CS 122A: Introduction to Data Management – Fall 2020

Homework 8: NoSQL (100 points)

Due Date: **Sunday, Dec 13 (6:00 PM)**

Submission

All HW assignments must be submitted online as a PDF through the associated dropbox on Gradescope. See the table below for the HW submission opportunities. Note that after 6:00 PM on Monday, December 13 no further HW submissions will be accepted. (We will be releasing the solution at that time so you can study for the Endterm exam.) Please strive to get your work in on time!

Date / Time	**Grade Implications**
Sunday, Dec 13 (6:00 PM)	Full credit will be available
Monday, Dec 14 (6:00 PM)	10 points will be deducted

NoSQL (SQL) [100 pts]

Congratulations! If you are reading this, you **must have successfully finished reading and running the exercises in the SQL++ Primer**. ***If you haven't, stop now and go do that first!*** Business is booming, and CheckedTweets.org needs to scale beyond MySQL's capabilities. In this assignment you will explore some of the virtues of "NoSQL" systems for applications like this. Enter AsterixDB...! As you read through the assignment, answer its questions, and write queries, you will want to refer to two other sources of information -- namely, the solution to HW #1 (the E-R schema) and the corresponding translated relational schema that you have used for all the SQL-based HW assignments. You may also find that the queries that you're asked to write here seem hauntingly familiar (*deja vu!*), so you can refer to the relevant earlier assignment(s) if you are wondering if your answers here are on track. (We will also specify the number of expected answers for each query, as usual, and show a sample output record to help you with that.)

Step 0. Load the Dataset

To save you time, we have transformed the relational CheckedTweets database schema into ADM (the Asterix Data Model). You need to download our scripts [HERE] (you must use your UCI account to download this) and populate the dataset. Unzip the 'hw8.zip' file; there should be 3 folders and 4 SQL++ scripts inside the folder like the following picture.

After you've started your sample AsterixDB cluster, the next thing you should do is run the initial test script. This will create the dataverse, datatypes, and datasets for this NoSQL translation of the schema, as well as load the test data. To do so, first open the `checkedtweets_test.sqlpp` file. **Replace the text "PATH_TO_HW8_FOLDER_GOES_HERE" to point to the above *hw8* folder.**

Windows Users: There is **NO** need to change '/'. The path you should use will look like this:
`("path"="///C:/Users/XXX/Downloads/hw8/data/test/rawtweet.json")`

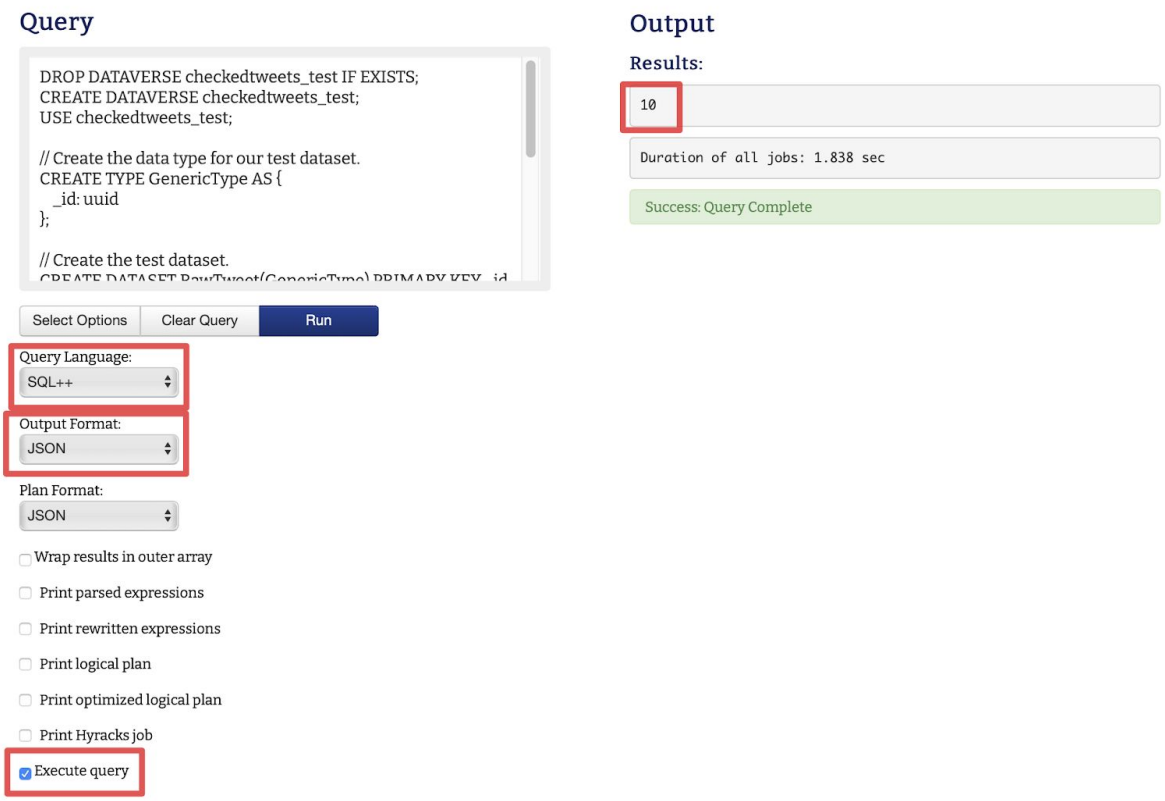
OSX, Linux Users: Change the **PATH_TO_HW8_FOLDER_GOES_HERE** to match with the location where you placed the hw8 folder. (E.g., `"path"="///Users/XXX/Downloads/hw8/data/test/rawtweet.json"`)

All users: Beware of errors in these paths -- particularly having the wrong number of leading slashes! Your load commands will fail if you get that wrong.

After you change the path, copy and paste the DDL you just edited into the Query box of the AsterixDB web-based query interface. Make the following adjustments to the settings right below the query box:

1. Make sure that the `Query Language` selected there is "SQL++" (the default setting). For the curious, AQL is the now-deprecated original query language for AsterixDB.
2. Make sure that `Execute query` is checked at the very bottom (the default setting) and that the other options are unchecked.
3. Make sure that `Output Format` is `JSON` (the default setting).

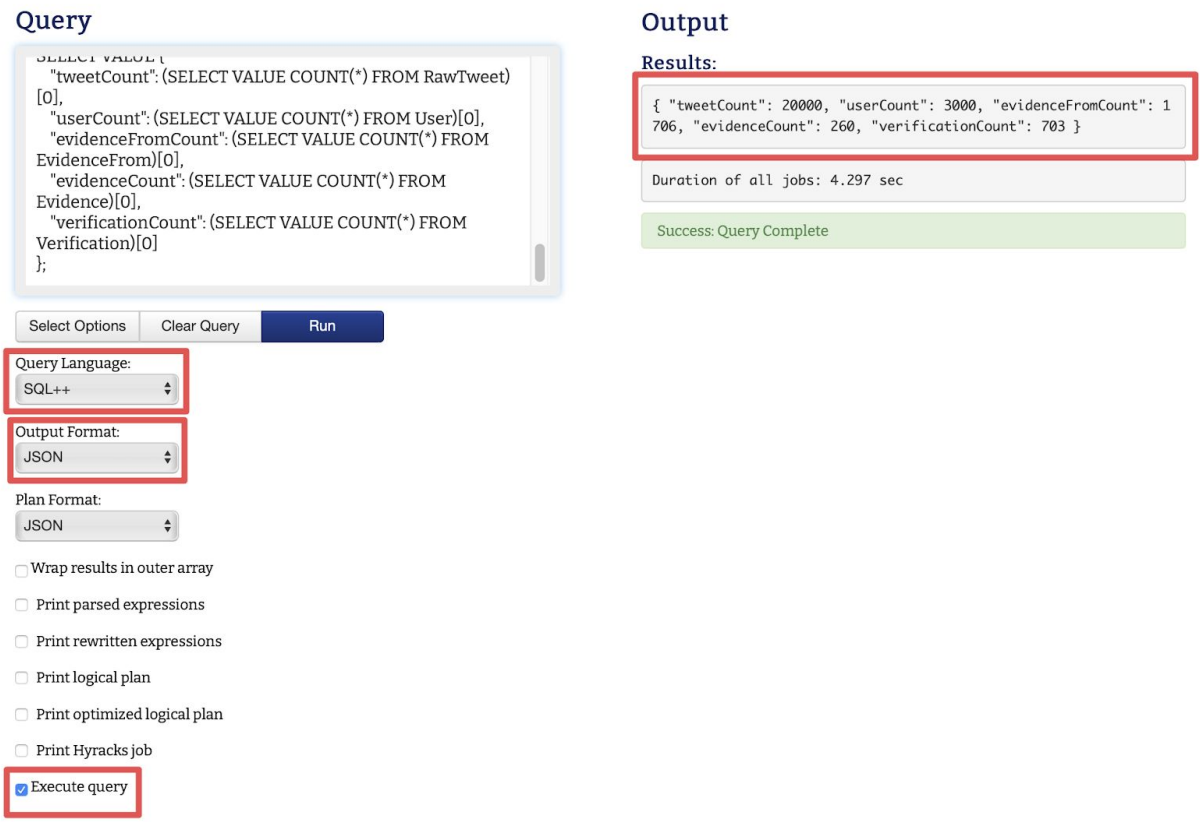
Finally, click 'Run'. After execution, you should see something that resembles the following result (with the number 10 under `Results`):



If the Output doesn't show "Success: Query Complete" and the query results listed here, please post your issue on Piazza. Again, beware of bad LOAD paths (too many or too few slashes). Please check for that error before reaching out for help on Piazza. ***NOTE: Under no circumstances should you try to load the full datasets until you get loading working perfectly here first!***

Once this test has executed successfully, great! You are now ready to populate the ***actual*** datasets. Open `checkedtweets_schema.sqlpp` and change the path's value to your downloaded 'hw8' folder just like what you did with the test dataset. Make sure this is the absolute path and ***very carefully*** change **ALL** 5 occurrences of the path variable like you did above. Each file will populate a dataset that we have previously created. (*Hint:* If you get the path variable replacement wrong, you will probably get an error message that says something like: `ASX3077: /wrong_path/hw8/n1nf/user.json: path not found [HyracksDataException]`.)

After changing all 5 of the path values, go ahead and copy and paste the **ENTIRE** content of `checkedtweets_schema.sqlpp` file into the query interface. As before, execute it by clicking `Run`. You should see something similar to the following result:



Again, if the Output doesn't show "Success: Query Complete", please post your issue on Piazza after first trying to resolve the problem yourself. Don't wait until you finish your script! (And if you later happen to see other students struggling on Piazza at this step -- students making the same mistake you did -- please chime in and give them a hand! Loading the data successfully isn't a part of the graded exercise, so helping one another out with the ***loading*** specifics will be lauded rather than disapproved of as being inappropriate collaboration. :-))

Compare the counts in the output above with your own output to make sure that you have the complete set of datasets. All datasets for the "schema`ed" dataverse reside in the `checkedtweets_schema` dataverse, so be sure to put the statement `USE checkedtweets_schema;` in front of each one of your queries when you run them. Here is an example of how that looks:

```
USE checkedtweets_schema;
SELECT *
FROM User U
WHERE U.user_id = 0;
```

Note: For this assignment, you must turn in a PDF that you have created by copying/pasting from the

query interface into a copy of the HW8 template and then PDF-printing the results. When you copy and paste your query and result, do ***not*** take a screenshot. Instead, use the text copy and paste feature.

NOTE: As you work on this HW, you should check the DDL of the schema-fied versions of the datasets to determine their attribute names and structures (for working with both versions).