Open Assignment

Oliver Foster – 2019-03-05

# Idea #1: Count Meetups Through Time

My first idea for analysis involves sending meetup data to a business intelligence software (such as Power BI or Tableau) in order to create a dashboard regarding the current state of arranged meetups for a given country. The steps to accomplish said idea would be as follows:

1. Set up a flume agent to collect the meetup data and store in HDFS. The flume agent should use a regex interceptor to only collect meetup data within the United States.
   * The idea here would be to test the student’s understanding of flume and regex. It should be up to them to create the appropriate regex rule to satisfy this filtering condition
2. Create a hive database and table to analyze the meetup data
3. Download the appropriate ODBC driver to connect the hive meetup table to your favorite BI software (in my case Power BI)
4. Connect the BI software to the hive table to extract the information collected from meetup
5. Group the meetup data by city and hour of day to count the number of meetups in each city every hour
6. Create a visualization that indicates how these meetups change through time
7. Try to detect some visual patterns

The idea behind this project is to showcase how flume can be used in the back-end to extract real value that can be leveraged by a front-end solution to create meaningful visualizations. Additionally, the student should be encouraged to be creative with their visualizations to extract some pattern seeing how the number of meetups in each city changes over the time of day.

# Idea #2: Predict Ideal Meeting Time

My second idea for big data analysis involves using meetup data to create a feature that should be used by Meetup itself: a machine-learning algorithm that can predict the ideal meetup time given the user IDs of each member proposed to be at said meeting. In order to limit how wide the dataset required woulf be for the regressor the student should focus on a single city (say New York). The steps involved should be as follows:

1. Set up a flume agent to collect the meetup data and store in HDFS
2. Create a Hive database and table to analyze the collected data
3. Use Sqoop to migrate the data from hdfs to a mysql database (or any database of choice)
4. Using python, connect to the mysql database and extract for each meeting the user IDs of each attendee and the time of said meeting for all users in New York
   * This might require some data cleaning in pandas
5. Binarize the data by giving a binary column for each user ID for each meeting event
6. Translate the time-of-meeting field into an ordinal datatype suited for a regression problem
7. Using sklearn train the repressor of their choice to predict the meeting time based on the binarized user IDs of the people present

The output of this activity is that now Meetup has a “time proposer”. This feature could be exposed to propose optimal meetup times to meeting organizers based on who they have invited to the meeting. While this regressor has just been programmed for users in New York, this approach can be applied to all cities.