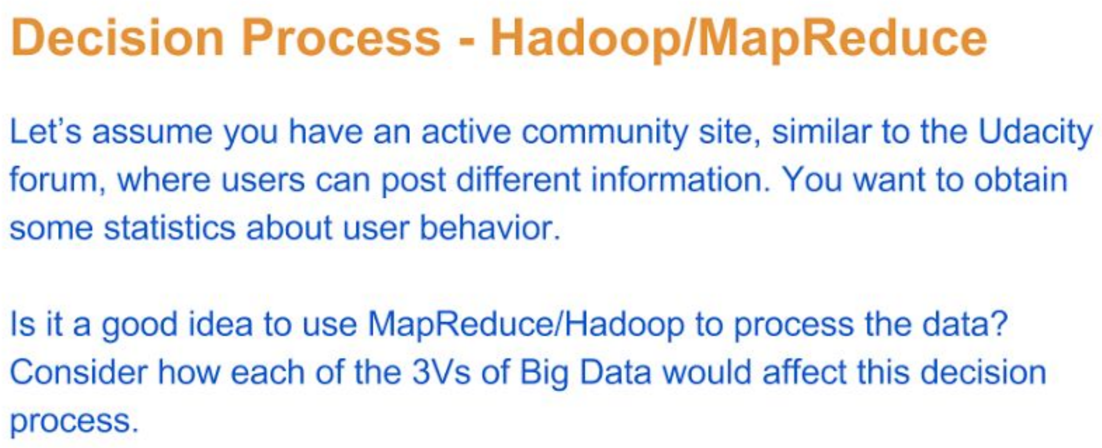
Student: Elderclei Regis Reami
E-mail: elderclei.reami@gmail.com
Course: Intro to Hadoop and MapReduce



Decision Process - Hadoop/MapReduce

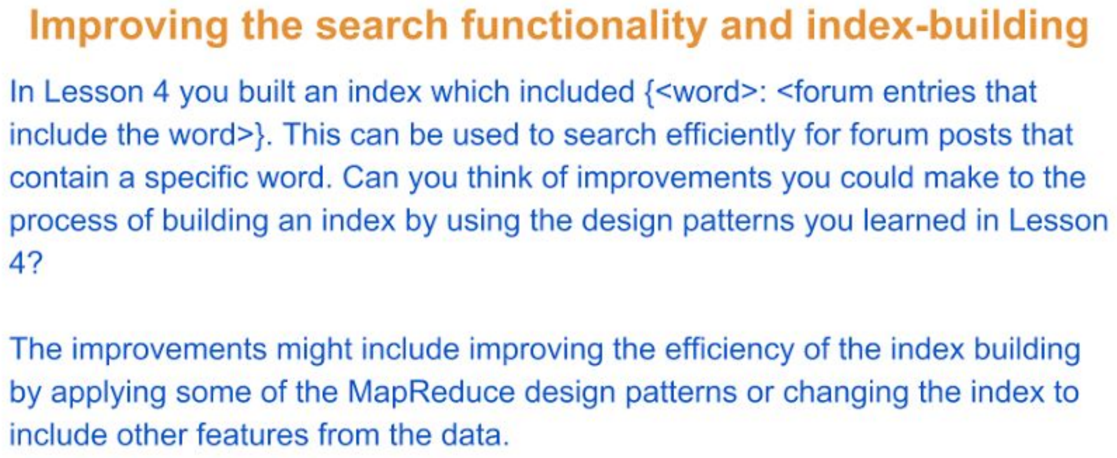
Let’s assume you have an active community site, similar to the Udacity forum, where users can post different information. You want to obtain some statistics about user behavior.

Is it a good idea to use MapReduce/Hadoop to process the data? Consider how each of the 3Vs of Big Data would affect this decision process.

It is useful to remember of three V's of Big Data (Volume, Velocity and Variety). Do you have many users, speaking different languages and using different media to post? Are there many simultaneous discussion threads with lots of posts and replies, which represent complex interactions inside the community? Even if the volume is not that big, Hadoop and MapReduce may be useful to understand clickstream and patterns of social interaction inside the community.

As the community grows, the velocity by which new interactions are created increases exponentially, so having information processed will be costly. Mapreduce will help in distributing and scaling information extraction in this scenario.

As it concerns to variety, the dataset is basically made of texts sent by authors, but new media could be added, for instance, images, audio, and videos, and other external references that could be analyzed by specialized data processors.



Improving the search functionality and index-building

In Lesson 4 you built an index which included {<word>: <forum entries that include the word>}. This can be used to search efficiently for forum posts that contain a specific word. Can you think of improvements you could make to the process of building an index by using the design patterns you learned in Lesson 4?

The improvements might include improving the efficiency of the index building by applying some of the MapReduce design patterns or changing the index to include other features from the data.

Creating an useful and flexible index requires lots of processing. For instance, one could include extra information that can be used to order results by relevance (words that appears in threads with a deeper hierarchy of comments and answers, or with a huge number of user interactions; words that appear together in tuples of 2, 3, 4, 5, and so on). It would also be useful to create a map of related results based on tuples; eliminate stop words from the index; associate index entries for singular, plural, or gender variation of words; generate entries for phonetic equivalents instead of the original words. All of this are examples of filtering, summarization, structure and input/output patterns.

Other questions about the dataset

What other questions do you think could be answered by using MapReduce on this dataset?

1) It's possible to process threads (questions, answers, comments) in order to detect uncommon activity, so as to have them analyzed by humans: is it a flamewar, is it off-topic, is it an extremely interesting topic?
2) Summary and filter interactions by author so as to detect possible moderators ou tutors for future courses
3) It's possible to analyse datetime data so as to find peaks or burst of usage, and create predictive data to be used by infrastructure engineers so as to provided more or less servers in a cloud installation of the forum.
4) Could process words used by an author in every and each of its interactions, so as to detect anomalies or abuse of an account.