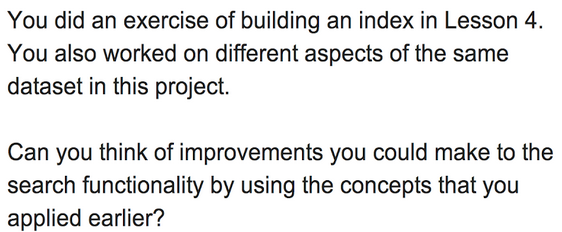


I can use the 3 Vs of Big Data to approach this decision.

I’m told that the data is stored in a RDBMS, so that means

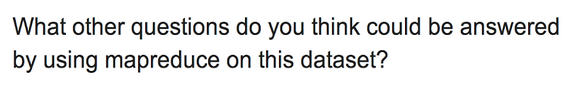
* its current volume fits into a traditional relational DB
* data variety is managable by a traditional relational DB

I will look at the speed (velocity) at which my data grows and will need to be processed. If it will reach terabytes in the short term then I expect my current RDBMS to struggle so I will start migrating to Hadoop.



In lesson 4, we dealt with finding the frequency that a word appears in the forum (for eg, the word “fantastic”) and outputing the nodes where a certain word appears (for eg, the word “fantastically”)

One thing that would be interesting to look at is given one of the top tags, who in forum is the most active for that tag. Udacity may want to hire the most active contributor to say cs101, or perhaps connect the most active contributor to cs253 to reddit.



I’m interested in sentiment so I may look for the most used words in the forum, eyeball the list and keep the top 10 words that are either positive or negative. I may do this for a specific class (filtering by tag).

Also, I wonder if people who help others the most also get more help, so I need to

* identify users who are active in answering and commenting
* find the questions that they ask
* see how many replies they get (or how quickly are their questions resolved)