SRN | | | | | | | | | | | | |

PES UNIVERSITY

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE18CS322

December, 2021 : END SEMESTER ASSESSMENT (ESA)

UE18CS322 - Big Data

Time: 3 Hrs | Answer All Questions | Max Marks: 100

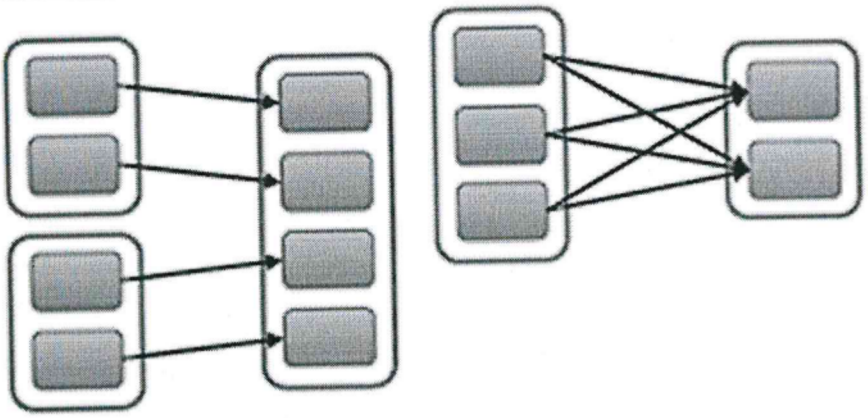
1.	a.	Consider an application where satellite images of city roads were captured every minute at various junction points of the city. What solution you suggest for the storage and processing of such data? Justify your choice with appropriate reasons.	5
	b.	We have rainfall data collected every day for last 50 years. We want to sort these values using a map reduce program, so that the operation can be done faster. Mention what will be the input and from where it will be taken for each of these functions – Mapper, Combiner and Reducer. Is it necessary to use Combiner function? Justify your answer.	5
	c.	What is data locality? How it is relevant in Hadoop Map Reduce Job?	5
	d.	Consider we have a text file with all the customer information of size 20 GB to be stored on HDFS. Assuming HDFS is having 6 data nodes, compute how many data blocks will be created and how the data gets stored across these nodes? (Assume Hadoop 2.X for default block size)	5
2.	a.	List any five tools of Hadoop ecosystem along with their purpose.	10
	b.	What is sparse matrix notation and where it is useful? Consider the matrix M =	10

20	0	0	0	0
0	0	4	1	0
0	0	0	10	2
0	17	0	0	0
0	0	0	1	5

Write the sparse matrix representation for the above matrix. (Assume row and column no. Starts with Zero)

3.	a.	What is a lazy computation in terms of Spark RDD? How fault tolerance is supported in Spark?	5
	b.	For page rank computation among Hadoop, Spark and RDBMS, which framework you would prefer? Justify your choice.	5
	c.	What is the difference between Jobs and Tasks in YARN? What are the different types of task failures?	5
	d.	Why columnar databases are preferred for analytical queries over HDFS? HBase is composed of three types of servers in a master slave type of architecture. What are the three types of servers?	5
4.	a.	In a Streaming Spark program that is receiving a stream stocks with key value pairs of the form <stock, number of stocks>, consider the statement stocks.c(Minutes(10), Seconds(1)) maxByValueAndWindow is a new function that computes the maximum of the value in a	4

SRN | | | | | | | | | | | |

		window similar to countByValueAndWindow. How can this be implemented in Streaming Spark to compute this quantity efficiently; i.e., without recomputing the max for every window from the beginning?	
	b.	 Fig (a) Fig (b) Above diagram indicates some concept in Big Data. Illustrate what concept it is illustrating and what is its importance?	8
	d.	Given the numbers [41, 23, 101, 44, 65, 90], using a k-means implementation using Map-Reduce, show what the output of the Mapper and reducer will be after the first iteration? Highlight the keys and values of the Mapper output and output of the reducer. Assume that we are clustering the set into two clusters (33 and 54) as our first estimates of the centroids.	8
5.	a.	With a neat block diagram, explain how the Hadoop job is handled by YARN in a Hadoop cluster. Clearly describe the roles of Application Master (AM) and Resource Manager (RM) in the process.	10
	b.	Given the following hash functions used with a bloom filter initialized with 10 buckets. Identify if the numbers **93** and 87 are recognized as a ***seen*** number by the bloom filter. Hash function 1 – (x*19)% 10 Hash function 2 = (x * 5) % 10 Hash function 3 = (x * 7) % 10 1 1 0 1 0 1 1 1 0 0 0 1 2 3 4 5 6 7 8 9	10