**BigData And Hadoop**

**Project1.2 -** Titanic Data Analysis

**Task1**

**Dataset Description:**

The data set contains information about passengers who boarded Titanic ship.

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

**Problem Statement:**

Write a MapReduce/Pig program to find the average fare of each class.

**Solution**

**Code files are as follows:**

Mapper class: Task1Mapper.java

Reducer class: Task1Reducer.java

Combiner class: Task1Reducer.java

Driver class: Task1Driver.java

**Solution logic:**

We need to find average fare for each class.

In the data, class is present in column 3 and fare is present in column 10 of the records.

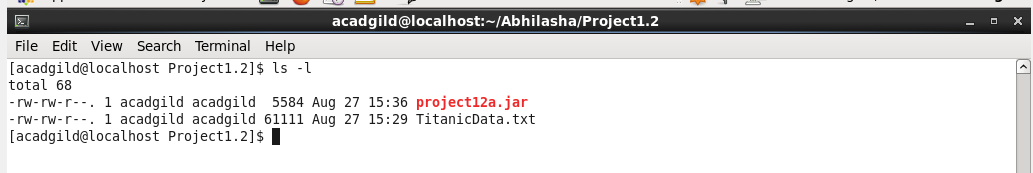
In mapper, we emit the class as the key and the fare as the value. Class is used as key as we need to find average fare for every class and so we will group by class.

In reducer, we find the average fare for every class, i.e., we sum up the fare for all the records of a class and divide it by number of records.

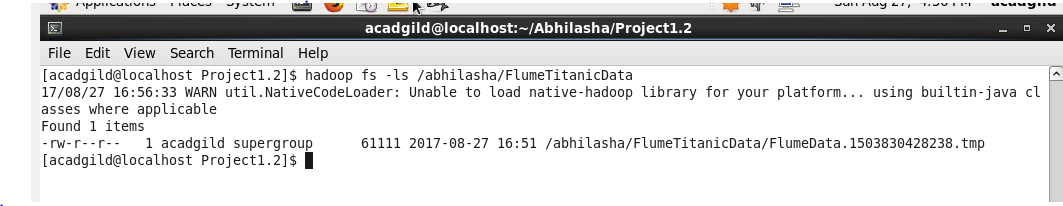
We have used combiner as well, to improve performance, which is same as the reducer class.

**Snapshots of the output are as follows:**

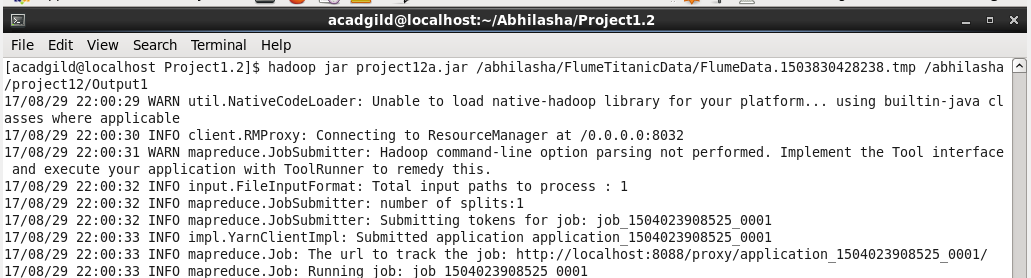
1. The jar is compiled and placed at location ‘/home/acadgild/Abhilasha/Project1.2’. Name of the jar is project12a.jar



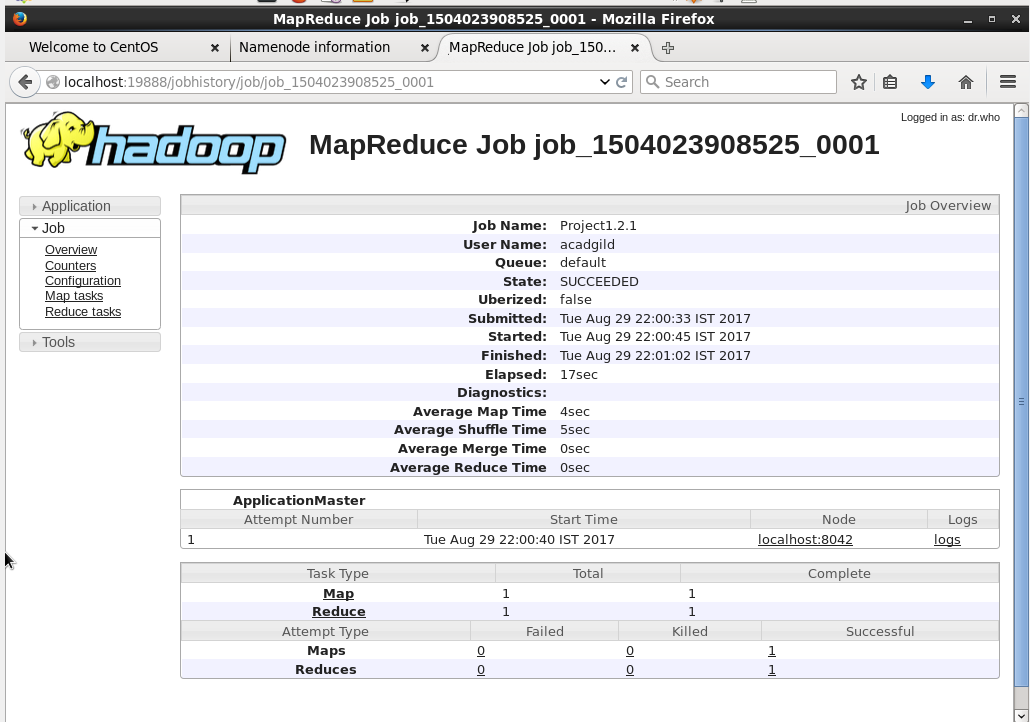
1. Dataset to be used is put into hdfs using flume at the location ‘/abhilasha/FLumeTitanicData’ as follows:



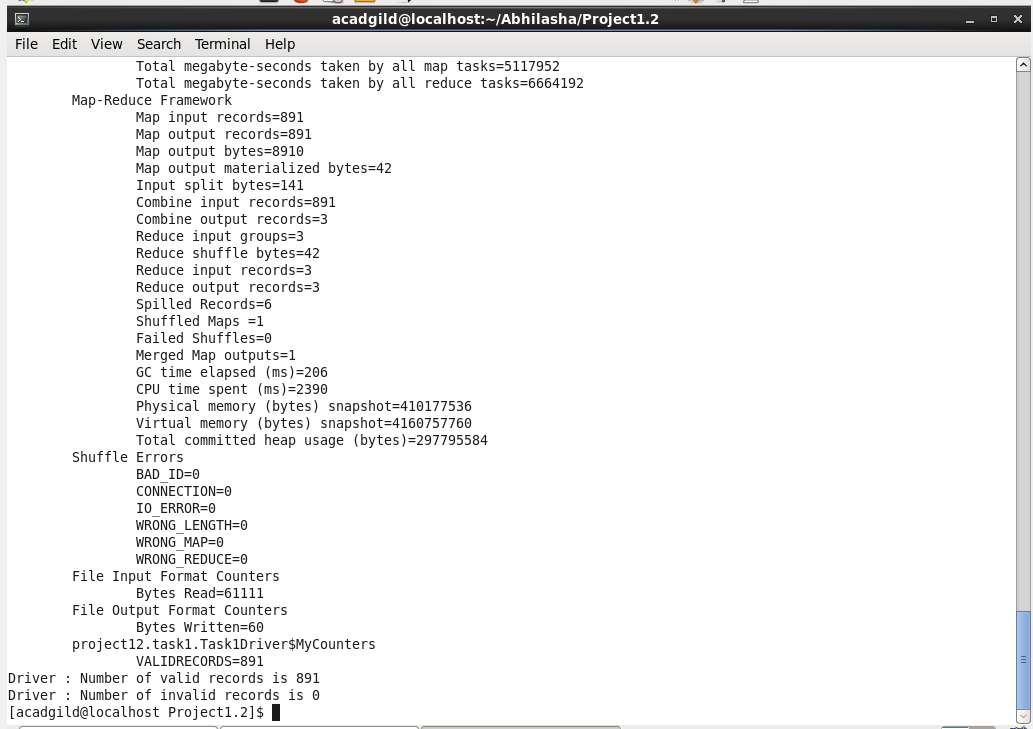
1. We now execute the map-reduce code as follows. In this, we are using the dataset put in hdfs using flume as input. The output folder mentioned is ‘/abhilasha/project12/Output1’



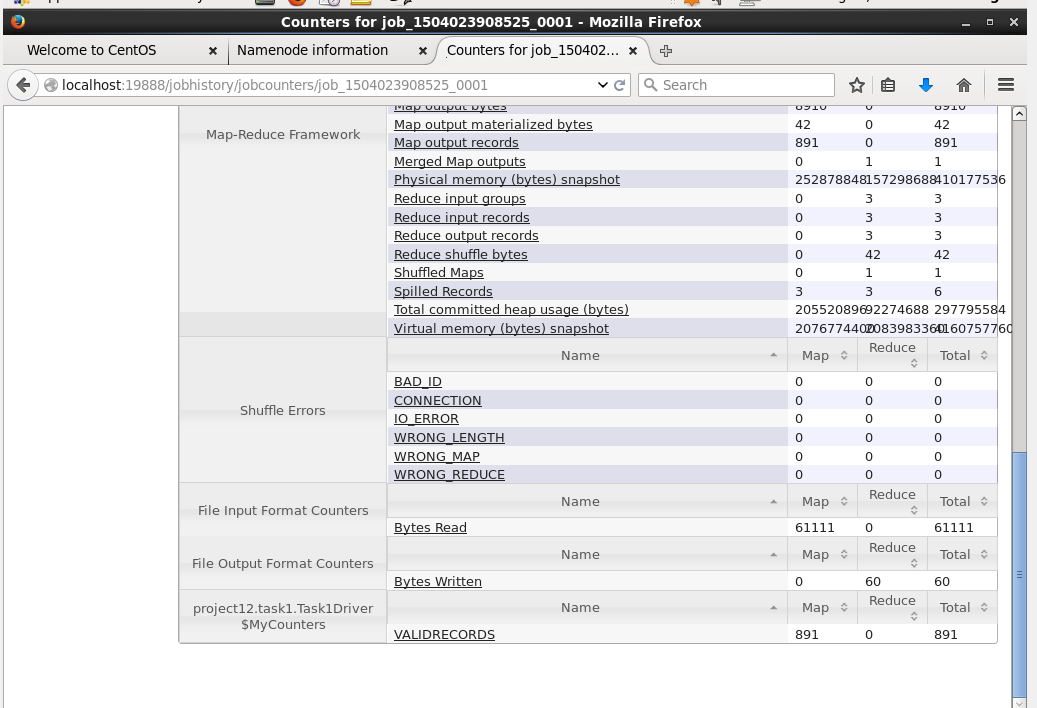
1. Job’s successful completion is shown on the job history server as follows:



1. We used custom counters in the code, their values are as shown in the screen shot. We used these counters to find number of valid and invalid records. Invalid are the ones that have missing columns in the record or that have empty value in column of fare and class. No invalid records were found in this dataset.



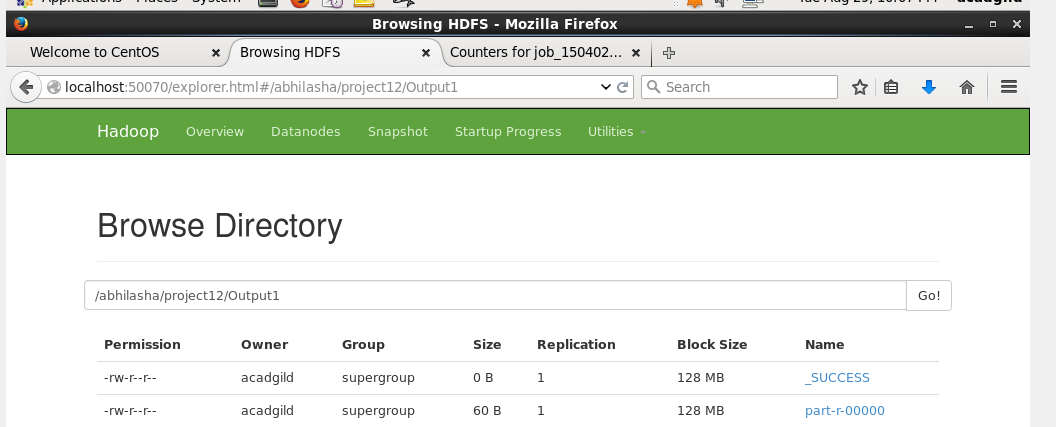
These custom counters are also visible on history server in Counters section as follows:



1. Listing the directory Output1, which contains output of map-reduce as follows:



We can also see this from HDFS UI as follows:



1. The output of the map-reduce program is as follows:

We have used cat command to display the output file ‘/abhilasha/project12/Output1/part-r-00000’ on the terminal.

Format of the output is

<Class> <Count>

