UE17CS313#1550#42

SRN

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE17CS313

December 2019: END SEMESTER ASSESSMENT B.TECH. 5 SEMESTER

FINAL EXAM

UE17CS313: BIG DATA

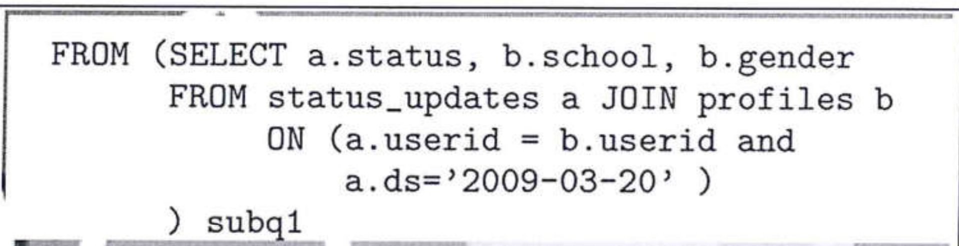
Time: 2 Hrs | Answer All Questions | Max Marks: 60

Instructions:

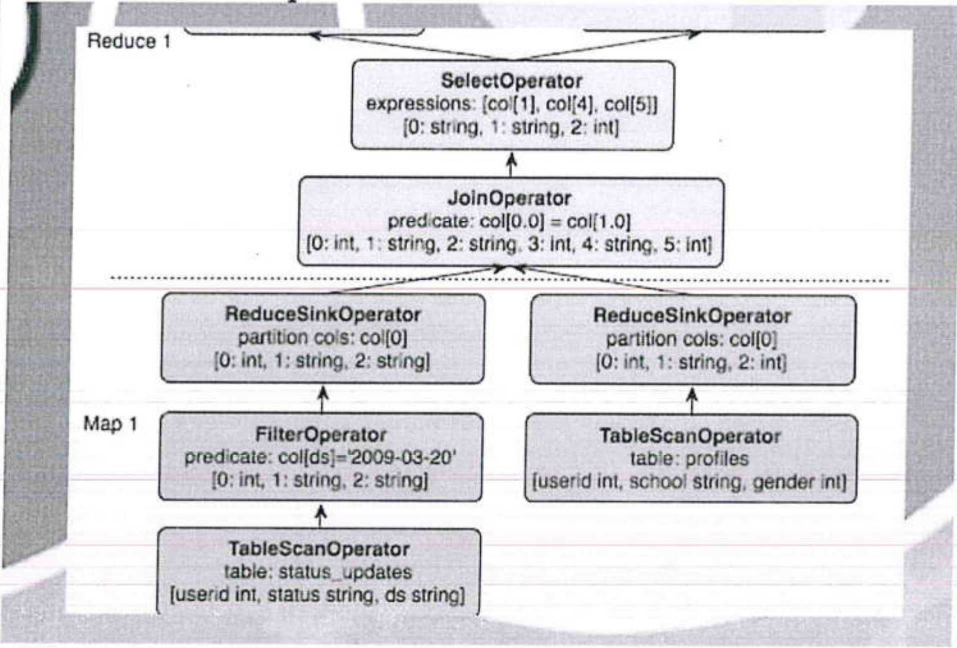
- For problems, show the working. Providing just the answer is not acceptable. Clearly state your assumptions.
- One memory recall sheet (A4) is permitted in your own handwriting.

1			
	a)	During an HDFS Read, describe the interaction between the HDFS client and the other components of the HDFS system	3
	b)	For doing a sort in MapReduce, describe the input to the Mapper, the output of the Reducer, and the partitioning function	3
	c)	How does MapReduce in Hadoop handle failure of a node?	1
	d)	Suppose the matrix M = $\begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}$ is to be multiplied by the vector V = $\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$ (a) Show the sparse matrix representation of M (b) If the multiplication is to be done by MapReduce with the vector V stored in memory, show the output of the Mapper and the output of the reducer	2+3
	e)	Consider the fragment from the compilation of a Hive query shown below	1+2

SRN



```
FROM (SELECT a.status, b.school, b.gender
      FROM status_updates a JOIN profiles b
           ON (a.userid = b.userid and
               a.ds='2009-03-20' )
      ) subq1
```



Reduce 1
SelectOperator
expressions: [col[1], col[4], col[5]]
[0: string, 1: string, 2: int]
JoinOperator
predicate: col[0.0] = col[1.0]
[0: int, 1: string, 2: string, 3: int, 4: string, 5: int]
ReduceSinkOperator
partition cols: col[0]
[0: int, 1: string, 2: string]
ReduceSinkOperator
partition cols: col[0]
[0: int, 1: string, 2: int]
Map 1
FilterOperator
predicate: col[ds]='2009-03-20'
[0: int, 1: string, 2: string]
TableScanOperator
table: profiles
[userid int, school string, gender int]
TableScanOperator
table: status_updates
[userid int, status string, ds string]

(a) Which part of the Hive query is executed in the step marked **FilterOperator**?

(b) The record output by the **JoinOperator** has 6 fields. What are likely to be their contents? Hint: look at [0: int, 1: string, 2: string, 3: int, 4: string, 5: int]

2	a)	Consider a Bengaluru smart city application that monitors traffic between two different junction points in Bengaluru every 1 minute and stores these on log files on a server. Each data point is variable length and approximately 1KB in size. Each data point contains the following information *<location, timestamp, source, destination, vehicleCount>* and there are 4096 locations from which traffic data is gathered. You have written map reduce code to identify the most congested locations and need to export the data out to an SQL database every day. Which tools will you use to import the data into HDFS? and which tool will you use to export the data out to the SQL database?	3
	b)	"Acyclic data flow is inefficient for applications that repeatedly reuse a working set of data" (i)Give an example of acyclic data flow framework (ii) Why is it inefficient for the case mentioned?	3
	c)	Some of the parameters that characterize a Spark RDD are - its Partitions, the dependencies of the RDD on its parent, and partitioner function. Consider a Spark application that performs a *reduceByKey()* transformation. What will be the parameters for such a *reduceByKey*. transformation in Spark ? For your reference a solution for filterRDD has been provided. Use that as a template to determine	3

SRN

the solution for a reduceByKey RDD

	Example for FilterRDD
Partitions	*same as parent*
Dependencie s	*1-1 with parent*
Partion Function	*None*

	d)	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, USN, course, studentname => (course, 1) }`
		`.reduceByKey(_ + _).count()`
			3

	e)	Join operations in Spark can result in narrow or wide dependencies depending on what factor? Illustrate with a diagram and example how join operations can have either narrow or wide dependencies.	3

3	a)	Consider the statement hashTags.window(Minutes(10), Seconds(1)) in a streaming spark program. Assume that hashTags is a stream. What are the RDDs produced by this statement?	2
	b)	Consider the statement hashtags.countByValueAndWindow(Minutes(10), Seconds(1)) in a streaming Spark statement. What are the two operations performed by Streaming Spark to compute this quantity efficiently?	3
	c)	In Kafka, how is a topic implemented in the system?	3
	d)	In a web site, suppose we want to use a sample to determine the distribution of the time between successive visits by users of the web site. How should the sample be drawn?	3
	e)	Consider the statement SELECT * FROM A JOIN B ON A.x=B.x WHERE A.z=500 Assume that the probability of the join condition being satisfied is *p*, the probability of the where clause being satisfied is *q*, and the number of records in tables A and B are N_A and N_B, respectively (a) What is the communication complexity of the statement if the join operation is done first followed by a select (b) What is the communication complexity of the statement if the select operation is done first followed by a join	2+2

4	a)	Given the numbers [201, 54, 43, 102, 151, 99, 88], 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 using the first two numbers (201 and 54) as our first estimates of the centroids	3
	b)	Given a very large dataset (containing over 100 million images of monuments), you are building a machine learning pipeline to classify a given monument image as either Ancient or	3

SRN □□□□□□□□□□□□

		Modern . You are considering to use the SVM model. (i) Illustrate with a diagram the ML pipeline construction using transformers, estimators and evaluators using ML Lib. Show what will happen in each step	
	c)	You have over a 100 machines in a YARN cluster and have created two queues reserving 30% of resources for queue 1 and 70% of resources for queue 2. You submit a MapReduce job to process 5GB of data on HDFS for queue 1 and observe that the system starts all the mappers immediately. How many machines are required for processing map jobs if each machine can run 1 map job at a time. Can you conclude as to which scheduler is being used by YARN? Justify	3
	d)	"Mesos delegates control over scheduling to the *frameworks* using an abstraction called *resource offers*. Each framework consists of a *scheduler* and an *executor*". What are *frameworks* and *resource offers*?	3
	e)	A Rutgers University study examining decision making process during emergency found that a large number of tweets originated from Manhattan during Hurricane Sandy. Can we conclude that Manhattan was the most affected area? Why/Why not?	3