Big Data Hadoop—Real World Project—Social Media

Analyze data set from Stack Exchange

As part of a recruiting exercise of the biggest social media company, they asked candidates to analyze data set from Stack Exchange. We will be using similar data set to arrive at certain key insights.

Download the data set from the following link:

http://www.ics.uci.edu/~duboisc/stackoverflow/answers.csv

The data set contains the following attributes:

qid: Unique question id

i: User id of questioner

qs: Score of the question

qt: Time of the question (in epoch time)

tags: a comma-separated list of the tags associated with the question. Examples of tags are ``html'', ``R'', ``mysql'', ``python'', and so on; often between two and six tags are used on each question.

qvc: Number of views of this question (at the time of the datadump)

qac: Number of answers for this question (at the time of the datadump)

aid: Unique answer id

j: User id of answerer

as: Score of the answer

at: Time of the answer (in epoch time)

We need to arrive at following results:

- Top 10 most commonly used tags in this data set.
- Average time to answer questions.
- Number of questions which got answered within 1 hour.
- Tags of questions which got answered within 1 hour.

The total time expected to complete this task is 8 hours.

Hint: Use the techniques taught in lesson 03 to load data sets in HDFS. For using Pig/Hive, recall the techniques mentioned in Lesson 07/08.

This project should be performed using CloudLab.