Places Help
1496223809202-0
1496421544560-0
1496421876645-0
1496421953990-0
1496422393293-0
1496422414467-0
1496423002131-0
1496424301517-0
1496424619564-0
1496425204352-0
1496425751872-0
1496425978419-0
1496426267789-0
1496426894617-0
1496427392223-0
1496427419400-0
1496427460728-0
1496427560744-0
1496429600794-0
1496429686970-0
1496431307733-0
1496431796382-0
hadoop-acadgild
hsperfdata_acadgild
Jetty_0_0_0_0_8042_node____19tj0x
Jetty_0_0_0_0_8088_cluster____u0rgz3
keyring-b0JdZx
keyring-Br3RJM
spooldir
File Edit View Places Help
studentDS.COMPLETED
TitanicData.txt
spooldir "TitanicData.txt" selected (59.7 KB)
To direct input to this VM, click inside or press Ctrl+G.
Type here to search
8:27 PM
21-Jun-2017


Acadgild_64bit - VMware Workstation
File Edit View VM Tabs Help
Acadgild_64bit
Applications Places System
Sun Jun 11, 3:47 PM
acadgild
tmp
File Edit View Places Help
1496431307733-0
1496431796382-0
1496431971757-0
1496432015910-0
1496432109203-0
flume
hadoop-acadgild
hsperfdata_acadgild
Jetty_0_0_0_0_8042_node____19tj0x
Jetty_0_0_0_0_8088_cluster____u0rgz3
Jetty_0_0_0_0_50070_hdfs____w2cu08
Jetty_0_0_0_0_50075_datanode___hwtdwq
Jetty_0_0_0_0_50090_secondary___y6aanv
keyring-b0JdZx
keyring-Br3RJM
keyring-EKddEw
keyring-f6Jzxh
keyring-Fv5JAK
keyring-KQF4Ed
keyring-LVWzTb
keyring-lxcutV
keyring-NYw2zy
keyring-w0tlmY
keyring-XLMCQD
orbit-acadgild
orbit-gdm
pulse-RONXUceKDOqT
spooldir
spooldir
File Edit View Places Help
studentDS.COMPLETED
TitanicData.txt.COMPLETED
spooldir 2 items, Free space: 8.5 GB
targetdir
temp236811427
temp803011607
temp-127310301
temp-144962196
temp-179865562
temp-330024873
temp-880364443
To direct input to this VM, click inside or press Ctrl+G.
Type here to search
12:26 PM
23-Jun-2017

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.

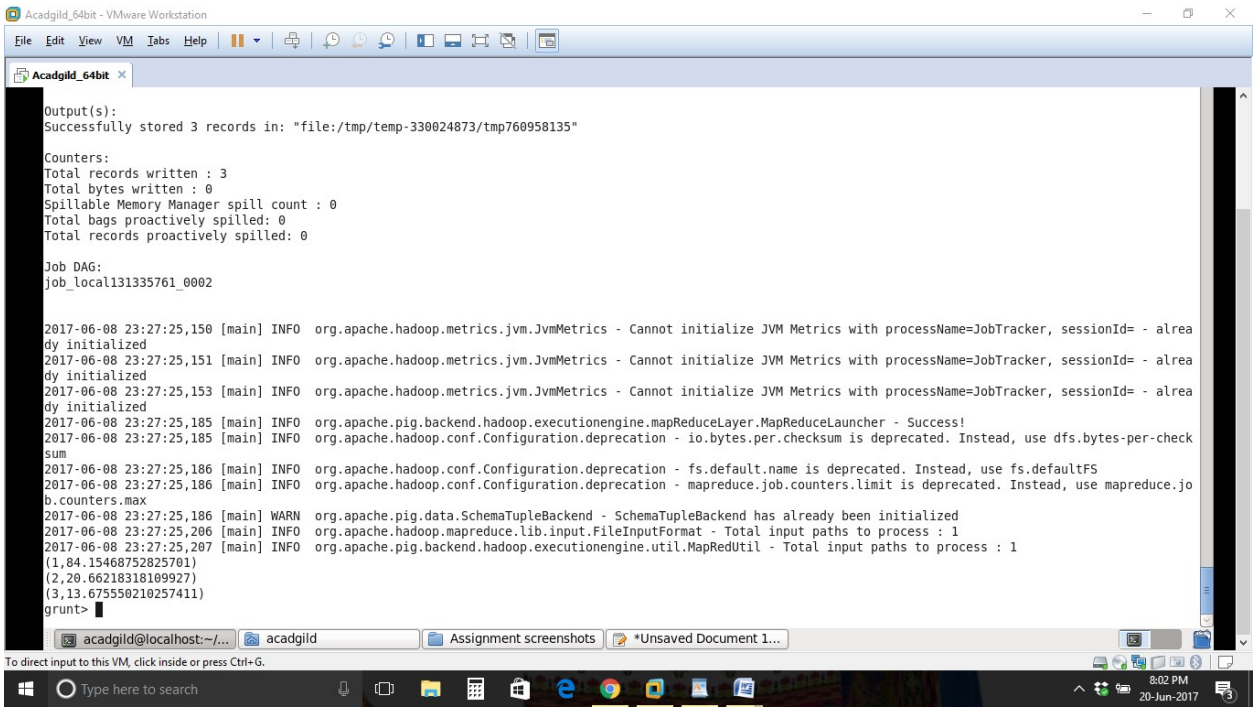
fare_class = foreach titanic_data generate class,fare**;(generate only class & fare details)**

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



Problem Statement

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

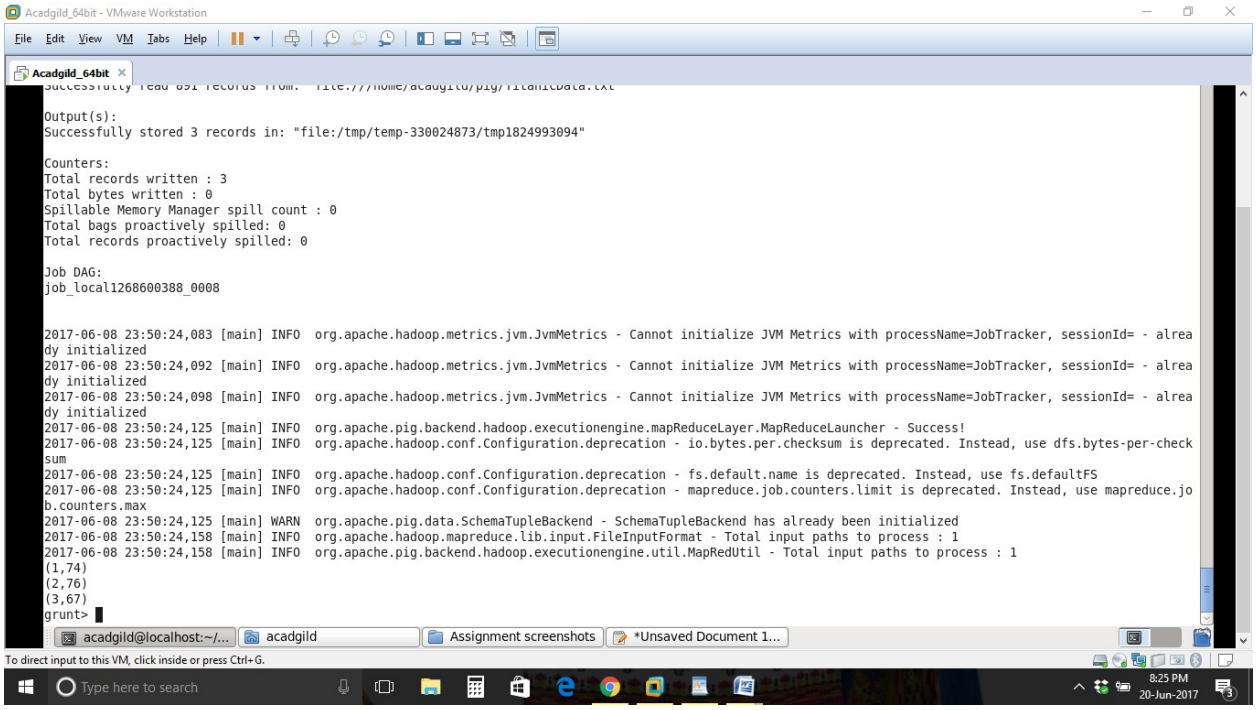
alive_southampton = filter titanic_data by embarked matches'S' and survived==1; **(filter data of passengers who embarked in Southampton and are alive)**

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



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)

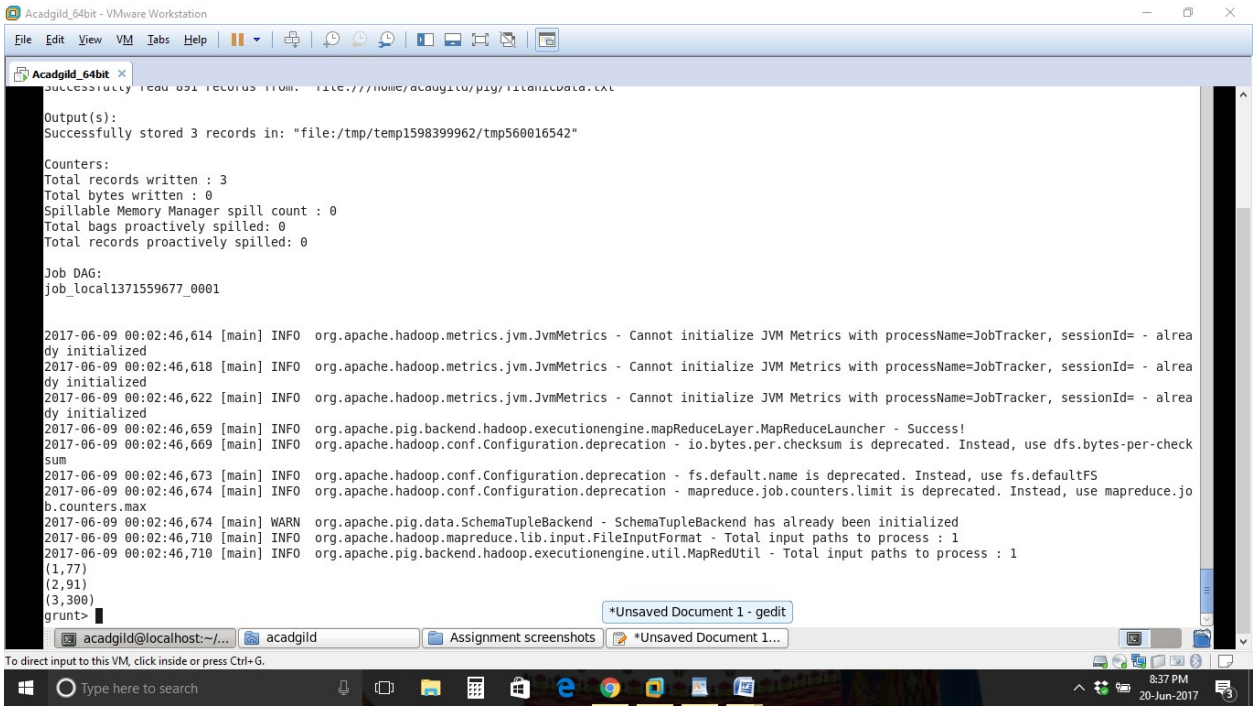
male_pass_died = filter pass_died_data by gender matches'male';
(filtering details of only male passenger)

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



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