BDA1 - Spark - Exercises

In this set of exercises you will work exclusively with Spark. This means that in your programs, you only need to create the `SparkContext`.
In a number of exercises you will be asked to calculated temperature averages (daily and monthly).
These are not always computed according to the standard definition of 'average'. In this domain the daily average temperature is calculated by averaging the daily measured maximum and the daily measured minimum temperatures. The monthly average is calculated by averaging the daily maximums and minimums for that month. For example, to get the monthly average for October, take maximums and minimums for each day, sum them up and divide by 62 (which is the same as taking the daily averages, summing them up and divide by the number of days). [1]

Assignments

1) What are the lowest and highest temperatures measured each year for the period 1950-2014. Provide the lists sorted in the descending order with respect to the maximum temperature. In this exercise you will use the *temperature-readings.csv* file.
    a) Extend the program to include the station number (**not the station name**) where the maximum/minimum temperature was measured.
    b) (not for the SparkSQL lab) Write the non-parallelized program in Python to find the maximum temperatures for each year without using Spark. In this case you will run the program using:
    `python script.py`
    This program will read the local file (not from HDFS). The local file is available under `/nfshome/hadoop_examples/shared_data/temperatures-big.csv`.
    How does the runtime compare to the Spark version? Use logging (add the --conf spark.eventLog.enabled=true flag) to check the execution of the Spark program. Repeat the exercise, this time using *temperatures-big.csv* file available on hdfs. Explain the differences and try to reason why such runtimes were observed.

2) Count the number of readings for each month in the period of 1950-2014 which are higher than 10 degrees. Repeat the exercise, this time taking only distinct readings from each station. That is, if a station reported a reading above 10 degrees in some month, then it appears only once in the count for that month.
    In this exercise you will use the *temperature-readings.csv* file.
    The output should contain the following information:
    Year, month, count

[1] Note: In many countries in the world, the averages are calculated as discussed. However, in Sweden, daily and monthly averages are calculated using Ekholm-Modéns formula which in addition to minimum and maximum daily temperature also takes into account readings at specific timepoints, the month as well as the longitude of the station. For more information check (in Swedish):
http://www.smhi.se/kunskapsbanken/meteorologi/hur-beraknas-medeltemperatur-1.3923

3) Find the average monthly temperature for each available station in Sweden. Your result should include average temperature for each station for each month in the period of 1960-2014. Bear in mind that not every station has the readings for each month in this timeframe.
   In this exercise you will use the *temperature-readings.csv* file.
   The output should contain the following information:
   Year, month, station number, average monthly temperature

4) Provide a list of stations with their associated maximum measured temperatures and maximum measured daily precipitation. Show only those stations where the maximum temperature is between 25 and 30 degrees and maximum daily precipitation is between 100 mm and 200 mm.
   In this exercise you will use the *temperature-readings.csv* and *precipitation-readings.csv* files.
   The output should contain the following information:
   Station number, maximum measured temperature, maximum daily precipitation

5) Calculate the average monthly precipitation for the Östergotland region (list of stations is provided in the separate file) for the period 1993-2016. In order to do this, you will first need to calculate the total monthly precipitation for each station before calculating the monthly average (by averaging over stations).
   In this exercise you will use the *precipitation-readings.csv* and *stations-Ostergotland.csv* files. HINT (not for the SparkSQL lab): Avoid using joins here! *stations-Ostergotland.csv* is small and if distributed will cause a number of unnecessary shuffles when joined with precipitation RDD. If you distribute *precipitation-readings.csv* then either repartition your stations RDD to 1 partition or make use of the collect to acquire a python list and broadcast function to broadcast the list to all nodes.
   The output should contain the following information:
   Year, month, average monthly precipitation

6) Compare the average monthly temperature (find the difference) in the period 1950-2014 for all stations in Östergotland with long-term monthly averages in the period of 1950-1980. Make a plot of your results.
   HINT: The first step is to find the monthly averages for each station. Then, you can average over all stations to acquire the average temperature for a specific year and month. This RDD/Data Frame can be used to compute the long-term average by averaging over all the years in the interval.
   The output should contain the following information:
   Year, month, difference