Distributed architectures for big data processing and analytics

July 20, 2020

Student ID _____

First Name _____

Last Name _____

The exam lasts **2 hours**

Part I

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

1. (2 points) Consider the input HDFS folder *myFolder* that contains the following two files:

   - log2018.txt
     - log2018.txt contains the following three lines
       Paolo,Turin
       Luca,Rome
       Giovanni,Turin
   - log2019.txt
     - log2019.txt contains the following two lines
       Paolo,Turin
       Matteo,Rome

   Suppose that you are using a Hadoop cluster that can potentially run up to 10 instances of the mapper class in parallel. 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 string associated with the name of a country while the value part is always 1):

   ("Turin", 1)
   ("Rome", 1)
   ("Turin", 1)
   ("Turin",1)
   ("Rome,1)

   Suppose the reduce method of the reducer class sums the values associated with each invocation of the reduce method. Suppose the number of instances of the reducer class is set to 7. Consider all the 7 instances of reducer class.

   Overall, how many times is the reduce method invoked?

a) 10

b) 7

c) 5

d) 2

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

```
temperatureRDD = sc.textFile("temperatures.txt")

# Select the lines with temperature > 30°C
highTempRDD = temperatureRDD.filter(lambda line: float(line.split(",")[1])>30)

# Select the lines with temperature < -10°C
lowTempRDD = temperatureRDD.filter(lambda line: float(line.split(",")[1])<=-10)

# Print on the standard output of the driver the total number of input lines
print("Total bumber of lines: " + str(temperatureRDD.count()))

# Print on the standard output of the driver the number of lines with
# temperature > 30°C
print("Number of lines with temp. > 30°C: " + str(highTempRDD.count()))

# Print on the standard output of the driver the number of lines with
# temperature < -10°C
print("Number of lines with temp. < -10°C: " + str(lowTempRDD.count()))
```

Suppose the input file temperatures.txt is stored in an HDFS folder. Suppose to execute 1 time this Spark application. Which one of the following statements is true?

a) This application reads 1 time the content of temperatures.txt

b) This application reads 3 times the content of temperatures.txt

c) This application reads 5 times the content of temperatures.txt

d) This application reads 6 times the content of temperatures.txt

Part II

PoliMobile is an e-commerce company that sells smartphones around the world. The management of PoliMobile is interested in analyzing the purchases of their smartphones. The computed statistics are based on the following input data sets/files.

- SmartphoneModels.txt
  - SmartphoneModels.txt is a text file containing the catalog of available smartphone models (more than 1000 smartphone models). The file contains one line for each smartphone model.
  - Each line of SmartphoneModels.txt is related to one model of smartphone and has the following format
    - MID,ModelName,Brand
      where, *MID* is the smartphone model identifier, *Name* is its name and *Brand* is its brand.

    - For example, the line

      MID10,Moto G,Motorola

      means that the smartphone model ***Moto G*** (this is the name of the smartphone model) is identified by the code ***MID10*** and its brand is ***Motorola***.

- Purchases.txt
  - Purchases.txt is a text file containing the list of purchases of the last 20 years (more than 10000000 lines). Every time a customer buys a smartphone a new line is appended at the end of Purchases.txt.
  - Each line of Purchases.txt is related to one purchase and has the following format
    - MID,CustomerId,Date,Price

      where *CustomerId* is the identifier of the customer who bought a smartphone associated with model *MID* on date *Date*. The cost of the purchase is *Price*.

    - For example, the line

      MID10,Cust1,,2017/05/02,119

      means that customer ***Cust1*** bought a smartphone on May 2, 2017 and the model identifier of the purchased smartphone is MID10. The cost of the purchase is ***119***€

Exercise 1 – MapReduce and Hadoop (7 points)

The managers of PoliMobile are interested in performing some analyses about their smartphones.

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

1. *Brands with one single smartphone model*. The application selects the brands associated with one single smartphone model. Store in the output HDFS folder the selected brands and for each of them store also the identifier (MID) of the associated smartphone model. Each output line contains one pair Brand\t*t*MID, one line per selected brand (the separator is the tab symbol).

   For instance, suppose that the brand Motorola produces only one smartphone model, and specifically the one with model identifier MID10. In that case Motorola will be selected and the following line will be stored in the output folder:

   Motorola\t*t*MID10

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

- 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 next two specific multiple-choice questions 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:

  - 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 get and set methods. I suppose they are "automatically defined"

Exercise 1 - Number of instances of the reducer - Job 1 - MapReduce and Hadoop (0.5 points)

Select the number of instances of the reducer class of the first Job

(a) 0

(b) exactly 1

(c) any number >=1

Exercise 1 - Number of instances of the reducer - Job 2 - MapReduce and Hadoop (0.5 points)

Select the number of instances of the reducer class of the second Job

(a) One single job is needed for this MapReduce application

(b) 0

(c) exactly 1

(d) any number >=1

Exercise 2 – Spark and RDDs (19 points)

The managers of PoliMobile are interested in performing some analyses related to the sales of their smartphones.

The managers of PoliMobile asked you to develop one single application to address all the analyses they are interested in. The application has three arguments: the input file Purchases.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 code, to address the following points:

1. *Smartphone models with the highest income in 2019.* The application selects the model identifier (MID) of the smartphone model associated with the highest income in year 2019. For each smartphone model, the income in year 2019 of that smartphone model is given by the sum of the prices of the purchases in 2019 associated with that specific smartphone model. The application stores in the first HDFS output folder the MID(s) of the selected smartphone model(s). **Note that** you may have more than one smartphone model associated with the highest income in year 2019, i.e., you may have more than one line in the returned output. The output contains one line for each of the selected smartphone models (one MID per line).

2. *Most purchased smartphone model in at least two years.* Consider the purchases of the last ten years (i.e., from year 2010 to year 2019). The application selects the smartphone models that have been the most purchased smartphone model in at least two years (from year 2010 to year 2019). For each year, the most purchased smartphone model in that year is the one associated with the maximum number of purchases in the considered year (we recall you that each line of Purchases.txt is associated with one purchase). The application stores in the second HDFS output folder the identifiers (MIDs) of the selected smartphone models.

   Note that you may have more than one smartphone model associated with the maximum number of purchases in each year.

   Suppose sc (Spark Context) and spark (Spark Session) have been already set.