

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.

I need to know the amount of data I have, in other words, the total size of this database to decide whether this is a big data problem.

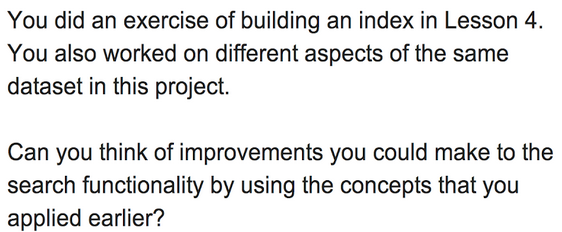
<https://www.udacity.com/course/viewer#!/c-ud617/l-306818608/e-309273651/m-309273654>

at 0:12, Sarah says “A list of purchases at a single store is almost certainly small enough to be easily handled by a traditional relational database system -- or even just a spreadsheet. Orders from hundreds of stores nationwide, though, could start to overwhelm traditional systems. Likewise, information about a single person’s stock portfolio is a small and easily managed chunk of data. But data on trades across the entire NYSE for a year will run into tens or hundreds of terabytes -- and that’s where traditional systems really do start to struggle.”

my interpretation of this is it’s basically saying “if the data fits in a relational data base, then it’s not a big data problem”

Firstly, I think it’s important to think about what I may want to do with such data. I may want to filter for those posts that contain the word “New York”, I may want to find the top 10 most active users, I may want to index my data, etc etc.

And these fit the Mapreduce design patterns covered in lesson 4.



Student times: what times of day are they the most active

Post and Answer: is there a correlation between the length of a post and the length of answers

Top Tags: top tags by the number of questions they appear in

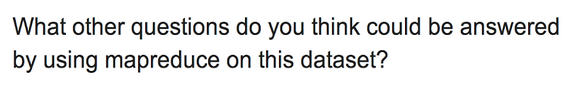
Study Groups: are there people who are already collaborating

In lesson 4:

* found the frequency of a word
* the nodes a certain word appeared in

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).

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)