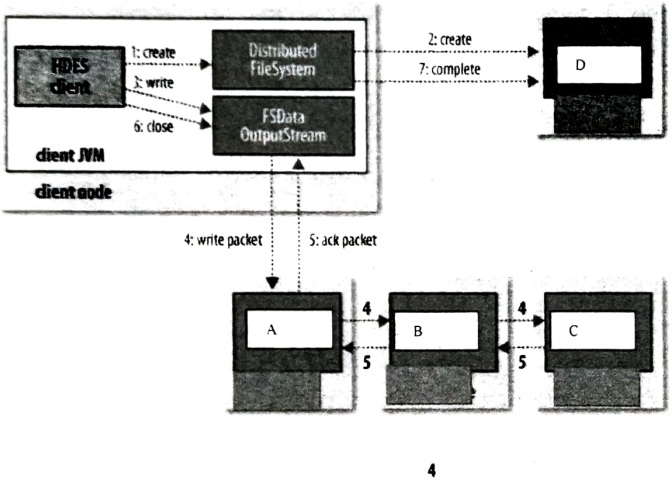
PES University, Bengaluru
(Established under Karnataka Act No. 16 of 2013)

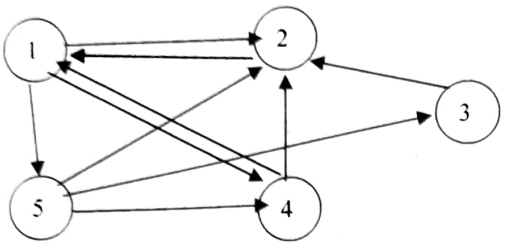
UE18CS322

## OCTOBER 2020: IN SEMESTERASSESSMENT B Tech SEMESTER TEST – 1

## UE18C322(4 credit subject) - BIG DATA

| Time: 2 Hrs | Answer All Questions | Max Marks: 60 |
|---|---|---|
| | • **Please answer all questions in sequence.**<br>• **Show working for the problems and state your assumptions; giving just the solution will not get any marks** | |

| | | | |
|---|---|---|---|
| 1. | a) |  <br>Given above is the sequence diagram for HDFS write operation. The diagram was drawn by student during the class but he/she forgot to name the boxes A, B, C and D. You have to label the 4 components A, B, C, D. Also, the student is not sure why the write packet is forwarded to B via A and not directly written to B. Provide a justification for this operation. | 4 |
| | b) | *"Secondary namenodes are introduced in HDFS to help tolerate failures"*. Is this statement true/false? Please justify your answer with a reasoning. Also which of the 4Vs of Big Data characteristics does storing data in HDFS address. | 4 |
| | c) | An analysis of data on drowning based deaths and ice cream sales shows that as ice cream sales increase, the number of drowning based deaths also increase. What pitfall of data analysis does this indicate and what do you think is the reason? (2-3 lines) | 2 |
| 2. | | Consider that you are given a file containing the annotated form of the Ramayana which runs into 900MB as a text file. The Ramayana is broken into 7 major Kaandas (books) namely the *Baala, Ayodhya, Aranya, Kishkindha, Sundara, Yuddha, Uttara kaandas*. Unfortunately the sentences of this text file are not in order and are jumbled up. However, each sentence of the text is stored in the format *<kaandaname>, <sentence>* (assume there are no commas in the sentence). For example, a short section is of the form<br>*Yuddha, Hanuman set off to bring back the Sanjeevani plant*<br>You have been asked to process this file using Map Reduce using Hadoop v2. Answer the following questions providing justifications. Credit will be awarded only if the justification is right. | |
| | a) | Write MR pseudo code to identify the number of sentences in each *kaanda?* Identify the intermediate keys and final keys. | 4 |
| | b) | How many mappers and reducers would be used for processing this? Will a combiner help to | 4 |

| | | | |
|---|---|---|---|
| | | improve the performance? | |
| | c) | One of the mappers is progressing slowly. How does the Hadoop YARN framework respond to this? | 2 |
| 3. | |  <br>The above graph shows the a sample of the internet of 5 pages with the edges representing the links between the pages. | |
| | a) | Express the graph as an adjacency matrix and a sparse transition matrix stored on HDFS in CSV format. | 4 |
| | b) | If the HDFS file containing keys 1,5 are in block 1 and 2,3,4 are in block 2, then compute page rank for one iteration using Map Reduce. Show keys of mappers and reducers and how the keys are transferred between the 2 mappers and the reducer(Assume a single reducer is used. Assume initial page rank of all the pages is 1 and this is already available to the mapper.) | 4 |
| | c) | If you need to export the page rank obtained out to an SQL database after the computation, which tool would you use? | 2 |
| 4. | | Consider the following tables<br>PatientInfo<br>PatientID \| Phone# \| Date<br>FirstLevelContact<br>Date \| Phone# \| PatientID<br>that were generated by a ContactTracingApp for finding out the first level contacts made by a patient. PatientInfo has information about the id of the tested patient, their phone number and the date on which they tested positive for a disease while FirstLevelContact contains PatientId of the tested patient, the phone number of the primary contact and the date on which the contact was made. The data is stored as CSV files on HDFS and runs into a few GB each. | |
| | a) | Write MR pseudo code to identify the number of superspreaders, which is defined as the number of patients who have more than 20 first level contacts. Show intermediate key-value pairs. | 4 |
| | b) | How many map-reduce steps do you require to generate the output? | 2 |
| | d) | If the FirstLevelContact were stored on HBase by using PatientID as the key and with range partitioning for 4000 keys with 500keys per region, show how the data will be spread across different region servers. | 4 |
| 5. | | Consider a file that contains many lines and you have been given the following Apache Spark code to process this file<br>`val lines = sc.textfile("harrypotter.txt")`<br>`val l = lines.filter("Hagrid")`<br>`val m = l.map(s => s.length)`<br>`val mcache = m.cache()`<br>`val totalLen = mcache.reduce((a, b) => a+b)` | |
| | a) | Classify the following operations into transformations and actions in the code? Filter, map, cache and reduce. | 4 |
| | b) | Where will transformations and actions be computed? Spark Driver or Worker. Give a brief | 4 |

| | | | |
|---|---|---|---|
| | | explanation of where the operations will be executed and what is the trigger. | |
| | c) | What does this code do? | 2 |
| | | | |
| 6 | a) | What are narrow and wide dependencies in Spark? With an example demonstrate how you can convert a wide dependency of join to a narrow dependency. | 4 |
| | b) | Consider a file poem.txt with the following lines<br>*Where the mind is without fear and the head is held high;*<br>*Where knowledge is free;*<br>*Where the world has not been broken up into fragments by narrow domestic walls;*<br>*Where words comé out from the depth of truth;*<br>*Where tireless striving stretches its arms towards perfection;*<br>*Where the clear stream of reason has not lost its way into the dreary desert sand of dead habit;*<br>*Where the mind is led forward by thee into ever-widening thought and action*<br>*Into that heaven of freedom, my Father, let my country awake.*<br>Being processed by the following code<br>`val lines = sc.textfile("poem.txt")`<br>`val l = lines.flatmap(line=>line.split(" "))`<br><br>Will the flatmap result in a narrow or a wide dependency? What will the RDD contain for the first line of the file? | 4 |
| | c) | What is reducer size and replication rate with respect to the communication cost model of complexity? Give an example to illustrate | 2 |