The Course Project

**Plagiarism and cheating will not be tolerated for this individual and independent project assignment. All submissions will undergo a thorough plagiarism check. Sharing your work with other students, copying others' work, or using AI tools such as ChatGPT is strictly prohibited. Any instances of plagiarism or cheating will result in a grade of 0 for the entire project submission.**

The course project consists of four parts. Parts 1 and 2 require the use of the project data zip file, while parts 3 and 4 require the use of a text file named “skyceiling.txt”.

Part 1:

The first part is to develop a Mapper and Reducer application to calculate the *range* (the difference between max and min values) of *sky ceiling height* (meters) for *each observation month* from NCDC records (note: 99999 indicates missing value, and [01459] indicate good quality value).

Part 2:

The second part is to develop a python application that can be implemented in PySpark to calculate the *average visibility distance* (meters) for *each USAF weather station ID* from NCDC records (note: 999999 indicates missing value, and [01459] indicate good quality value).

Part 3:

The third part is to load the text file into Pig and get the range of sky ceiling height for each USAF weather station ID.

Part 4:

The fourth part is to load the text file into Hive and get the average sky ceiling height for each USAF weather station ID.

You need to turn in:

**1)** Part 1:

a. *if you are using JAVA to develop the Mapper and Reducer applications:* the three java files (mapper, reducer and main);

b. *if you are using Hadoop streaming jar and developing two python programs (mapper python file and reducer python file)*: the two python files (mapper and reducer);

c. *if you are using mrjob library and developing one python program with two functions:* the python file (with the mapper and reducer functions);

2) Part 2: the python program you developed;

**3)** the commands from converting java files into a Jar file to running the Jar file in Hadoop, or the commands to execute the python files in Hadoop and in Spark;

**4)** the step by step commands and screenshots of solutions from all the parts;

The original dataset for this project is available on Blackboard.