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

PES UNIVERSITY

PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

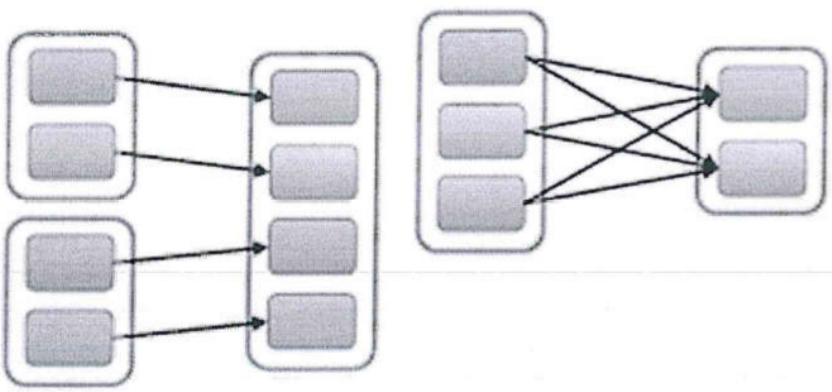
UE17CS313

END SEMESTER ASSESSMENT (ESA) – Backlog
Department of CSE , December, 2020
UE17CS313 - Big Data

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

1.	a.	Consider an application where traffic images 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.	Assume a map reduce program for sorting operation – mention what will be the input and from where it will be taken for each of these functions – Mapper, Combiner and Reducer.	5
	c.	In Hadoop Map Reduce Job, which of the tasks (Map or Reduce or both) have the advantage of data locality? Justify your answer.	5
	d.	Consider the data file of size 10 GB to be stored on HDFS. Assuming HDFS is having 6 data nodes, show how the data gets stored across these nodes, in terms of data blocks. (Assume Hadoop 2.X for default block size)	5
2.	a.	What tools of Hadoop ecosystem can be used to analyze the web server logs?	5
	b.	What are structured and unstructured data? For each of the following data storage organizations - mention what kind of data can be stored and queried. Give your reasons for this choice, along with the type of the DBMS. i. SQL Database ii. HBASE iii. HIVE iv. HDFS	5
	c.	If we are to write a map-reduce code for a sort function, then suggest what computations should be done in the mapper and reducer functions. Suggest which function should have the logic for sorting part. What would happen if a map task fails to complete?	5
	d.	Consider the matrix M = <table><tr><td>2</td><td>0</td><td>0</td><td>0</td></tr><tr><td>0</td><td>0</td><td>4</td><td>1</td></tr><tr><td>0</td><td>0</td><td>0</td><td>10</td></tr><tr><td>0</td><td>7</td><td>0</td><td>0</td></tr></table> What is the best way to represent this matrix in Big Data perspective? Give that representation. Justify your answer. (Assume row and column no. Starts with Zero)	5
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

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

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 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?	5
	b.	 Fig (a) Fig (b) A student studying spark took down the above figures but forgot to label the context in which he had taken down the figures. Can you provide a suitable explanation of what these figures represent?	5
	c.	How can we find the maximum number of stocks sold in any window using Streaming Spark in the program in 4(b); illustrate using pseudo code?	5
	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. If the MR program is running on two nodes, what will you use as intermediate keys for the computation?	5
5.	a.	What are the short comings of Hadoop 1.X in terms of resource management? Which component performs the equivalent of JobTracker in Hadoop v2? Which component performs the equivalent of TaskTracker in Hadoop v2?	5
	b.	"In YARN, Application Masters codify their request for resources using a Resource Request that contains – no of containers, resources per container, locality preferences and priority of requests within the application" What is meant by resources per container? Why is this required? What are locality preferences? Which YARN scheduler uses a FIFO ordering for scheduling jobs?	5
	c.	What are speculative duplicates in Hadoop? What component of Hadoop ecosystem provides this feature?	5
	d.	Let us assume that you have data about students in the format <USN, course, marks> where marks is out of 100. You need to compute the total marks obtained by a student using Spark dataframes. Write pseudo code to solve the problem.	5