MapReduce

Suppose that you have a dataset from a sport retailer/shop that has several branches in in different cities. The data set contains the following information: retailer name, invoice data, city name, product name, price per product unit, and the amount sold per product.

Here is the sample of the data

Retailer, Invoice Date, City, Product, Price per Unit, Units Sold
Foot Locker, 1/2/20, New York, Men's Athletic Footwear, $50.00, 1,000
Foot Locker, 1/4/20, New York, Women's Athletic Footwear, $45.00, 850
Walmart, 4/19/20, New York, Men's Street Footwear, $60.00, 1,200
Walmart, 4/21/20, New York, Women's Street Footwear, $50.00, 925
Sports Direct, 7/25/20, Houston, Men's Street Footwear, $25.00, 850
Sports Direct, 8/7/20, Houston, Men's Athletic Footwear, $40.00, 900

Using HDFS commands , show how to do the following

- create the input directory; name it ‘sportRetailer’
- the sample data above, add it to a text file name it ‘salesData.txt’, put this file in a directory called ‘input’ in the sportsRetailer on HDFS
- browse the content of the salesData.txt on HDFS

Write a MapReduce job that reads the input data and produces the following “retailer, city”, totalSalesAmount, and the output must be sorted in descending order based on sales amount

The output must be stored in a director called ‘output’ in the ‘sportRetailer’ directory

Using HDFS commands, show to do the following:

- run the job
- browse the content of ‘output’ directory
- show the content of the output files

What to submit

- A pdf files the contains HDFS commands and the full code of the MapReduce job

- You need to record your work with your voice, discussing everything you were asked to do above; HDFS commands, discuss the code, show how to run the job. **The recording must not exceed 5 minutes**

- For the MapReduce part you can write code on any IDE, **and you must run it on the VM.**

    - If it not on the VM you loose **three marks**, but still your code has to be in Java project that has all dependencies.