**Stream Processing with Spark Streaming**

1.What are the parameters defined to specify window operation

a. Window length, sliding interval

b. State size, window length

c. State size, sliding interval

d. None of the above

2.Which of the following is true for stateful transformation?

a. The processing of each batch has no dependency on the data of previous batches.

b. Uses data or intermediate results from previous batches and computes the result of the current batch.

c. Stateful transformations are simple RDD transformations.

d. None of the above

3. Which of the following is not output operation on Dstream

a. SaveAsHadoopFiles

b. SaveAs TextFiles

c. ReduceByKeyAndWindow

4. How many tasks does Spark run on each partition

a. one Task

b. Any no of tasks

5. How many messages are included in one RDD if the log messages are received at the rate of four messages per second and the batch interval is five seconds.in the DStream?

a.10

b.20

c.30

d.40

6.What are the basic components of any Spark application?

a.Driver and Executor

b.Driver, Executor and Receiver

c.Coordinator, Worker and Input Sources

d.Driver, Executor and Checkpoints

7.Which of the following options represent the objectives of Clustering?

A. Maximize intra-cluster similarity

B. Maximize inter-cluster differences

C. Minimize inter-cluster similarity

D. Maximize intra-cluster differences

E. All of the above

8.Explain in brief. What is Forgetfulness Metric? Why is it Important in ML? What are the 2 ways it can be applied in K-Means Clustering Algorithm?

9.What should be the decay factor if we want to use only the most recent data?

A. 1

B. 0.5

C. Any value between 0 and 1

D. 0

10.What are the ways to specify the half-life?

11.Advantages of checkpoint are:

a. limiting state re-computation in case of failure

b. resilient spark applications

c. fault tolerance for driver jobs

D . all of these

12\_\_\_\_\_\_\_\_\_is the open standard for token based authorization which allows a user's account to be used by a 3rd party without exposing password.

13.Spark streaming provides exactly how many number of semantics for processes?

a.4

b.1

c.2

d.0

14The coordinator process which executes the user program?

a.Executor

b.Driver

c. Worker

d. Slave

Answers:

1. A
2. B
3. C
4. A
5. B
6. B
7. A, B, C
8. Measure to balance the importance of new data vs old data. Makes algorithm adaptive to changing dataset. By specifying Decay factor or Half life
9. D
10. Number of batches and number of data points
11. A, C
12. Oauth
13. B
14. B