CPRE/SE 419 Software Tools for Large Scale Data Analytics

Spring 2023

Take-home Final Exam

Due: Wednesday, May 10, at 11:59PM

Preamble:

Your final exam has two categories of questions. Specifically:

- The first category consists of two sub-categories: (A) quick-answer-problems (5 questions); and (B) "definition-style" problems (4 questions) in which you need to concisely demonstrate an understanding (and capability of using) certain concepts/abstractions; call them – "first-round interview" questions.
- The second category (4 questions) consists of actual problems for which you need to demonstrate an understanding of a particular methodology/paradigm and apply it to specific problem settings.

The points total to 107 (7 extra credits, randomly scattered throughout the problems).

You can work in teams of two students (or, if prefer to work solo – that is fine too), and *please do not forget to put the names of all the team-members with your answer.*

Please make sure that your solutions are typed, and upload the file in the Canvas for the corresponding assignment (and do not forget to put the names of the team-members).

Good luck!

Part I.A – Provide brief/concise answer (with justification):

1. (5 pts.) What are the long running daemons in YARN?
2. (5 pts.) In the context of CAP theorem for distributed data systems, give an example of a database that belongs to the "CP" part of the spectrum.
3. (5 pts.) Explain briefly the difference between NoSQL and NewSQL.
4. (5 pts.) What are the main features of EDA (Event Driven Architectures)
5. (5 pts.) Explain briefly the *turnstile* models of data streams.

Part I.B – Explain the following (please try not to be overly-verbose (i.e., sufficiently detailed, but concise discussions)):

1. (9 pts.) Explain the concept of a *partitioner* in MapReduce (i.e., what is its use), and give an example of types of Partitioners readily available for MR jobs in HDFS.

2. (9 pts.) In the context of distributed databases, explain the Bully Algorithm for coordinator selection.

3. (9 pts.) Define the concept of RDD in Spark. What are its main benefits?

4. (9 pts.) What is the difference between tumbling window and sliding window?

Part III – Algorithmic questions

1. (14 pts.) Consider the following input stream:
   A, B, C, E, A, A, A, D, F, E, F, F, F, B, B, C, C, D, F, F, F,
   What is the outcome of executing Misra-Gries algorithms for selecting the (approximate) heavy hitters in the stream, when the capacity of the map is limited to c = 3.

   Justify/explain your answer.

2. (13 pts.) Consider a scenario in which a distinct sampling is required, for a sample of size k=3. Assume that following are the values for the hash function h(A) = 0.7; h(B) = 0.8; h(C) = 0.4; h(D) = 0.9; h(E) = 0.3. Show the content of the sample for each arrival of an element in the following stream: A, D, D, A, B, A, A, A, C, C, E, A, B, A, D.

3. (14 pts.) Consider the composite event C ≡ A; B (i.e., C is detected whenever an occurrence of the primitive event "A" is followed by an occurrence of the primitive event "B"). Consider the following stream of events – where the first parameter is the primitive event, and the 2nd parameter is the time-instant of its detection (for example: e(B,4) denotes that the primitive event B was detected at time-stamp 4):
   e(A,1), e(A, 2), e(A,3), e(B,4), e(A,5), e(B,6)

   How many detections of "C" will occur under the *recent* context and how many under *chronicle* context for selecting the participating primitive events (illustrate/justify your answer)?

4. (5 pts.) For this problem, you will need to have each member of a team work on it separately, and you will need to provide the individual answers! The purpose is to combine your skills in file-searching/scanning and user's preferences matching. Specifically, you are to execute the following request:
   a. Go through the "Contacts" in your mobile phone;
   b. Select your favorite person for executing the activity in Step III below;
   c. Upon completing all the other problems from this assignment, call and tell your favorite person[1] from Step II above in a loud and clear manner: "I am done with 419!!!"
   d. Write the answer that you received.

[1] It is acceptable that for this problem you use a person from the same household, for as long as the notification (after scanning and matching) is executed via phone.