CPRE/SE 419 Software Tools for Large Scale Data Analytics

Spring 2022

Homework 2

Due: Tuesday, April 5, 11:59PM

Yuichi Hamamoto

Preamble:
The purpose of this homework is for you to practice/recap the main concepts from Spark; distributed databases, and SQL vs. NoSQL paradigms.

Problem Set

1. (13pts.) Describe briefly the (difference between) *transformations* and *actions* in Spark.

   Transformations reduces new RDD(s) from the existing RDD(s). Action return a result such as count().

2. (13pts.) Describe briefly the concept of RDD in Spark.

   RDD is the fundamental data structure in Spark which is fault-tolerant record of data distributed to multiple nodes.

3. (15pts.) Describe briefly the concept of *false cycle* in distributed transaction management

   False cycle is a cycle has a problem in the order of process. For example, T_1 <- T_2 <- T_3 and it was distributed as S_1: T_1 <- T_2 and S_2: T_2 <- T_3. In this scenario, S_1 should wait for S_2 but because they are in the different nodes, they would not notice it unless from the coordinator.

4. (15pts.) What are the benefits and drawbacks of the *Majority Protocol* for distributed lock management?

   The benefit of Majority Protocol is that is can be used even when some one sites are not available. The drawbacks are that it needs 2(n/2+1) message for handling the lock/unlock requests and the potential deadlock.

5. (14pts.) Describe briefly the difference between NoSQL and NewSQL type of databases.

   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.

6. (14pts.) Describe briefly the *CAP theorem.* Provide an example of a NoSQL database from the "AP" part of the spectrum.

   CAP theorem claims that a distributed system can satisfy any of two CAP (Consistency, Availability and Partition Tolerance but not all three. The example would be web caching which has the functionality of conflict resolution.

7. (19pts.) Recall the Bully-Algorithm for selecting a coordinator in distributed replicated DB. How does it behave *when a failed site subsequently recovers*?

   The failed site also starts an election which can ensure that the highest ID is always elected.

What to turnin: Typed solutions are strongly preferred. If, for whatever reason, you are prevented from using any editor, then we may accept hand-written solution – provided that they are legible.
(yes, the points do add up to 103%, so 3% extra credit...).

This is assignment can be done in teams of two students (of course, we will honor individual submissions). While one submission per team is enough – please make sure to indicate the names of both team members (or, to explicitly state that this was an individual assignment) in the preamble of the document with your solutions.