NAME _____ ROLLNUMBER _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Big Data**	**Course Code:**	**CS636**
Program:	**MS(Computer Science)**	**Semester:**	**Spring 2018**
Duration:	**180 Minutes**	**Total Marks:**	**45**
Paper Date:	**21-5-2018**	**Weight**	**40%**
Section:	**-**	**Page(s):**	**8**
Exam:	**Final Exam**		

Instruction/Notes:

1. Solve the exam in the space provided. **Rough sheet will not be checked.**
2. Don't use pencil.
3. All CODES should be properly commented

Question 1: *[1+4+5 marks]*

i. **What are the Broadcast variables?**

ii. What is the difference between **SPARK Driver, Master, Executor and Cluster Manager.**

iii. Write **Spark Code** that input a text file containing temperature readings of last 50 years and for each year output the total readings, the min, the max and the average readings. Write an efficient SPARK code that do not unnecessarily scan the data.

Question 3: *[5 marks]* Write **Spark Code that i**nput a graph in form of the adjacency list and find all pairs of nodes that directly link to each other.

Input:
A-> B, C B-> A, E C-> F E-> A, C F-> A, C, E
Output:
A, B C, F

Question 4**: **[10 marks] Write a Spark code to multiply two extremely huge sparse matrices. Each Matrix is saved in a separate file in form RowNo, ColumnNo and Value. You can assume that total number of rows and columns of each matrix are given as input. You have to save the output in MN.txt.

Sample Input

M.txt

```
1,1 1
1,2 5
1,3 7
2,1 2
2,2 3
2,3 1
```

N.txt

```
1,1 1
1,2 5
2,1 2
2,2 3
```

Hint: We have done matrix multiplication in Map Reduce. You can adopt similar logic for writing matrix multiplication in spark.

NAME _____ ROLLNUMBER _____

Question 5: [10 marks] **MAP REDUCE**

You are given a huge file of student records. We want you to group the users on basis of their year and performance and create a separate file for each group. ***The grouping would be done as follows:***

The number of groups to be formed will be given as input by the user (One can pass it as input argument at runtime along with the jar file). The number of years in each group will be determined as follows: **(MaxYear – MinYear)\ Number of groups**

Input File	**Working**
17-3021 Sara GPA=3.9 Age 22 16-4021 Ali GPA=2.1 Age 22 01-3121 Zara GPA=2.79 Age 22 04-4221 Ahmad GPA=2.31 Age 22 05-1121 Toba GPA=3.0 Age 22 16-4021 Farah GPA=2.51 Age 22 16-3021 Hamza GPA=3.19 Age 22 12-4021 Haseeb GPA=1.72 Age 22 09-3021 Sara GPA=3.7 Age 22 08-4021 Ali GPA=2.3 Age 22	**Max year in this File = 2017** **Min-year in this File = 2001** **Let say input for number of groups is 5** **(2017-2001)/5 = 3** So, each group will have students from three years starting from 2001. *Last group can accommodate the extra years.* In this case the output should be in **5 files** File1: Contains data of students from year 2001-2003 File2: Contains data of students from year 2004-2006 File3: Contains data of students from year 2007-2009 File4: Contains data of students from year 2010-2012 File5: Contains data of students from year 2013-2017 All these files will contain data sorted by the year and for each year sorted by the GPA.

Write efficient MapReduce algorithm for the above task. Do give the code of Combiner and etc. if you think they are required. No information regarding Max and Min value is given as input

Question 6: [10 marks] ***Write an <u>efficient Map Reduce Algorithm</u> for the following problem***

A new very dangerous computer virus is launched that is very difficult to detect. The virus can remain dormant for days before it corrupts the hard disk and delete all the important data.

The virus spread through emails. If a computer receives an email from an infected computer, then it has a very high chance of getting infected. Let say, the infected computer A sends an email to computer B. Then B might be infected and in turn if B send the email to C then C might also be infected.

You are given a huge log file of emails send from one computer to another. In-addition to this you are given a list L, a list of computers that are infected with the virus. You task is to find out all the computers that might be infected.

Input Log File	***List of Infected Computers***	***Final Output file of your Map Reduce should contain***
Source: A Destination: B Source: A Destination: C Source: B Destination: C Source: C Destination: D Source: E Destination: D Source: E Destination: F Source: D Destination: K Source: K Destination: L Source: E Destination: G	***A*** ***B***	***A, B, C, D, K, L***