

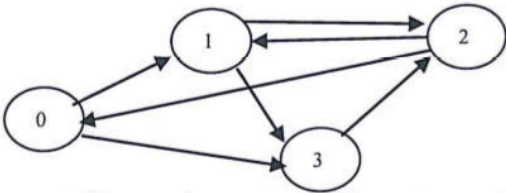


## DECEMBER 2020: END SEMESTER ASSESSMENT (ESA) B TECH V SEMESTER

## UE18CS322 – Big Data

Time: 3 Hrs	Answer All Questions in the order they are asked	Max Marks: 100
-------------	--	----------------

1	a)	Suggest two reasons why the amount (Volume) of data in Big Data is exponentially increasing and is larger in comparison to the traditional data with examples.	5
	b)	You have setup a new cluster of Hadoop v2 and are in process of inserting the first file into HDFS and the size of the file is 171MB. Outline the insertion process in terms of what changes will occur on the namenode files and how many datanodes are used for the storage of this file. Assume 3 replicas for HDFS.	5
	c)	Consider that you are given a file containing the annotated form of the Mahabharata which runs into 4GB as a text file. The Mahabharata is broken into 18 chapters of parvas and each parva had many shlokas. Different shlokas were given to different scholars for translation to English and each shloka and its translation were entered into a web page that accepted data in the following format and stored it on a text file <i>Parva Number, Shloka Number, Translation</i> And hence were in random order in the file "Mahabharata.txt" which was stored on HDFS. Design a MapReduce program to sort all the shlokas and their translations in the right order both based on the Parva and the shloka within it. You need not write entire map-reduce code, but need to identify the intermediate keys of the mapper and corresponding values. The keys at the reducer and the corresponding output values. Do you need anything else to make this work? How many mappers do you expect to start if you are using Hadoop v2?	5
	d)	A student taking notes on YARN drew the diagram but forgot to mark the various components. Identify where application master, node manager and resource manager would run? If we are to run 2 map reduce jobs, each requiring 3 mappers and 1 reducers, how many tasks, node managers, resource managers and application masters would run if we assume that there are only 3 worker nodes and 1 master in the cluster. <div style="text-align: center; margin-top: 20px;"> <div style="border: 1px solid black; padding: 5px; width: 100px; margin: 0 auto 10px auto;">Master</div> <div style="display: flex; justify-content: space-around; margin-top: 20px;"> <div style="border: 1px solid black; padding: 5px; width: 100px;">Worker A</div> <div style="border: 1px solid black; padding: 5px; width: 100px;">Worker B</div> <div style="border: 1px solid black; padding: 5px; width: 100px;">Worker C</div> </div> </div>	5

2	a)	Consider the following graph of web pages connected to each other  Represent this graph as sparse adjacency matrix and a transition matrix. Assume that each of the nodes in this transition matrix are distributed in an HDFS file across 2 nodes such that if a destination node of an edge is even it is on node 0 and if it is odd it is on node 1. Perform a page rank computation of one iteration assuming all initial page ranks are 1. Show intermediate keys to mappers and values and input/output keys of the reducer assuming a single reducer. How many mappers will be started for the 2 <sup>nd</sup> iteration of the page rank and why?	10																			
	b)	Given a table for IPL that has the following schema - <i>&lt;match#, ball#, batsman name, runsscore&gt;</i> you need to determine the total #runs scored by all the batsmen. Write MR pseudocode as a graph showing the sequence of map-reduce steps that will be used to find this? You need not write actual MR code	5																			
	c)	Given the following data about players – “Rohit Sharma is a right handed opening batsman for India” and “Hardik Pandya is hard hitting middle order batsman with a strike rate in excess of 150 and a right arm medium fast bowler” and you need to store this data in a columnar database like HBase. Design this in terms of column families and columns and show how the two descriptions given above will be stored in the database	5																			
3	a)	What is <i>Type inferencing</i> in Scala? Give an example to illustrate	5																			
	b)	Consider the following Spark Code. What does the following code do? If this code is distributed across 3 workers, and if worker 3 dies after executing the filter function, how does Spark recover? val lines = sc.textfile("ipl_commentary.txt") val l = lines.filter("OUT") val pairs = l.map(x.split(" ")(0), x)) val m = pairs.countByKey()	5																			
	c)	For the <i>reduceByKey</i> transformation in Spark, what would be (i) #partitions (ii) preferred location of each partition (iii)dependencies (iv) partitioner?	5																			
	d)	Why does the communication cost model of complexity only consider input size and not output size? What does wall clock time required for a map-reduce job depend on?	5																			
4	a)	What does the following code do? Clearly indicate what the + and the – operators are used for? playerScores.reduceByKeyAndWindow(_+_, _-_, Minutes(5), ...)	5																			
	b)	What are partitions in Kafka? How is fault tolerance of partitions achieved?	5																			
	c)	Given the following hash functions used with a bloom filter initialized with 10 buckets. Identify if the number 23 is recognized as a <i>seen</i> number by the bloom filter. Hash function 1 – (x*17) % 10 Hash function 2 = (x *3) % 10 <table border="1" data-bbox="243 1751 1353 1824"><tr><td>1</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td></tr><tr><td>0</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td></tr></table>	1	0	0	1	0	1	1	1	0	0	0	1	2	3	4	5	6	7	8	9
1	0	0	1	0	1	1	1	0	0													
0	1	2	3	4	5	6	7	8	9													
d)	As per the Flajolet Martin Algorithm, the probability that at least one element has a tail length <i>r</i> is $1 - e^{-m^r}$ . What is tail length? Is this statement true under all conditions? If not, then provide arguments to support the condition in which this is true?	5																				



5	a)	How are the number of centroids in a MR version of the k-means algorithm related to the number of reducers? The k-means algorithm terminates when the centroids do not change across iterations. However, the previous centroids are in the reducer of the previous iteration while the current centroids are computed on the reducer of the current iteration. The reducers of the previous iteration and current iteration may be running on different machines. How can this comparison be performed? If one of the mappers that is used for computing the centroids crashes, will we have to recompute all the centroids from the beginning?	10
	b)	Consider a scenario where we are receiving retail sales data of purchase made on a site like Amazon. A machine learning algorithm has to be run on this data by copying it to the big data compute engines, refined and then used on their production servers during the weekend. Based on the tools learnt in the course, design the entire workflow pipeline and show what tools you will use for the same?	6
	c)	Give two examples of Big Data solutions for Deep Learning?	4