PES UNIVERSITY

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE19CS322

DECEMBER 2021: END SEMESTER ASSESSMENT (ESA) B TECH V SEMESTER

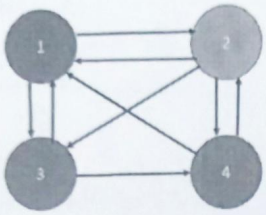
UE19CS322 – BIG DATA

Time: 3 Hrs	Answer All Questions Clearly mark the question number/part and write the entire answer for a subquestion together.	Max Marks: 100

1	a)	Volume, Veracity, Variety and Velocity are considered to be the 4V's of Big Data. Give an example to demonstrate each of these. Also, give two reasons as to why Volume has become a challenge in Big Data when compared to traditional data.	6
	b)	"Combiners are considered as mini-reducers in Map Reduce" – What are combiners and why are they considered as mini reducers. A map reduce programmer submitted a MR program that had optimized the MR program by replacing the reducer by a combiner. Is this guaranteed to give correct results? Justify	4
	c)	What is HDFS – briefly explain any two features of HDFS? Why is the HDFS architecture termed a master-slave architecture – briefly explain the major different components and their functionality? Discuss the motivation behind this architectural organization? (3+5+2)	10
2	a)	1. Which of the following components does not belong to the Hadoop eco-system? a. Sqoop b. Pig c. Mahout d. Dbase 2. Workflow does not involve a. Flow of work from consumer to producer b. Steps to be run c. Executing the specified steps identified d. Error handling 3. Apache Oozie a) Schedules the different types of jobs b) Runs different types of jobs c) Manages different types of jobs d) All the above 4. Identify the odd one. These are needed to install Ambari in a Cluster a) Stack of Services b) Services c) Components making up the service d) CLI for installation of the clusters	4

SRN

b) Write down the adjacency matrix and transition matrix(columns as source and rows as destination) of the following directed graph and compute page rank using Map Reduce for 1 iteration assuming that initial page ranks of each node is ¼. Please use the format shown below to show the working of the Map reduce program. Assume that the transition matrix is stored on HDFS in two blocks – columns 1 and 2 are stored on Node 1 and columns 3 and 4 on Node 2 8



Node 1 Map Intermediate key	Node 1Map Intermediate value	Node 2 Map Intermediate key	Node 2 Map intermediate values
Reduce input key	Reduce input value	Reduce output key	Reduce output values

c) Given that you have the following data stored on HDFS using CSV files 8

TaxPaid Table

Name	PAN #	Date	Tax Paid

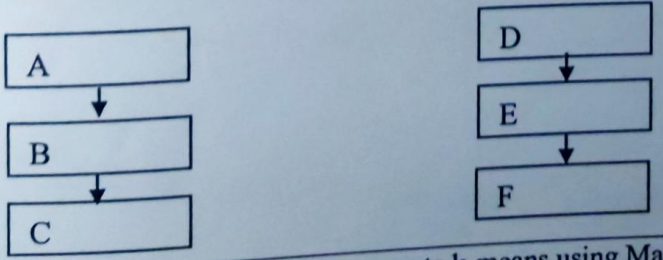
BankDetails

PAN #	State in which tax paid	Bank account.

You need to design a Map-reduce program to compute the total tax paid per state. Which relational operations will you require to use to perform this computation. How many map-reduce steps will you require to perform the computation(2 marks) and what will be relational operation performed? (4 marks) Express your solution in the following format

MR Step no	Relational Operational operations performed	Intermediate key value pairs; Intermediate keys are output of mapper	Final key-values; final keys are output of reducer.

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

3	a)	Which of the following statements is true/false about Apache Spark? Mark out solution as <subpart#, T/F> 1) Spark RDDs are stored in disk between transformations F 2) reduceByKey is a transformation 3) RDDs are stored in the memory of spark workers 4) Spark workers wait for a threshold number of transformations before executing them in Lazy execution F 5) RDDs are useful for storing lineage information 6) Spark is faster than Hadoop as it performs its operations in memory.	6
	b)	What is lazy evaluation in Spark? Explain with an example. (4) Why does Spark opt for this approach? (2)	6
	c)	What are narrow and wide dependencies and how are they useful in Spark Scheduling? Given the following operations, what type of dependency will they result in? (4+4) - reduceByKey - map - filter - join	8
4	a)	Your team is tasked to write a stateful streaming spark application for measuring pollution in the environment and need to compute *TotalPollutionLevel* (TPL) in the atmosphere using a given formula on a stream of data coming from sensors. The application needs to keep track of the maximum TPL seen over the month in different parts of the city. You are presented with two designs – (A) which stores the max TPL in a global variable, (B) which stores max TPL in a local file. You reject both the designs as they have a flaw. Point out the flaw in the above two designs.(4) What would be your alternative design?(2)	6
	b)	What are topics and partitions in Apache Kafka? In Kafka, partitions are replicated for fault tolerance. Briefly explain the replication model used in Kafka? (4+4)	8
	c)	Given the sequence of hashes of items in a stream as follows – 0x32, 0x44, 0x1e, 0x40, 0x38, use the Flajolet Martin algorithm to estimate the total numbers of unique elements in the stream. Comment on the accuracy of the estimate as why it is or it is not accurate.	6
5	a)	Given below is a Spark MLlib workflow designed by an engineer to use RandomForest to classify a set of tweets into different categories like Sports, Politics, Entertainment etc.. The boxes marked A, B and C represent a training pipeline while the boxes marked D, E and F represent a testing pipeline in the design. Unfortunately, the designer forgot to mention which boxes will be transformers, estimators and evaluators. The input is a set of labelled tweets.Complete the design by marking the individual blocks and indicate what operation needs to be done at each step in both pipelines in order to achieve the goals. Assume that input is a set of labelled tweets.  	10
	b)	Outline the algorithm used to compute k-means using Map-Reduce. (4)Suggest one optimization to improve the performance of map-reduce.(2) Use the algorithm to compute one iteration of k-means for the following data { 48, 49, 67, 90, 19, 105,130} using 48 and 49 as initial estimates. Assume k = 2.(4)	10