The Course Project

The course project includes 4 parts.

Part 1:

The first part is to develop a Mapper and Reducer application to calculate 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 2:

The second part is to develop a Mapper and Reducer application to retrieve USAF weather station ID and sky ceiling height (meters) from NCDC records (note: 99999 indicates missing value, and [01459] indicate good quality value) and then write the USAF weather station ID and sky ceiling height data into a text file.

Part 3:

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

You need to turn in:

**1)** a. *if you are using JAVA to develop the Mapper and Reducer applications:* the three java files (mapper, reducer and main) for both part 1 and part 2;

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) for both part 1 and part 2;

c. *if you are using mrjob library and developing one python program with two functions:* the python file (with the mapper and reducer functions) for both part 1 and part 2;

**2)** 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;

**3)** the screenshot of output from part 1;

**4)** the text file including USAF weather station ID and sky ceiling height data created by you;

**5)** the screenshot of the text file being created;

**6)** the Pig commands and the screenshot of the final Pig output showing the average sky ceiling height for each USAF weather station ID;

**7)** the Hive commands and the screenshot of the final Hive output showing the highest and lowest sky ceiling height for each USAF weather station ID.

The original dataset for this project is available on Blackboard.