Distributed architectures for big data processing and analytics

September 1, 2022

Student ID _____

First Name _____

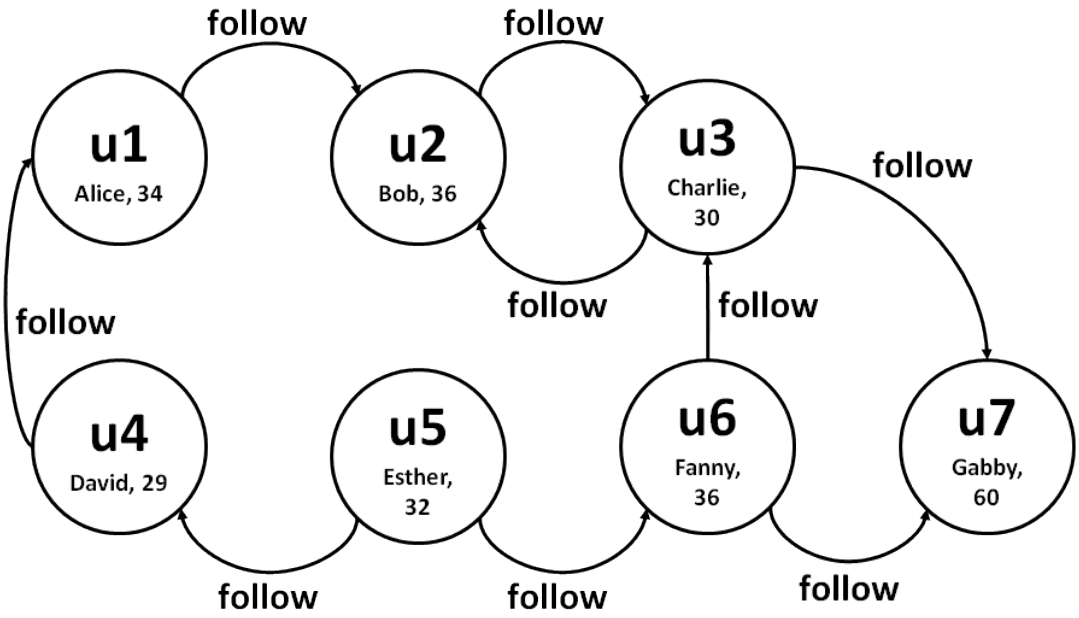
Last Name _____

The exam lasts **90 minutes**

Part I

Answer to the following questions. There is only one right answer for each question.

1. (2 points) Consider the following graph and suppose g is its instantiation in GraphFrame.



Suppose the following command is executed on g:

motifs = g.find("(v1)-[]->(v2); (v1)-[]->(v3); !(v2)-[]->(v3)")

Which one of the following statements is **true**?

a) One of the rows stored into the Dataframe motifs is

v1	v2	v3
[u4, David, 29]	[u1, Alice, 34]	[u5, Esther, 32]

b) One of the rows stored into the Dataframe motifs is

v1	v2	v3
[u5, Esther, 32]	[u4, David, 29]	[u6, Fanny, 36]

c) One of the rows stored into the Dataframe motifs is

v1	v2	v3
[u6, Fanny, 36]	[u3, Charlie, 30]	[u7, Gabby, 60]

d) One of the rows stored into the Dataframe motifs is

v1	v2	v3
[u3, Charlie, 30]	[u2, Bob, 36]	[u3, Charlie, 30]

2. (2 points) Consider the input HDFS folder *myFolder* that contains the following two files:

   - ProfilesItaly.txt
     - the text file ProfilesItaly.txt contains the following three lines
       Paolo,Rome
       Luca,Rome
       Giovanni,Rome
   - ProfilesFrance.txt
     - the text file ProfilesFrance.txt contains the following two lines
       Paolo,Paris
       Luis,Nice

   Suppose that you are using a Hadoop cluster that can potentially run up to 4 instances of the mapper class in parallel. Suppose the HDFS block size is 1024MB. Suppose to execute a MapReduce application for Hadoop that analyzes the content of *myFolder*. Suppose the map phase emits the following key-value pairs (the key part is a city while the value part is the length of the city name):

   (“Rome”, 4)
   (“Rome”, 4)
   (“Rome”, 4)
   (“Paris”, 5)
   (“Nice”, 4)

Suppose the number of instances of the reducer class is set to 2 and suppose the reduce method of the reducer class sums the values associated with each key and emits one pair (city, sum values) for each key. Suppose the following pairs are emitted, overall, by the reduce phase:

("Rome", 12)
("Paris", 5)
("Nice", 4)

Considering all the instances of the mapper class, overall, how many times is the ***map method*** invoked?

a) 2

b) 3

c) 4

d) 5

Part II

PoliDataCenters is an international organization that monitors the usage of GPUs around the world. PoliDataCenters is composed of several companies and computes a set of statistics about the data centers and the GPUs used in each data center. The analyses are performed by considering the following input data sets/files.

- Companies.txt
  - Companies.txt is a text file containing the list of companies involved in the PoliDataCenters organization. Each line of Companies.txt is associated with one company and contains its profile. The number of companies is more than 50.
  - Each line of Company.txt has the following format
    - CodC,CompanyName,NumberOfEmployees

      where *CodC* is the company identifier, *CompanyName* is its name, and *NumberOfEmployees* is its current number of employees.

    - For example, the following line

      C31,Politecnico di Torino,4508

      means that the company with id ***C1*** is called ***Politecnico di Torino*** and its current number of employees is ***4508***.

- DataCenters.txt
  - DataCenters.txt is a text file containing the list of data centers monitored by PoliDataCenters. One line for each data center is stored in DataCenters.txt. The number of managed data centers is more than 10000. Each data center is owned by one single company and each company can own several data centers (at least one).
  - Each line of DataCenters.txt has the following format
    - CodDC,CodC,Size,City,Country,Continent

      where *CodDC* is the data center identifier, *CodC* is the identifier of the company owning the data center, *Size* is the size of the data center in square meters, *City* is the city where the data center is located, *Country* is the country where the data center is located, and *Continent* is the continent where the data center is located (according to the seven-continent model).
    - For example, the following line

      DC21,C31,100,Turin,Italy,Europe

      means that the dataset center with id ***DC21*** is owned by the company with id ***C31***, its size is ***100*** square meters, and it is located in the city of ***Turin***, which is in ***Italy***, which is part of ***Europe***.

- GPUs.txt
  - GPUs.txt contains the information about the GPUs installed in each data center. Each GPU is installed in one and exactly one data center.
  - Each line of GPUs.txt has the following format
    - CodG,Type,CodDC
      where *CodG* is the identifier of the GPU, *Type* is its type, and *CodDC* is the identifier on the data center in which the GPU is installed.
    - For example, the following line

      G104, NVIDIA RTX A5000,DC21

      means that the GPU with id ***G104*** is an ***NVIDIA RTX A5000*** and is installed in the data center with id ***DC21***.

Exercise 1 – MapReduce and Hadoop (8 points)

Exercise 1.1

The managers of PoliDataCenters are interested in performing some analyses about the companies of the organization.

Design a single application, based on MapReduce and Hadoop, and write the corresponding Java code, to address the following point:

1. *Companies with many large data centers in both Europe and North America*. This application selects the identifiers of the companies with (i) at least 10 data centers in Europe, each one larger than 10000 square meters, and (ii) at least 10 data centers in North America, each one larger than 10000 square meters. Both constraints must be satisfied. Store in the output HDFS folder the identifiers of the selected companies (one CodC per output line).

Suppose that the input is DataCenters.txt and has been already set. Suppose that also the name of the output folder has been already set.

- **Write your code on your papers.**
- Write only the content of the Mapper and Reducer classes (map and reduce methods. setup and cleanup if needed). The content of the Driver must not be reported.
- Use the following two specific multiple-choice questions (**Exercises 1.2 and 1.3)** to specify the number of instances of the reducer class for each job.
- If your application is based on two jobs, specify which methods are associated with the first job and which are associated with the second job.
- If you need personalized classes, report for each of them:
  - the name of the class
  - attributes/fields of the class (data type and name)
  - personalized methods (if any), e.g., the content of the toString() method if you override it

Answer the following two questions to specify the number of jobs (one or two) and the number of instances of the reducer classes.

Exercise 1.2 - Number of instances of the reducer - Job 1
Select the number of instances of the reducer class of the first Job

- (a) 0
- (b) exactly 1
- (c) any number >=1 (i.e., the reduce phase can be parallelized)

Exercise 1.3 - Number of instances of the reducer - Job 2
Select the number of instances of the reducer class of the second Job

- (a) One single job is needed
- (b) 0
- (c) exactly 1
- (d) any number >=1 (i.e., the reduce phase can be parallelized)

Exercise 2 – Spark (19 points)

The managers of PoliDataCenters asked you to develop one single application to address all the analyses they are interested in. The application has five arguments: the input files Companies.txt, DataCenters.txt, and GPUs.txt, and two output folders “outPart1/” and “outPart2/”, which are associated with the outputs of the following points 1 and 2, respectively. Specifically, design a single application, based on Spark RDDs or Spark DataFrames, and write the corresponding Python code, to address the following points:

1. *Minimum data center size and maximum data center size of large companies with data centers in many European cities.* This first part of the application considers only the companies with at least 200 employees and data centers in at least 10 distinct European cities. Then, for each of those companies, it selects the size of its smallest European data center and the size of its largest European data center. Store the result in the first HDFS output folder (one output line for each company with at least 200 employees and data centers in at least 10 distinct European cities). Each output line has the following format:

   (CodC, CompanyName, Minimum European data center size, Maximum European data center size)

2. *Companies with the same number of NVIDIA RTX A5000 in all data centers.* This second part of the application considers all companies and all data centers. For each company, it computes the number of GPUs of type NVIDIA RTX A5000 in each data center of that company and selects that company only if in all its data centers the number of NVIDIA RTX A5000 GPUs is the same. Store in the second HDFS output folder the identifiers of the selected companies (one CodC per output line). Pay attention that if a company, in all its data centers, has zero NVIDIA RTX A5000 GPUs it must be selected.

Example Point 2

- *Toy example.* For the sake of simplicity, suppose that there are only three companies, which are identified by C1, C2, and C3. Suppose that
  - C1 owns two data centers (DC11 and DC12)
    - In DC11 there are 10 NVIDIA RTX A5000 GPUs
    - In DC12 there are 10 NVIDIA RTX A5000 GPUs
  - C2 owns three data centers (DC21, DC22, and DC23)
    - In DC21 there are 10 NVIDIA RTX A5000 GPUs
    - In DC22 there are 15 NVIDIA RTX A5000 GPUs
    - In DC23 there are 15 NVIDIA RTX A5000 GPUs
  - C3 owns two data centers (DC31 and DC32)
    - In DC31 there are 0 NVIDIA RTX A5000 GPUs
    - In DC32 there are 0 NVIDIA RTX A5000 GPUs

  In the case of this small running example, the second part of this application **selects C1** and **C3** and stores these two identifiers in the second output folder. **C2 is discarded** because the number of GPUs is not the same in all its data centers.

- **Write your code on your papers.**
- You do not need to write imports. Focus on the content of the main method.
- Suppose both **SparkContext sc** and **SparkSession ss** have been already set.
- Suppose the following variables have been already set:
  - compsPath='Companies.txt'
  - dataCentersPath='DataCenters.txt'
  - gpusPath= GPUs.txt'
  - output1='outPart1/'
  - output2='outPart2/'