Jake Mathura
COMP 590

Spark Program Instructions

I made two spark programs, one finds the absolute minimum by year, and the other finds the absolute maximum temperature by year. For both of these programs, the input is going to be a small sample of global land temperature data. The input file for both of these programs is named Spark_Sample_Input and the output file for both programs will be named Spark_Sample_Output.

Instructions:

1. First, we need to place our Sample_Input_File in the input hdfs directory using the following command:
    a. hdfs dfs -put Spark_Sample_Input /user/jakemath/input
2. Next, we will run spark_javac.sh using the next command:
    a. ./spark_javac.sh OverallMinLandTemp
    b. For the OverallMaxLandTemp, simply replace OverallMinLandTemp with OverallMaxLandTemp
3. Next, we must send the application to the cluster by running spark_run_cluster.sh:
    a. ./spark_run_cluster.sh OverallMinLandTemp /user/jakemath/input/Spark_Sample_Input /user/jakemath/output
    b. Similar to step 2, just replace the program names for the Max program
4. Next, we need to merge our output using the following command:
    a. hadoop fs -getmerge /user/jakemath/output Spark_Sample_Output
5. Finally, we need to sort the output in order for it to be valuable:
    a. sort Spark_Sample_Output > Sorted_Spark_Sample_Output

Important Note:

- In the program, we are unable to simply skip when the min/max temperature is not provided for a certain year (1750-1850), so those values are set as the Integer.MAX_VAL for the min program and Integer.MIN_VAL for the max program. These values are then filtered in post processing steps.