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

PES UNIVERSITY

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE14CS314

END SEMESTER ASSESSMENT(ESA) B. TECH 5th Semester

UE14CS314 - BIG DATA

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

Instructions:

- Provide short answers and to the point (max 3-4 lines) for all questions. Long answers when not asked for will not be looked upon favorably
- For problems, show the working. Providing just the answer is not acceptable.
- Answer all questions in the order they are asked. Leave space and come back if you need time to think
- One memory recall sheet (A4) is permitted in your own handwriting. Please attach that to your answer booklet

1		Provide short answers to the following questions. In case of True/False, please provide justification	10x2 =20
	a)	Which YARN scheduler uses a FIFO ordering for scheduling jobs?	
	b)	What security measures does Twitter take to deal with Photographs?	
	c)	What is the role of the SerDe component in HIVE?	
	d)	Why is a columnar database more suited for Analytics applications?	
	e)	What is a DRM in Mahout?	
	f)	If a Bloom filter determines that a key is in a set (e.g, a userid is in a set of known non-spam userids) then it is not necessary to look up the key from a backup store. True/False	
	g)	What are the function of a *supervisor* in Storm?	
	h)	What is the *batch size* in Streaming Spark?	
	i)	Which of the following operations is not supported by PNUTS? Why? Get, Set, Scan, Join	
	j)	How is load balancing achieved in PNUTS?	

2	a)	Suppose a stream of temperatures generated by weather monitoring stations is available. The records are in the form <time, location, temperature> which gives the time, location and value of temperature. For example, <2016_5_12.16:55:00, Bangalore-1, 25> means that at 16:55:00 on 12 May 2016, the temperature at the Bangalore-1 station was 25 degrees. Draw the Storm topology needed for a Storm program that continuously outputs the value of the maximum temperature for each location for the day whenever it changes. Assume that the computation is too large to be handled by a single bolt. Show (i) how data is transferred between the various elements in the topology (ii) any data structures needed (iii) computation carried out in each bolt and output.	2+1+2
	a)	Explain the working of a Bloom filter with multiple hashes. Explain (i) data structures needed (ii) how the filter is initialized (iii) how it is used	2+1+2
	b)	Suppose the program of 2(a) is to be done using Streaming Spark and MapReduce. The	2+1+2

PTO

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

		volume of the stream is too large to be handled by a single mapper or reducer. Draw the Streaming Spark workflow diagram needed for a program that continuously outputs the value of the current maximum temperature for each location for the day whenever it changes. Show (i) The incoming stream and conversion to an RDD stream (ii) What the mapper does (iii) any data structures needed (iv) computation carried out by the reducer(s).	+2
	c)	What is lineage in Spark? How is it used to handle failure of a task?	1+2

3	a)	Given that you have a HIVE installation with 2 tables in it -- **customers** and **purchases**. The customers table has the following structure <cust_name, cust_id, dept> and the purchases table has the structure <purchase_id, cust_id, store, amount> and you run the following SQL query on it: **select** dept, count(1) **from** customers **group by** dept; (i)What does this query do? (ii)How many map-reduce jobs will this require?	2+3
	b)	Physically, HBase is composed of three types of servers in a master slave type of architecture. (i) What are the three types of servers? (ii) Why is Hbase a master-slave type of architecture?	3+2
	c)	Cluster the following data into two clusters using a k-means map-reduce implementation. Assume that the first two points form the initial cluster centres. (i)Perform two iterations. For each iteration, show what the map and reduce stages will output. (ii) What will you use as a distance measure Data: 90, 80, 43, 26, 18, 75 Express the solution in the following tabular format	5
	d)	"In YARN, Application Masters codify their request for resources using a ResourceRequest that contains – no of containers, resources per container, locality preferences and priority of requests within the application". (i)What is meant by resources per container? Why is this required?(ii)What are locality preferences?	3+2

Iteration 1 Map output	Iteration 1 Reduce Output	Iteration 2 Map Output	Iteration 2 Map Output