The Course Project

The course project includes 3 parts. The first part is to develop a Python application to retrieve Year and Temperature from original NCDC records (i.e., the dataset we are using for this class) and then write the Year and Temperature data into a text file. The second part is to load the text file into Pig and get the highest and lowest temperatures for each year. The third part is to load the text file into Hive and get the average temperature for each year.

You need to turn in 1) Python files (mapper and reducer), 2) the commands for executing the Python application in Hadoop, 3) the text file including Year and Temperature data created by you, 4) the screenshot of the text file being created, 5) the screenshot of the final Pig output showing the year and the highest and lowest temperatures, and 6) the screenshot of the final Hive output showing average the year and temperature.

The original dataset for this project is available on Canvas.