Lab Test - Introduction to Hadoop

You are given a file that is a snapshot of NYSE daily data. The file is located at

Firstly, wipe off your VM and start a new one.

They consist of the following fields:

The first three columns are:
exchange, *symbol* and *date*, which records the stock exchange firm, stock symbol and the date of the entry.

The other columns are the data, which includes:
open, *high*, *low*, *close*, *volume*, and *adj_close*.

Task 1:

- Import the data file using Hive
- Set the columns and their Column Type accordingly.

Answer the following question:

Q1. How many stocks (i.e. symbols) are listed in NYSE?
Q2. What is the average, minimum and maximum of open and close for stock “CRT”?
Q3. Which stock has the highest opening price?
Q4. Which stock has the lowest opening price?
Q5. What is the total volume transacted for “CVA”

Email the query for all 5 questions.

Task 2:

- Load “nyse_daily.csv” using Pig
- Group the records by Symbol (email the top 10 entries).
- Group the records by the Symbol, and their maximum volume (email your top 10 entries)
- Email your final pig script.