Solution CA3

X13112406 (Navjot Singh Virk)

**Using the Train Dataset (train.csv)**

**How was the dataset prepared for MapReduce**

Downloaded the file and took its first row that is its header and put it to use in script files header too allow/help read the csv file (dataset)

*# Data header: "PassengerId" "Survived" "Pclass" "Name" "Sex" "Age" "SibSp" "Parch" "Ticket" "Fare" "Cabin" "Embarked"*

**The Pattern you are using and why ?**

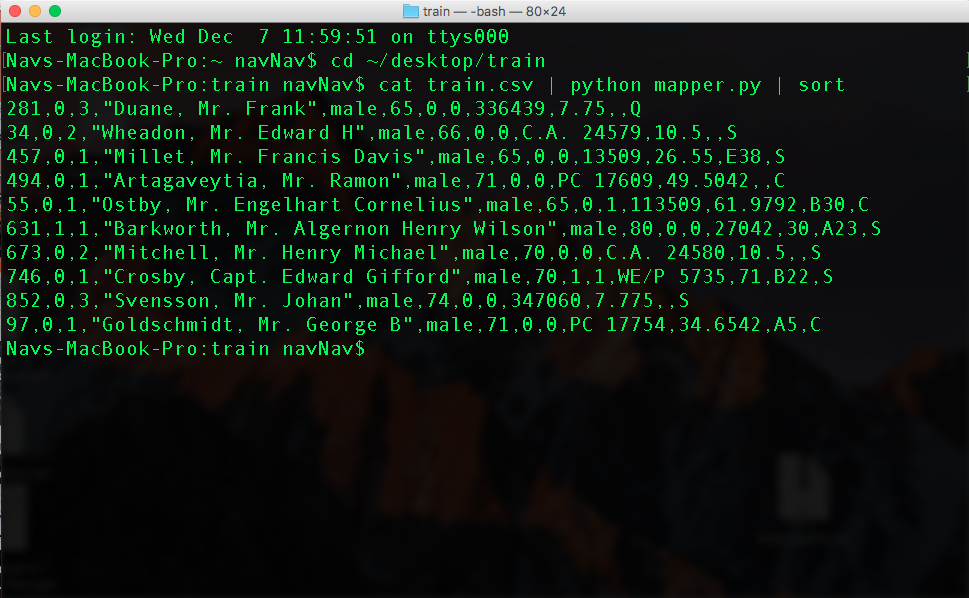
The pattern I am using is finding top 10 passengers by age, because doing this will help us know that what is the age group of the passengers that is travelling the most.

**What was result of executing the pattern ?**

Results were as follows

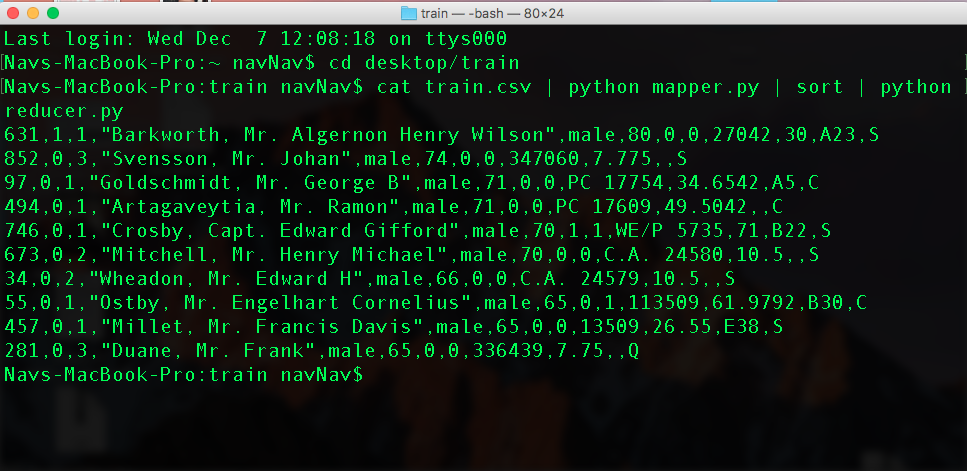
Passing dataset through mapper (the mapper maps the top 10 passengers by age)

cat train.csv | python mapper.py | sort



Passing File through reducer.py

cat train.csv | python mapper.py | sort | python reducer.py



And the results are outputted to a file with this command

cat train.csv | python mapper.py | sort | python reducer.py >> outputResults.txt

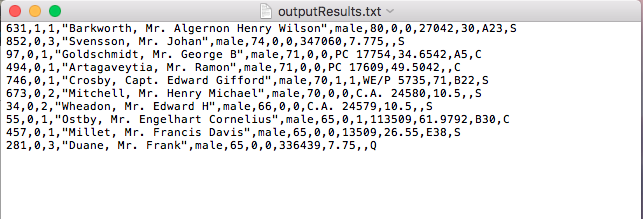
How would this pattern be used for data analytics project on this data set ?

In general i.e in real life, finding the top 10 passengers by age will allow us to know –

1. Which passenger travelled the most
2. Their age group (hence we can target their age group for sales)
3. Can provide them with custom offers and much more

So, in project this kind of knowledge will be of great use.

Output



more Screenshots are listed above.