SRN □□□□□□□□□□□□

PES UNIVERSITY	**PES University, Bengaluru** (Established under Karnataka Act No. 16 of 2013)	UE16CS313 CS313

December 2018: END SEMESTER ASSESSMENT B.TECH. 5th SEMESTER

FINAL EXAM

UE16CS313 - BIG DATA

Time: 2 Hrs | Answer All Questions | Max Marks: 60

Instructions:

- Answer questions in the space provided
- For problems, show the working. Providing just the answer is not acceptable.
- One memory recall sheet (A4) is permitted in your own handwriting.

Name	
USN	
Signature	
Marks	

			5x3=15
1		Answer in 2-3 lines.	
	a)	YARN is configured with 2 queues each sharing 50% of the resources. A job submitted to YARN seems to be using all the resources as it is the only job on the system. Which scheduler do you expect YARN to be using and why?	3
	b)	What is Flume? Where will you use it?	3
	c)	Why do scala programs not have to deal with concurrency issues?	3

SRN

	d)	What are reducer size and replication rate with respect to Big Data Algorithm complexity?What will be the value of replication rate for a map-reduce word count application without using combiners	3
	e)	In Spark MLLib, what are transformers, estimators and evaluators?	3

2	a)	Given two tables stored on HIVE 1. party (contains party_name, consituency_name) 2. constituency(consituency_name, improvement_description, year, date) - logs every improvement made in a constituency We need to find out the total number of improvements made by different political parties in a 2018. For example, how many improvements did ModernParty make in 2018. (i) design a HIVE SQL query to achieve this objective(exact syntax is not neessary. Important to get the logic right). (ii) illustrate through a DAG, how HIVE will translate this query through a series of Map Reduce steps to execute this query.	1+4

SRN □□□□□□□□□□□□

	b)	How would you store the following unstructured data onto a columnar database such as HBase? Design the column families required and illustrate your design by working out the column families/columns for the statements given below. "2.0 is a science-fiction action film starring Rajinikanth released in November 2018." "Maze Runner is a fast paced action Hollywood action movie directed by Joe Wright received poor reviews from the critics"	5

c) Some of the parameters that characterize a Spark RDD are - its Partitions, the dependencies of the RDD on its parent, the computation that is performed on the partition, the preferred locations for a partition and partitioner function. Consider a Spark application that performs a *groupByKey()* transformation.

What will be the parameters for such a groupByKey. transformation in Spark ? For your reference a solution for filterRDD has been provided. Use that as a template to determine the solution for a groupByKeyRDD

3+2

	Example for FilterRDD	Solution for groupByKeyRDD
Partitions	*same as parent*	
Dependencies	*1-1 with parent*	
Compute	*compute parent and then filter it*	
Preferred Locations	*ask parent(none)*	
Partition function	*none*	

SRN □□□□□□□□□□□

3	a)	Elections are being held in India in 2019 and the analysis is being partially automated. Each polling booth generates at every 1 minute interval, a data in the following format - <polling booth, constituency_name, partyname, #votes>. Such a tuple will be generated for each party. You would like to use Apache Storm to compute total number of votes currently cast per party in every constituency. Design a Storm topology to do same indicating Streams, Spouts and Bolts for the above problem. Indicate what type of a grouping you will use at each stage along with a justification.	5
	b)	If we try to cluster the data in the following dataset {23, 32, 82, 99, 8, 105} using kmeans algorithm(using 23 and 32 as initial guesses of centroids) and we are using an iterative MapReduce implementation of Kmeans. Enumerate the key-value pairs at the end of the first two iterations of kmeans for the Map Phase. Also compute the new centroids in reduce phase.	6

	c)	Write spark streaming pseudo code to measure the average number of votes polled over a stream of values for a window of 10 minutes and moving the window by 2 minutes for the problem given in 3(a).	4

SRN □□□□□□□□□□□□

4	a)	Given the following hash functions f1(x) = (x+3) % 10 and f2(x) = (x*7) %10. You are given a bloom filter with 10 buckets. Your bloom filter is initialized with two non spam values 32 and 25. Determine whether the number 31 is a spam. Show the working.	5

	b)	What is meant by *lazy execution* in Spark? For the following Spark code identify the transformations and actions and illustrate how it will be lazily evaluated
		`data.map{ case (id, age, gender, profession,zipcode) => (gender, 1) }`
		`.reduceByKey(_ + _).count()`
			5

	c)	What is the difference between Jobs and Tasks in YARN? What are the different types of task failures?	5

SRN □□□□□□□□□□□□

Additional Space to be used in case of mistakes

SRN

SRN □□□□□□□□□□□□