L5: Data Streaming Management Systems

Assignment

**Introduction:**

The goal of this assignment is to illustrate the different concepts and ideas about processing of continuous queries, e.g.: techniques such as sliding windows or sampling.

We will provide instructions to connect to the Twitter API and continuously read a stream of tweets associated to a certain account.

The students will be required to implement a script that performs continuous update of some Continuous Queries fixed by us. We will define the queries in natural language, so the students will also have to translate them into some CQ language (although they are not required to parse the language, their program will just implement continuous update of their interpretation of the queries).

**Preliminary Readings:**

1. Models and Issues in Data Stream Systems (Babcock, Babu, Datar, Motwani, Widom)

(Section 6 is not necessary).

**Research Questions**

1. Think about the following three situations:
   1. Analyze the data of recent power usage statistics reported to a power station and adjust the power generate rate if is necessary.
   2. A school need to know data about their students, and query about courses, age, and instructors

Which system would you choose for each situation? If you choose DSMS, explain in detail how you implement it.

1. Explain three of the most typical strategies for CQ processing, and for each of them, think of scenarios where it would be more convenient to apply it.
2. Mention the main characteristics of a Query Scheduler in a Data Stream Management System. What do you think is the biggest problem of the scheduling in comparison with Data Base Management System?

Lab assignment:

Continuous Querying over Twitter Stream

**Readings:**

* Realtime Analytics on Streaming Twitter Data
  + <https://www.wakari.io/sharing/bundle/wakari_demo/realtime_twitter_analysis>

**Requirements**

* 1. Python 2.7
  2. **Libraries**
  3. Auth and consumer key of a twitter acount in <https://dev.twitter.com/>

**Development**

1. Write queries in CQL (the objetive is understand the queries, but the script isn’t in SQL)
2. Implement query processing script in python over Tweeter Data Stream. (Using the model in “**Realtime Analytics on Streaming Twitter Data”**)
3. Give the answer of the following **queries**.
4. Think four queries over de Twitter Data Stream. Do it assuming diferents flow rate, and time computing time.
5. Conclusions

**Queries:**

* percentage of tweets mentioning “Yes” during 2 mins, from all tweets, without restrictions.-32
* (sliding windows) Use sliding windows, and calculate the percentage of twetts that mentioning “. Use a windows size of 300 tweets.-8
* (batching) percentage of tweets mentioning “No” during 2 mins. Assumes that the counting function is very slow rate, use buffers (length 100) for the elements and compute the query answer using each tweet once.-1960
* (sampling) precentage of tweets mentioning “Hi” during 2 mins. Assumes that the update function is slow. Update the list with a sample of the elements. (e.g: one of one hundred).-2009

**Useful Resources:**

1. <https://dev.twitter.com/>
   1. h[ttps://dev.twitter.com/docs/api/1.1/get/search/tweets](https://dev.twitter.com/docs/api/1.1/get/search/tweets)
   2. [https://dev.twitter.com/docs/entities#tweets](https://dev.twitter.com/docs/entities" \l "tweets)
2. https://code.google.com/p/python-twitter/

**Future Readings:**

1. Creating a Python Script for Twitter Search
   1. <http://coding2day.com/TwitterPython.pdf>