Revision 3/3/2022

COMS 363 Spring 2022

Assignment III Part 2

Percentage in your final grade: 5%
Objectives:

1. Practice ETL (Extract-Transfer-Load) to bring data into a relational DBMS. This process is often very time-consuming and involves eliminating inconsistency in the data.
2. Practice writing SQL DML and stored procedure for a larger database.

Submission requirements

The file <netid>Q4.sql where <netid> is replaced with your university netid.

Include the comment indicating your authorship of the work, for example,

--Author: Wallapak Tavanapong

4. (39) Working with a large database.

Execute ProjectDDL.sql to create a 'project' database. Execute ProjectInsert.sql that uses bulk loading to load the database with tweets of state legislators and presidential candidates for the 2020 presidential election. See the help file for the explanation about the ProjectInsert.sql and the lecture on this. If the import is successful, there are 81906 rows of tweets, 13517 users, 61957 rows of mentions, 40260 rows of tweet associations with URLs, and 52240 rows of tweets associations with hashtags.

A Twitter user account has the following properties: name, screen name, the number of followers, the number of people this user follows, subcategory, category, and the state the user lives. The screen name cannot be null and is unique among all users. The subcategory indicates the political party to which the user belongs. The values of this attribute are "GOP," "Democrat," "na," or null. The values of the category attribute are "senate_group", "presidential_candidate", "reporter", "Senator", "General", or null. Presidential candidate accounts are not associated with any state. We use "na" as the value of the state attribute for the user account without an associated state.

Deliverable: <netid>Q4.sql that has all the stored procedures for all the questions.

Suggestion: Write an SQL query first. As the database is relatively large, your query may not return if you do not put the right join condition. Test that the query works first. Then, put the query inside a stored procedure. Test each stored procedure individually before putting all of them in the file <netid>Q4.sql. It is easier to debug everything this way.

a. (19 points) Write a stored procedure called **influentialUsers** that accepts two parameters: k users of type integer and the name of the political party of type varchar(80). This procedure shows k users of the specified party who have the most followers. Show the user's screen name, the user's party, and the number of followers in descending order. Recall that the subcategory attribute indicates the political party to which a user belongs.

Output of call call influentialUsers(5, 'GOP');

screen_name	subcategory	numFollowers
realDonaldTrump	GOP	8062804
marcorubio	GOP	1356595
RealBenCarson	GOP	1324661
tedcruz	GOP	1097745
JebBush	GOP	599221

b. (13 points) Write a stored procedure called **influentialTweet** that accepts three parameters: k of type integer, the month of type integer, and the year of type integer. This procedure shows k rows of tweet text, the retweet count of that tweet, the name, and the category of the user who posts the tweet in descending order of their retweet count. Only consider tweets posted in the given month of the given year. The example below shows the requested information for the tweets posted in February 2016.

Output of call infleuentialTweet(5,2, 2016)

texts	retweetCt	user_name	category
RT @ow: please enjoy this fox confu...	75555	Jessiehellmann	reporter
America s first black president cannot...	44461	BernieSanders	presidential_candidate
RT @BernieSanders: America s first bl...	44461	HillaryClinton	presidential_candidate
RT @BernieSanders: America s first bl...	44456	RepMarkCardenas	House_representative
RT @JebBush: America. https://t.co/...	27460	jasonmdstein	reporter

c. (13 points) Write a stored procedure called **mostMentioned** that accepts four parameters: k of type integer, the name of the party of type varchar(80, the month of type integer, and the year of type integer. This procedure returns the tweet handles of *k* users who were mentioned the most in tweets of users of the given party. Consider only the tweets posted in the given month and given year. Show the user's screen name, user's state, and the list of the screen name of the user(s) who mentioned this user in descending order of the number of tweets mentioning this user.

Output of call mostMentioned (5, 'Democrat', 1, 2016)

Hint: The function group_concat() is useful for creating a list of values.

mentionedUser	postingUsers
BernieSanders	BernieSanders,NebraskaDems,nvdem...
POTUS	BernieSanders,CTSenateDems,FlaDe...
onetoughnerd	MISenDems
TheDemocrats	BernieSanders,Deldems,FlaDems,NCD...
HillaryClinton	BernieSanders,HillaryClinton,Nebraska...