Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2021-2022

Comprehensive Test
(EC-3 Regular)

Course No.	: DSECL ZC556
Course Title	: Stream Processing and Analytics
Nature of Exam	: Open Book
Weightage	: 40%
Duration	: 2 Hours
Date of Exam	: ~~06/03/2021 or 19/03/2021 (FN/AN)~~

No. of Pages	= 4
No. of Questions	= 4

Note to Students:

1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1. When you decide to implement your own Bloom filter, you need to understand the main formulas relating important parameters impacting the design of Bloom filter, so that you can optimally configure the Bloom filter. Consider the following notation for the four parameters of the Bloom filter:

- f = the false positive rate
- m = number of bits in a Bloom filter
- n = number of elements to insert
- k = number of hash functions

The formula that determines the false positive rate as a function of other three parameters is as follows (*Formula 1*):

$$f \approx \left(1 - e^{-\frac{nk}{m}}\right)^{k}$$

a) For each of the following pair of bits-per-element value and number of hash functions, compute the value of “f”. **Show all the necessary calculations. [10]**

- Bits-per-element: 5,6,8,10
- Number of hash functions: 1 to 10

k m/n	1	2	3	4	5	6	7	8	9	10
5	F?	F?	F?	F?	F?	F?	F?	F?	F?	F?
6	F?	F?	F?	F?	F?	F?	F?	F?	F?	F?
8	F?	F?	F?	F?	F?	F?	F?	F?	F?	F?
10	F?	F?	F?	F?	F?	F?	F?	F?	F?	F?

b) Plot the graph of “f” against Bits-per-element and number of hash functions. **[2]**

c) What is impact of change in Bits-per-element on the false positive rate? **[1]**

d) Is there any relevant relationship that exhibit between the number of hash functions and false positive rate? **[1]**

e) If the optimal *k* for a particular bits-per-element is given by following formula, then for Bits-per-element value of 7, what is optimal number of hash functions are required? **[1]**

$$k_{opt} = \frac{m}{n}\ln 2$$

Q2. The weighted moving average (WMA) is generalization of the standard moving average that uses different weights for each of the elements in the window. This collection of weights is known as “kernels”. Consider the following implementation of WAM algorithm.

```
Public class WMA {
        Double [] kernel;
        Double [] values;
        Double kernelSum = 0;
        Int k = 0;
        Long N = 0;

        Public WMA (double [] kernel) {
                this.kernel = kernel;
                for ( double j : kernel) kernelSum += j;
                values = new double[kernel.length];
        }
        Public double observe (double x) {
                Values [k++] = x;
                If ( k == values.length ) k = 0;
                N++;
                If ( N < kernel.length) return Double.NaN;
                Double y = 0;
                For ( int i = 0; i < kernel.length; i++)
                        Y += kernel[i] * values [ (k+i) % values.length];
                return y/ kernelSum;
        }
}
```

Assume the kernel weights are given as 1, 2, 3, and 1.

a) Compute the WMA for each of the following “x” values. Show all the necessary calculations. **[7]**
   x = 1, 2, 3, 4, 5, 6, 7, 8
b) Discuss the impact of this algorithm in off-line and online processing environments**. [1]**

Q3. Consider the following stream of events coming from a truck. Periodically these events are received and processed on the server side for doing some rum time analytics as shown in the query below.

Event Stream data

ID	Event	Processing Time	Status	Qty
	Time			
T1	11.2	12	Moving	2
T2	11.15	12	Moving	3
T3	11.09	12	Moving	1
T1	11.5	12	Moving	2
T2	11.45	12	Static	3
T3	11.39	12	Broken	1
T4	11.19	12	Moving	2
T1	12.2	1	Moving	2
T2	12.15	1	Static	3
T3	12.09	1	Broken	1
T4	11.49	1	Moving	2
T1	12.5	1	Broken	2
T2	12.45	1	Moving	3
T3	12.39	1	Static	1
T4	12.37	1	Moving	2

The structured streaming Query –
inputDataFrame.groupBy(Status).window(120 minutes).count(Qty)

Showcase the content of following tables when the query is executed at 12.00 and 1.00 PM respectively. **[7]**

a) Input table
b) Result table
c) Output when mode is respectively
    i. Complete
    ii. Update
    iii. Append

Q.4. Look at following stream of data values with time stamp. **[1 + 2 + 1 + 1 + 3 + 2 = 10]**

Value	34	67	-6	78	34	12	90	45	12
time	0	0	1	1	2	2	2	3	3

(For a, b, c) If a count based tumbling window is defined with eviction policy set to 4,

a) How many windows will be processed for above stream of data values?
b) What will be the difference between the first and last data value in the second window?
c) If the trigger policy is set to 2, then how many times the code will be executed for query?

(For d, e) If a time based tumbling window is defined,

d) With eviction policy set to 3 seconds, how many windows will be processed for the above stream of data values?
e) If the trigger policy and eviction policy both are set to 1 second, what will be the average values (for each window)?
f) If a sliding window is defined, with slide interval 1 second and window length 1 seconds, what are the different windows that will be visible for the given streaming data values?

****