Electrical and Computer Engineering / Software Engineering

CPRE/SE 419 Software Tools for Large Scale Data Analytics

Spring 2022

Homework 3

Due: Monday, April 18, 11:59PM

Preamble:
The purpose of this homework is for you to practice streaming data processing.

Problem Set

1. (25 pts.) Would you say that MapReduce is a good paradigm for streaming data processing?

2. (30 pts.) Describe briefly the difference between a *tumbling* window and a *sliding* window.

3. (20 pts.) Describe briefly the notion of *distinct sampling* from a stream.

4. (45 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.

NOTE: you can work in teams of up to 2 students for this assignment. One submission per team is sufficient – however, make sure that the preamble lists the names of both team-members.