Electrical and Computer Engineering / Software Engineering

CPRE/SE 419 Software Tools for Large Scale Data Analytics

Spring 2022

Take-home Final Exam

Due: Wednesday, May 11, at “high-noon” (12:01PM)

Yuichi Hamamoto

Preamble:

Your final exam has two categories of questions. Each part/category is on a separate page – more specifically:

- The first category consists of two sub-categories: (A) quick-answer-problems (6 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 (3 questions) consists of actual problems for which you need to demonstrate an understanding of a particular methodology/paradigm/algorithm and apply it to specific problem settings.

The points total to 106 (6 extra credits, randomly scattered).
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 both team-members with your answer.*
Please make sure that your solutions are typed, and upload the file in the Canvas for the corresponding assignment.

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

1. (5 pts.) Explain briefly the concept of *shuffling* in MR.

   Scuffling is that mapper maps the input to the reducer as their intermediate output.

2. (5 pts.) Recall the CAP theorem for distributed data systems. Give an example of a database that belongs to the "CP" part of the spectrum.

   The CAP, or Consistency, Availability, and Partition Tolerance, theorem claims that any two of them can be satisfied but not all of them at the same time. An example of a database that belong to the CP part of the spectrum is MongoDB, which is a NoSQL.

3. (6 pts.) What is the benefit of providing the type when defining the schema (i.e., when loading the data) in Pig Latin.

   The benefit of providing the type helps pig process data loading quickly. Pig does not require the data types but then it need to determine the data type by itself and will take extra time.

4. (6 pts.) Explain briefly the difference between NoSQL and NewSQL.

   NoSQL is not relational, but NewSQL is relational as it has SQL as the primary mechanism for the application interaction. NewSQL has ACID support while NoSQL rather provide CAP.

5. (5 pts.) Explain the difference between an *online* algorithm and *streaming* algorithm.

   Streaming algorithms make an action when a group of points arrives where online process each point as soon as it arrives.

6. (5 pts.) Explain briefly the difference between *cash-register* and *turnstile* models of data streams.

   The difference between them is that register models always have partial range value (c[x]) as positive where turnstile models can have a negative value. Therefore, how we reconstruct the signal is different. For turnstile one, we add or subtract the contribution of the streaming item where for register one, we add.

Part I.B – Explain the following (please try not to be overly-verbose – i.e., provide sufficient details to justify your answer, but be concise discussions):

1. (10 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.

   The concept of a partitioner in MR is mapping the key to a specific reduce task, so that it can possibly split task evenly and improve the efficiency.

2. (10 pts.) Consider the scenario of distributed concurrency management under Majority Protocol. How many locks are needed for an operation to be allowed to execute? What is the main drawback of this protocol?

   n/2 – 1 locks are needed. The drawback is that it can potentially cause deadlock because each operation needs this many locks.

3. (9 pts.) What are the main benefits of RDDs in Spark?

   Because RDD is read only and will be distributed, it can process the data parallelly and safely. So that, it can handle huge data quickly.

4. (9 pts.) Consider a scenario in which the EPS has defined composite events that need to be detected by combining multiple streams of (primitive) events. Describe briefly what is the meaning of a *consumption context for the primitive events* upon detection of a composite event. Describe in detail the *chronicle context*.

   The consumption context is the primitive events to consume while it is evaluating a composite event. Based on the oldest components, chronicle context determines the parameters of compositive events. After that, it deletes the consumed components

Part III – Algorithmic questions

1. (14 pts.) Consider two files:
    I. temperature.txt, in which the lines have structure (temp_reading, city, time)
    II. geo-data.txt, in which the lines are of the form (city, country, city_population)

Write a Pig Latin code which will report the average temperature per country, but only for the countries in which that average temperature is higher than 65K.

Assumption: the output will be stored in ‘/ouput’

```
--load files
temp = LOAD ‘/temperature.txt’ USING PigStorage(‘ ‘) AS (temp: double, reading: chararray, city: chararray, time: chararray);
geo = LOAD ‘/geo-data.txt’ USING PigStorage(‘ ‘) AS (city: chararray, country: chararray, city_population: long);

--intersect
joined = JOIN temp BY city, geo BY city

--keep attributes
countries= FOREACH joined GENERATE country, temp;

--group and count
groups = GROUP countries BY country;
ave = FOREACH groups GENERATE group, AVE(temp) AS average;

output = FILTER ave BY (average > 65);

STORE ouput INTO ‘/ouput’ USING PigStorage(‘\t’);
```

2. (14 pts.) Consider a scenario in which a distinct sampling is required, for a sample-limit 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.

arriving element	sample 1	sample 2	sample 3
A	A(0.7)		
D	A(0.7)	D(0.9)	
D	A(0.7)	D(0.9)	
A	A(0.7)	D(0.9)	
B	A(0.7)	D(0.9)	B(0.8)
A	A(0.7)	D(0.9)	B(0.8)
A	A(0.7)	D(0.9)	B(0.8)
A	A(0.7)	D(0.9)	B(0.8)
C	A(0.7)	C(0.4)	B(0.8)
C	A(0.7)	C(0.4)	B(0.8)
E	A(0.7)	C(0.4)	E(0.3)
A	A(0.7)	C(0.4)	E(0.3)
B	B(0.8)	C(0.4)	E(0.3)
A	A(0.7)	C(0.4)	E(0.3)
D	D(0.9)	C(0.4)	E(0.3)

3. (8 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 for each!* The purpose is to combine your skills in file-searching/scanning and user's preferences matching. Specifically, you are to execute the following request:

   I. Go through the "Contacts" in your mobile phone;

   II. Select the person with whom you have had most calls exchanged (regardless of who was the caller) in the past two weeks.

   III. 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!!!"

   IV. Report the answer.

I called my father and told me "good job, good luck on other exam."

[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.