Description

The goal of this assignment is to implement two MapReduce programs in Java (using Apache Hadoop). Specifically, your MapReduce jobs will analyzing a data set consisting of New York City Taxi trip reports in the Year 2013.

The data set itself is a set of simple text files. Each taxi trip report is a different line in a file. The attributes present on each line of the files are, in order:

attribute	description
`medallion`	an md5sum of the identifier of the taxi - vehicle bound (Taxi ID)
`hack_license`	an md5sum of the identifier for the taxi license (driver ID)
`vendor_id`	identifies the vendor
`pickup_datetime`	time when the passenger(s) were picked up
`payment_type`	the payment method -credit card or cash
`fare_amount`	fare amount in dollars
`surcharge`	surcharge in dollars
`mta_tax`	tax in dollars
`tip_amount`	tip in dollars
`tolls_amount`	bridge and tunnel tolls in dollars
`total_amount`	total paid amount in dollars

The data files are in comma separated values (CSV) format. This is about 20 GB of data in all. For testing and development, a super-small subset of the first file (only 1000 lines) is available.

1. Task 1

Write a MapReduce program that checks all of the files and computes the total amount of revenue (total dollars) for each date present in the data set.

2. Task 2

Write a MapReduce program that computes 5 taxi drivers who had the most revenue in the data set.