**Titanic Data Analysis**

**1. Associated Data Files**

https://drive.google.com/file/d/0ByJLBTmJojjzNmV0dk1EMmwwQ1U/view?usp=sharing

**DATA SET DESCRIPTION**

Column 1 : PassengerId

Column 2 : Survived (survived=0 & died=1)

Column 3 : Pclass

Column 4 : Name

Column 5 : Sex

Column 6 : Age

Column 7 : SibSp

Column 8 : Parch

Column 9 : Ticket

Column 10 : Fare

Column 11 : Cabin

Column 12 : Embarked

**2. Problem Statements**

You can use any of the technologies like Map Reduce, Pig or Hive of your choice.

Note: You need to copy the data set into HDFS using flume and send the screen shot of that with the project solution

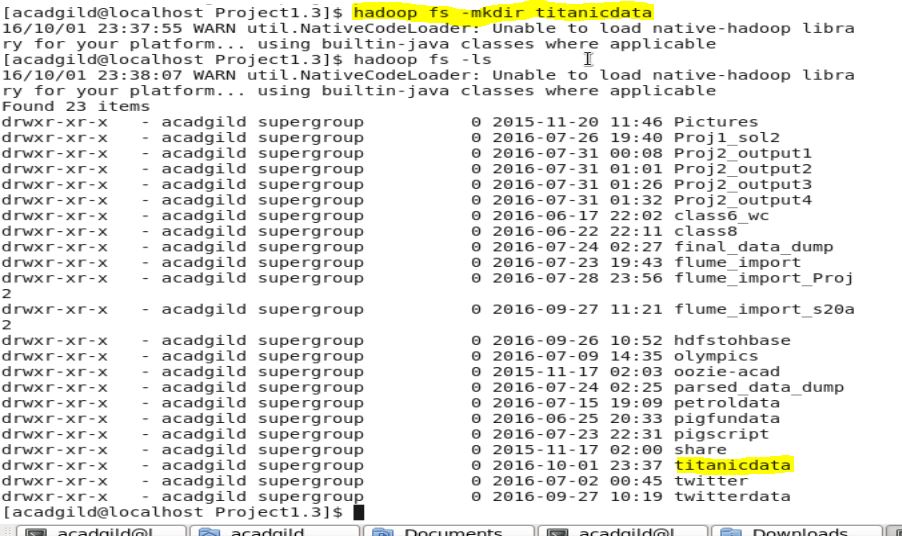
1. In this problem statement we will find the average fare of each class.

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

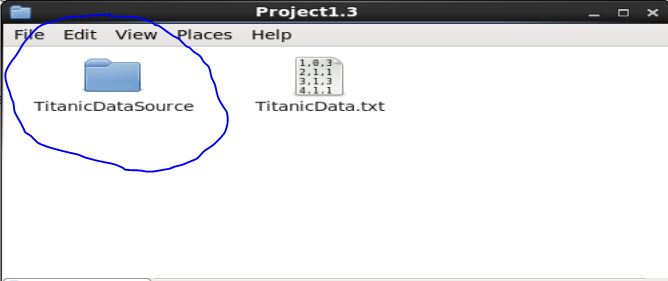
3. In this problem statement we will find out number of male and female people died in each class.

**Solution**

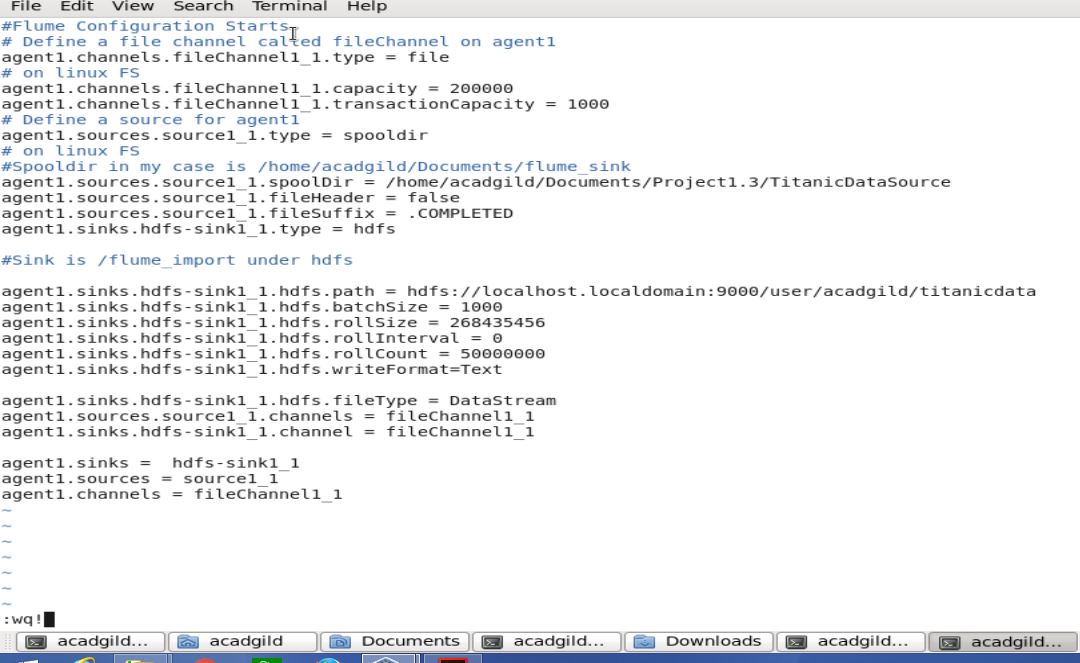
Created directory "titanicdata" to store the data in HDFS, which will act as a sink while transferring data using flume.

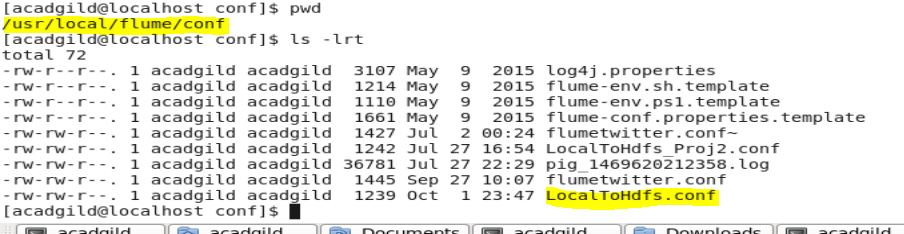
****

Created "TitanicDataSource" folder on local drive which will act as a source to transfer data using flume.



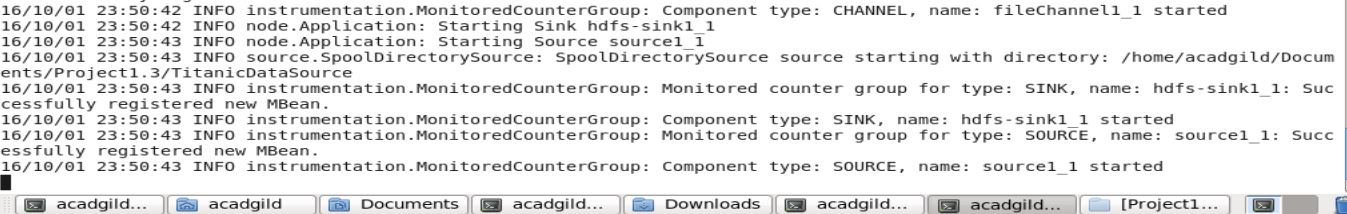
Created configuration file for data transfer using flume.



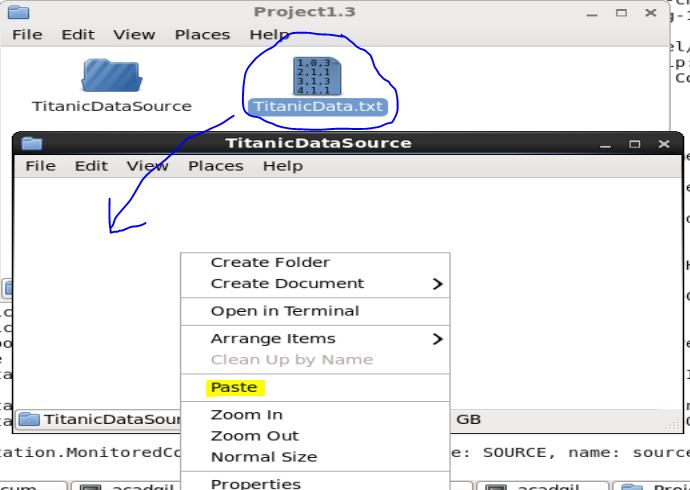


Started flume agent.

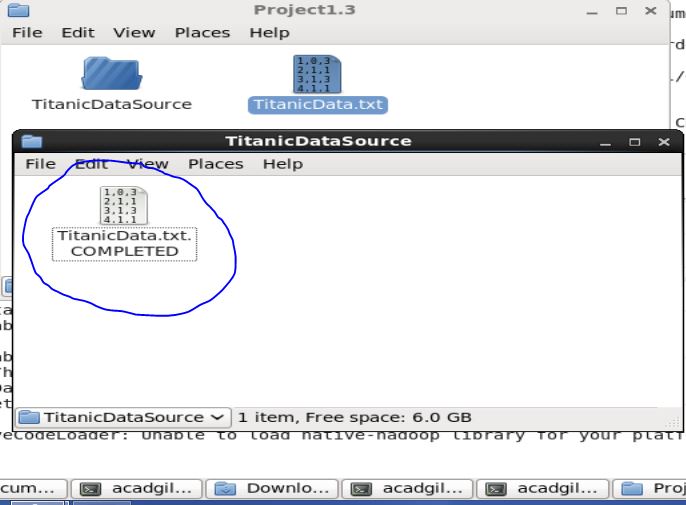
C:\Users\RAKESH\Desktop\Screenshot\Project1.3\S9.JPG



Copied the data file to source folder.



The data file's extension changed to COMPLETED as mentioned in configuration file. It means that data has loaded into HDFS.



**Problem 1:**

Created the below Pig script and ran using following command in MapReduce mode:

**C:\Users\RAKESH\Desktop\Screenshot\Project1.3\S13.JPG**

Pig Script:

rawdata = LOAD '/user/acadgild/titanicdata' USING PigStorage(',') AS (passengerid: int,survived: int,pclass: int,name: chararray,sex: chararray,age: int,sibsp: int,parch: int,ticket: chararray,fare: float,cabin: chararray,embarked: chararray);

generate\_class = FOREACH rawdata GENERATE pclass,fare;

group\_class = GROUP generate\_class by pclass;

avg\_fare = FOREACH group\_class GENERATE group as pclass, AVG(generate\_class.fare);

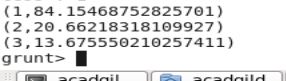
dump avg\_fare;

Output:

(1,84.15468752825701)

(2,20.66218318109927)

(3,13.675550210257411)

****

**Problem 2:**

Created the below Pig script and ran using following command in MapReduce mode:

C:\Users\RAKESH\Desktop\Screenshot\Project1.3\S14.JPG

Pig Script:

rawdata = LOAD '/user/acadgild/titanicdata' USING PigStorage(',') AS (passengerid: int,survived: int,pclass: int,name: chararray,sex: chararray,age: int,sibsp: int,parch: int,ticket: chararray,fare: float,cabin: chararray,embarked: chararray);

generate\_filter = FILTER rawdata by survived == 0 AND embarked == 'S';

generate\_pclass = FOREACH generate\_filter GENERATE pclass;

group\_pclasss = GROUP generate\_pclass by pclass;

count\_people = FOREACH group\_pclasss GENERATE group as pclass, COUNT(generate\_pclass.pclass);

dump count\_people;

Output:

(1,53)

(2,88)

(3,286)



**Problem 3:**

Created the below Pig script and ran using following command in MapReduce mode:

C:\Users\RAKESH\Desktop\Screenshot\Project1.3\S15.JPG

Pig Script:

rawdata = LOAD '/user/acadgild/titanicdata' USING PigStorage(',') AS (passengerid: int,survived: int,pclass: int,name: chararray,sex: chararray,age: int,sibsp: int,parch: int,ticket: chararray,fare: float,cabin: chararray,embarked: chararray);

generate\_died = FILTER rawdata by survived == 1;

generate\_class = FOREACH generate\_died GENERATE pclass,sex;

filter\_male = FILTER generate\_class by sex == 'male';

group\_male = GROUP filter\_male by pclass;

count\_male = FOREACH group\_male GENERATE group as pclass,COUNT(filter\_male.sex) AS m;

filter\_female = FILTER generate\_class by sex == 'female';

group\_female = GROUP filter\_female by pclass;

count\_female = FOREACH group\_female GENERATE group as pclass,COUNT(filter\_female.sex) AS f;

join\_count = JOIN count\_male by pclass, count\_female by pclass;

final\_count = FOREACH join\_count GENERATE count\_male::pclass,m,f;

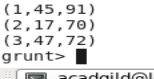
dump final\_count;

Output:

(1,45,91)

(2,17,70)

(3,47,72)

****