UNIVERSITY OF PETROLEUM AND ENERGY STUDIES, DEHRADUN

End Semester Examination, Dec 2021

Course: Disk Based Processing **Semester: V**

Program: B.Tech (CSE-Big Data) **Time: 03 hrs**

Course Code: CSBD 3001 **Max. Marks: 100**

Section A

1. **Each question will carry 4 marks**

(5Qx 4M = 20 Marks)

S. No.	Question	CO
Q1	Mention what is the difference between an RDBMS and Hadoop using only five points?	**CO3**
Q2	Explain what does the conf.setMapper Class do?	**CO1**
Q3	Explain what happenes when Data node fails.	**CO1**
Q4	Mention what are the main configuration parameters that user need to specify to run MapReduce Job?	**CO2**
Q5	What is speculative execution in hadoop?	**CO3**

SECTION B

1. Each question will carry 10 marks
2. In Question 4 there are two parts containing two questions each. Attempt only one part.

(4Qx10M = 40 Marks)

Q1	Differentiate Reducer and Combiner in Hadoop MapReduce. (**10 Marks)**	**CO2**
Q2	Define HDFS? Explain in brief about the basic building blocks of Hadoop? **(4+6 Marks)**	**CO2**
Q3	What are the advantages of In memory architecture over disk based processing? Explain by listing the fundamental properties of both Disk based processing and In memory architecture. **(10 Marks)**	**CO3**
Q4	**Part 1** a) Define the role of combiner and partitioner in a map reduce application. **(5 Marks)** b) What are the various functions of name node? **(5 Marks)** **OR** **Part 2** a) List out the limitations of Hadoop V-1. Explain how Hadoop V-2 has addresses those limitations. **(5 Marks)** b) What is YARN and its key compenents? What is the Resource Manager in YARN **(3+2 Marks)**	**CO1**

SECTION C

1. Each Question carries 20 Marks.
2. Instruction: Write long answer.
3. In Question 2, attempt only one in 2(a) and 2(b)

(2Qx 20M= 40 Marks)

Q1	If 8TB is the available disk space per node (10 disks with 1 TB, 2 disk for operating system etc. were excluded.). Assuming initial data size is 600 TB. How will you estimate the number of data nodes (n)? **(20 Marks)** Consider following factors while your calculation • The actual size of data to store • At what pace the data will increase in the future • Replication factor plays an important role – default 3x replicas • Hardware machine overhead (OS, logs etc.) • Intermediate mapper and reducer data output on hard disk • Space utilization • Compression ratio	**CO4**
Q2	a) Suppose you have to develop an application platform for Amazon prime where real time streaming of data can done. Propose a solution for this task. **(20 Marks)** **OR** b) There is a YARN cluster in which the total amount of memory available is 40GB. There are two application queues, ApplicationA and ApplicationB. The queue of ApplicationA has 20 GB allocated, while that of ApplicationB has 8GB allocated. Each map task requires an allocation of 32GB. How will the fair scheduler assign the available memory resources under the DRF (Dominant Resource Finder) Scheduler? **(20 Marks)**	**CO4**

End