Big Data Processing - End Sem Spring 2022

1. Roll Number *

18EE35014

2. The keys of a paired RDD has two fields. In order to re-partition the RDD what would be the best approach from among the options given below?

○ Use the spark default partitioner on the first field of the key

○ Use the spark default partitioner on the second field of the key

○ Use a custom partition function on any one of the fields

● Use a custom partition function involving both the fields

3. In spark pagerank algorithm, how would you like to store the pagerank score so that all the tasks can access the data?

○ In the driver

○ Create a new RDD for storing pagerank values

● Use a broadcast variable table

○ Use an accumulator variable table

4. Consider two rdds R1={(1,2),(2,3)} and R2 = {(3,4)}. How many elements the new rdd produced by the following operation will have?

R1. Cogroup(R2)

● 3

○ 2

○ 1

○ 0

5. What precaution needs to be taken before the following operation is performed on an RDD D?

D.groupByKey().map(row => (row._1, row._2.reduce(addFunc))).collect()

Assume that the volume of the data produced by the code is M GB.

○ Make sure the cluster RAM size is higher than M

○ Make sure the node executing the driver has hard disk size higher than M

● Make sure the node executing the driver has RAM size higher than M

○ Make sure the node executing the driver has combined hard disk and RAM size higher than M

6. In spark, when does the maximum data shuffle take place?

○ For narrow transformations

● For wide transformations

○ Both narrow transformations and wide tranformations are comparable

○ Depends on the data distribution

7. An RDD R is persisted with the operation: R. MEMORY_ONLY_2. During computation, a node has failed. Which of the following steps Spark is going to take?

○ It will re-compute R from scratch.

○ It will shuffle the data from one node to another and then run the other operations

○ It will repartition the data that was residing in the failed node

● It will execute the operations on the duplicate copy of the data directly

8. In spark, when does the maximum data shuffle take place?

○ For narrow transformations

● For wide transformations

○ Both a and b are comparable

○ Depends on the data distribution

9. Consider a graph which is denoted by g, where the edges are labelled with relationships. You want to count the total number of "friend" relationships between pairs of vertices. Choose the correct option to display the total count.

● g.edges.filter("relationship = 'friend'").count()

○ count(g.edges.filter("relationship = 'friend'"))

○ g.count(edges.filter("relationship = 'friend'"))

○ g.edges.count(filter("relationship = 'friend'"))

10. An iterative spark code performs only two operations in every iteration: The first is map() and the next is groupByKey(). If the number of iterations is set to 10, how many times the data will be shuffled?

○ 20 times

●

● 10 times

○ Unknown

○ 1 time

11. In spark, when a transformation operation is applied, which one of the following components performs the actual operations?

● Executors

○ The driver

○ The cluster manager

○ Both driver and the executors

12. An RDD R takes 10 GB of total storage. R is persisted using the command: R. MEMORY_AND_DISK_2. The available main memory of the cluster for storing RDD is 6 GB. How much disk this operation is going to consume?

○ 4 GB

○ 6 GB

● 14 GB

○ 8 GB

13. Which one of the following statements is true for spark?

● RDD can’t be changed in-place

○ RDD can’t be destroyed once created

○ RDDs can only hold basic data types

○ RDDs must always contain key-value pairs

14. How does spark create RDDs from continuous stream?

● By creating micro-batches in a time interval

○ By storing data into hard disk of a dedicated server and then reading from it

○ By storing data into distributed storage and reading from it

○ None of the above

15. Which of the operations from among reduceByKey(), combineByKey(), and lookup() benefit from partitioning?

○ Only reduceByKey()

○ Only lookup()

● All of them

○ None of them

16. You need to count the number of hashtags from tweets in last one hour. Which one of the following options would be the best choice?

● Use Dstreams API

○ Use structured API

○ Both can be used

○ None of the above

17. A paired RDD P = [('a',2) , ('a',4) , ('a' , 8), ('j',3),('j' , 7), ('r',1), ('r',3), ('r',5), ('r',6), ('c' ,9)] is given.

After applying the following transformation operation on P,

output = P. **A().** mapValues(lambda values : sum (values))

we have got the output i.e. **[('j' ,10) ,('r' , 15) , ('a' , 14) , ('c',9)].**

Then, A() is

● groupByKey()

○ groupByValue()

○ reduceBykey()

○ reduceByValue()