**Part I**

Now it's time for you to have a go at this. For starters you will have to work with the same data set - sales data, that we discussed in lessons. You will have to write some Mappers and Reducers yourself and then answer the questions about data that follow. You will have to do the data processing on your local pseudo-distributed cluster, but you will be able to see if your solution was correct by submitting your results to our system.

The three questions that you have to answer about this data set are:

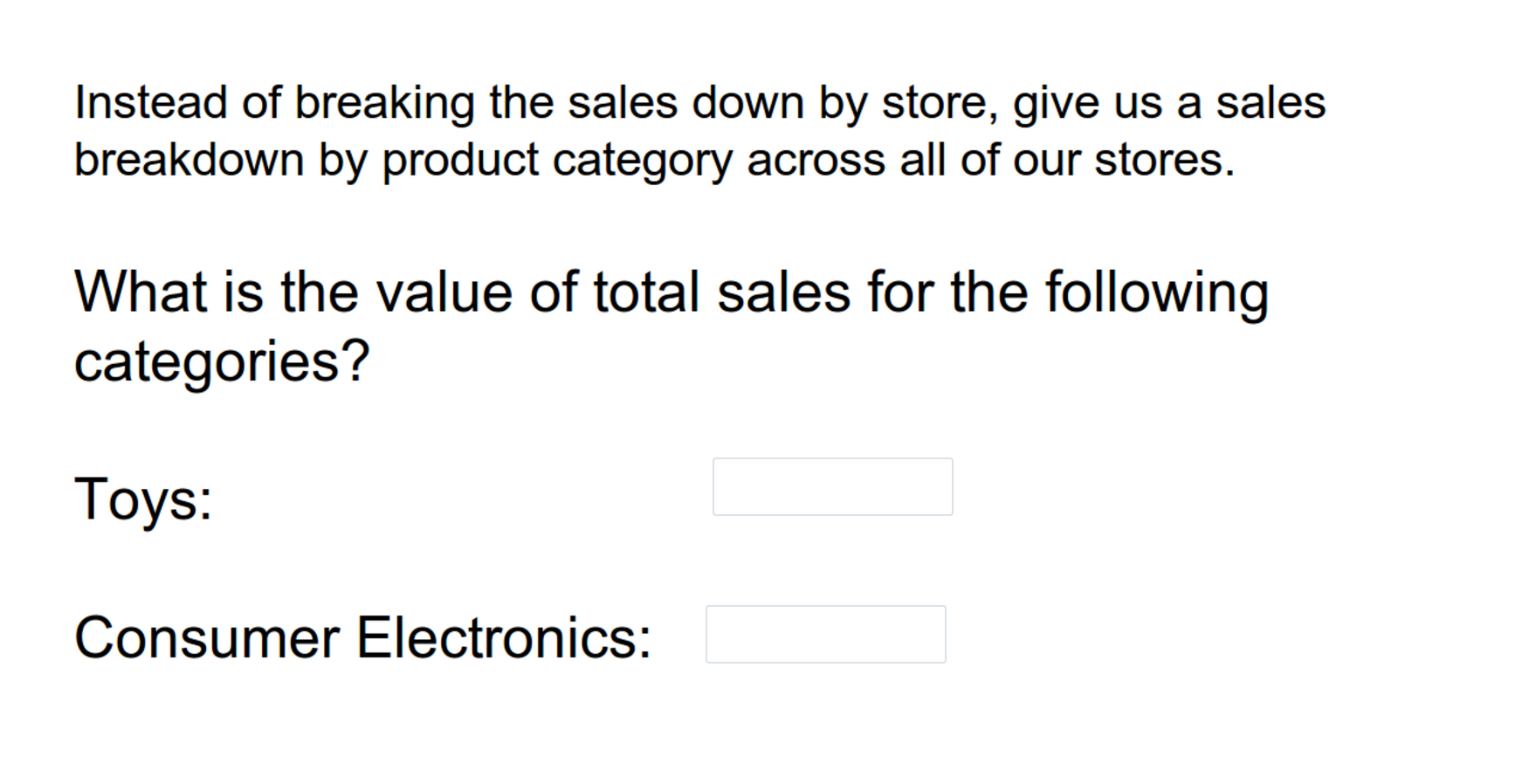
* Instead of breaking the sales down by store, instead give us a sales breakdown by product category across all of our stores.
* Find the monetary value for the highest individual sale for each separate store.
* Find the total sales value across all the stores, and the total number of sales. Assume there is only one reducer.

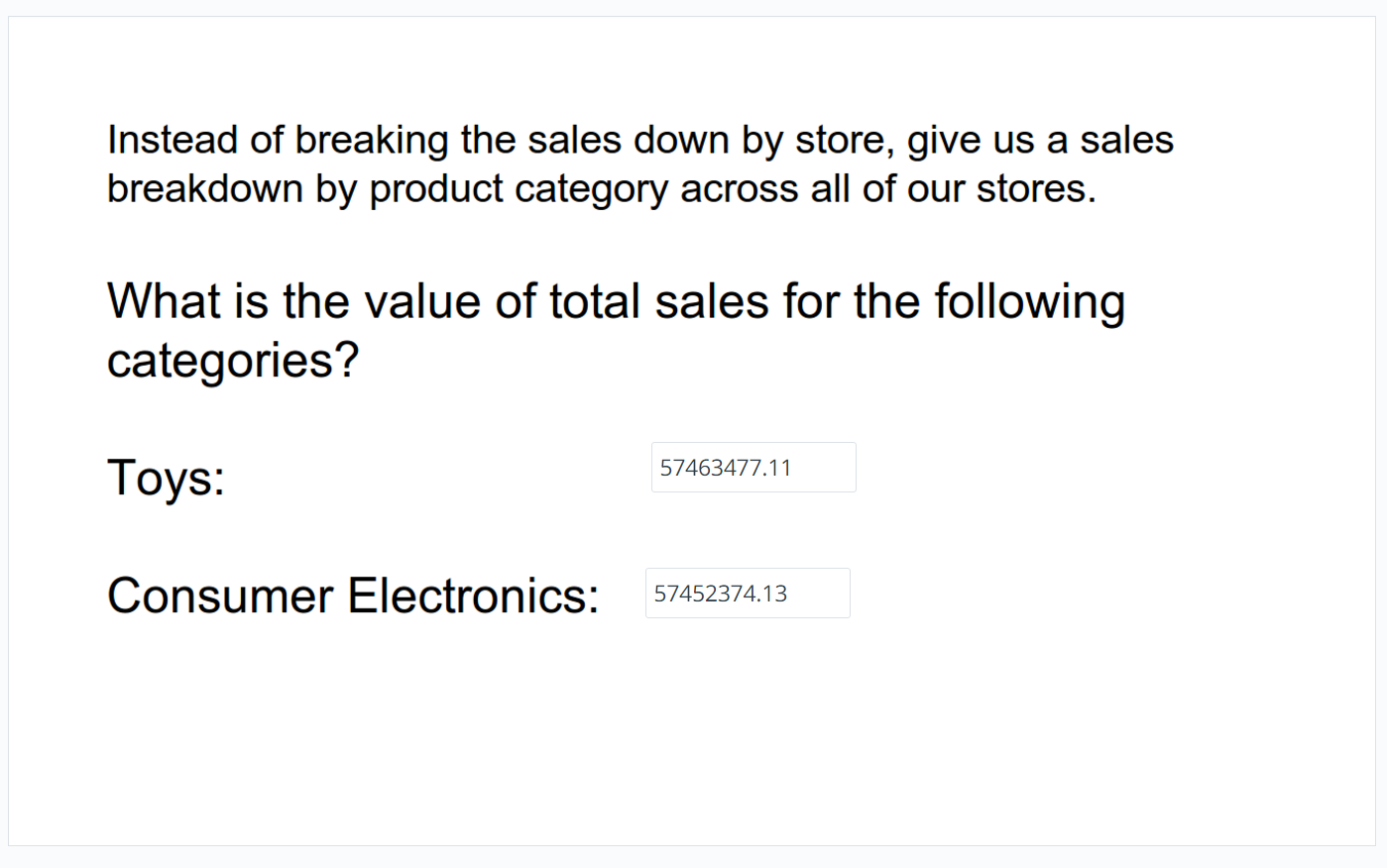
When you have finished writing running your mapreduce jobs, press 'Next' to submit and check the answers.

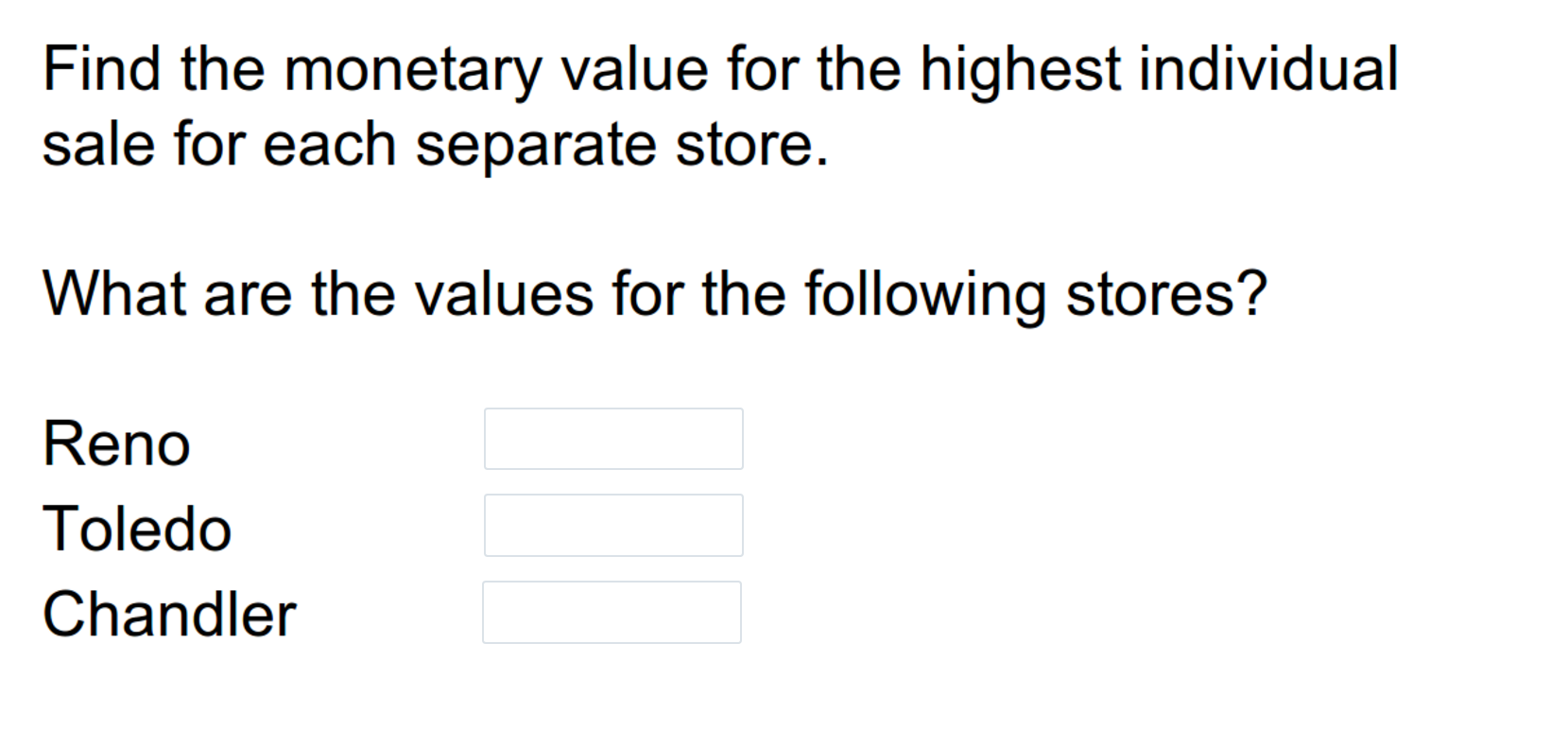
You can read instructions on how to download and run the virtual machines [**here**](https://docs.google.com/document/d/1v0zGBZ6EHap-Smsr3x3sGGpDW-54m82kDpPKC2M6uiY/edit?usp=sharing).

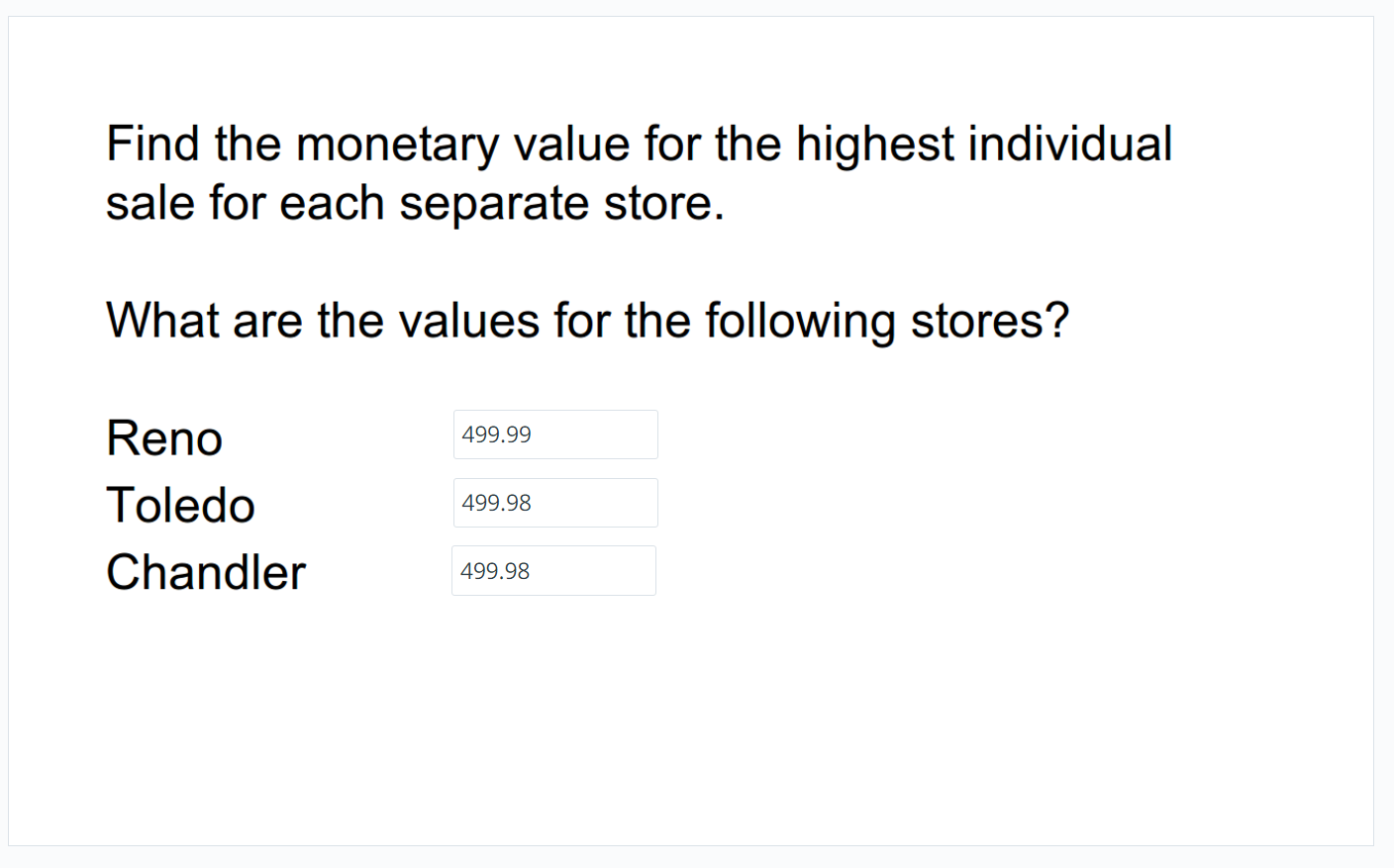
You can see the input setup example [**here**](https://classroom.udacity.com/courses/ud617/lessons/308873795/concepts/3095085570923), and the commands for running a job [**here**](https://classroom.udacity.com/courses/ud617/lessons/308873795/concepts/3093825950923).

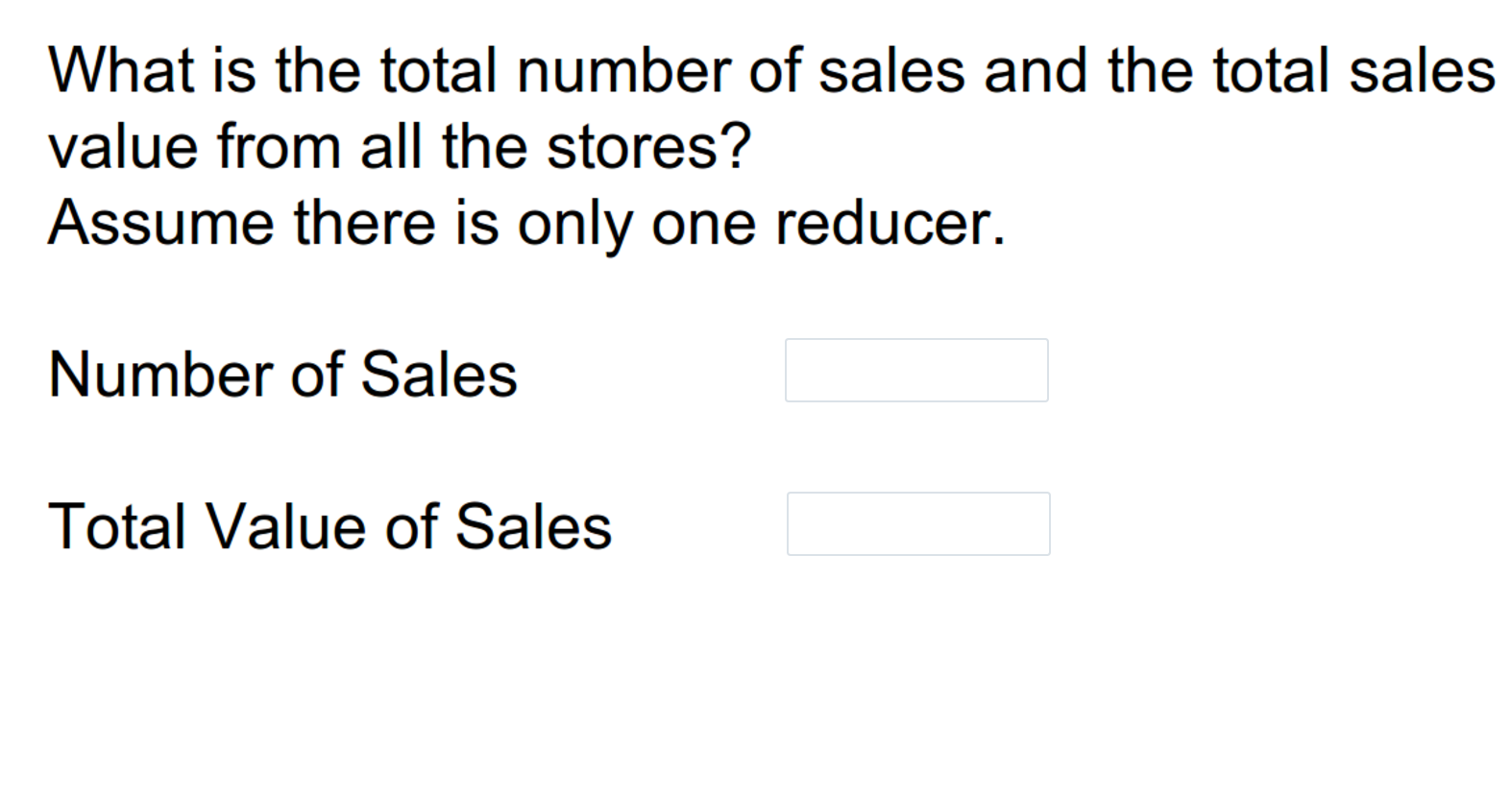
See the [**Python Reference**](https://www.udacity.com/wiki/cs101/%3A-python-reference) doc from [**Intro to Computer Science**](https://www.udacity.com/course/cs101) for a review of Python basics and syntax.

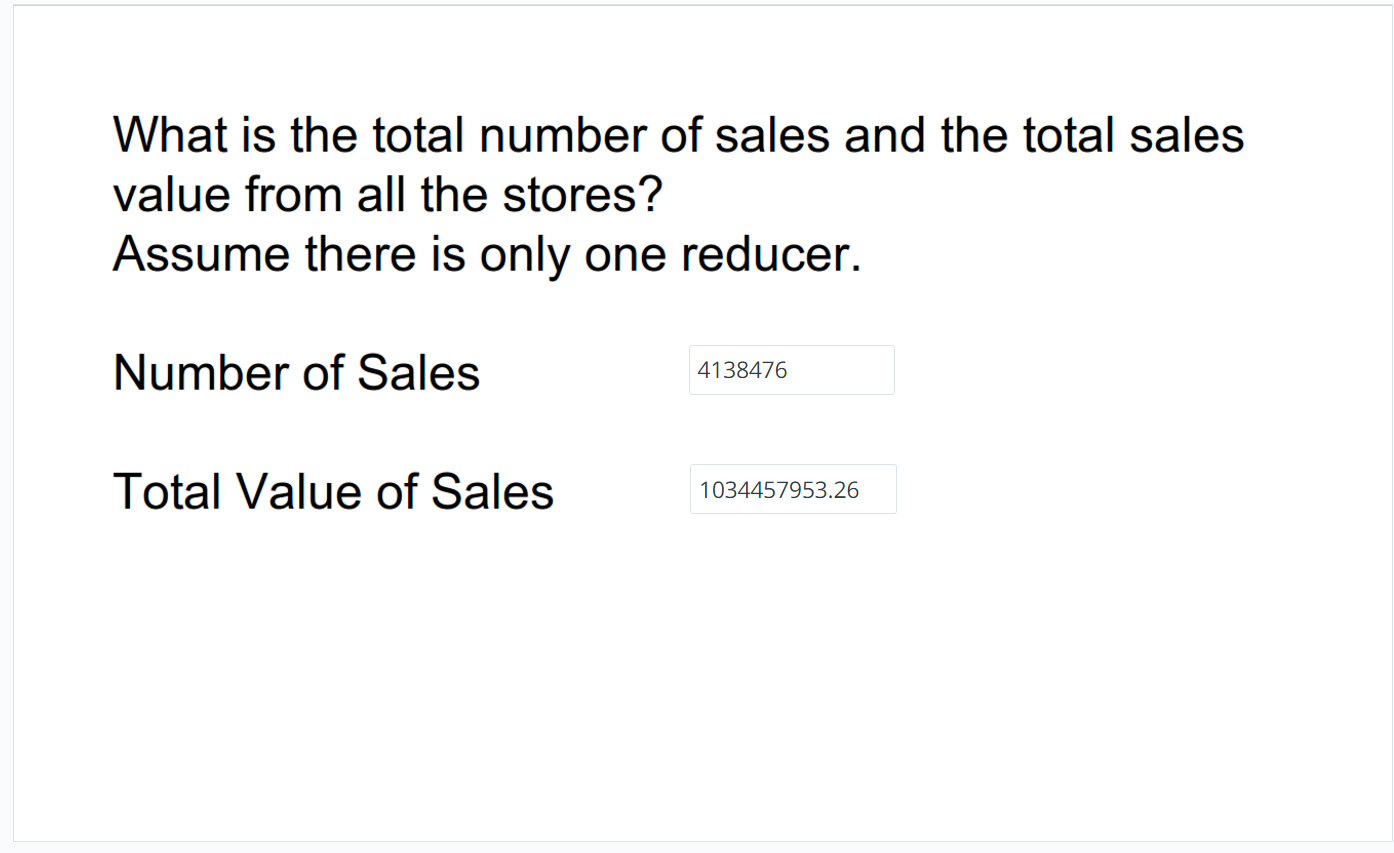












**Part II**

We've talked about how Hadoop works, you've seen Hadoop code, and you've written some code of your own using our store sales data. Now we'd like you to solve some problems using a different dataset on your own. **You'll have to write your Mappers and Reducers from scratch**; please use Python. You will have to do the data processing on your local pseudo-distributed cluster, but you will be able to see if your solution was correct by submitting your results to our system.

The data set we're using is an anonymized Web server log file from a public relations company whose clients were DVD distributors. The log file is in the udacity\_training/data directory, and it's currently compressed using GnuZip. So you'll need to decompress it and then put it in HDFS. If you take a look at the file, you'll see that each line represents a hit to the Web server. It includes the IP address which accessed the site, the date and time of the access, and the name of the page which was visited.

The logfile is in [**Common Log Format**](http://en.wikipedia.org/wiki/Common_Log_Format):

10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/lowpro.js HTTP/1.1" 200 10469

%h %l %u %t \"%r\" %>s %b

Where:

* %h is the IP address of the client
* %l is identity of the client, or "-" if it's unavailable
* %u is username of the client, or "-" if it's unavailable
* %t is the time that the server finished processing the request. The format is [day/month/year:hour:minute:second zone]
* %r is the request line from the client is given (in double quotes). It contains the method, path, query-string, and protocol or the request.
* %>s is the status code that the server sends back to the client. You will see see mostly status codes 200 (OK - The request has succeeded), 304 (Not Modified) and 404 (Not Found). See more information on status codes [**in W3C.org**](http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html)
* %b is the size of the object returned to the client, in bytes. It will be "-" in case of status code 304.

For each of the problems, we would like you to write a MapReduce job to solve the problem and when you have done that you should be able to answer the question we are going to ask you.

See the [**Python Reference**](https://www.udacity.com/wiki/cs101/%3A-python-reference) doc from [**Intro to Computer Science**](https://www.udacity.com/course/cs101) for a review of Python basics and syntax.

