[Fall-2021]

HW4

Deadline: November 18th 11:59PM ET (8:59PM PT)

General Description:

In this homework, you are tasked to create a Hadoop MapReduce application to find the maximum temperature **in every day** of the years 1901 and 1902 from the NCDC weather records.

Your application should be deployed on Google Cloud Platform.

The data are stored using a line-oriented ASCII format, in which each line is a record. The yearMonthDay information is located between the 15th and 23rd byte of each record.

```
006701199099999 19500515 07004...9999999N9+00001+99999999999...
004301199099999 19500515 12004...9999999N9+00221+99999999999...
004301199099999 19500515 18004...9999999N9-00111+99999999999...
004301265099999 19490324 12004...0500001N9+01111+99999999999...
004301265099999 19490324 18004...0500001N9+00781+99999999999...
```

yearMonthDay (15-23) temperature (87-92) quality (93)

Your application should filter out invalid records. Invalid records are defined as:

- Records with temperature value of 9999 (missing reading).
- Records with quality flag other than: 0, 1, 4, 5 or 9

Your final key-value results should consist of yearMonthDay and the maximum temperature for that day. For example, (you may get different temperature readings)

```
19021008        72
19021015        67
19021022        61
19021029        78
```

You will find the dataset at https://github.com/mohamedfarag/ncdc-data

Your application should read the input from HDFS and store the output to HDFS.

When your application completes, merge all the results to one file and store it on the local cluster.

Add a copy of the results file to your GitHub repository

Submission Guidelines:

- Post URL for your GitHub repository to Canvas. Make sure to keep your GitHub repository public.
- Your GitHub repository should contain the following
    1. Copy of Your mapper.py code (or equivalent in another programming language) (25% of total grade)
    2. Copy of Your reducer.py code (or equivalent in another programming language) (25% of total grade)
    3. Screenshot of the execution of Hadoop MapReduce Job in the terminal (25% of total grade)
    4. Copy of your output file (after merging) containing the results (25% of total grade)

Don’t forget to disable billing after you finish using GCP

Extra Credit:

- (+20%) Build GUI to upload the two data files from your local machine to GCP bucket automatically without having
    - To provide this functionality, submit a video recording showing your bucket before the upload, the upload functionality from your application, and after the upload from your GUI.

Common Penalties:

- Your GitHub repository is not public: 100% reduction (won’t be graded)
- Late submissions on Canvas or GitHub: 100% reduction (won't be graded)