National University of Computer and Emerging Sciences, Lahore Campus

Course:	Big Data	Course Code:	CS636
Program:	MS(Computer Science)	Semester:	Fall 2018
Duration:	80 Minutes	Total Marks:	25
Paper Date:	22-Oct-16	Weight	
Section:	-	Page(s):	4
Exam:	Midterm		

Instruction/Notes: Solve the mid in the space provided. Rough sheet will not be checked.

Question 1: *[13 marks]* You are given a database of the "TheBlogs" website. Recently, the traffic to the website has increased a lot and the owner wants you to use Map Reduce framework for analyzing the data of the website and for performing some basic operations. The database consists of the following tables:

Table Name	Table Attributes (columns)	Table Size (or size of the file that store the table data)
USER	userID, userName, email, age, gender	10 GB
TOPIC	topicID, topicName, topicDetail topic table include the various subject on which a user can write a blog like computer science, mathematics, art etc	1MB
BLOG	blogId, blogName, userID, topicID here userID is a foreign key and indicate that the user who is the writer of the blog	20 GB
COMMENT	commentID, blogID, UserID, commentText This table keeps the information of the comments given by the various users on the different blogs.	2TB

The initial cluster that was setup consist of 1 Master Node and 5 Slave nodes. Each node is a commodity hardware with 16GB RAM. The hdfs block size is 128MB and hdfs data replication is set to 2. The size of hard disk attached to each node is 2TB.

a) Find the names of the users who have written blogs on atleast 10 different topics (the ouput should have userName, topicName and blogName). For this we need to join three tables USER, BLOG and TOPIC.

Specify the relational operator you need in addition to JOIN and WHY	

Specify you will perform two two-way joins or one three way-join. **WHY?** Also Specify the JOINS you will perform (reduce side (1-1, 1-M or M-N), sort-merge join, hash-join or in memory striped variant. And **WHY?**	
Specify the number of map-reduce jobs you will need for all the operations. And WHY? You can build query tree to explain.	

b) We wish to analyze the data and identify the topics that interest female users. As a first step we decided to list the topicID, blogName, and commentText of each comment given by the female user.

To do so you need to join three tables BLOG, COMMENT and USER.

Specify the relational operator you need in addition to	

JOIN and WHY	
Specify the JOINS you will perform (reduce side (1-1, 1-M or M-N), sort-merge join, hash-join or in memory striped variant. And WHY? Specify you will perform two two-way joins or one three way-join. Why?	
Specify the number of map-reduce jobs you will need for all the operations. And WHY? You can build query tree to explain.	

***Question 2**: [12 marks]*

Develop an **efficient** MapReduce program to find pair of users who have written at least one blog on the same topic. For this question you would only need the BLOG table. Your output would be pair of users who have written one or more blogs on the same topic

U1, U2
U1, U9
U2, U7

Hint: Convert the BLOG table to the form**: topic-Id u1, u2, u9**

that is topicId followed by the list of users who have written a blog on that topic. And then use pairs approach on this to get the desired result.