19COC105 Cloud Computing Coursework

Module leader: Dr Posco Tso
2pm on Monday 25 Nov 2019 – Individual work only

Coursework Specification

You are a newly hired Hadoop MapReduce professional and you are now asked by your company Loughborough Technology Limited to write a MapReduce program to process a dataset from National Centers for Environmental Information (NCEI) dataset archive[1].

The data is organised in CSV formatted files and an extract, which lists the location ID, date (yyyymmdd), element, value and measurement flag, is shown below. The dataset may contain more than 5 columns, you should refer to the dataset's *readme*[2] document for the detailed interpretation of each column.

US1ILMG0006	20180101	PRCP	0	N
US1ILMG0006	20180101	SNOW	0	N
ASN00015643	20180101	TMAX	401	a
ASN00015643	20180101	TMIN	234	a

Your tasks is to plot (with the tool of your choice) the difference between the maximum and the minimum temperature in Oxford (UK000056225) and Waddinton (UK000003377) for each day in 2018 in degrees Celsius. For the ease of plotting the data, you are asked to write a Hadoop MapReduce program (in the language of your choice) to generate a one-column CSV file, for each location, which contains one value for each day from *2018.csv* in the NCEI dataset. Your program should take input from and produce output to Google Storage.

Submission

Coursework type: Individual work only. Electronic submissions must be uploaded to module Learn page by 2pm on Monday 25 Nov 2019

1. Source code: You only need to submit the code written by you. Your code should be clearly commented and readable. Upload a **zipped document** to Learn.
2. A short report (max. 4 pages, min. 11 point font size) in **PDF format**: Every thing, including references, if any, counts towards the page limit, and your report should contain the following:
   - List of your Google Storage path(s) to the input and output directories in `gs://path_to_io` format. Make sure they are read-only accessible to the public.
   - Describe the challenges you have faced during the development of your MapReduce algorithm and how you have overcome these challenges.
   - Provide the plots of your result.
   - Discuss how your MapReduce program's execution time can be improved.

[1] NCEI Datasets `https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/`

[2] NCEI Datasets Readme `https://www1.ncdc.noaa.gov/pub/data/ghcn/daily/readme.txt`

Assessment

- Implementation (Source code, 60%)
  - Your implementation will be marked against two criteria: Efficiency of your algorithm (40%), Comments/Readability (20%).
- Report (40%)
  - Google storage paths that are read-only accessible to the public (10%).
  - Challenges and solutions of MapReduce algorithm development (10%).
  - The plots of the results produced by your MapReduce program (10%).
  - The discussion of potential improvement of your MapReduce program execution time (10%).