Apache Spark Questions

1)Hadoop combiner functionally in spark using

- combineByKey()
- combineMap()
- reduceByKey()

2)If number of N (nodes in cluster) increases then impact of broadcast variable on network

- Increases as N increases
- No change
- Depends on broadcast variable memory
- One more option

3)Valid Window () operation in spark streaming.

4)For different code snippets which option gives fewer shuffles (2 questions)

5) For different code snippets which option takes less memory. (related to joins)

6)One questions related to partitionor() of Rdd what we need

- Checkpointing
- Persist()
- Unpersist() old data
- One more option

7)We cannot change the Rdd as they are immutable but in which situation values of that may Rdd changes.

- updateValue()
- when partition is lost when slide interval is not mentioned
- declare Rdd with var
- one more partition related option

8)Word count in java

9)Which operation require checkpoint->spark streaming operations

10)Some simple questions ->code snippet is given and ask which line represents Rdd actions, transformations, base Rdd etc.

11) **java.net. BindException: Address already in** use ->how to solve this select one option

12)one question related to read and write data in amazon S3

13)one question graphX -> code snippet is given and ask for output.

14)Parallelize a collection and apply keyBy() in that groupBy and ask for the output of the code snippet.

15)which one is not a persistence level

- MEMORY_ONLY
- MEMORY_CACHE
- ….

16)One theory question related to closures: **NotSerializableException** occurs , how to solve this.

17)Find the output of a parallelize () collection with map and flat map after Cartesian product.

18)Which spark SQL code in Scala give same result in python (it mainly contains groupBY (), orderBy (), sum (), count () and joins)

20)UpdateStateByKey() ->Given proper code snippet ask for what this code is doing in options.

21)Spark Sql code snippet using registerTempTable and then apply select query and from options select which one is giving same result (options contains dataframe. select (). where () and they will also change the order like dataframe. where (). select ())

22)One question from Spark Mllib (k-means)