I used Apache Pig since it provides an alternative language interface to programming MapReduce in Java. It also comes with lots of handy User-Defined Functions (UDFs) such as *sessionize*, which I used to sessionize user's logs by time window.

I downloaded the Hortonworks Data Platform (HDP) and used **Grunt shell** to interactively execute my scripts.

First, I read the log file into *PigStorage* and kept only three fields I thought were useful in this task: timestamp, client ip, and request url. It makes sense to consider different ports with the same ip as the same client. However, it is not enough to identify distinct user by using ip addresses; it would be more reliable to include **cookie** to track each session.

For all 4 questions, I used the sessionized logs `(iso_time, unix_time, user_ip, visited_url, session_id)` in my preprocessing step. I then carry on some specific operations such as *GROUP BY*, *AVG* to compute the needed statistics.

The results of my script:

```
The average session times in minutes: (1.7)
The number of unique urls hit per session: (8.3)
The top 5 most engaged users with their longest session length in minutes:
(52.74.219.71,   34.4860)
(119.81.61.166,  34.4808)
(106.186.23.95,  34.4793)
(125.19.44.66,   34.4786)
(125.20.39.66,   34.472)
```

The difficulty in doing this task is that sometimes it is hard to debug for the errors happened way back. One mistake I made was to parse the log file using comma as delimiter, but the original data was actually separated by space. I later encoutered some NULL error and thought I had to deal with null values in my dataset. It took me some time to figure out I did not parse the data correctly. For Question 4, I tried to use some complicated operations, like two nested *FOREACH* operations. It failed. I then learned to break down the task into more straighforward steps and simplified the operations.