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?

   I would say MapReduce is a good paradigm for streaming data processing. For instance, Hadoop DataStream is existing utility allows to create/run map reduce and it uses Unix stream as the interface. It is designed for streaming data processing and works well.

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

   Tumbling repeat non-overlapping interval whereas sliding overlaps. For example, when the window size 3 and input series is [1,2,3,4,5,6], tumbling moves 123->456 where sliding moves 123->234-> etc.

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

   Distinct sampling is getting sample from the set of distinct identifiers within the 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.

   n = 21

Stream = A, B, C,

1	1	1
A	B	C

Stream = A, B, C, E,

Stream = A, B, C, E, A, A, A, D, F,

3	1	1
A	D	F

Stream = A, B, C, E, A, A, A, D, F, E,

2		
A		

Stream = A, B, C, E, A, A, A, D, F, E, F, F, F, B, B,

2	3	2
A	F	B

Stream = A, B, C, E, A, A, A, D, F, E, F, F, F, B, B, C, C,

1		
F		

Stream = A, B, C, E, A, A, A, D, F, E, F, F, F, B, B, C, C, D, F, F,F,

4	1	
F	D	

The number of decrements to any counter is no more than n/21

⇨ There's no outcome

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.