Big Data: Architectures and Data Analytics

September 19, 2016

Student ID _____

First Name _____

Last Name _____

The exam is **open book** and lasts **2 hours**.

Part I

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

1. (2 points) Consider the cache "mechanism" of Spark. Which one of the following statements is true?

   a) Caching an RDD is always useful

   b) An RDD that is used only one time in an application must always be cached

   c) Caching an RDD that is used multiple times in an application could improve the efficiency of the application (in terms of execution time).

   d) An RDD must never be cached by using the MEMORY_ONLY storage level

2. (2 points) Consider the HDFS file words.txt. The size of words.txt is 1060MB. Suppose that you are using an Hadoop cluster that can potentially run up to 5 mappers in parallel and suppose to execute the word count application, based on MapReduce, by specifying words.txt as input file. Which of the following values is a proper HDFS block size if you want to "force" Hadoop to run 5 mappers in parallel when you execute the word count application by specifying words.txt as input file?

   a) Block size: 256MB

   b) Block size: 530MB

   c) Block size: 1024MB

   d) Block size: 5300MB

Part II

PoliAgency is an environmental agency that monitors and analyzes pollutant concentrations by means of a network of sensors located in several cities of Italy. The analyses of interest are based on the following data sets/files.

- ReadingsPerMonitoringStations.txt
  - Readings_MonitoringStations.txt is a textual file containing the historical information about the PM10 and PM2.5 concentrations collected by the monitoring stations.
  - The sampling rate is 15 minutes (i.e., every 15 minutes each monitoring station collects the values of PM10 and PM2.5 and a new line for each monitoring station is inserted in Readings_MonitoringStations.txt)
  - Each line of the file has the following format
    - monitoringStationId,date,hour,minute,PM10value,PM2.5value

      where *monitoringStationId* is a monitoring station identifier, *PM10value* is the value of PM10 collected by station *monitoringStationId* at time *date-hour-minute*, and *PM2.5value* is the value of PM2.5 collected by station *monitoringStationId* at time *date-hour-minute*.

    - For example, the line

      station1,2016/03/07,17,15,29.2,15.1

      means that ***station1*** collected the following values on ***March 7, 2016*** at ***17:15***: ***PM10=29.2, PM2.5=15.1***

- MonitoringStations.txt
  - MonitoringStations.txt is a textual file containing the list of available monitoring stations with the associated characteristics
  - The file contains one single line for each monitoring station
  - Each line of the file has the following format
    - monitoringStationId,latitude,longitude,cityArea,city

      where *monitoringStationId* is the identifier of the station, *latitude* and *longitude* are the GPS coordinates of the station, area is the city area in which the station is located, and *city* is the name of the city in which the station is located.

    - For example, the line

      station1,45.4781N,9.2273E,AreaB,Milan

      means that the GPS location of ***station1*** is ***(45.4781N,9.2273E)***, and the station is located in the ***AreaB*** of ***Milan***.

Exercise 1 – MapReduce and Hadoop (10 points)

The managers of PoliAgency are interested in understanding how frequently PM10 is greater than PM2.5 in each station, considering only the readings associated with year 2013. A monitoring station is defined as a "***frequently high PM10 station***" if for that station the number of times PM10 was greater than PM2.5 is at least equal to 30 during year 2013 (i.e., the number of readings with PM10 greater than PM2.5 is at least equal to 30 for the considerer station in year 2013).

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

A. *Frequency of PM10 greater than PM2.5 for the "frequently high PM10 stations" in year 2013.* Specifically, considering only the reading associated with year 2013, the application must count for each station the number of times PM10 was greater than PM2.5 and store the result in an HDFS folder only for those stations with a number of times at least equal to 30. Each line of the output file has the format
monitorignStationId,number of times PM10 was greater than PM2.5

The name of the output folder is one argument of the application. The other argument is the path of the input file ReadingsPerMonitoringStations.txt.

Exercise 2 – Spark and RDDs (17 points)

The managers of PoliAgency are interested in identifying highly polluted stations. Specifically, a monitoring station is classified as a "***highly polluted station***" if the PM10 value is greater than a user provided threshold *PM10th_limit* more than 45 times in one year (i.e., 45 readings are characterized by a PM10 value greater than *PM10th_limit* in one year for the considered monitoring station). The analysis is based on the historical data stored in ReadingsPerMonitoringStations.txt.

Another analysis of interest is the identification of the "always polluted" stations. A station is defined as an "***always polluted station***" if the value of PM2.5 was always above a user provided threshold *PM2.5th_limit* in that station (based on the historical data stored in ReadingsPerMonitoringStations.txt).

The managers of PoliAgency asked you to develop an application to address the analyses they are interested in.

The inputs of the application are the file ReadingsPerMonitoringStations.txt, two output folders (associated with the outputs of points A and B of this exercise), and the thresholds

PM10th_limit and *PM2.5th_limit*. Inputs and outputs are specified as arguments of the application.

Specifically, design a single application, based on Spark and RDDs, and write the corresponding Java code, to address the following points:

A. *Select the list of highly polluted stations*. Specifically, the application must select the "*highly polluted stations*", based on the definitions reported above, considering only the readings associated with year 2015, and store the stationIds of the selected stations in an HDFS folder. The name of the output folder is one argument of the application.

B. *Select the list of always polluted stations*. Specifically, the application must select the *always polluted stations* based on the definition reported above, and store the stationIds of the selected stations in an HDFS folder. The name of the output folder is one argument of the application.