Big Data: Architectures and Data Analytics

September 6, 2022

Student ID _____

First Name _____

Last Name _____

Part I

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

1. (2 points) Consider the following Spark application.

```
import ….;
public class SparkDriver {
    public static void main(String[] args) {
        // Create a configuration object and set the name of the application
        SparkConf conf = new SparkConf().setAppName("Spark Code");

        // Create a Spark Context object
        JavaSparkContext sc = new JavaSparkContext(conf);

        // Read a first input file
        JavaRDD<String> Temp1RDD = sc.textFile("Temperature1.txt");

        // Print on the standard output the number of elements of Temp1RDD
        System.out.println(Temp1RDD.count());

        // Read a second input file
        JavaRDD<String> Temp2RDD = sc.textFile("Temperature2.txt");

        // Print on the standard output the number of elements of Temp2RDD
        System.out.println(Temp2RDD.count());

        // Create an RDD that contains the intersection of Temp1RDD and
        // Temp2RDD
        JavaRDD<String> IntersectionRDD = Temp1RDD.intersection (Temp2RDD);

        // Print on the standard output the number of elements of IntersectionRDD
        System.out.println(IntersectionRDD.count());

        sc.close();
    }
}
```

Suppose the input files Temperature1.txt and Temperature2.txt are read from HDFS. Suppose this Spark application is executed only 1 time. Which one of the following statements is true?

a) This application reads the content of Temperature1.txt 1 time and the content of Temperature2.txt 1 time.
b) This application reads the content of Temperature1.txt 2 times and the content of Temperature2.txt 1 time.
c) This application reads the content of Temperature1.txt 2 times and the content of Temperature2.txt 2 times.
d) This application reads the content of Temperature1.txt 3 times and the content of Temperature2.txt 3 times.

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:
       Kevin,Rome
       Luca,Rome
       Candy,Naples
   - ProfilesFrance.txt
     - the text file ProfilesFrance.txt contains the following two lines:
       Paolo,Paris
       Kevin,Nice

Suppose that you are using a Hadoop cluster that can potentially run up to 2 instances of the mapper class in parallel. Suppose the HDFS block size is 2048MB. 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)
("Naples", 6)
("Paris", 5)
("Nice", 4)

Suppose the number of instances of the reducer class is set to 3 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", 8)
("Naples", 6)
("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

MonitorDataCenters is an international organization that monitors the power consumption of data centers around the world. MonitorDataCenters is funded by several companies and computes a set of statistics about the data centers of the funding companies. 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 funding the MonitorDataCenters organization. Each line of Companies.txt is associated with one company and contains its profile. The number of companies is more than 100.
  - Each line of Company.txt has the following format
    - CodC,CompanyName,Headquarters-Country

      where *CodC* is the company identifier, *CompanyName* is its name, and *Headquarters-Country* is the country in which the headquarters of the company is located.
    - For example, the following line

      C12,Databricks,United States of America

      means that the name of the company with id ***C12*** is ***Databricks***, and its headquarters is in the ***United States of America***.

- DataCenters.txt
  - DataCenters.txt is a text file containing the list of data centers monitored by MonitorDataCenters. 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. Each company owns several data centers (at least one).
  - Each line of DataCenters.txt has the following format
    - CodDC,CodC,City,Country,Continent

      where *CodDC* is the data center identifier, *CodC* is the identifier of the company owning the data center, *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,C12,Nice,France,Europe

  means that the dataset center with id ***DC21*** is owned by the company with id ***C12*** and it is located in the city of ***Nice***, which is in ***France***, which is part of ***Europe***.

- DailyPowerConsumption.txt
  - DailyPowerConsumption.txt contains the information about the daily power consumption of the data centers monitored by MonitorDataCenters.
  - Each line of DailyPowerConsumption.txt has the following format
    - CodDC,Date,kWh
      where *kWh* is the kilowatt-hours consumed by the data center identified by the id *CodDC* in the date *Date*.
      Each line of DailyPowerConsumption.txt is uniquely identified by the primary key (CodDC,Date), i.e., each combination (CodDC,Date) occurs at most once in DailyPowerConsumption.txt.
      The format of the date is YYYY/MM/DD.

    - For example, the following line

      DC21,2021/01/12,360

      means that the power consumption of data center ***DC21*** on ***January 12, 2021****,* was ***360*** *kWh*.

Exercise 1 – MapReduce and Hadoop (8 points)

Exercise 1.1

The managers of MonitorDataCenters are interested in performing some statistics.

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

1. *Data centers with an increasing number of dates with high power consumption (kWh>1000) in the last year.* The application considers only the lines of DailyPowerConsumption.txt associated with high power consumption (kWh greater than 1000) and selects the data centers for which the number of dates with high power consumption in the year 2021 is greater than the number of dates with high power consumption in the year 2020. Store the identifiers (CodDC) of the selected data centers in the output HDFS folder.

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

- **Write the 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
    - do not report the get and set methods. Suppose they are "automatically defined"

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 MonitorDataCenters 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 DailyPowerConsumption.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, and write the corresponding code, to address the following two points:

1. *Dates with high power consumption in many data centers.* The first part of this application selects the dates on which at least 90% of the data centers consumed at least 1000 kWh (at least 1000 kWh in each data center). Store the selected dates in the first HDFS output folder (one date per line).

2. *Continent(s) with the highest average power consumption per data center in the year 2021 and the highest number of data centers.* This second part of the application selects the continent(s) with (i) the highest average power consumption per data center in the year 2021 and (ii) the highest number of data centers. This second part of the application selects only the continent(s) satisfying both constraints. Given a continent, its average power consumption per data center in 2021 is the sum of the power consumption of all its data centers in 2021 divided by the number of data centers in that continent. In case of a tie, the application selects all the continents that satisfy both constraints. Store the selected continent(s) in the second HDFS output folder (one continent per output line).

   Note. Suppose all data centers had a power consumption greater than zero in 2021.

- **Write the code on your papers.**
- You do not need to report imports. Focus on the content of the main method.
- Suppose both JavaSparkContext sc and SparkSession ss have been already set.