WEEK 15 QUIZ SOLUTION

1. The basic abstraction provided by Spark Streaming is called
   a. RDD
   b. Spark Core
   c. *DStream
   d. StreamingContext

   Explanation: DStream is a basic abstraction provided by Spark Streaming

2. Which of the following is not a high-level construct in Spark?
   a. DataFrame
   b. Dataset
   c. *RDD
   d. Spark SQL

   Explanation: RDD is a low-level construct in Spark.

3. In Spark-streaming application DStream can consist of:
   a. Multiple RDDs, all must be of same size
   b. A Single huge RDD
   c. Multiple RDDs, all might be of different size
   d. Multiple Stream of DataFrames/Datasets

   Explanation: DStream consist on multiple RDDs and its size can be different. Remember tap and balloon example.

4. Let's consider you have a streaming application which runs for 2 hours and batch interval size is 10 minutes, During the course of entire streaming application how many rdds will be created?
   a. 12
   b. 20
   c. 5
   d. 3

   Explanation: interval is 10 minutes and application run for 2 hours (120 minutes) so in total 120/10=12 RDDs will be created

5. In Spark Streaming module, transformations are applied by developers on:
   a. ***DStream**
   b. All RDDs
   c. Underlying Entities

   Explanation: For Spark streaming applications, Developers have to use DStream.

6. Spark is truly a streaming engine.
   a. TRUE
   b. ***FALSE**

   Explanation: Spark internally forms smaller batches and executes hence it can't be said as true streaming engine.

7. Which is true about Stateful transformations? (Choose possible options.)
   a. ***Stateful transformations operate on entire DStream**
   b. Stateful transformations always operate on RDD generated outside a batch interval
   c. Stateful transformations operate on entire source file
   d. ***Stateful transformations operate on a window of entities in a DStream**

   Explanation: True statements are A and D.

8. Which of these is not true in case of updateStateByKey. (Choose possible options.)
   a. It can find cumulative count of words on a DStream
   b. ***Replacing it with reduceByKey will still give same results**
   c. Checkpointing directory should be configured
   d. ***Function specified in updateStateByKey takes one parameter as input**

   Explanation: False statement are B and D

i. ReduceByKey is stateless and updateStateByKey is stateful, hence thess metgods can not be replaced to obtain same results.
ii. updateStateByKey accepts a function as parameters which has previous and current state.

9. Suppose we want to generate word counts over the last 30 seconds of data, every 10 seconds. Which of the transformations we can use here after applying a flatMap and map transformation on the DStream?
   a. reduceByKey
   b. updateStateByKey
   c. *reduceByKeyandWindow
   d. reduceByWindow

   Explanation: reduceByKeyandWindow supports stateful and sliding window and can be applied after flatMap and map transformation on the DStream

10. Which of these transformations does not require a pair RDD to operate upon?
   a. reduceByKeyAndWindow
   b. countByKey
   c. reduceByKey
   d. *reduceByWindow

   Explanation: B and D does not require a pair RDD to operate upon

11. Real processing of spark streaming data starts
   a. When we create a spark streaming context ssc
   b. When we set up all the transformations in the application
   c. *When we call ssc.start()
   d. when we use ssc.awaitTermination()

   Explanation: Spark streaming starts when we call ssc.start()

12. updateStateByKey is a stateful transformation and it requires 2 steps: 1. Define a state to start with. 2. A function to update the state
    a. True
    b. False

    Explanation: updateStateByKey is a stateful transformation and it requires 2 states previous and current state.

13. countByWindow will count the number of lines in the window.
    a. *TRUE
    b. FALSE

    Explanation: countByWindow will count the number of lines in the window

14. Internally, a DStream is represented by a big RDD
    a. TRUE
    b. *FALSE

    Explanation: DStream creates multiple small RDDs as per the interval.

15. The number of receivers must be more than the number of cores allocated to the Spark Streaming application.
    a. TRUE
    b. ***FALSE**

    Explanation: The number of cores allocated to the Spark Streaming application must be more than the number of receivers. Otherwise, the system will receive data, but not be able to process it.