2. Problem Statement

1) What is the distribution of the total number of air-travelers per year

2) What is the total air distance covered by each user per year

3) Which user has travelled the largest distance till date

4) What is the most preferred destination for all users.

*val travel = sc.textFile("file:///home/cloudera/18/S18\_Dataset\_Holidays.txt").map(x => {*

*val row = x.split(",");*

*(row.apply(0), row.apply(1), row.apply(2), row.apply(3), row.apply(4).toInt, row.apply(5).toInt)*

*})*

*val f1 = travel.map(x => x.\_6 -> 1).groupByKey().map(x => x.\_1 -> x.\_2.sum)*

*f1.collect.foreach(println)*

*val f2 = travel.map(x => x.\_1 -> x.\_5).groupByKey().map(x => x.\_1 -> x.\_2.sum)*

*f2.collect.foreach(println)*

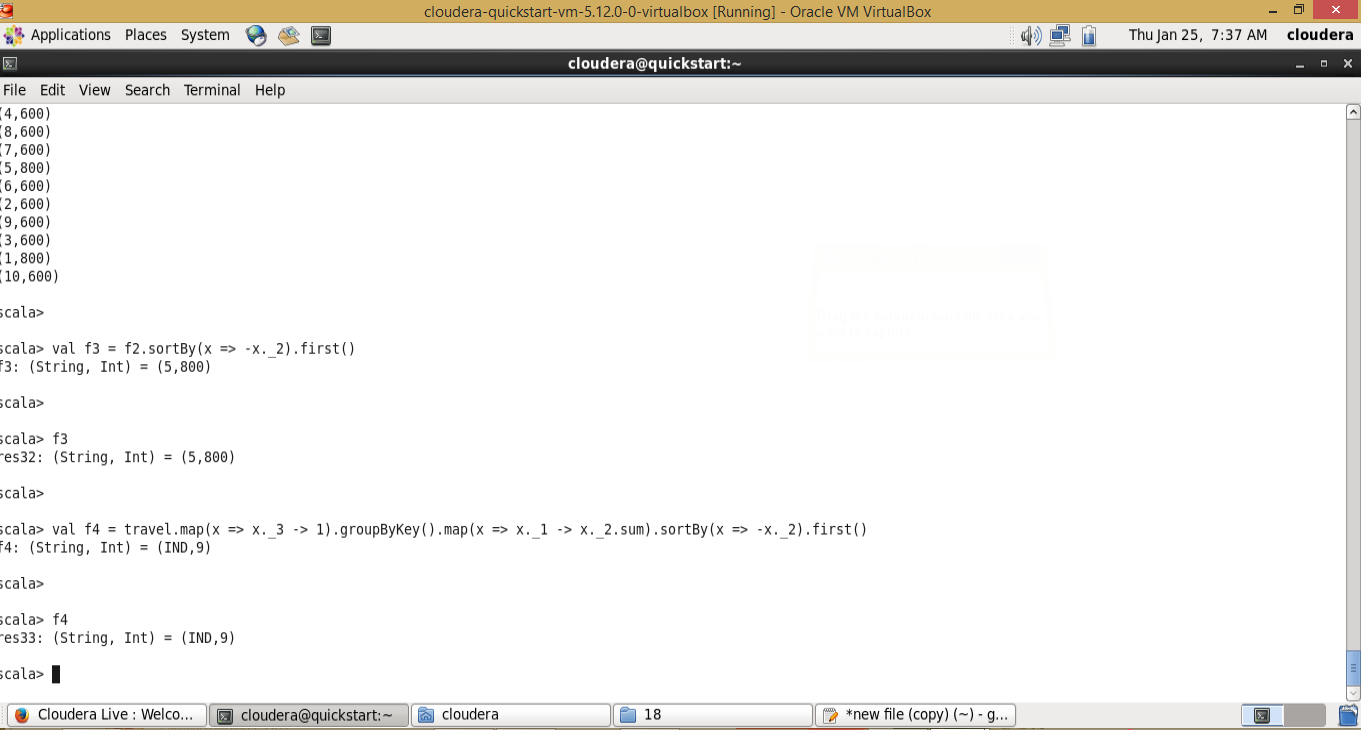
*val f3 = f2.sortBy(x => -x.\_2).first()*

*f3*

*val f4 = travel.map(x => x.\_3 -> 1).groupByKey().map(x => x.\_1 -> x.\_2.sum).sortBy(x => -x.\_2).first()*

*f4*





Use below link to download the dataset:

https://drive.google.com/drive/folders/0B\_P3pWagdIrrVThBaUdVSUtzbms